TECHNICAL WORKING PAPER SERIES

SELECTION BIAS ADJUSTMENT IN TREATMENT-EFFECT MODELS AS A METHOD OF AGGREGATION

Robert A. Moffitt

Technical Working Paper No. 187

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 1995

This is a revised version of a paper presented at the August 1995 Meetings of the American Statistical Association in Orlando, Florida. This paper is part of NBER's research program in Public Economics. Any opinions expressed are those of the author and not those of the National Bureau of Economic Research.

NBER Technical Working Paper #187
October 1995

SELECTION BIAS ADJUSTMENT IN TREATMENT-EFFECT MODELS AS A METHOD OF AGGREGATION

ABSTRACT

The aim of this note is to interpret estimation of the conventional treatment-effect selection-bias model in econometrics as a method of aggregation and to draw the implications of this interpretation. In addition, the paper notes the connection of this interpretation with an older style of analysis using grouped data and illustrates the aggregation analogy with examples from the literature. The estimation technique used to illustrate the points is the method of instrumental variables.

Robert A. Moffitt
Department of Economics
Johns Hopkins University
Baltimore, MD 21218
and NBER

The aim of this note is to interpret estimation of the conventional treatment-effect selection-bias model in econometrics as a method of aggregation and to draw the implications of this interpretation. In addition, the paper notes the connection of this interpretation with an older style of analysis using grouped data and illustrates the aggregation analogy with examples from the literature. The estimation technique used to illustrate the points is the method of instrumental variables.

I. THEORY

The Model. To illustrate the issues in their simplest form, we consider the following conventional selection-bias econometric model of treatment effects:

$$(1) \qquad y_i = \beta + \alpha d_i + \epsilon_i$$

$$(2) \qquad d_i^* = \delta + \gamma z_i + \nu_i$$

$$(3) \qquad d_i = 1[d_i^* \geq 0]$$

where y_i is an outcome variable of interest, d_i is a binary treatment dummy whose effect α is the object of interest, z_i is an identifying variable which is also assumed to be a dummy variable, and 1 is the indicator function. Because there are no other covariates in the model and because both d_i and z_i are binary, the linearity assumptions in (1)-(2) are entirely unrestrictive. Although constant treatment effects are assumed, replacing α by an individual-specific heterogeneous response parameter (Heckman and Robb, 1985; Bjorklund and Moffitt, 1987) would change nothing of substance in what follows except the interpretation of the estimated treatment effect (Imbens and Angrist, 1994).

The selection-bias problem arises when $E(d_i\epsilon_i|z_i) \neq 0$ or, equivalently, $E(\nu_i\epsilon_i) \neq 0$. In this case the least-squares estimator

$$\alpha^{LS} = \bar{y}_{d1} - \bar{y}_{d0} \tag{4}$$

where $\bar{y}_{dm}$ is the sample mean of y_i taken over observations with d_i=m, is biased and inconsistent for α. Conditioning on z_i in (1) will not remedy the problem because the bias arises from the correlation of ϵ_i with the unobservable ν_i, not z_i (i.e., the bias is nonignorable).

Instrumental Variables. There are several alternative methods of consistently estimating α (Heckman and Robb, 1985). The aggregation interpretation can be most easily seen within the framework of instrumental variables (IV) estimation. If z_i is used as the instrument, the IV estimator is

$$\alpha^{IV} = \frac{\bar{y}_{z1} - \bar{y}_{z0}}{\bar{d}_{z1} - \bar{d}_{z0}} \tag{5}$$

where $\bar{y}_{zm}$ and $\bar{d}_{zm}$ are the means of y_i and d_i, respectively, taken over observations for which z_i=m. This estimator is equivalent to the two-stage least-squares version in which d_i is first regressed on z_i and then y_i is regressed on the predicted values of d_i from the first-stage regression. The estimator is consistent under a variety of well-known assumptions, the most important of which for present purposes is that $E(z_i\epsilon_i)=0$. Also necessary is that the denominator of (5) not have expectation zero.

Aggregation Interpretation. That the IV estimator can be interpreted as an estimator based on aggregated data follows immediately from (5), which is the coefficient from a least-squares regression of the group means $\bar{y}_{zm}$ on the group means $\bar{d}_{zm}$. The means of the outcome variable for each z-group are weighted averages of outcome means within the groups:

$$\bar{y}_{zm} = \bar{d}_{zm}\,\bar{y}_{zmd1} + (1-\bar{d}_{zm})\,\bar{y}_{zmd0} \tag{6}$$

where $\bar{y}_{zmdm'}$ is the mean of the outcome variable for those observations with $z_i=m$, $d_i=m'$.

Although the IV estimator has not been viewed in this fashion in the literature, it has long been recognized in the aggregation literature that aggregation is equivalent to IV where the grouping indicator is the instrument. In general, applying IV to a regression of y_i on a variable x_i using a set of grouping indicators z_{ij} as instruments--each z_{ij} set equal to 1 if observation i is in group j--is equivalent, in the two-stage version, of regressing y_i on predicted first-stage values of x_i, which are the group means the group means $\bar{x}_j$. But because a regression of the individual y_i on $\bar{x}_j$ yields the same coefficient as a regression of the group means $\bar{y}_j$ on $\bar{x}_j$, the IV estimator can be seen to be equivalent to aggregation.

The interpretation of IV as an aggregator can be further related to an ANOVA interpretation of IV, for the IV estimator is based on the between-group covariance of y_i and d_i, not its within-group covariance (where z_i defines the groups). It can be easily shown that

$$\alpha^{LS} = k\alpha^{IV} + (1-k)\alpha^{W} \tag{7}$$

where k is the fraction of the variation in d_i that arises from the between (i.e., from $\bar{d}_{z1}-\bar{d}_{z0}$) and α^W is the estimate obtained from within-z variation, equal to

$$\alpha^{W} = p(\bar{y}_{z1d1} - \bar{y}_{z1d0}) + (1-p)(\bar{y}_{z0d1} - \bar{y}_{z0d0}) \tag{8}$$

where p is a weight determined by the fraction of the variance in d_i in each of the two z-groups. The within estimator α^W arises from a regression of y_i on d_i, conditioning on z_i. The IV estimator thus subtracts out the contribution of the within-group variance to the estimate of α, leaving only the contribution arising from between-group aggregates.

Implications. One implication of this interpretation is that IV suffers a loss of efficiency under the null of no selection bias. The efficiency loss is identical to that arising from aggregation in general. Assuming equal sample sizes for the $z_i=1$ and $z_i=0$ groups to avoid side issues, the efficiency loss is:

$$(9) \qquad \frac{Var(\alpha^{LS})}{Var(\alpha^{IV})} = \frac{n \sum_m (\bar{d}_{zm}-\bar{d})^2}{\sum_i (d_i-\bar{d})^2} = \frac{.25(\bar{d}_{z1}-\bar{d}_{z0})^2}{\bar{d}(1-\bar{d})}$$

where n is the common group sample size and $\bar{d}$ is the overall mean of d_i. The RHS of is the ratio of the between-group variation in d_i to the total variation in d_i, and equals the parameter k in (7) assuming equal sample sizes. The fraction necessarily lies between 0 and 1.

A different issue pertains to the absolute level of precision in IV when the error term ϵ in (1) is correlated within z-groups. Such correlation is most commonly represented in econometrics by the random effects model, $\epsilon_{im} = \mu_m + \eta_{im}$. In the presence of such effects the variance of the group mean error term in IV is $Var(\epsilon_m) = Var(\mu_m) + Var(\eta_{im})/n$ and hence does not approach zero as n increases. Ignoring μ_m results in spuriously low standard errors (Moulton, 1986). However, this problem affects least-squares precision as well.

A second implication of the aggregation interpretation is that bias from group-level omitted variables is worse under IV than least squares, just as such bias is made worse by aggregation in general. If the true individual-level equation is

$$(10) \qquad y_{im} = \beta + \alpha d_{im} + \gamma w_m + \epsilon_{im}$$

for individual i in z-group m, omitting w_m from (10) biases the least-squares estimator of α by less than omitting w_m from a regression of $\bar{y}_m$ on $\bar{d}_m$ biases the IV estimator of α. The magnitude of the differential bias again follows directly from ANOVA considerations. In fact, assuming equal sample sizes for the $z_i=1$ and $z_i=0$ groups, the differential omitted-variable bias is

$$(11) \qquad \frac{\text{Bias}(\alpha^{LS})}{\text{Bias}(\alpha^{IV})} = \frac{.25(\bar{d}_{z1}-\bar{d}_{z0})^2}{\bar{d}(1-\bar{d})}$$

i.e., identical to (9).

II. ILLUSTRATIONS

An Historical Note. The aggregation implicit in IV bears a close relationship to the historical practice of using aggregate data in quasi-experimental program evaluations. Modern evaluation textbooks such as Rossi and Freeman (1993, pp.335-361) discuss the use of time-series, regression-discontinuity, simple before-and-after, and related designs, which be traced back at least to the literature on the study of educational reforms (Campbell and Stanley, 1963). While this literature has rarely formulated the problem in IV or simultaneous-equation terms, the frequent use of aggregate data on outcomes and treatment rates across states, cities, and over time is equivalent to IV. The use of aggregate data for program evaluation, common in the 1960s and early 1970s, lost favor in the research community as individual data came to be preferred. The typical critique of aggregate data is that the results are nonrobust and "extremely sensitive to specification errors" (Rossi and Freeman, 1993, p.335). The IV method represents a partial return to this approach.

While there is nothing in the IV approach that requires aggregation over area groups or time periods as instruments (rather than over individual or family characteristics), many applications in economics and other social sciences lead naturally in this direction for two reasons. First, studies of the effects of government policies on behavior naturally tend to this approach because such policies generally apply to individuals within governmentally-defined jurisdictions such as cities, school districts, counties, states, and countries. In other cases, government policies (e.g., tax policies) affect well-defined subgroups of the population differently than others, leading to aggregation over those groups. Second, in economic applications, the behavior of interest often takes place in the context of markets, which are by definition at a higher level of aggregation than the individual or firm and are often thought to be exogenous to the individual. Both area-defined units as well as industry aggregates lend themselves naturally to use under this rationale.

The following two illustrations are chosen because they use area-defined instruments. However, the studies couch their analyses in individual-level terms even though the selection-bias correction implicitly aggregates the analysis. One set of studies uses government program variation by area and the other uses private-market prices by area.

Two Studies of the U.S.D.A. Food Programs. Burghardt and Devaney (1995) studied the effects of the USDA School Lunch Program (SLP) on dietary intake of children. The study sample included approximately 3300 students in 329 schools. In each school there was a mix of participants and non-participants, for only some students were eligible and only some of those eligible chose to enroll in the program. The selection bias problem arises because those children who participate may like different types of food than those children who do not enroll, implying a difference in dietary intake even in the absence of the SLP. Although a variety of instruments for participation were examined, the preferred instrument was the price of an SLP lunch. Those prices were uniform within schools but varied across schools, and should naturally be expected to affect participation but not dietary intake directly.

In the context of the aggregation framework, the selection bias adjustment in this case implicitly represented a 329-observation regression of mean school dietary intake on the school participation rate (though scaled in price terms). This interpretation leads to at least two issues. One is whether school prices were set in a way that was responsive to school-specific dietary preferences; more generally, the mechanism determining price-setting and its source of variation can be seen to be an important issue. A second is whether sufficient other school and neighborhood characteristics were included to ensure that the school participation rate was not inadvertently picking up the effects of other factors.

Very little data were collected on price determination and relatively few school and neighborhood characteristics were obtained to examine these issues in depth. Because the evaluation design was framed in individual terms (i.e., by a comparison of participants to nonparticipants) rather than school-level terms, the data needs appropriate to the selection-bias corrections were not anticipated.

A second USDA program of significant policy interest is the Women, Infants, and Children (WIC) program, which provides supplemental foods, nutrition education, and health-care referrals to low-income pregnant women and their children up to age 5. Federal guidelines set income limits for participation but states have some authority in the setting of priority assignments for receipt of services. Participation is also voluntary on the part of recipients.Gordon and Devaney (1992) review several studies of the effect of the program on child birthweight. Selection bias was thought to occur because WIC participants might have lower birthweight even in the absence of the program than nonparticipants. The studies used a variety of candidate instruments, including individual characteristics but also state-level program variables such as WIC expenditures and the number of WIC clinics in the state.

From the aggregation perspective the latter type of selection bias adjustment implicitly involves a 51-observation regression of mean state birthweight on state WIC participation rates, scaled by WIC expenditures and the number of clinics in the state. From this perspective, the same two issues arise as in the SLP evaluation. First, it is natural to wish to investigate why WIC expenditures and the number of clinics vary across states, and to determine whether they vary in part because of the WIC caseload. Alternative WIC program characteristics such as federal funding levels and priority assignments might be more exogenous. Second, it is natural to seek to investigate whether sufficient additional state-level variables should be included to be able to convincingly control for other factors affecting the birthweight in a state that may be correlated with the WIC program variables.

Health Production Functions. There is a considerable literature on the estimation of the effect of health inputs on health status (e.g., the effect of going to the doctor on health). Selection bias is likely because those who use more health inputs presumably are in worse health. Here economic theory provides a strong case that market prices for the inputs should affect the level of inputs but should otherwise have no direct effect on health status. Since prices must be measured at the market level, selection-bias adjustment implicitly must require aggregation to the market level.

An early example is Rosenzweig and Schultz (1983), who estimate (among many other things) the effect of smoking behavior of a pregnant woman on the birthweight of her child. The authors had information on up to 10,000 women but used 51 state-level instruments for prices of cigarettes, milk, and other variables. The analysis therefore implicitly represented a 51-observation regression of mean state birthweight on mean state smoking rates, scaled by the price variables. More additional state-level variables were included in the model than is typical in the literature, which may have lessened any omitted variables problem. Nevertheless, a general issue in the approach concerns the source of price variation across states and whether it is demand or supply based.

III. REFERENCES

Bjorklund, A. and R. Moffitt. "Estimation of Wage Gains and Welfare Gains in Self-Selection Models." Review of Economics and Statistics 69 (February 1987): 42-49.

Burghardt, J. and B. Devaney. "Background on the School Nutrition Dietary Assessment Study." The American Journal of Clinical Nutrition 61 (Suppl 1995): 178S-181S.

Campell, D. and J. Stanley. Experimental and Quasi-Experimental Designs for Research. Chicago: Rand-McNally, 1963.

Gordon, A. and B. Devaney. "Selection Bias in Estimating WIC Impacts." Mathematica Policy Research Memorandum to Janet Tognetti, USDA, December 1992.

Heckman, J. and R. Robb. "Alternative Methods for Evaluating the Impact of Interventions." In Longitudinal Analysis of Labor Market Data, eds. J. Heckman and B. Singer. Cambridge: Cambridge University Press, 1985.

Imbens, G. and J. Angrist. "Identification and Estimation of Local Average Treatment Effects." Econometrica 62 (March 1994): 467-475.

Moulton, B. "Random Group Effects and the Precision of Regression Estimates." Journal of Econometrics 32 (August 1986): 385-397.

Rosenzweig, M. and T.P. Schultz. "Estimating a Household Production Function: Heterogeneity, the Demand for Health Inputs, and Their Effects on Birth Weight." Journal of Political Economy 91 (October 1993): 723-746.

Rossi, P. and H. Freeman. Evaluation: A Systematic Approach. Fifth Edition. Newbury Park: Sage Publications, 1993.