Simulating the Effects of Selection Bias in the Minorities at Risk Project.

Andreas Beger[*]
Florida State University

18 April 2008

The Minorities at Risk (MAR) Project collects data on approximately 300 ethnic groups and is one of the few data sets on ethnic groups. However, it suffers from selection bias since it collects data on groups that are deemed to be "at risk", and consequently its use has been limited. However, while the selection bias in MAR limits the types of inferences that can be drawn using it and distorts estimates of the causal effect of variables, causal inferences made using MAR are not fundamentally flawed. The reason is that the selection bias in MAR will likely weaken coefficient estimates and inflate standard errors, thus providing harder hypothesis tests and a conservative bias that increases the chances that we falsely reject true hypotheses.

[*] Ph.D. Student, Department of Political Science, Florida State University, 113 Collegiate Loop, Tallahassee, FL 32306-2230. abeger@fsu.edu.

1 Introduction

"The Minorities at Risk (MAR) Project is a university-based research project that monitors and analyzes the status and conflicts of politically-active communal groups in all countries with a current population of at least 500,000."[1] Communal groups are considered to be politically active or "at risk" by either of the following criteria: (1) "the group collectively suffers, or benefits from, systematic discriminatory treatment vis-a-vis other groups in a society", or (2) "the group is the basis for political mobilization and collective action in defense or promotion of its self-defined interests[2]". These selection criteria for minority group inclusion lead to selection bias and the consequent problems for any type of research that attempts to generalize beyond minorities that are at risk.

In regard to the two forms of inference discussed in King, Keohane and Verba (1994), descriptive and causal inference, the selection bias in MAR is a major problem for descriptive inferences, but not necessarily for causal inferences. Descriptive inferences drawn from the sample in MAR are not generalizable to the population of ethnic groups in the world if the selection process in MAR is correlated with whatever dependent variable we are interested in (King, Keohane and Verba 1994, 141). As long as the population we are trying to generalize to is something other than the set of minorities at risk, MAR data, without some type of correction, will potentially lead to biased descriptive inferences.

Causal inferences are probably the more common type of inference that empirical research is interested in making. Because of the selection issue in MAR, many quantitative researchers reject the use of its data for the study of ethnic conflict, rebellion, protest, and similar topics. For example, Fearon (2003, 196) presents an alternative and more inclusive set of ethnic groups that was motivated in part because of the selection bias in MAR. Hug (2003) develops an alternative estimator that may mitigate some of the selection bias in MAR and similar datasets, but in either case the implication is that it is inappropriate to just use data like MAR without considering the selection bias. I will present the results of several simulations below to show that while the selection bias certainly is unfortunate, and limits the types of inferences we can draw, *the selection bias in MAR does not fundamentally undermine a researcher's ability to make valid causal inferences using MAR data*. More specifically, the simulations presented here suggest that in applied research using MAR data, the selection bias will influence causal inferences in two ways: (1) it will attenuate coefficient estimates, i.e. weaken estimated causal effects (King, Keohane and Verba 1994), and (2) disproportionately increase standard errors even after reduced sample sizes are taken into account.

[1] From the MAR website at <http://www.cidcm.umd.edu/mar/>. Accessed 15 February 2008.

[2] From the MAR website. <http://www.cidcm.umd.edu/mar/about.asp>. 15 February 2008.

2 Nature of selection process in MAR

Selection bias is a result of nonrandom selection. We have some population of units that we wish to study, i.e. all communal groups in all countries in the world, but we lack data for all of these groups and rather study a sample drawn from that population. No one knows exactly how many ethnic groups there are in the world (maybe a couple thousand), but all lists of ethnic groups are samples of this larger population of ethnic groups.[3] Fearon (2003) has more than 800 groups in his list, whereas MAR collects data on around 300. The problem of selection bias arises if the process used to include groups in the MAR sample is systematically related to variables that are of interest to a particular research project (King, Keohane and Verba 1994, 128-149). Since the Minorities at Risk project includes communal groups in its sample that are politically active or "at risk", based on two criteria: (1) "the group collectively suffers, or benefits from, systematic discriminatory treatment vis-a-vis other groups in a society", or (2) "the group is the basis for political mobilization and collective action in defense or promotion of its self-defined interests", the selection process (being at risk) is probably related to most dependent variables that we might use with MAR data.[4] Therefore selection bias will be an issue for most research that uses MAR.

Here is a more specific example to illustrate why selection bias will be an issue. Assume we have a dichotomous dependent variable y that measures whether an ethnic group engaged in violent rebellion in a given year or not. Let us also assume that this variable is partly a function of group size, i.e. what proportion of a country's population consists of members of that group. The larger a group's relative size, the more likely it is to engage in violent rebellion against the state. Let x denote this variable. Furthermore, there is another dichotomous variable, s, that measures whether a minority group was at risk in a given year or not. Ordinarily, we would test for whether a relationship exists between x and y by estimating a probit regression where:

$$Pr(y = 1|x) = f(x, \varepsilon_Y)$$

or the probability that $y = 1$ is a function of x as well as some stochastic error term ε_Y. To make it easier to illustrate the subsequent argument, let's deal with the propensity of a minority group to engage in violent rebellion instead, so that y^*, the propensity for violent rebellion becomes a linear function of x:[5]

$$y^* = \beta_{Y0} + \beta_{Y1}x + \varepsilon_Y \tag{1}$$

[3] In fact, it probably is impossible to construct an exhaustive list of ethnic groups given how arbitrary (and fluid) the concept is (Hug 2003, 269). Fearon (2003) and Chandra (2006) have nice discussions about the complexity of an operational and conceptual definition of what constitutes and ethnic group. In any case, the implication is that *all* lists of ethnic groups probably suffer from some degree of selection bias.

[4] From the MAR website. <http://www.cidcm.umd.edu/mar/about.asp>. 15 February 2008.

[5] One can derive the probit (and logit) estimators exactly by thinking of some continuous latent variable that measures the propensity of an ethnic group to engage in violent rebellion (King 1989, 110-115).

However, we only observe y and x for ethnic groups that are included in the MAR sample, i.e. groups that are at risk. Let s be a dichotomous variable that denotes whether a group is at risk or not, and as before, let s^* be a latent variable that captures the propensity for an ethnic group to be at risk. The selection problem comes into play because for most dependent variables (and independent variables) that we can imagine using with MAR data, the correlation with s will not be zero. Specifically, there are three potential sources of selection bias that I can think of.

First, most independent variables that are related to our dependent variable of interest probably are also related to the likelihood that an ethnic group is at risk. In our example, ethnic groups that are larger are more likely to engage in violent rebellion, but they are also more likely to be included in the MAR sample because they are more likely to clear the population threshold MAR requires for inclusion and because larger ethnic groups that are politically active are easier to identify than smaller ethnic groups. This constitutes non-random selection, but it is actually *not* a source of selection bias because we already include x in the estimation of y and thus in effect control for x's role on the selection process (King, Keohane and Verba 1994, 137). Thus casual inferences about other variables will not be biased from this source. The next two problems, however, will bias causal inferences about other variables.

Second, most dependent variables are probably also related to s, i.e. $\text{cor}(y,s) \neq 0$. Ethnic groups that engage in violent rebellion are much more likely to be considered at risk (one would think they actually all are considered at risk given their population size exceeds the MAR threshold of 500,000) than ethnic groups that do not, even once we take the effect of group size on y into account.

Third, unobserved factors (that are captured in the error term ε_Y) that make it more likely that $y = 1$ for an ethnic group with a given group size probably will also make it more more likely that $s = 1$, i.e. that the group is observed in the MAR sample. In other words, the error components of the selection process and of process that results in y are correlated. With these three claims we can thus write the process that produces s as:

$$Pr(s = 1|x,y) = f(x,y,\varepsilon_S)$$

Where $\text{cor}(\varepsilon_Y, \varepsilon_S) \neq 0$. If we instead look at the propensity that an ethnic group is at risk, s^*, this becomes:

$$s^* = \beta_{S0} + \beta_{S1}x + \beta_{S2}y^* + \varepsilon_S$$

Note that we can substitute equation 1 for y^*:

$$s^* = \beta_{S0} + \beta_{S2}\beta_{Y0} + (\beta_{S1} + \beta_{S2}\beta_{Y1})x + \beta_{S2}\varepsilon_Y + \varepsilon_S \qquad (2)$$

Thus even if the two error terms were not correlated by themselves, if s is a function of our dependent variable y, the new error term of s^*, which equals $\beta_{S2}\varepsilon_Y + \varepsilon_S$, would be correlated with the error term for the outcome equation, ε_Y. Furthermore, equation 2 implies that as long as $\text{cor}(\varepsilon_Y, \varepsilon_S) \neq 0$ or $\beta_{S2} \neq 0$, our dependent variable y will be correlated with the selection mechanism based on s and the error terms for the outcome (y) and selection (s) equations will be correlated.

Here is the reasoning for this claim. As a first step, let's assume that there is some sort of underlying latent propensity for minority groups to be at risk that translates into the real world binary outcome of a group being either "at risk" or 'not at risk", based on whether the latent propensity is above or below some threshold τ. For convenience let us also assume that this threshold τ is 0.[6] MAR selects its sample of groups based on this propensity, i.e. groups that have a value above 0 are part of the sample, groups with a propensity value below zero are not part of the sample. If the dependent variable in a study was not systematically correlated with this propensity for being at risk, the selection bias in MAR would not present any problems. However, for most dependent variables like engaging in violent rebellion, protest, etc., we probably believe that the propensity for violent rebellion, etc. is systematically correlated with the propensity of being at risk. Minority groups that are at risk are more likely to engage in rebellion, etc., than groups that are not at risk. For such dependent variables, MAR selects a sample on the basis of a process that is correlated with our dependent variable.

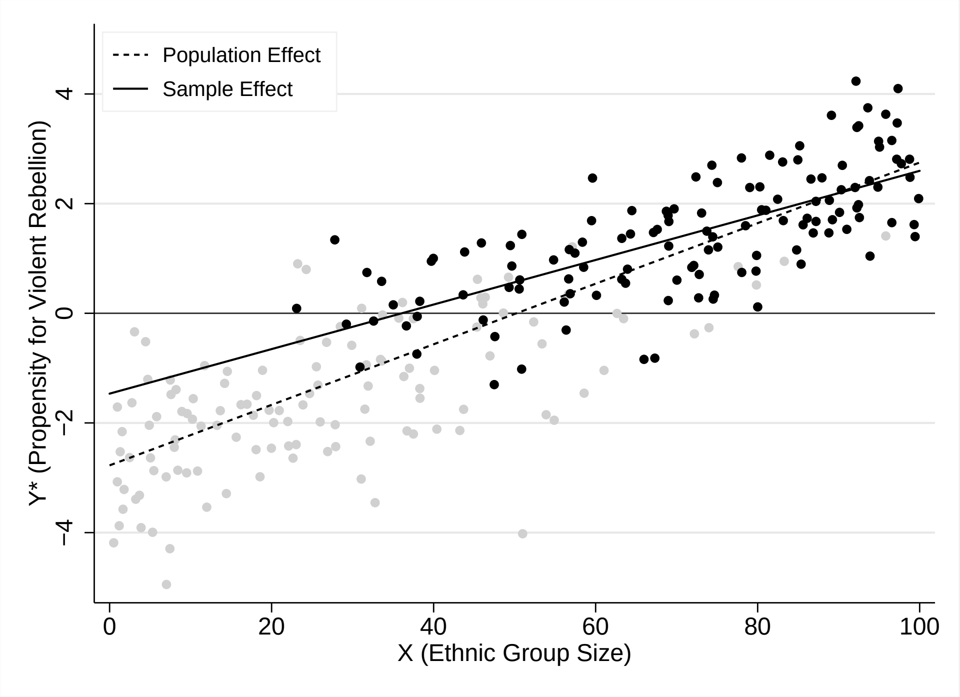
3 Effects of MAR selection bias

I have argued that the selection mechanism in MAR is correlated with most dependent variables we might use with it. King, Keohane and Verba (1994, 128-132) directly discuss the effects of such selection bias: "*any selection rule correlated with the dependent variable attenuates estimates of causal effects on average*" (emphasis in original). They proceed to illustrate this claim with a figure that shows the effects of truncation based on the dependent variable (King, Keohane and Verba 1994, Figure 4.1, 131). I have replicated a similar figure that specifically illustrates the effect that selection on a variable (propensity of being at risk) that is *correlated* with our dependent variable (propensity for violent rebellion) has on estimates of the relationship between an explanatory variable and violent rebellion. The simulations used to produce this figure also suggest that standard errors will be disproportionately increased after taking the reduced sample size into account. This latter claim is explored in more detail further below.

Figure 1 shows a hypothetical world of minority groups. The x-axis shows some (uniformly

[6] We essentially do the same when we use probit and logit regression models since both can be derived from a latent variable model, and where τ is also assumed to be equal to 0.

Figure 1: Selection Bias in Minorities at Risk



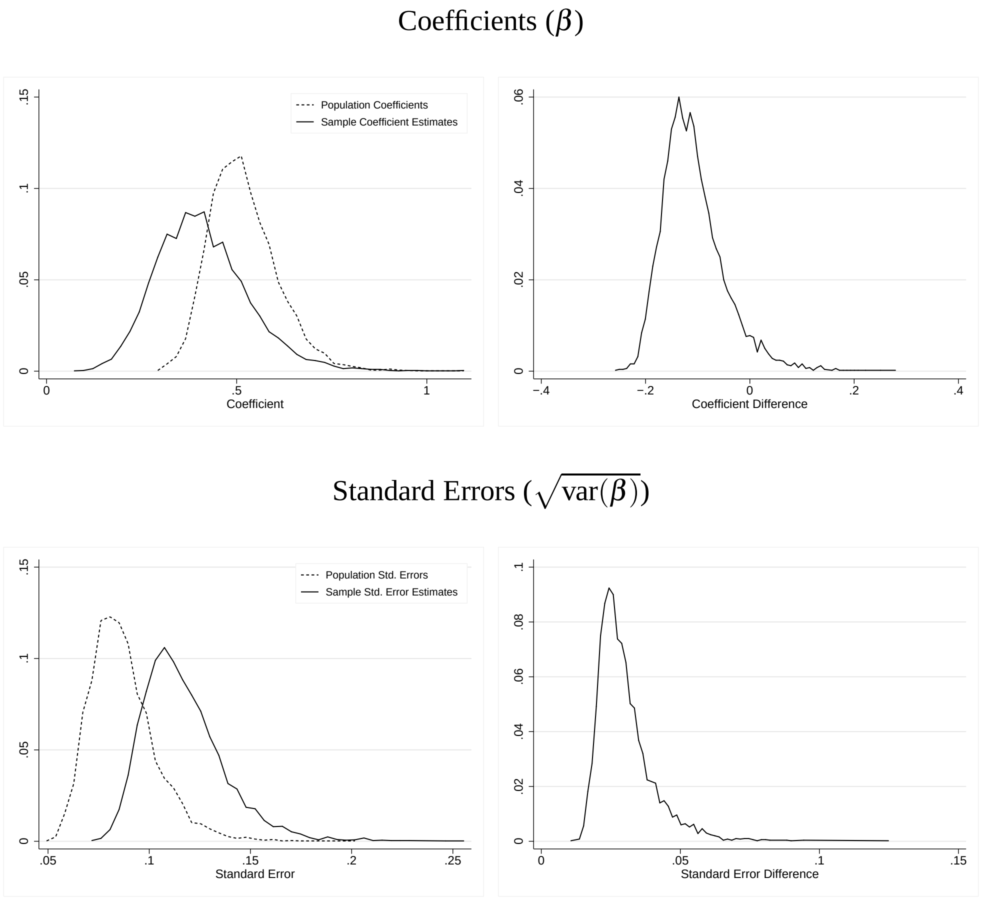
distributed) explanatory variable that is continuous and ranges from 0 to 10. The y-axis shows the propensity for being at rebellion.[7] Minority groups that are above 0 on this latent propensity are engaged in violent rebellion, all others are not. The black and grey dots show individual observations and constitute the full population of minority groups in this example. In the full population, the explanatory variable is positively related to the propensity for violent rebellion, albeit with a normally distributed error term ($\varepsilon \sim N(0,1)$). The dashed line shows this relationship (or one could run a regression of the propensity for violent rebellion on the explanatory variable).

Now let's introduce a selection mechanism. The black dots show the sample that is drawn from the full population of minority groups after we select from another variable, the propensity for being at risk, that is correlated with the propensity for violent rebellion (cor=0.8). The solid line shows the estimated relationship between the explanatory variable and the propensity for violent rebellion in the new (systematically biased) sample. The coefficient estimate in the (biased) sample is weaker (closer to zero) than the 'true' coefficient in the full population of minority groups. Thus a statistical test of the hypothesis that the explanatory variable is positively related to the (propensity for) violent rebellion would

[7] One could do what I am about to do just as well with a binary dependent variable, but it is graphically a lot clearer when we use the the latent propensity instead.

be harder to meet in the selected sample than in the full population.

Figure 2: Coefficient and Standard Error distributions for MAR scenario.



Coefficients (β)

Standard Errors ($\sqrt{\mathrm{var}(\beta)}$)

Repeating the process of random draws that was used to generate figure 1 does not change the main conclusion that coefficient estimates in the sample selected from a process that is correlated with the propensity for violent rebellion will be systematically weaker (biased towards zero) than those in the full population. To substantiate this claim, figure 2 shows the distribution of coefficient estimates and standard error estimates for the sample and full population that I obtained after repeating the process used to generate figure 1 several thousand times. The first figure on the top left shows the distributions of coefficients for the full population and sample respectively. Since each iteration of the simulated data produces both a full population and sample coefficient that are unique to that iteration, just looking at the distribution of coefficients after several thousand iterations is potentially misleading (since it assumes that each coefficient pair for a sample and full population is independent from each other). Therefore the second graph on the top right of figure 2 shows a distribution of the difference between the population and sample coefficients. Negative values indicate that the sample coefficient was smaller than the population coefficient. While in a few cases the sample coefficient was larger than the population coefficient, in most iterations the sample coefficient was smaller than the population coefficient given the selection processed outline above.

The bottom two graphs in figure 2 show the corresponding information for the standard errors of each coefficient estimate. In this case, the standard errors for the sample were larger than the population standard errors in every single iteration.

Since we should expect standard errors to be larger in a sample due to the lower number of observations anyways, I also conducted some simulations to see whether standard errors in samples of the same size from a population of the same size were different depending on whether the selection process was random or correlated with the dependent variable.

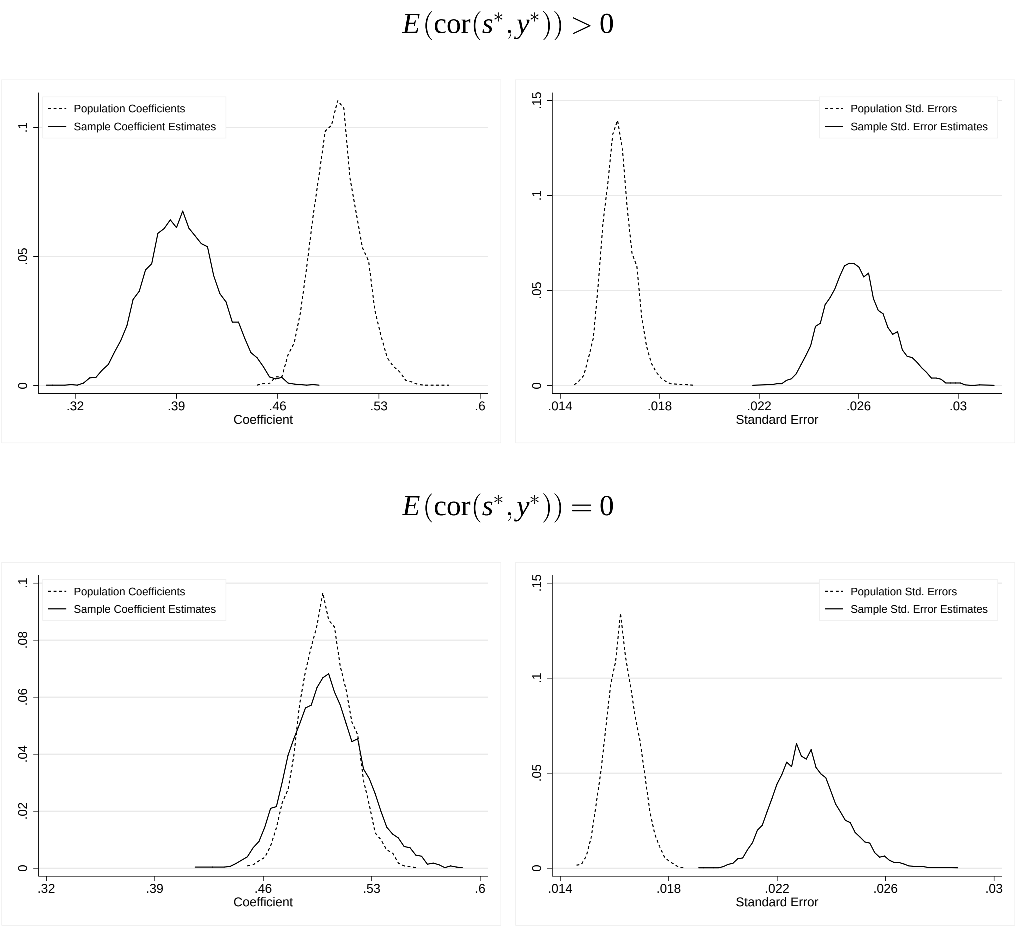
Figure 3 shows the resulting distributions of coefficients and standard errors. The two graphs on the top show the coefficients and standard errors from the sample and full population when the correlation between the the dependent variable and selection process is larger than 0. The two graphs on the bottom show the same for the situation where the correlation is 0, i.e. in the case where the sample consists of observations that were randomly selected from population. The population of simulated data was of equal size in both cases and the sample size for both instances is on average the same as well. As a comparison of the two graphs on the rights shows, while standard errors are always larger in the samples drawn from the full population, they are still slightly larger, on average, when the selection process is positively correlated with the dependent variable. Thus it seems that the standard errors in a sample that suffers from selection bias might be unduly large after accounting for the increase in standard errors due to the smaller number of observations used to estimate a regression.[8]

4 Conclusion

These simulations support the claim by King, Keohane and Verba (1994) and others that selection on a process positively correlated with the dependent variable of interest will bias causal inferences (i.e. coefficient estimates) towards zero, thus making standard hypothesis tests harder to meet. Furthermore, the simulations suggest that there are two reasons for this. First, the selection process outlined here will weaken relationships that exist in the full population and bias them towards zero in the sample of ethnic groups in MAR. This is exactly the point made by King, Keohane and Verba (1994, 129-132) in regard to se-

[8] I imagine this might depend heavily on what the data look like. In this case, we know that there is a linear relationship between *x* and *y* with a normally distributed error term. This probably will rarely be the case in real world data.

Figure 3: Coefficient and Standard Error distributions under nonrandom and random selection.



$E(\text{cor}(s^*, y^*)) > 0$

$E(\text{cor}(s^*, y^*)) = 0$

lection related to the dependent variable. Second however, the selection process may also inflate standard errors disproportionately in relation to the reduction in sample size (i.e. the increase in standard errors purely due to the lower number of observations in the sample).[9] Since standard hypothesis tests rely on the ratio of coefficients to standard errors, the type of selection process likely to occur in MAR thus biases causal inference towards a null finding in two ways, by reducing coefficient magnitudes, and by increasing standard errors.

Using just current MAR data, empirical analyses are more likely to reject hypothesis that are empirically supported in the full population. This implies that any statistical relationships that do occur in the MAR sample of ethnic groups should be generalizable to the

[9] Given the assumptions about the distribution of the error term I made in the simulations here.

full population of ethnic or minority groups in the world. Of course it also implies that there may be some relationships that exist in the full population but fail to meet standard hypothesis tests within the MAR sample.

5 Appendix

5.1 How figure 1 was produced

1. Draw 250 observations from a uniform distribution to generate x and the same number from a normal distribution to generate ε_Y and ε_S, where $\text{cor}(\varepsilon_Y, \varepsilon_S) = 0.25$.
2. Generate the propensity for violent rebellion with the function $y^* = \beta_{Y0} + \beta_{Y1}x + \varepsilon_Y$.
3. Generate the propensity for being at risk with the function $s^* = \beta_{S0} + \beta_{S1}x + \beta_{S2}y^* + \varepsilon_S$.
4. Estimate the relationship between x and propensity for violent rebellion in the full population (dashed line).
5. Drop observations that have a propensity for being at risk that is below zero (grey dots; $y^* \leq 0$).
6. Reestimate the relationship between the explanatory variable and propensity for violent rebellion (solid line).

5.2 How figures 2 and 3 were produced

1. Draw 2500 observations from a uniform distribution to generate x and the same number from a normal distribution to generate ε_Y and ε_S, where $\text{cor}(\varepsilon_Y, \varepsilon_S) = 0.25$.
2. Generate the propensity for violent rebellion with the function $y^* = \beta_{Y0} + \beta_{Y1}x + \varepsilon_Y$.
3. Generate the propensity for being at risk with the function $s^* = \beta_{S0} + \beta_{S1}x + \beta_{S2}y^* + \varepsilon_S$.
4. Estimate the relationship between x and propensity for violent rebellion in the full population using probit.
5. Reestimate the relationship between the x and propensity for violent rebellion when $y^* > 0$ using probit.
6. Save desired quantities.
7. Repeat starting at (1) for 5000 iterations.

5.3 Some disclaimers

Some preliminary cautions or ideas (i.e. stuff that needs more work):

1. If the error term is normally distributed in the full population, then drawing a biased sample will also lead to heteroscedasticity, i.e. the variance of the error term will not be constant anymore.

2. To what extent the selection bias is a problem depends on how highly correlated the selection process is with the dependent variable. In this example, the correlation is fairly high, but lower correlations will mitigate the problems of selection bias.

5.4 Additional files

`selectionbias.do` - STATA do file to replicate the figures.
`selectionbias.txt` - Log file for the do file above.

References

Chandra, Kanchan. 2006. “What is Ethnic Identity and Does It Matter?” *Annual Review of Political Science* 9: 397–424.

Fearon, James D. 2003. “Ethnic and Cultural Diversity by Country.” *Journal of Economic Growth* 8: 195–222.

Hug, Simon. 2003. “Selection Bias in Comparative Research: The Case of Incomplete Data Sets.” *Political Analysis* 11(3): 255–274.

King, Gary. 1989. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*. Ann Arbor: Michigan University Press.

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