Known Unknowns and Unknown Knowns

Incorporating Uncertainty in Second-Stage Estimation

Christopher N. Lawrence

Department of Political Science, Tulane University
New Orleans, Louisiana 70118-5698
c.n.lawrence@gmail.com

Abstract

Recent political science research has seen a resurgence in interest in estimating latent variables (including ideal points, political sophistication, and democratization) using item-response theory modeling and other factor-analytic techniques. Yet, despite these models offering advantages over summated scales other techniques, one key advantage–the estimation of the uncertainty of estimates of the latent variable–is often discarded in second-stage analysis, such as efforts to explain roll-call voting behavior or incorporating estimated knowledge into explanations of voter decision-making.

Here I demonstrate a technique known as simulation-extrapolation estimation (SIMEX) for incorporating uncertainty into these estimates, and compare estimates using standard estimators such as ordinary least squares linear regression and maximum-likelihood probit regression to their SIMEX counterparts using latent-variable estimates with both low (estimates of legislator ideal points) and high (estimates of voter sophistication) variance. These results demonstrate the value of including known error variance in second-stage estimates without resorting to the use of structural-equation model approaches.

1. Background

One of the more desirable trends in recent political science research has been an increasing concern with quantifying measurement error when determining the likely quantities of variables that cannot be directly observed, such as legislators' and Supreme Court justices' ideal points (Clinton, Jackman and Rivers 2004; Martin and Quinn 2002), the level of democracy of a state (Bollen 1993; Treier and Jackman 2005), and the level of political sophistication of voters (Levendusky and Jackman 2003; Lawrence 2003, 2007). Most of these measurement approaches are based on the use of item-response theory models (Johnson and Albert 1999), a factor-analytic technique originating in psychological and educational testing which has been shown to have application to the measurement of a wide variety of other phenomena as well.

However, this concern with measurement error has not been matched with a concern for carrying through the measurement error in subsequent analysis, despite evidence that inferences may be subject to unknown bias in the absence of accounting for this error; see e.g. Martin and Quinn (2005). The most common approach in the literature is to use the mean (or, occasionally, median) estimate of the ideal point (or level of sophistication or democracy) recovered from the IRT procedure as either an independent or dependent variable in subsequent analysis.

I demonstrate a technique known as simulation-extrapolation estimation (or SIMEX), developed by Cook and Stefanski (1994) and implemented in the R `simex` package (Lederer and Küchenoff 2006). This procedure produces nearly asymptotically unbiased and efficient estimates for both linear regression and common generalized linear models, through either the jackknife or an asymptotic approximation.

2. Application: Legislator Ideal Points

The ideal points of legislators would appear to be a case where error variances would not be problematic; more than one commentator has described this application of the IRT model as a classic case of a repeated measurement with very similar items, so we would expect low variances.[1] After 548 votes in the 110th Congress, the mean standard deviation of the estimated ideal points for House members is 0.09, producing an estimated 95% credible interval of ± 0.177. However, this small degree of mean variation obscures the fact that some members have high estimated variances;[2] it also is non-negligible, as the mean difference between many legislators is within the estimated error variance.

Below are comparisons between standard (OLS) and SIMEX jackknife estimates of models seeking to predict member ideal points; i.e. with the estimated ideal points as the *dependent* variable:

	OLS estimates			SIMEX estimates		
	Coef.	Std. Err.	p	Coef.	Std. Err.	p
(Intercept)	-2.389	0.097	$< .001$	-2.394	0.094	$< .001$
Member black	-0.081	0.081	n.s.	-0.087	0.082	n.s.
Member Hispanic	-0.226	0.091	$\approx .013$	-0.212	0.091	$\approx .020$
Bush vote % '04	0.025	0.002	$< .001$	0.025	0.002	$< .001$
Member female	-0.133	0.054	$\approx .013$	-0.133	0.054	$\approx .013$
Member GOP	1.448	0.057	$< .001$	1.447	0.057	$< .001$
Dist. minority %	0.001	0.001	n.s.	0.001	0.001	n.s
Minority % $\times$ GOP	0.002	0.002	n.s.	0.002	0.002	n.s.
Female $\times$ GOP	0.125	0.093	n.s.	0.131	0.094	n.s.

Table 1: *Explaining ideal points of House members, 110th Congress; linear regression estimates.*

In this case, we do not see very large differences in the estimates; the coefficient of the "Member Hispanic" variable decreases somewhat, but the other estimates remain fairly stable.

In some instances where ideal points are used as an explanatory variable, however, the differences are more pronounced, as in the table below, which presents results from a probit model of voting on final passage of legislation that would allow federal funding of more forms of embryonic stem cell research:

	ML estimates			SIMEX estimates		
	Coef.	Std. Err.	p	Coef.	Std. Err.	p
(Intercept)	1.139	1.032	n.s.	0.705	1.149	n.s
Member female	0.539	0.286	$\approx .059$	0.536	0.289	$\approx .064$
Member GOP	0.661	0.526	$\approx .209$	0.918	0.604	$\approx .129$
Member ideology	-1.686	0.383	$< .001$	-1.882	0.453	$< .001$
Bush vote % '04	-0.031	0.015	$\approx .047$	-0.026	0.016	$\approx .118$

Table 2: *Explaining voting on stem-cell research legislation (roll call 1-443, 110th Congress); probit estimates.*

The rows highlighted in the table show notable changes in coefficient effects. The most noteworthy change is in the last row of the table, where the effect of constituency support for the president meets the traditional two-tailed .05 significance test in the ML probit model but fails even a .10 significance test (or a .05 single-tailed test) when estimated with the SIMEX procedure. In other words, a researcher examining the ML estimates would conclude that the legislator was voting at least in part based on constituency preferences (as reflected by their support for Bush in the 2004 election, when he took a public position in opposition to federal funding for most stem cell research), while the SIMEX estimates incorporating the uncertainty in legislator ideology indicate that we can only be confident that their ideology affected legislators' votes.

3. Application: Voter Sophistication

By contrast with the case above, estimates of voter sophistication tend to have high measurement error, in large part because most surveys administered to voters have relatively few (if any) knowledge items, despite some evidence that "test fatigue" is not problematic (Delli Carpini and Keeter 1996). I previously developed knowledge measures from the American National Election Study and Dutch Parliamentary Election Study containing 12–44 items, varying between surveys; compared with the nearly 550 items (and growing) in a single session of the House, these surveys provide a paucity of items from which we can recover estimated sophistication levels. In the case of the 2000 NES, with 28 items, the mean standard deviation of respondents' estimated sophistication was 0.332, corresponding to a 95% credible interval of approximately ± 0.651; the *range* of the estimated means for respondents was only 4.5, indicating a great deal of imprecision.[3]

Below are comparisons between models estimated using standard probit and SIMEX seeking to explain voters' perceptions of whether or not the presidential election was likely to be close:

	ML estimates			SIMEX estimates		
	Coef.	Std. Err.	p	Coef.	Std. Err.	p
(Intercept)	0.238	0.235	n.s.	0.310	0.249	n.s.
Sophistication	0.260	0.100	$\approx .009$	0.313	0.117	$\approx .008$
Age	0.007	0.003	$\approx .008$	0.007	0.003	$\approx .012$
Education	0.139	0.034	$< .001$	0.135	0.034	$< .001$
PID Strength	-0.023	0.049	n.s.	-0.035	0.052	n.s.
Income	0.015	0.014	n.s.	0.013	0.014	n.s.
South	0.019	0.096	n.s.	0.019	0.096	n.s.
Soph. $\times$ PID Str.	-0.091	0.046	$\approx .046$	-0.107	0.053	$\approx .043$

Table 3: *Explaining pre-election beliefs about election closeness in 2000; probit estimates.*

While the differences in the model estimates are not dramatic, the SIMEX estimates clearly suggest a stronger effect of political sophistication than the ML estimates based on means, both in terms of the linear effect and its interactive effect with strength of party identification.

4. Conclusion

In this poster I have demonstrated the value of incorporating known measurement error in second-stage estimation when using estimates recovered from IRT models (or other techniques) using a simulation-extrapolation procedure established in the statistical literature and available in a number of statistical packages, including Stata and R.

However, the SIMEX technique could be improved to better account for *individual-level* uncertainty in the estimates; the procedure in R presently assumes that uncertainty is constant across observations, but the IRT procedure clearly indicates that this is not the case. I plan to extend the technique to allow for the uncertainty to vary across observations in a future iteration of this research.

References

Barone, Michael and Richard E. Cohen. 2006. *The Almanac of American Politics, 2006*. Washington, D.C.: National Journal.

Bollen, Kenneth. 1993. "Liberal Democracy: Validity and Method Factors in Cross-National Measures." *American Journal of Political Science* 37:1207–30.

Burns, Nancy, Donald R. Kinder, Steven J. Rosenstone, Virginia Sapiro and the National Election Studies. 2002. *American National Election Study, 2000: Pre- and Post-Election Survey [Computer file]*. Ann Arbor, Mich.: University of Michigan, Center for Political Studies/Inter-university Consortium for Political and Social Research. ICPSR Study No. 3131.

Clinton, Joshua, Simon Jackman and Douglas Rivers. 2004. "The Statisical Analysis of Roll Call Data." *American Political Science Review* 98:355–37.

Cook, J.R. and L.A. Stefanski. 1994. "Simulation-Extrapolation Estimation in Parametric Measurement Error Models." *Journal of the American Statistical Association* 89:1314–28.

Delli Carpini, Michael X. and Scott Keeter. 1996. *What Americans Know about Politics and Why It Matters*. New Haven, Conn.: Yale University Press.

Jackman, Simon. 2007. *pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory, Stanford University*. R package version 0.92.
URL: *http://pscl.stanford.edu*

Johnson, Valen E. and James H. Albert. 1999. *Ordinal Data Modeling*. New York: Springer-Verlag.

Lawrence, Christopher N. 2003. *The Impact of Political Sophistication on the Decision-Making Processes of Voters*. PhD thesis. Oxford, Miss.: University of Mississippi.
URL: *http://www.cnlawrence.com/papers/dissertation.pdf*

Lawrence, Christopher N. 2007. "Should Voters Be Encyclopedias? The (Mis?)estimation of Voter Sophistication Based on Political Knowledge Items." Presented at the 2007 annual meeting of the Midwest Political Science Association, Chicago, Ill.
URL: *http://www.cnlawrence.com/papers/irt-soph.pdf*

Lederer, Wolfgang and Helmut Küchenoff. 2006. *simex: SIMEX and MCSIMEX algorithms for measurement error models*. R package version 1.2.

Levendusky, Matthew S. and Simon D. Jackman. 2003. "Reconsidering the Measurement of Political Knowledge." Working Paper.
URL: *http://www.stanford.edu/class/polisci353/2004spring/reading/levendusky_final.pdf*

Lewis, Jeffrey. 2007. "110th Congress roll call data.". Retreived June 26, 2007.
URL: *http://adric.sscnet.ucla.edu/rollcall/*

Martin, Andrew D., and Kevin M. Quinn. 2007. *MCMCpack: Markov chain Monte Carlo (MCMC) Package*. R package version 0.8-2.
URL: *http://mcmcpack.wustl.edu*

Martin, Andrew D. and Kevin M. Quinn. 2002. "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999." *Political Analysis* 10:134–53.

Martin, Andrew D. and Kevin M. Quinn. 2005. "Can Ideal Point Estimates be Used as Explanatory Variables?" Unpublished working paper.
URL: *http://mqscores.wustl.edu/media/resnote.pdf*

R Development Core Team. 2007. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.
URL: *http://www.R-project.org*

Treier, Shawn and Simon Jackman. 2005. "Democracy as a Latent Variable." Updated version of a paper prepared at the 2003 Annual Meeting of the Society for Political Methodology, Minneapolis, Minnesota.
URL: *http://www.arches.uga.edu/ satreier/DemocracyAsLatentVariable.pdf*

[1] The ideal point estimates in this section were estimated using the `ideal` procedure in the `pscl` package for R (Jackman 2007; R Development Core Team 2007) using data from the 110th Congress provided online by Lewis (2007); only a single dimension was estimated. Additional covariates were added to the roll-call data using the *Almanac of American Politics* (Barone and Cohen 2006).

[2] Notably, the speaker of the House, Nancy Pelosi (D-Ca.), in addition to being the member with the left-most mean ideal point, has an estimated standard deviation of 0.66, due to the tradition of the speaker rarely voting on the House floor. Members who died in office, who were expelled, or are running for other office and frequently absent from Congress will also have high variance estimates.

[3] Estimates were produced using the `MCMCirtKd` procedure in the MCMCpack package for R (Martin, and Quinn 2007) using data from the 2000 American National Election Study (Burns et al. 2002).