Advance Access publication November 20, 2009

Political Analysis (2010) 18:78–102
doi:10.1093/pan/mpp033

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

Simon Hug
Département de science politique, Faculté des sciences économiques et sociales, Université de Genève, 40, Bd du Pont d'Arve, 1211 Genève 4, Switzerland
e-mail: simon.hug@unige.ch (corresponding author)

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 binary 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, Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics* 87:239–69) allows for correcting these biases in binary probit models. Empirical examples based on reanalyses of models explaining the occurrence of rebellions and civil wars demonstrate the problem that comes from neglecting these misclassifications.

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 some problems, which are inconsequential in classical linear regression, are much more serious in nonlinear models. For instance, although 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, although 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).

Author's note: 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 available data used in this paper and to the anonymous reviewers and Dominic Senn for helpful comments on an earlier version of this paper.

[1] See also the more general and very instructive discussion of omitted variable biases provided by Clarke (2005).

© The Author 2009. Published by Oxford University Press on behalf of the Society for Political Methodology.
All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

Neglect of these issues in much of the 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. Although 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 throughout political science research, these measurement problems are, however, quite explicit, and, in spite of this, 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 binary 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 3, I discuss an estimator proposed by Hausman, Abrevaya, and Scott-Morton (1998) to address the problem of misclassification in a binary probit setting. Although 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 4, 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" (MAR) (Gurr 1993) that engage in protest (Regan and Norton 2005) and on civil wars. Section 5 concludes.

2 Misclassifications in Political Science

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 endemic. 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 suffer from misclassification. If we compare such a rather central question in people's lives 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. Consider only recall questions on vote choices.

Also in research not relying on survey data, misclassifications are likely. For instance, research on 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 country is embroiled in a civil war). Under the hardly outlandish assumption that the underlying

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

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

continuous indicator is measured with error, there is a strictly positive probability that a war is coded as a peaceful period or vice versa.[4]

Similarly and relatedly, if from a set of groups like the 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 missed some groups (e.g., Hug 2003; Fearon 2006) and this data are aggregated to the level of states, misclassifications will be the result.

Hence, misclassifications are very likely in much 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) proposed an estimator that allows the direct correction of possible misclassifications. In both Monte Carlo simulations and empirical examples, they demonstrate how even small amounts of misclassification affect the estimated coefficients, even if the misclassification is unrelated to any of the independent variables.[5] Their estimator explicitly models the probability of misclassification in a *probit* setup. In a simple probit model, the log-likelihood function is simply

$$L(\boldsymbol{\beta}|y,\boldsymbol{x}) = \sum_{i=1}^{n} \{y_i \ln \Phi(\boldsymbol{x'}_i\boldsymbol{\beta}) + (1-y_i)\ln(1-\Phi(\boldsymbol{x'}_i\boldsymbol{\beta}))\}, \tag{1}$$

where y is the observed dichotomous outcome, $\boldsymbol{x}$ is a vector of explanatory variables, and $\boldsymbol{\beta}$ the vector of coefficients to be estimated. If α_0 corresponds to the probability that the unobserved $y_i = 0$ is classified as a 1 and α_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:

$$\begin{aligned} L(\alpha_0, \alpha_1, \boldsymbol{\beta}|y,\boldsymbol{x}) = \sum_{i=1}^{n} \{ & y_i \ln(\alpha_0 + (1-\alpha_0-\alpha_1)\Phi(\boldsymbol{x'}_i\boldsymbol{\beta})) \\ & + (1-y_i)\ln(1-\alpha_0-(1-\alpha_0-\alpha_1)\Phi(\boldsymbol{x'}_i\boldsymbol{\beta}))\}\,. \end{aligned} \tag{2}$$

For identification of this model α_0, alpha$_1 \in [0, 1)$ and α_0 + alpha$_1 \leqslant 1$ has to hold (Hausman, Abrevaya, and Scott-Morton 1998).[6] It is easy to see that equation (2) reduces to equation (1) if $\alpha_0 = \alpha_1 = 0$. Maximizing equation (2) yields not only estimates for the coefficients $\boldsymbol{\beta}$ but also for the amount of misclassification in the data set through the values

[4] Obviously, given that the binary-dependent variable is underpinned by a continuous indicator, it might be advisable to consider this explicitly in the empirical model to be tested. Given that a large literature on wars and civil wars refrains from this and estimates simply binary logit or probit models, I do not pursue this avenue here. I wish, however, to thank an anonymous reviewer for alerting me to this alternative possibility.

[5] The reason for this is easily understood if we consider marginal effects of particular variables in nonlinear models. The marginal effect of a particular variable corresponds not to the estimated coefficient, as in a linear regression, but is a function of all estimated coefficients. Hence, a wrongly estimated constant term, for instance, affects all estimated effects. See Hausman (2001) for a more general discussion of mismeasured variables.

[6] Lewbel (2000) derives some more general conditions under which this model, also including explanatory variables, is identified.

of α_0 and α_1.[7] Although Hausman, Abrevaya, and Scott-Morton (1998) report estimates for a model employing this setup, they also suggest that both α_0 and α_1 may depend on some exogenous variables (i.e., z_0, z_1):[8]

$$\begin{aligned} \alpha_0 &= f(z_0) \\ \alpha_1 &= f(z_1). \end{aligned} \tag{3}$$

As for the estimates of α_0 and α_1 in the original formulation (equation [1]) of Hausman, Abrevaya, and Scott-Morton (1998), constraints need to be set such that these values remain in the interval [0, 1). As with regression models with dichotomous variables, the most convenient specification is either the logit transformation or the cumulative density function of the normal curve.[9]

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 equation (3) the misclassification probabilities are a function of exogenous variables z_0, z_1. These variables may easily be part of the vector of explanatory variables of the probit model $\boldsymbol{x}$, but again the parameters associated with equation (3) can only be estimated because the functional form differs from the way in which these explanatory variables affect the likelihood.

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 the presence of misclassification. The equation they employ to generate the simulated data set for the Monte Carlo simulations is the following:

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

x_1 and ϵ are drawn from a normal distribution with mean 0 and variance 1, whereas x_2 and x_3 are random draws from a uniform distribution over the unit interval. A certain percentage, namely 2%, 5%, or 20% 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 misclassification come much closer to the true values.

These Monte Carlo simulations are limited in several ways, which make them only partly relevant for typical political science research. More specifically, data sets with 5000 observations or more are not the modal category in political science research using binary probit models. For this reason, I extend these simulations in various ways by using exactly the same setup as shown in equation (4). First, I carried out the Monte Carlo

[7] Given the rather simple structure of the likelihood function, the estimator is easy to implement. Code for *Gauss* and *R* are available from the author upon request.

[8] This possibility for introducing explanatory variables for the amount of misclassification might even be used to bring to bear case-specific information, as suggested by Gordon and Smith (2005).

[9] 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. In that latter case, I use the normal cumulative density function.

simulations for smaller data sets, namely for samples of 1000, 2000, 3000, 4000, and 5000 observations. Second, although Hausman, Abrevaya, and Scott-Morton (1998) kept the amount of misclassifications 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 $\alpha_0 = \alpha_1$ and under the assumption that $\alpha_0 \neq \alpha_1$. Finally, since the proposed estimator also allows the amount of misclassification to depend on exogenous variables, I also carried out Monte Carlo simulations with $\alpha_0 = f(z_0)$ and $\alpha_1 = f(z_1)$.

Figure 1 reports the first set of results for the simulations in which the two probabilities of misclassification α_0 and α_1 are set equal to each other and only one probability of misclassification is estimated.[10] For each estimated coefficient (see the four columns in Fig. 1), I depict the root of the mean squared error (RMSE)[11] both for a simple probit and the model proposed by Hausman, Abrevaya, and Scott-Morton (1998). The rows in Fig. 1 correspond to the four different levels of misclassification assumed, namely 0, 0.02, 0.05, and 0.2. Not surprisingly, the RMSEs increase when we move from the upper to the lower rows in Fig. 1. At the same time, the RMSEs 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 model of Hausman, Abrevaya, and Scott-Morton (1998) becomes preferable to the simple probit model if the probability of misclassification is at least 0.05 (third and fourth row of panels in Fig. 1). Then, however, whether the RMSEs of the probit model are higher or not depend on the sample size and the coefficient considered. Interestingly enough, although the RMSEs of the intercept (β_0) and β_2 are systematically the largest, it is especially for the estimates of β_1 and β_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 α_0 and α_1. In almost all cases, when at least one of the two probabilities is at least 0.05, the RMSEs, especially for larger sample sizes, are smaller for the constant term as estimated by the Hausman, Abrevaya, and Scott-Morton (1998) estimator than the one estimated by probit.[12] The advantage of this estimator becomes even more obvious if we look at cases where one of the misclassification probabilities, namely α_1, is equal to 0.2 (see Fig. 2).

What is striking in the results depicted in Fig. 2 is that for two estimated coefficients, namely β_1 and β_3, independent of the sample size, the RMSE of the estimator proposed by Hausman, Abrevaya, and Scott-Morton (1998) is systematically smaller than the one for the probit estimator. On the other hand, this is never the case for the RMSEs for the constant

[10] 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. Although for all settings of the parameters, 1000 data sets were drawn, the results presented here rely only on the set of estimations that converged. In the Appendix, I provide more details on the number of replications and the simulation results in general. The estimation employed the Broyden–Fletcher–Goldfarb–Shanno algorithm in *Gauss*, which seemed to perform best. Below I discuss in more detail these problems of convergence in the context of the empirical applications.

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

[12] Given that this result is of lesser significance, I refrain from reporting it in more detail graphically here. All results of the Monte Carlo simulations appear, however, in the Appendix.

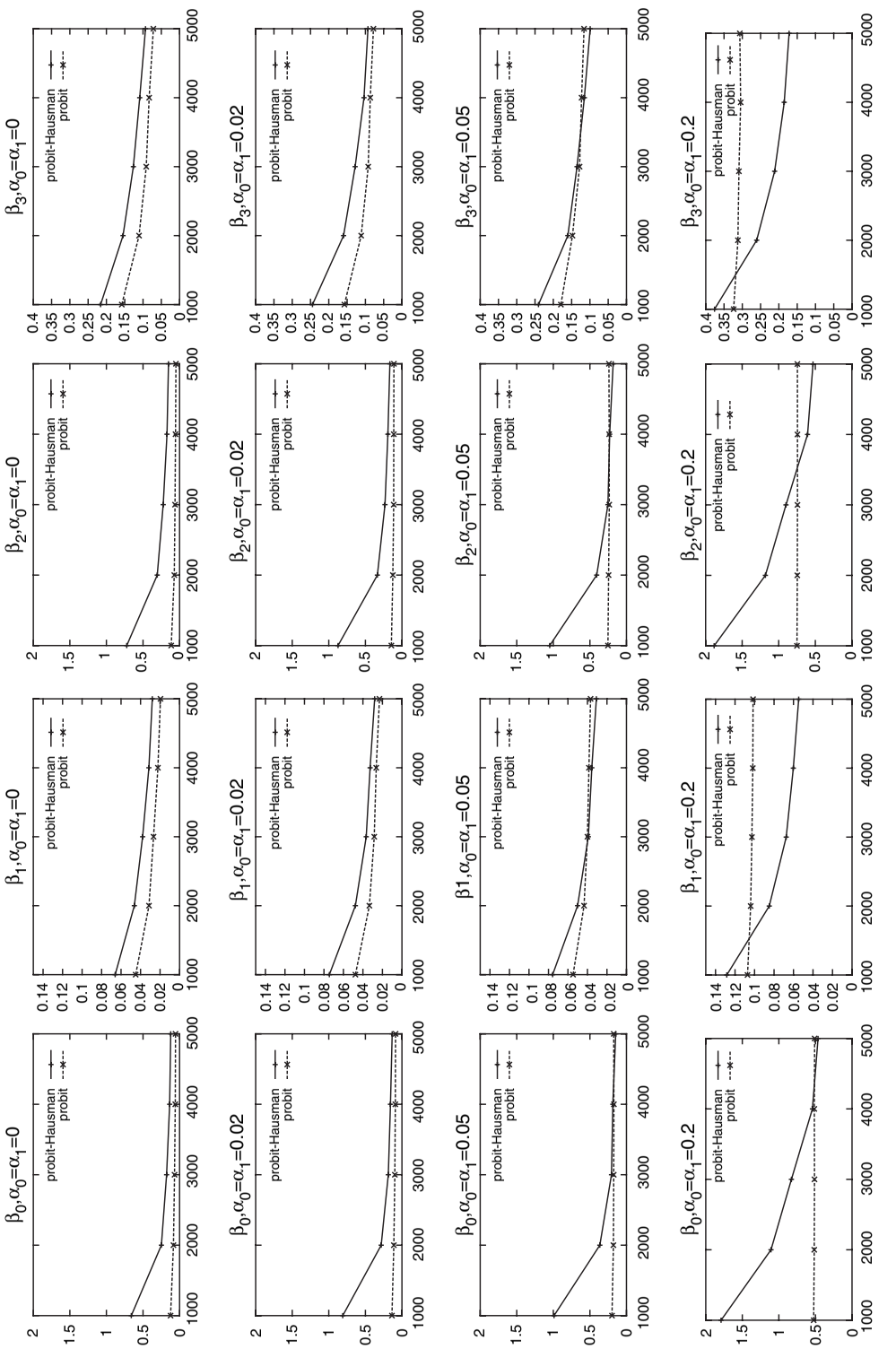

Fig. 1 MC results: $\alpha_0 = \alpha_1$, one coefficient estimated.

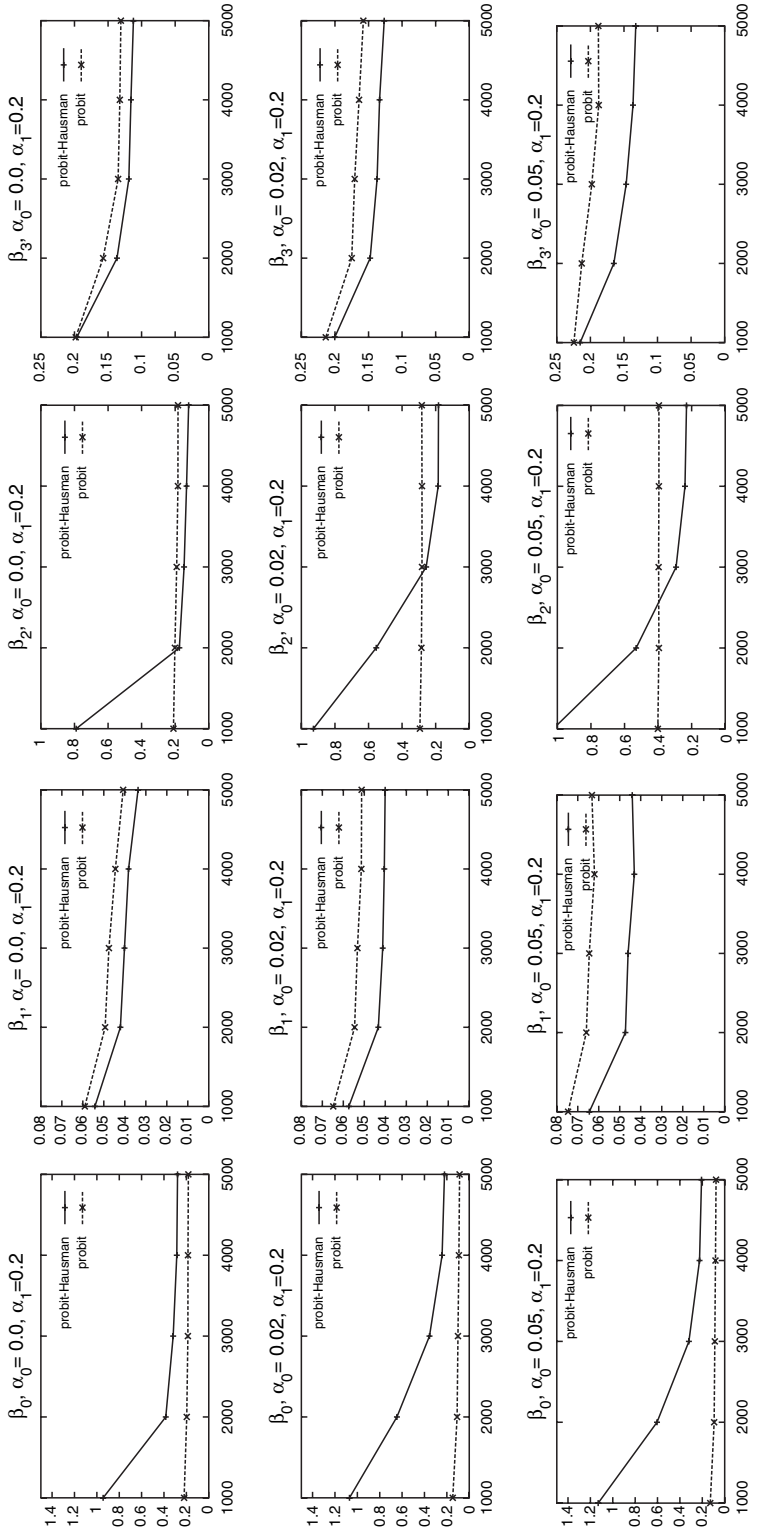

Fig. 2 MC results: $\alpha_0 \neq \alpha_1$, one coefficient estimated.

β_0 and for the remaining slope coefficient (β_2) this only occurs if the sample sizes are larger. 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, and this seems to hold also in smaller samples.

Resorting to the exact same setup, namely letting the two probabilities of misclassification vary independently of each other across the four selected values, I estimated models where both probabilities were coefficients. If the two probabilities are identical, the RMSEs 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 RMSEs of the estimator proposed by Hausman, Abrevaya, and Scott-Morton (1998) (mostly of the constant) is less biased than 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.[13]

To assess the estimator's performance when the probability of misclassification depends on an explanatory variable, I used the following setup for each of the two probabilities:[14]

$$\alpha_{\cdot} = \alpha_a \times (0.5 + x_1) + \theta, \tag{5}$$

where α_a varied across the four values above and θ was drawn from $N(0, 1)$.[15]

The various panels in Fig. 3 report the results for the cases where α_0 depends on x_1 as specified in equation (5), and α_a takes on the three values used above, whereas α_a for α_1 is equal to 0. The results depicted in Fig. 4 are generated in the same fashion, but with α_0 and α_1 inversed. It is apparent in both figures that already with 5% misclassification, the RMSEs of the estimator proposed by Hausman, Abrevaya, and Scott-Morton (1998) for some coefficients yield less biased estimates than those of the simple probit model. If the amount of misclassification is rather large, the differences become quite large and appear even for smaller sample sizes. Hence, for many situations where we expect the probability of misclassification to depend on exogenous variables, the estimator proposed by Hausman, Abrevaya, and Scott-Morton (1998) provides improved estimates.

4 Empirical Examples

To illustrate the performance of proposed estimator by Hausman, Abrevaya, and Scott-Morton (1998), I employ it on two studies dealing with protests of MAR (Gurr 1993) and civil wars.[16] 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 MAR data (Gurr 1993), aggregated, 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[17]

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

[14] Hence, here, the z_0 and z_1 from equation (3) correspond to an included independent variable in the model, namely x_1.

[15] Strictly speaking, this setup does not guarantee that $\alpha \in [0, 1]$. The few values that failed to fall into this interval were recoded to the closest boundary value. Given that this equation is only used for the generation of the simulated data, this fails to have any impact on the results reported below.

[16] In the empirical applications, the problem of convergence in the maximum-likelihood estimation also appeared. Using several sets of starting values for the parameters to be estimated, however, always allowed obtaining estimates.

[17] 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.

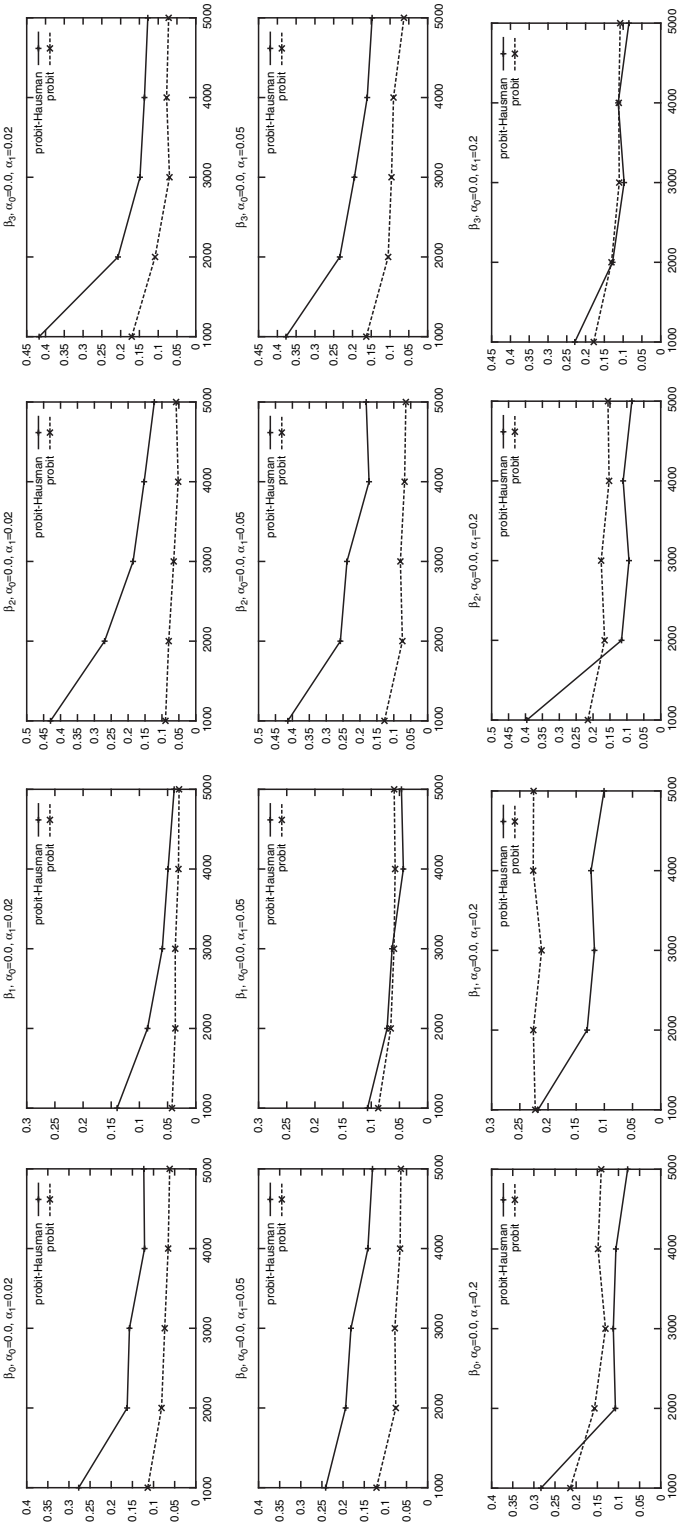

Fig. 3 MC results: α_0 as a function of x_1.

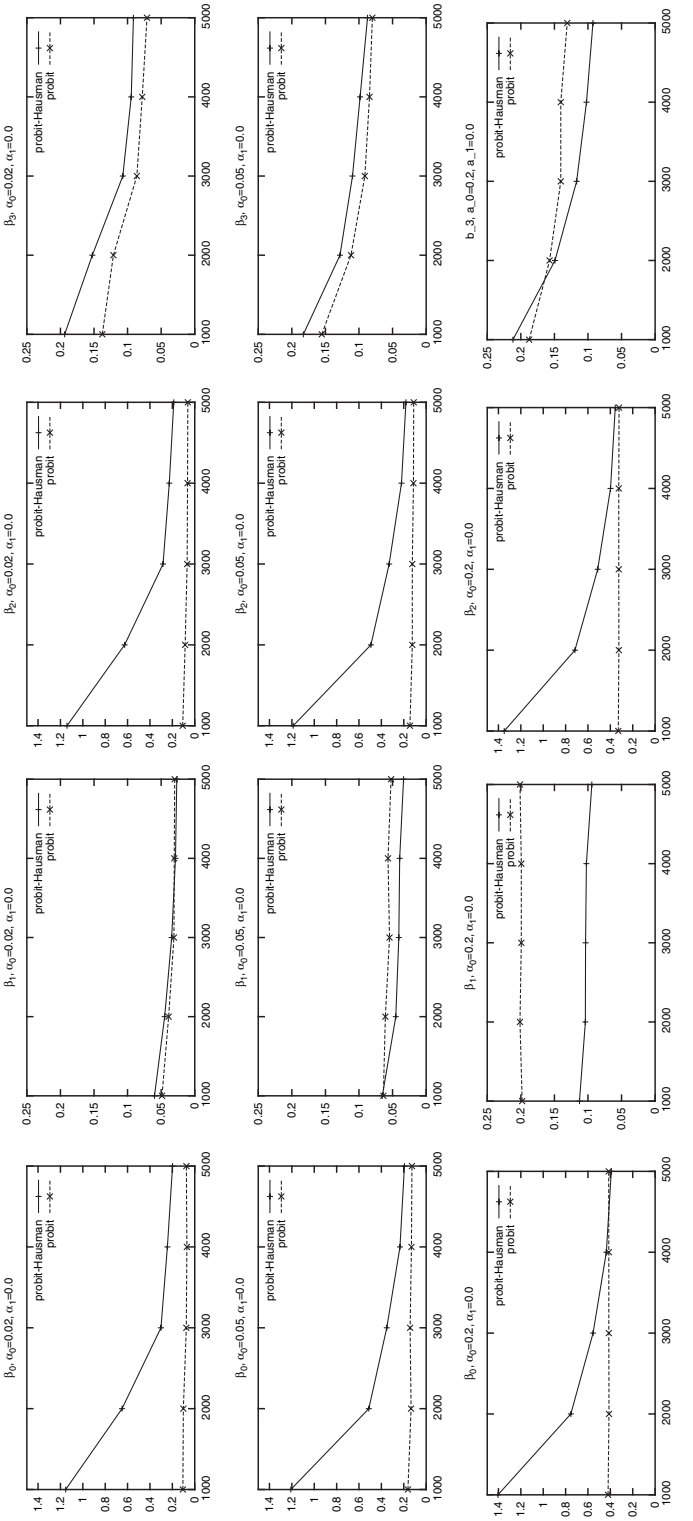

Fig. 4 MC results: α_1 as a function of x_1.

and based on this code whether a minority is engaged in protests, 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 occurred or not. As explanatory variables Regan and Norton (2005) use discrimination, political repression (lagged), extractable resources, per capita gross domestic product (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 splines as well as a counter for the number of years since the last event.

Although 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 protest.[18] Substantively, the results obviously fail to differ from the logit results. Discrimination, per capita GDP, the log of the population size and ethnolinguistic fragmentation positively and statistically significantly (though only moderately for the first variable) affect the outbreak of protest. Repression decreases the probability of such an outbreak, though not statistically significantly, whereas the effect of democracy, as measured by the Polity IV scale, is curvilinear and statistically significant.

When allowing for the possibility of misclassification but assuming that the two probabilities take the same value (column 5 in Table 1), I find a sizeable probability of misclassification of almost 1%.[19] 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), with the exception of the effect of discrimination on protest onset, which completely loses its statistical significance. This is due to a reduced size of the coefficient and an increased value for the standard error.

As seen in the Monte Carlo simulations, estimating an identical probability of misclassification, even if the probabilities differ, is often advisable. Here, however, I also wish to check what happens if individual probabilities are estimated separately (columns 2 and 3 in Table 1) or jointly (column 4 in Table 1). In the case where only the probability that a peaceful year is miscoded as a year with a protest onset, this estimated probability is again quite large, namely 0.008. The differences in the other estimated coefficients are, however, rather small. The probability that a year with a protest was miscoded as a peaceful year is considerably smaller (column 3 in Table 1), approaching zero. Not surprisingly, here, the estimated coefficients barely differ compared with those reported in column 2.

If both probabilities of misclassification are estimated separately in the same model (column 4 in Table 1), I find stronger changes. First of all, the two probabilities of misclassification are quite sizeable with the first one again reaching 0.008. With regard to the coefficients for the substantive variables, quite a few notable changes appear. Discrimination appears to have a much weakened effect when misclassification is taken into account, as is the case for the effect of per capita GDP. Finally, I also report results of an estimation where the two misclassification probabilities are assumed to be equal (column 5 in Table 1). These results, in substantive terms, barely differ from those obtained when allowing the two probabilities to vary separately.

To substantially interpret these results, however, the nonlinear nature of the model suggests using additional information, for instance, maximal effects for the variables in

[18] 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.

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

Table 1 Misclassification: protest (Regan and Norton 2005)

Variables	*(1)* *Probit, b (SE)*	*(2)* *Probit, b (SE)*	*(3)* *probit, b (SE)*	*(4)* *probit, b (SE)*	*(5)* *probit, b (SE)*
Discrimination	0.067 (0.037)	0.064 (0.042)	0.067 (0.036)	0.004 (0.048)	0.028 (0.046)
Per capita GDP	−0.068 (0.084)	−0.050 (0.092)	−0.050 (0.080)	−0.051 (0.098)	−0.061 (0.101)
Lagged political repression	−0.131 (0.065)	−0.147 (0.073)	−0.130 (0.064)	−0.122 (0.075)	−0.146 (0.078)
Extractable resources	0.005 (0.136)	0.086 (0.151)	0.059 (0.133)	0.007 (0.156)	−0.004 (0.163)
Log population size	−0.060 (0.043)	−0.059 (0.049)	−0.051 (0.042)	−0.070 (0.051)	−0.064 (0.052)
Polity IV democracy scale	−0.040 (0.046)	−0.022 (0.049)	−0.016 (0.046)	−0.026 (0.062)	−0.024 (0.055)
Polity IV democracy scale^2	0.002 (0.002)	0.001 (0.002)	0.000 (0.002)	0.001 (0.003)	0.001 (0.003)
Ethnolinguistic fragmentation	−0.002 (0.002)	−0.003 (0.002)	−0.002 (0.002)	0.000 (0.002)	0.000 (0.003)
Peace-years	−2.925 (0.125)	−2.804 (0.089)	−2.551 (0.051)	−3.036 (0.089)	−3.039 (0.084)
Spline 1	−0.256 (0.005)	−0.238 (0.003)	−0.216 (0.002)	−0.259 (0.003)	−0.259 (0.003)
Spline 2	0.055 (0.001)	0.049 (0.000)	0.046 (0.000)	0.054 (0.000)	0.053 (0.000)
Spline 3	−0.013 (0.002)	−0.010 (0.001)	−0.011 (0.001)	−0.010 (0.001)	−0.010 (0.001)
Constant	3.250 (0.965)	3.050 (1.059)	2.820 (0.918)	3.252 (1.197)	3.252 (1.170)
$\sqrt{\alpha_0}$		0.092 (0.020)		0.096 (0.019)	
$\sqrt{\alpha_1}$			0.005 (0.208)	0.030 (0.210)	
$\sqrt{\alpha_0} = \sqrt{\alpha_1}$					0.094 (0.018)
Log likelihood	−326.515	−315.315	−322.000	−312.522	−312.810
n	2019	2019	2019	2019	2019

Table 2 Maximal effects (model 4): protest (Regan and Norton 2005)

	Probit			*Probit with misclassification*		
		Quantiles			*Quantiles*	
Variables	*Mean*	*2.5%*	*97.5%*	*Mean*	*2.5%*	*97.5%*
Discrimination	0.043	−0.004	0.093	0.003	−0.062	0.064
Per capita GDP	−0.049	−0.169	0.087	−0.031	−0.170	0.098
Lagged political repression	−0.280	−0.629	−0.002	−0.260	−0.652	0.019
Extractable resources	−0.000	−0.051	0.043	0.001	−0.054	0.052
Log population size	−0.110	−0.272	0.055	−0.120	−0.313	0.067
Polity IV democracy scale	−0.034	−0.090	0.021	−0.023	−0.093	0.046
Ethnolinguistic fragmentation	−0.035	−0.111	0.027	0.001	−0.075	0.069

Note. For the calculation of the maximal effects, 1000 simulations were carried out by holding all other variables at their means, except the dichotomous variable extractable resources and the peace-years related variables, which were all set to zero.

the models.[20] I report these for models 1 (simple probit) and 4 in Table 2. Over the board one notes that these maximal effects are in absolute values much smaller for the estimates of the Hausman, Abrevaya, and Scott-Morton (1998) model than those of the probit model. When also considering the confidence bands for these marginal effects, the differences for the discrimination variable are especially notable. Hence, I also depict the densities of these simulated marginal effects in Fig. 5. Although the density for the marginal effects from the

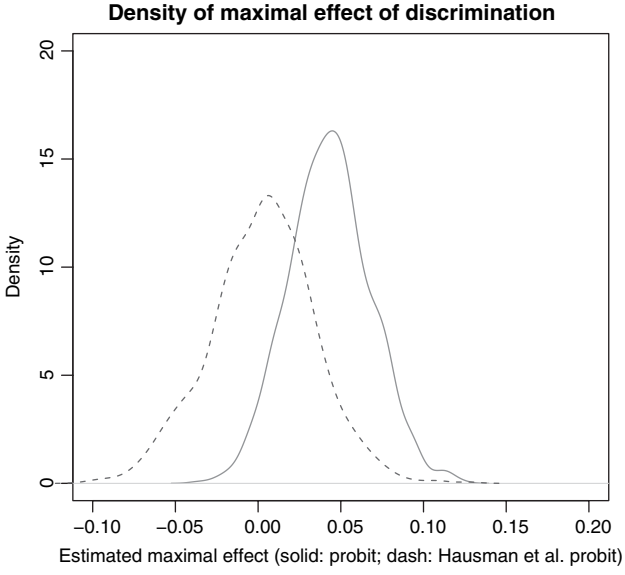

Fig. 5 Simulated maximal effect for discrimination.

[20]To simulate these effects, 1000 draws from the distribution of parameters were drawn, and setting all continuous variables to their mean and all remaining variables to 0, I calculated the maximal effect for the variable of interest. The same procedure was followed for calculating all maximal effects.

Table 3 Misclassification: Fearon and Laitin (2003)

Variables	*(1)* *Probit, b (SE)*	*(2)* *Probit, b (SE)*	*(3)* *Probit, b (SE)*
Prior war	−0.391 (0.130)	−0.586 (0.241)	−0.389 (0.130)
Per capita income$_{t-1}$	−0.135 (0.028)	−0.150 (0.064)	−0.135 (0.028)
Log(population)$_{t-1}$	0.108 (0.031)	0.169 (0.058)	0.107 (0.031)
Log(mountainous terrain)	0.091 (0.034)	0.144 (0.071)	0.091 (0.034)
Noncontiguous state	0.179 (0.122)	0.271 (0.171)	0.178 (0.122)
Oil exporter	0.352 (0.123)	0.539 (0.204)	0.352 (0.123)
New state	0.757 (0.163)	1.027 (0.261)	0.757 (0.163)
Instability	0.259 (0.101)	0.389 (0.164)	0.257 (0.101)
Democracy$_{t-1}$(polity)	0.008 (0.007)	0.015 (0.013)	0.008 (0.007)
Ethnic fractionalization	0.087 (0.157)	0.092 (0.252)	0.085 (0.157)
Religious fractionalization	0.128 (0.209)	0.310 (0.371)	0.135 (0.209)
$\Phi^{-1}(\alpha_0)$		−2.261 (0.187)	
Per capita income		−0.093 (0.052)	
$\Phi^{-1}(\alpha_1)$			−3.759 (16.646)
Per capita income			0.142 (1.013)
Constant	−3.223 (0.303)	−4.302 (0.807)	−3.221 (0.303)
Log likelihood	−481.419	−479.972	−481.416
n	6327	6327	6327

probit model barely includes the value 0, the density from the Hausman, Abrevaya, and Scott-Morton (1998) estimates is centered on zero. Hence substantively, once taking into account misclassification we would presume that discrimination has no effect on protest onset, whereas a simple binary probit would lead us to believe that the probability of protest can be increased by more than 4% through maximum discrimination.

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.[21] As civil war is coded a violent conflict inside a state in which at least 1000 battle deaths are deplored in 1 year. In Table 3 (column 1), I first report a replication of the base model of Fearon and Laitin (2003), which they estimate as a logit model, estimated as a probit model. Estimating the various misclassification probabilities as in the previous example yielded predicted probabilities indistinguishable from zero.[22] Despite these small probabilities, 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

[21]Country-years in which a civil war is coded as ongoing are dropped from the analysis.
[22]For this reason, I refrain from reporting these results here.

Table 4 Maximal effects (model 2): Fearon and Laitin (2003)

	Probit			*Probit with misclassification*		
		Quantiles			*Quantiles*	
Variables	*Mean*	*2.5%*	*97.5%*	*Mean*	*2.5%*	*97.5%*
Prior war	0.000	0.000	0.001	0.000	0.000	0.001
Per capita income$_{t-1}$	0.000	0.000	0.000	0.003	0.000	0.000
Log(population)$_{t-1}$	0.007	0.002	0.019	0.004	0.000	0.021
Log(mountainous terrain)	0.005	0.002	0.010	0.002	0.000	0.011
Noncontiguous state	0.001	0.000	0.002	0.000	0.000	0.002
Oil exporter	0.001	0.000	0.004	0.000	0.000	0.003
New state	0.005	0.001	0.017	0.002	0.000	0.012
Instability	0.001	0.000	0.003	0.000	0.000	0.004
Democracy$_{t-1}$(polity)	0.002	0.000	0.005	0.001	0.000	0.004
Ethnic fractionalization	0.000	0.000	0.001	0.000	0.000	0.002
Religious fractionalization	0.000	0.000	0.002	0.000	0.000	0.002

Note. For the calculation of the maximal effects, 1000 simulations were carried out by holding all other variables at their means, except the dichotomous variable set at values to generate the lowest predicted probabilities, that is, zero for all variables except prior war (set at one).

countries than in rich ones. Ideally, a better measure for the quality of the sources employed for particular countries and years should be used, here, but the simple GDP per capita indicator seems to be a sufficiently good proxy for the quality of reporting.

In column 2 of Table 3, 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 the probability of misclassification decreases with higher GDPs per capita, and this effect reaches statistical significance. Substantively, these estimated coefficients suggest that the amount of misclassification in the poorest countries in the sample is approximately 1%. The coefficients for the misclassification model and especially a likelihood ratio test comparing this model with the one estimated by Fearon and Laitin (2003) suggest that the second type of misclassification is not affected by GDP per capita. Given that only the first type of misclassification seems to be affected by GDP per capita, I report in Table 4 again the maximal effects for all the independent variables.

The results reported in Table 4 suggest again that when considering misclassification, the substantive effects of the various variables are considerably smaller. The most interesting concerns the mountainous terrain, whose maximal effect is cut in half when considering possible misclassification (Table 4). This reduction of the effect also appears in Fig. 6. The density of this effect from the simple probit model would lead us to believe that whether or not a country has a large amount of mountainous terrain significantly affects the likelihood of civil war. Considering the results from the model with misclassification would question, however, this inference. The density for this effect is heavily concentrated on zero, suggesting that mountainous terrain has no effect.

5 Conclusions

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 well known

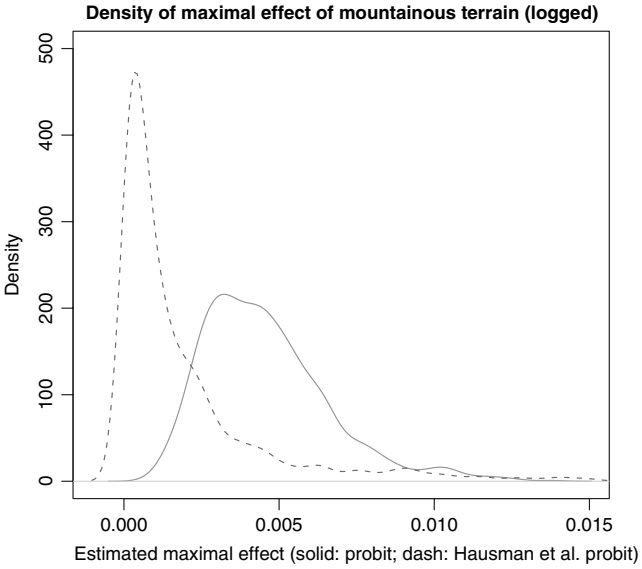

Fig. 6 Simulated maximal effects for mountainous terrain.

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, affects in most cases all estimated coefficients, even in the most innocuous looking cases (Hausman 2001).

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 redress of this problem in binary probit models. In Monte Carlo simulations, I demonstrated that, provided a researcher works with a sizeable sample, the corrections proposed by Hausman, Abrevaya, and Scott-Morton (1998) clearly yield estimates with smaller bias than 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 probability of misclassification.

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

Funding

Partial funding by the Swiss National Science Foundation through the NCCR project *Challenges to democracy in the 21st century* is gratefully acknowledged.

Appendix

Table A1 Descriptive statistics for reanalyses of Regan and Norton (2005)

Variables	*Minimum*	*Mean*	*Maximum*	*SD*	*n*
Protest onset	0	0.293	1.000	0.455	2019
Discrimination	0	1.970	4.000	1.702	2019
Per capita income	5.737	8.107	9.771	0.861	2019
Lagged political repression	1	2.383	9	1.147	2019
Extractable resources	0	0.288	1	0.453	2019
Log population	12.319	16.169	20.918	1.464	2019
Polity IV democracy scale	0	10.753	20	7.712	2019
Polity IV democracy scale^2	0	175.076	400	169.639	2019
Ethnolinguistic fractionalization	1	42.631	93	29.039	2019

In Table A1, I report the descriptive statistics for the example based on Regan and Norton (2005), whereas Table A2 does the same for the analysis based on Fearon and Laitin (2003). Tables A3–A7 report the results of the Monte Carlo simulations (RMSEs) on which the figures in the main text are based.

Table A2 Descriptive statistics of Fearon and Laitin (2003)

Variables	*Minimum*	*Mean*	*Maximum*	*SD*	*n*
Civil war onset	0	0.017	1	0.128	6327
Prior war	0	0.134	1	0.341	6327
Per capita income$_{t-1}$	0.048	3.636	53.901	4.352	6327
Log(population)	5.403	9.065	14.029	1.460	6327
Log(% mountainous terrain)	0	2.175	4.557	1.411	6327
Noncontiguous state	0	0.178	1	0.383	6327
Oil exporter	0	0.128	1	0.334	6327
New state	0	0.026	1	0.158	6327
Instability	0	0.146	1	0.353	6327
Democracy(polity)	−10	−0.396	10	7.554	6327
Ethnic fractionalization	0.001	0.389	0.925	0.286	6327
Religious fractionalization	0	0.366	0.783	0.219	6327

Table A3 RMSE for estimates under the assumption $\alpha_0 = \alpha_1$

	1000		*2000*		*3000*		*4000*		*5000*	
	Hausman	*Probit*	*Hausman*	*Probit*	*Hausman*	*Probit*	*Hausman*	*Probit*	*Hausman*	*Probit*
					$\alpha_1 = \alpha_0 = 0$					
n	463		467		493		463		483	
β_0	0.660	0.123	0.249	0.083	0.177	0.069	0.135	0.059	0.120	0.055
β_1	0.066	0.045	0.046	0.031	0.038	0.027	0.031	0.022	0.028	0.020
β_2	0.722	0.114	0.305	0.073	0.224	0.062	0.172	0.055	0.151	0.050
β_3	0.216	0.158	0.155	0.111	0.127	0.091	0.109	0.083	0.093	0.072
$\alpha_1 = \alpha_0$	0.164		0.171		0.178		0.181		0.183	
					$\alpha_0 = 0.0, \alpha_1 = 0.02$					
n	426		446		441		427		435	
β_0	0.729	0.119	0.244	0.087	0.172	0.072	0.160	0.060	0.142	0.057
β_1	0.061	0.044	0.044	0.032	0.033	0.026	0.030	0.023	0.024	0.020
β_2	0.769	0.104	0.268	0.080	0.178	0.066	0.167	0.054	0.146	0.050
β_3	0.220	0.156	0.138	0.107	0.114	0.093	0.091	0.078	0.089	0.074
$\alpha_1 = \alpha_0$	0.165		0.173		0.178		0.181		0.182	
					$\alpha_0 = 0.0, \alpha_1 = 0.05$					
n	413		401		404		389		408	
β_0	0.716	0.119	0.332	0.094	0.215	0.082	0.181	0.080	0.160	0.068
β_1	0.060	0.044	0.039	0.031	0.032	0.028	0.025	0.024	0.024	0.022
β_2	0.718	0.117	0.319	0.087	0.182	0.075	0.143	0.069	0.119	0.069
β_3	0.200	0.156	0.136	0.116	0.101	0.091	0.090	0.085	0.077	0.076
$\alpha_1 = \alpha_0$	0.163		0.172		0.178		0.182		0.183	
					$\alpha_0 = 0.0, \alpha_1 = 0.2$					
n	328		305		258		235		207	
β_0	0.944	0.220	0.384	0.197	0.318	0.186	0.286	0.186	0.279	0.182
β_1	0.054	0.059	0.042	0.049	0.040	0.048	0.038	0.045	0.034	0.041
β_2	0.790	0.210	0.175	0.202	0.147	0.191	0.132	0.184	0.120	0.183
β_3	0.197	0.198	0.137	0.157	0.119	0.135	0.116	0.132	0.112	0.131
$\alpha_1 = \alpha_0$	0.168		0.177		0.182		0.186		0.185	
					$\alpha_0 = 0.02, \alpha_1 = 0.0$					
n	568		646		705		739		743	
β_0	0.848	0.147	0.371	0.119	0.179	0.108	0.144	0.103	0.129	0.103
β_1	0.077	0.048	0.051	0.031	0.041	0.027	0.035	0.024	0.033	0.021
β_2	0.919	0.134	0.434	0.108	0.250	0.098	0.206	0.090	0.190	0.092
β_3	0.240	0.156	0.159	0.109	0.130	0.087	0.116	0.081	0.109	0.077
$\alpha_1 = \alpha_0$	0.149		0.157		0.163		0.167		0.167	
					$\alpha_1 = \alpha_0 = 0.02$					
n	574		621		669		691		725	
β_0	0.810	0.137	0.284	0.111	0.187	0.097	0.157	0.091	0.134	0.088
β_1	0.074	0.048	0.048	0.033	0.037	0.028	0.033	0.026	0.028	0.023
β_2	0.874	0.144	0.336	0.125	0.232	0.113	0.192	0.112	0.166	0.110
β_3	0.245	0.157	0.160	0.111	0.128	0.092	0.104	0.087	0.093	0.078
$\alpha_1 = \alpha_0$	0.149		0.159		0.165		0.167		0.170	
					$\alpha_0 = 0.02, \alpha_1 = 0.05$					
n	589		603		613		657		682	
β_0	0.952	0.118	0.309	0.092	0.195	0.079	0.154	0.077	0.131	0.073

Continued

Table A3 (continued)

	1000		*2000*		*3000*		*4000*		*5000*	
	Hausman	*Probit*	*Hausman*	*Probit*	*Hausman*	*Probit*	*Hausman*	*Probit*	*Hausman*	*Probit*
β_1	0.065	0.048	0.042	0.039	0.034	0.033	0.028	0.028	0.025	0.027
β_2	0.988	0.165	0.322	0.150	0.195	0.142	0.158	0.140	0.130	0.140
β_3	0.209	0.160	0.141	0.120	0.115	0.111	0.097	0.091	0.081	0.085
$\alpha_1 = \alpha_0$	0.149		0.162		0.166		0.170		0.171	
					$\alpha_0 = 0.02$, $\alpha_1 = 0.2$					
n	483		460		460		488		481	
β_0	1.070	0.148	0.646	0.109	0.356	0.101	0.242	0.092	0.221	0.086
β_1	0.057	0.065	0.043	0.055	0.041	0.053	0.040	0.051	0.040	0.051
β_2	0.928	0.294	0.554	0.285	0.257	0.282	0.185	0.281	0.183	0.283
β_3	0.200	0.214	0.148	0.175	0.137	0.171	0.134	0.164	0.127	0.158
$\alpha_1 = \alpha_0$	0.156		0.166		0.170		0.175		0.177	

Table A4 RMSE for estimates under the assumption $\alpha_0 = \alpha_1$

	1000		*2000*		*3000*		*4000*		*5000*	
	Hausman	*Probit*	*Hausman*	*Probit*	*Hausman*	*Probit*	*Hausman*	*Probit*	*Hausman*	*Probit*
					$\alpha_0 = 0.05$, $\alpha_1 = 0.0$					
n	717		795		847		893		918	
β_0	0.922	0.229	0.420	0.213	0.193	0.210	0.166	0.213	0.143	0.209
β_1	0.093	0.048	0.064	0.037	0.049	0.033	0.045	0.030	0.039	0.028
β_2	1.056	0.201	0.523	0.190	0.305	0.187	0.267	0.185	0.238	0.183
β_3	0.284	0.157	0.191	0.124	0.162	0.108	0.147	0.097	0.132	0.089
$\alpha_1 = \alpha_0$	0.129		0.135		0.141		0.142		0.144	
					$\alpha_0 = 0.05$, $\alpha_1 = 0.02$					
n	714		772		844		874		907	
β_0	0.974	0.218	0.327	0.206	0.189	0.197	0.157	0.197	0.151	0.193
β_1	0.078	0.051	0.056	0.039	0.046	0.035	0.038	0.034	0.036	0.031
β_2	1.074	0.228	0.416	0.209	0.271	0.206	0.238	0.204	0.218	0.207
β_3	0.261	0.168	0.181	0.124	0.143	0.114	0.132	0.100	0.109	0.100
$\alpha_1 = \alpha_0$	0.129		0.138		0.143		0.143		0.146	
					$\alpha_1 = \alpha_0 = 0.05$					
n	677		774		816		850		871	
β_0	0.990	0.198	0.364	0.178	0.205	0.174	0.187	0.172	0.147	0.174
β_1	0.076	0.055	0.050	0.043	0.039	0.040	0.036	0.038	0.031	0.037
β_2	1.049	0.253	0.407	0.244	0.254	0.235	0.230	0.237	0.182	0.240
β_3	0.241	0.179	0.160	0.148	0.135	0.128	0.115	0.123	0.099	0.116
$\alpha_1 = \alpha_0$	0.131		0.140		0.144		0.146		0.150	
					$\alpha_0 = 0.05$, $\alpha_1 = 0.2$					
n	580		601		659		713		725	
β_0	1.130	0.129	0.605	0.094	0.320	0.088	0.225	0.082	0.207	0.077
β_1	0.064	0.075	0.047	0.066	0.046	0.065	0.043	0.062	0.044	0.063
β_2	1.025	0.402	0.531	0.395	0.293	0.397	0.240	0.395	0.230	0.395

Continued

Table A4 (continued)

	1000		*2000*		*3000*		*4000*		*5000*	
	Hausman	*Probit*	*Hausman*	*Probit*	*Hausman*	*Probit*	*Hausman*	*Probit*	*Hausman*	*Probit*
β_3	0.215	0.225	0.165	0.213	0.147	0.198	0.136	0.188	0.132	0.188
$\alpha_1 = \alpha_0$	0.137		0.144		0.153		0.156		0.157	
					$\alpha_0 = 0.2, \alpha_1 = 0.0$					
n	723		827		898		919		951	
β_0	1.410	0.635	0.744	0.632	0.619	0.630	0.470	0.629	0.375	0.631
β_1	0.345	0.073	0.189	0.065	0.165	0.063	0.146	0.062	0.138	0.062
β_2	2.341	0.498	1.240	0.500	1.022	0.495	0.806	0.494	0.681	0.496
β_3	1.347	0.222	0.627	0.207	0.526	0.194	0.457	0.190	0.431	0.188
$\alpha_1 = \alpha_0$	0.065		0.060		0.057		0.054		0.052	
					$\alpha_0 = 0.2, \alpha_1 = 0.02$					
n	777		861		908		946		964	
β_0	1.542	0.619	0.689	0.619	0.527	0.621	0.438	0.622	0.379	0.620
β_1	0.303	0.076	0.185	0.071	0.166	0.067	0.144	0.068	0.137	0.065
β_2	2.280	0.525	1.198	0.520	0.966	0.523	0.814	0.525	0.710	0.523
β_3	0.955	0.243	0.622	0.213	0.513	0.207	0.466	0.204	0.432	0.201
$\alpha_1 = \alpha_0$	0.064		0.063		0.055		0.054		0.049	
					$\alpha_0 = 0.2, \alpha_1 = 0.05$					
n	800		879		928		964		973	
β_0	1.484	0.604	0.919	0.604	0.598	0.602	0.394	0.602	0.396	0.601
β_1	0.331	0.082	0.191	0.075	0.146	0.074	0.136	0.073	0.126	0.072
β_2	2.295	0.564	1.419	0.563	0.981	0.563	0.767	0.564	0.727	0.563
β_3	1.009	0.250	0.603	0.231	0.466	0.227	0.421	0.223	0.389	0.221
$\alpha_1 = \alpha_0$	0.064		0.059		0.054		0.049		0.047	
					$\alpha_1 = \alpha_0 = 0.2$					
n	729		841		892		936		957	
β_0	1.791	0.524	1.108	0.516	0.829	0.514	0.541	0.515	0.465	0.511
β_1	0.128	0.107	0.085	0.104	0.067	0.103	0.060	0.102	0.055	0.101
β_2	1.884	0.751	1.183	0.746	0.901	0.744	0.606	0.744	0.532	0.745
β_3	0.376	0.324	0.261	0.312	0.212	0.309	0.185	0.305	0.172	0.307
$\alpha_1 = \alpha_0$	0.065		0.059		0.060		0.052		0.051	

Table A5 RMSE for estimates under the assumption $\alpha_0 \neq \alpha_1$

	1000		*2000*		*3000*		*4000*		*5000*	
	Hausman	*Probit*	*Hausman*	*Probit*	*Hausman*	*Probit*	*Hausman*	*Probit*	*Hausman*	*Probit*
					$\alpha_0 = \alpha_1 = 0.0$					
n	368		401		381		388		386	
β_0	0.947	0.113	0.382	0.086	0.165	0.066	0.133	0.059	0.119	0.056
β_1	0.229	0.044	0.134	0.032	0.113	0.027	0.090	0.022	0.081	0.019
β_2	1.495	0.104	0.735	0.082	0.534	0.062	0.450	0.055	0.406	0.048
β_3	0.793	0.144	0.427	0.114	0.340	0.086	0.276	0.076	0.239	0.069
α_0	0.054		0.042		0.036		0.031		0.029	
α_1	0.228		0.187		0.172		0.153		0.146	

Continued

Table A5 (continued)

	1000		*2000*		*3000*		*4000*		*5000*	
	Hausman	*Probit*	*Hausman*	*Probit*	*Hausman*	*Probit*	*Hausman*	*Probit*	*Hausman*	*Probit*
					$\alpha_0 = 0.0, \alpha_1 = 0.02$					
n	378		361		391		396		406	
β_0	0.956	0.111	0.291	0.083	0.175	0.066	0.150	0.062	0.124	0.060
β_1	0.254	0.045	0.141	0.031	0.109	0.026	0.091	0.022	0.086	0.019
β_2	1.754	0.103	0.738	0.074	0.531	0.062	0.455	0.060	0.423	0.048
β_3	1.078	0.156	0.481	0.105	0.332	0.088	0.274	0.078	0.253	0.070
α_0	0.054		0.040		0.036		0.033		0.029	
α_1	0.221		0.189		0.171		0.151		0.147	
					$\alpha_0 = 0, \alpha_1 = 0.05$					
n	381		425		397		412		434	
β_0	0.924	0.130	0.326	0.093	0.181	0.086	0.144	0.075	0.139	0.066
β_1	0.270	0.044	0.144	0.035	0.111	0.027	0.094	0.025	0.082	0.024
β_2	1.695	0.121	0.754	0.088	0.536	0.080	0.464	0.073	0.422	0.070
β_3	0.984	0.155	0.471	0.117	0.330	0.097	0.296	0.084	0.255	0.074
α_0	0.054		0.040		0.035		0.032		0.031	
α_1	0.227		0.185		0.160		0.151		0.140	
					$\alpha_0 = 0.0, \alpha_1 = 0.2$					
n	383		467		462		488		454	
β_0	1.160	0.219	0.307	0.196	0.294	0.192	0.182	0.187	0.158	0.184
β_1	0.382	0.058	0.170	0.049	0.134	0.046	0.105	0.044	0.094	0.045
β_2	2.082	0.208	0.812	0.203	0.677	0.190	0.520	0.186	0.477	0.188
β_3	1.306	0.190	0.532	0.161	0.422	0.142	0.326	0.140	0.297	0.133
α_0	0.046		0.035		0.031		0.028		0.026	
α_1	0.197		0.167		0.154		0.137		0.130	
					$\alpha_0 = 0.02, \alpha_1 = 0.0$					
n	410		418		441		442		490	
β_0	0.706	0.136	0.265	0.113	0.182	0.108	0.155	0.106	0.129	0.101
β_1	0.288	0.045	0.141	0.033	0.106	0.028	0.087	0.023	0.077	0.021
β_2	1.558	0.125	0.711	0.101	0.546	0.097	0.430	0.095	0.386	0.090
β_3	0.979	0.146	0.428	0.117	0.348	0.088	0.267	0.081	0.234	0.071
α_0	0.054		0.043		0.036		0.032		0.028	
α_1	0.228		0.187		0.162		0.143		0.136	
					$\alpha_0 = \alpha_1 = 0.02$					
n	431		473		459		462		497	
β_0	0.981	0.129	0.432	0.110	0.195	0.098	0.158	0.089	0.144	0.088
β_1	0.260	0.047	0.140	0.032	0.097	0.028	0.088	0.026	0.079	0.024
β_2	1.763	0.138	0.791	0.126	0.504	0.114	0.449	0.111	0.393	0.112
β_3	0.976	0.156	0.426	0.118	0.312	0.094	0.266	0.090	0.240	0.079
α_0	0.054		0.042		0.036		0.031		0.029	
α_1	0.220		0.177		0.153		0.146		0.133	
					$\alpha_0 = 0.02, \alpha_1 = 0.05$					
n	445		499		485		552		527	
β_0	0.846	0.123	0.371	0.094	0.189	0.082	0.172	0.074	0.147	0.069
β_1	0.225	0.049	0.136	0.036	0.114	0.032	0.094	0.029	0.081	0.029
β_2	1.468	0.163	0.755	0.149	0.562	0.144	0.483	0.138	0.415	0.136

Continued

Table A5 (continued)

	1000		*2000*		*3000*		*4000*		*5000*	
	Hausman	*Probit*	*Hausman*	*Probit*	*Hausman*	*Probit*	*Hausman*	*Probit*	*Hausman*	*Probit*
β_3	0.906	0.168	0.441	0.118	0.372	0.106	0.292	0.097	0.248	0.090
α_0	0.054		0.039		0.036		0.032		0.029	
α_1	0.210		0.174		0.160		0.142		0.131	
					$\alpha_0 = 0.02, \alpha_1 = 0.2$					
n	446		523		542		559		577	
β_0	1.293	0.144	0.566	0.108	0.342	0.096	0.209	0.095	0.172	0.092
β_1	0.344	0.064	0.174	0.058	0.126	0.053	0.109	0.052	0.087	0.052
β_2	2.335	0.296	1.023	0.286	0.688	0.284	0.550	0.280	0.443	0.281
β_3	1.575	0.211	0.598	0.170	0.406	0.162	0.344	0.160	0.283	0.159
α_0	0.048		0.037		0.032		0.029		0.026	
α_1	0.192		0.165		0.142		0.135		0.122	

Table A6 RMSE for estimates under the assumption $\alpha_0 \neq \alpha_1$

	1000		*2000*		*3000*		*4000*		*5000*	
	Hausman	*Probit*	*Hausman*	*Probit*	*Hausman*	*Probit*	*Hausman*	*Probit*	*Hausman*	*Probit*
					$\alpha_0 = 0.05, \alpha_1 = 0.0$					
n	427		491		480		485		510	
β_0	1.128	0.224	0.476	0.218	0.302	0.215	0.186	0.211	0.163	0.209
β_1	0.306	0.047	0.146	0.038	0.108	0.033	0.088	0.032	0.084	0.029
β_2	2.050	0.197	0.827	0.193	0.607	0.186	0.464	0.185	0.431	0.182
β_3	1.245	0.170	0.432	0.120	0.364	0.106	0.298	0.092	0.255	0.089
α_0	0.057		0.045		0.038		0.034		0.032	
α_1	0.222		0.180		0.159		0.144		0.136	
					$\alpha_0 = 0.05, \alpha_1 = 0.02$					
n	457		501		532		521		543	
β_0	1.129	0.208	0.554	0.201	0.295	0.194	0.182	0.195	0.174	0.193
β_1	0.241	0.051	0.157	0.039	0.114	0.036	0.095	0.033	0.087	0.031
β_2	1.820	0.228	0.968	0.209	0.623	0.204	0.473	0.206	0.446	0.203
β_3	0.925	0.170	0.504	0.128	0.351	0.116	0.297	0.106	0.258	0.100
α_0	0.056		0.045		0.039		0.034		0.032	
α_1	0.210		0.176		0.160		0.140		0.133	
					$\alpha_0 = \alpha_1 = 0.05$					
n	473		518		518		541		577	
β_0	1.328	0.196	0.501	0.187	0.214	0.171	0.186	0.173	0.178	0.167
β_1	0.366	0.052	0.154	0.046	0.110	0.040	0.095	0.036	0.080	0.037
β_2	2.263	0.248	0.880	0.249	0.559	0.236	0.484	0.237	0.413	0.236
β_3	1.224	0.174	0.481	0.137	0.358	0.125	0.283	0.120	0.243	0.119
α_0	0.052		0.044		0.036		0.034		0.030	
α_1	0.210		0.171		0.151		0.136		0.126	
					$\alpha_0 = 0.05, \alpha_1 = 0.2$					

Continued

Table A6 (continued)

	1000		*2000*		*3000*		*4000*		*5000*	
	Hausman	*Probit*	*Hausman*	*Probit*	*Hausman*	*Probit*	*Hausman*	*Probit*	*Hausman*	*Probit*
n	492		556		613		634		674	
β_0	1.611	0.131	0.484	0.097	0.338	0.088	0.289	0.078	0.209	0.077
β_1	0.479	0.072	0.189	0.068	0.134	0.065	0.110	0.063	0.095	0.063
β_2	2.844	0.400	0.987	0.393	0.752	0.395	0.611	0.391	0.485	0.397
β_3	1.594	0.220	0.610	0.205	0.522	0.191	0.350	0.192	0.298	0.190
α_0	0.047		0.039		0.034		0.030		0.029	
α_1	0.188		0.155		0.140		0.132		0.123	
	$\alpha_0 = 0.2, \alpha_1 = 0.0$									
n	476		502		514		533		523	
β_0	1.577	0.648	0.887	0.640	0.603	0.632	0.325	0.634	0.402	0.636
β_1	0.363	0.071	0.175	0.065	0.133	0.063	0.112	0.062	0.102	0.062
β_2	2.420	0.504	1.265	0.502	0.927	0.495	0.624	0.498	0.650	0.498
β_3	1.178	0.215	0.551	0.201	0.399	0.196	0.347	0.190	0.323	0.184
α_0	0.064		0.053		0.047		0.041		0.039	
α_1	0.192		0.163		0.142		0.131		0.126	
	$\alpha_0 = 0.2, \alpha_1 = 0.02$									
n	458		505		542		523		526	
β_0	1.646	0.626	0.980	0.625	0.638	0.621	0.492	0.621	0.305	0.619
β_1	0.524	0.072	0.188	0.070	0.133	0.070	0.115	0.068	0.099	0.065
β_2	4.448	0.527	1.430	0.524	0.929	0.526	0.765	0.523	0.564	0.518
β_3	3.605	0.238	0.625	0.213	0.416	0.210	0.373	0.203	0.303	0.199
α_0	0.063		0.052		0.046		0.042		0.039	
α_1	0.186		0.157		0.141		0.130		0.115	
	$\alpha_0 = 0.2, \alpha_1 = 0.05$									
n	479		526		567		591		557	
β_0	1.968	0.610	1.183	0.606	0.676	0.607	0.426	0.598	0.310	0.603
β_1	0.611	0.081	0.190	0.076	0.152	0.074	0.121	0.073	0.102	0.072
β_2	4.030	0.565	1.646	0.563	1.050	0.563	0.739	0.562	0.578	0.562
β_3	2.793	0.243	0.639	0.234	0.474	0.219	0.376	0.229	0.327	0.216
α_0	0.062		0.053		0.048		0.041		0.037	
α_1	0.177		0.148		0.139		0.125		0.115	
	$\alpha_0 = \alpha_1 = 0.2$									
n	451		571		607		628		704	
β_0	9.257	0.521	1.251	0.520	0.805	0.511	0.714	0.514	0.622	0.511
β_1	9.970	0.103	0.278	0.100	0.201	0.103	0.242	0.101	0.139	0.101
β_2	54.519	0.743	1.832	0.747	1.288	0.743	1.435	0.743	0.927	0.743
β_3	31.885	0.320	0.941	0.302	0.661	0.311	0.846	0.303	0.418	0.306
α_0	0.058		0.055		0.045		0.047		0.041	
α_1	0.157		0.136		0.131		0.124		0.117	

Table A7 RMSE for estimates with α_0 or α_1 as a function of x_1

	1000		*2000*		*3000*		*4000*		*5000*	
	Hausman	*Probit*	*Hausman*	*Probit*	*Hausman*	*Probit*	*Hausman*	*Probit*	*Hausman*	*Probit*
					$a\alpha_0 \times (0.5 + x_1[j])0.02000$, $\alpha_1 = 0.0$					
n	162		165		145		143		170	
β_0	1.154	0.106	0.649	0.101	0.302	0.076	0.244	0.071	0.199	0.075
β_1	0.060	0.049	0.045	0.039	0.034	0.031	0.029	0.031	0.027	0.030
β_2	1.141	0.107	0.626	0.085	0.283	0.067	0.227	0.065	0.188	0.062
β_3	0.194	0.138	0.153	0.121	0.107	0.086	0.095	0.078	0.091	0.071
$a\alpha_0$	0.059		0.050		0.038		0.036		0.031	
$a\alpha_1$	0.031		0.020		0.016		0.014		0.012	
					$a\alpha_0 \times (0.5 + x_1[j])0.050000$, $\alpha_1 = 0.0$					
n	191		184		206		213		229	
β_0	1.209	0.162	0.512	0.135	0.352	0.143	0.234	0.131	0.195	0.127
β_1	0.065	0.064	0.045	0.061	0.041	0.055	0.039	0.057	0.034	0.052
β_2	1.188	0.145	0.493	0.123	0.331	0.124	0.221	0.114	0.181	0.111
β_3	0.183	0.155	0.128	0.112	0.110	0.091	0.098	0.084	0.087	0.080
$a\alpha_0$	0.040		0.047		0.042		0.037		0.033	
$a\alpha_1$	0.027		0.021		0.020		0.020		0.019	
					$a\alpha_0 \times (0.5 + x_1[j])0.200000$, $\alpha_1 = 0.0$					
n	178		310		362		345		343	
β_0	1.404	0.420	0.753	0.412	0.555	0.414	0.436	0.414	0.392	0.415
β_1	0.113	0.198	0.104	0.201	0.104	0.199	0.102	0.199	0.094	0.201
β_2	1.348	0.329	0.716	0.324	0.513	0.325	0.399	0.325	0.356	0.322
β_3	0.212	0.188	0.149	0.157	0.117	0.140	0.102	0.140	0.093	0.131
$a\alpha_0$	0.076		0.065		0.061		0.058		0.060	
$a\alpha_1$	0.076		0.072		0.078		0.078		0.074	
					$\alpha_0 = 0$, $a\alpha_1 \times (0.5 + x_1[j])0.020000$					
n	46		61		47		54		71	
β_0	0.277	0.114	0.162	0.081	0.157	0.074	0.121	0.065	0.123	0.062
β_1	0.140	0.042	0.086	0.036	0.059	0.036	0.049	0.031	0.038	0.030
β_2	0.430	0.090	0.270	0.080	0.185	0.065	0.153	0.052	0.123	0.058
β_3	0.417	0.170	0.207	0.108	0.149	0.070	0.137	0.077	0.128	0.072
$a\alpha_0$	0.174		0.139		0.115		0.092		0.088	
$a\alpha_1$	0.029		0.029		0.023		0.017		0.017	
					$\alpha_0 = 0$, $a\alpha_1 \times (0.5 + x_1[j])0.050000$					
n	64		66		78		79		76	
β_0	0.241	0.121	0.194	0.075	0.181	0.077	0.141	0.065	0.131	0.063
β_1	0.107	0.088	0.072	0.066	0.063	0.060	0.043	0.058	0.046	0.059
β_2	0.413	0.128	0.258	0.075	0.238	0.080	0.173	0.068	0.182	0.064
β_3	0.377	0.163	0.234	0.105	0.195	0.096	0.161	0.091	0.148	0.063
$a\alpha_0$	0.159		0.140		0.142		0.105		0.106	
$a\alpha_1$	0.043		0.030		0.029		0.028		0.025	
					$\alpha_0 = 0$, $a\alpha_1 \times (0.5 + x_1[j])0.200000$					
n	13		34		16		92		55	
β_0	0.283	0.214	0.107	0.157	0.112	0.131	0.106	0.148	0.078	0.141
β_1	0.218	0.223	0.131	0.226	0.118	0.211	0.124	0.226	0.101	0.226
β_2	0.396	0.215	0.116	0.166	0.094	0.176	0.111	0.153	0.085	0.156
β_3	0.229	0.178	0.128	0.131	0.098	0.110	0.113	0.112	0.085	0.108
$a\alpha_0$	0.196		0.070		0.071		0.072		0.060	
$a\alpha_1$	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'Economie et de Statistiques* 55–56:243–75.

Achen, Christopher H. 2005. Let's put garbage-can regressions and garbage-can probits where they belong. *Conflict Management and Peace Science* 22:327–39.

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* 42:1260–88.

Clarke, Kevin. 2005. The phantom menace: Omitted variable bias in econometric research. *Conflict Management and Peace Science* 22:341–52.

Fearon, James D. 2006. Ethnic mobilization and ethnic violence. In *The oxford handbook of political economy,* ed. Barry R. Weingast and Donald A. Wittman, 852–68. Oxford, UK: Oxford University Press.

Fearon, James D., and David D. Laitin. 2003. Ethnicity, insurgency, and civil war. *American Political Science Review* 97:1–17.

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, Sweden.

Gleditsch, Nils Petter, Peter Wallensteen, Mikael Eriksson, and Margareta Sollenberg, and Håvard Strand. 2002 Armed conflict 1946-2001: A new dataset. *Journal of Peace Research* 39:615–37.

Gordon, Sanford C., and Alastair Smith. 2005. Qualitative leverage and the epistemology of expert opinion. *Political Analysis* 13:280–91.

Gurr, Ted Robert 1993. *Minorities at risk. A global view of ethnopolitical conflict*. Washington, DC: United States Institute of Peace Press.

Hausman, Jerry. 2001. Mismeasured variables in econometric analysis: Problems from the right and problems from the left. *Journal of Economic Perspectives* 15:57–67.

Hausman, Jerry, Jason Abrevaya, and Fiona Scott-Morton. 1998. Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics* 87:239–69.

Hug, Simon. 2003. Selection bias in comparative research. The case of incomplete datasets. *Political Analysis* 11:255–74.

Lee, Lung-Fei. 1982. Specification error in multinomial logit models. *Journal of Econometrics* 20:247–58.

Lewbel, Arthur. 2000. Identification of the binary choice model with misclassification. *Econometric Theory* 16:603–09.

Regan, Patrick M., and Daniel Norton. 2005. Greed, grievance, and mobilization in civil wars. *Journal of Conflict Resolution* 49(3):319–36.

Yatchew, Adonis, and Zvi Griliches. 1985. Specification error in probit models. *Review of Economics and Statistics* 67:134–39.