

**Concerns with Endogeneity in Statistical Analysis:
Modeling the Interdependence *Between* Economic Ties and Conflict**

Richard J. Timpone
Department of Political Science
The Ohio State University

In their respective and intersecting literatures, the level of economic interdependence among nations and the likelihood of their engaging in conflict are both treated as dependent or endogenous variables that are determined within the system of interest. While economic ties and conflict are each important dependent variables, as the contributions in this volume demonstrate, there remain a number of debates regarding the relationship between the two. Does economic interdependence decrease the likelihood of conflict between states? Do political-military power relations and conflict themselves affect economic ties? While traditional approaches to testing hypotheses such as these often treat the relevant factor of interest as the dependent variable and the other as an independent variable in an equation, the fact that both factors are endogenously determined in a larger system of equations has consequences for empirical models that deserve more consideration.

How one theorizes about this causal relationship has important implications for the proper formulation of quantitative analyses. As will be discussed, the statistical concerns regarding the modeling of these forces are not isolated to theories of reciprocal causation, but may require further justification even in ‘simpler’ models because both economic ties and military conflict can be considered to be endogenous in a larger system. Including an endogenous factor as an explanatory variable in a regression model may create problems for accurately understanding relationships of interest because of the possibility of biased results. This is true for ordinary least squares, probit, logit models and their extensions.

As William of Occam stated, “Non sunt multiplicanda entia praeter necessitatem”, we should not multiply the complexity of our models beyond what is necessary (Starfield, Smith and Bleloch 1990). While this is often misinterpreted as forcing scholars to choose the simplest possible model, it actually requires scholars to develop theories and models that are both simple *and* that accurately reflect the system being examined. Some analyses may have cut too deeply with ‘Occam’s Razor’ by ignoring the fact that both economic and military relations are endogenous and it may be necessary to add additional complexity for accurate understanding.

In this chapter, I will discuss why simplifications may lead to biases in models testing theories that consider reciprocal causation as well as those that do not. Several extensions to traditional single equation approaches (i.e. ordinary least squares, probit and logit models) allow for different types of measurement of the endogenous factors. Some of these are relatively simple to implement while others are more complex and point to future directions that may prove fruitful for methodological attention. While technical matters regarding the choice of estimator will be discussed, it is important to acknowledge concerns that exist with the ‘solutions’ used to address the concerns in practice as well. The chapter will therefore follow the discussion of different statistical estimators with caveats about the practical applications of these solutions.

1. Concerns with Traditional Single Equation Approaches

1.1 Concerns That May Arise Without Reciprocal Causation

Much research that serves as the foundation for the debates addressed in this volume examines models of conflict where measures of economic interdependence are included as independent variables to test whether such economic ties have an effect on the likelihood of conflict. Ignoring the numerous types of economic links, I will simplify the discussion of the

models by dubbing these simply as ‘trade’. In this way, a statistical model may be set up along the line of:

$$\mathbf{conflict} = f_1(\mathbf{trade}, \mathbf{X}_1, \boldsymbol{\epsilon}_1) \quad \mathbf{eq. 1a}$$

In equation 1a, conflict is a function of trade, a set of exogenous variables \mathbf{X}_1 and the stochastic error term ϵ_1 . Unlike endogenous variables that are determined within the system, exogenous variables are assumed to be determined outside of it and uncorrelated with the error term.

Traditional regression models assume all independent variables are exogenous, and as we will see there are reasons to be concerned when we use endogenous explanatory variables, like trade, in equation 1a.

There are numerous debates in the literature regarding how to analyze equation 1a in practice. These include choices in operationalizing ‘conflict’ and ‘trade’, as well as the set of exogenous factors included and the functional forms of the relationships. The questions of the interdependence between trade and conflict that I focus on in this chapter all deal with the assumption of statistical models that all independent variables are uncorrelated with the error term in the estimated equation. Concerns with the appropriateness of this assumption exist *regardless* of whether or not one theorizes reciprocal causation when endogenous explanatory variables are included.

Even though we simply treated trade as an independent variable in equation 1, we know that it is a function of other factors. We can therefore set up a second equation for trade:

$$\mathbf{trade} = f_2(\mathbf{X}_2, \boldsymbol{\epsilon}_2) \quad \mathbf{eq. 1b}$$

In equation 1b, trade is a function of a set of exogenous factors \mathbf{X}_2 as well as a stochastic error term ϵ_2 . The sets of exogenous factors \mathbf{X}_1 and \mathbf{X}_2 are likely to overlap, as we know that numerous factors, such as geographic proximity, are clearly related to both trade and conflict.

In this case, equations 1a and 1b together form model 1. If a researcher is only interested in modeling conflict, it *may* be appropriate to look at equation 1a in isolation, *but this is not always the case*. Equations 1a and 1b form a triangular/hierarchical model. This simply describes how the endogenous factors, trade and conflict, are related within the larger system. At the top of the hierarchy, the systematic component of trade is only a function of exogenous factors, \mathbf{X}_2 , that are assumed uncorrelated with the errors in the model by definition. Going lower in the hierarchical system, conflict is a function of the earlier endogenous factor trade as well as the set of exogenous variables, \mathbf{X}_1 .

In this hierarchical model, the equation for trade can be estimated in isolation without any corrections for the problems discussed here, since all of the independent variables in equation 1b are exogenous and uncorrelated with the error. The key to whether or not equation 1a can be estimated in isolation revolves around the relationship between ε_1 and ε_2 , and by extension ε_1 and trade. While the variables that comprise \mathbf{X}_1 are not correlated with ε_1 , does this also hold for trade? It is possible that ε_1 and ε_2 are uncorrelated, in which case a recursive system exists. If a system of equations is recursive (hierarchical with uncorrelated errors), each equation can be estimated in isolation (using ols, logit, probit etc. as appropriate) without concern over bias caused by the inclusion of the endogenous explanatory variables.

Even if a researcher is only interested in modeling conflict, i.e. equation 1a, a correlation between the error terms leads to bias. This is likely to occur if important factors systematically related to both trade and conflict are left out of their respective equations. These would then become part of the error terms and lead to a correlation between them. Since trade is a function of ε_2 , the correlation between the errors will lead trade to also be correlated with ε_1 . An alternative way to consider this is that, because trade is related to the omitted factors in equation

1a, it will therefore be correlated with the error term that includes the factors . For this reason, the potential problem for estimating equation 1a in isolation (in this hierarchical case) is often discussed with issues of model specification and the general problem of omitted variable bias. Many standard texts include discussions of model specification, including the problems caused by omitted variables and ways to test for them (i.e. Green 2000; Gujarati 1995). While theory leads to concerns about endogeneity, these texts also discuss tests (i.e. the Hausman test and extensions) that can be applied to determine if an explanatory variable is exogenous when the relationship between continuous variables is examined.

The accuracy of estimation results therefore is a function of model specification, and the variables included for theoretic and control purposes. Estimation of equation 1a in isolation can only avoid potential bias due to correlation between trade and ε_1 if the set of exogenous variables \mathbf{X}_1 incorporate all of those factors that are directly related to both trade and conflict. This is true whether the researcher is interested in the full system or only interested in modeling conflict and ignores equation 1b entirely.

1.2 Concerns When Reciprocal Causation is Theorized or Tested

As was seen in the previous section, even if a scholar is only interested in modeling one of the areas and assumes that the direction of causality only flows in one direction, consideration of the fuller system is needed to insure that the estimation of the individual equations is unbiased. In that case, the problem could in theory be handled with adequate statistical controls in the set of exogenous variables included in the estimation. Adopting theories of reciprocal causation always requires consideration of the fuller system for accurate estimation even if a researcher is primarily interested in what is causing one of the dependent variables, trade or

conflict. In this case, the system of equations would not be hierarchical and would be represented by different functions than those discussed above. Now we would have:

$$\begin{aligned} \mathbf{conflict} &= g_1(\mathbf{trade}, \mathbf{X}_1, \mathbf{d}_1) \\ \mathbf{trade} &= g_2(\mathbf{conflict}, \mathbf{X}_2, \mathbf{d}_2) \end{aligned} \quad \mathbf{model\ 2}$$

The two equations that comprise model 2 form a system where conflict and trade are reciprocally related to each other and are functions of exogenous factors, \mathbf{X}_1 and \mathbf{X}_2 , and stochastic error terms, δ_1 and δ_2 respectively. With the hierarchical assumption, bias exists if the error terms are correlated. The problem of bias is more serious in cases of reciprocal causation. Even if the errors in model 2 were uncorrelated, δ_1 is a component of conflict that is causally related to trade, thus trade will be correlated with δ_1 . The same is true for conflict and δ_2 given its chain of causality. Even if the errors were uncorrelated across the equations in model 2, reciprocal causation leads to correlation between the endogenous factors on the right hand side of each equation with its respective error term. While such correlation was a potential problem with the hierarchical system in model 1 that could potentially be addressed with improved specification, it is endemic to non-hierarchical models that cannot be as simply corrected.

2. Addressing these Statistical Issues

There are a number of different estimators that have been developed to address the issue of endogenous variables being correlated with the errors in models. Here, I will focus on the extensions of linear regression and basic maximum likelihood models which are the most common statistical tools used to test these hypotheses. While I introduce a number of approaches briefly in this chapter, a number of texts provide fuller discussions of the problems and basic techniques to address them (i.e. Greene 2000, Gujarati 1995) Even with these

regression extensions, differences and difficulties emerge depending on the nature of the measurement of the dependent variables, i.e. whether they are continuous or categorical. All of the various approaches have certain commonalities including the need for instrumental variables that provide additional information to obtain consistent estimates of the parameters of interest. These are variables that are uncorrelated with the error terms and allow for identification of the estimators. Given the centrality of these variables, special attention will be given to caveats that affect the alternative approaches in practice.

Extending traditional ordinary least squares (OLS) regression and maximum likelihood approaches, several estimators have been developed to obtain consistent estimates for equations in broader systems like those in models 1 and 2. While full information estimation of an entire system of equations simultaneously is possible, given their greater simplicity and desirable statistical properties, limited information estimators are most common (Greene 2000, pp. 693-696 discusses full information estimation of systems of equations). The limited information estimators allow for the estimation of a single equation in a system, for instance equation 1a, and correct for problems where the endogenous variables serving as independent variables, trade in that case, are (or may be) correlated with the error term. Thus, these are equation by equation estimators that are appropriate whether you are interested in modeling all equations or only one in a broader system. These approaches may be used for any non-recursive system, i.e. if the system is not hierarchical, or if hierarchical, when the errors are correlated.

2.1 Problems Where Both Endogenous Variables are Continuous

Using the limited information approach to obtaining consistent estimates of a single equation underlies the frequently used statistical methods of instrumental variable analysis or

two stage least squares (2SLS) for handling continuous variables. In these approaches, instrumental variables, that is, variables uncorrelated with the error term, are used to provide information in place of the endogenous variables on the right hand side of the equation.

In 2SLS, the researcher first creates predicted values of right hand side endogenous terms using all of the exogenous variables in the system (the union of the sets of variables in \mathbf{X}_1 and \mathbf{X}_2 in the two models discussed earlier). This first stage equation is known as the reduced form equation and because the exogenous variables are uncorrelated with the error terms, the predictions created from them will likewise be uncorrelated with the errors. The prediction from the reduced form equation is plugged into the second stage analysis in place of the actual endogenous variable. The results of this analysis will produce consistent estimates of the model parameters (as long as quality instrumental variables are available, as discussed in section 3 of the chapter) but the standard errors will be wrong because they are based on the predicted values of the endogenous explanatory factor. The instrumental variable approach in the next paragraph is preferable when all endogenous variables are continuous in an equation. However, the manual two stage approach introduced here will serve as the foundation for extensions into problems where the nature of the measures creates additional complexity.

While 2SLS is one approach to using additional information for consistent parameter estimation, this could also be handled in one step with an instrumental variable estimator that folds the two stages into a single estimation by extending the matrix algebra used in ordinary least squares. While the parameter estimates of each approach are consistent, the advantage of this instrumental variable estimator over the manual two stage estimation is correct estimation of the standard errors whereas in the manual two stage calculation these must be corrected. Many statistical textbooks provide overviews of 2SLS/instrumental variable estimation extensions to

OLS such as Hanushek and Jackson (1977), Greene (2000), and Gujarati (1995). Also, the formulas for correcting the standard errors in the manual application of 2SLS in this case are known and can be implemented as well (see for instance Achen 1986, Gujarati 1995).

While the basic statistical concerns and the 2SLS/instrumental variable corrections may be familiar to many with advanced statistical training, the situation is rarely so easily handled in the modeling of economic interdependence and conflict as discussed thus far. The reason for this is that the desirable statistical property of consistency assumes that we are dealing with continuous variables. While some economic data may fit this assumption, many common operationalizations of conflict do not (such as dichotomous measures of engaging in militarized disputes). Even the economic measures may violate this if we consider factors like involvement in trading agreements (this is true whether we use these as our sole measures of economic interdependence or build larger systems of equations with more complex interrelations between the factors). With non-continuous variables, parameter estimates may not be consistent even with large samples and quality instrumental variables.

Unlike the 2SLS/instrumental variable corrections for endogeneity with continuous variables, extensions for the types of variables common in studies of economic interdependence and conflict are not as straightforward. We will continue with the limited information approach of estimating each equation in a broader system and correcting for the problems caused by endogenous explanatory variables. While the earlier discussions explain the conceptual solution for two continuous variables, it is worthwhile to consider separately 3 distinct cases, (1) continuous dependent variable/non-continuous endogenous explanatory variable; (2) non-continuous dependent variable/continuous endogenous explanatory variable; and (3) the case where neither of the endogenous variables on the right hand (explanatory) or left hand side

(dependent variable in the equation) is continuous. For simplicity in the following discussions and given some of the common operationalizations of variables of interest (MIDs, trade organizations etc.), I will treat the non-continuous variables as dichotomous. The core extension will be two stage probit least squares in these conditions although other approaches are appropriate in some conditions as well.

2.2 Dichotomous Dependent Variable, Continuous Endogenous Explanatory Variable

The simplest approach to dealing with any endogenous dichotomous variable (either explanatory or explained) is to extend ordinary least squares to a linear probability model. In fact, given the complexity of preferable corrections in some of the conditions, this has often been the approach endorsed and implemented. Running ordinary least squares on a dichotomous variable results in inefficient estimates by definition, since the structure of the estimation is guaranteed to create heteroscedasticity (Aldrich and Nelson 1984; Long 1997). Just as weighted least squares can be implemented to adjust for the heteroscedasticity, in this case additional corrections can be made if one uses 2SLS. Achen (1986) details the necessary corrections to extend the previous logic to a generalized two stage least squares framework. While the easiest to estimate, problems of functional form have been the key reason for moving from linear probability models to logit and probit treatments of dichotomous dependent variables in general. Moving from linear probability models to these maximum likelihood estimators leads to added complexity when addressing issues of endogeneity. While extensions of the 2SLS approach have been developed, they require manual estimation of the multiple stages and corrections to produce consistent parameter estimates and correct standard errors.

The first extension to the 2SLS/instrumental variable approach that can deal with the case where the dependent variable is dichotomous and the endogenous explanatory variable is continuous is known as two stage probit least squares (2SPLS). Again, this could apply to either of the equations explaining conflict in model 1 or 2 regardless of theories of reciprocal causation. Because the endogenous explanatory variable is continuous, the first stage is the same as traditional 2SLS models. An OLS reduced form equation for this factor is created using all exogenous variables in the system (the union of the sets of \mathbf{X}_1 and \mathbf{X}_2 in models 1 or 2). The predicted value is used in place of the original variable and probit is then run on the second stage dependent variable. An additional re-scaling step is necessary to produce consistent parameter estimates. The statistical derivation and practical application of this technique is described in detail elsewhere (Achen 1986, p 48-50; Alvarez and Glasgow 2000; Maddala 1983, p 244-245). The main drawback of 2SPLS is that the standard errors produced are biased and their correction is very difficult. In section 2.5, I discuss how the 2SPLS results may be used by addressing this limitation with bootstrapped standard errors.

The problems in calculating the standard errors provided one motivation for the Alvarez and Glasgow (2000) discussion of two stage conditional maximum likelihood (2SCML) whose technical derivation and comparison with other estimators is provided by Rivers and Vuong (1988). Rivers and Vuong (1988) developed the 2SCML estimator as an alternative to 2SPLS to allow for consistent estimation where the dependent variable is dichotomous and the independent variable is continuous (see Alvarez and Glasgow 2000 for a less technical introduction and application of this technique). While also a consistent estimator of the model parameters, Rivers and Vuong (1988) demonstrate that it has advantages in efficiency as well as calculation although Alvarez and Glasgow's (2000) simulations show that the 2SPLS approach is slightly

superior in reducing the bias in the coefficient estimates caused by the endogeneity. In spite of this, the accuracy in estimates of the standard errors lead Alvarez and Glasgow (2000) to support use of the 2SCML estimator.

The 2SCML approach begins like the 2SLS and 2SPLS approaches with the estimation using the endogenous continuous explanatory variable as a dependent variable in a first stage OLS reduced form regression model using all exogenous variables in the system. In the second stage, the residuals of the first stage model are included in the probit model as an additional variable along with the original endogenous regressor. In addition to producing consistent estimates and accurate standard errors, this also allows for the statistical testing of endogeneity by comparing the log of the likelihood functions calculated with and without the residuals included in the second stage (again, Alvarez and Glasgow 2000 provide a fuller overview of this estimator).

2.3 Continuous Dependent Variable, Dichotomous Endogenous Explanatory Variable

The correction for endogenous explanatory variables (regardless of whether or not one treats the relationship as reciprocal) is also more complicated than the traditional instrumental variable/2SLS approaches if the variable on the right hand side of the equation is dichotomous. Again, the simplest approaches would treat the dichotomous variable as a linear probability model and extend the corrections in traditional 2SLS with weighted least squares. As discussed earlier, these further extensions do not address the problems of functional form that come from the use of the linear and additive specification of OLS with a dichotomous dependent variable (in the first stage in this case).

The 2SPLS approach for obtaining consistent parameter estimates under these conditions is similar to the dichotomous dependent variable/continuous endogenous explanatory variable condition although the reduced form estimation and re-scaling are different (Maddala 1983, pp 244-245). In the first stage, the reduced form model is estimated for the dichotomous endogenous variable. Instead of OLS, a probit model is conducted using all exogenous variables in the system (again the union of the variables that comprise \mathbf{X}_1 and \mathbf{X}_2).

There are different ways to derive the probit specification, and one is to consider the dichotomous variable to be generated from an unobserved continuous index function, where a value of zero is observed if the value of this unobserved variable is below a threshold and 1 if it is above. Predicted values from probit models, as in the first stage here, produce 'z' values on this unobserved continuum. While normally we transform these raw values into probabilities for substantive interpretation, the raw untransformed continuous z value predictions would be used in place of the endogenous variable in the second stage. The rescaling is different than the dichotomous dependent variable/continuous endogenous explanatory variable situation as the parameter estimates on the exogenous explanatory variables in the second stage do not need rescaling (unlike the dichotomous dependent variable case), although the parameter on the purged endogenous indicator would require a transformation for substantive interpretation.

As in the earlier discussion of the 2SPLS approach, the standard errors in the estimates in the second stage equation of interest will be wrong and difficult to correct. Developing 2SPLS estimators that incorporate bootstrapped standard errors would be even more useful in this case since there is not another simple approach to correcting the results under these conditions (without moving to the assumptions of the linear probability model). Issues that methodologists will need to consider when creating models of this sort will be discussed in section 2.5.

2.4 Problems Where Both Endogenous Variables are Dichotomous

Finally, a researcher may be dealing with cases where both the economic and conflict variables are measured dichotomously. This would be the case, for instance, if involvement in trade agreements and the emergence of militarized disputes were examined. The estimation of the 2SPLS can be extended to the case where both endogenous variables are dichotomous as well (Maddala 1983, p 245-247). Again, these solutions employ a limited information equation by equation approach and do not require a theory of reciprocal causation. The first stage of estimation in this case follows the previous example where the endogenous explanatory variable was dichotomous. The reduced form equation is estimated in this case using probit analysis and the untransformed predicted z value from this stage is used in place of the variable of concern in the second stage. The dependent variable in the second stage is also dichotomous and this is also estimated using probit analysis.

While these first steps are straightforward, the rescaling is more complex for both the exogeneous and endogenous independent variables and distinct adjustments are required. Even if this two stage approach is adopted and the consistent estimates for the second stage are obtained in a cleanly interpretable manner, the standard errors remain wrong and their correction is extremely difficult in practice.

2.5 Extending 2SPLS with Bootstrapped Standard Errors

Two stage least squares/instrumental variable estimation is appropriate when the dependent and explanatory endogenous variables are both continuous, and 2SCML can address the additional complexity caused by having a dichotomous dependent variable and continuous explanatory endogenous variables. Throughout the different conditions, 2SPLS is discussed as

the logical extension of the 2SLS approach. While parameters require different re-scaling transformations for substantive interpretation, the problems in estimating the standard errors have been discussed as the biggest limitation for this estimator. One solution is to use the consistent 2SPLS parameter estimates along with bootstrapped standard errors.

Bootstrapping is a statistical technique where the sampling distributions for the parameter estimates of interest are simulated through an iterative process (see Mooney and Duval 1993, Mooney 1996 for useful introductions to the topic). This approach allows for the creation of confidence intervals for statistics where the sampling distributions are unknown or, as in the case of 2SPLS, very difficult to estimate. Bootstrapped estimates of standard errors are computationally demanding to produce but do not make the restrictive assumptions common in analytic estimates used for statistical inferencing.

Bootstrapping the standard errors in 2SPLS will require integrating a number of statistical issues. First, the uncertainty of the first stage predictions will need to be incorporated into the second stage bootstrapped results. A possibility may be to incorporate the logic of variable imputation strategies for the first stage along with the bootstrapping of the second to estimate the standard errors (King, Honaker, Joseph and Scheve 2001). While multiple imputation strategies may prove useful for addressing the uncertainty in the first stage predictions, previous work that has explored the potential of using different data sources for the first and second stages may also provide guidance for developing accurate bootstrapped standard errors for the second stage (see for instance Franklin 1990; Gelman, King and Liu 1998). The Gelman, King and Liu (1998) work on imputation in a hierarchical framework may be particularly useful in this case. In addition to the uncertainty, incorporating the appropriate scaling corrections and dealing with the nature of the measures will need to be handled as well. The statistical foundations that exist in

these areas along with the dramatic increase in computing power provides more promise that general solutions to these issues can be developed that will aid substantive investigations.

2.6 *Covariance Structure Models as an Alternative Approach*

Thus far, this chapter has focused on extensions of traditional regression models (OLS, probit etc.) to handle the concerns raised when models contain multiple endogenous factors. It is worth briefly noting that a number of other approaches exist as well. Researchers building time series models will likely incorporate techniques such as vector autoregression (VAR) and examine the nature of interdependence using granger causality tests. While VAR models are relatively simple to estimate, they are sometimes criticized as being less theoretic as they focus on forecasting and treat all variables as endogenous (simple introductions to these techniques are found in many texts such as Gujarati 1995). I will leave the debates over these techniques for elsewhere, but highlight the issue for scholars who are using time series data in their research.

Even without moving to different types of data, such as time series, other approaches can be used for investigation. One such approach uses structural equation models (SEMs) estimated by examining covariance structures as performed by programs like Lisrel, EQS and Amos. These models often are comprised of two parts, a measurement component that deals with the construction of latent variables, and a structural model that examines the relationships among the latent variables and other indicators (Hoyle 1995). These models are likely to be most familiar to political scientists with some interest in political psychology where these techniques are more common.

While these SEMs are designed to explicitly examine systems of equations with potentially complex inter-relationships, like simple OLS models they are linear and additive and

the problems found in 2SLS models based on functional form may apply here as well.

Extensions to traditional SEMs incorporate the logic used above in describing one manner that the probit model could be derived. In the probit model, one could consider the dichotomous indicator a blunt measure of the unobserved continuous index function. In SEM extensions, this logic has been adopted for the specification of the latent variables which have been formulated to behave as the index functions that underlie models like probit. Xie (1989) details these extensions and programs like Mplus (Muthén and Muthén 1998) can handle such operationalizations directly. Thus, researchers trained in structural equation model estimation using programs like Lisrel can consider this alternative approach to address the concerns of analyzing a system of equations without the traditional limiting assumptions as well.

3. Caveats Regarding the Implementation of Corrections

The critical issue for addressing the concerns of endogenous explanatory variables that are (or may be) correlated with the error term is that in order to correct the potential problems, adequate instrumental variables are necessary. In order to obtain consistent estimates of model parameters in these cases, we thus far assumed that enough information that fulfilled fairly strict assumptions existed. This information is provided by instrumental variables that are uncorrelated with the error term. To aid estimation, instrumental variables are needed that are directly related to the endogenous explanatory variable but not the dependent variable. It is the leverage from these variables that allows for the consistent estimation in the instrumental variable, two stage and even the structural equation modeling approaches. If such information does not exist, unique solutions can not be obtained. If variables that do not fulfill the necessary assumptions are used in their place, the desirable qualities of our estimators will not hold.

Some scholars feel that, given the complexity of inter-relationships in political science, true instrumental variables that are uncorrelated with the error terms may be so rare that estimators such as those discussed throughout this chapter may have limited utility in practice. Although this chapter has focused on the technical issues involved in creating models with desirable properties under different conditions, even after estimators are developed and widely available, researchers may still find their research problem intractable because of the difficulty in adequately identifying their model. While the caveats and difficulties involved in actually correcting the problems when explanatory variables are correlated with the errors in models must be acknowledged, simply ignoring the biases that these problems produce is not a viable solution.

Bartels (1991) has demonstrated the consequences of failing to satisfy statistical assumptions. If the variables used to identify the estimation are only 'quasi-instrumental' and approximately uncorrelated with the error, the results of the second stage/instrumental variables estimation will not be consistent. While one may use a two stage estimation and give the appearance of addressing the problems, if the instruments used for identification are correlated with the error, not only is the appearance deceptive, but the 'corrected' results could be worse than a model that ignores the endogeneity in the first place. Bartels' conclusions reinforce the point that there are serious consequences to violating the assumptions of a statistical estimator and care is needed for identifying equations.

The issue of identification requires that the indicators that are used as instruments are not only uncorrelated with the error term, but also that they are meaningfully related to the endogenous variable for which they are serving as a proxy. Since the instruments are, in essence, standing in for the actual variables that are correlated with the error, variables that are

very weakly related to the endogenous explanatory variable will not aid in estimation. In fact, Bound, Jaeger and Baker (1995) show that even if the instruments are truly exogenous and uncorrelated with the error, as the R^2 in the first stage estimation goes to zero, the bias in the ‘corrected’ results approaches that of an uncorrected OLS model. Thus weak instruments can lead to inconsistencies as well. The fit between the instrumental variables that are providing the additional information and the original endogenous variable is critical. Bound, Jaeger and Baker (1995) further show that the weaker the relationships between the instruments and the variable they are standing in for, the larger the consequences for even small amounts of ‘quasi-instrumentality’ and correlation with the error.

While the need for variables that are related to the endogenous variable they are serving as instruments for makes obvious intuitive sense, the logical question is ‘how strong does this relationship need to be?’ Bollen (1996) provides a rule of thumb that a first stage R^2 below .1 will lead to poor 2SLS results in most cases. Bartels (1991) explains that even a high R^2 in the first stage reduced form equation is not adequate to insure that the instruments are up to the task they are fulfilling though.

The key to understanding the problems associated with the fit of first stage estimates can be seen by emphasizing again the role that the instruments play in the reduced form equation. Recall that the proxies created from the first stage for the endogenous indicator are functions of the sets of exogenous variables in all of the equations in the system. In the models used to illustrate the problem earlier in the chapter, a first stage reduced form equation of trade would include the union of variables in the sets of \mathbf{X}_1 and \mathbf{X}_2 (those directly causally related to conflict and trade respectively). However, it is not enough to have a reduced form equation that has strong predictive ability. What is critical is that those exogenous factors in the first stage that are

not also in the second stage of the estimation predict well. In the example, this means that one needs to be concerned with how well the variables that comprise \mathbf{X}_2 , that are not also in the set of \mathbf{X}_1 , predict trade. The problem of focusing on the overall model fit of the reduced form equation can be seen clearly in the extreme case. If the sets of variables in \mathbf{X}_1 and \mathbf{X}_2 are the same, the predicted value for trade would be a perfect linear function of the exogenous variables in the second stage and the model of conflict could not be estimated. In this extreme case, not enough information would be available to obtain unique estimates. What is needed are strong predictors of trade in \mathbf{X}_2 that are not also in \mathbf{X}_1 . Bartels (1991) shows that the partial R^2 based on these items is more important to determine if there is enough information for stable results. Standard texts spend a substantial time on the question of identification of the equations which deals explicitly with the question of how much information is available to obtain unique results.

While this discussion has focused on creating an instrument for the endogenous variable of trade in an equation determining conflict, the same issues arise in the second model in creating a quality measure for the endogenous conflict variable in the second stage equation for trade. The caveats illustrate the difficulty in implementing some of the techniques described earlier in the chapter in models of economic interdependence and conflict in general. What is needed is a set of variables that can be considered truly uncorrelated with the error in an equation and is strongly related only to the first stage dependent variable. Variables that are related to both endogenous factors do not provide additional information to aid in the estimation and quasi-exogenous and weak indicators are also problematic for the estimation. In addition to other exogenous variables, lagged endogenous variables are often considered useful instruments in these first stage equations. While common, one must still be cautious that these are truly uncorrelated with the error terms as well. This section has demonstrated the difficulty in finding

quality variables to identify the system. However, it is still worth attempting to overcome the obstacles rather than to ignore the potential bias that results from running standard uncorrected single equation estimators in models 1 and 2 that disregard the potential correlation between the endogenous explanatory variables and the error term in each equation.

4. Conclusion

In models examining the relationship between economic interdependence and conflict, the fact that measures of both types of factors can be treated as endogenous has troubling negative consequences for the traditional single equation models used to test our hypotheses. As shown in the first section, whether one theorizes the relationship to be one of reciprocal causation or not, the possibility exists that an endogenous factor included as an explanatory variable will be correlated with the error term and so lead to biased estimates of the substantive parameters of interest.

If continuous measures of both the economic and conflict factors are available, traditional instrumental variable/2SLS approaches can address some of the statistical problems involved. Given that common measures often do not fit this measurement scheme, the problem becomes more complex. Approaches such as 2SPLS and 2SCML are possible solutions, depending on the nature of measures used for analysis. Extensions to 2SPLS that bootstrap the standard errors may prove to be very useful in the future. Also, while the discussion in this chapter has focused on the relationships among continuous and dichotomous variables, obtaining general solutions for ordinal and nominal variables will be fruitful for numerous research problems as well.

The technical issues of how to obtain consistent estimators and accurate standard errors are often the focus of the discussions dealing with systems of equations. The problems of

identification for the actual application of these techniques also deserve serious consideration for political scientists. The need for sets of variables that are uncorrelated with the error terms and directly related to only one of the endogenous variables in a meaningful way, may in the long run continue to be a serious hurdle even after the other technical issues have been addressed. Following up on the caveats introduced in this chapter will be useful for those interested in actually implementing these techniques in the future.

While technical and practical difficulties exist in addressing the concerns raised by endogenous explanatory variables, ignoring the problems caused is inappropriate. Given interest in understanding how economic interdependence and conflict are linked, as well as how other factors are related to each of these, scholars need to work to minimize the biases that result when basic statistical assumptions are violated. Thus, following the logic of William of Occam, the solution may be to produce more complex models that accurately reflect the true richness of the international system. This chapter points to some issues and possible solutions to move research further down this path.

Bibliography

- Achen, Christopher H. 1986. *The Statistical Analysis of Quasi-Experiments*. University of California Press.
- Aldrich, John H., and Forrest D. Nelson. 1984. *Linear Probability, Logit, and Probit Models* (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-045), Sage.
- Alvarez, R. Michael, and Garrett Glasgow. 2000. "Two-Stage Estimation of Non-Recursive Choice Models." *Political Analysis* 8(Spring): 147-165.
- Bartels, Larry M. 1991. "Instrumental and 'Quasi-Instrumental' Variables." *American Journal of Political Science* 35: 777-800.
- Bollen, Kenneth A. 1996. "An Alternative Two Stage Least Squares (2SLS) Estimator For Latent Variable Equations." *Psychometrika* 61(March): 109-121.
- Bound, John, David A. Jaeger, and Regina M. Baker. 1995. "Problems with Instrumental Variables Estimation when the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak." *Journal of the American Statistical Association* 90(June): 443-450.
- Franklin, Charles H. 1990. "Estimation Across Data Sets: Two-Stage Auxiliary Instrumental Variables Estimations (2SAIV)." *Political Analysis – Volume 1, 1989*. University of Michigan Press.
- Gelman, Andrew, Gary King, and Chuanhai Liu. 1998. "Not Asked and Not Answered: Multiple Imputation for Multiple Surveys." *Journal of the American Statistical Association* 93(September): 846- 857.
- Greene, William H. 2000. *Econometric Analysis*, Fourth edition. Prentice Hall.
- Gujarati, Damodar. 1995. *Basic Econometrics*, Third edition. McGraw-Hill.
- Hanushek, Eric A. and John E. Jackson. 1977. *Statistical Methods for Social Scientists*. Academic Press.
- Hoyle, Rick H, editor. 1995. *Structural Equation Modeling: Concepts, Issues, and Applications*. Sage.
- King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review* 95(March): 49-69.

- Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Sage.
- Maddala, G.S. 1983. *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge University Press.
- Mooney, Christopher Z. 1996. "Bootstrap Statistical Inference: Examples and Evaluations for Political Science." *American Journal of Political Science* 40(May): 570-602.
- Mooney, Christopher Z., and Robert D. Duval. 1993. *Bootstrapping: A Nonparametric Approach to Statistical Inference* (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-095), Sage.
- Muthén, Linda K. and Bengt O. Muthén. 1998. *Mplus: The Comprehensive Modeling Program for Applied Researchers*. Muthén and Muthén.
- Rivers, Douglas, and Quang H. Vuong. 1988. "Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models." *Journal of Econometrics* 39: 347-366.
- Starfield, Anthony M., Karl A. Smith, and Andrew L. Bleloch. 1990. *How to Model It: Problem Solving for the Computer Age*. McGraw-Hill.
- Xie, Yu. 1989. "Structural Equation Models for Ordinal Variables." *Sociological Methods and Research* 17(May): 325-352.