Taming the Selection Bias
An Application to Compliance with International Agreements

Hyeran Jo*

September 12, 2008

Abstract

This paper has two goals. One is to examine the existing statistical techniques to correct for the selection bias with a particular attention to Heckman-type models and matching. After the survey of relevant models and their assumptions, I offer a practical advice in choosing one model over the other. The second goal is to consider the question of bias-variance tradeoff, which has been one of the weakest links in the matching literature when dealing with data of a fixed sample size. Matching is designed to recover balance for causal inference but the efficiency loss should be carefully considered since the matching leaves some subsamples unmatched. Using the dataset on the compliance with international agreements (Simmons 2000; Von Stein 2005; Hopkins and Simmons 2005), I demonstrate the consequence of this bias-variance tradeoff in estimating the average treatment effect.

*Assistant Professor. Department of Political Science, Texas A&M University, College Station, TX 77843–4348. Email: `hyeranjo@politics.tamu.edu`. Phone: 979–845–3038. This paper is prepared for the presentation at the Visions in Methodology (VIM) Conference at the Ohio State University, Oct 2–4, 2008. Earlier version of this paper was presented at the 2005 Annual Meeting of MPSA, EITM section. This is preliminary draft. Please do not cite without author's permission. I thank Ben Hansen and Dan Hopkins for helpful discussion.

Contents

1 Introduction

What is the best way to deal with the selection problem that is present in political science? Politicians select themselves into the employment of certain political strategies; judges select themselves into the cases; survey responders sometimes choose to be silent[1]; countries choose their battlegrounds.[2] This paper first characterizes the nature of the selection problem and reviews existing statistical methods employed in political science literature. I then focus on the specific problem of bias-variance tradeoff, which has serious implications for observational studies with relatively small and fixed samples.

2 Selection: Problems and Solutions in Political Science

This section defines the selection problem, provides the typology of selection problem, and reviews existing methods as practiced in political science literature.

2.1 Nature of the Selection Problem: What is the Selection Bias, Anyway?

Selection problem is present when the population of interest presents a systemic bias to the inference.[3] Let y be the observed outcome, x the coveriates, and z is the treatment actually received. The problem is that we want to make inference about $P(y \mid x)$ but actually the data we observe has a structure of $P(y, z \mid x)$.[4]

[1]See Berinksy 2004 for instance.

[2]See Huth 1998 how countries are selected into border disputes; See Achen and Snidal 1989 for the difficulty of empirical tests of military deterrence since those countries that have been deterred do not enter into the sample.

[3]For good reviews of selection problems and related statistical evaluations, see Manski 2004 for the characterization of the problem; Achen 1986 and Breen 1996 for the typology of selection problems; Geddes 2007 in the context of comparative politics; Ho, Imai, King, Stuart 2008, Heckman 2008, and Todd 2006 for statistical methods; Langbein and Felbinger 2006 for selection problem in the context of policy analysis.

[4]This formal characterization is due to Manski 2004.

In this paper, selection problem refers to a problem, encompassing any systematic data–generating process that can hurt inference by introducing bias. All the selection problems are not borne equal, however. The sources of selection are diverse from non-response in surveys, to missing data, and to strategic selection.[5] Normally, the literature identifies two kinds of selection problems. One is endogeneity (non-random assignment to treatment) and the other is sample selection problem (partial observability problem; censoring).[6] Some selection problems are strategic – that is, deliberate choice of political actors[7]; others are sampling issues, where a subset of the data is systematically excluded due to a particular attribute.[8] Selection problems categorized in these types are related to other kinds of statistical problems. Missing data problems are shown to be equivalent to selection problems.[9] Omitted variable bias is shown to be equivalent to selection problem.[10]

2.2 Existing Statistical Models Dealing with Selection Bias

Dominant approaches to correct for selection bias include Heckman selection (with various names such as two-stage model, instrumental variable approach), matching, and bivariate probit.[11]

Heckman-type selection model considers selection in two stages. We are interested in the outcome equation $y = x\beta + u$ but account for the selection process of $z = w\alpha + e$, where x and w are covariates and u and e are random disturbances. The goal of matching

[5] Many game-theoretic models produce counterfactual scenarios of strategic selection, generating empirical implications for selection problem. See for example, Schultz 2001.

[6] See Achen 1986 for the distinction between the two kinds of selection problems.

[7] See for instance Vreeland and Preworski 2002 on the problem of partial observability in IMF agreements.

[8] Breen 1996 classifies three sets of sampling issues of censored, sample selected, and truncated data.

[9] Missing data are usually counterfactual outcomes that we are interested in. Manski 2004 formally shows the equivalence of missing data problems with selection problem.

[10] Heckman 1979

[11] Other models include control function methods, quite common in econometrics literature. I do not review this because it is not used in political science literature. For control function methods, see Todd 2006 and Heckman and Navarro 2003. Some works address the issue of selection in specific class of models. Boehmke, Morey, and Shannon 2006 deals with non-random selection in duration models. Semykina and Wooldridge 2007 consider estimation of panel data models with sample selection. Su 2008 considers matching in the context of multilevel data.

is similar – correcting for the selection process to the treatment. Matching aims to reduce bias and estimate the Average Treatment Effect (ATE): $E(Y_1 - Y_0 \mid T = 1)$. The basic idea of matching is to pre-process the data and re-sample such that treated and control groups achieve similar propensity to the treatment. Bivariate probit model with sample selection is applicable to discrete choice models and estimates two equations $y_1 = \beta X + \epsilon$ and $y_2 = \beta X + \epsilon$ where y_1 and y_2 are latent variables and y_1 is observed only if $y_2 = 1$. Both Heckman selection models and bivariate probit models assume the bivariate normal distributions for the disturbances in two equations. Matching in contrast does not have distributional assumption and is thus called a non-parametric method.

Table 1: Practice of Political Science Dealing with Selection Bias: Examples

Methods	Data Type	Research Topic	Examples
Heckman Selection	Observational	Compliance with Intl Agreements	Von Stein 2006
Matching	Observational	Compliance with Intl Agreements	Simmons and Hopkins 2006
Matching	Field Experiment	Campaign Voter Turnout	Hansen and Bowers 2006; Arceneaux, Gerber and Green 2006
Heckman Selection	Survey	Public Opinion	Berinsky 2004
Bivariate Probit	Observational	IMF	Vreeland and Preworski 2002; Boehemke et al. 2006

2.3 Practice in Political Science: A Brief Survey

To gauge the practice of selection problems, I looked at major journals (*APSR*, *AJPS*) and the working papers in *PolMeth* archive. The search words include selection bias, selection problem or matching. Some examples are presented in Table 1. A noticeable trend is that different kinds of data are subject to different degrees of selection problem. For instance, experimental data rarely touches upon selection problem as the experiments are conducted in controlled environments.[12] Survey researchers are not too concerned about selection problem when the design of survey questions can screen out selection problems, except the situations of non-compliance.[13] Observational studies or quasi-experimental data is most susceptible to selection problems.

[12] This is not to say selection problem does not exist in experimental data. See Keele, McConnaughy, and White 2008 for the non-randomization in experimental data.

[13] See Imai 2008 for instance.

3 Choice of Selection Models

The choice of various methods to control for selection effects involves careful examination of underlying assumptions. My goal in this section is to examine the assumptions and discuss the practical implications of the employment of each method.

3.1 Weighing the Assumptions

Recent works started to compare the performance of various selection models.[14] The verdict of favoring one model over the other is not out yet. The performance of each model is generally tested with Monte Carlo simulations, where the test is conducted with a sample data with a known true estimate. However, the merits of model assumptions differ across different data problems, and therefore, it is difficult to conclude that one method is superior to another in every situation.

The major theoretical difference lies in what each method assume about the unobservables (or missing variables). Matching assumes that the selection is only on observables. This assumption is frequently called the ignorability (or unconfoundedness or exogeneity) assumption.[15] In contrast, Heckman selection models account for unmeasured factors that are related to the outcome, and therefore rely on instrumental variables to get at the unobservables. This assumption about unobservables define the weaknesses and strengths for each model.

Heckman-type selection models allow for a test for selection bias but it requires strictly exogenous instruments, which can be taxing in many empirical cases.[16] The error terms in

[14] Arceneaux, Gerber and Green 2006 compares matching and regression estimators, showing matching can produce biased results. Freedman and Sekhon 2008 point out numerical problems of biprobit likelihood function. Lo 2008 runs Monte Carlo simulation, reporting bivariate probit better recovers true parameters than matching and that selection models perform better than matching when the effect of unobservables are present.

[15] Ignorability assumption has been suggested as the Achilles heels of matching methods but recently sensitivity tests for this ignorability assumption are being developed. See Ichino, Mealli and Nannicini (2005) for instance.

[16] See Sartori 2000 for instance

Table 2: Potential Strengths and Weaknesses of Matching and Heckman Selection Models

Matching	Heckman Selection Models
Non–parametric (S)	Distributional Assumption (W) (bivariate normality)
Strengthen Internal Validity (S)	Exclusion restriction (W) Finding instruments
Ignorability Assumption (W) Selection on Observables	Selection and Outcome Equations (S)
Bias–Variance Tradeoff (?)	

selection and outcome equations are assumed to be normally distributed while matching is non-parametric.

As a practical matter, matching is good for looking at the data in a new light. It is beneficial to check the balance with matching. The logic of matching allows us to think about the ways to think about counterfactuals up front. Grouping countries or individuals into control and treatment groups can also be fruitfully used in qualitative case studies.[17] One practical advantage of Heckman selection models is that they let us model the selection process explicitly, although finding the right instrument is often a great challenge.

Table 2 summarizes the weaknesses and strengths of Matching and Heckman selection models. It discusses both theoretical and practical concerns in using either method. The last item, bias–variance tradeoff, is discussed in the following section.

3.2 Practical Advice for Selecting Methods

This section provides a practical guide on selecting among the existing methods, based on the assumptions built into each statistical model discussed in the previous section. The examination of the existing methods suggests several research strategies one can take in

[17] For instance, the unmatched sample can be considered as outliers in the spirit of Lieberman 2005. See Sekhon 2007 for the detailed discussion on the link between qualitative studies and matching methods.

practice. For instance, when the violation of ignorability assumption is highly suspect, one would use Heckman selection. This suspicion may arise when researchers think potential variables that may affect the behavior but does not have good handle on the measure.

The steps involve checking the assumptions of the above models as well as the nature of the data. The following questions should be asked either simultaneously or sequentially:

1) *Is data missing at random? Can a meaningful sample be constructed in any way?* The first decision of missing data is based on the knowledge of substantive aspects of the data. This question should be asked at the data construction or collection stage. This question will be helpful in identifying the nature of the selection problem, and the analysts would be able to control the collection of data if possible.

2) *Is the selection on observables only?* When a researcher is confident that the covariates at hand account for the treatment, then one can choose matching. If a researcher has some doubts about the unobservable factors affecting the treatment,[18] then matching may not be a right method to choose.[19] One wants to check at this point whether Heckman selection model is warranted by checking ρ.

3) *Can an instrumental variable that can satisfy exogeneity assumption be easily found?* This happens rarely, but if an instrumental variable can be justified, a researcher may opt for Heckman-type models. The easiness of finding instrumental variable can be one of the crucial criteria.[20]

[18]akin to the problem of distinguishing between unknown unknowns and known unknowns, a la Donald Rumsfeld

[19]Lo 2008 nicely shows via Monte Carlo simulation that matching performs well with low values of ρ, but as mentioned previously, the test for ignorability assumption is being developed as in Ichino, Mealli, and Nannicini 2008.

[20]Gilligan and Sergenti 2008 for instatnce defend their choice of matching over other methods based on the substantive theoretical reason that no adequate instrumental variable can be found that can influence selection equation but not the outcome equation.

4 Bias–Variance Tradeoff in Matching Algorithms

I now focus on the problem of bias–variance tradeoff in matching (Rubin 1976; Dehejia and Wahba 1999; Smith 1997; Abadie and Imbens 2007) – a crucial consideration in estimating the effect of treatment in fixed samples. As a researcher, we want to reduce bias and have lower variance. However, in many cases, lower variance and reduced bias are two competing demands in pre-processing the data. In what follows, I demonstrate this tradeoff in the context of art 8 compliance and discuss the implications of the tradeoff for observational studies in general.

The central concern in matching is the reduction of bias in making inferences about the effect of treatment. This logic of preferring less bias to less variance is illustrated by Rubin (2006) "First, since it is generally not wise to obtain a very precise estimate of a drastically wrong quantity, the investigator should be more concerned about having an estimate with small bias than one with small variance. Second, since in many observational studies the sample sizes are sufficiently large that sampling variances of estimators will be small, the sensitivity of estimators to biases is the dominant source of uncertainty."

The recent work by Icaus, King, and Porro (2008) develops a method to balance the tradeoff, assuming that "the data are typically plentiful." However, what happens if the sample size is small?[21] Pre–processing invariably leaves unmatched observations. Dropping observations is not especially appealing to international relations scholars, although the ultimate goal of causal inference is to obtain an unbiased estimate. The limitations of the data in international relations research is clear: the sample size is usually fixed, one cannot have additional experiments, and sometimes outliers are important.

So, what does this bias–variance trade–off tell us about the ways to deal with selection problem? This paper does not provide the answer as to when a researcher relies on one method over the other. Rather, in what follows, I demonstrate that the bias–variance

[21] The consistency and efficiency of large sample properties are proved in Imbens and Abadie 2006 where they develop what they call "bias-corrected matching" that corrects for bias without efficiency loss.

trade–off harms the inference in an unexpected way in observational data with fixed sample. Researchers using matching should well recognize the trade–off between efficiency and bias with a fixed sample size. As a practical consideration therefore, one should report efficiency reduction as well as bias reduction, using various matching algorithms.

4.1 Empirical Test of the Bias and Efficiency: *Article VIII* Data Set

I examine the bias–variance tradeoff with the issue of compliance with the Article VIII capital account restriction, a point of debate in the series of articles in *American Political Science Review* (Simmons 2000; Von Stein 2005; Simmons and Hopkins 2005). The data set is the record of current account restriction behavior of 133 countries between the time periods of 1962-1997. The data has been recorded unevenly and the average observation window is six years. In matching analysis, treated groups are the countries who adopted article 8 in their five year window (Simmons and Hopkins 2006) and control groups are the ones that did not sign article 8 in their fifth year window.

What complicates the treatment is the fact that countries voluntarily sign on to (or select themselves into) the article 8 agreement with the International Monetary Fund(IMF).[22] We are interested in the effect of the signing article 8. However,just by looking at the relationship between the restriction behavior and potential factors that affect the behavior would not be the right way to quantify the effect of article 8. It is because some countries' initial propensity to enter into the agreement is different. If countries expect that they will have easy time complying with the capital account restriction, they are more likely to sign article 8.

Then the question is whether article 8 has its teeth given the fact that countries can self-select into the international agreement. To characterize the problem fully resorting to the description of selection problem *a la* Manski 2004, $Pr(y, z|x)$, the behavior of interest y is the restriction behavior; the covariates that affect the restriction behavior, X include

[22] This selection bias is common to the study of the effect of international agreements as shown in Mitchell and Hensel 2007.

political and economic variables, such as economic openness, volume of capital reserves, and type of political regime. The treatment actually received, z is the acceptance of article 8.

The quantity of interest is the restriction behavior of countries. The table on the next page shows the results of various matching algorithms, compared to the estimates reported by Von Stein 2005 and Hopkins and Simmons 2005. Depending on the matching algorithm, the magnitude of the treatment effect varies and substantive conclusions change.[23] The raw model reported in Simmons 2000 finds the average 18% reduction in current account restriction behavior. Heckman-selection model run by Von Stein 2005 reports almost no effect of article 8 after the signing year. The exact match conducted by Simmons and Hopkins 2005 finds about 8% increase in the reduction behavior, indicating a constraining effect of article 8. The optmatch run by the author returns almost no effect of article 8. We cannot tell clearly which is the right estimation because of bias-variance tradeoff I highlighted before. If one favors bias (which we usually aim for in point estimation), we should buy the result by Simmons and Hopkins. However, if we do not want to sacrifice too much efficiency, we may conclude that the treaty does not have effect.

[23]Exact match was performed with .25 caliper (one standard deviation) propensity score. [24] Optmatch is the algorithm developed by Hansen 2004 and was performed with the treated-to-control ratio of 1:2. Fullmatch is a greedy matching algorithm with replacement, which is the extreme case of using all the data available.

Table 3: Estimating the Magnitude of Treaty Effect of Art 8

Author	method	Reduction of current account restriction behavior (with 95% confidence interval)
Simmons 2000	Time–series cross–section logit model	Average effect 18%
Von Stein 2005	Heckman selection	Signatories 13% less likely to restrict than non-signatories for the first year of signing no effect post-signing years
Hopkins & Simmons 2005	Exact Match (44 treated, 44 control)	Signing Year 17.7% (.7, 35.6) Year after signing 24.2% (3.9, 43.1)
Jo 2006	optmatch (66 treated, 132 control)	signing year: 9% (.2, 16) Year after signing:7% (.8, 14)
	fullmatch (66 treated, 1634 control)	signing year: 10% (8.4, 12.2) Year after signing: 7% (4.9, 9.8)

In terms of bias-efficiency tradeoff, observe that optmatch and fullmatch present tighter confidence interval, compared to exact match. This means optmatch and fullmatch reduced variance and increased efficiency by using more samples.

Fullmatch sacrifices a lot of bias but optmatch fares well. The example shows that in the case of fixed small sample observational data, optmatch performs better in balancing sharper estimates and adequate bias. In contrast, Exactmatch sacrifices too much efficiency while Fullmatch penalizes bias harshly.[25]

The empirical example of art8 offers a cautionary tale that the magnitude of average treatment effect may differ depending on how matches are made. The average treatment effect using various matching algorithms should be reported because of the sensitivity of results depending on the bias–variance tradeoff. Although Monte Carlo simulations have been conducted to report the performance of each matching algorithm, the particulars of each data set (e.g. distribution structure of covariates) may respond to each algorithm differently. The analysis presented above reinforces the idea we bring in more engaging discussion about efficiency in matching of observational studies of a limited sample size.

5 Conclusion

It would be ideal to be in control over data collection as in randomized experiments or some surveys. In observational data however one does lose control over such manipulation and has to choose one method over the other to take care of selection bias. This paper considered the choice sets available to applied researchers and proposed a sequential research strategy based on the assumptions proposed in the literature: 1) collecting the data to avoid potential selection problem by constructing a fictitious or likely sample, 2) matching, to estimate the treatment effect and to re-group the data into treated and control groups with reduced bias, 3) Heckman selection models, if selection on observables is suspect, 4) comparing

[25] I should note the recent development in other matching algorithms having this trade–off in mind, including GenMatch (Sekhon 2008) and Synthetic Matching (Hainmueller 2008).

the estimates from various models, including various matching algorithms. In any route a researcher takes, it will require a lot of data-looking, robust tests, and sensitivity analysis.

References

[1] Achen,Christopher. 1986. *The statistical analysis of quasi-experiments*. Berkeley: University of California Press.

[2] Berinsky, Adam. 2004. *Silent Voices: Public Opinion and Political Participation in America.* Princeton University Press.

[3] Arceneaux, Kevin, Alan S. Gerber, and Donald P. Green. 2006. "Comparing Experimental and Matching Methods using a Large–Scale Voter Mobilization Experiment." *Political Analysis*. 14(Winter): pp. 37-62.

[4] Abadie, A. and G. Imbens. 2006. "Large Sample Properties of Matching Estimators for Average Treatment Effects." *Econometrica* 74(1). pp. 235–267.

[5] Boehemke, Frederick, Daniel Morey, Megan Shannon. 2006. "Selection Bias and Continuous-Time Duration Models: Consequences and a Proposed Solution." *American Journal of Political Science*. 50 (1). pp. 192–207.

[6] Breen, Richard. 1996. *Regression models : censored, sample selected or truncated data.* Thousand Oaks, CA. : Sage Publications.

[7] Freedman, David, and Jasjeet Sekhon. 2008. "Endogeneity in Probit Response Models" *Working Paper.* UC Berkley.

[8] Geddes, Barbara. 1990. "How the Cases You Choose Affect the Answers You Get: Selection Bias in Comparative Politics." *Political Analysis*. 2. pp. 131–150

[9] Hainmueller, Jens, and Alexis Diamond. 2006. "Beyond Balance: A Placebo Test for Matching Estimators." *Paper presented at the Midwest Political Science Association.*

[10] Hansen, Ben. 2004. "Full Matching in an Observational Study of Coaching for the SAT." *Journal of the American Statistical Association*. 99(467). pp. 609–618.

[11] Heckman, James, and Edward Vytlacil. 2008. "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation" in *Handbook of Econometrics*. Vol.6B. pp.4779–4867.

[12] Heckman, James. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47(1): pp. 153–161.

[13] Hainmueller, Jens. 2008. "Synthetic Matching for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies." *Working Paper*. Harvard University.

[14] Ho, Daniel, Kosuke Imai, Gary King, and Elizabeth Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference," *Political Analysis* 15(3): pp. 199–236, http://gking.harvard.edu/files/abs/matchp-abs.shtml.

[15] Gilligan, Michael and Ernest Sergenti. 2008. "Do UN Interventions Cause Peace? Using Matching to Improve Causal Inference" *Working Paper* New York University.

[16] Iacus, Stefano, Gary King, Giuseppe Porro. 2008. "Matching for Causal Inference Without Balance Checking." *Paper Presented at the 2008 PolMeth Conference, Ann Arbor, MI.*

[17] Ichino, Andrea, Fabrizia Mealli, and Tommaso Nannicini. 2008. "From Temporary Help Jobs to Permanent Employment: What Can We Learn from Matching Estimators and their Sensitivity?" *Journal of Applied Econometrics*. 23(3). pp. 305–327.

[18] Imai, Kosuke. 2008. "Sharp Bounds on the Causal Effects in Randomized Experiments with "Truncation–by–Death" *Statistics & Probability Letters*. 78(2) pp. 144–149.

[19] Keele, Luke, Corrine McConnaughy, and Ismail White. 2008. "Matching in Randomized Experiments" Paper presented at the 2008 PolMeth Conference.

[20] Langbein, Laura and Claire Felbinger. 2006. *Public Program Evaluation: A Statistical Guide.* M.E.Sharpe. Armonk, NY.

[21] Lo, Nigel. 2008. "Selecting How to Handle Selection Effects." *Working Paper* Emory University.

[22] Mitchell, Sara and Paul Hensel. 2007. "International Institutions and Compliance with Agreements" *American Journal of Political Science.* 51(4). pp.721-737.

[23] Rubin, Donald B. 2006. *Matched Sampling for Causal Effects.* Cambridge, England: Cambridge University Press.

[24] Schultz, Ken. 2001. “Looking for Audience Costs” *Journal of Conflict Resolution* 45(1). pp. 32–60.

[25] Sekhon, Jasjeet. 2007. “The Neyman-Rubin Model of Causal Inference and Estimation via Matching Methods” To appear *The Oxford Handbook of Political Methodology.* ed. by Janet Box-Steffensmeier, Henry Brady, and David Collier.

[26] Semykina, Anastasia, and Jeffrey Wooldridge. 2005. “Estimating Panel Data Models in the Presence of Endogeneity and Selection: Theory and Application.” *Working Paper* Michigan State University.

[27] Todd, Petra. 2006.“Evaluating Social Programs with Endogenous Program Placement and Selection of the Treate” To appear *Handbook of Agricultural Economics*. Vol.4.

[28] Sartori, Anne. 2003. “An Estimator for Some Binary-Outcome Selection Models Without Exclusion Restrictions.” *Political Analysis*. 11. pp. 111–138.

[29] Przeworski, Adam and James Raymond Vreeland. 2002. “A Statistical Model of Bilateral Cooperation.” *Political Analysis*. 10 (2) pp. 101–112.