Bayesian versus maximum likelihood estimation of treatment effects in bivariate probit instrumental variable models [*]

Florian M. Hollenbach[†]
Department of Political Science
Texas A&M University

Jacob M. Montgomery
Department of Political Science
Washington University in St. Louis

Adriana Crespo-Tenorio
Lead Researcher
Facebook

January 24, 2017

Abstract

Bivariate probit models are a common choice for scholars wishing to estimate causal effects in instrumental variable models where both the treatment and outcome are binary. However, standard maximum likelihood approaches for estimating bivariate probit models are problematic. Numerical routines in popular software suites frequently generate inaccurate parameter estimates, and even estimated correctly, maximum likelihood routines provide no straightforward way to produce estimates of uncertainty for causal quantities of interest. In this note, we show that adopting a Bayesian approach provides more accurate estimates of key parameters and facilitates the direct calculation of causal quantities along with their attendant measures of uncertainty.

[*] We are grateful for helpful comments from Kosuke Imai, Kevin Quinn, Marc Ratkovic, Justin Esarey, and a helpful audience at Washington University in St. Louis. Previous versions of this paper were presented at the 2013 Annual Meeting of the Midwest Political Science Association in Chicago, the 2013 Summer Meeting of the Society of Political Methodology at the University of Virginia, and the 2016 Annual Meeting of the Southern Political Science Association in San Juan, Puerto Rico.

[†] Corresponding Author

1. INTRODUCTION

In many areas of social science inquiry, hypotheses cannot be tested with randomized treatments and laboratory-levels of control. Scholars must instead draw inferences based on either observational data or experiments conducted with imperfect levels of compliance. As a consequence, a great deal of attention has gone into developing methods to generate unbiased estimates of causal effects in the presence of unmeasured confounders that influence both outcomes and the probability of treatment. In particular, a frequent choice is applying instrumental variables analysis to generate unbiased estimates of causal effects.

In this note, we show that Bayesian estimation methods for bivariate probit models are superior to traditional maximum likelihood (ML) routines. Specifically, Bayesian methods provide more accurate estimates of key model parameters in realistic settings. Further, the Bayesian approach facilitates the direct calculation of important causal quantities of interest along with measures of uncertainty. After presenting the bivariate probit instrumental variables model, we report results from two studies contrasting the performance of Bayesian and ML estimation methods. First, using simulated data, we show that Bayesian bivariate probit models significantly outperform the ML routine in terms of bias and coverage, especially in the presence of weak instruments or small sample sizes. Second, we compare ML and Bayesian estimates using the Sondheimer and Green (2010) data set, which estimates the effect of educational attainment on voter turnout, and show that the differences in estimation methods can affect substantive conclusions in a real-world application.

2. BIVARIATE PROBIT MODELS FOR ESTIMATING CAUSAL EFFECTS

While instrumental variable (IV) models are increasingly common in political science (c.f., Sovey and Green 2010, 194), some researchers find the standard approach—two stage least squares—too inflexible and difficult to adjust to handle common data-types. In particular, social scientists frequently encounter instances where both the endogenous regressor and the outcome variable are dichotomous. In these cases, some authors have suggested a bivariate probit model. For instance, Angrist and Pischke (2009) argue that, ''Bivariate probit probably qualifies as harmless in the sense that it's not very complicated and easy to get right using packaged software routines'' (p. 201).

Indeed, bivariate probit models have appeared regularly in prominent political science work in the past (e.g., Ansolabehere, Iyengar and Simon 1999; Gerber and Green 2000; Arceneaux and Nickerson 2009), and continue to be used regularly even in the discipline's most prestigious journals (Christin and Hug 2012; Maves and Braithwaite 2013; Pierskalla and Hollenbach 2013; Fuhrmann and Sechser 2014; Cederman, Hug and Wucherpfennig 2015). Most often this is justified by the desire to appropriately handle an endogenous regressor when both the instrument and the outcome are binary. Wucherpfennig, Hunziker and Cederman (Forthcoming), for instance, state that, ''Because the dependent variable (conflict) and endogenous regressor (inclusion) are both binary, a nonlinear model is appropriate. Thus, we rely on a seemingly unrelated bivariate probit estimator. This is a framework suitable for two processes with dichotomous outcomes for which the error terms are correlated'' (pg. 11).

In theory, bivariate probit models represent a modest adjustment to standard IV models. Yet, correctly implementing the model is surprisingly challenging. First, some research suggests that the maximum-likelihood (ML) bivariate probit algorithms in common software suites can be inaccurate (e.g, Freedman and Sekhon 2010). Second, due to the nonlinearity in the assumed data generating process, estimates for treatment effects are difficult to calculate directly from standard model outputs. Quantities of interest such as the average treatment effect (ATE) require further calculations. More critically, uncertainty estimates for these quantities are not directly available.

2.1. *Approach*

We consider the problem of identifying a causal effect in the context of the potential outcomes framework introduced by Rubin (1974). Let $y \in \{0, 1\}$ be a binary outcome, $T \in \{0, 1\}$ be a binary treatment variable, and $\mathbf{x_i}$ be a vector of covariates (including a constant) for observation $i \in [1, 2, \ldots, n]$. We observe an outcome y_i which is a realization of one of two potential outcomes (y_{i0}, y_{i1}) corresponding to the two possible levels of treatment, (T_{i0}, T_{i1}). Thus, y_{i0} and y_{i1} are random variables such that

$$\begin{aligned} y_{i0} &= 1(\mathbf{x}_i \boldsymbol{\gamma}_2 + \eta_i > 0) \\ y_{i1} &= 1(\beta + \mathbf{x}_i \boldsymbol{\gamma}_2 + \eta_i > 0), \end{aligned} \tag{1}$$

where η_i is a symmetric unobserved error term centered at zero, and $1(\cdot)$ is the standard indicator function. The observed outcome, y_i, is given by $y_i = (1 - T_i)y_{i0} + T_i y_{i1}$ and the average treatment effect (ATE) is $\mathbb{E}(y_1) - \mathbb{E}(y_0) \equiv \Delta$. Assuming that $f(\cdot|\boldsymbol{\mu})$ is the probability density function (PDF) for the error term η, $F(\cdot|\boldsymbol{\mu})$ is the cumulative density function (CDF), and $\boldsymbol{\mu}$ represents any parameters defining $f(\cdot)$, then $\Delta = F(\beta + \mathbf{x}\boldsymbol{\gamma}_2|\boldsymbol{\mu}) - F(\mathbf{x}\boldsymbol{\gamma}_2|\boldsymbol{\mu})$.

Assuming that the unobserved errors are conditionally independent of the treatment, $\eta_i \perp T_i | \mathbf{x}_i$, the ATE can be estimated using standard regression techniques (Rosenbaum and Rubin 1983). However, in many settings, this assumption does not hold; some factors may affect both the treatment assignment and the outcome. In this scenario, estimates from standard linear models are known to be inconsistent.

To generate consistent estimates of the true causal effect, we introduce a variable,[1] $\mathbf{z}$, that is related to T but is independent of y conditioned on the covariates $\mathbf{x}$. With some additional assumptions, it is then possible to apply methods from structural equation modeling to generate consistent and efficient estimates of causal effects (Chib 2003). The basic approach is to solve a system of simultaneous equations,

$$\begin{aligned} T_i &= 1(\mathbf{x}_i' \boldsymbol{\gamma}_1 + z_i \pi + \eta_{1i} > 0) \\ y_i &= 1(\mathbf{x}_i' \boldsymbol{\gamma}_2 + T_i \beta + \eta_{2i} > 0), \end{aligned} \tag{2}$$

[1] For the sake of clarity, we restrict ourselves to the case where there is only one instrument and one treatment of interest.

where $(\eta_{1i}, \eta_{2i})' \sim f^{(2)}(\boldsymbol{\eta}|\boldsymbol{\mu})$ is some bivariate PDF defined by parameters $\boldsymbol{\mu}$.

2.2. *Model specification and likelihood*

Correctly estimating Model (2) will, given additional assumptions, allow us to correctly estimate the various model parameters and thus the causal quantities of interest (Angrist, Imbens and Rubin 1996; Chib 2003). We make four assumptions. Assumption 1 states that the instrument has a conditional effect on the endogenous treatment, or more formally $\text{COV}(z, T|\mathbf{X}) \neq 0$. Assumption 2, commonly referred to as the exclusion restriction, states that knowledge about the data-generating process of the instrument will not effect the outcome except through the treatment. Assumption 3, the stable unit treatment value assumption (SUTVA), states that the potential outcomes for each person are unrelated to the treatment received by other observations. Finally, Assumption 4 states that the unmeasured errors are distributed according to a bivariate normal distribution symmetric around the vector $\mathbf{0}$ with covariance matrix $\boldsymbol{\Omega}$.[2]

Letting $\phi^{(2)}(\cdot)$ be a bivariate normal PDF and $\Phi^{(2)}(\cdot)$ be the CDF, we can formalize this model by letting $(\eta_{1i}, \eta_{2i})' = (\epsilon_{1i}, \epsilon_{2i})' \overset{iid}{\sim} \phi^{(2)}(\mathbf{0}, \boldsymbol{\Omega})$ in Equation (2). To identify the model, we set $\omega_{11} = \omega_{22} = 1$ and $\omega_{12} = \omega_{21} = \rho \in [0, 1]$, which means that $\boldsymbol{\Omega}$ is a two-dimensional correlation matrix.

Applying this to Equation (1), we obtain the marginal probabilities

$$\begin{aligned} \text{Pr}_0(y_{i0} = 1|\mathbf{x}_i, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \beta, \boldsymbol{\Omega}) &= \Phi(\mathbf{x}_i'\boldsymbol{\gamma}_2, 1) \\ \text{Pr}_1(y_{i1} = 1|\mathbf{x}_i, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \beta, \boldsymbol{\Omega}) &= \Phi(\mathbf{x}_i'\boldsymbol{\gamma}_2 + \beta, 1). \end{aligned} \tag{3}$$

[2] Assumption 4 is stronger than the standard monotonicity assumptions required for two stage least squares models (2SLS) (Angrist, Imbens and Rubin 1996). However, our purpose here is not to compare the performance of Bayesian bivariate probit models with 2SLS models, but rather to contrast it with the performance and flexibility of ML bivariate probit models, which also require this assumption. We do, however, provide an evaluation of 2SLS on the simulated data in section E of the online Appendix.

Likewise, we can calculate the marginal probabilities for T, $\Pr(T_i = 1|\mathbf{x}_i, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \beta, \boldsymbol{\Omega}) = \Phi(\mathbf{x}_i\prime\boldsymbol{\gamma}_1 + z_i\pi)$. Finally, the likelihood can then be expressed as:

$$\begin{aligned} L(y, T|\mathbf{x}, z, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \beta, \boldsymbol{\Omega}) = {} & \prod_{i\in\mathcal{F}} \Phi^{(2)}\left(\begin{pmatrix} -(x_i\boldsymbol{\gamma}_1 + z_i\pi) \\ (2y_i - 1)(x_i\boldsymbol{\gamma}_2) \end{pmatrix}, \boldsymbol{\Omega}\right) \\ & \times \prod_{i\in\mathcal{G}} \Phi^{(2)}\left(\begin{pmatrix} x_i\boldsymbol{\gamma}_1 + z_i\pi \\ (2y_i - 1)(x_i\boldsymbol{\gamma}_2 + \beta) \end{pmatrix}, \boldsymbol{\Omega}\right), \end{aligned} \tag{4}$$

where $\mathcal{F}$ is the subset where we observe $T_i = 0$, and $\mathcal{G}$ represents the subset where $T_i = 1$.

2.3. *Bayesian estimation*

The model in Equation (4) can, in principle, be estimated using one of several possible ML routines. However, previous work by Freedman and Sekhon (2010) suggests that these ML routines should be used with caution:

> In multidimensional problems, even the best numerical analysis routines find spurious maxima for the likelihood function. Our models present three kinds of problems: (1) flat spots on the log likelihood surface, (2) ill-conditioned maxima, where the eigenvalues of the Hessian are radically different in size, and (3) ill-conditioned saddle points with one small positive eigenvalue and several large negative eigenvalues. (Freedman and Sekhon 2010, p 144)

Thus, we turn to a Bayesian approach for estimating this model, which sometimes offers superior performance for higher-dimensional models with multiple maxima (e.g., Jackman 2000) and which can incorporate prior structure to improve numerical performance where the likelihood in isolation is difficult to characterize.

The Bayesian model is implemented in the `Stan` modeling language using no-U-turn sampling (Hoffman and Gelman 2014). This implementation proved superior in terms of sampling time and convergence speed to alternatives. Full replication code is provided in the online Appendix, which also provides additional discussion of the model.

2.4. *Estimating quantities of interest*

To calculate the actual treatment effects of interest, we first estimate the interim quantity $\Delta_i^{(ATE)} = \Phi(\mathbf{x}_i'\boldsymbol{\gamma}_2 + \beta) - \Phi(\mathbf{x}_i'\boldsymbol{\gamma}_2)$. We evaluate this quantity at the observed value of the covariates $\boldsymbol{x}_i$. We then calculate $\Delta^{(ATE)} = \sum_{i=1}^{n} \Delta_i^{(ATE)}/n$. Crucially, calculating $\Delta^{(ATE)}$ for each draw from the posterior provides a posterior distribution for the ATE.

Another quantity of interest is the average effect of the treatment on the treated (ATT), defined as $\Delta^{(ATT)} = E[y_1|T = 1] - E[y_0|T = 1]$. We can again calculate interim value

$$\Delta_i^{(ATT)} = \frac{\Phi^{(2)}\left(\begin{pmatrix} x_i\gamma_1 + z_i\pi \\ x_i\gamma_2 + \beta \end{pmatrix}, \boldsymbol{\Omega}\right) - \Phi^{(2)}\left(\begin{pmatrix} x_i\boldsymbol{\gamma}_1 + z_i\pi \\ x_i\gamma_2 \end{pmatrix}, \boldsymbol{\Omega}\right)}{\Phi(\mathbf{x}_i'\boldsymbol{\gamma}_1 + z_i\pi)}$$

evaluated for the observed values of $\mathbf{x}_i$ and z_i. The ATT is then $\sum_{i=1}^{n} \Delta_i^{(ATT)}/n$.

In a supplementary online Appendix we show detailed calculations for the local average treatment effect (LATE). This is the effect of the treatment on the subpopulation that complies with the instrument.

3. SIMULATION STUDY

In this section, we compare the performance of the Bayesian and ML variants of the bivariate probit models using simulated data. Specifically, we compare the results from the `Stan` model shown in the online Appendix with the `biprobit` command in `Stata` version 14.[3]

For this study, x is a matrix of an intercept and one independent variable. Both the independent variable in x and the instrument are drawn from standard normal distributions. Using this data, we generate the treatment and outcome exactly as assumed

[3] We specify Gaussian priors with mean zero and standard deviation 2.5 on the structural parameters of the model. In the online Appendix we consider alternative prior specifications, including a noninformative uniform prior.

Table 1. Parameter Values for Data Simulation

Variable Name	Values considered				
N	50	250	500	1000	
ρ	0.25	0.5	0.75		
π	0.5	1	1.5	2.0	
β	0	1.0	2.0	3.0	4.0

Table 2. Example Simulation

	First Stage			Second Stage		
	Intercept	*X*	*Instrument*	*Intercept*	*X*	*Treatment*
True Values	1.5	-1	0.5	-2	2.5	0
ML Estimate	2.71	-1.47	0.85	-0.64	2.01	-1.82
ML 95% CI	(0.78, 4.64)	(-2.68, -0.25)	(0.13, 1.57)	(-2.11, 0.83)	(0.82, 3.19)	(-3.01, -0.63)
Bayesian Estimate (median)	2.50	-1.39	0.73	-2.06	2.70	-0.69
Bayesian 95% CI	(1.38,4.33)	(-2.50,-0.58)	(0.14,1.41)	(-4.20,-0.26)	(1.45,4.62)	(-2.24,0.98)

in Equation (2). We fix the γ parameters such that $\gamma_1 = (1.5, -1)$ and $\gamma_2 = (-2, 2.5)$. However, we vary the correlation between the first and second stage (ρ), the number of observations (N), the instrument strength (π), and the treatment coefficient (β). Table 1 displays the parameter values considered. For each of the 240 parameter combinations, we simulate 500 independent data sets.

We begin with a single example. Table 2 displays the results for both the `Stata` ML estimation and the Bayesian estimation for one data set (N=50, $\rho = 0.25$). The top row displays the true parameter values for the simulated data, while the remaining rows show the ML and Bayesian estimates along with their their 95% confidence intervals and credible intervals respectively.[4] The standard ML estimates in Table 2 would lead a researcher to conclude that the treatment of interest has a substantively strong and statistically significant negative effect on the outcome. Instead, the actual treatment effect is exactly zero, which is correctly included in the 95% CI for the Bayesian estimation. This

[4] In all our simulations, we use the median posterior parameter estimates.

pattern, where the ML method provides inaccurate point estimates as well as inadequate coverage relative to the Bayesian method, can be generalized by considering a wider set of simulated data sets.[5]

To compare the methods, we calculate the mean squared error (MSE) as well as 95% coverage probabilities for the average treatment effect.[6] Each frame of Figure 1 displays the MSE on the average treatment effect (ATE) for the ML and Bayesian methods partitioned by sample size.[7] The gray vertical bars are the *difference* between the MSE for each approach. The figure reveals two important patterns. First, the MSE of the Bayesian estimator is *always* smaller than that of the ML estimator, even with modestly large sample sizes and strong instruments. Second, both estimators perform better with stronger instruments (from left to right along the x-axis) and larger sample sizes (plots going from the top left to bottom right). However, when the instrument is weak or sample sizes are smaller, the

[5] An alternative approach would be to calculate standard errors using the non-parametric bootstrapping method as implemented in `Stata`. However, in our experience this often leads to absurdly large confidence intervals while doing nothing to improve the accuracy of the point estimates. For the example data set analyzed in Table 2, for instance, the 95% bootstrapped confidence interval for the second-stage coefficients were [-2516.4, 2515.1], [-1712.2, 1716.2], and [-1390.2, 1386.6], for the intercept, the covariate (x), and the treatment respectively.

[6] The 95% confidence intervals for the ATE and LATE estimates for the Stata ML model are created via the parametric bootstrap. We are grateful to an anonymous reviewer for suggesting this approach. Details are presented in Section I of the Online Appendix. Note that these general results hold when analyzing only the β parameter from Equation (3) using the usual asymptotic standard errors (results not shown).

[7] In section E of the Appendix we present the results for the local average treatment effect (LATE).

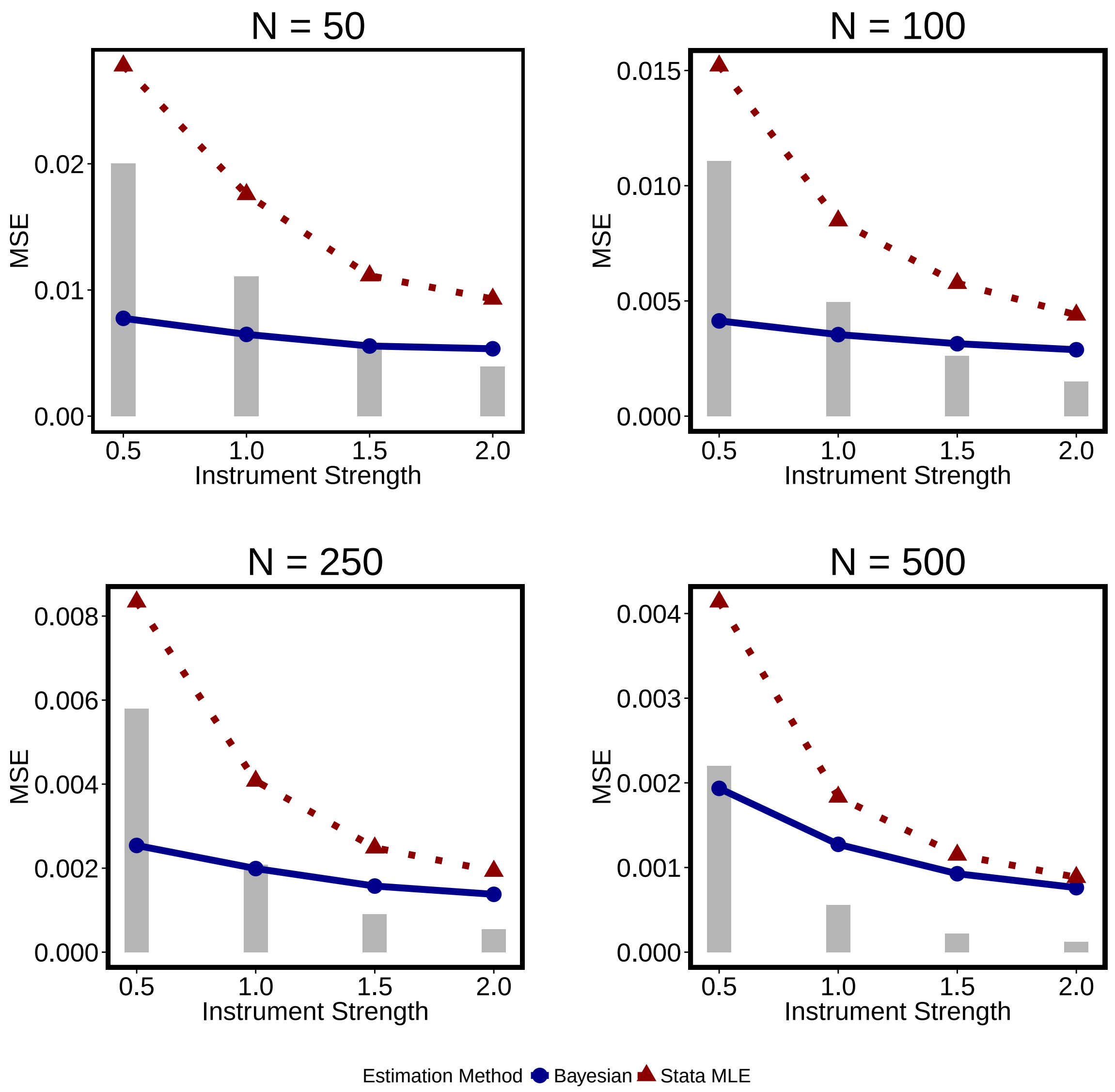

Figure 1. This Figure compares the mean squared error (MSE) on the average treatment effect (ATE) for the Bayesian and ML estimates on the simulated data sets. Gray vertical rectangles show the difference between the two MSEs at each parameter value. The Bayesian estimator outperforms the ML estimate in terms of the MSE across *all* parameter settings, and this pattern is particularly clear with small samples and weak instruments.

relative advantage of the Bayesian estimator is greater.[8]

Figure 2 shows the coverage probabilities for the 95% confidence or credible intervals on the ATE. Specifically, it shows the share of data sets at each parameter combination for which the 95% confidence or credible intervals include the true parameter value.

[8] In the online Appendix, we show that this pattern holds using mean absolute error.

Ideally, we would expect this to happen 95% of the time. Figure 2 again shows the Bayesian approach to be preferable, and the differences between the methods are especially pronounced for weak instruments and small samples. Bayesian CIs are somewhat conservative, with the maximum and minimum coverage probabilities being 0.98 and 0.95 respectively. More broadly, the coverage rates fall between 94% and 97% for 11 of the 16 different parameter combinations shown in the figure. In contrast, the ML coverage probabilities are too small. The smallest coverage rate is 0.67 ($\pi = 0.5$, $N = 100$) while the maximum is 0.94 ($\pi = 2$, $N = 500$). Indeed, the coverage rate for the ML exceeds 90% for only 6 of the 16 parameter combinations in Figure 2.

4. THE EFFECT OF EDUCATION ON VOTER TURNOUT

In this section, we compare the Bayesian and ML bivariate probit models to estimate the causal effect of education on turnout. To address this question, Sondheimer and Green (2010) (henceforward SG) leverage three randomized experiments aimed at stimulating educational attainment (e.g., Heckman et al. 2010; Barnett 2010; Kahne and Bailey 1999; Hanushek 1999): the High/Scope Perry Preschool Experiment (Perry), the I Have A Dream program (IHAD), and the Student-Teacher Achievement Ratio (STAR) experiment. The basic approach is to consider these experiments as encouragement studies, where educational attainment is the actual causal variable of interest and the experimental assignments are instruments. SG then match data from participants to turnout in federal elections between 2000 and 2004, where the outcome variable is simply an indicator for voting in any election in this period. The goal is to estimate the effect of educational attainment on political participation, which may be confounded by socioeconomic status (Tenn 2007; Berinsky and Lenz 2011).

There are two features of this data, which is displayed in Table 3, that are relevant for our purposes. First, the sample sizes are small for the Perry (N=123) and IHAD (N=58) experiments. Second, high school graduation rates in the STAR experiment were very high in absolute terms across treatment conditions (90.08% in the treatment group and

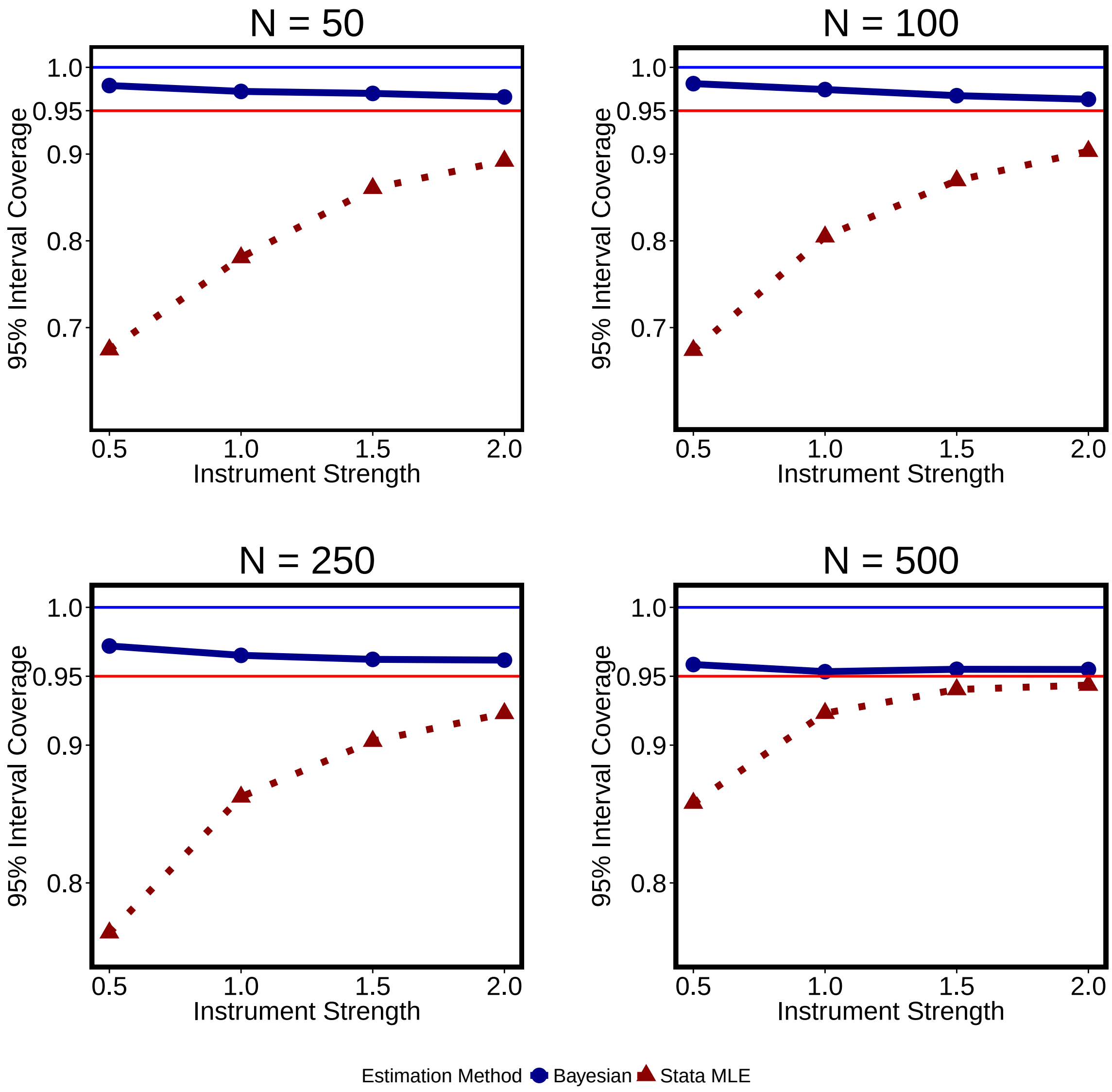

Figure 2. This Figure compares the 95% coverage probabilities on the ATE for both the Bayesian and ML estimators. The Bayesian estimator is much closer to this value for nearly all parameter combinations, although it is somewhat conservative. The coverage probabilities for the ML model are too small, especially with weak instruments or small sample sizes.

84.97% in the control group), which implies a relatively weak instrument.

The ML bivariate probit results from SG are shown in the columns labeled ML in Table 4. The columns labeled "Bayesian" show these estimates using the Bayesian approach. Note that the coefficients estimated in the Bayesian models differ considerably from the ML estimates in the second stage. In the STAR model, for instance, the Bayesian coefficient measuring the impact of education on turnout is less than half the size of the ML estimate.

Table 3. Description of educational experiment data from Sondheimer and Green (2010)

	Perry	IHAD	STAR
Treatment			
HS Graduation	65.00%	78.95%	90.08%
Turnout	18.33%	42.11%	46.83%
N	60	19	252
Control			
HS Graduation	44.44%	61.54%	84.97%
Turnout	12.70%	33.33%	42.22%
N	63	39	559

The Bayesian estimates are also substantially smaller for the IHAD and Perry models where the sample sizes are small. In all, the Bayesian estimates for the effect of education on turnout are all positive. However, they suggest more modest effect sizes and are more likely to include zero in the 95% CI than originally reported.

The differences between the ML and Bayesian models persist when we pool across studies. SG construct a precision-weighted estimate of the coefficient measuring the effect of education on turnout on the probit scale.[9] This is shown on the right column of Table 4 and suggests that there is strong evidence in favor of the effect of high-school graduation on turnout. The Bayesian models, however, only weakly support this conclusion. Figure 3 displays the posterior estimates for the ATEs (left panel) and ATT (right panel) for each study, as well as the pooled posterior density.[10] The posterior mean pooled ATE is 0.24, which suggests that the causal effect of high school graduation on voting in a federal election is a 24% increase. However, there is significant uncertainty. The posterior probability that the pooled ATE is less than zero, $\Pr(\Delta^{(ATE)} \leq 0)$, is approximately 0.08. The right

[9] For the sake of comparison, we create "pooled" estimates for this coefficient using the same method for the Bayesian estimates. However, our preferred method of pooling estimates is shown below.

[10] Pooled estimates are calculated by taking a sample-size weighted average from each draw from the posterior.

Table 4. Bivariate probit regression and Bayesian IV results for the downstream effects of educational attainment on turnout

	Perry		IHAD		STAR		Pooled	
	ML	*Bayesian*	*ML*	*Bayesian*	*ML*	*Bayesian*	*ML*	*Bayesian*
HS Graduation Level								
Intercept	-0.14	-0.13	0.30	0.30	1.04	1.04		
	(0.16)	(0.16)	(0.19)	(0.20)	(0.07)	(0.06)		
Treatment	0.53	0.51	0.51	0.50	0.25	0.25		
	(0.22)	(0.23)	(0.76)	(0.38)	(0.13)	(0.13)		
Voter Turnout Level								
Intercept	-1.6	-1.35	-1.01	-0.40	-1.85	-0.91		
	(0.48)	(0.37)	(0.72)	(0.48)	(0.82)	(0.52)		
HS Graduation	1.08	0.61	1.05	0.09	1.95	0.86	1.35	0.48
	(0.83)	(0.60)	(1.03)	(0.68)	(0.92)	(0.63)	(0.60)	(0.36)
N	123	123	58	58	811	811	992	992
ρ		-0.06		-0.07		-0.15		
		(0.35)		(0.38)		(0.33)		

ML columns replicate Table 5 in Sondheimer and Green (2010). Coefficients are ML estimates and Bayesian posterior means. In parenthesis, we report bootstrapped standard errors for ML estimates and posterior standard deviations for Bayesian estimates. Pooled estimates are calculated by creating precision weighted coefficients.

panel shows similar estimates for the effect of the treatment on the treated.

5. CONCLUSION

In this note, we argue that a Bayesian approach to estimating bivariate probit instrumental variables is superior to standard maximum likelihood routines. Broadly speaking, Bayesian and ML approaches to inference provide virtually identical parameter estimates. However, in the case of bivariate probit instrumental variable models, this appears not to be the case. First, numerical issues with the likelihood surface appear to prevent standard routines from locating the true ML estimator (Freedman and Sekhon 2010). The Bayesian model, however, is better able to recover the true model parameters in many settings. Specifically, our simulation study showed that in settings more typical to actual applied research in the social sciences—where sample sizes are often modest and instruments are often quite weak—the Bayesian approach is more likely to provide accurate inferences.

Second, in an IV setting, the specific parameter values from the bivariate probit model

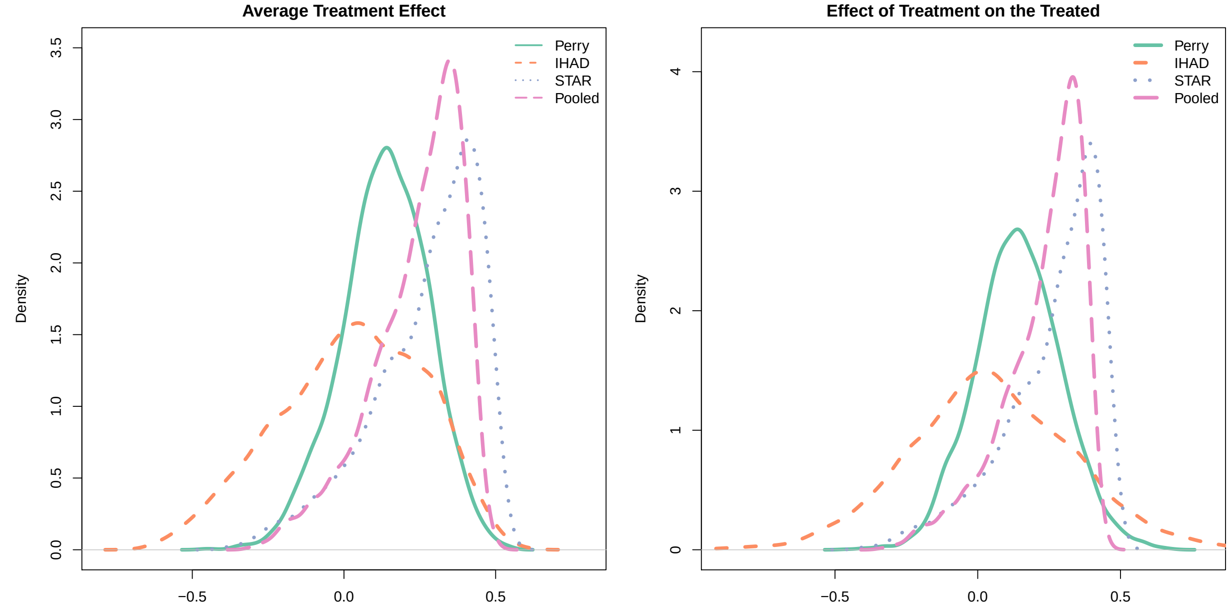

Figure 3. Bayesian treatment effect estimates for the effect of education on voter turnout. These posterior densities of the treatment effects for each data set were obtained by calculating the ATE and ATT at each iteration of the MCMC algorithm. Based on the pooled posterior density for each, P(ATE$\leq$ 0.082) and P(ATT$\leq$ 0.086). The pooled estimates are an average of the estimated ATE and ATT at each MCMC draw weighted by sample size.

are not of direct interest, making interpretation difficult. The Bayesian approach makes it simple to easily calculate key causal quantities of interest along with appropriate measures of uncertainty. We illustrate the advantage of this flexibility by re-estimating the effect of educational attainment and voter turnout (Sondheimer and Green 2010). We provide formulas for the most common quantities of interest in the online Appendix.

In all, the analyses above illustrate how the Bayesian approach can provide more reliable and accurate estimates of treatment effects, and are more flexible for calculating treatment effects in the presence of binary outcomes than traditional ML approaches. Before concluding, however, it is worth noting one important limitation. Specifically, analysts should realize that *any* bivariate probit model requires more restrictive assumptions for estimating causal quantities than alternative approaches such as two-staged least squares (2SLS). By assuming the joint normality of the error distribution, bivariate probit models are less agnostic than 2SLS, which assumes only monotonicity. Thus, while the bivariate probit model allows us to directly handle the dichotomous nature of the data, it comes at the cost of additional parametric assumptions.

With that said, one of the advantages of adopting the Bayesian framework is that it is possible to extend the model in a variety of ways, leveraging the extensive array of tools available in the Bayesian literature, to address these concerns. Chib and Greenberg (2007) provide a variant of the model used here under non-parametric assumptions, and Chib, Greenberg and Jeliazkov (2009) extend this further to allow for non-random sample selection. However, although far more flexible, the models just cited also make specific distributional assumptions about the joint distribution of the treatment and the outcome that may not always be appropriate. In theory, however, the modeling framework can accommodate flexible joint distributions, such as Dirichlet process mixtures (Chib and Greenberg 2010), although we are aware of no such implementations in the literature. A particularly promising approach is provided by Ratkovic and Shiraito (2014), who specify a Bayesian model to directly model compliance status for each observation.

Nonetheless, bivariate probit models are still routinely estimated in the social sciences, and the straightforward Bayesian implementation described here offers significant practical advantages over the standard ML techniques now dominant in the literature. With large samples and strong instruments, all approaches will provide nearly identical answers. Yet, researchers are seldom blessed with this kind of data. Under more realistic circumstances with finite samples and weak instruments, the Bayesian implementation we describe promises to aid researchers in making more accurate, efficient, and interpretable estimates of treatment effects in the context of unmeasured confounders.

References

Angrist, Joshua D., Guido W. Imbens and Donald B. Rubin. 1996. ''Identification of Causal Effects Using Instrumental Variables.'' *Journal of the American Statistical Association* 91(434):444--455.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Ansolabehere, Stephen D., Shanto Iyengar and Adam Simon. 1999. ''Replicating Experiments Using Aggregate and Survey Data: The Case of Negative.'' *American Political Science Review* 93(4):901--909.

Arceneaux, Kevin and David W. Nickerson. 2009. ''Who Is Mobilized to Vote? A Re-Analysis of 11 Field Experiments.'' *American Journal of Political Science* 53(1):1--16.

Barnett, W. Steven. 2010. ''Benefit-Cost Analysis of the Perry Preschool Program and Its Policy Implications.'' *Educational Evaluation and Policy Analysis* 7(4):333--342.

Berinsky, A.J. and G.S. Lenz. 2011. ''Education and Political Participation: Exploring the Causal Link.'' *Political Behavior* 33(3):357--373.

Cederman, Lars-Erik, Simon Hug and Andreas Schadel Julian Wucherpfennig. 2015. ''Territorial Autonomy in the Shadow of Conflict: Too Little, Too Late?'' *American Political Science Review* 109(2):354--370.

Chib, S., E. Greenberg and I. Jeliazkov. 2009. ''Estimation of Semiparametric Models in the Presence of Endogeneity and Sample Selection.'' *Journal of Computational and Graphical Statistics* 18(2):321--348.

Chib, Siddhartha. 2003. On Inferring Effects of Binary Treatments with Unobserved Confounders. In *Bayesian Statistics*, ed. J.M. Bernardo, M.J. Bayarry, J.O. Berger and A.P. Dawid. Vol. 7 Oxford: Oxford University Press pp. 65--84.

Chib, Siddhartha and Edward Greenberg. 2007. ''Semiparametric Modeling and Estimation of Instrumental Variable Models.'' *Journal of Computational and Graphical Statistics* 16(1):86--114.

Chib, Siddhartha and Edward Greenberg. 2010. ''Additive Cubic Spline Regression with Dirichlet process mixture errors.'' *Journal of Econometrics* 156(2):322--336.

Christin, Thomas and Simon Hug. 2012. ''Federalism, the Geographic Location of Groups, and Conflict.'' *Conflict Management and Peace Science* 29(1):93--122.

Freedman, David A. and Jasjeet S. Sekhon. 2010. ''Endogeneity in Probit Response Models.'' *Political Analysis* 18(2):138--150.

Fuhrmann, Matthew and Todd S. Sechser. 2014. ''Signaling Alliance Commitments: Hand-Tying and Sunk Costs in Extended Nuclear Deterrence.'' *American Journal of Political Science* 58(4):919--935.

Gerber, Alan S. and Donald P. Green. 2000. ''The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment.'' *American Political Science Review* 94(3):653--663.

Hanushek, Eric A. 1999. ''Some Findings From an Independent Investigation of the Tennessee STAR Experiment and From Other Investigations of class size effects.'' *Educational Evaluation and Policy Analysis* 21(2):143--163.

Heckman, James J, Seong Hyeok Moon, Rodrigo Pinto, Peter A Savelyev and Adam Yavitz. 2010. ''The Rate of Return to the HighScope Perry Preschool Program.'' *Journal of Public Economics* 94(1):114--128.

Hoffman, Matthew D. and Andrew Gelman. 2014. ''The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo.'' *The Journal of Machine Learning Research* 15(1):1593--1623.

Jackman, Simon. 2000. ''Estimation and Inference via Bayesian Simulation: An Introduction to Markov Chain Monte Carlo.'' *American Journal of Political Science* 44(2):375--404.

Kahne, Joseph and Kim Bailey. 1999. ''The Role of Social Capital in Youth Development: The Case of ''I Have a Dream'' Programs.'' *Educational Evaluation and Policy Analysis* 21(3):321--343.

Maves, Jessica and Alex Braithwaite. 2013. ''Autocratic Institutions and Civil Conflict Contagion.'' *Journal of Politics* 75(2):478---490.

Pierskalla, Jan H. and Florian M. Hollenbach. 2013. ''Technology and Collective Action: The Effect of Cell Phone Coverage on Political Violence in Africa.'' *American Political Science Review* 107(207--224).

Ratkovic, Marc and Yuki Shiraito. 2014. ''Strengthening Weak Instruments by Modeling Compliance.'' Paper presented at the 2014 annual meeting of the Midwest Political Science Association.

Rosenbaum, P.R. and Donald B. Rubin. 1983. ''Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome.'' *Journal of the Royal Statistical Society. Series B (Methodological)* 45(2):212--218.

Rubin, Donald B. 1974. ''Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.'' *Journal of Educational Psychology* 66(5):688--701.

Sondheimer, Rachel Milstein and Donald P. Green. 2010. ''Using Experiments to Estimate the Effects of Education on Voter Turnout.'' *American Journal of Political Science* 54(1):174--189.

Sovey, Allison J. and Donald P. Green. 2010. ''Instrumental Variables Estimation in Political Science: A Readers' Guide.'' *American Journal of Political Science* 55(1):188--200.

Tenn, S. 2007. ''The Effect of Education on Voter Turnout.'' *Political Analysis* 15(4):446--464.

Wucherpfennig, Julian, Philipp Hunziker and Lars-Erik Cederman. Forthcoming. ‘‘Who Inherits the State? Colonial Rule and Postcolonial Conflict.’’ *American Journal of Political Science* .