

# Registration and Voting under Rational Expectations: The Econometric Implications

Christopher H. Achen  
Department of Politics and  
Center for the Study of Democratic Politics  
Princeton University  
Princeton, NJ 08544  
achen@princeton.edu

July 7, 2008

## **Abstract**

Alone among modern democracies, the United States makes voter registration a personal responsibility rather than a governmental function. In almost all states, registration deadlines occur well before elections. Failure to register by the deadline makes the probability of voting exactly zero. This sequential feature of the registration and voting decisions has been skipped over by most researchers, who simply ignore registration. Others, notably Timpone (1998), have used the seemingly appropriate Heckman-style selection model, but have arrived at findings difficult to believe. This paper investigates the appropriate choice of a registration model under a rational expectations assumption about the desire to vote, showing that, rather surprisingly, conventional selection models will generally perform less well than ignoring the selection effect of registration entirely. However, neither is quite correct. Finally then, the paper proposes and tests a flexible model for registration as a step toward substantively appropriate joint modeling of registration and voting.

# Introduction<sup>1</sup>

Voter turnout is central to the theory of democratic government. Without equal participation, elections lose legitimacy. Many believe that uneven participation also biases government policy (Lijphart 1997, Griffin & Newman 2005). Thus uneven turnout across the population has concerned political scientists for a long time (Merriam and Gosnell 1924, Arneson 1925, Gosnell 1927, Tingsten 1937), and prominent political scientists continue to study it empirically and to test our theoretical notions against the evidence (e.g., Kelley *et al.* 1967, Wolfinger & Rosenstone 1980, Verba *et al.* 1995, Blais 2000).

Turnout cannot be understood without prior study of the voter registration process. Every democracy maintains lists of eligible voters. In rural areas where everyone is known to the neighbors, the lists may be informal. Far more commonly, though, the electoral register is a written or electronic document maintained by government authorities. Those not on the list at election day cannot vote. In every country, the lists are fallible to a greater or lesser degree.

In the United States, voter rolls are a state responsibility by national constitutional provision. Each state sets its own rules for registering voters, with most requiring that citizens register themselves at least one month before the election at which they wish to vote. Becoming registered is the citizen's responsibility, typically requiring a trip to city hall, a phone request that an application be mailed, or the downloading of a registration form. In principle, most states offer a chance to register when a driver's license is obtained, but in practice, enforcement of the rule may be uneven.<sup>2</sup> Only three states have election-day registration or no registration at all. The result is that many American citizens are not registered, amounting to perhaps 25% of the eligible adult population.<sup>3</sup>

---

<sup>1</sup>Copyright by the author. A preliminary and partial version was presented at the Annual Meeting of the Midwest Political Science Association, Chicago, Illinois, April 3-6, 2008. My special thanks to Aya Kachi for her professional and cheerful research assistance. The paper would have been impossible without her assistance. My appreciation also to Larry Bartels for a lengthy conversation about how the topic should be approached. Andre' Blais suggested the topic, and his advice and encouragement have been helpful from the start. Simon Jackman gave helpful advice. I also thank Bruce Willsie of Labels and Lists, whose generous assistance to his alma mater has made possible the acquisition of the New Jersey dataset used in this paper.

<sup>2</sup>In a colorful state like New Jersey, as the author learned in person, a new resident with a valid out-of-state drivers license can pay \$10 at the Department of Motor Vehicles to "skip all that." "All that" includes both the driver's examination on New Jersey law and the voter registration form. A February 2007 survey (Department of the Public Advocate, 2008) showed that only 8% of New Jersey drivers license applicants were given the opportunity to register to vote. Reform efforts are underway. Reform efforts are always underway in New Jersey.

<sup>3</sup>No one knows how many Americans are eligible to vote but not registered. First, counting the eligible is not easy due to the uncertainties of Census population estimates and the exclusion of felons in many states (McDonald and Popkin 2001). Second, counting the registered is no easier: Voter registration lists are notoriously heavy on "deadwood" (people who have died or moved away). Alternate sources also fail: Surveys overstate both voting and registration (Uhlener 1989).

In the Census Bureau's 2000 CPS election module, 19% of the population said that they had failed to

Statistical studies of voter registration and turnout are far too numerous to cite. (A recent review is Highton 2004.) But with few exceptions (for example, Uhlaner 1989, Jackson 1996), they have ignored the registration decision and focused directly on the decision to vote. The many studies showing that age, partisanship, and interest in the campaign make people more likely to vote are based on such models. But in the United States, no such specification can be correct. Growing older or becoming a partisan has no effect at all on the turnout of those who are not registered. The marginal effect is exactly zero. In spite of many methodological studies attempting to improve our statistical analyses of turnout (for example, Nagler 1994, Achen and Sinnott 2009), this aspect of the turnout problem has received little attention.

Those unfamiliar with selection bias might imagine that voter turnout studies could be confined to the registered voters. After all, it might be thought, what interests us is the effect of explanatory variables on the *registered*. Why not study just them? The difficulty, as Heckman (1979) pointed out, is that this approach will distort actual causal effects. Variables that powerfully influence turnout may appear to be unimportant, while those that do not matter may loom large (Achen 1986, chap. 4). In any case, so long as some unmeasured factors influence both registration and turnout, biases will occur.

The apparent solution is to model both the selection and outcome equations in the manner pioneered by Heckman and extended to the bivariate probit case by Dubin and Rivers (1989/1990). Timpone (1998) was the first to apply these models to registration and voting, modeling them as two successive decisions, with voting being conditional on registration. This article was an important step forward and a brave effort to be unreservedly commended. However, selection models rely heavily on their assumptions, and they are less robust than many other econometric techniques.

Indeed, Timpone's(1998) findings are odd. Strength of partisanship powerfully influences the decision to register, for example, as one would expect. But it virtually disappears as a cause of voting. As Figure 1 illustrates, other customarily powerful influences on turnout also suffer dramatic and implausible declines in their estimated effects when Timpone's model replaces conventional probit. And the estimated correlation between the unmeasured factors in Timpone's registration and voting equations is negative (albeit non-significant), meaning that unmeasured individual differences that *increase* the desire

---

vote due to being unregistered. Other non-voters were probably unregistered without realizing it. Uhlaner (1989, 79) uses the vote and registration validation studies by the National Election Studies (NES) in 1980 and 1984 to estimate the unregistered at about 25% of the population, though that figure may be too high due to some states' poor recordkeeping or too low due to non-response to the NES by those with little interest in politics, a disproportionate number of whom are probably unregistered. Figure 4 below suggests that 30% is a better estimate for a transient state like New Jersey with unusually many non-native born citizens. Thus on average over the entire U.S., the 25% figure is my own rough estimate for recent decades.

to register probably *reduce* the likelihood of voting. All these results raise questions.

### Timpone's Estimated Impacts on Turnout: Conventional Probit vs. Selection-Corrected Probit

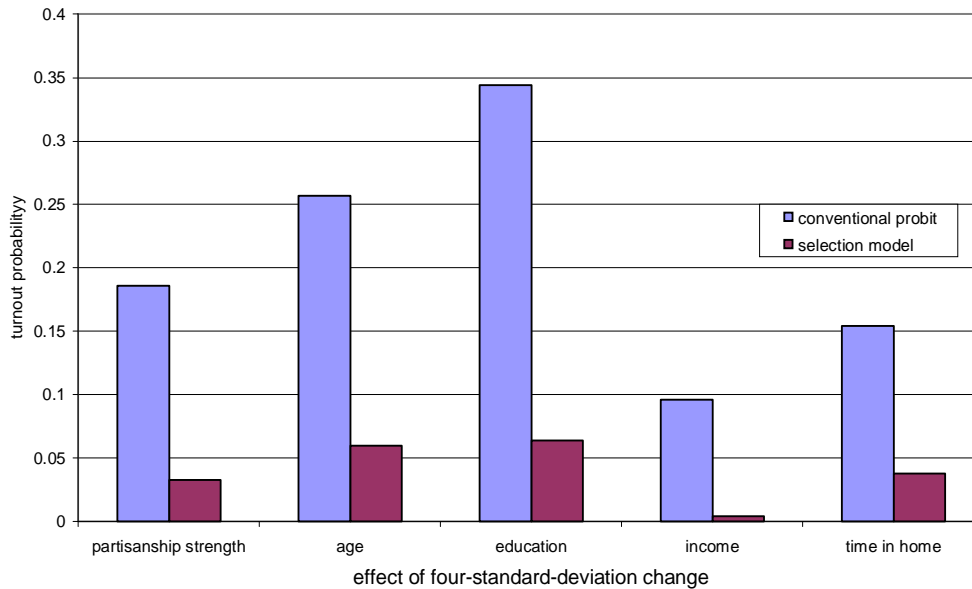


Figure 1.

Identification in selection bias models typically requires that some variable that influences selection have no effect on the outcome variable. Timpone used state-level differences in registration laws to generate the exclusion restriction: These laws presumably influence registration but not turnout. However, it is not obvious that such laws are exogenous where turnout is concerned. High-turnout states are those with a civic-minded culture and low corruption; they tend to have very liberal laws. (In the extreme cases, Minnesota and Wisconsin permit election-day registration, and North Dakota has no voter registration at all.) Lower-turnout states are more restrictive. Timpone does not employ state fixed effects, and his estimated impact of the differences in state laws is small even without them. It is not clear how well identification has been achieved, and the large standard errors in the outcome equation reinforce those concerns.

More importantly, Timpone’s selection equation is written as if everyone were required to register at each election. In his specification, the survey respondent’s current values on the explanatory variables influence the decision to register, and then they influence the decision to vote. But registration does not work that way for anyone but the newest voters. At the 2008 American presidential election, for example, some older voters will appear at the polls eligible to vote because they have been registered since 1952. In that year, they may have been excited about Dwight Eisenhower’s candidacy, and that got them to city hall to sign up. Modeling their registration status in 2008 using their current age and candidate enthusiasm does not get their life history right, and it constitutes a potentially serious specification error.

In sum, most statistical studies of American voter turnout have failed to deal with voter registration in a plausible way. Most studies ignore it, which would seem to induce bias. Timpone’s (1998) sophisticated attempt to model registration and voting jointly faced deep challenges of identification and specification, leading to results that seem doubtful. Perhaps for that reason, his work has had no successors, leaving us stranded econometrically. The remainder of this paper is “an attempt to fashion a tool” (Bentley 1967 [1908], unpaginated preface)—a tool that builds on substantive theoretical foundations in the EITM spirit (Granato and Scioli 2004).

## Empirical Foundations for the Registration Decision

Why do people register to vote? The standard view is that most people register because they anticipate that they will want to vote. The same underlying considerations drive both decisions. Squire, Wolfinger and Glass (1987, 51) give this reason for analyzing turnout directly, remarking that “registration and turnout are almost the same thing,” an insightful remark to which we return below.<sup>4</sup> Indeed, registration spikes right before presidential elections, illustrating the important role that major elections play in getting people registered. The New Jersey monthly registration counts for 2000-2006 are shown in Figure 2, with the two large spikes both occurring at the deadline for registration in presidential years, and the more intensely fought 2004 election showing the larger spike. The registration deadline in that year was October 4. Those who registered in September and the first four days of October, 2004, constituted more than a quarter (25.6%) of all voters registering since the previous presidential election in 2000.<sup>5</sup> People register primarily because they want to

---

<sup>4</sup>These authors also examine registration.

<sup>5</sup>This estimate is slightly too high, since it is based on the registration files in May 15, 2007. Some people who registered in 2001-2003 had already left the voter rolls by 2007, and thus are inappropriately excluded from the count of all who registered between 2000 and 2004. Back of the envelope calculations suggest that the correct proportion of new registrants is therefore probably one or two percentage points lower than the 25.6% given in the text.

vote.<sup>6</sup>

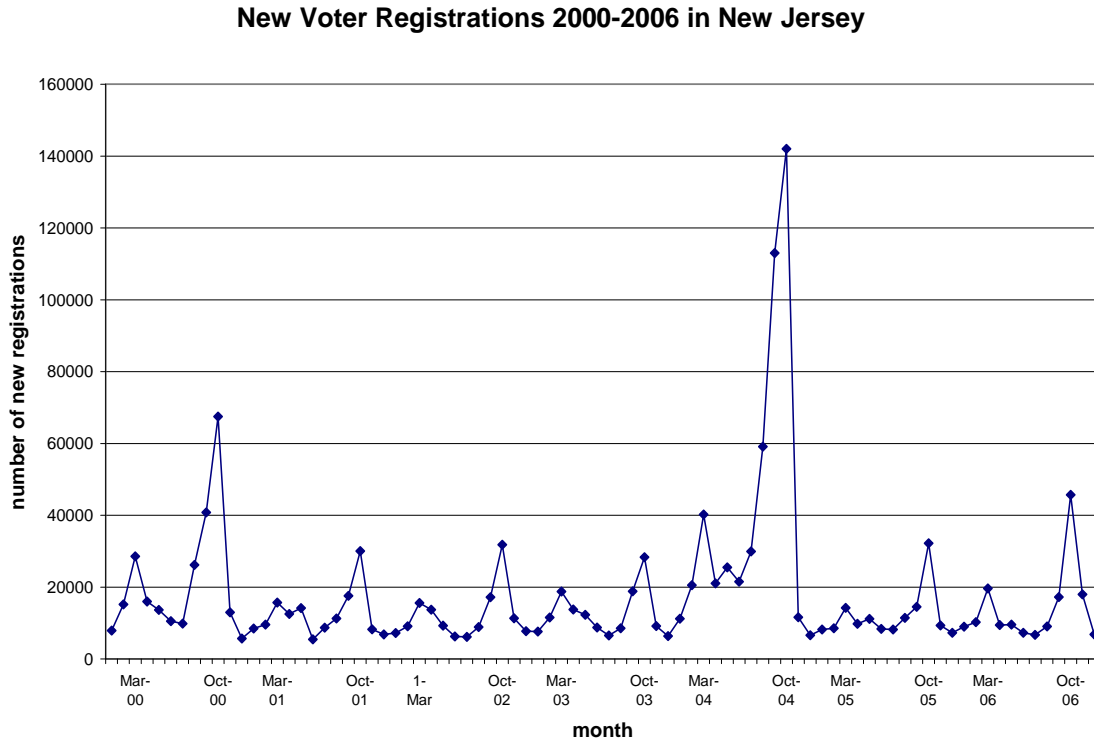


Figure 2.

Of course, registering is usually more difficult than voting, so that the two acts are not identical. Every election day, some ineligible people wish they had registered. But we surely expect the same motivational variables to influence both decisions in similar ways.

If differences in cost were the only distinctions between registration and voting, we would expect that nearly all registrants, particularly those who register right before a major election, would always vote. If their motivation to vote was higher than the cost of registering, it should also be higher than the more minor cost of voting. However, this

---

<sup>6</sup>Of course, party and interest group mobilization gets some people registered near elections. But as Figure 2 below shows, recent registrants get to the polls at a slightly higher rate than those who registered in prior months or years, indicating either that they are primarily self-motivated or that the mobilizers can substitute their own motivation for that of the registrants.

simple rational-choice syllogism, which would be true in a static world, turns out to be false in a more dynamic reality. Figure 3 shows the 2004 turnout rates among New Jerseyans who registered in each month of 2004 prior to the election. About one in every six new registrants (17%) failed to appear at the polls, only a bit better than the 19% of other registrants who failed to appear. Obviously, between registration and election day, other factors intervened—perhaps a loss of enthusiasm for the candidates, unexpected business travel, the birth of a child, family illness, or just forgetfulness. (See Sinnott 2009 for a list of the colorful reasons that registered Irish voters give for not voting.) Party workers who helped with registration may also have failed to be equally helpful with voting. In all such cases, we need to allow for unmeasured factors influencing the turnout decision that were not present when the citizen chose to register.

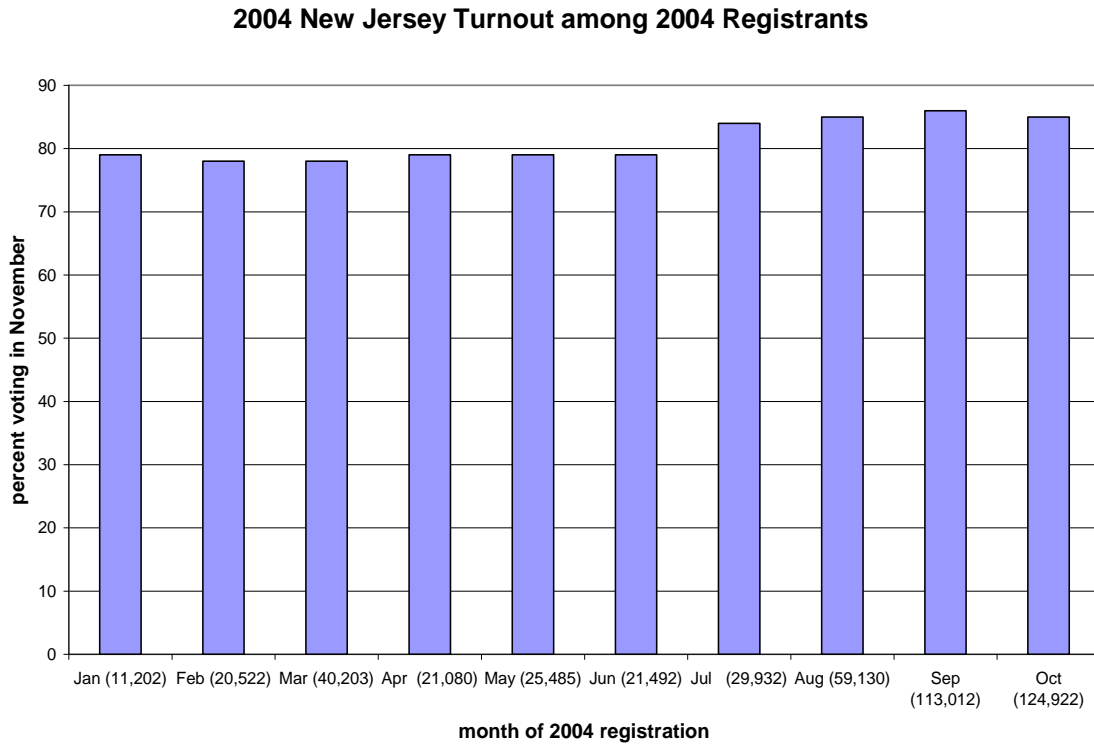


Figure 3.

Third, the most dramatic aspect of voter registration is its relationship to age. Rela-

tively few young people are registered; large majorities of the middle-aged are. Figure 4 shows the proportion of New Jersey citizens registered by age.<sup>7</sup> The strong upward trend demonstrates that the social norm of registering is learned, that learning can require 20 years or more before it takes hold, and that some citizens never learn the norm.

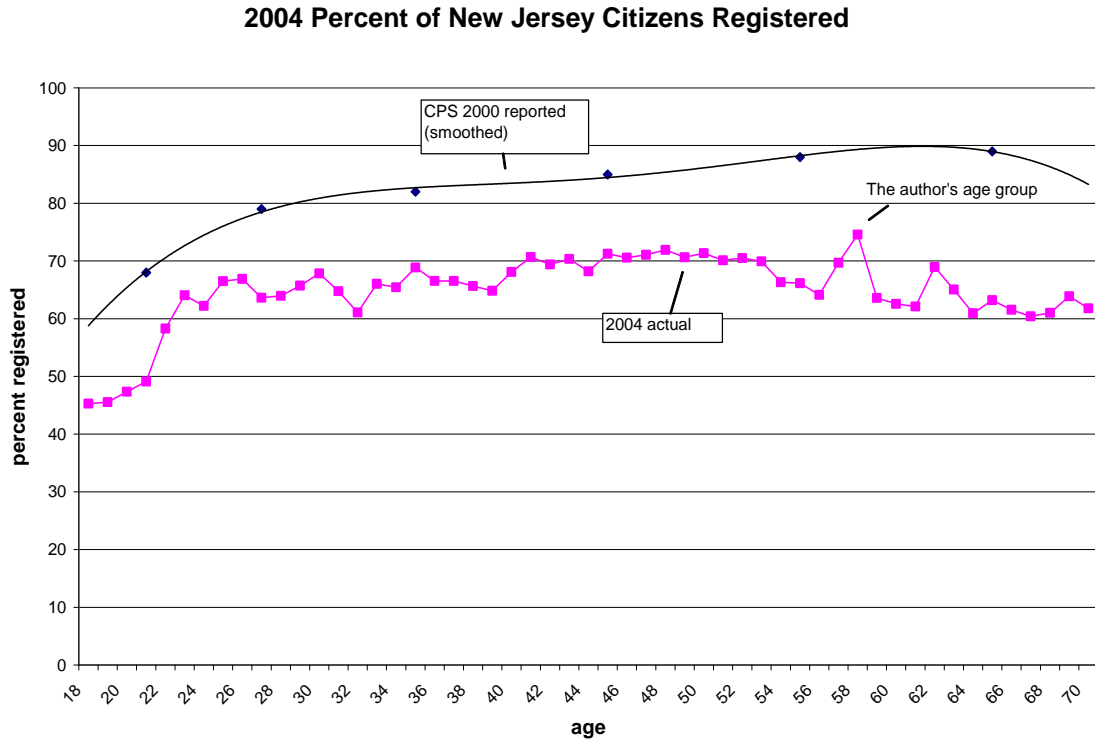


Figure 4.

<sup>7</sup>This figure was computed from the official New Jersey registration list in March, 2007. The firm Labels and Lists then attempted to delete “deadwood” (those who had died or moved), based on Social Security Administration death notices and United States Postal Office change-of-address forms. Labels and Lists’s record of actual New Jersey voters by age is the basis for the numerators in Table 1. The denominator for each age group is taken from the U.S. Census Bureau’s PUMS file for 2004, with non-citizens removed. Because the PUMS file is a sample, and because the data are taken from 2004 rather than the 2007 base for the registered voters, the denominators are excellent estimates but not exact counts.

Felons are ineligible to vote in New Jersey, though they can regain their voting rights after a certain number of years. They are relatively few in number as a proportion of the eligible population, and in any case, no count of them by age exists. Thus I have made no attempt to remove them from the denominators in Figure 3.

**New Jersey 2006 Registered Partisans by Age**  
(n = 4,135,104)

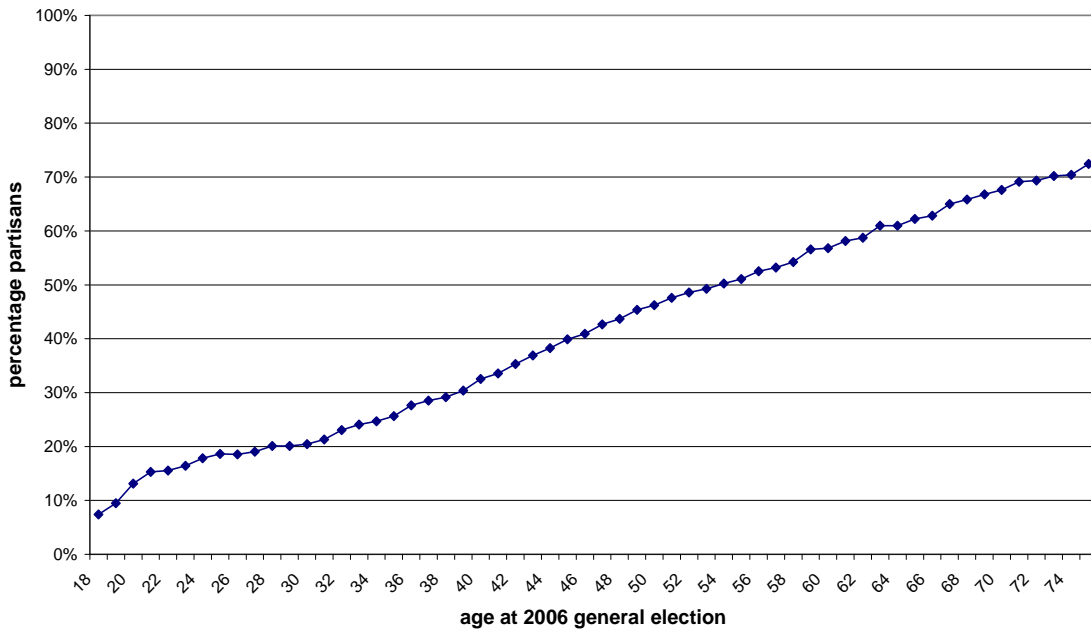


Figure 5.

Unlike the claims in *The American Voter* (Campbell *et al.* 1960, 496), the increase in registration with age is not due entirely to the growth in partisanship over the lifecycle. Partisanship does indeed rise with age, as *The American Voter* and many studies since have demonstrated. Figure 5 shows the corresponding relationship in the current New Jersey voter registration file. But Independents are influenced by age, too. As Figure 6 shows, even among Independents who have gotten themselves registered, voting rates rise with age.<sup>8</sup> Nor is this a peculiar feature of New Jersey politics, like “walking around money” on election day. The same pattern appears in the Annenberg 2000 election study data, even when “Independent” is defined as a pure Independent, with no partisan leanings

<sup>8</sup>Since the party affiliation of unregistered citizens is unknown, the New Jersey voting file cannot be used to assess how registration rates among Independents rise over time. Raising similar problems, the CPS survey by the Census Bureau does not ask party affiliation.

toward either party. Just as the tide raises all boats, age raises voting rates in all party identification groups.

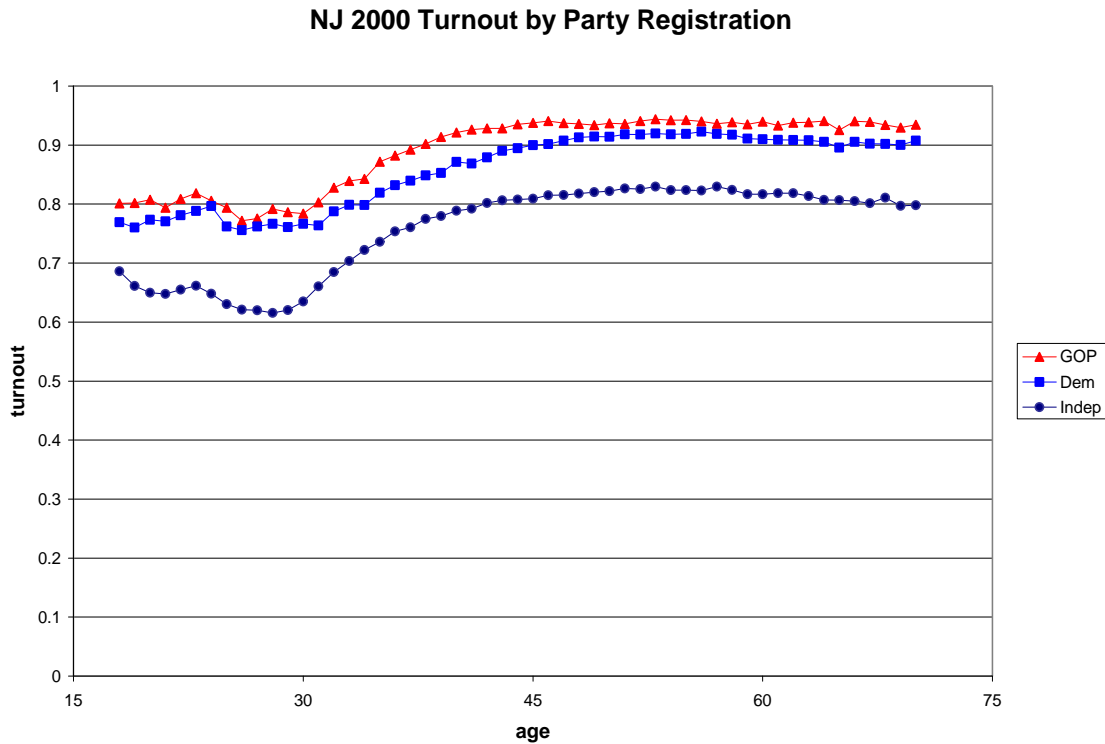


Figure 6.

Fourth and finally, a sensible statistical model must take account of the deleterious impact on voter registration of residential mobility. Americans change their living place at rates that would be astonishing in Europe. Figure 7 shows the percentage of New Jersey residents who reported in the 2000 Census that they had lived somewhere else five years before. Especially among young adults, mobility is the norm. American voters must re-register each time they move, and it is well established that mobility reduces registration (Squire, Wolfinger, and Glass 187). Thus the American tendency to pull up stakes and try

the luck someplace else is a key factor in reducing American turnout.

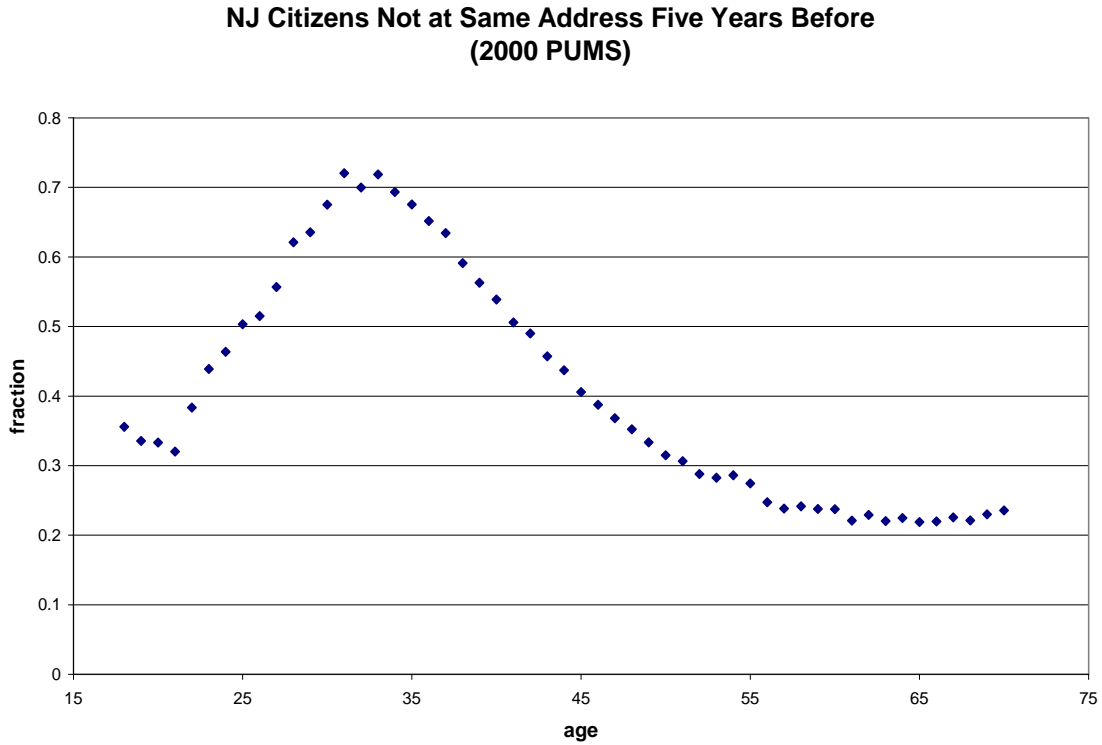


Figure 7.

## Do Conventional Registration Selection Models Work in Theory?

How can the available sources be used to assess the impact on turnout of the registration hurdle? To do so, we make the rational expectations assumption that the decision to register is based on the expectation of wanting to vote in subsequent elections. Thus to derive the registration equation, we must begin with the voting decision. Suppose then that the citizen's value of voting in subsequent elections  $y_{2i}^*$  is given by the standard linear regression equation:

$$y_{2i}^* = \beta_{02} + X_{1i}\beta_1 + u_{2i} \quad (1)$$

where for all  $i$  ( $i = 1, \dots, n$ ),  $\beta_{02}$  is a constant term,  $X_{1i}$  is a  $k$ -dimensional row vector of explanatory factors for individual  $i$ ,  $\beta_1$  is the corresponding coefficient (column) vector, and  $u_{2i}$  is a disturbance term distributed normally (Gaussian) and independently of  $X_{1j}$  for all  $j$ :  $u_{2i}|X_{1j} \sim N(0, \sigma_2^2)$ .

Now suppose that the citizen's information set includes all aspects of Equation (1) except the disturbance term, which represents factors unforeseeable in advance (such as whether it will be raining on election day, or whether the citizen will be in the hospital that day). However, the citizen also knows that registering is harder than voting by the amount  $\delta$ . Hence the expected value of registering is the same as the value of voting, less the extra cost of registering:

$$E(y_{2i}^* - \delta) = \beta_{02} - \delta + X_{1i}\beta_1 = \beta_{01} + X_{1i}\beta_1 \quad (2)$$

where  $\beta_{02} = \beta_{01} + \delta$ . Adding a standard normal disturbance  $u_{1i}$ , independent of all  $X_{1j}$ , to fix the utility scale and account for unmeasured factors unrelated to the desire to vote gives the value of registering  $y_{1i}^*$  as:

$$y_{1i}^* = \beta_{01} + X_{1i}\beta_1 + u_{1i} \quad (3)$$

As a starting point for the analysis, we assume that  $\text{cov}(u_{1i}, u_{2i}) = 0$ , so that the unmeasured factors inducing the citizen to register in, say, the February of ten years ago are uncorrelated with the unmeasured factors that propel her to the polls this current November. As Figure 3 confirms, there is no perfect correlation between the registration decision, even a registration decision just one month before the election, and the actual vote on election day.

This last equation embodies the critical rational-expectations assumption of this paper: *Any variable that influences voting, and that is foreseeable at the time of registration, will influence registration, too.* The fundamental explanatory factors will influence both in the same way. Thus there are no obvious exclusion restrictions in the turnout equation—variables that influence registration but not turnout. Differences in the institutional features of registration machinery are the only plausible exclusions (as in Timpone 1998), but these are under suspicion as endogenous, as noted above, and they do not apply to within-state samples. In sum, this is a difficult econometric context for conventional analyses.

Now both registration  $y_{1i}$  and turnout  $y_{2i}$  are dichotomous variables, and the underlying utilities that determine them are not observed. Instead we observe only the registration decision, and then, only if the citizen is registered, we observe the turnout decision<sup>9</sup>. For

---

<sup>9</sup>That is, if the citizen is not registered, we do not observe whether the citizen had a (frustrated) desire

registration:

$$y_{1i} = \begin{cases} 1 & \text{if } y_{1i}^* \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

and for turnout:

$$y_{2i} = \begin{cases} 1 & \text{if } y_{1i} = 1 \text{ and } y_{2i}^* \geq 0 \\ 0 & \text{if } y_{1i} = 1 \text{ and } y_{2i}^* < 0 \\ \text{unobserved} & \text{if } y_{1i} = 0 \end{cases} \quad (5)$$

Now to identify a selection model, it is necessary in practice to find a measured variable that influences registration but not turnout.<sup>10</sup> But that is not easily done: Rational-expectations registration implies that the same variables causing the vote will also cause registration. Thus the possible exclusion restrictions are likely to be false.

Suppose then that we divide  $X_{1i}$  into those variables  $x_{1i}$  believed (correctly) to influence both registration and turnout, and those variables  $x_{2i}$  believed (falsely) to influence only registration:

$$y_{1i}^* = \beta_{01} + x_{1i}\beta_{11} + x_{2i}\beta_{12} + u_{1i} \quad (6)$$

In this spirit, suppose that the analyst mistakenly believes that the correct model for the vote excludes  $x_2$ , so that the assumed model for the desire to vote  $y_{2i}^*$  (before any conditioning on prior registration) is:

$$y_{2i}^* = \hat{\beta}_{02} + x_{1i}\hat{\beta}_{11} + \hat{u}_{2i} \quad (7)$$

What happens when a selection model is applied? Assessing the resulting biases is not easy in this two-equation nonlinear system with dichotomous outcomes. To get some intuition, we will analyze the simpler, but closely related, case in which  $y_{2i}^*$  is observed. (Note that if the specification were correct, the same coefficients would emerge from this version of the model as in the bivariate probit setup, since maximum likelihood is consistent for both.) In that instance, the standard Heckman two-step correction can be applied; it is known to be consistent when the specification is correct. What does it converge to in this case, when the specification is wrong in the manner we expect for registration and voting data?

Equations (3) and (4) define a standard probit setup for the registration decision. Then

---

to vote on election day, or whether she remained unmotivated.

<sup>10</sup>“In practice,” since in principle, if the distributional assumptions are correct, the two equations are identified solely “from the curvature,” i.e., from the nonlinearity in the probit functions. But since that nonlinearity is very slight except for probabilities near zero or one, and since our confidence in the normality assumptions is weak (and thus we do not know the true shape of the nonlinearity), achieving formal identification without an exclusion restriction risks serious biases in the estimates.

for the turnout decision, it follows that if the value of voting  $y_{2i}^*$  were observed, its proper specification conditional on the selection process defined by the decision to register would be:

$$y_{2i}^* = \beta_{02} + x_{1i}\beta_{11} + x_{2i}\beta_{12} + u_{2i} \quad (8)$$

Note that no term involving the expected value of the selection equation (registration) enters here, since by assumption,  $\text{cov}(u_{1i}, u_{2i}) = 0$ .<sup>11</sup> However, the analyst knows nothing about that, and so estimates a conventional two-step selection model with  $x_{2i}$  excluded and the usual inverse Mill's ratio included:

$$y_{2i}^* = \hat{\beta}_{02} + x_{1i}\hat{\beta}_{11} + \hat{\gamma}\lambda_i + \hat{u}_{2i} \quad (9)$$

where as usual,  $\lambda_i = \phi_i/\Phi_i$ , and  $\phi_i$  and  $\Phi_i$  are the density and cdf, respectively, of the standard normal distribution, each evaluated at  $z = \beta_{02} + X_{1i}\beta_{11}$ . In addition,  $\gamma = \rho\sigma_2^2$ , where  $\rho$  is the Pearson correlation between  $u_{1i}$  and  $u_{2i}$ . Here the true value of  $\rho$  is zero, so that in actuality  $\gamma = 0$ .<sup>12</sup>

What happens when Equation (9) is estimated by the customary ordinary least squares (OLS) method? To achieve analytic tractability, we approximate  $\lambda_i$  by  $\kappa^*(1 - \beta_{01} + x_{1i}\beta_{11} + x_{2i}\beta_{12})$ , where  $\kappa^*$  is a scaling factor. As Figure 8 shows, this approximation is quite close for selection (registration) probabilities  $\Phi_i(z)$  between 10% and 90% (see also Achen 1986,

---

<sup>11</sup>A rigorous rational expectations framework would derive the vanishing correlation of the disturbances from the observation that if the correlation were non-zero, that fact would be taken into account at the time of registration. The intercept in the registration equation would be adjusted to accommodate the expected value of  $u_2$ . The new disturbance terms would then be uncorrelated.

<sup>12</sup>Note that in practice,  $\lambda_i$  would be estimated, not known. Asymptotically this makes no difference to the coefficient estimates, and so the “hat” on  $\hat{\lambda}_i$  has been suppressed here and in Appendix 1 for simplicity.

104.)

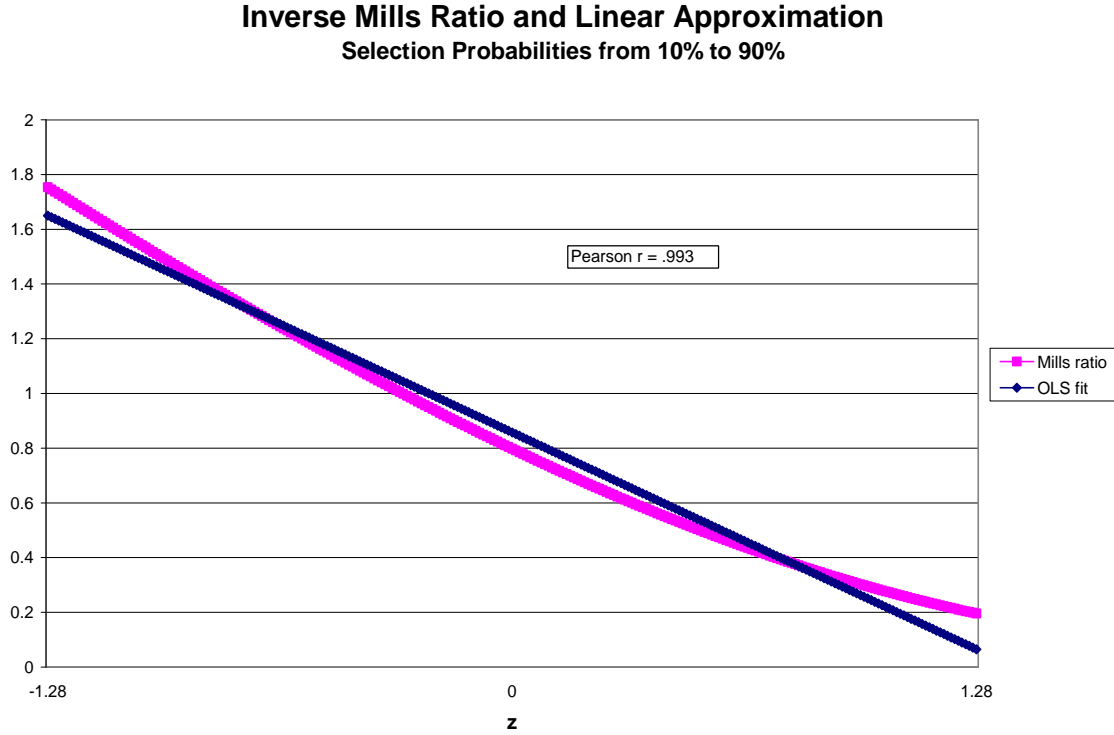


Figure 8.

(10)

Letting  $\hat{\kappa} = \kappa^*/\hat{\gamma}$  and rewriting Equation (8) then gives:

$$y_{2i}^* \approx \hat{\beta}_{02} + x_{1i}\hat{\beta}_{11} + \hat{\kappa}(1 - \beta_{01} - x_{1i}\beta_{11} - x_{2i}\beta_{12}) + \hat{u}_{2i} \quad (11)$$

Here OLS will find the best linear fit, using the true value of  $y_{2i}^* = \beta_{02} + x_{1i}\beta_{11} + x_{2i}\beta_{12} + u_{2i}$ . In particular,  $\hat{\kappa}$  will be chosen to pick up the effect of the (falsely) excluded variables in  $x_{2i}$ , which requires  $\hat{\kappa} = -1$ . But then to maintain the correct weight on  $x_{1i}$ ,  $\hat{\beta}_{11}$  must be set to zero. Thus intuition suggests, and the appendix demonstrates formally, that to the degree of approximation represented by Equation (10):

$$E \begin{bmatrix} \hat{\beta}_{11} \\ \hat{\kappa} \end{bmatrix} \approx \begin{bmatrix} 0 \\ -1 \end{bmatrix} \quad (12)$$

That is, the implied correlation between the disturbances will be highly negative, and the estimated effect of the included variables will be near zero.<sup>13</sup> That is qualitatively what Timpone (1998) found, using the slightly more complex but entirely parallel estimator in which  $y_{2i}$  is observed but  $y_{2i}^*$  is not.<sup>14</sup>

As an irony, note that if the entire problem of selection is ignored, as so many analysts have done following the wise advice in Squire, Wolfinger and Glass (1987, 51), then the dependent variable being analyzed is  $y_{3i}$ , a dichotomous variable signifying whether the citizen voted or did not, with the latter category including both registered and unregistered nonvoters. Consider the stylized case in which each voter has had many prior opportunities to register, so that anyone who wishes ever to vote has at some prior point had sufficient enthusiasm to register, while those who never wish to vote have not. In that case, registration and voting are synonymous. As Figure 4 shows, that assumption is quite consistent with the evidence for New Jerseyans over age 35. By the rational-expectations hypothesis, the explanatory variables for registration and voting are identical. Then the usual probit analysis of the vote decision, with the selection problem due to registration ignored entirely, is the correct approach and will give the true coefficients asymptotically! Meantime, the seemingly more sophisticated conventional modeling of the selection process will bias the turnout coefficients badly.

Now of course, the statistical setup used here to illustrate the point, while it is a conventional selection model, is quite imperfect as a representation of registration and voting: The true registration and voting decisions do not occur in immediate temporal proximity even for the young, and they may occur at quite different stages of life for the middle-aged.<sup>15</sup> Plus the correlation between the disturbances is not zero in practice. And there are other real-world differences from the theoretical setup used here. Thus our stylized finding that justifies ignoring registration entirely (for those past young adulthood), as most turnout researchers do, is not precisely correct, and a better estimator is needed. One such approach is set out below. First, however, we explore the possibility that ignoring registration is a better approach than using doubtful exclusion restrictions to estimate conventional selection models. The next section tests this judgment with turnout data from the Dakotas.

---

<sup>13</sup>Less consequentially, the intercept is also biased, in a way easily seen from Equation (9).

<sup>14</sup>With some additional assumptions, such as  $x_{2i}$  being distributed as a multivariate normal, parallel versions of all these results would be derivable from Timpone's specification.

<sup>15</sup>The woman in front of me at the polling place during the New Jersey presidential primary in February, 2008, asked to vote in the Democratic primary. She was told that she could not, since she was a registered Republican. "But I registered twenty years ago," she said. "I was a completely different person then."

## Do Registration Selection Models Work in Practice?

North Dakota and South Dakota are vertically adjacent American states of the Great Plains. They are each agricultural regions, with little immigration and few minorities. In the 2000 U.S. Census, North Dakota was 91% non-Hispanic white, while the corresponding figure for South Dakota was 87%. Less than 2% of each state was foreign-born. The only sizable minority is the Native American population, amounting to 5% of North Dakota and 8% of South Dakota.

The two states are also ethnically homogeneous due to the settlement of the (united) Dakota Territory by northern European immigrants, especially Germans and Norwegians, whose descendents constitute more than half of each state. For the same reason, more than half of each state is either Catholic or Lutheran, and those two denominations are represented in approximately equal numbers in both states. Per capita income (in thousands) in 1999 was \$17.8 in North Dakota and \$17.6 in South Dakota, both well below the national average. In sum, these two states are unusually closely matched, particularly with respect to their white populations.

In one respect, however, the two states differ. South Dakota has a conventional American voter registration system. North Dakota abolished voter registration after World War II and currently has no registration system at all. Any citizen can show up at the polls on election day and vote if she can demonstrate that she has been a resident for 30 days. Thus a comparison of the two states offers an attractive opportunity to assess how well our various estimators perform. North Dakota's turnout can be modeled without reference to the selection effect of registration, since it has no registration. In South Dakota, alternate estimators can be applied to see how well they produce something like the North Dakota pattern in explaining desire to vote. Of course, no one expects the two states to have precisely identical coefficients in properly estimated vote equations, but they should be close. As a hint of what to expect, Figure 9 shows the turnout pattern by age for each state. The sample sizes permit only suggestive inferences, but the young and early middle-aged seem to vote more in North Dakota. There is no apparent substantial difference among those

over 50.

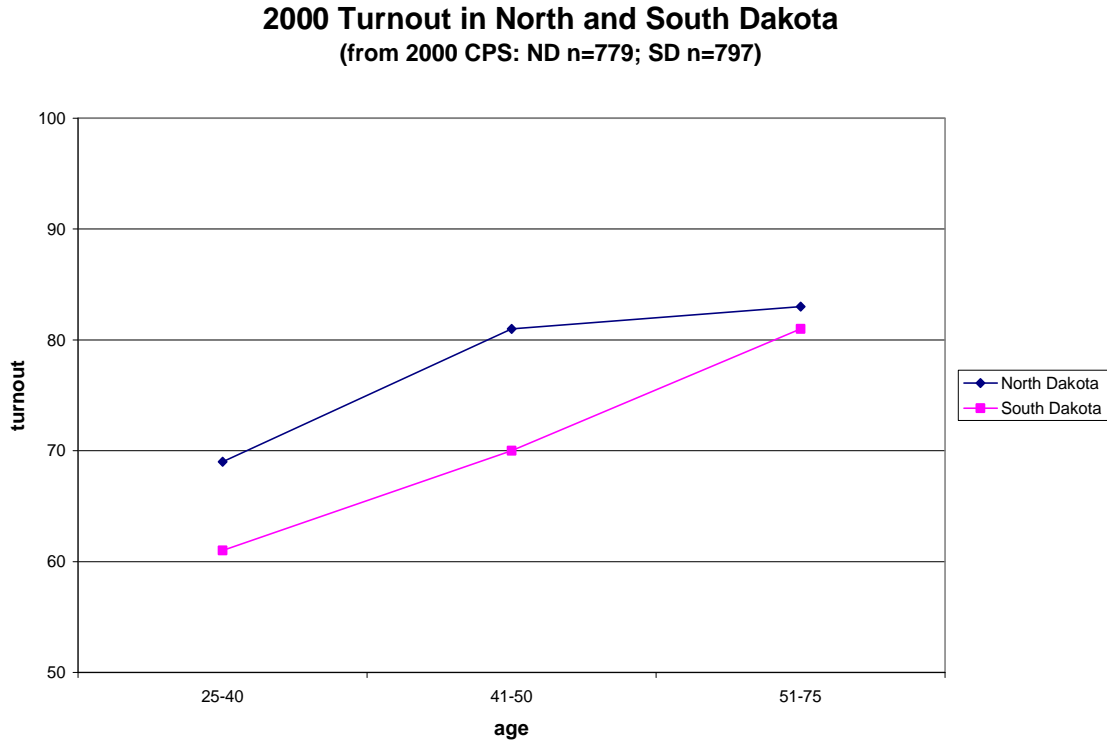


Figure 9.

For the following investigations, we use the Census Bureau’s 2000 Current Population Survey for each state. The sample is confined to native-born non-Hispanic whites to eliminate the minor differences in minority and immigrant populations and to enhance the causal homogeneity of the sample. Those under the age of 25 are dropped due to the complexities of voter turnout for those away at college or in the armed forces. Those older than 70 are also dropped because American turnout begins to drop in the early seventies due to frailty, and we wish to avoid specifications that require medical diagnoses, for which political scientists are unqualified. Those who refused to answer the voter turnout question were eliminated. To preserve sample consistency with subsequent estimations for which residential mobility is needed, four North and South Dakota citizens missing that information were also dropped.

In North Dakota, there were 783 respondents meeting these constraints, of whom 602 voted (77%).<sup>16</sup> Once again, there is no reason to worry about the selection effect of registration in this state. We began by estimating the effects of the usual powerful proxies for turnout, age, age-squared, and education (six categories),<sup>17</sup> on North Dakota voter turnout, using probit analysis. However, neither here nor in any other analyses of this paper was the customary age-squared term statistically significant in North or South Dakota, and its joint effect with age was sometimes nonsensical. It has therefore been dropped in what follows.

Denoting age by A and education by E, the result is (with conventional asymptotic standard errors in parentheses, log likelihood = -384.91941):

$$\begin{aligned} \text{ND Turnout} = & -2.0468 + .02755 * A + .4159 * E & (13) \\ & (.3451) \quad (.004873) \quad (.05439) \end{aligned}$$

This set of estimates confirms the pattern of Figure 9, with all coefficients comfortably statistically significant. First, age makes some difference in North Dakota, though less than in other parts of the U.S. Moving from age 25 to age 70 changes the predicted value by about one and a quarter units on the probit scale—the difference between 60% and 93% turnout, for instance. Education is a powerful predictor of turnout in this state with no registration requirement, both substantively and statistically, thus contradicting the claims of some qualitative students of turnout who have argued that class bias in turnout stems from registration requirements. Here the implied effect of moving from one end to the other on the six-point education scale is more than two units—the difference between 50% and 98% turnout.

In South Dakota, on the other hand, registration is required to vote. As a first step, we ignore that fact, as turnout researchers tend to do, and simply report the same probit estimates as in the North Dakota case, using a turnout variable that makes no distinction between abstention by those who were registered and abstention by those who were not registered. In South Dakota, the sample consists of 800 native-born non-Hispanic white individuals aged 25-70, of whom 561 voted (70%).<sup>18</sup> The result of the probit estimation is

---

<sup>16</sup>Among the 210 North Dakota abstainers, 89 said that they did so because they weren't registered! These citizens were disproportionately, though by no means entirely, relatively recent arrivals in the state. Thus as other studies have demonstrated, self-reported registration has many of the same difficulties as self-reported turnout.

<sup>17</sup>The categories are (1) 0-8 years, (2) 9-11 years, (3) high school graduate or equivalency certificate, (4) 1-3 years college, including associate degree, (5) college graduate, and (6) graduate degree.

<sup>18</sup>The average effect over all white non-Hispanic citizens of having registration is not a very interesting statistic here, since the impact almost certainly varies by age, education, partisanship, and other important factors. But it could be estimated by a matching method, and since the two states have very similar

(log likelihood =-435.55098):

$$\begin{aligned} \text{SD Turnout} = & -2.1686 + .03201 * A + .3440 * E & (14) \\ & (.2821) \quad (.004308) \quad (.04510) \end{aligned}$$

The remarkable aspect of these South Dakota estimates is their closeness to those of Equation (13). None of the coefficients is distinguishably different statistically from its North Dakota counterpart, and the implied effects are very similar in spite of the distinction between their registration requirements. The largest difference is that the point estimate for education is lower in South Dakota, but this is just the reverse of what we would expect if registration requirements were keeping its less educated citizens from the polls, which suggests that the state coefficient differences may just be estimation noise. In sum, these are the sorts of estimates we would expect if we had estimated the correct South Dakota selection model for registration, and then estimated its vote equation conditional on registration. The coefficients look like North Dakota's, as they should if we had corrected for selection. But in fact, this is the wrong model: We have ignored selection entirely.<sup>19</sup>

What happens if we estimate a selection model for South Dakota, in parallel to Timpone (1996)? With the correct selection model, we hope to get a vote equation that is nearly the same as the North Dakota vote equation, where selection is not an issue. To estimate such a model, we need an exclusion restriction, but we might argue in the informal style of 1980s structural equation estimation, for example, that residential mobility affects the probability of registering, but in a presidential election, it should have no effect on turnout, since people will want to vote regardless of whether they have moved. This assumption is false, as Figure 10 demonstrates: Residential mobility damages both registration and turnout. Hence the setup here approximates the model considered in the text earlier and in the appendix, in which a fallacious exclusion restriction is employed to identify a

---

demographics, would necessarily be approximately 4 percentage points, the raw difference in non-Hispanic white turnout between the two states. This estimate is approximately the same, though a bit higher, than Hanmer's (2004) estimate of the effect of election day registration.

<sup>19</sup>Uhlaner (1989) estimates an ordered probit model whose three categories are (1) not registered, (2) registered but did not vote, and (3) voted. When estimated with South Dakota data, its coefficients are quite close to those of this equation, and in that sense, it, too, comes closer to getting the right answer than the seemingly more sophisticated selection model presented below.

two-equation selection model.

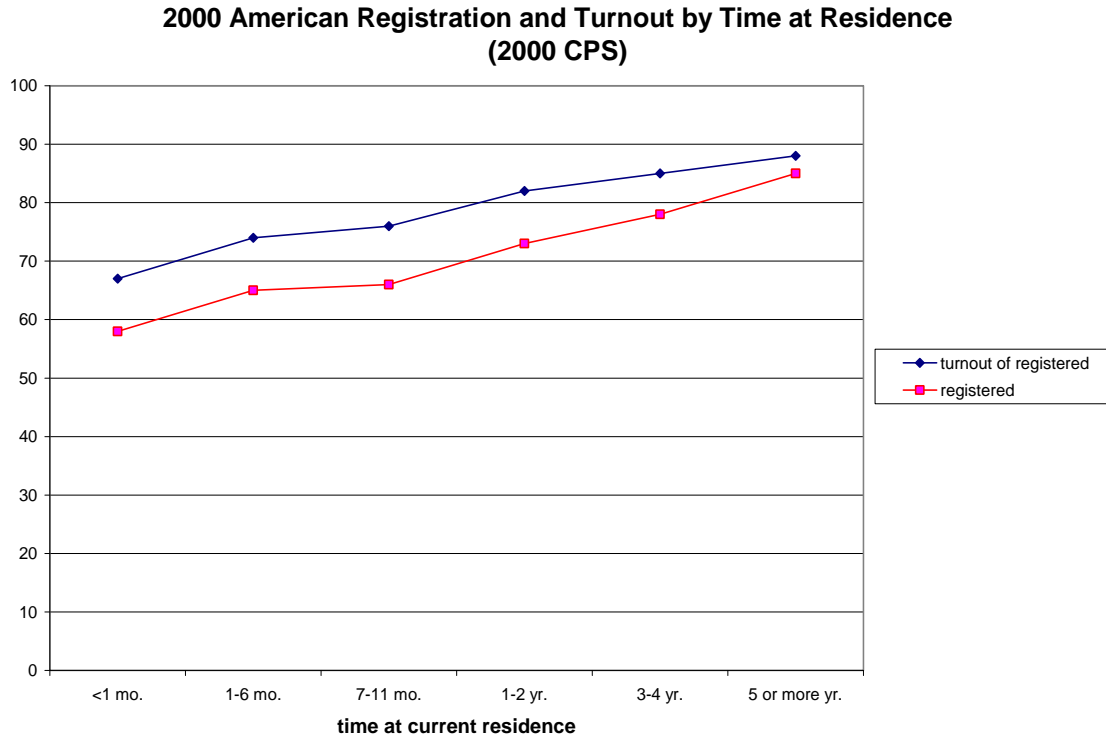


Figure 10.

The South Dakota sample consists of the same 800 observations as before, of whom 150 are unregistered and thus censored. Four non-voters who did not know or refused to say whether they were registered have been coded as registered.<sup>20</sup> We follow the Census Bureau in defining mobility as a six-category variable ranging from less than one month in the current residence through five years or more.<sup>21</sup> Denoting residential mobility by  $M$ , the resulting registration (selection) equation is (log likelihood = -566.9928)<sup>22</sup>:

<sup>20</sup>Counting them as unregistered or dropping them makes little difference. Either way, the coefficients in the outcome equation become uniformly a bit smaller, strengthening the argument made below.

<sup>21</sup>The categories are (1) less than a month, (2) 1-6 months, (3) 7-11 months, (4) 1-2 years, (5) 3-4 years, (6) five years or more.

<sup>22</sup>The model is the bivariate probit selection model pioneered by Dubins and Rivers (1989/1990), as implemented in the procedure *heckprob* in the STATA computing package (version 9.2).

$$\begin{aligned} \text{SD Registration} = & -2.9981 + .02873 * A + .4139 * E + .2274 * M & (15) \\ & (.3459) \quad (.005175) \quad (.05251) \quad (.04127) \end{aligned}$$

Recall that this equation is not the one affected by the bias calculations in the prior section: Its coefficients should be sensible. Indeed, it looks much like the turnout equation for South Dakota, just as the rational expectations idea suggests. Here residential mobility clearly matters, but the age coefficient is within a standard error of its value in the vote equation; and the education coefficient is about 20% larger, not statistically significantly so.

What this equation is most like, however, is the North Dakota vote equation. Apart from the lower intercept in South Dakota, the coefficients are virtually identical. Again, if North and South Dakota native-born whites are similar, and if the argument of the previous section is correct, then the vote equation in a state without registration should be nearly identical to the registration equation in a state that requires it. The difference should be a lower intercept in the registration equation, reflecting the greater cost of registering relative to voting. That is exactly what we find here.

Next, the outcome equation for the South Dakota vote, with the selection effect of registration taken into account:

$$\begin{aligned} \text{SD Turnout} = & .3481 + .008941 * A + .1272 * E & (16) \\ & (.3179) \quad (.004884) \quad (.05300) \end{aligned}$$

Although both substantive coefficients are statistically significant, their estimated effects are about 1/3 of their effects in North Dakota, and the differences from North Dakota are statistically significant. In addition, the estimated correlation between the disturbances of the selection and outcome equations is  $\hat{\rho} = -.9996$ , again a wildly negative correlation in a context in which the same unmeasured factors must surely be influencing both registration and turnout in the same direction.

Let us summarize these results. First, naively ignoring the selection effect of registration in South Dakota gave plausible estimates. Second, when the bivariate probit selection model was applied, the consequence was vanishing coefficients in the turnout equation and the negative correlation between the disturbances, just as Timpone (1996) found. And third, good estimates from the naive specification and mis-estimates from the seemingly sophisticated model are qualitatively just what the previous section of this paper predicted would occur if the decision to register is made because people intend to vote.

Thus doing what researchers have long been doing—ignoring registration—makes more sense than any statistical correction we currently have. Registration and voting are so closely related that we bias our findings only somewhat in doing so. However, as Figures 6 and 10 demonstrate, and as the differences between Equations (14) and (16) for South Dakota hint, the two decisions are not identical. Turnout varies with partisanship and age even among the registered. Furthermore, even among the registered, mobility reduces turnout, as the increasing gap toward the bottom left of Figure 10 shows. This suggests the non-obvious idea that over and above its effect on registration, residential mobility is a main factor keeping even registered voters from getting to the polls: Testing that notion would require us to tease apart mobility’s impact on registration from its impact on the vote net of registration, i.e., to estimate a good selection model. With our current tools, we cannot. Thus in the longer term, modeling registration and turnout jointly will be needed for full confidence in our understanding of American voting. Thus we need a new model of the registration decision, one that takes account of the multiple opportunities to register that citizens enjoy. The remainder of this paper works toward that intermediate goal. Subsequent versions of the paper will attempt joint modeling of registration and voting in a full selection model. But the critical first step is to get the registration model right.

## A Statistical Model for American Voter Registration

As noted earlier, a key feature of the voluntary American registration process is that at each election, most citizens have had many prior opportunities to become registered. If, as Figure 2 suggests, we should make the unit of observation the four-year interval leading up to each presidential election, then citizens 21 years of age or younger at the time of a presidential election have had one interval in which to register, those 22-25 years of age have had two such intervals, and so on. People 54 years of age and older have had 10 or more quadrennial intervals in which to become registered. Those 74 and older have had 15 prior opportunities.<sup>23</sup> Hence modeling the registration decision in a multivariate probit setup, under which the citizen is unregistered if she has registered at none of the prior opportunities, would require computing integrals of 10 or more dimensions for a substantial fraction of the population. Due to unobserved heterogeneity, which is substantial in registration and turnout models (see below), the registration decisions would be correlated in unmeasured ways. Thus the integrals would not simplify to products of univariate integrals.

The resulting probabilities certainly appear computable by simulated moments if the number of unique combinations of the covariates is not too large.<sup>24</sup> However, as will be seen,

---

<sup>23</sup> Actually, 74 year olds in 2008 have had only 14 prior opportunities in the current U.S., since the voting age had not yet changed to 18 when these citizens began voting. The change occurred in 1971.

<sup>24</sup> I thank Simon Jackman for a consultation on this point.

the simplest multivariate probit models would not fit the data; mixtures of densities would have to be used. And the entire registration model would have to be linked to another statistical equation, too, where the vote decision is modeled, probably with a nonlinear right-hand-side specification. With contemporary computing power available to individual researchers, estimation of such a model would certainly not occur in the blink of an eye. This is not an attractive prospect for day-to-day applied work, where models must be estimated many times to find our inevitable specification mistakes. The fact that the probit model is justified more by theoretical charm and long familiarity than by any verifiably close tracking of human choices strengthens this point.

Instead, we proceed using beta distributions to represent the distribution of unobserved heterogeneity in registration probabilities. The key first step in the modeling is to recognize the multiple opportunities each citizen has to register. Let  $\pi_i$  be the probability that an eligible citizen  $i$  desires to register at trial  $n$  ( $n > 0$ ), where  $\pi_i$  is assumed fixed over time and where  $n$  is interpreted as a count of the presidential elections at which the citizen has been eligible to vote. We refer to the  $n$ th election as “the current election.” Then the number of prior registration opportunities includes the current election, since registration for a given election precedes voting. We define  $y_{1i}$  as the citizen’s registration status at the current election:

$$y_{1i} = \begin{cases} 1 & \text{if citizen is registered at time } n \\ 0 & \text{otherwise} \end{cases} \quad (17)$$

We assume that the registration decisions at each time period is mutually independent of the others conditional on  $\pi_i$ , so that the appropriate distribution is binomial. We suppose that the citizen is currently registered if she has desired to register at the current period or at any prior time period. Hence a citizen encountering her  $n$ th presidential election is *unregistered* if she has never desired to register, which occurs with probability:

$$\Pr(y_{1i} = 0 \mid \pi_i, n) = (1 - \pi_i)^n \quad (18)$$

Now suppose that the empirical distribution of  $\pi_i$  is  $Beta(a, b)$ , so that its density is  $\pi_i^a(1 - \pi_i)^b/B(a, b)$ , where the beta function  $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$  and  $\Gamma(\cdot)$  is the gamma function. Then the proportion of citizens not registered at time  $n$  would be given by:

$$\Pr(y_{1i} = 0 \mid n) = \int_0^1 (1 - \pi_i)^n \pi_i^a(1 - \pi_i)^b/B(a, b) d\pi_i \quad (19)$$

But this equation may be rewritten as:

$$\Pr(y_{1i} = 0 | n) = \frac{B(a, b + n)}{B(a, b)} \int_0^1 \pi_i^a (1 - \pi_i)^{b+n} / B(a, b + n) d\pi_i \quad (20)$$

The integral is just the area under a  $Beta(a, b + n)$  density, and thus it equals one. It follows that the proportion not registered at time  $n$  is:

$$\begin{aligned} \Pr(y_{1i} = 0 | n) &= \frac{B(a, b + n)}{B(a, b)} \\ &= \frac{\Gamma(a + b) \Gamma(b + n)}{\Gamma(b) \Gamma(a + b + n)} \end{aligned} \quad (21)$$

This is a standard Bayesian result for the beta–binomial model. Prior empirical applications include Wiley *et al.* (1989), who develop the model to assess heterogeneous probabilities of acquiring HIV/AIDS in repeated sexual contact with an infected partner.

It follows easily from Equation (21) that  $\lim_{n \rightarrow \infty} \Pr(y_{1i} = 0 | n) = 0$ . That is, under this model the fraction unregistered declines monotonically to zero as people age. Everyone becomes a registered voter. Figure 4 makes it clear that the reality in New Jersey is quite different: Registration rises rapidly in the first few elections, indicating that many people in the population have relatively large  $\pi_i$ , but the fraction registered quickly asymptotes, leaving a substantial minority of the population untouched. For a variety of reasons, some people do not wish to be registered and virtually never experience a desire to register, no matter how long they live.

We take account of people not socialized into norms of democratic citizenship in the same way Wiley *et al.* (1989) might have coped with couples in their sample who were both uninfected: Their probability of acquiring HIV from contact with each other is zero. We apply the same logic to registration. Thus setting the unreachable fraction of citizens to  $k$ , and using the previous beta–binomial model only for the remaining  $1 - k$  fraction of the population, gives:

$$\Pr(y_{1i} = 0 | n, k) = k + (1 - k) \frac{\Gamma(a + b) \Gamma(b + n)}{\Gamma(b) \Gamma(a + b + n)} \quad (22)$$

This gives the probability of registration as:

$$\Pr(y_{1i} = 1 | n, k) = 1 - k - (1 - k) \frac{\Gamma(a + b) \Gamma(b + n)}{\Gamma(b) \Gamma(a + b + n)} \quad (23)$$

Note that these last two convenient results correspond to the difficult 10– or 15–dimension Gaussian integrals discussed earlier. The computational ease of the Beta mixing distribution for a binomial (the beta is a “conjugate prior”) recommends its use in problems of

this kind, as researchers studying human choices over time have long known (for example, Massy *et al.* 1970, 60-79 and ff.).<sup>25</sup>

The next step is to postulate a link function for the parameters of the beta-binomial model, so that the probabilities can be expressed as functions of covariates. If there were just one trial per citizen ( $n = 1$ ), then the previous equation reduces to:

$$\Pr(y_{1i} = 1 | n = 1, k) = (1 - k) \frac{a}{a + b} \quad (24)$$

It would be natural to model the fraction  $a/(a + b) = 1/(1 + b/a)$  with a logit link, setting the odds ratio  $b/a = \exp(-X_i\beta)$ , for a vector of covariates  $X_i$  and coefficient vector  $\beta$ . When everyone is susceptible to becoming registered ( $k = 0$ ), this is, in fact, just the usual logit model. Note that a parameter defined as  $c = a + b$  would not be identified when  $n = 1$  for all observations; however, it is identified here because  $n$  varies. Thus we have three parameters:  $k$ ,  $c$ , and  $b/a$ , with the latter modeled as  $\exp(-X_i\beta)$ . Solving gives  $b = c/[1 + \exp(X_{1i}\beta_1)]$ .

Define the usual logit function as  $\text{logit}(z) = 1/[1 + \exp(-z)]$ , so that  $c \text{logit}(-X_{1i}\beta_1) = b$ . Then substituting into Equation (23) gives the likelihood of observing a registered citizen:

$$\Pr(y_{1i} = 1 | n, k, X_i) = 1 - k - (1 - k) \frac{\Gamma[c] \Gamma[c \text{logit}(-X_{1i}\beta_1) + n]}{\Gamma(c + n) \Gamma[c \text{logit}(-X_{1i}\beta_1)]} \quad (25)$$

and similarly for  $\Pr(y_{1i} = 0 | n, k)$ . This is the general version of the specification for the registration decision used in this paper.

The model could be extended by letting  $k$  be a logit function of covariates, for instance. Graphical exploration of CPS samples for 2000, 2002, and 2004 seemed to show only weak differences in  $k$  (asymptotic non-registration rates) across education and age groups, however, so that the simpler model with fixed  $k$  was provisionally adopted.<sup>26</sup> Of course,  $n$  is just the number of trials and thus should not be made to depend on covariates.

We now apply this model to the South Dakota CPS data to assess its fit compared to that of conventional probit analysis.

---

<sup>25</sup>This setup gives the distribution of  $\pi_i$  as a mixture of two beta distributions, one of them a mass function at the point 0. Pairs or triples of non-degenerate betas could also be used to extend the generality of the model.

<sup>26</sup>The model's assumptions are meant for a relatively stable period like postwar America, in which turnout forces vary only modestly from one presidential election to the next. When turnout surges because a realignment is underway during the Depression, for instance,  $k$  would not be fixed, certainly not across elections and perhaps not across subgroups. It would be useful to explore whether the higher turnouts of the Fifties and Sixties have detectable effects on the current registration rates of older citizens.

## The Beta-Binomial Model Applied to South Dakota

South Dakota has one of the highest voter turnout rates among the American states, and its older citizens are registered at even higher rates. Thus it is no surprise that several rounds of estimates with the model of Equation (25) gave consistent estimates for  $k$  of less than 2%, often less than 1%. To simplify the model, and to ease comparison with probit analysis,  $k$  was constrained to zero in the estimations that follow.

First we fit a simple age-and-education specification to South Dakota citizens' registration decisions. To maintain consistency with the model of the previous section, age is coded as the number of presidential elections the citizen has experienced (including the current one), which is widely thought to be what age proxies for.<sup>27</sup> With the exception of the small hiccup occurring because the voting age was changed from 21 to 17 in 1971, the number of prior presidential elections is very nearly a linear function of age, so that the two codings are essentially equivalent for assessing fit.

To avoid all the issues raised by residential mobility, we begin by comparing probit and beta-binomial fits among those South Dakotans who have been at their current residence five years or longer. Table 1 lists the resulting coefficients from a linear probit model vs. those from Equation (25).<sup>28</sup>

---

<sup>27</sup>One can see the step function in young people's turnout in the New Jersey voter file for 2004. The number of presidential elections one has experienced matters more than age *per se*.

<sup>28</sup>Estimates were computed in STATA using the *lf* option in the *ml* command. Thus standard errors are computed numerically, exploiting the linear-in-parameters feature of Equation (25).

Table 1. Voter Registration in South Dakota  
 Citizens at current residence 5 years or more  
 (CPS 2000 data. Standard errors in parentheses. N = 467.)

	<u>probit</u>	<u>beta-binomial</u>
prior pres. elections (“age”)	.1242 (.03269)	
education	.4503 (.07502)	.4727 (.08996)
constant	-1.2320 (.3772)	-2.1905 (.3536)
$\alpha+\beta$		7.9728 (7.7956)
log-likelihood	-139.4	-137.8

Both models produce sensible coefficients, but the beta-binomial fit is a bit better. All the coefficients in each model are statistically significant except that for  $c = \alpha + \beta$  in the beta-binomial model. This parameter is an estimate of the variability in  $\pi_i$ . The point estimate indicates a moderate amount of heterogeneity in one-time registration probabilities within South Dakota, as Figure 11 illustrates for the case  $c = 8$  and  $\alpha/(\alpha + \beta) = .5$ . That is, the figure shows the estimated range of variation in  $\pi_i$  in South Dakota for a group of residentially stable citizens whose age and education predict a one-time registration probability of 50%. However, the estimate for  $c$  is quite noisy, and thus the figure is no more than suggestive. What it suggests, though, is that unmeasured heterogeneity may be considerable.

**Density Function for Beta(4,4)**

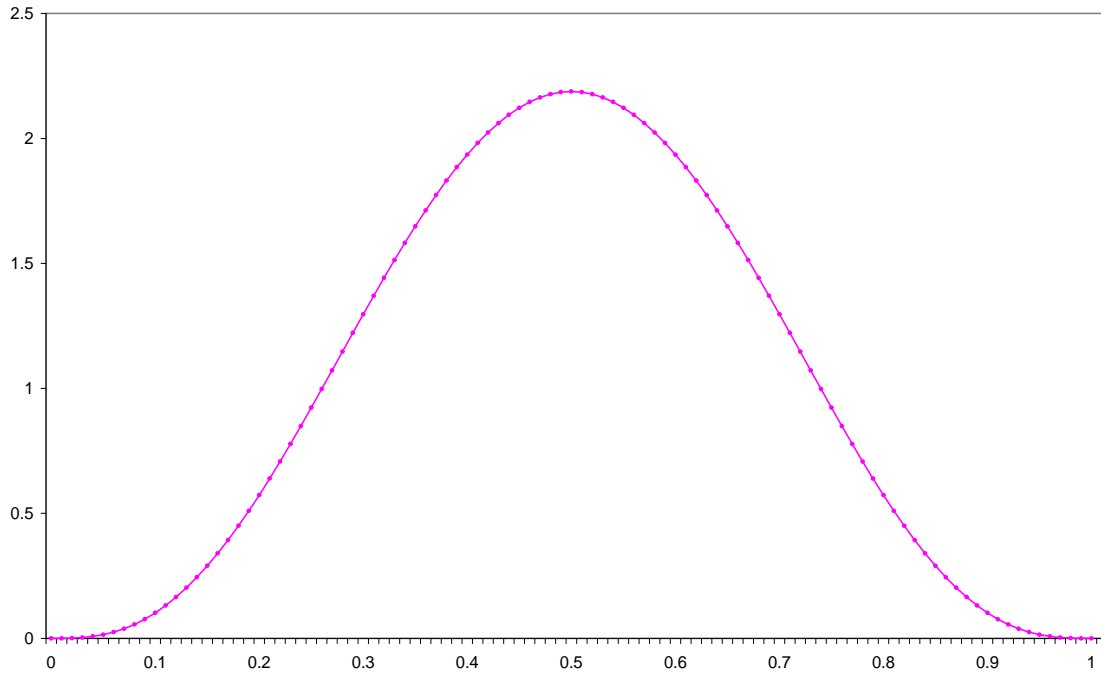


Figure 11.

Finally, we construct estimates for the full South Dakota sample, taking mobility into account. As a first approach in the spirit of the beta-binomial model, we suppose that residential mobility resets the citizen's count of the number of registration opportunities. That is, a middle-aged citizen approaching her first presidential election at a new address is in the same position as a young voter: Both are encountering their initial opportunity to register. The CPS data do not allow us to learn precisely how many presidential elections the citizen has experienced at the current address. As the most straightforward guess, we assume that everyone at their current address for two years or fewer is at their first election, those at their current address three to four years are experiencing their second, and those at their current address five years or more are experiencing the number that have occurred since they turned 18 (or 21 for older citizens). The CPS codings allow no finer distinctions

or better guesses.<sup>29</sup>

With this definition, Table 2 gives the estimation results for probit and the beta-binomial model in South Dakota.

**Table 2. Voter Registration in South Dakota**  
(CPS 2000 data. Standard errors in parentheses. N = 800.)

	<u>probit</u>	<u>beta-binomial</u>
prior pres. elections (adj. for mobility)	.1387 (.01674)	
education	.3987 (.05133)	.5915 (.07533)
constant	-1.1855 (.2186)	-2.1905 (.3536)
$\alpha+\beta$		1.7134 (.2926)
log-likelihood	-326.1	-324.5

Now all parameters are well estimated, and again the beta-binomial model is performing a bit better. Here the parameter  $c$  is much better determined and its point estimate much smaller than in the previous table. The amount of heterogeneity implied is quite large. In particular, a  $Beta(1, 1)$  density is uniform.

Thus the beta-binomial model shows promise as a model of registration. It is easily computable and requires no simulation. In a first test in one American state, it fit somewhat better than the usual garbage-can probits that dominate the literature. Of course, more needs to be done with data sets from different sources and places. It would be particularly useful to study actual voter files, since they avoid the problem of over-reporting that plagues surveys (though the CPS is less afflicted by this problem than most). Unfortunately, the

---

<sup>29</sup>Both probit and the beta-binomial model do better with this recoded prior presidential elections variable than with its original version based solely on age. In addition, alternate recodings of that variable did slightly less well than the obvious first idea set out in the text.

voter files convey no information about time at residence. Thus no test is perfect, and triangulation will be necessary.

## Conclusion

This paper began by postulating a rational–expectations model for registration: People register because they expect to want to vote in the future. Hence at this level of abstraction, exactly the same variables that predict turnout will predict registration. It follows that when the actual covariance between the disturbances of the registration and turnout equations is zero, then one is better off ignoring the selection problem and just running probit equations for the entire sample, with turnout as the dependent variable.

In this same situation, seemingly appropriate Heckman–style selection models will tend to generate incorrect zero coefficients in the vote equation, and highly negative correlation estimates of the correlation between the disturbances. This paper showed that prior attempts along these lines by Timpone (1998) illustrated precisely these problems. And an empirical comparison of North and South Dakota here produced again the same pattern of erroneous estimates.

A new selection equation for registration was derived and applied to South Dakota. It fit better than garbage–can probits with the same number of coefficients. The next step will be to test the model on other datasets, and then combine it with the vote equations set out in Achen & Sinnott (2009) to get a full selection and outcome model. A copula will be used to link the two equations (Nelson 2000).

Of course, the registration model proposed here is just one way to model the American registration process. Other possibilities exist. Neither they nor the present effort have undergone stress testing with a variety of datasets. In the short run, then, empirical researchers are wise to continue doing what most of them have done, which is to ignore registration and just model turnout directly, with non–registered citizens counted as abstainers. If everyone under the age of 35 and everyone who has moved in the past four years is excluded from the sample, thus giving everyone who wishes to register at least two presidential elections in which to do so, sensible coefficients should result and biases will be minimized. At the present stage of our knowledge, no better statistical advice can be given for working with observational data. Lab and field experiments present their own difficulties of validity and inference as well. Much remains to be done before reliable quantitative evidence can be brought to bear on the turnout problem.

Finally, younger voters are the source of most American abstention. They present all sorts of difficulties in addition to registration failures—frequent moves, attendance at schools distant from their homes, military service, and others. They need to be studied

separately. Carrying out a substantial national turnout study directed specifically at young people is a key priority if low American turnout rates are to be understood and improved. That analysis will present a dramatic new range of econometric challenges.

## Appendix: Proof of Equation (11)

Standardizing all variables to mean zero (thus eliminating the constant term), write Equation (8) in standard matrix form as:

$$y_2^{**} = X_1\beta_{11} + X_2\beta_{12} + u_2 \quad (26)$$

where  $y_2^{**} = y_2^* - E(y_2^*)$ , with  $y_2^*$  and  $y_2^{**}$  column vectors of observations on  $y_{2i}^*$  and  $y_{2i}^{**}$  respectively,  $X_1$  and  $X_2$  are matrices of observations on  $x_{1i} - E(x_{1i})$  and  $x_{2i} - E(x_{2i})$ ,  $u_2$  is the vector of disturbances  $u_{2i}$ , and  $\beta_{11}$  and  $\beta_{12}$  are coefficient vectors as before.

Now turn to Equation (10). First let  $q_i^* = 1 - \beta_{01} - x_{1i}\beta_{11} - x_{2i}\beta_{12}$ . Let  $q^*$  be the vector with elements  $q_i^*$ , and define  $q = q^* - E(q^*)$ . It follows that  $q = -X_1\beta_{11} - X_2\beta_{12}$ , where all variables are standardized to mean zero. Note that  $q = -y_2^{**}$ . Then Equation (10), the linear approximation to the researcher's erroneously estimated equation, may be written in mean-deviated form as:

$$\begin{aligned} y_2^{**} &\approx X_1\hat{\beta}_{11} + \hat{\kappa}q + \hat{u}_2 \\ &= X_1\hat{\beta}_{11} - \hat{\kappa}y_2^{**} + \hat{u}_2 \end{aligned} \quad (27)$$

Applying OLS to the latter equation gives, by the usual formula:

$$\begin{aligned} \begin{bmatrix} \hat{\beta}_{11} \\ \hat{\kappa} \end{bmatrix} &= \begin{bmatrix} X_1'X_1 & -X_1'y_2^{**} \\ -y_2^{**'}X_1 & y_2^{**'}y_2^{**} \end{bmatrix}^{-1} \begin{bmatrix} X_1'y_2^{**} \\ -y_2^{**'}y_2^{**} \end{bmatrix} \\ &= \frac{1}{s^2} \begin{bmatrix} s^2(X_1'X_1)^{-1} + bb' & b \\ b' & 1 \end{bmatrix} \begin{bmatrix} X_1'y_2^{**} \\ -y_2^{**'}y_2^{**} \end{bmatrix} \end{aligned} \quad (28)$$

with  $b = (X_1'X_1)^{-1}X_1'y_2^{**}$  and  $s^2 = y_2^{**'}y_2^{**} - y_2^{**'}X_1(X_1'X_1)^{-1}X_1'y_2^{**}$ . Multiplying out in Equation (28) then gives:

$$\begin{aligned}
\hat{\beta}_{11} &= \frac{1}{s^2} [s^2(X_1'X_1)^{-1}X_1'y_2^{**} + by_2^{*'}X_1(X_1'X_1)^{-1}X_1'y_2^{**} - by_2^{*'}y_2^{**}] \\
&= \frac{bs^2 + b(-s^2)}{s^2} \\
&= 0
\end{aligned} \tag{29}$$

Similarly, from Equation (28):

$$\begin{aligned}
\hat{\kappa} &= \frac{y_2^{*'}X_1(X_1'X_1)^{-1}X_1'y_2^{**} - y_2^{*'}y_2^{**}}{s^2} \\
&= -1
\end{aligned} \tag{30}$$

## References

- [1] Achen, Christopher H. 1986. *The Statistical Analysis of Quasi-Experiments*. Berkeley: University of California Press.
- [2] Achen, Christopher H., and Richard Sinnott, eds. 2009. *Voter Turnout in Multi-Level Systems*. Book manuscript in preparation.
- [3] Arneson, Ben A. 1925. Non-Voting in a Typical Ohio Community. *American Political Science Review* 19,4 (Nov.): 816-825.
- [4] Bentley, Arthur F. 1967 [1908]. *The Process of Government*. Cambridge, Massachusetts: Harvard University Press.
- [5] Blais, André. 2000. *To Vote or Not to Vote*. Pittsburgh: University of Pittsburgh Press.
- [6] Campbell, Angus, Philip E. Converse, Warren E. Miller, and Donald E. Stokes. 1960. *The American Voter*. New York: Wiley.
- [7] Department of the Public Advocate. 2008. New Jersey to Press Forward with Motor Voter Implementation. Press release. Trenton: State of New Jersey, March 24, 2008. [http://www.state.nj.us/publicadvocate/news/2008/approved/080324\\_nvra\\_mou.html](http://www.state.nj.us/publicadvocate/news/2008/approved/080324_nvra_mou.html)
- [8] Dubin, Jeffrey A., and Douglas Rivers. 1989/1990. Selection Bias in Linear Regression, Logit and Probit Models. *Sociological Methods and Research* 18: 360-390.
- [9] Griffin, John D., and Brian Newman. 2005. Are Voters Better Represented? *Journal of Politics* 67,4 (Nov.): 1206-1227.
- [10] Gosnell, Harold. 1927. *Getting Out the Vote*. Chicago: University of Chicago Press.
- [11] Granato, Jim, and Frank Scioli. 2004. Puzzles, Proverbs, and Omega Matrices: The Scientific and Social Significance of Empirical Implications of Theoretical Models (EITM). *Perspectives on Politics* 2, 2: 313-323.
- [12] Hanmer, Michael J. 2004. From Selection to Election and Beyond: Understanding the Causes and Consequences of Electoral Reform in America. Doctoral dissertation, Political Science, University of Michigan.
- [13] Heckman, James. 1979. Sample Selection Bias as a Specification Error. *Econometrica* 47: 931-959.
- [14] Highton, Benjamin. 2004. Voter Registration and Turnout in the United States. *Perspectives on Politics* 2 (Sept.): 507-515.

- [15] Jackson, Robert A. 1996. A Reassessment of Voter Mobilization. *Political Research Quarterly* 49: 331-349.
- [16] Kelley, Stanley, Richard E. Ayres, and William G. Bowen. 1967. Registration and Voting: Putting First Things First. *American Political Science Review* 61,2 (Jun.): 359-379.
- [17] Lijphart, Arend. 1997. Unequal Participation: Democracy's Unresolved Dilemma. *American Political Science Review* 91,1 (Mar.): 1-14.
- [18] McDonald, Michael P., and Samuel L. Popkin. 2001. The Myth of the Vanishing Voter. *American Political Science Review* 95: 963-974.
- [19] Massy, William F., David B. Montgomery, and Donald G. Morrison. 1970. *Stochastic Models of Buying Behavior*. Cambridge, Massachusetts: MIT Press.
- [20] Merriam, Charles E. and Harold F. Gosnell. 1924. *Non-Voting: Causes and Methods of Control*. Chicago: University of Chicago Press.
- [21] Nagler, Jonathan. 1994. Scobit: An Alternative Estimator to Logit and Probit. *American Journal of Political Science* 38,1 (Feb.): 230-255.
- [22] Nelson, Roger B. 2007. *An Introduction to Copulas*. 2nd ed. New York: Springer.
- [23] Sinnott, Richard. The Irish Case. In Achen and Sinnott 2009.
- [24] Squire, Peverill, Raymond E. Wolfinger, and David P. Glass. 1987. Residential Mobility and Voter Turnout. *American Political Science Review* 81 (Mar.): 45-66.
- [25] Timpone, Richard J. 1998. Structure, Behavior, and Voter Turnout in the United States. *American Political Science Review* 92: 145-158.
- [26] Tingsten, Herbert. 1937. *Political Behavior*. London: P. S. King.
- [27] Uhlaner, Carole Jean. 1989. Turnout in Recent Presidential Elections. *Political Behavior* 11: 57-79.
- [28] Verba, Sidney, Kay Lehman Schlozman, and Henry E. Brady. 1995. *Voice and Equality*. Cambridge, Massachusetts: Harvard University Press.
- [29] Wiley, J.A., S.J. Herschkorn, and N.S. Padian. 1989. Heterogeneity in the Probability of HIV Transmission per Sexual Contact: The Case of Male-to-Female Transmission in Penile-Vaginal Intercourse. *Statistics in Medicine* 8,1 (Jan.): 93-102.

- [30] Wolfinger, Raymond E. and Steven J. Rosenstone. 1980. *Who Votes?* New Haven: Yale University Press.