2002

An Analysis of the Effectiveness of Supplemental Instruction: The Problem of Selection Bias and Limited Dependent Variables

Tyler J. Bowles  
_Utah State University_  

Jason Jones  
_Utah State University_

Follow this and additional works at: https://digitalcommons.usu.edu/eri

**Recommended Citation**  
https://digitalcommons.usu.edu/eri/247

This Article is brought to you for free and open access by the Economics and Finance at DigitalCommons@USU. It has been accepted for inclusion in Economic Research Institute Study Papers by an authorized administrator of DigitalCommons@USU. For more information, please contact digitalcommons@usu.edu.
AN ANALYSIS OF THE EFFECTIVENESS OF SUPPLEMENTAL INSTRUCTION: THE PROBLEM OF SELECTION BIAS AND LIMITED DEPENDENT VARIABLES

by

TYLER J. BOWLES
JASON JONES

Department of Economics
Utah State University
3530 Old Main Hill
Logan, UT 84322-3530

June 2002
AN ANALYSIS OF THE EFFECTIVENESS OF SUPPLEMENTAL INSTRUCTION: THE PROBLEM OF SELECTION BIAS AND LIMITED DEPENDENT VARIABLES

Tyler J. Bowles, Associate Professor
Jason Jones, Research Assistant

Department of Economics
Utah State University
3530 Old Main Hill
Logan, UT 84322-3530

The analyses and views reported in this paper are those of the author(s). They are not necessarily endorsed by the Department of Economics or by Utah State University.

Utah State University is committed to the policy that all persons shall have equal access to its programs and employment without regard to race, color, creed, religion, national origin, sex, age, marital status, disability, public assistance status, veteran status, or sexual orientation.

Information on other titles in this series may be obtained from: Department of Economics, 3530 University Boulevard, Utah State University, Logan, Utah 84322-3530.

Copyright © 2002 by Tyler J. Bowles and Jason Jones. All rights reserved. Readers may make verbatim copies of this document for noncommercial purposes by any means, provided that this copyright notice appears on all such copies.
Single equation regression models have been used rather extensively to test the effectiveness of Supplemental Instruction (SI). This approach, however, fails to account for the possibility that SI attendance and the outcome of SI attendance are jointly determined endogenous variables. Moreover, the standard approach fails to account for the fact that these two endogenous variables are categorical. This paper presents and applies a simultaneous equation, limited dependent variable model of SI effectiveness. Our analysis suggests that results from applying this type of model may differ markedly from the traditional statistical models applied in SI research. Specifically, our results suggests that students with below average academic ability are more likely to attend SI and that common measures of student ability included in single equation models fail to adequately control for this characteristic. Therefore, single equation OLS models may underestimate SI effectiveness.
AN ANALYSIS OF THE EFFECTIVENESS OF SUPPLEMENTAL INSTRUCTION: THE PROBLEM OF SELECTION BIAS AND LIMITED DEPENDENT VARIABLES

Introduction

A number of researchers have attempted to evaluate the effectiveness of Supplemental Instruction (SI), a student academic assistance program used at many colleges and universities (see Blanc, Debuhr, and Martin, 1983; Congos and Schoeps, 1993; Kochenour et al., 1997; and Schwartz, 1992). The statistical model used in this research has been various versions of the following:

\[ \text{Outcome}_i = \alpha_1 + \alpha_2 \text{SI Attendance}_i + \sum_{i=1}^{N} \beta_i X_i + \epsilon_i \]

where \( \text{Outcome} \) = a measure of student performance that is potentially influenced by SI attendance (e.g., class score); \( \text{SI Attendance} \) = a measure of SI attendance; \( X_i \) = explanatory variables designed to control for self-selection bias (e.g., ACT or SAT score); and \( \epsilon \) = the random error term.

In the education literature on program effectiveness, however, single equation models and the technique used to apply those models, ordinary least-squares regression analysis (OLS), have been criticized (e.g., Weiler and Pierro, 1988; and Willis and Rosen, 1979). The reasons are that single equation models fail to adequately account for self-selection bias and that OLS is inappropriate since program attendance and/or the measure of outcome are not continuous but

---

1 See Congos and Schoeps (1999) for a concise description of the standard "Inferential Statistical" model that has been used to test the effectiveness of SI.
rather categorical variables. Below, we first discuss these two problems in the context of SI and then present an alternative model and estimation technique.

**The Problem of Selectivity Bias**

Students choose to attend SI. If students who attend SI receive, on average, a higher grade in the course than the average of students who do not attend, it may be because better students choose to attend. Researchers, therefore, have attempted to control for this self selection bias by including in equation (1) variables that measure student characteristics that may affect course grades other than SI attendance (e.g., SAT and ACT scores, high school GPA, etc.) (see Congos and Schoeps, 1999). The problem with this approach, generally, is that there are unmeasurable or unobserved variables that affect both SI attendance and outcome that are left out of equation (1). This results in SI attendance being correlated with the error term; therefore, OLS will yield a biased estimate of the effect of SI attendance on class grade.

For example, assume that there is an unmeasurable variable, Z (e.g., intrinsic motivation), that affects both outcome and SI attendance and is left out of equation (1). By definition, the error term would account for the effect of Z on outcome. In this situation, any observed correlation between outcome and SI attendance may be partially explained by the missing variable Z.

A number of researchers have commented on the self selection problem in so-called treatment studies and introduced appropriate models. The seminal article in the education literature was Willis and Rosen (1979). They were interested in the effect of educational attainment on lifetime income. They applied a simultaneous equations model (hereafter referred

---

2 Weiler and Pierro (1988) note the extreme difficulty of including all variables that may affect outcome and status in single equation models designed to test program effectiveness.
to as the Treatment Model) where the two jointly determined (i.e., endogenous) variables were lifetime income and educational attainment. They estimated the coefficients of the model using the now famous Heckman technique (Heckman, 1979). More recently, Murnane, Newstead, and Olsen (1985) applied a similar model and estimation technique to test whether student test scores were affected by private versus public school attendance. Finally, Weiler and Pierro (1988) apply the Treatment Model and the Heckman technique to estimate the parameters of a two-equation system where the endogenous variables are measures of whether students continue with their education and their initial enrollment status (i.e., part or full time).

All of these models use a system of two equations to control for selectivity bias. In the first equation the dependant variable is a binary variable (i.e., 1 or 0) representing whether or not the specific individual was in the program (i.e., whether the individual was “treated”). The dependent variable in the second equation is outcome. Of course, one of the explanatory variables in the second equation is the binary dependent variable of the first equation. This Treatment Model is written as follows:

\[
A_i = \alpha_i + \sum_{i=1}^{N} \beta_i W_i + \mu_i
\]

\[
Y_i = \theta_i + \theta_2 A_i + \sum_{i=1}^{N} \delta_i X_i + \epsilon_i
\]

where \(A_i\) = the binary variable that measures program participation; \(Y_i\) = outcome measure; \(W_i\) = exogenous variables that determine \(A_i\); and \(X_i\) = exogenous variables that determine \(Y_i\). The fact that one or both of the dependent variables in the above system is binary or otherwise

---

3 This type of model has become the standard approach for testing program effectiveness (see Greene, 2000, section 20.4.4).
limited means that simple application of OLS in the two-stage least squares technique is inappropriate. This problem is addressed below.

**The Problem of Limited Dependant Variables**

An additional problem of applying OLS to a single equation model in which the dependent variable is some measure of outcome is that often this variable is discrete. For example, if the outcome variable is class GPA, the only values that may be observed are 0, 1, 2, 3, and 4. The typical regression (i.e., OLS) is inappropriate in this context. Furthermore, in the context of the Treatment Model (i.e., equations (2) and (3)) $A_i$ is binary. If the dependent variable is binary, it is easily shown that the error term is heteroscedastic and inferences made based on an OLS regression are suspect. More importantly, there is nothing inherent in the OLS algorithm that restricts the expected value of $A_i$, $E(A_i)$, to be in the 0 to 1 interval. Given the limitations of OLS in the context of discrete dependent variables, this "... model is becoming less frequently used except as a basis for comparison to some other more appropriate models" (Greene, 2000, p. 813).

The appropriate model to apply when the dependent variable is limited depends on the specific nature of this variable. If the dependent variable is a simple dummy variable indicating the presence (or lack thereof) of a characteristic (e.g., SI attendance), the standard model is referred to as the probit model. In the context of the Treatment Model, $A_i$ is indeed a dummy variable. The probit model imposes the following structure:

---

4 See Greene (2000), chapter 19 for an in-depth discussion of discrete dependent variable models. Also see Maddala (1983).

\( \text{Prob}(A_i = 1) = F\left( \alpha + \sum_{i=1}^{N} \beta_i W_i \right) / \sigma \)

where \( \text{Prob}(A_i = 1) \) is the probability of individual \( i \) being in the treatment group; \( F = \) the standard normal cumulative density function; and \( \sigma = \) the standard deviation of the standard normal distribution.

This structure limits the expected value of \( A_i \) to be in the interval 0 - 1. Moreover, \( A_i \) is a nonlinear function of the \( X \)'s and, hence, equation (4) cannot be estimated using OLS. Maximum likelihood estimation (MLE) is the technique normally applied to estimate probit models (Greene, 2000, Ch. 19). This is the approach taken by Weiler and Pierro (1988) to estimate an equation where the dependent variable was whether or not a student was enrolled in college at some point in time.

Often, the outcome variable in a Treatment Model also is discrete but not necessarily binary. In testing the effect of SI attendance, a possible choice would be class grade point average on the four-point scale (i.e., 0, 1, 2, 3, and 4). This type of dependent variable is referred to as count data and is estimated using the Poisson regression model (see Greene, 2000, Section 19.9). In terms of equation (3) of the Treatment Model that we have been working with, the Poisson regression model imposes the following structure:

\( \ln \lambda_i = \sum_{i=1}^{N} \delta_i X_i \)

where \( \lambda_i \) must satisfy the following equation: \( \ln \lambda_i = \sum_{i=1}^{N} \delta_i X_i \). Again, MLE generally is applied to estimate the parameters of this nonlinear model.

**A Model of SI Effectiveness**

Based on the previous analysis, a model for testing the effectiveness of SI must allow
(1) for the fact that SI attendance and outcome are jointly determined variables, and (2) for the fact that one or both of these variables is discrete. We propose the following model:

\[ SI \text{ Attendance}_i = \alpha_i + \sum_{i=1}^{N} \beta_i W_i + \mu_i \]

\[ Class \text{ Grade}_i = \theta_1 + \theta_2 SI \text{ Attendance}_i + \sum_{i=1}^{N} \delta_i X_i + \epsilon_i \]

where the \( W \)'s and \( X \)'s are sets of explanatory variables for \( SI \text{ Attendance} \) and \( Class \text{ Grade} \), respectively.\(^6\) Table 1 provides a list of the variables included in each set. The model is estimated using a variation of the Heckman technique.\(^7\) The probit structure was imposed on equation (6), and this equation was estimated using MLE. Results from this probit estimation include a parameter, lambda, that captures the correlation between SI attendance and the error term \( \epsilon \) (i.e., the effects of the missing variables in (7)).\(^8\) This parameter, along with the Poisson structure, is imposed on equation (7), and this equation is estimated using MLE to obtain unbiased estimates of \( \theta_2 \), the effect of SI attendance on class grade.\(^9\)

For comparison, we present two sets of results based on two different models. In model 1, the OLS Model, the sample selection and limited dependent variable problems are ignored and equation (7) is taken as a stand-alone model and estimated using OLS. Model 2,\(^6\)

\(^6\) Since the proposed estimation techniques are nonlinear, the standard exclusionary restrictions for identification are not necessary. That is, the set of variables in \( X \) can be the same as the set in \( W \). However, Ashenfelter, Levine, and Zimmerman (2003) recommend that \( W \) include variables not present in \( X \).

\(^7\) See Ashenfelter, Levine, and Zimmerman (2003, Section 16.4.1) for an overview of the general Heckman technique. See Greene (1998, Section 28.2.2) for a more detailed explanation of the specific Heckman technique that we have applied.

\[^8\] Specifically, the probit estimation provides \( \hat{\lambda} = \left( \frac{\sum_{i=1}^{N} \hat{\beta}_i W_i}{\sum_{i=1}^{N} \hat{\beta}_i W_i} \right) \) where \( f \) and \( F \) are the normal density function and normal cumulative density function, respectively.

\(^9\) This technique is in the spirit of the two-stage least squares technique commonly applied to estimate the parameters of systems of equations (i.e., models with two or more endogenous variables).
referred to as the Treatment Model, considers (6) and (7) to be a simultaneous system with limited dependent variables. The parameters of this system of equations are estimated in a two-step process: a probit model applied to equation (6) and a Poisson regression then applied to equation (7). Table 2 presents the results.

Results from the OLS Model indicate that SI attendance improves a student’s grade, ceteris paribus. Indeed, the OLS Model indicates that SI attendance improves a student’s grade by approximately 0.50 grade points on a 4.0 scale. A priori, one might expect that this OLS coefficient overestimates the effect of SI attendance based on the argument that there are missing variables in equation (7) that are positively correlated with both class grade and SI attendance. The Treatment Model results, however, suggest a different conclusion.

For the dependent variable class grade, two sets of coefficients are reported in Table 2 for the Treatment Model. This is because the Poisson model is nonlinear; therefore, the coefficient estimates from this model are not equal to the expected change in the dependent variable for a given unit change in the explanatory variable as is the case in the OLS Model. In a nonlinear model, the marginal effect of a unit change in an explanatory variable depends upon the specific values of the explanatory variables. For comparison to the OLS Model, marginal effects of changes in the explanatory variables of the Treatment Model have been calculated based on the mean values of the explanatory variables.

The estimated marginal effect of SI attendance on course grade in the Treatment Model is substantially different than in the OLS Model: 1.352 versus 0.492 in the OLS Model. Although the coefficient on SI attendance is not statistically significant, a comparison of these two models suggests that the OLS model underestimates the effect of SI attendance. A possible reason for
this can be seen in the last row of Table 2. A negative coefficient on lambda\textsuperscript{10} means SI attendance and ε in equation (7) are negatively correlated. Equivalently, this means that there are unmeasured variable(s) in equation (7) that affect class grade in the opposite direction than they affect SI attendance. The OLS Model fails to account for this possibility and, therefore, underestimates the affect of SI attendance on class grade.

A plausible story that fits with the above empirical results is that inherently less able students are more likely to attend SI and that the control variables (e.g., ACT scores and GPAs) included in equation (7) do not adequately measure student ability. This unmeasurable inherent ability is positively correlated with SI attendance and negatively correlated with class score.\textsuperscript{11}

Our results are consistent with other studies that compare Treatment Effect Models to single equation OLS models in testing the effect of some program or treatment where sample selection is a problem. For example, in testing the effect of initial status (i.e., full- vs. part-time student) on educational persistence, Weiler and Pierro (1988) found that the sign of the coefficient on initial status changed when moving from a single equation model to system of equations model similar to the model used in the instant research. Indeed, in discussing applications of the Treatment Effects Model, Greene (2000, p. 934) notes that such applications have “called into question the interpretation of a number of received studies” that have not controlled for sample selection within a simultaneous equation, limited dependent variable model.

\textsuperscript{10} Lambda is the parameter from the probit regression that becomes an explanatory variable in the Poisson regression (see footnote 10). It is analogous to a set of instrumental variables applied in two-stage least squares.

\textsuperscript{11} It follows that although researchers may not be able to adequately measure and control for academic abilities, students are aware of their abilities and the less able attend SI.
Summary and Conclusions

Colleges and universities with SI are interested in testing the effectiveness of this program. Simple descriptive statistics that compare the class grades of participants to nonparticipants are of little value since academic abilities of the two separate groups may be different. Consequently, single equation regression models that include measures of student ability and are estimated using OLS have been used by many researchers to control for self-resolution bias. This paper has shown that the single equation OLS model is often inadequate and has discussed and applied a simultaneous, limited dependent variable model.

Our results suggest that students with below average academic ability are more likely to attend SI and that common measures of student ability included in single equation models fail to adequately control for this characteristic. Therefore, single equation OLS models may underestimate SI effectiveness. We recognize that our results may be unique to our sample, however, our main point is that more sophisticated and appropriate models than single equation OLS exist for testing the effect of program participation were sample selection bias and limited dependent variables are present. Moreover, our analysis suggests that results from applying a simultaneous equations-limited dependent variable approach may differ markedly from the traditional statistical models applied in SI research.
References


Table 1. List of Variables Included in the Two Sets of Explanatory Variables $W$ and $X$

**Variables in $W$**

24 dummy variables indicating whether a student was in a particular SI class (e.g., Economics, Biology, College Algebra)

HSGPA = student’s high school GPA

SEXDUM = a dummy variable for sex, male = 1

COLLEGE = the natural log of the student’s college GPA weighted by the number of college credits the student had earned

**Variables in $X$**

24 dummy variables (see above)

ACT = student’s composite score on the ACT exam

COLLEGE = (see above)
<table>
<thead>
<tr>
<th>Variable</th>
<th>OLS Model</th>
<th>Treatment Model</th>
<th>Class Grade (Marginal Effects)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Class Grade</td>
<td>SI Attendance</td>
<td>Class Grade</td>
</tr>
<tr>
<td>n = 3,645</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>-0.702*</td>
<td>0.234*</td>
<td>-0.860*</td>
</tr>
<tr>
<td></td>
<td>(0.147)</td>
<td>(0.066)</td>
<td>(0.246)</td>
</tr>
<tr>
<td>College</td>
<td>0.479*</td>
<td>0.010</td>
<td>0.220*</td>
</tr>
<tr>
<td></td>
<td>(0.019)</td>
<td>(0.008)</td>
<td>(0.022)</td>
</tr>
<tr>
<td>ACT</td>
<td>0.101*</td>
<td>---</td>
<td>0.365*</td>
</tr>
<tr>
<td></td>
<td>(0.004)</td>
<td></td>
<td>(0.004)</td>
</tr>
<tr>
<td>SI Attendance#</td>
<td>0.492*</td>
<td>---</td>
<td>0.889</td>
</tr>
<tr>
<td></td>
<td>(0.040)</td>
<td></td>
<td>(0.592)</td>
</tr>
<tr>
<td>HSGPA</td>
<td>---</td>
<td>0.088*</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.015)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEXDUM</td>
<td>---</td>
<td>-0.035*</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lambda</td>
<td>---</td>
<td>---</td>
<td>-0.407</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Standard errors are in parentheses.

*Estimate p ≤ 0.10.

#SI Attendance = 1 if student attended SI three or more times; 0 otherwise.