Zero the Hero: Upgrading Targeted Surveys to Case-Control Designs

Katerina Vrablikova, University of Mannheim

Richard Traunmüller, Goethe University Frankfurt

A very first draft of a paper prepared for the ECPR General Conference, Université de Montréal 26 - 29 August 2015.

Abstract

Targeted surveys of participants are increasingly popular in research on political activism. While this strategy solves problems of national surveys by effectively reaching rare types of activists, they also suffer a major drawback: since they only include participants and no "zeros", i.e. non-participants, they are ill suited to test the determinants of activism. We propose "case-control designs", widely used in epidemiology, to expand targeted surveys to a more powerful design by supplementing the cases with „zeros", i.e. eligible controls that allow us to test causal effects. In this regard, the case-control design can smoothly upgrade more qualitatively oriented studies and connect them to quantitative approaches, taking advantage of the strength of both. Using the example of protest participation at a recent anti-austerity demonstration, we illustrate the necessary steps of our approach: a) the definition and sampling of participants, b) the selection of zeros and c) the analysis of case-control data.

Keywords: Contentious Politics, Methods, Political Methodology, Political Participation, Social Movements

Introduction

Causal explanation is one of the most important goals in social sciences. Together with examination of causal mechanisms, testing causal effects (i.e. that A caused B) is the key fundament of causal inference (King, Keohane, and Verba 1994; Brady and Collier 2010; George and Bennett 2005). For scholars of social movements and contentious politics it is not that easy to design their studies to test causal effects because often times we cannot fully rely on classical procedures advised in sociology or political science textbooks. The challenge lies in the fact that a lot of phenomena we aim to study, such as movements, revolutions, participants at demonstrations or political violence, have a character of rare cases. Since there are too few of rare cases or they are extremely diluted in source populations, they are not reachable by conventional research designs that are usually used for testing causal effects, such as quantitative studies based on random sampling or a comparative method of a few cases.

So far social movement and contentious politics studies have not used research designs that would allow full-fledged testing of causal effects in rare cases. This article offers a solution by following the strategy of so called case-control designs. Case-control designs have been widely used already for decades in epidemiology (Armenian 2009; Schlesselman 1982) and have been applied in econometrics under the label of choice or response-based sampling (Manski 1995; Manski and Lerman 1977). Social science in general and social movement literature in particular have overlooked this powerful and efficient research design. Most of the few existing social science literature on case-control designs are rather technically oriented and focus on data analysis and estimation (King and Zeng 2001; King and Zeng 2004; Lacy 1997). The goal of this article is to provide the social movement community with an accessible guide on how to practically conduct a case-control study. The article does not aim to bring a new insights into case-study designs or provide readers with detailed statistical background of the data analysis. The article first explains why standardly used designs are not efficient for studying rare cases and explains the logic of a case-control design. Then using a model example of a case-control study that is interested in why people protest at anti-austerity demonstrations the article shows three stages of the case-control study: specification of a problem and sampling of cases, sampling of controls, and analysis of case-control design data.

Provided the researchers are interested in testing causal effects, case-control design is preferable solution for rare cases to other strategies available among the causal effect approaches (Armingeon 2002). Specifically, the main advantages are: 1. Well-designed case-control study allows testing more general theories that are not restricted only to sub-populations (which are usually not as much theoretically interesting as more general comparative framework). 2. Case-study studies are in essence mixed method designs that combine advantages of deep examination of interesting cases (rich descriptive analysis, process tracing, and theory development) with quantitative testing of causal effects. 3. The practical employment of a case-control study is easy and low-cost. Since social movement researchers usually already study their positive cases in much detail, the only additional work needed is to supplement the existing data with a few standardized information on negative cases.

We do not argue that case-control design is an ideal method without any negatives. Its quantitative part shares most of the shortcomings usually related to quantitative methods (limited option to study causal mechanisms and equifinality, higher chance to arrive in validity problems, and often inability to capture time development because of the cross-sectional character of the data). Case-control design also has some specific problems (Schlesselman 1982; King, Keohane, and Verba 1994, 141). The biggest challenge is to appropriately sample controls. If this is done incorrectly, the causal inference is biased. Also the design by its nature does not allow descriptive inference on the distribution of the dependent variable in the general population. The limitation of data generated by case-control designs is also the fact that the dataset cannot be used for other analytic purposes but the analysis of the one dependent variable, for which it was collected. However, being aware of these limits, we think that the case-control design can be an effective and simple way to improve on the current practice in the contentious politics and social movement literature.

Causal effects and rare cases

In the view of the causal effect approach the social scientific explanation rests on the counterfactual notion of causality. It considers a factor to be causal if a case is exposed to a treatment and shows an outcome, which the same case would not show if it did not receive the treatment (King, Keohane, and Verba 1994, 70). Observational studies try to approximate this logic by controlled comparison of real-world cases that experienced the treatment with different cases that did not happen to experience it. Ordinarily, random sampling in quantitative studies or selection of a few cases on independent variables in a comparative method are used to carry out this controlled comparison and test causal effects (Lijphart 1971; King, Keohane, and Verba 1994).

These standardly used designs cannot be effectively applied to study causes of rare phenomena, such as revolutions, protestors or political violence. Being a rare case means that only a small number of units display positive outcomes while a much larger number of them displays the negative outcome (King and Zeng 2001; Armenian 2009). Skockpol, for instance, identified six social revolutions within the whole world's history (Skocpol 1979). Similarly, there were only around 60.000 people in Oslo taking part at the demonstration against the war in Iraq on 15 February 2003 diluted in 4,5 million of Norwegian population (i.e. around 1.3 percent) (Verhulst 2010, 16–17). In this situation, the standard procedures of random sampling within a population or selecting a few cases on their independent variables as in the classical comparative method is not effective. The reason is that when positive cases are really rare, it is very likely that the final sample will not include any or too few of them.

Consider, for example, participants at illegal demonstrations. Nationally representative surveys, which are standardly used to study individual political participation, show that not more than one percent of publics in Western democracies declare participation in illegal protest activities (Teorell, Torcal, and Montero 2007, 339). National random samples usually include around 1000 cases to represent the whole population. As there is only one percent of protestors at illegal demonstrations only around ten cases of participants in illegal activities are covered by the nationally representative survey data, which is too few for effective analysis. Similarly, using the whole population of all units (both cases and controls) is in the case of rare events not a feasible solution. For instance, collecting all necessary data on all country-year units all over the world in the last 2000 years to study determinants of ten revolutions makes data collection and data management difficult and unnecessarily expensive.

Other designs that most social movement studies use are not very efficient either and have some important limitations. A lot of studies use "no-variance" designs that include data only on positive cases (such as case studies of existing movements, successful revolutions or protestors at demonstrations) and lack information on negative cases (Porta 2014; Flam 2001; Jasper and Poulsen 1995; Walgrave and Rucht 2010). While only-positive cases studies are undoubtedly excellent for a number of research goals[1], such designs usually do not provide a strong test of causal effects because they lack the control/negative cases. Congruence method, which tests whether a particular value of the independent variable in one case corresponds to the value of a dependent variable as expected by the theory, might be good at eliminating false hypotheses, however, usually it is not very strong to disentangle alternative explanations (George and Bennett 2005, chapter 9; Van Evera 1997, 31–32). Similarly, only-positive cases design can well test necessary condition, but it cannot test sufficient conditions (Brady and Collier 2010, 146; Seawright 2002). Put bluntly, showing that most of the revolutions happened in countries that faced external threat does not prove that external threat is a (sufficient) cause of revolution. The point is that even states not experiencing revolutions can be facing external threats or that other factors are, are in addition to external threat, needed to induce revolution.

When examining causal effects, the lack of variation on the dependent variable (i.e. the lack of zeros) can hardly be compensated by other research strategies. In order to demonstrate a causal effect, one usually needs to show both, that the positive cases experienced the cause

[1] In addition to rich description of particular cases, these studies are very strong particularly when used for theory development or reformulation. They are especially useful for making causal inference based on testing theoretically specified causal mechanisms. Only-positive cases designs can be also used to test necessary conditions or can be used to examine anomalies. Under specific conditions (extreme and rare values), they also can be used for testing causal effects (George and Bennett 2005; Van Evera 1997).

and that the negative cases did not experience it. Therefore, the core message is that studying only positive cases is usually not enough to establish (controlled) causal effects.[2]

Although some social movement studies use variation designs that include cases and controls, they usually heavily restrict their sour population and use sub-samples (McAdam 1986; Corrigall-Brown et al. 2009). However, this has has important consequences for the results. For instance, McAdam's study (1986) on risk activism compares participants of the Freedom summer ride with people, who applied to the program but did not show up. The problem is that application for the Freedom summer ride program is very likely to correlate with a number of important predictors of political action: ideological affiliation, political interest, political efficacy, individual resources etc.[3] These factors are kept constant or their variation is restricted by the design. Because of that the McAdam's study (1986: 64) cannot answer the general question mentioned by in the introduction: "Why does one individual get involved while another remains inactive?" It only explains why people, who have already applied to take part in a risk activism, actually participate in the freedom summer ride. This is much less general and less theoretically interesting question.[4] Also, though the study claims that social networks play the most important role for participation in risk activism (McAdam 1986), it cannot really prove it as other explanations are underestimated because of the restricted design.

While in theory the necessity of having nonbiased samples representing the general population (not restricted to sub-populations) seems to be simple and clear, it is much more challenging to find an effective way how to implement it in practical research of rare cases. In the following we introduce the case-control design as a simple way to deal with these challenges and to improve the current research practice in social movements and contentious politics that enables valid and efficient analysis of causal effects.

Case control research design

Case-control design can be defined as "a comparison of a group of persons with a certain outcome or condition with another group of persons who do not have that outcome or condition. The comparison is done for a number of determinants and potential exposures" (Armenian 2009, 19–20). The primary purpose of case-control design is to assure that rare cases, which are hard to reach and analyse with other designs, will be well covered by this design and then not rare anymore.

The crucial difference to classical research design strategies is the sampling on the dependent variable (King, Keohane, and Verba 1994, 141). The selection of units proceeds in two

[2] Trying to overcome the lack of negative cases in the research design by generation of variation among positive cases does not solve the problem either. For example, some studies might compare existing movements across countries, examine over-time variation in the amount of protest during a revolution or explain the number of demonstrations participants at the demonstration attended in the past. However, the questions these designs can answer are not the question of why movements emerge, why protest emerges, or why people protest, but different questions that address the variation covered by the studies: why movements differ, why protest fluctuates during revolution, why protestors differ in the number of demonstrations they attend.

[3] Another problem is the fact that the extent of the restriction is unknown as the selection on the dependent variable is not based on some theoretically relevant and crucial variable for which this restriction could be theoretically justified, see below more on this issue.

[4] Given the fact that even applicants for the Freedom summer ride are a tiny and special group of people, it might seem more theoretically valuable and informative to know, why people apply for the risk activism. Application seems to be much more important threshold to overcome when moving from category of non-participant to participant in risk activism than making the decision to participate among applicants.

separate procedures, which are also often related to different data sources: it samples separately the rare (positive) cases and separately the controls (negative cases). Such sampling creates outcomes that are binary variables having values one for the (rare) cases and zero for the controls. The crucial condition that needs to be met for the case-control design (and the selection on the dependent variable) to allow non-biased and effective causal inference is that both, cases and controls, need to come from the same source population. This is the population, which is at risk of the development of the outcome, i.e. population that had a chance for the exposure (Armenian 2009; Schlesselman 1982, 44). The implication of this condition is that one of the outcome values (case or controls) cannot be restricted/biased (i.e. cannot correlate with potential independent variables) while the other outcome value does not have this restriction/bias.

The process of conducting a case-control study can be distinguished into three steps (Armenian 2009): 1] Problem specification and sampling of cases, 2] selection of controls and 3] analysis of case-control data. We will demonstrate all of these three steps using the example of study aiming to infer determinants of participation at anti-austerity protest (i.e. to test causal effects). The research question our exercise asks is: Why do people participate in anti-austerity protest? From the perspective of the national population, protestors at anti-austerity demonstrations are in general rare cases not going above a few percent in most of the countries. As a case representing an anti-austerity demonstration we pick an anti-austerity march called “Stop the Government” that was organized on17th December 2012 in Prague by a coalition including major trade unions and leftist anti-austerity groups and had a turnout of around ten thousand people.

1. Problem specification and sampling of cases

Scholars of contentious politics and social movements typically start with interesting positive cases: the mobilization of the Occupy movement in Western democracies, historical revolutions or - as in our example - the participants in an anti-austerity demonstration. Starting with interesting cases is similar to how researchers in epidemiology proceed, whose research is very much problem driven. They start their inquiry with the detection of a potentially few cases suffering from a particular disease.

Note that at this stage the case-control studies in fact begin as (qualitative) case studies: The research is inductively driven and consists of a deep exploration of specific cases (Armenian 2009). This only-positive cases stage allows social movement scholars carrying out field research that helps develop theories on conceptualization of the cases, generate explanatory theories and allows specification and testing of causal processes and mechanisms. The crucial advantage is that vast majority of social movement and contentious politics studies already has this part of the case-control design, hence the already gathered data and research work done can be effectively used further and extended to case-control design.

The main task in the first phase of extending these cases studies into case-control studies is to specify the problem under study not as a case but as a variable and determine the units of analysis. This process of “moving from cases to variables” has been a basic rule of comparative approach (Przeworski and Teune 1970). When interested in a question on why women’s movements mobilize researchers should switch from a perspective of a movement as a case and start using the perspective of a variable, i.e. try to explain existence, which is the positive outcome of a variable, versus non-existence of the women’s movement, which is the negative outcome of the variable.

“Turning the case into a variable” concerns the unit on which this variable should be measured and thus, ultimately, the population under study. Just as in epidemiology it is helpful to think of the outcomes of the dependent variable as ‘incidents’ or ‘events’, i.e. things that happen to certain entities within a certain period of time they are observed or that are ‘at risk’ of experiencing it. In other words, it is useful to define the units of analysis in terms of entity by time. In our example of protestors at the anti-austerity demonstration, the most straightforward unit of analysis is individual people at the time of the demonstration. Specifically, the outcome variable that we want to explain is the difference between people who took part at the demonstration on this day (rare positive cases) and those who did not (negative/control cases).

Other examples of research objects in social movements, however, are not that straightforward. A lot of options are possible and selection of units can be a challenge. For instance, researchers interested in explaining the occurrence of women’s movements have a number of options: short term or long term differences among times when there was a women’s movement and when was not in one geographical context, differences across geographical contexts, such as countries or regions where there is movement and where there is not, or a combination of both, such as regions in years.

Importantly, the decision on the units of analysis, i.e. among what units one intends to explain the difference, is a conceptual task, dependent on our research questions and theories we aim to test (Brady and Collier 2010, 51). For instance, if one asks “Why did some countries experience mobilization of women’s movement?” or want to test explanation based on political opportunity structure argument about varied effect of the state political institutions, which differ mainly across countries but not that much over time, then the suitable unit of analysis is a country. If the question is “Why did the women’s movement appear in the specific time?”, the units of analysis are probably years. Notice that although the examination begins with more or less the same positive case – the existing women’s movement, the different specification of units is related to different research questions and will reveal different causal effects.

Having defined the unit of analysis (citizens, countries or years), the next question is what the source population of these units is. The decision on the source population is, again, a theoretical decision depending on our particular research question and on how general our findings should be. Most often the advice of how to achieve unbiased inference is to specify the relevant target population as a whole, i.e. to specify and sample of cases and controls so that they are not correlated with any potential causes of the outcome. However, there are some instances when focus on more restricted sub-population is reasonable (Armenian 2009). For instance, a qualitative study of terrorist attacks could show that terrorism in democratic and non-democratic regimes are considerably distinct phenomena and probably have different sources. In this situation, it is plausible to divide existing terrorist attacks into two sub-populations (revolutions in non/democratic countries) and proceed the two studies separately. Another example are situations when some independent variables are strongly correlated with the outcome so that it is hard to detect smaller nuances in effects of other factors or when there is some crucial subgroup of exposure for which we want to explore the exact mechanism in more detail (Armenian 2009, 42; King, Keohane, and Verba 1994). In these situations, both selection on the dependent variable and selection on the explanatory variable is possible.

In our case, it is reasonable to define our target population as the Czech population at the time around the demonstration. National populations are the traditional target population in comparative politics studies. We can expect that for instance Americans did not have the same chance to experience the outcome, i.e. to be exposed to the causes, such as being recruited and to have developed attitudes to the topic. Also not that many foreigners probably attended the demonstration hence the coverage of national variations would not add that much value.

What positive cases should be included into the analysis? Recall that the selection of cases should well represent the population of cases. In principle, a non-biased sample either requires taking all cases or some random selection procedure. In practice, researchers of social movements will probably use the first strategy and take all existing cases since the incidence of phenomena of interest is rare in absolute terms (e.g. only ten revolutions in the history).[5]

Not all rare events and cases in social movement studies are rare in terms of absolute numbers. Though protestors at a particular demonstration are rare in relation to the general population, their absolute number can be quite high and count up to tens of thousands of participants. In this situation gathering data on all cases would not be technically possible or at least not cost-effective. A solution advised by quantitative scholars is random sampling.

This is the strategy we use in our example as there were around ten thousand people attending the Stop the Government march. We applied random sampling to select our (positive) cases of protestors at the anti-austerity march. The data gathering followed a standardized sampling procedure of a protest survey method (Walgrave and Verhulst 2011) and was done within a research project "Caught in the Act of Protest: Contextualizing Contestation" (CCC, (van Stekelenburg et al. 2012). Protest surveying lies in systematic random sampling of protestors during the protest event and distribution of a pre-paid postal questionnaires. The sampling of respondents was done by two pointers, who started at the beginning of the march moving backwards on the two sides of the march. The pointers followed systematic sampling procedure to selecting respondents from the half of the march on their side (both pointers counted every second row and selected two respondents from each row based on the initial estimate of the size of the march of 8000 participants) and sent interviewers to approach the selected individual. There were teams of five interviewers with each pointer. Interviewers distributed the core pre-paid postal survey questionnaire to respondents, who agreed to take the survey booklet with them, fill out and send back (refusal rate nine percent). Since our source population is defined as the Czech general population, the questionnaires were only in Czech and people not speaking Czech were not sampled. The whole march was covered using this strategy and 636 postal survey questionnaires were distributed. The response rate was 16 percent.[6]

2. Sampling zeros

The second step necessary for testing causal effects in case-control designs is sampling of zeros, i.e. controls or negative cases. The selection of controls is considered to be the crucial

[5] However, even in the seemingly easy situation when all cases are sampled, there is a danger of selection bias. The list from which we take the cases, may not include all existing cases or the list itself could correlate with explanatory variables. For instance, when studying what determined people to commit political suicides during the communist regime, reliance on reports in that time-media might cause bias in our sample. Geddes points at the same problem of selection bias in comparative politics, which tends to examine only well-known cases or cases coming only from one region (Geddes 2003, 89–106).

[6] The assessment of non-response bias is done using data from brief face-to-face structured interviews carried out with every sixth respondent before they got the postal questionnaire. As the response rates for this face-to-face interview are usually high (here 89 percent) they provide a better picture on the character of the population of the demonstration participants.

and the most demanding challenge of the case-control design. The zeros constitute the comparative framework of the study and are inevitably related to what results one gets. In particular, it is important to establish comparability of cases and controls by avoiding or at least reducing the pitfalls of *selection bias*, *confounding bias*, and *information bias* (Wacholder et al. 1992).

A primary challenge in selecting comparable controls, as already said, is the identification of the source population that both, cases and controls come from. Units that do not belong to the population are units that even in theory could have never been cases, i.e. could have never been exposed (Armenian 2009). A classic example from epidemiology is that only women are part of the population when studying ovarian cancer because men cannot ever have this disease. To determine what controls to include we must apply the same criteria that we used for definition of cases. Hence, if we used some restriction criteria, such as terrorist attacks only in democracies or Czech anti-austerity protestors, we should apply the same criteria to define our controls and use negative cases only from democracies or the Czech population.

Notice that the specification of the population in too general way does not necessarily mean that our results will be biased. The redundant cases can be later handled in the analysis; either excluded or controlled. If, for instance, we would include in our study Americans as controls, we would have to exclude them from the analysis or ad a dummy variable for Americans into the analysis. If the data do not include a variable distinguishing the correct population from the redundant cases (Czechs and Americans), who did not have opportunity for exposure, we underestimate the relationships and our results will show smaller effects (Armenian 2009, 42). All in all, too brought definition of our population does not cause that much harm to the results at least in a way that it will not favour confirmation of our theories (it will probably underestimate them). However, more precise specification of our population helps avoid the unnecessary collection of data, which is in practical research an important aspect.

By contrast inappropriate restriction of our population can be a fatal problem as it cannot be fixed later in the analysis and can result into biased inference. When specifying the sub-population of controls we have to be careful not to rule out some explanations, i.e. definition of the controls should not be correlated with potential independent variables. This could, for instance, happen if we sampled our controls from people, who signed up to receive newsletters from the organizers of the march Stop the government. The problem is that positive cases did not have this restriction (cases were not only protestors receiving newsletters). This definition of controls inserts bias as receiving a newsletter probably correlates with predictors of protest: ideological affiliation, social networks, political interest etc. Hence, results from such study would probably be biased and would underestimate the effect of the above mentioned factors.[7]

A practical hint to define comparable controls and determine theoretically interesting source population of our case-control study in general is to perceive controls as “normal” or “average” units that did not experience the outcome. In our example, the unit of analysis is individual people coming from the Czech general population, i.e. all people speaking Czech who in theory could have come to the Prague anti-austerity demonstration if they experienced

[7] Notice that re-specification of the target population would not solve the problem. For instance, we could sample the controls from the list of trade union member and after that also apply this restriction to our cases (i.e. to use only protestors, who declared trade union membership). However, in this case the research question would be different (not “Why do people participate in anti-austerity protest?” but “Why do trade union members participate in anti-austerity protest?”) and the results would be less general, applicable only a sub-population of anti-austerity protestors, who are trade union members.

the causes: could have been recruited by trade unions or citizen movements, could have had supportive political attitudes, individual resources, etc.

Having decided on the target population or the sub-population, again several strategies how to sample controls from this population are available. A strategy which is often applied to positive cases, i.e. to take all cases, is with controls counterproductive or impossible. Recall that our cases are rare cases, i.e. there exist incomparably much more units without outcome. In order to make effective examination of causal effects, we do not have to survey the whole Czech population, as in our model example, or gather data on all year-country units in last 2000 years to have eligible controls for revolutions. All we need is a non-biased sample from this population.

Probably the most common strategy is random sampling. We used this strategy to sample our controls as well. Specifically, we use data representative of the Czech population from the Czech wave of the European Social Survey (ESS) 2012 that selects individuals by random probability method (fielded in January 2013). A number of questions asked in the ESS overlap with questions asked in the protest survey, which is necessary for the case-control design as there need to be the same measures for cases and controls. Moreover, the survey includes a question on participation in protest in last twelve months. Thanks to this question we can distinguish people, who might have taken part at the Stop the Government march (though it is very unlikely that the ESS would cover any of the participants).

In some situations random sampling within the target population is not possible because of costs reason or because the list of units, from which to do the random sample, is missing. In this situation, controls can be taken from a sub-population restricted by a variable, which does not correlate with the outcome (Schlesselman 1982, 77). Though results from such study are not generalizable in a statistical way to the target population, the results on causal effects are valid and would probably hold also in other cases. The only challenge here is to find the variable, which does not correlate or does not condition the effect of our explanatory variables.

3. Statistical analysis of case-control data

Table 1 gives the summary statistics for the two separate samples of cases (protestors at the Stop the government march) and controls (ESS data representative to the Czech general population) we have collected in the two previous steps.

– Table 1 –

Once the tasks of sampling ones and zeros have been accomplished, we are ready to proceed with the data analysis. A statistical models that is most commonly used to analyse case-control data is a standard logistic regression that specifies the probability of a one given the values of a set of explanatory variables. Here it is important to note that more advanced statistical analysis of case-control data will only be valid with some minor corrections. Fortunately, these are easy to implement and readily available in widely used statistical softwares, such as STATA or R (King and Zeng 2001; Imai, King, and Lau 2008). From a practical point of view, the analysis of case-control data is hence very convenient as it does not require any extra or advanced statistical skills because the main method rests on the widely used logistic regression.

Two corrections of the standard logistic regression that are necessary to analyse case-control data are the 1) case-control correction for selecting on the dependent variable and in some situations (but not always) 2) rare events correction for the rare event character of the data.

The first correction is related to the basic challenge of inference in case-control designs that lies in the fact that case-control studies themselves do not reveal the distribution of ones in the population (Manski 1995, 81). Without this information causal effects can only be expressed in relative but not absolute terms. In logistic regression, this means that the selection on the dependent variable biases the constant term. However, the other regression coefficients as such are not biased and provide valid estimates of the effects.

The interpretation of results from logistic regression is not usually based on the coefficients that are in terms of the log odds. Often quantities of interest that are easy to understand, such as predicted probabilities, are calculated. However, to calculate predicted probabilities, valid estimates for the constant term are needed. We can easily correct the constant term using a simple correction formula if we have knowledge on the distribution of the dependent in the population from sources external to the data (Manski and Lerman 1977; Prentice and Pyke 1979).[8] More advanced techniques allow us to proceed even in cases when we have no or only partial knowledge of this distribution (Manski 1995; King and Zeng 2004).

In our case, a variety of sources provide estimates about the number of protesters at the anti-austerity demonstration and we can use this external information to correct the constant term accordingly (see table 1). Estimates are provided by five different organizers of the demonstration, the police, as well as the CCC research team that collected the protest data. Unsurprisingly, given the different incentives of these actors, these estimates differ substantially from each other with the highest estimate of 15´000 participants coming from two organizers. The research team´s estimate is far more conservative and counts a third of this number, i.e. 5`000 protesters at the Stop the Government march. Given this discrepancy we will calculate the case-control correction for different estimates and compare the results.

– Table 2 –

The second correction relates to the situation with extremely low proportion of positive cases. As already explained, the main purpose of the case-control design is to sample rare cases in a way that they are not rare anymore and that the data include a balanced number of cases and controls. This is the case in our example as well where we combined N= 93 protesters captured in the protest survey with N = 1263 controls taken from the Czech sample of the ESS 2012.[9] However, there can be situations that despite using the case-control design, the proportion of cases in the dataset will still be very low. This happens when the phenomena are very rare in absolute terms (e.g. only 15 revolutions ever took place). Here even when using case-control design, which enables to cover all existing cases, we will not be able to increase the proportion of cases in the final sample. The reason is that the sample of controls should be representative to the population of controls and hence a larger number of controls needs to be usually sampled.

[8] The formula for the case-control correction of the constant term α, known as prior correction, is given by $\alpha - \log(\ ((1-\tau) \ /\tau)* \ (p/(1-p)) \ )$, where τ is the prior fraction of ones taken from external sources and p is the fraction of ones in the data.

[9] These numbers refer to cases and controls for which full information on all the covariates is available. Since this example mainly serves for the purpose of illustration we deal with the problem of missing data by list-wise deletion. In a more serious setting one would want to use multiple imputations to address this issue.

The general problem of logistic regression with very low proportion of cases is that the probability of ones will be underestimated and, conversely, the probability of zeros overestimated. Unlike in the case of case-control correction (solving the selecting on the dependent variable), the low proportion of cases in the data, i.e. having way more zeros than ones, biases all coefficients in a logistic regression and leads to higher variances (King and Zeng 2001). And even if we correct for the above described bias the derived quantity of interest, i.e. the predicted probability would still have problems because it ignores the uncertainty in the coefficients estimates.

Again, this bias can be easily corrected by taking the uncertainty into account (King and Zeng 2001). Following analysis will show the rare events correction but we expect it to only marginally improve our estimates, as they are not really rare in the sample. Table 3 presents the uncorrected logistic regression results of our Czech example (I) and compares them to the results when correcting for selecting on the dependent variable (II), the rare event character of the data (III) or both (IV). Note that all variables except for the dummy variables have been standardized by dividing by two standard deviations to facilitate the comparison (Gelman and Hill 2007).

– Table 3 –

– Figure 1 –

The results in Table 3 model I, i.e. the naïve logit model, suggest the following about the determinants of protest participation at the Stop the Government march. Whereas gender does not matter, the protest participation at this anti-austerity demonstration decreases with age and increases with education. Interestingly, individual income and employment status are not statistically related to the participation at this event. With regards to political attitudes, political interest is the most important predictor for protest participation. Those who trust the parliament, are more rightist in their political ideology and those who are satisfied with democracy were less likely to join the demonstration. Finally, union membership is clearly related to a higher probability of protest participation which suggests the successful mobilization by trade unions.

Before we turn to the substantive effect sizes of these results, we quickly illustrate how the correction for the case-control design and the rare-events character of the data affects the results. Overall, the effects are minor. Correcting for the case-control design using the estimates of the number of protesters as provided by the research team only changes the constant term, just as it should. The constant decreases from -5.4 to -10.4 and thus considerably corrects the probability of participation downward (from $\log^{-1}(-5.4) = 0.004$ to $\log^{-1}(-10.4) = 0.00003$ for an average person with all the covariates in the model set to zero). The rare-event correction has two effects. It slightly decreases (unbiases) most of the logit coefficients and it yields smaller standard errors, although this latter effect is hardly worth the mention. Finally, model IV combines both corrections and figure 1 shows how it compares to the naïve logit model. Again, the major difference lies in the constant term which is markedly smaller in the corrected model. Other results do not show much difference across the four ways of estimation.

Since the coefficients of logistic regressions are unintuitive to interpret they do not represent the endpoint of a statistical analysis of case-control designs. Instead they serve as intermediary quantities from which we can derive meaningful quantities of interest that are easy to understand and communicate. In our case, this would involve the average probability of participating in the anti-austerity march for different categories of the independent variables. Our illustration shows the average probability of protest participation for trade union members and non-members, the difference and the respective uncertainties in these probabilities (see figure 2).

– Figure 2 –

Here the differences across the results from the different ways of estimation are enormous. In the un-corrected or naïve logistic regression non-members of trade unions are predicted to participate at the Stop the Government march with a probability of about 6 percent and trade union members with a probability of roughly 17 percent. Clearly, these estimates are unrealistically high and stem from the fact that we do not know the true fraction of protesters in the population. When we correct our logistic regression models by incorporating the number of protesters, these probabilities drop considerably. The averaged predicted probability of protest participation lies at 0.2 percent for people who are not member of a trade union and at 0.9 percent for those who are union members.

Note that also differences in predicted probabilities that show the relative effect vary. According to the naïve logistic regression model without any correction, the probability of participation increases by 11 percent if a non-member of trade unions becomes a member. However, the results from the logistic model with corrections show that the increase in participation probability is only .7 percent. This demonstrates that relatively speaking trade union members were much better mobilized than non-members but at the same time these numbers reflect the fact that protestors at the anti-austerity march still present rare cases in absolute terms.

References

Armenian, Haroutune. 2009. *The Case-Control Method: Design and Applications*. Oxford University Press.

Armingeon, Klaus. 2002. “The Effects of Negotiation Democracy: A Comparative Analysis.” *European Journal of Political Research* 41 (1): 81–105.

Brady, Henry E., and David Collier. 2010. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Second Edition edition. Lanham, Md: Rowman & Littlefield Publishers.

Corrigall-Brown, Catherine, David A. Snow, Kelly Smith, and Theron Quist. 2009. “Explaining the Puzzle of Homeless Mobilization: An Examination of Differential Participation.” *Sociological Perspectives* 52 (3): 309–35.

Flam, Helena. 2001. *Pink, Purple, Green: Women’s, Religious, Environmental and Gay/lesbian Movements in Central Europe Today*. East European Monographs.

Geddes, Barbara. 2003. *Paradigms and Sand Castles: Theory Building and Research Design in Comparative Politics*. University of Michigan Press.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

George, Alexander L., and Andrew Bennett. 2005. *Case Studies and Theory Development in the Social Sciences*. Fourth Printing edition. Cambridge, Mass: The MIT Press.

Imai, Kosuke, Gary King, and Olivia Lau. 2008. “Toward a Common Framework for Statistical Analysis and Development.” *Journal of Computational and Graphical Statistics* 17 (4): 892–913.

Jasper, James M., and Jane D. Poulsen. 1995. “Recruiting Strangers and Friends: Moral Shocks and Social Networks in Animal Rights and Anti-Nuclear Protests.” *Social Problems* 42 (4): 493–512.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, N.J: Princeton University Press.

King, Gary, and Langche Zeng. 2001. “Logistic Regression in Rare Events Data.” *Political Analysis* 9 (2): 137–63.

———. 2004. *Encyclopedia of Biopharmaceutical Statistics*. New York: Marcel Dekker.

Lacy, Michael G. 1997. “Efficiently Studying Rare Events: Case-Control Methods for Sociologists.” *Sociological Perspectives* 40 (1): 129–54.

Lijphart, Arend. 1971. “Comparative Politics and the Comparative Method.” *American Political Science Review* 65 (03): 682–93.

Manski, Charles F. 1995. *Identification Problems in the Social Sciences*. Harvard University Press.

Manski, Charles F., and Steven R. Lerman. 1977. “The Estimation of Choice Probabilities from Choice Based Samples.” *Econometrica* 45 (8): 1977–88.

McAdam, Doug. 1986. “Recruitment to High-Risk Activism: The Case of Freedom Summer.” *American Journal of Sociology* 92 (1): 64–90.

Porta, Donatella della. 2014. *Social Movements in Times of Austerity. Bringing Capitalism Back In*. Polity Press.

Prentice, R. L., and R. Pyke. 1979. “Logistic Disease Incidence Models and Case-Control Studies.” *Biometrika* 66 (3): 403–11.

Przeworski, Adam, and Henry Teune. 1970. *The Logic of Comparative Social Inquiry*. Wiley-Interscience.

Schlesselman, James J. 1982. *Case-Control Studies: Design, Conduct, Analysis*. 1 edition. New York: Oxford University Press.

Seawright, Jason. 2002. “Testing for Necessary And/or Sufficient Causation: Which Cases Are Relevant?” *Political Analysis* 10 (2): 178–93.

Skocpol, Theda. 1979. *States and Social Revolutions: A Comparative Analysis of France, Russia and China*. Cambridge: Cambridge University Press.

Teorell, Jan, Mariano Torcal, and José R Montero. 2007. “Political Participation: Mapping the Terrain.” In *Citizenship and Involvement in European Democracies: A Comparative Analysis*, edited by Jan W. van Deth, José R Montero, and Westholm, Anders, 334–58. London; New York: Routledge.

Van Evera, Stephen. 1997. *Guide to Methods for Students of Political Science*. Cornell University Press.

van Stekelenburg, Jacquelien, Stefaan Walgrave, Bert Klandermans, and Joris Verhulst. 2012. “Contextualizing contestation. Framework, design and data.” *Mobilization* 17 (3): 249–62.

Verhulst, Joris. 2010. “February 15, 2003: The World Says No to War.” In *The World Says No to War: Demonstrations against the War on Iraq*, edited by Stefaan Walgrave and Dieter Rucht, 1–19. Minneapolis: Univ Of Minnesota Press.

Walgrave, Stefaan, and Dieter Rucht. 2010. *The World Says No to War Demonstrations against the War on Iraq*. Minneapolis: University of Minnesota Press.

Walgrave, Stefaan, and Joris Verhulst. 2011. “Selection and Response Bias in Protest Surveys.” *Mobilization: An International Quarterly* 16 (2): 203–22.

Table 1: Summary statistics of cases and controls

	CCC Protest Survey				*European Social Survey 2012*			
	Stop the Government				*Czech sample*			
	Mean	SD	Min.	Max.	Mean	SD	Min.	Max.
Participation at Stop the Government Demonstration	1.0	--	1	1	0.0	--	0	0
Male	0.65	--	0	1	0.50	--	0	1
Age in years	50.4	15.9	20	80	47.9	16.4	14	90
Education	7.4	2.4	3	11	5.0	2.4	1	11
Income1	0.17	--	0	1	0.20	--	0	1
Income2	0.49	--	0	1	0.49	--	0	1
Income3	0.33	--	0	1	0.31	--	0	1
Unemployed	0.02	--	0	1	0.05	--	0	1
Political interest	3.3	0.7	2	4	2.0	0.7	1	4
Trust in parliament	0.02	--	0	1	0.19	--	0	1
Left-right-ideology	3.2	1.6	1	8	5.1	2.5	0	10
Democratic satisfaction	3.0	1.9	1	10	5.0	2.4	0	10
Trade union	0.23	--	0	1	0.06	--	0	1

Table 2: Estimates of the Number of Protestors at the Stop the Government Demonstration

Actor	*Estimate of Number of Protestors*
Organizer 1	10´000
Organizer 2	15´000
Organizer 3	12´500
Organizer 4	15´000
Organizer 5	10´000
Police	10´000
CCC Research Team	5´000

Table 1: Results of logistic regressions for protest participation at the Stop the Government demonstration

	I.		II.		III.		IV.	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE
Constant	-5.40	(0.54)	-10.43	(0.54)	-5.16	(0.53)	-10.14	(0.53)
Male	-0.01	(0.31)	-0.01	(0.31)	-0.01	(0.31)	0.00	(0.31)
Age in years	-0.71	(0.36)	-0.71	(0.36)	-0.68	(0.36)	-0.68	(0.36)
Education	1.81	(0.32)	1.81	(0.32)	1.74	(0.31)	1.74	(0.31)
Income2	0.12	(0.44)	0.12	(0.44)	0.10	(0.43)	0.08	(0.43)
Income3	0.00	(0.51)	0.00	(0.51)	-0.01	(0.51)	-0.03	(0.51)
Unemployed	0.35	(0.82)	0.35	(0.82)	0.52	(0.81)	0.54	(0.81)
Political interest	3.85	(0.40)	3.85	(0.40)	3.71	(0.40)	3.69	(0.40)
Trust in parliament	-2.06	(0.66)	-2.06	(0.66)	-1.84	(0.65)	-1.82	(0.65)
Left-right-ideology	-0.75	(0.32)	-0.75	(0.32)	-0.72	(0.32)	-0.70	(0.32)
Democratic satisfaction	-1.60	(0.39)	-1.60	(0.39)	-1.54	(0.38)	-1.54	(0.38)
Trade union	2.04	(0.44)	2.04	(0.44)	1.98	(0.44)	1.98	(0.44)
Case-Control-Correction	*no*		*yes*		*no*		*yes*	
Rare-Events-Correction	*no*		*no*		*yes*		*yes*	

Note: Standardized logit coefficients (mean-centred and divided by two standard deviations) and standard errors in parentheses.

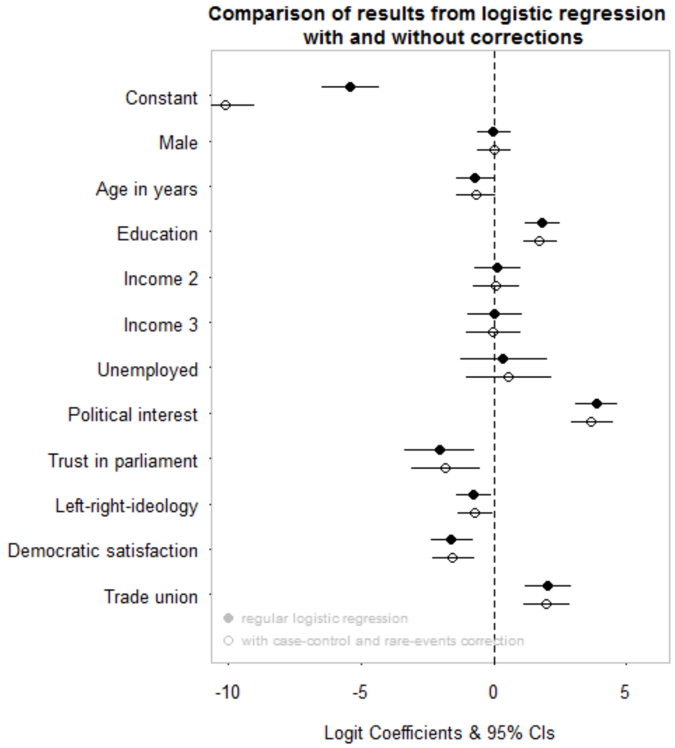

Figure 1: Visual comparison of the results of models I and IV in table 3. Dots are standardized logit coefficients and line segments represent 95% confidence intervals. Explanatory variables are considered “statistically significant” if the confidence intervals do not include the null (the dotted vertical line). As intended, the corrections mainly change the constant term and thus correct for the fraction of ones in the population. In this case we use the estimation of 5´000 participants provided by the CCC research team.

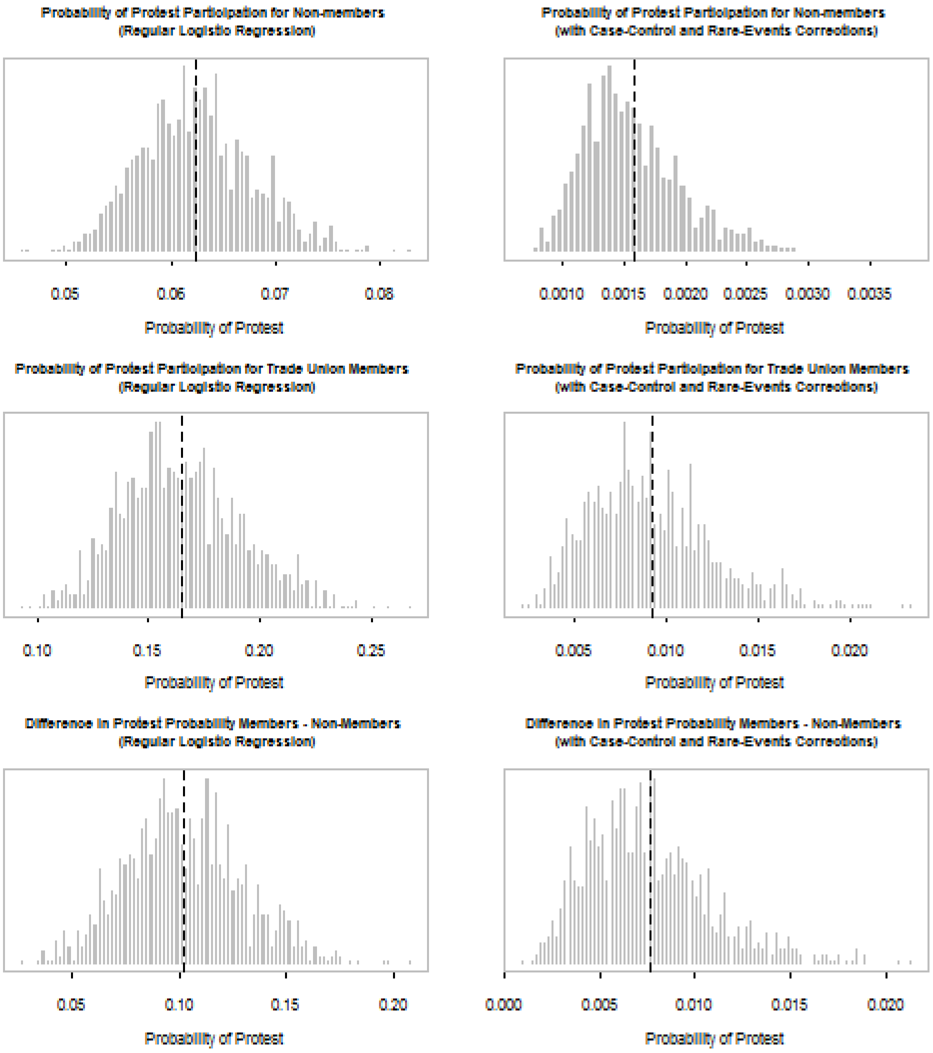

Figure 2: Averaged predicted probabilities for protest participation at the anti-austerity demonstration for non-members and members of trade unions with simulated uncertainties. Derived from the estimates of model I and model IV in table 3. Clearly, the corrected model which gives much smaller protest probabilities and differences between members and non-members than the naïve logistic regression model.