CIS
uni | eth | zürich
Center for Comparative and International Studies (CIS)
Working Paper No 20, 2006
Published by the Center for Comparative and International Studies (ETH Zurich and University of Zurich)

The Effect of Misclassifications in Probit Models: Monte Carlo Simulations and Applications

by Simon Hug
University of Zurich

Universität Zürich
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

The Effect of Misclassifications in Probit Models Monte Carlo Simulations and Applications*

Simon Hug†
CIS, IPZ, Universität Zürich

First version: February 2005, this version: August 7, 2006

Abstract

The increased use of models with limited-dependent variables has allowed researchers to test important relationships in political science. Often, however, researchers employing such models fail to acknowledge that the violation of some basic assumptions has in part difference consequences in nonlinear models than in linear ones. In this paper I demonstrate this for probit models in which the dependent variable is systematically miscoded. Contrary to the linear model, such misclassifications affect not only the estimate of the intercept, but also those of the other coefficients. In a Monte-Carlo simulation I demonstrate that a model proposed by Hausman, Abrevaya and Scott-Morton (1998) allows for correcting these biases. Empirical examples based on re-analyses of models explaining the occurrence of rebellions and civil wars demonstrate the problem that comes from neglecting these misclassifications

* This paper draws in part on work carried out with Thomas Christin, whom I wish to express my gratitude for extremely helpful research assistance. Thanks are also due to James Fearon and Patrick Regan for making data used in this paper available and to Dominic Senn for helpful comments on an earlier version of this paper. Future revised versions of the paper as well as the data used in the analyses will be available on the web at http://www.ipz.unizh.ch/personal/hug

† Center for Comparative and International Studies; Institut für Politikwissenschaft; Universität Zürich; Hirschengraben 56; 8001 Zürich; Switzerland; phone +41 (0)44 634 50 90/1; fax: +41 (0)44 634 5098; email: hug@pwi.unizh.ch

1 Introduction

Research in political science has seen a considerable increase in the use of models with limited-dependent variables. Probit and logit models, even of the multinomial variety, have become the mainstay in many subfields, as have duration models, etc. When using such nonlinear models many scholars seem to neglect, however, that many problems which are inconsequential in the classical linear regression are much more serious in the nonlinear models. For instance, while the omission of variables in a linear regression fails to affect the estimated effect for the included variables as long as the former are uncorrelated with the latter, this does generally not hold in nonlinear models (see for instance Lee, 1982; Yatchew and Griliches, 1985).[1] Similarly, while in a linear model measurement error in the dependent variable only affects the precision with which the effect of our independent variables can be determined and possibly the estimate of the intercept, the same problem may bias our estimated effects in a nonlinear model (see Hausman, Abrevaya and Scott-Morton, 1998; Abrevaya and Hausman, 1999; Hausman, 2001).

Neglecting these issues in much research in political science is problematic. Quite clearly theories in political science are hardly sufficiently developed to guide us to completely specified empirical models to avoid the problem of misspecification.[2] Similarly, few are the situations in which we can be sure that our limited-dependent variable is measured without error. While the former problem is largely linked to the theoretical level and a series of specification tests exist for nonlinear models (see for instance Yatchew and Griliches, 1985), the latter problem relates much more to problems of measurement at the empirical level. In many contexts of political science research these measurement problems are, however, quite transparent, and all the same scholars refrain from considering them in earnest. Hence, in the present paper I discuss one particular type of measurement problem, namely misclassification in limited-dependent models in general and probit models in particular.

In the next section I state more formally the problem of misclassification and provide a series of examples where such misclassification is to be expected. In section three I discuss an estimator proposed by Hausman, Abrevaya and Scott-

[1] See also the more general discussion of omitted variable biases in Clarke (2005).

[2] **?** and Clarke (2005) discuss these problems in a more general context.

Morton (1998) to address the problem of misclassification in a probit setting. While these authors provide initial Monte Carlo simulations for their model, I extend their work to cover a broader range of situations to offer insights on when it is advisable to use their model to correct for misclassifications. In section four I provide an application of the empirical model demonstrating that taking into account misclassification may help avoid biases in our inferences in research on minorities at risk that engage in rebellion and on civil wars. Section five concludes.

2 Misclassifications in political science

In a classical linear regression framework miscodings and measurement error are part and parcel of the error term of the theoretical model. Hence, to assess the effect of miscodings and measurement error it suffices to evaluate the basic assumptions of the classical linear regression model. Three of the basic assumptions of the classical linear regression involve this error term (U_i) (e.g., Hanushek and Jackson, 1977; Gujarati, 1995, 60-63):

- $E(U_i) = 0 \forall i$
- $cov(U_i X_i) = 0 \ \forall i$
- $var(U_i) = \sigma^2 \ \forall i$

While violations of these assumptions from the classical linear regression model carry over more or less also to models of limited-dependent variables, violating the first one has more dramatic consequences. More precisely, while systematic measurement error leads to an expected value of the error term different from zero and thus a biased estimate of the constant term, in a nonlinear model, all our estimates become inconsistent (e.g., Hausman, Abrevaya and Scott-Morton, 1998; Hausman, 2001).

Considering the type of data that is often used in political science research in conjunction with models with limited-dependent variables, it is obvious that misclassifications and measurement error are paramount. For instance, Hausman, Abrevaya and Scott-Morton (1998) use as empirical example to illustrate their estimator for misclassification a model trying to explain job changes. As they

show with panel survey data, recall questions on job tenure often provide biased information. Hence, models attempting to estimate the effect of various factors on job change will be suffering from misclassification. If we compare such a rather central question in people's life with responses to survey questions often employed in political science research we can be sure that the problem of misclassification is widespread and the effects consequential.

Also in research not relying on survey data, misclassifications are likely. For instance, research into wars in general and civil wars in particular rely on the number of battle-deaths per year to decide whether a violent conflict is a war (or civil war) or not. An often employed rule is to consider as war (or civil war) a conflict with at least 1000 battle-deaths per year.[3] Hence, starting from a continuous indicator (number of battle-deaths) a dichotomous indicator is formed, which shows whether, for instance, two countries are at war (or a county is embroiled in a civil war). Under the hardly outlandish assumption that the underlying continuous indicator is measured with error, there is a strictly positive probability that a war is coded as a peaceful period or vice-versa.

Similarly and relatedly, if from a set of groups like the "minorities at risk" (MAR) information at the level of states is generated (e.g., presence or not of minorities), misclassifications are possible. More precisely, if the MAR data collection effort might have missed some groups (e.g. Hug, 2003) and this data is aggregated to the level of states, misclassifications will be the result.

Hence, misclassifications are very likely in much of political science research employing models with limited-dependent variables. Whether using survey data or data generated from continuous variables summarized in dichotomous indicators, misclassifications are likely to occur.

3 A model of misclassification and Monte Carlo simulations

To address the problem of misclassifications in a probit model, Hausman, Abrevaya and Scott-Morton (1998) propose an estimator that allows directly to correct for possible misclassifications. In both Monte Carlo simulations and empirical ex-

[3]In research on civil wars more recent work relies on a threshold of 25 battle-deaths (e.g. Gleditsch, Wallensteen, Eriksson, Sollenberg and Strand, 2002; Gates and Strand, 2004). Obviously, even at this lower level, measurement error is still possible, and misclassifications likely.

amples they demonstrate how even small amounts of misclassification affect the estimated coefficients, even if the misclassification is unrelated to any of the independent variables.[4] Their estimator relies on explicitly modeling in a *probit* setup the probability of misclassification. In a simple *probit*-model the log-likelihood function is simply

$$L(b|y,x) = \sum_{i=1}^{n}\{y_i ln\Phi(x_i'b) + (1-y_i)ln(1-\Phi(x_i'b))\} \quad (1)$$

where y is the observed dichotomous outcome, x a vector of explanatory variables and b the coefficients to be estimated. If a_0 corresponds to the probability that the unobserved $y_i = 0$ is classified as a 1 and a_1 corresponds to the probability that the unobserved $y_i = 1$ is classified as a 0, Hausman, Abrevaya and Scott-Morton (1998) derive the following log-likelihood function:

$$\begin{aligned} L(a_0,a_1,b|y,x) = \sum_{i=1}^{n}\{&y_i ln(a_0 + (1-a_0-a_1)\Phi(x_i'b)) \\ &+(1-y_i)ln(1-a_0-(1-a_0-a_1)\Phi(x_i'b))\} \end{aligned} \quad (2)$$

It is easy to see that equation 2 reduces to equation 1 if $a_0 = a_1 = 0$. Maximizing equation 2 yields estimates for the coefficients b but also for the amount of misclassification in the dataset through the values of a_0 and a_1. While Hausman, Abrevaya and Scott-Morton (1998) report estimates for a model employing this setup, they also suggest that both a_0 and a_1 may depend on some exogenous variables:

$$\begin{aligned} a_0 &= f(z_0) \\ a_1 &= f(z_1) \end{aligned} \quad (3)$$

As for the estimates of a_0 and a_1 in Hausman, Abrevaya and Scott-Morton's (1998) original formulation (equation 1), constraints need to be set such that these values remain in the interval $[0,1]$. As with regression models with dichotomous

[4]See Hausman (2001) for a more general discussion of mismeasured variables.

variables, the most convenient specification is either the logit transformation or the cumulative density function of the normal curve.[5]

What is also readily transparent is that the identification of the parameters to be estimated is only secured through the assumed functional form. More precisely, estimating the two additional parameters in equation 2 is only possible because they enter additively to then multiply the expression with the cumulative normal density. The same holds if as specified in equations 3 the misclassification probabilities are a function of an exogenous variable z. This variable may easily be part of the vector of explanatory variables of the probit model x, but again the parameters associated with equations 3 can only be estimated because the functional form differs from the way in which these explanatory variables affect the likelihood $y = 1$.

Despite this limitation Hausman, Abrevaya and Scott-Morton (1998) report encouraging results from Monte-Carlo simulations demonstrating that the proposed estimator performs much better than simple probit estimations in presence of misclassification. The equation they employ to generate the simulated dataset is the following:

$$\begin{aligned} y &= -1 + 0.2 \times x_1 + 1.5 \times x_2 - 0.6 \times x_3 + \epsilon \\ y^o &= 1 \text{ if } y > 0 \\ y^o &= 0 \text{ else} \end{aligned} \tag{4}$$

x_1 and ϵ are drawn from a normal distribution with mean 0 and variance 1, while x_2 and x_3 are random draws from a uniform distribution over the unit interval. A certain percentage, namely 2, 5, or 20 percent of the observed y^o (both 0s and 1s) were then randomly recoded. The simulations performed by Hausman, Abrevaya and Scott-Morton (1998) with a sample of 5000 observations then clearly show that the estimated coefficients taking into account the problem of missclassification come much closer to the true values.

Since these Monte-Carlo simulations are limited in several ways, I extend these simulations by using exactly the same setup as shown in equation 4. First, I carried out the Monte-Carlo simulations for smaller datasets, namely for samples of

[5]Below I also use the absolute value of the estimated parameter to ensure positive values. This, however, only works if no explanatory variables are used to explain the probability of misclassification.

1000, 2000, 3000, 4000, and 5000 observations. Second, while Hausman, Abrevaya and Scott-Morton (1998) kept the amount of missclassifications for both types at the same level in their simulations and only estimated one coefficient, I allow both coefficients in equation 2 to take on the three values reported above and in addition the value 0. For each possible permutation I then estimated the model both under the assumption that $a_0 = a_1$ and under the assumption that $a_0 \neq a_1$. Finally, since the proposed estimator also allows the amount of misclassification to depend on exogenous variables, I also carried out Monte-Carlo simulations with $a_0 = f(z_0)$ and $a_1 = f(z_1)$.

Figure 1: MC results: $a_0 = a_1$, one coefficient estimated

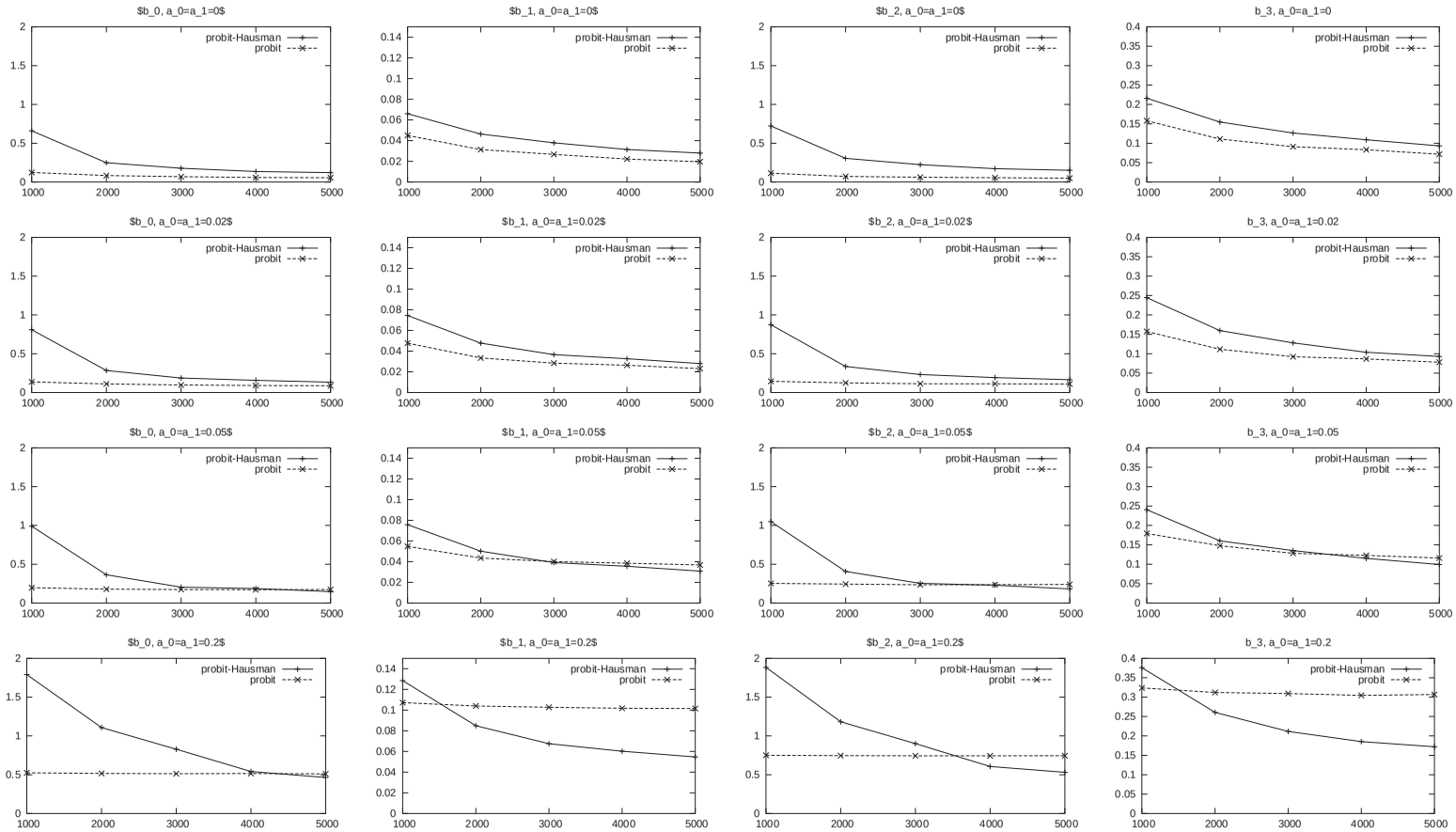


Figure 1 reports the first set of results for the simulations in which the two probabilities of misclassification a_0 and a_1 are set equal to each other and only one probability of misclassification is estimated.[6] For each estimated coefficient (see the four columns in figure 1) I depict the root of the mean-squared error $(rmse)$[7] both for a simple probit and the model proposed by Hausman, Abrevaya and Scott-Morton (1998). The rows in figure 1 correspond to the four different levels of misclassification assumed, namely 0, 0.02, 0.05 and 0.2. Not surprisingly, the $rmse$s increase when we move from the upper to the lower rows in figure 1. At the same time the $rmse$s of the model proposed by Hausman, Abrevaya and Scott-Morton (1998) become comparatively speaking better than the ones of the probit model. The various panels show also, however, that more generally the Hausman, Abrevaya and Scott-Morton's (1998) model becomes preferable to the simple probit model if the probability of misclassification is at least 0.05 (third and fourth row of panels in figure 1). Then, however, whether the $rmse$s of the probit model is higher or not depends on the sample size and the coefficient considered. Interestingly enough, while the $rmse$ of the intercept (b_0) and b_2 are systematically the largest, it is especially for the estimates of b_1 and b_3 that the correction proposed by Hausman, Abrevaya and Scott-Morton (1998) is a clear improvement, even for smaller sample sizes of 2000 observations or more.

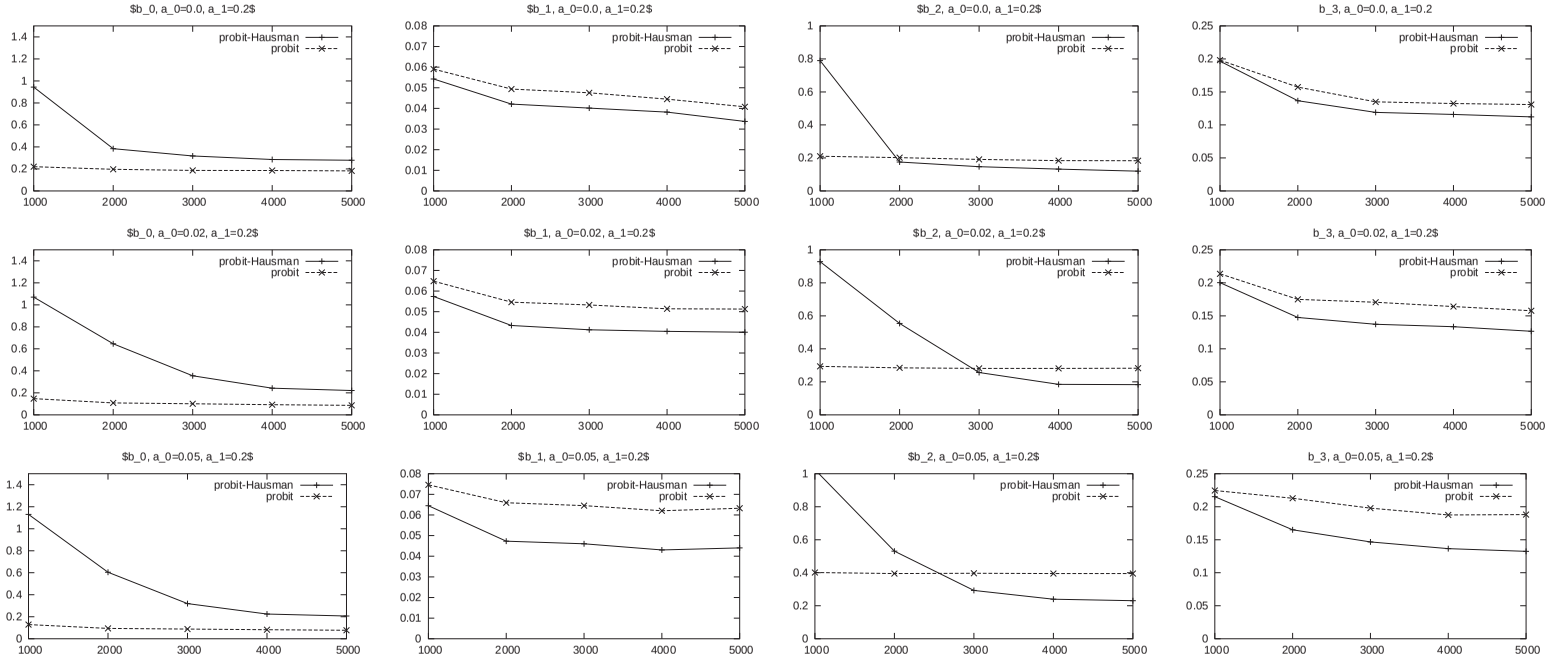
To assess the sensitivity of this estimator to other sets of probabilities of misclassifications I carried out Monte Carlo simulations for all possible combinations of the four values for a_0 and a_1. In almost all cases, when at least one of the two probabilities is at least 0.05, the $rmse$s, especially for larger sample sizes, are smaller for the constant term as estimated by the Hausman, Abrevaya and Scott-Morton (1998) than the one estimated by probit.[8] The advantage of this estimator becomes even more obvious if we look at cases where one of the misclassification probabilities, namely a_1, is equal to 0.2 (see figure 2).

[6]Estimating this model is not as straightforward as it seems, given that the parameters are only identified through the functional form. Convergence in the maximum likelihood estimations depends strongly on the starting values and is often difficult to achieve. While for all settings of the parameters 1000 datasets were drawn, the results presented here rely only on the set of estimations which converged. In the appendix I provide more details on the number of replications and the simulation results in general.

[7]The mean squared error is simply the variance of the estimated coefficient plus its bias squarred.

[8]Given that this result is of lesser significance I refrain from reporting it in more detail graphically, here.

Figure 2: MC results: $a_0 \neq a_1$, one parameter estimated



What is striking in the results depicted in figure 2 is that for two estimated coefficients, namely b_1 and b_3, independent of the sample size the $rmse$ for Hausman, Abrevaya and Scott-Morton's (1998) estimator is systematically smaller than the one for the probit estimator. On the other hand this is never the case for the $rmse$s for the constant b_0 and only for larger sample sizes for the remaining slope coefficient (b_2). This suggests that if at least one type of misclassification is rather important, then even estimating a model where it is assumed that both probabilities are equal can yield less biased estimates even in smaller samples.

Resorting to the exact same setup, namely letting vary the two probabilities of misclassification independently from each other across the four selected values, I estimated models where both probabilities were coefficients. If the two probabilities are identical, the $rmse$s for all coefficients from the probit estimates are systematically lower for the sample sizes considered in the Monte Carlo simulations. If the two misclassification probabilities differ from each other, the $rmse$s of Hausman, Abrevaya and Scott-Morton's (1998) estimator (mostly of the constant) beats the one of the probit model for large sample sizes as long as at least one of the probabilities exceeds the value of 0.02.[9]

To assess the estimator's performance when the probability of misclassification depends on an explanatory variable I used the following setup for either of the two probabilities:

$$a_{\cdot} \quad = \quad a_a \times (0.5 + x_1) + \theta \tag{5}$$

where a_a varied across the four values above and θ was drawn from $N(0, 1)$.[10]

[9]Given that these results are substantially less interesting I refrain from reporting them in detail here.

[10]Strictly speaking, this setup does not guarantee that $a_{\cdot} \in [0, 1]$, but the way in which the Monte Carlo simulations are set up, this fails to have an impact since values below or above the boundaries of the unit interval are implicitly brought to the closest boundary value.

Figure 3: MC results: a_1 as a function of z

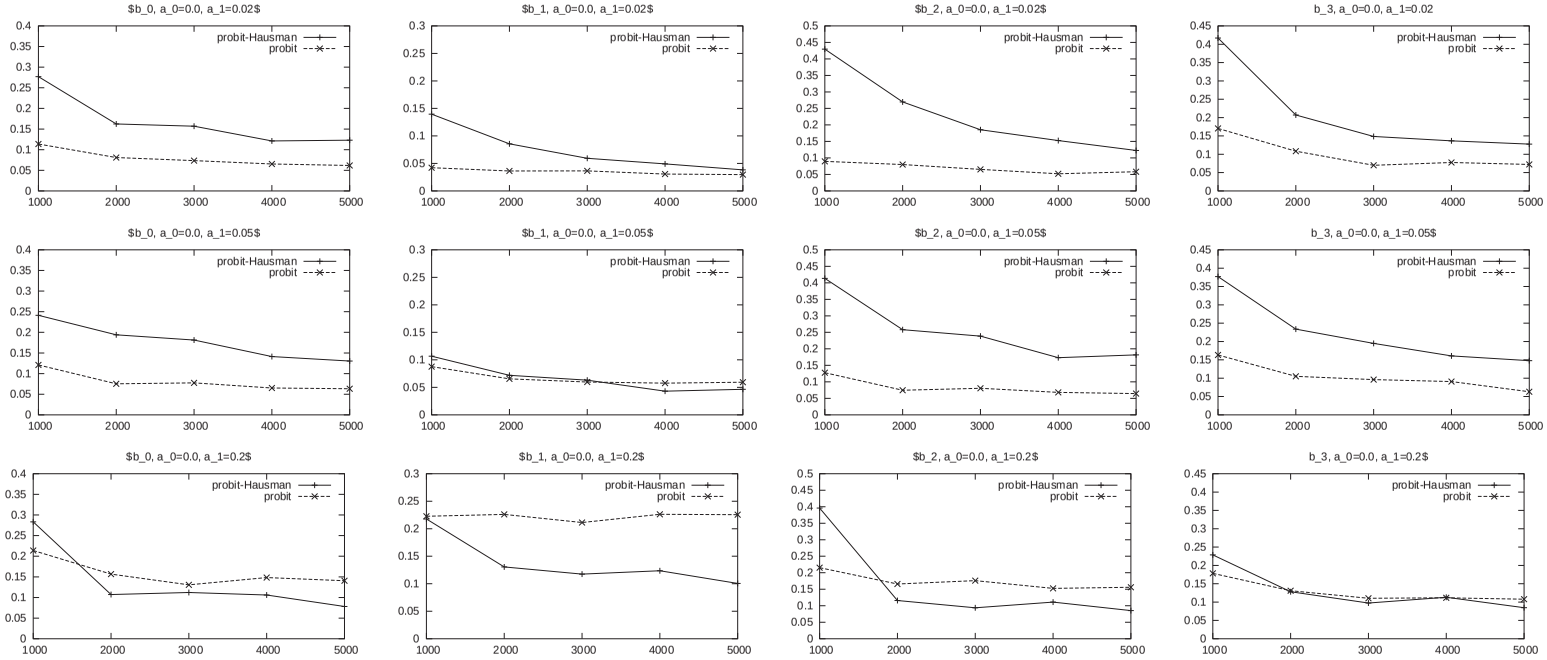
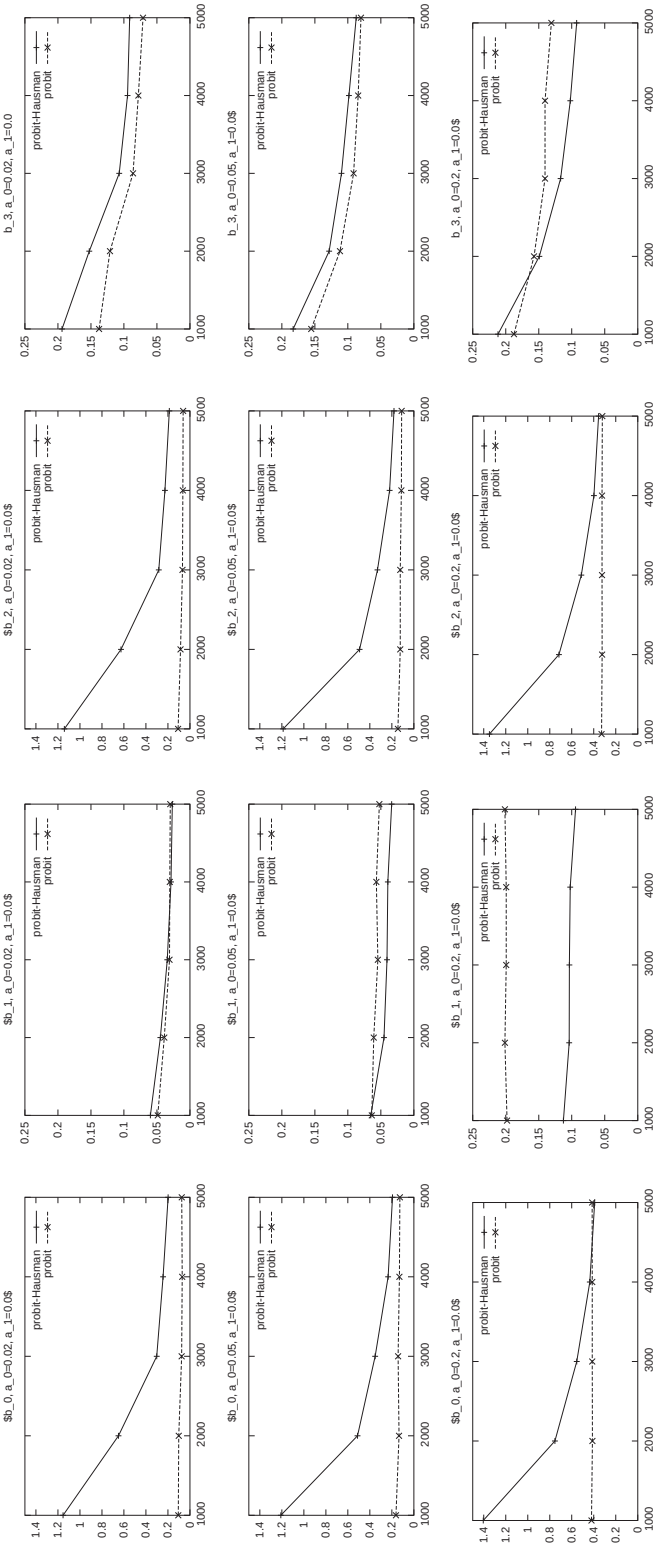


Figure 4: MC results: a_1 as a function of z



The various panels in figure 3 report the results for the cases where a_0 depends on x_1 as specificed in equation 5, and a_a takes on the three values used above, while a_a for a_1 is equal to 0. The results depicted in figure 4 are generated in the same fashion, but with a_0 and a_1 inversed. It is apparent in both figures that already with 5 percent misclassification the $rmse$s of Hausman, Abrevaya and Scott-Morton's (1998) estimator for some coefficients beats the ones of the simple probit model. If the amount of misclassification is rather large, the differences become even quite large and appear even for smaller sample sizes. Hence, even for many situations where we expect the probability of misclassification to depend on exogenous variables Hausman, Abrevaya and Scott-Morton's (1998) estimator provides improved estimates.

4 Empirical examples

To illustrate the performance of Hausman, Abrevaya and Scott-Morton's (1998) proposed estimator, I employ it on two studies dealing with rebellions and civil wars. The first study by Regan and Norton (2005) proposes an empirical model to assess how various factors influence the outbreak of protest, rebellions and civil wars. To test this empirical model they employ the "minorities at risk" data (MAR) (Gurr, 1993), aggregate it, however, to the level of country-years. More precisely, they create a summary indicator for each minority based on variables measuring protest and rebellious behavior in the MAR data,[11] and based on this code whether a minority is engaged in demonstrations, rebellions or a civil war. Aggregating this to the country level allows the authors to have a dichotomous indicator for each county-year showing whether a protest, rebellion, or civil war had occurred or not. As explanatory variables Regan and Norton (2005) use discrimination, political repression (lagged), extractable resources, per capita GDP, population size, regime type, and ethnolinguistic fractionalization. To account for possible time-dependencies, the authors follow Beck, Katz and Tucker (1998) and use cubic-splies as well as a counter for the number of years since the last event.[12]

[11] Regan and Norton (2005, 327) give detailed instructions on how they constructed this summary indicator as well as their three dichotomous variables for protest, rebellions, and civil war.

[12] Employing both a time-counter and cubic-splies is not exactly common practice, but since Regan and Norton (2005) employ it in their work, I follow their example.

While Regan and Norton (2005) estimate their model as a logit, I report the results of a probit model in column 1 of table 1 for the onset of a rebellion.[13] Substantively the results obviously fail to differ. Discrimination, per capita GDP, the log of the population size and ethnolinguistic fragmentation positively and statistically significantly affect the outbreak of rebellions. Repression decreases the probability of such an outbreak, though not statistically significantly, while the effect of democracy, as measured by the Polity IV scale, is curvilinear and statistically significant.

Table 1: Misclassification: Regan and Norton (2005)

	(1)	(2)	(3)	(4)	(4)
	probit	probit	probit	probit	probit
	b	b	b	b	b
variables	(s.e)	(s.e.)	(s.e.)	(s.e.)	(s.e.)
discrimination	0.267	0.323	0.276	0.329	0.519
	(0.026)	(0.039)	(0.027)	(0.035)	(0.076)
per capita GDP	0.251	0.512	0.277	0.276	0.451
	(0.054)	(0.090)	(0.059)	(0.066)	(0.145)
lagged political repression	-0.024	-0.088	-0.026	-0.070	-0.235
	(0.041)	(0.059)	(0.043)	(0.050)	(0.094)
extractable ressources	0.069	0.017	0.052	0.106	0.198
	(0.087)	(0.123)	(0.093)	(0.106)	(0.204)
log population size	0.134	0.199	0.145	0.168	0.351
	(0.027)	(0.041)	(0.029)	(0.036)	(0.081)
Polity IV democracy scale	0.17	0.289	0.189	0.203	0.382
	(0.033)	(0.050)	(0.035)	(0.039)	(0.081)
Polity IV democracy scale^2	-0.007	-0.012	-0.007	-0.008	-0.015
	(0.002)	(0.002)	(0.002)	(0.002)	(0.004)
ethnolinguistic fragmentation	0.004	0.005	0.003	0.005	0.007
	(0.001)	(0.002)	(0.002)	(0.002)	(0.004)
peaceyears	0.087	0.308	0.106	0.124	0.724
	(0.008)	(0.038)	(0.011)	(0.012)	(0.101)
spline1	0.012	0.024	0.013	0.014	0.044
	(0.001)	(0.003)	(0.001)	(0.002)	(0.006)
spline2	-0.013	-0.024	-0.014	-0.016	-0.041
	(0.002)	(0.003)	(0.002)	(0.002)	(0.006)
spline3	0.008	0.013	0.008	0.009	0.021
	(0.002)	(0.002)	(0.002)	(0.002)	(0.004)
constant	-5.671	-9.622	-6.192	-6.390	-12.278
	(0.613)	(1.158)	(0.684)	(0.740)	(1.740)
$\lvert a_0\rvert = \lvert a_1\rvert$		0.044			
		(0.008)			
$\lvert a_0\rvert$			0.010		0.027
			(0.005)		(0.005)
$\lvert a_1\rvert$				0.157	0.272
				(0.044)	(0.029)
log-likelihood	-766.624	-736.319	-761.547	-755.041	-697.530
n	2019	2019	2019	2019	2019

When allowing for the possibility of misclassification but assuming that the

[13] I estimated the same models also for the two other dependent variables used by Regan and Norton (2005), but refrain from reporting these results here. The reason for this omission is that the results reported here are the most illustrative for the effect of misclassification.

two probabilities take the same value (column 2 table 1), I find a sizeable probability of misclassification of 0.044.[14] The other estimated coefficients of the model also undergo some changes. These fail, however, to affect the substantive conclusions reached by Regan and Norton (2005). The most interesting changes are the doubling of the size of the coefficient for GDP per capita and the quadrupling of the coefficient for repression. The latter effect, given that the standard error increases less dramatically, almost reaches statistical significant.

As seen in the Monte Carlo simulations, estimating an identical probability of misclassification, even if the probabilites differ, is often advisable. Here, however, I also wish to check what happens if individual probabilities are estimated separately (columns 3 and 4 in table 1) or jointly (column 5 in table 1). In the case where only the probability that a peaceful year is miscoded as a year with a rebellion, this estimated probability is quite small, namely 0.01. As the Monte Carlo simulations suggested, with such small probabilities the efficiency gain of the Hausman, Abrevaya and Scott-Morton (1998) estimator is very small if existent at all. Hence, it hardly surprises that the changes in the estimates are vanishingly small and in no case affect the substantive conclusions. The probability that a year with a rebellion was miscoded as a peaceful year is considerably larger (column 4 in table 1), namely 0.157. Not surprisingly, several estimated coefficients for the substantively interesting variables approach the ones reported in column 2. Hence, again the effect of repression appears stronger and almost reaches statistical significance.

Finally, if both probabilities of misclassification are estimated separately in the same model (column 5 in table 1), I find still stronger changes. First of all, the two probabilities of misclassification are quite sizeable with the second one reaching 0.272. With regard to the coefficients for the substantive variables, quite a few notable changes appear. Discrimination appears to have a much strengthened effect when misclassification is taken into account, as is the case for the effect of per capita GDP. While in the original model the effect of political repression failed to reach statistical significance, this is no longer the case if misclassification is accounted for. Reversely, while the effect of ethnolinguistic fragmentation had a statistically significant effect in the original model, this is no

[14]For this estimation I used as specification the absolute value of the parameter to constrain the parameter to strictly positive values. In this particular instance this estimation strategy performed reasonably well.

longer the case when misclassification is considered. On the other hand the effect of discrimination is considerably strengthened, while the curvilinear relationship of the POLITY IV democracy scale is reduced. Hence, simply by accounting for the possibility of misclassifications in the dependent variable, some of the results of Regan and Norton's (2005) analysis are either strengthened or substantively changed. Quite clearly, then, accounting for misclassification is of considerable importance.

To illustrate the way in which explanatory variables for misclassification may affect results of empirical analyses I turn to the second example. Fearon and Laitin (2003) assess in a simple empirical model, how various explanatory factors contribute to explaining the onset of civil wars. For this they create a data-set where each observation corresponds to a country-year and the dependent variable takes the value of 1 if a civil war starts in a particular year.[15] As civil war is coded a violent conflict inside a state in which at least 1000 battle-deaths are deplored in one year. In table 2 (column 1) I first report a replication of Fearon and Laitin's (2003) base-model, which they estimate as a logit model, estimated as a probit model and with slightly updated data.[16]

In columns 2 and 3 of table 2 I report the results of estimations where one of the probabilities of misclassification is estimated with a single paramater of the cumulative normal density function. For both types of misclassification the estimated parameter is negative and quite large, which indicates that the probability of both types of misclassification is quite small (0.000000220905 respectively 0.0000607697).[17]

[15]Country-years in which a civil war is coded as ongoing are dropped from the analysis.

[16]I wish to thank James Fearon for making available this updated dataset with a few changes in the codings of some civil wars. The results hardly differ and no substantive conclusion is affected.

[17]These tiny probabilities combined with the fact that the likelihood function is actually smaller for the models reported in columns 2 and 3, compared to the model in column 1 suggests that the probability of missclassification is actually 0. This conjecture is confirmed if the misclassification parameter is not estimated as cumulative normal distribution but as absolute value. The estimated value is zero, but given this, the Hessian cannot be inverted. For this reason I refrain from reporting these results here.

Table 2: Misclassification: Fearon and Laitin (2003), updated data

	(1) probit b	(2) probit b	(3) probit b	(4) probit b	(5) probit b	(6) probit b	(7) probit b	(8) probit b
variables	(s.e)	(s.e)	(s.e.)	(s.e.)	(s.e.)	(s.e.)	(s.e.)	(s.e.)
Prior war	-0.373	-0.339	-0.339	-0.395	-0.465	-0.601	-0.450	-0.525
	(0.129)	(0.126)	(0.126)	(0.141)	(0.167)	(0.255)	(0.138)	(0.171)
Per capita income$_{t-1}$	-0.135	-0.131	-0.131	-0.129	0.073	0.272	-0.101	0.067
	(0.028)	(0.028)	(0.028)	(0.028)	(0.091)	(0.135)	(0.032)	(0.091)
log(population)	0.102	0.101	0.101	0.102	0.124	0.201	0.111	0.133
	(0.031)	(0.031)	(0.031)	(0.032)	(0.041)	(0.072)	(0.035)	(0.042)
log(mountainous	0.084	0.088	0.088	0.081	0.097	0.210	0.088	0.098
terrain)	(0.034)	(0.034)	(0.034)	(0.035)	(0.042)	(0.093)	(0.037)	(0.044)
Noncontiguous	0.210	0.200	0.199	0.223	0.290	0.391	0.405	0.515
state	(0.120)	(0.120)	(0.120)	(0.124)	(0.165)	(0.220)	(0.163)	(0.217)
Oil exporter	0.335	0.321	0.321	0.337	0.362	0.605	0.228	0.192
	(0.123)	(0.123)	(0.123)	(0.125)	(0.167)	(0.241)	(0.137)	(0.175)
New state	0.747	0.747	0.747	0.750	0.881	1.321	0.727	0.818
	(0.163)	(0.163)	(0.163)	(0.165)	(0.235)	(0.377)	(0.172)	(0.223)
Instability	0.260	0.251	0.251	0.233	0.334	0.482	0.273	0.336
	(0.101)	(0.101)	(0.101)	(0.110)	(0.133)	(0.227)	(0.106)	(0.134)
Democracy	0.007	0.006	0.006	0.008	0.008	0.012	0.009	0.009
(Polity)	(0.007)	(0.007)	(0.007)	(0.007)	(0.009)	(0.013)	(0.008)	(0.009)
Ethnic	0.109	0.128	0.129	0.148	0.126	0.207	0.021	0.037
fractionalization	(0.157)	(0.156)	(0.156)	(0.167)	(0.191)	(0.326)	(0.173)	(0.210)
Religious	0.091	0.072	0.070	0.008	0.149	0.453	-0.031	0.025
fractionalization	(0.207)	(0.206)	(0.209)	(0.336)	(0.257)	(0.462)	(0.216)	(0.282)
constant	-3.150	-3.151	-3.150	-3.142	-3.287	-5.235	-3.105	-3.310
	(0.301)	(0.300)	(0.300)	(0.320)	(0.390)	(1.078)	(0.330)	(0.406)
$\Phi(a_0)$		-4.869		-1.172		-1.926		
		(14.817)		(0.849)		(0.177)		
Per capita income				-4.243		-0.347		
				(5.303)		(0.180)		
$\Phi(a_1)$			-4.299		-0.251	-0.472	1.185	0.615
			(105.654)		(0.668)	(0.712)	(0.517)	(0.683)
Per capita income					0.335	0.406		0.284
					(0.088)	(0.101)		(0.093)
Eastern							-1.277	-0.592
Europe							(0.659)	(0.452)
Latin							-1.570	-0.651
America							(0.685)	(0.442)
Subsahara							-4.716	-1.159
Africa							(24.916)	(0.682)
Asia							-1.853	-0.691
							(0.702)	(0.436)
Northafrica							-4.293	-1.199
Middle East							(25.594)	(0.611)
log-likelihood	-486.231	-491.229	-491.228	-485.153	-482.951	-478.831	-481.841	-479.186
χ^2 model 1		-9.995	-9.994	2.156	6.561	14.800	8.781	14.090
df		1	1	2	2	4	6	7
p				0.340	0.038	0.005	0.186	0.050
χ^2 model 3					16.555		18.775	24.084
df					1		5	6
p					0		0.002	0.001
χ^2 model 5						8.239		7.529
df						2		4
p						0.016		0.110
n	6327	6327	6327	6327	6327	6327	6327	6327

Despite these small probabilites it might be the case that some systematic

features explain the probability of misclassification. To assess this I allow the probability of misclassification to depend on the GDP per capita. The argument for this is that reports on battle-deaths, which are used to determine whether a civil war occurs or not, are likely to be much more imprecise in poor countries than in rich ones. In column 4 the results appear for a model where the probability of a peaceful year to be miscoded as a year of civil-war onset is allowed to vary. The estimates suggest that this probability of misclassification decreases with higher GDPs, but this effect fails to reach statistical significance. A likelihood ratio test comparing this model to the one estimated by Fearon and Laitin (2003), confirms, that we cannot maintain the hypotheses that this probability of misclassification is related to the GDP and different from 0. When I allow the other probability of misclassification to vary as a function of GDP per capita, however, I find a statistically significant effect both for the estimated coefficient and the likelihood ratio test. The effect of GDP per capita is, however, counterintuitively positive. One explanation for this might be, that reports from poorer countries on battle-deaths are much more imprecise and exaggerate the number of casualties.

If I allow both probabilities of misclassification to vary with GDP per capita, both estimated coefficients for the latter variable are statistically significant. Their substantive effect remains, however, the same. In these two latter models the estimated coefficients of some of substantive variables also undergo some changes. The most notable, not completely unexpected, is the effect of the lagged GDP, which turns from negative to positive.[18]

As the previous analyses suggested, it is mostly the probability of misclassifying a civil war onset as a peaceful year that seems to matter, I propose two last models where this probability depends on the region to which a country belongs. Using as omitted category Western Europe, I estimate the effect of five dichotomous indicators on the misclassification probability. Contrary to expectation, all of these estimated coefficients are negative, suggesting that the probability of misclassification is highest in countries of Western Europe (column 7). Not surprisingly, when controlling in addition for GDP, these differences decrease quite dramatically. While the individual estimates for the various dichotomous indica-

[18]These results illustrate the limitation of the model, since the parameters are identified only through the assumed functional form. Given that the underlying theoretical model is hardly solidly specified, it remains debatable whether GDP per capita affects civil war onset or the likelihood of misclassification.

tors are quite imprecise, jointly they reach statistical significance as the likelihood ratio test shows. The substantively interesting results in these last two models is that the economic indicators loose much of their explanatory power. In the very last model GDP as no longer a statistically significant effect on civil war onset, and the same holds in the last two models for whether a country exports oil or not. While the diminished effect of GDP is certainly linked to the fact that its unlagged value appears as explanatory variable for the probability of misclassification, the results certainly question the predominance of economic variables in explaining civil war onsets.

5 Conclusion

Too often researchers in political science employing models for limited-dependent variables fail to acknowledge that violations of assumptions that are rather innocuous in the classical linear regression model may have much more dramatic effects. It is (should be) well known that the effect of omitted variables is quite different in nonlinear models than in linear ones. Similarly, measurement error, or misclassification in limited-dependent variables affect in most cases all estimated coefficients, even in the most innocuously looking cases.

In this paper I discussed various cases in which we would expect misclassifications and presented a model proposed by Hausman, Abrevaya and Scott-Morton (1998) which allows to address this problem in probit models. In Monte Carlo simulations I was able to demonstrate that, provided that a researcher works with a sizeable sample, the corrections proposed by Hausman, Abrevaya and Scott-Morton (1998) clearly outperform a simple probit estimation. This even holds if the amount of misclassification is rather limited. Similarly, the Monte Carlo simulations suggest that even if the two possible probabilities of misclassification differ, a joint estimation under the assumption that they are equal is often an improvement over probit estimates. The same also holds for situations where we expect exogenous variables to affect the likelihood of misclassification.

I illustrated the estimator discussed in two empirical examples related to rebellions and civil wars. In both cases addressing the issue of possible misclassification suggested that systematic measurement seems present in both cases. In addition, the corrections changed some of the substantive results of the original analyses. Combined with the insights from the Monte Carlo study this suggests

that researchers should pay much more attention to this potential problem. As I noted in the paper, in many areas where political scientists employ models for limited-dependent variables, misclassifications are very likely.

Appendix

In table 3 I report the descriptive statistics for the example based on Regan and Norton (2005), while table 4 does the same for the analysis based on Fearon and Laitin (2003). Tables 5-9 report the results of the Monte Carlo simulations ($rmse$s) on which the figures in the main text are based.

Table 3: Descriptive statistics for reanalyses of Regan and Norton (2005)

Variable	Min	Mean	Max	Std. Dev.	n
Rebellion	0	0.245	1	0.430	2019
Discrimination	0	1.970	4	1.702	2019
Per capita income	5.737	8.107	9.771	0.861	2019
Repression$_{t-1}$	1	2.383	9	1.147	2019
Extractables	0	0.288	1	0.453	2019
Log population	12.319	16.169	20.918	1.464	2019
Democracy	0	10.752	20	7.712	2019
Democracy^2	0	175.076	400	169.639	2019
Ethnolinguistic fractionalization	1	42.631	93	29.039	2019

Table 4: Descriptive statistics Fearon and Laitin (2003)

Variable	Min	Mean	Max	Std. dev.	n
Civil War onset	0	0.02	1	0.13	6610
Prior war	0	0.14	1	0.34	6610
Per capita income$_{t-1}$	0.05	3.65	66.74	4.54	6373
log(population)	5.40	9.05	14.03	1.46	6585
log(% mountainous terrain)	0	2.18	4.56	1.40	6610
Noncontiguous state	0	0.17	1	0.38	6610
Oil exporter	0	0.13	1	0.34	6610
New state	0	0.03	1	0.17	6610
Instability	0	0.15	1	0.35	6596
Democracy (Polity)	-10	-0.48	10	7.51	6541
Ethnic fractionalization	0	0.39	0.93	0.29	6610
Religious fractionalization	0	0.37	0.78	0.22	6610

Table 5: RMSE for estimates under the assumption $a_0 = a_1$

	1000		2000		3000		4000		5000	
	Hausman	probit	Hausman	probit	Hausman	probit	Hausman	probit	Hausman	probit
					$a_1 = a_0 = 0$					
n	463		467		493		463		483	
b_0	0.660	0.123	0.249	0.083	0.177	0.069	0.135	0.059	0.120	0.055
b_1	0.066	0.045	0.046	0.031	0.038	0.027	0.031	0.022	0.028	0.020
b_2	0.722	0.114	0.305	0.073	0.224	0.062	0.172	0.055	0.151	0.050
b_3	0.216	0.158	0.155	0.111	0.127	0.091	0.109	0.083	0.093	0.072
$a_1 = a_0$	0.164		0.171		0.178		0.181		0.183	
					$a_0 = 0.0, a_1 = 0.02$					
n	426		446		441		427		435	
b_0	0.729	0.119	0.244	0.087	0.172	0.072	0.160	0.060	0.142	0.057
b_1	0.061	0.044	0.044	0.032	0.033	0.026	0.030	0.023	0.024	0.020
b_2	0.769	0.104	0.268	0.080	0.178	0.066	0.167	0.054	0.146	0.050
b_3	0.220	0.156	0.138	0.107	0.114	0.093	0.091	0.078	0.089	0.074
$a_1 = a_0$	0.165		0.173		0.178		0.181		0.182	
					$a_0 = 0.0, a_1 = 0.05$					
n	413		401		404		389		408	
b_0	0.716	0.119	0.332	0.094	0.215	0.082	0.181	0.080	0.160	0.068
b_1	0.060	0.044	0.039	0.031	0.032	0.028	0.025	0.024	0.024	0.022
b_2	0.718	0.117	0.319	0.087	0.182	0.075	0.143	0.069	0.119	0.069
b_3	0.200	0.156	0.136	0.116	0.101	0.091	0.090	0.085	0.077	0.076
$a_1 = a_0$	0.163		0.172		0.178		0.182		0.183	
					$a_0 = 0.0, a_1 = 0.2$					
n	328		305		258		235		207	
b_0	0.944	0.220	0.384	0.197	0.318	0.186	0.286	0.186	0.279	0.182
b_1	0.054	0.059	0.042	0.049	0.040	0.048	0.038	0.045	0.034	0.041
b_2	0.790	0.210	0.175	0.202	0.147	0.191	0.132	0.184	0.120	0.183
b_3	0.197	0.198	0.137	0.157	0.119	0.135	0.116	0.132	0.112	0.131
$a_1 = a_0$	0.168		0.177		0.182		0.186		0.185	
					$a_0 = 0.02, a_1 = 0.0$					
n	568		646		705		739		743	
b_0	0.848	0.147	0.371	0.119	0.179	0.108	0.144	0.103	0.129	0.103
b_1	0.077	0.048	0.051	0.031	0.041	0.027	0.035	0.024	0.033	0.021
b_2	0.919	0.134	0.434	0.108	0.250	0.098	0.206	0.090	0.190	0.092
b_3	0.240	0.156	0.159	0.109	0.130	0.087	0.116	0.081	0.109	0.077
$a_1 = a_0$	0.149		0.157		0.163		0.167		0.167	
					$a_1 = a_0 = 0.02$					
n	574		621		669		691		725	
b_0	0.810	0.137	0.284	0.111	0.187	0.097	0.157	0.091	0.134	0.088
b_1	0.074	0.048	0.048	0.033	0.037	0.028	0.033	0.026	0.028	0.023
b_2	0.874	0.144	0.336	0.125	0.232	0.113	0.192	0.112	0.166	0.110
b_3	0.245	0.157	0.160	0.111	0.128	0.092	0.104	0.087	0.093	0.078
$a_1 = a_0$	0.149		0.159		0.165		0.167		0.170	
					$a_0 = 0.02, a_1 = 0.05$					
n	589		603		613		657		682	
b_0	0.952	0.118	0.309	0.092	0.195	0.079	0.154	0.077	0.131	0.073
b_1	0.065	0.048	0.042	0.039	0.034	0.033	0.028	0.028	0.025	0.027
b_2	0.988	0.165	0.322	0.150	0.195	0.142	0.158	0.140	0.130	0.140
b_3	0.209	0.160	0.141	0.120	0.115	0.111	0.097	0.091	0.081	0.085
$a_1 = a_0$	0.149		0.162		0.166		0.170		0.171	
					$a_0 = 0.02, a_1 = 0.2$					
n	483		460		460		488		481	
b_0	1.070	0.148	0.646	0.109	0.356	0.101	0.242	0.092	0.221	0.086
b_1	0.057	0.065	0.043	0.055	0.041	0.053	0.040	0.051	0.040	0.051
b_2	0.928	0.294	0.554	0.285	0.257	0.282	0.185	0.281	0.183	0.283
b_3	0.200	0.214	0.148	0.175	0.137	0.171	0.134	0.164	0.127	0.158
$a_1 = a_0$	0.156		0.166		0.170		0.175		0.177	

Table 6: RMSE for estimates under the assumption $a_0 = a_1$

	1000		2000		3000		4000		5000	
	Hausman	probit	Hausman	probit	Hausman	probit	Hausman	probit	Hausman	probit
					$a_0 = 0.05, a_1 = 0.0$					
n	717		795		847		893		918	
b_0	0.922	0.229	0.420	0.213	0.193	0.210	0.166	0.213	0.143	0.209
b_1	0.093	0.048	0.064	0.037	0.049	0.033	0.045	0.030	0.039	0.028
b_2	1.056	0.201	0.523	0.190	0.305	0.187	0.267	0.185	0.238	0.183
b_3	0.284	0.157	0.191	0.124	0.162	0.108	0.147	0.097	0.132	0.089
$a_1 = a_0$	0.129		0.135		0.141		0.142		0.144	
					$a_0 = 0.05, a_1 = 0.02$					
n	714		772		844		874		907	
b_0	0.974	0.218	0.327	0.206	0.189	0.197	0.157	0.197	0.151	0.193
b_1	0.078	0.051	0.056	0.039	0.046	0.035	0.038	0.034	0.036	0.031
b_2	1.074	0.228	0.416	0.209	0.271	0.206	0.238	0.204	0.218	0.207
b_3	0.261	0.168	0.181	0.124	0.143	0.114	0.132	0.100	0.109	0.100
$a_1 = a_0$	0.129		0.138		0.143		0.143		0.146	
					$a_1 = a_0 = 0.05$					
n	677		774		816		850		871	
b_0	0.990	0.198	0.364	0.178	0.205	0.174	0.187	0.172	0.147	0.174
b_1	0.076	0.055	0.050	0.043	0.039	0.040	0.036	0.038	0.031	0.037
b_2	1.049	0.253	0.407	0.244	0.254	0.235	0.230	0.237	0.182	0.240
b_3	0.241	0.179	0.160	0.148	0.135	0.128	0.115	0.123	0.099	0.116
$a_1 = a_0$	0.131		0.140		0.144		0.146		0.150	
					$a_0 = 0.05, a_1 = 0.2$					
n	580		601		659		713		725	
b_0	1.130	0.129	0.605	0.094	0.320	0.088	0.225	0.082	0.207	0.077
b_1	0.064	0.075	0.047	0.066	0.046	0.065	0.043	0.062	0.044	0.063
b_2	1.025	0.402	0.531	0.395	0.293	0.397	0.240	0.395	0.230	0.395
b_3	0.215	0.225	0.165	0.213	0.147	0.198	0.136	0.188	0.132	0.188
$a_1 = a_0$	0.137		0.144		0.153		0.156		0.157	
					$a_0 = 0.2, a_1 = 0.0$					
n	723		827		898		919		951	
b_0	1.410	0.635	0.744	0.632	0.619	0.630	0.470	0.629	0.375	0.631
b_1	0.345	0.073	0.189	0.065	0.165	0.063	0.146	0.062	0.138	0.062
b_2	2.341	0.498	1.240	0.500	1.022	0.495	0.806	0.494	0.681	0.496
b_3	1.347	0.222	0.627	0.207	0.526	0.194	0.457	0.190	0.431	0.188
$a_1 = a_0$	0.065		0.060		0.057		0.054		0.052	
					$a_0 = 0.2, a_1 = 0.02$					
n	777		861		908		946		964	
b_0	1.542	0.619	0.689	0.619	0.527	0.621	0.438	0.622	0.379	0.620
b_1	0.303	0.076	0.185	0.071	0.166	0.067	0.144	0.068	0.137	0.065
b_2	2.280	0.525	1.198	0.520	0.966	0.523	0.814	0.525	0.710	0.523
b_3	0.955	0.243	0.622	0.213	0.513	0.207	0.466	0.204	0.432	0.201
$a_1 = a_0$	0.064		0.063		0.055		0.054		0.049	
					$a_0 = 0.2, a_1 = 0.05$					
n	800		879		928		964		973	
b_0	1.484	0.604	0.919	0.604	0.598	0.602	0.394	0.602	0.396	0.601
b_1	0.331	0.082	0.191	0.075	0.146	0.074	0.136	0.073	0.126	0.072
b_2	2.295	0.564	1.419	0.563	0.981	0.563	0.767	0.564	0.727	0.563
b_3	1.009	0.250	0.603	0.231	0.466	0.227	0.421	0.223	0.389	0.221
$a_1 = a_0$	0.064		0.059		0.054		0.049		0.047	
					$a_1 = a_0 = 0.2$					
n	729		841		892		936		957	
b_0	1.791	0.524	1.108	0.516	0.829	0.514	0.541	0.515	0.465	0.511
b_1	0.128	0.107	0.085	0.104	0.067	0.103	0.060	0.102	0.055	0.101
b_2	1.884	0.751	1.183	0.746	0.901	0.744	0.606	0.744	0.532	0.745
b_3	0.376	0.324	0.261	0.312	0.212	0.309	0.185	0.305	0.172	0.307
$a_1 = a_0$	0.065		0.059		0.060		0.052		0.051	

Table 7: RMSE for estimates under the assumption $a_0 \neq a_1$

	1000		2000		3000		4000		5000	
	Hausman	probit	Hausman	probit	Hausman	probit	Hausman	probit	Hausman	probit
					$a_0 = a_1 = 0.0$					
n	368		401		381		388		386	
b_0	0.947	0.113	0.382	0.086	0.165	0.066	0.133	0.059	0.119	0.056
b_1	0.229	0.044	0.134	0.032	0.113	0.027	0.090	0.022	0.081	0.019
b_2	1.495	0.104	0.735	0.082	0.534	0.062	0.450	0.055	0.406	0.048
b_3	0.793	0.144	0.427	0.114	0.340	0.086	0.276	0.076	0.239	0.069
a_0	0.054		0.042		0.036		0.031		0.029	
a_1	0.228		0.187		0.172		0.153		0.146	
					$a_0 = 0.0, a_1 = 0.02$					
n	378		361		391		396		406	
b_0	0.956	0.111	0.291	0.083	0.175	0.066	0.150	0.062	0.124	0.060
b_1	0.254	0.045	0.141	0.031	0.109	0.026	0.091	0.022	0.086	0.019
b_2	1.754	0.103	0.738	0.074	0.531	0.062	0.455	0.060	0.423	0.048
b_3	1.078	0.156	0.481	0.105	0.332	0.088	0.274	0.078	0.253	0.070
a_0	0.054		0.040		0.036		0.033		0.029	
a_1	0.221		0.189		0.171		0.151		0.147	
					$a_0 = 0, a_1 = 0.05$					
n	381		425		397		412		434	
b_0	0.924	0.130	0.326	0.093	0.181	0.086	0.144	0.075	0.139	0.066
b_1	0.270	0.044	0.144	0.035	0.111	0.027	0.094	0.025	0.082	0.024
b_2	1.695	0.121	0.754	0.088	0.536	0.080	0.464	0.073	0.422	0.070
b_3	0.984	0.155	0.471	0.117	0.330	0.097	0.296	0.084	0.255	0.074
a_0	0.054		0.040		0.035		0.032		0.031	
a_1	0.227		0.185		0.160		0.151		0.140	
					$a_0 = 0.0, a_1 = 0.2$					
n	383		467		462		488		454	
b_0	1.160	0.219	0.307	0.196	0.294	0.192	0.182	0.187	0.158	0.184
b_1	0.382	0.058	0.170	0.049	0.134	0.046	0.105	0.044	0.094	0.045
b_2	2.082	0.208	0.812	0.203	0.677	0.190	0.520	0.186	0.477	0.188
b_3	1.306	0.190	0.532	0.161	0.422	0.142	0.326	0.140	0.297	0.133
a_0	0.046		0.035		0.031		0.028		0.026	
a_1	0.197		0.167		0.154		0.137		0.130	
					$a_0 = 0.02, a_1 = 0.0$					
n	410		418		441		442		490	
b_0	0.706	0.136	0.265	0.113	0.182	0.108	0.155	0.106	0.129	0.101
b_1	0.288	0.045	0.141	0.033	0.106	0.028	0.087	0.023	0.077	0.021
b_2	1.558	0.125	0.711	0.101	0.546	0.097	0.430	0.095	0.386	0.090
b_3	0.979	0.146	0.428	0.117	0.348	0.088	0.267	0.081	0.234	0.071
a_0	0.054		0.043		0.036		0.032		0.028	
a_1	0.228		0.187		0.162		0.143		0.136	
					$a_0 = a_1 = 0.02$					
n	431		473		459		462		497	
b_0	0.981	0.129	0.432	0.110	0.195	0.098	0.158	0.089	0.144	0.088
b_1	0.260	0.047	0.140	0.032	0.097	0.028	0.088	0.026	0.079	0.024
b_2	1.763	0.138	0.791	0.126	0.504	0.114	0.449	0.111	0.393	0.112
b_3	0.976	0.156	0.426	0.118	0.312	0.094	0.266	0.090	0.240	0.079
a_0	0.054		0.042		0.036		0.031		0.029	
a_1	0.220		0.177		0.153		0.146		0.133	
					$a_0 = 0.02, a_1 = 0.05$					
n	445		499		485		552		527	
b_0	0.846	0.123	0.371	0.094	0.189	0.082	0.172	0.074	0.147	0.069
b_1	0.225	0.049	0.136	0.036	0.114	0.032	0.094	0.029	0.081	0.029
b_2	1.468	0.163	0.755	0.149	0.562	0.144	0.483	0.138	0.415	0.136
b_3	0.906	0.168	0.441	0.118	0.372	0.106	0.292	0.097	0.248	0.090
a_0	0.054		0.039		0.036		0.032		0.029	
a_1	0.210		0.174		0.160		0.142		0.131	
					$a_0 = 0.02, a_1 = 0.2$					
n	446		523		542		559		577	
b_0	1.293	0.144	0.566	0.108	0.342	0.096	0.209	0.095	0.172	0.092
b_1	0.344	0.064	0.174	0.058	0.126	0.053	0.109	0.052	0.087	0.052
b_2	2.335	0.296	1.023	0.286	0.688	0.284	0.550	0.280	0.443	0.281
b_3	1.575	0.211	0.598	0.170	0.406	0.162	0.344	0.160	0.283	0.159
a_0	0.048		0.037		0.032		0.029		0.026	
a_1	0.192		0.165		0.142		0.135		0.122	

Table 8: RMSE for estimates under the assumption $a_0 \neq a_1$

	1000		2000		3000		4000		5000	
	Hausman	probit	Hausman	probit	Hausman	probit	Hausman	probit	Hausman	probit
					$a_0 = 0.05, a_1 = 0.0$					
n	427		491		480		485		510	
b_0	1.128	0.224	0.476	0.218	0.302	0.215	0.186	0.211	0.163	0.209
b_1	0.306	0.047	0.146	0.038	0.108	0.033	0.088	0.032	0.084	0.029
b_2	2.050	0.197	0.827	0.193	0.607	0.186	0.464	0.185	0.431	0.182
b_3	1.245	0.170	0.432	0.120	0.364	0.106	0.298	0.092	0.255	0.089
a_0	0.057		0.045		0.038		0.034		0.032	
a_1	0.222		0.180		0.159		0.144		0.136	
					$a_0 = 0.05, a_1 = 0.02$					
n	457		501		532		521		543	
b_0	1.129	0.208	0.554	0.201	0.295	0.194	0.182	0.195	0.174	0.193
b_1	0.241	0.051	0.157	0.039	0.114	0.036	0.095	0.033	0.087	0.031
b_2	1.820	0.228	0.968	0.209	0.623	0.204	0.473	0.206	0.446	0.203
b_3	0.925	0.170	0.504	0.128	0.351	0.116	0.297	0.106	0.258	0.100
a_0	0.056		0.045		0.039		0.034		0.032	
a_1	0.210		0.176		0.160		0.140		0.133	
					$a_0 = a_1 = 0.05$					
n	473		518		518		541		577	
b_0	1.328	0.196	0.501	0.187	0.214	0.171	0.186	0.173	0.178	0.167
b_1	0.366	0.052	0.154	0.046	0.110	0.040	0.095	0.036	0.080	0.037
b_2	2.263	0.248	0.880	0.249	0.559	0.236	0.484	0.237	0.413	0.236
b_3	1.224	0.174	0.481	0.137	0.358	0.125	0.283	0.120	0.243	0.119
a_0	0.052		0.044		0.036		0.034		0.030	
a_1	0.210		0.171		0.151		0.136		0.126	
					$a_0 = 0.05, a_1 = 0.2$					
n	492		556		613		634		674	
b_0	1.611	0.131	0.484	0.097	0.338	0.088	0.289	0.078	0.209	0.077
b_1	0.479	0.072	0.189	0.068	0.134	0.065	0.110	0.063	0.095	0.063
b_2	2.844	0.400	0.987	0.393	0.752	0.395	0.611	0.391	0.485	0.397
b_3	1.594	0.220	0.610	0.205	0.522	0.191	0.350	0.192	0.298	0.190
a_0	0.047		0.039		0.034		0.030		0.029	
a_1	0.188		0.155		0.140		0.132		0.123	
					$a_0 = 0.2, a_1 = 0.0$					
n	476		502		514		533		523	
b_0	1.577	0.648	0.887	0.640	0.603	0.632	0.325	0.634	0.402	0.636
b_1	0.363	0.071	0.175	0.065	0.133	0.063	0.112	0.062	0.102	0.062
b_2	2.420	0.504	1.265	0.502	0.927	0.495	0.624	0.498	0.650	0.498
b_3	1.178	0.215	0.551	0.201	0.399	0.196	0.347	0.190	0.323	0.184
a_0	0.064		0.053		0.047		0.041		0.039	
a_1	0.192		0.163		0.142		0.131		0.126	
					$a_0 = 0.2, a_1 = 0.02$					
n	458		505		542		523		526	
b_0	1.646	0.626	0.980	0.625	0.638	0.621	0.492	0.621	0.305	0.619
b_1	0.524	0.072	0.188	0.070	0.133	0.070	0.115	0.068	0.099	0.065
b_2	4.448	0.527	1.430	0.524	0.929	0.526	0.765	0.523	0.564	0.518
b_3	3.605	0.238	0.625	0.213	0.416	0.210	0.373	0.203	0.303	0.199
a_0	0.063		0.052		0.046		0.042		0.039	
a_1	0.186		0.157		0.141		0.130		0.115	
					$a_0 = 0.2, a_1 = 0.05$					
n	479		526		567		591		557	
b_0	1.968	0.610	1.183	0.606	0.676	0.607	0.426	0.598	0.310	0.603
b_1	0.611	0.081	0.190	0.076	0.152	0.074	0.121	0.073	0.102	0.072
b_2	4.030	0.565	1.646	0.563	1.050	0.563	0.739	0.562	0.578	0.562
b_3	2.793	0.243	0.639	0.234	0.474	0.219	0.376	0.229	0.327	0.216
a_0	0.062		0.053		0.048		0.041		0.037	
a_1	0.177		0.148		0.139		0.125		0.115	
					$a_0 = a_1 = 0.2$					
n	451		571		607		628		704	
b_0	9.257	0.521	1.251	0.520	0.805	0.511	0.714	0.514	0.622	0.511
b_1	9.970	0.103	0.278	0.100	0.201	0.103	0.242	0.101	0.139	0.101
b_2	54.519	0.743	1.832	0.747	1.288	0.743	1.435	0.743	0.927	0.743
b_3	31.885	0.320	0.941	0.302	0.661	0.311	0.846	0.303	0.418	0.306
a_0	0.058		0.055		0.045		0.047		0.041	
a_1	0.157		0.136		0.131		0.124		0.117	

Table 9: RMSE for estimates with a_0 or a_1 as a function of x_1

	1000		2000		3000		4000		5000	
	Hausman	probit	Hausman	probit	Hausman	probit	Hausman	probit	Hausman	probit
	$aa_0 * (0.5 + x1[j])0.02000, a_1 = 0.0$									
n	162		165		145		143		170	
b_0	1.154	0.106	0.649	0.101	0.302	0.076	0.244	0.071	0.199	0.075
b_1	0.060	0.049	0.045	0.039	0.034	0.031	0.029	0.031	0.027	0.030
b_2	1.141	0.107	0.626	0.085	0.283	0.067	0.227	0.065	0.188	0.062
b_3	0.194	0.138	0.153	0.121	0.107	0.086	0.095	0.078	0.091	0.071
aa_0	0.059		0.050		0.038		0.036		0.031	
aa_1	0.031		0.020		0.016		0.014		0.012	
	$aa_0 * (0.5 + x1[j])0.050000, a_1 = 0.0$									
n	191		184		206		213		229	
b_0	1.209	0.162	0.512	0.135	0.352	0.143	0.234	0.131	0.195	0.127
b_1	0.065	0.064	0.045	0.061	0.041	0.055	0.039	0.057	0.034	0.052
b_2	1.188	0.145	0.493	0.123	0.331	0.124	0.221	0.114	0.181	0.111
b_3	0.183	0.155	0.128	0.112	0.110	0.091	0.098	0.084	0.087	0.080
aa_0	0.040		0.047		0.042		0.037		0.033	
aa_1	0.027		0.021		0.020		0.020		0.019	
	$aa_0 * (0.5 + x1[j])0.200000, a_1 = 0.0$									
n	178		310		362		345		343	
b_0	1.404	0.420	0.753	0.412	0.555	0.414	0.436	0.414	0.392	0.415
b_1	0.113	0.198	0.104	0.201	0.104	0.199	0.102	0.199	0.094	0.201
b_2	1.348	0.329	0.716	0.324	0.513	0.325	0.399	0.325	0.356	0.322
b_3	0.212	0.188	0.149	0.157	0.117	0.140	0.102	0.140	0.093	0.131
aa_0	0.076		0.065		0.061		0.058		0.060	
aa_1	0.076		0.072		0.078		0.078		0.074	
	$a_0 = 0, aa_1 * (0.5 + x1[j])0.020000$									
n	46		61		47		54		71	
b_0	0.277	0.114	0.162	0.081	0.157	0.074	0.121	0.065	0.123	0.062
b_1	0.140	0.042	0.086	0.036	0.059	0.036	0.049	0.031	0.038	0.030
b_2	0.430	0.090	0.270	0.080	0.185	0.065	0.153	0.052	0.123	0.058
b_3	0.417	0.170	0.207	0.108	0.149	0.070	0.137	0.077	0.128	0.072
aa_0	0.174		0.139		0.115		0.092		0.088	
aa_1	0.029		0.029		0.023		0.017		0.017	
	$a_0 = 0, aa_1 * (0.5 + x1[j])0.050000$									
n	64		66		78		79		76	
b_0	0.241	0.121	0.194	0.075	0.181	0.077	0.141	0.065	0.131	0.063
b_1	0.107	0.088	0.072	0.066	0.063	0.060	0.043	0.058	0.046	0.059
b_2	0.413	0.128	0.258	0.075	0.238	0.080	0.173	0.068	0.182	0.064
b_3	0.377	0.163	0.234	0.105	0.195	0.096	0.161	0.091	0.148	0.063
aa_0	0.159		0.140		0.142		0.105		0.106	
aa_1	0.043		0.030		0.029		0.028		0.025	
	$a_0 = 0, aa_1 * (0.5 + x1[j])0.200000$									
n	13		34		16		92		55	
b_0	0.283	0.214	0.107	0.157	0.112	0.131	0.106	0.148	0.078	0.141
b_1	0.218	0.223	0.131	0.226	0.118	0.211	0.124	0.226	0.101	0.226
b_2	0.396	0.215	0.116	0.166	0.094	0.176	0.111	0.153	0.085	0.156
b_3	0.229	0.178	0.128	0.131	0.098	0.110	0.113	0.112	0.085	0.108
aa_0	0.196		0.070		0.071		0.072		0.060	
aa_1	0.163		0.049		0.045		0.051		0.036	

References

Abrevaya, Jason and Jerry Hausman. 1999. "Semiparametric Estimation with Mismeasured Dependent Variables: An Application to Duration Models for Unemployment Spells." *Annales d'économie et de statistiques* 55-56:243–275.

Beck, Nathaniel, Jonathan Katz and Richard Tucker. 1998. "Taking Time Seriously: Time-Series-Cross-Section Analysis with a Binary Dependent Variable." *American Journal of Political Science* 42(4):1260–1288.

Clarke, Kevin. 2005. "The Phantom Menace: Omitted Variable Bias in Econometric Research." *Conflict Management and Peace Science* 22(4):341–352.

Fearon, James D. and David D. Laitin. 2003. "Ethnicity, Insurgency, and Civil War." *American Political Science Review* 97(1):1–17.

Gates, Scott and Havard Strand. 2004. "Modeling the Duration of Civil Wars: Measurement and Estimation Issues." Paper prepared for presentation at the Joint Session of Workshops of the ECPR, Uppsala.

Gleditsch, Nils Petter, Peter Wallensteen, Mikael Eriksson, Margareta Sollenberg and Håvard Strand. 2002. "Armed Conflict 1946-2001: A New Dataset." *Journal of Peace Research* 39(5):615–637.

Gujarati, Damodar N. 1995. *Basic Econometrics 3rd edition.* New York: McGraw-Hill.

Gurr, Ted Robert. 1993. *Minorities at Risk. A Global View of Ethnopolitical Conflict.* Washington: United States Institute of Peace Press.

Hanushek, Eric A. and John E. Jackson. 1977. *Statistical Methods for Social Scientists.* New York: Academic Press.

Hausman, Jerry. 2001. "Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left." *Journal of Economic Perspectives* 15(4):57–67.

Hausman, Jerry, Jason Abrevaya and Fiona Scott-Morton. 1998. "Misclassification of the Dependent Variable in a Discrete-Response Setting." *Journal of Econometrics* 87:239–269.

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

Lee, Lung-Fei. 1982. "Specification Error in Multinomial Logit Models." *Journal of Econometrics* 20:247–258.

Regan, Patrick M. and Daniel Norton. 2005. "Greed, Grievance, and Mobilization in Civil Wars." *Journal of Conflict Resolution* 49(3):319–336.

Yatchew, Adonis and Zvi Griliches. 1985. "Specification Error in Probit Models." *The Review of Economics and Statistics* 67(1):134–139.