SOCIAL SCIENCE RESEARCH **26,** 487–504 (1997)
ARTICLE NO. SO970606

Instrumental Variables for Logistic Regression: An Illustration

E. Michael Foster

School of Policy Studies, Georgia State University

The estimated effect of a regressor on an outcome is inconsistent when that regressor is determined simultaneously with that outcome. Instrumental variables estimation is a means of obtaining consistent parameter estimates in this situation. The best-known form of instrumental variables is two-stage least squares; unfortunately, this procedure cannot be simply extended to non-linear models such as logistic regression. Instrumental variables estimation, however, is still possible, and using the Generalized Method of Moments, this paper is the first to produce instrumental variables estimates for logistic regression. Obtaining these estimates is easy using widely available software. An illustrative example is provided. This methodology should be useful to social scientists familiar with 2SLS and logistic regression.

The fundamental requirement for the consistency of least squares estimation is that any regressors be uncorrelated with the error term in the model (Davidson and MacKinnon, 1993, p. 149). If such a correlation exists, ordinary least squares (OLS) estimates are not consistent. In essence, this means that the estimator is consistently ''off the mark,'' no matter how large the sample.

One situation where a regressor might be correlated with the error term involves a regressor that is endogenous or is simultaneously determined with the outcome.[1] This situation might occur when the regressor not only causes but also reflects the outcome of interest (or when both the outcome and the regressor are

Thanks to J. S. Butler, Yasuo Amemiya, Robert Saunders, Tom Doub, and Damon Jones for valuable comments on earlier drafts. The author also appreciates the helpful comments of three initial reviewers on an earlier draft and those of a very careful reviewer on this manuscript. The author is responsible for any remaining errors.

Please address correspondence and reprint requests to the author at School of Policy Studies, 35 Broad Street, Georgia State University, Atlanta, GA 30303-3083.

[1] Such a correlation also might exist if the regressor is measured imprecisely (Fuller, 1987). This is the common problem of attenuation—measurement error dilutes the estimated effect of the regressor on the outcome of interest. In the standard linear model, standard instrumental variables estimation handles either the simultaneity or measurement error problems. As Y. Amemiya (1990) shows, however, this is not the case for the non-linear model. In particular, special steps must be taken to handle the measurement error problem. These steps are worked out in detail for logistic regression in Foster (1996).

0049-089X/97 $25.00

correlated with other, unobserved factors). One can see why this correlation exists if one recognizes that the equation of interest is actually one of two. In the other (unspecified) equation, the troublesome regressor is the dependent variable. If the errors in the two equations are correlated, then unexplained factors affecting the outcome of interest are correlated with those determining the troublesome regressor. This implies that the latter is correlated with the error term in the equation of interest (Davidson and MacKinnon, 1993, p. 21ff).

As an example of where one might encounter this problem, consider the analysis of ''neighborhood effects''—the effect of neighborhood conditions (e.g., the opportunity to earn money through criminal activities) on an individual's behavior (e.g., high school completion) (Foster and McLanahan, 1996; Evans, Oates, and Schwab, 1992; Jencks and Mayer, 1989). Empirical assessments of neighborhood effects generally involve combining micro-level data (like the Panel Study of Income Dynamics) with data on census tracts from the decennial census. (See, for example, Brooks-Gunn, Duncan, Klebanov, and Sealand, 1993.) The latter are used to characterize the neighborhood (using, for example, the neighborhood dropout rate). The dependent variable in these analyses is the education of specific individuals included in the microlevel data.

One suspects that the neighborhood dropout rate is endogenous for at least two reasons. First, the behavior of peers may reflect as well as influence a given individual's likelihood of finishing high school. (This might involve the choice of peers in the first place.) A second reason is that a family's choice of neighborhood may reflect how well its children are doing in school. For these reasons, the neighborhood dropout rate is likely correlated with unobserved determinants of an individual's chance of finishing high school—in other words, individuals and their families living in high-dropout neighborhoods differ systematically from those living elsewhere. Simple analyses of the dropout rate may confound those differences with the effect of the neighborhood dropout rate per se.[2]

If a regressor is endogenous, simple OLS estimates of the effect of the variable of interest on the outcome are not consistent.[3] In the linear model, a standard solution to this problem is instrumental variables estimation (IVE) and is straightforward. When the model of interest is non-linear (such as logistic regression), instrumental variables estimation is possible but somewhat more difficult. This paper reviews the literature on instrumental variables for non-linear models and describes how these methods can be combined with logistic regression. It then illustrates those methods with an analysis of neighborhood effects.

The paper has three sections. The first provides a brief overview of instrumental variables. The second reviews instrumental variables in non-linear models and

[2] Another possibility is that the neighborhood dropout rate is correlated with other, unobserved determinants of an individual's likelihood of dropping out of high school. This seems particularly plausible if these factors also influence a family's choice of neighborhoods.

[3] Coefficient estimates for the other regressors are affected as well. The nature of this bias is uncertain and depends on the relationship between the other regressors and the outcome and on that between those regressors and the regressor correlated with the error term.

offers an overview of the Generalized Method of Moments (GMM). The third section applies the methods outlined in the second section to the analysis of neighborhood effects. A summary follows.

1. INSTRUMENTAL VARIABLES: A REVIEW

IVE projects both the explanatory and dependent variables onto a space spanned by the instrumental variables and minimizes the distance between the two. The resulting estimates are consistent but may be biased in small samples (Taylor, 1983; Phillips, 1983). (See Wonnacott and Wonnacott, 1979, for a good discussion of the geometry of instrumental variables.)

What does it mean to project the dependent and explanatory variables onto the space spanned by the instrumental variables? The result is a comparison across a different ''dimension'' (Moffitt, 1996). Consider an evaluation of the effect of a job training program on earnings. It seems likely that participation and earnings are simultaneously determined: individuals with better job prospects are less likely to participate. In this case a simple comparison of participants with non-participants will confound the effect of job training with systematic between-person differences.

What might serve as an instrument in this case? Moffitt (1996) provides a hypothetical example where data for such an analysis comes from two cities that differ in terms of the level of funding available for job training. In this case, city of residence would serve as an instrument. Instrumental variables estimation would involve between-city differences in earnings adjusted for between-city differences in participation in job training. The key comparison is no longer between workers who do and do not participate in job training; rather it is between cities where the extent of job training differs (Moffitt, 1996).

The key step in producing IV estimates involves selecting the instruments themselves. The variables selected should meet two requirements (Angrist, Imbens, and Rubin, 1996a; Heckman, 1996; Angrist and Imbens, 1995; Davidson and MacKinnon, 1993, p. 209ff). First, conditioning on the troublesome regressor, the instruments must be uncorrelated with the outcome of interest.[4] This rules out variables that influence the outcome and that should be included as regressors in the model but are arbitrarily omitted. A second requirement is that the instrument(s) be correlated with the troublesome regressor.[5] In a multivariate context

[4] Heckman (1996)'s comment on Angrist Imbens, and Rubin (1996a) highlights that only mean independence is necessary and not full independence (as assumed in the text and by Angrist, Imbens, and Rubin (1996a)). As pointed out by Angrist, Imbens and Rubin (1996b), however, this distinction is not meaningful in practice. If mean independence held but full independence did not, the instrument would be valid for an analysis of the dependent variable but not of a non-linear transformation of the dependent variable.

[5] An added assumption typically made in instrumental variable estimation is that the effect of the troublesome regressor does not vary across individuals. Heckman (1996) makes the point that this need not be the case, but this is true in the most applications of instrumental variable estimation. The discussion of GMM below and the empirical illustration that follows also rely on this assumption. This

where additional regressors are included in the analysis, these assumptions require that the instruments be uncorrelated with unobserved determinants of the outcome of interest and that the instrument influence the troublesome regressor, controlling for the other covariates.

What do these assumptions imply? In the analysis of job training, city of residence must be related to job training but must not be correlated with earnings, controlling for whether or not an individual received job training. This would not be the case if the local economy was booming in one city but not the other.

In many cases, these requirements are met by variables that affect the troublesome regressor but do not directly determine the dependent variable of interest. In an analysis of the effect of a woman's fertility on her work hours, for example, one might fear that the latter causes as well as reflects the former. As a solution, one might use the fertility of the woman's mother as an instrument. One could argue that the fertility of the woman's mother affects the woman's desired (and actual) fertility but does not influence her labor force behavior directly. In this case, the fertility of the woman's mother is not correlated with her labor force behavior, controlling for her actual fertility. Such an analysis would presume that the fertility of a woman's mother would affect her own fertility.

IVE has been used widely in linear models. Its use has been encouraged by the ease of using one of its forms, two-stage least squares (2SLS). Easy to apply, 2SLS involves the application of OLS twice. In the first stage, the troublesome explanatory variable is regressed on the other explanatory variables in the model as well as on the instruments. The resulting estimates are used to calculate a predicted value of the troublesome explanatory variable (e.g., predicted number of children). In the second stage of the analysis, the outcome (e.g., hours of work)

---

assumption has been the subject of some debate in the literature on treatment effects. That discussion is relevant here because that literature calls into question whether and in what sense standard instrumental variable estimates are causal. Angrist and Imbens (1995; Imbens and Angrist, 1994; Angrist, Imbens, and Rubin, 1996a; 1996b) argue that the presence of instruments alone is not enough to establish meaningful causality, at least in the Rubin (1974) sense. They set out additional assumptions (e.g., principally the stable unit treatment values assumption (SUTVA) and monotonicity) under which instrumental variable estimates can be interpreted as the local average treatment effect (LATE) in the case of a binary treatment (or as the average causal response (ACR) in the case of a treatment with variable intensity). In a comment on Angrist, Imbens, and Rubin (1996a), Heckman (1996) disagrees with the authors for several reasons. The thrust of his argument is that standard instrumental variable estimates are causal and in a way that is preferred. In particular, he argues that instrumental variable estimation produces an estimate of the average effect of ''treatment'' for individuals who are ''treated.'' The LATE (or ACR) estimates, in contrast, represent the estimated impact of treatment for individuals who would be induced to change treatment states by a change in the value of the instrument. Heckman argues that this parameter is of ''questionable value,'' principally because it is defined for a sub-population that is not observed. The direct implication of this is that the additional assumptions Angrist and colleagues make (monotonicity; full independence versus only mean independence, and zero correlation between the instrument and unobserved determinants of treatment status) (p. 460) are unnecessary.

is regressed on the predicted value of the problematic explanatory variable as well as the other regressors in the model. The instruments are not included as regressors in the second stage. By assumption they do not influence the outcome of interest directly. If they were included, perfect collinearity would exist: the predicted value of the troublesome regressor is a linear function of the other regressors and the instruments.

The mechanics of 2SLS illumine the two requirements for instrumental variables. If the troublesome regressor is not influenced by the instruments, then its predicted value is simply a function of variables already in the model. In that case, adding the predicted value of the troublesome regressor to the model creates perfect collinearity: it is a function of (only) variables already included as regressors. If the instruments are correlated with unobserved determinants of the outcome of interest, then the predicted value of the troublesome regressor will be correlated with the error term as well. This leaves the analyst back where he or she started.

Unfortunately, 2SLS produces consistent estimates only if the second-stage regression is linear[6] (T. Amemiya, 1985). In many cases the outcome of interest is dichotomous, requiring the use of logistic regression. In this case, one cannot simply replace the second-stage of 2SLS with a logit model. In particular, consistent parameter estimates will not be obtained if one simply runs logistic regression with the predicted value of the troublesome regressor included an explanatory variable.

Fortunately, even though 2SLS is inappropriate for non-linear models, IVE is still possible (T. Amemiya, 1985). One way to obtain these estimates is through the Generalized Method of Moments (GMM). In the next section, we review GMM and then discuss how one can use GMM to produce instrumental variables estimates in non-linear models (such as logistic regression).

2. INSTRUMENTAL VARIABLES IN A NON-LINEAR MODEL

Non-technical Introduction to GMM

GMM estimates are obtained by solving a set of estimator-defining equations for the parameters of interest. (See Hansen (1982, 1985) and Hansen and Singleton (1982). A good introduction to GMM can be found in Greene (1993). More detailed treatments are available in Davidson and MacKinnon (1993), Gallant (1987) and Gallant and White (1988).) Estimator-defining equations are sample counterparts of equalities involving population moments. (Typically, this involves population moments that equal zero.)

[6] See Davidson and MacKinnon (1993, p. 225). Consider the case where the troublesome regressor appears on the right-hand side as $e^{\beta X}$. The problem is that the projection of $e^{\beta X}$ onto the instrument space is not the same as projecting X onto the instrument space (creating $\hat{X}$) and calculating $e^{\beta \hat{X}}$.

OLS is a special case of GMM. Consider a simple linear model:

$$Y = X\beta + \epsilon. \tag{1}$$

Y is the regressand (or dependent variable) and is a $N \times 1$ vector; X, a $N \times K$ matrix of regressors; and ϵ, a $N \times 1$ vector of errors or unexplained determinants of Y. β is a $K \times 1$ vector of regression coefficients and represents the parameters of interest.

In this case, the estimator-defining equations are

$$\frac{1}{N}\sum_{t=1}^{N} x_{t,k}(y_t - X_t\hat{\beta}) = 0, \tag{2}$$

where $k = 1$ to K. X_t is a $1 \times K$ vector and represents the tth row of the X matrix. $x_{t,k}$ refers to the kth element of X_t and is the value of the kth explanatory variable for the tth individual. This set of K equations can be solved for the vector $\hat{\beta}$, the GMM estimates of β. These conditions (2) are implied by the first-order conditions for minimizing the sum of squares. Given that $e = Y - X\hat{\beta}$, it is apparent that estimation assumes the residual and the explanatory variables do not covary.

When X is endogenous or simultaneously determined with the outcome of interest, Eq. (2) will not hold. (See Davidson and MacKinnon, 1993, p. 21ff.) However, x can be replaced with instruments, w, producing

$$\frac{1}{N}\sum_{t=1}^{N} w_{t,k}(y_t - X_t\tilde{\beta}) = 0, \tag{3A}$$

where $k = 1$ to K. $w_{t,k}$ refers to the value of the kth instrument for the tth individual. As long as there are K of these instruments, one can solve these equations for the instrumental variables estimates ($\tilde{\beta}$). Equation (3A) represents the first-order conditions for instrumental variable estimation (Davidson and MacKinnon, 1993, p. 585). Explanatory variables that are not correlated with the error term can serve as their own instrument; for those variables, the corresponding members of Eqs. (2) and (3A) are identical.

Equation (3A) highlights the importance of the assumptions underlying instrumental variables estimation. If the instrument were correlated with ϵ in (1), then (3A) would not hold. In that case, solving (3A) for $\tilde{\beta}$ would produce incorrect estimates of that parameter. Estimation would not be possible if there were no relationship between the instruments and the troublesome regressor. Consider the case where there is one, potentially endogenous regressor and one instrument. In that case, assuming all variables are mean zero, (3A) can be rewritten as

$$\sum_{t=1}^{N} w_t y_t - \tilde{\beta}\sum_{t=1}^{N} w_t x_t = 0 \tag{3B}$$

and $\tilde{\beta}$ is calculated as

$$\tilde{\beta} = \frac{\sum_{t=1}^{N} w_t y_t}{\sum_{t=1}^{N} w_t x_t}. \tag{3C}$$

If x and w do not covary, then (3C) cannot be evaluated.

In many cases, the number of instruments exceeds the number of X variables. In an analysis of the effect of fertility on labor force behavior, one also might use the total fertility of a woman's mother-in-law as an additional instrument. This second instrument produces an extra equation in (3A). In this situation, the model is said to be overidentified; there are more instruments and estimator-defining equations than are necessary to produce parameter estimates. The additional elements of (3A) that correspond to the ''surplus'' instruments correspond to what are called overidentifying restrictions.

When the model is overidentified, rather than solving the system of equations, GMM estimates are obtained by minimizing the following criterion function:

$$e^T W(W^T W)^{-1} W^T e. \tag{4A}$$

L is the number of instruments ($L > K$), e is the $N \times 1$ vector of residuals, W is the $N \times L$ matrix of instruments. (One also can obtain estimates in the just-identified case ($L = K$) in this manner as well. If the model is just-identified, then the minimized value of the criterion function is 0.) The properties of the resulting estimates are well-established. GMM estimators are asymptotically normal and have a covariance matrix equal to $[G^T W(W^T W)^{-1} W^T G]^{-1}$, where G is -1 times the derivative of the right-hand-side of (1) with respect to β.

Note that in the estimation below the criterion function is modified somewhat. In particular, if the residuals are heteroskedastic (as one would expect with a dichotomous outcome), one can improve the efficiency of the estimates by minimizing

$$e^T W(W^T \Omega W)^{-1} W^T e, \tag{4B}$$

where Ω is the variance–covariance matrix of the residuals. Minimizing (4B) requires an estimate of $W^T \Omega W$, and a standard procedure is to use White's (1980) heteroskedasticity-consistent covariance matrix estimator. This matrix allows for very general forms of heterogeneity. In this case, estimation proceeds in two stages. An estimate of the weighting matrix is obtained in a first stage; (4B) is minimized in the second, producing efficient estimates of the coefficients of interest. In this case, the variance covariance matrix of the parameter estimates is $[G^T W(W^T \Omega W)^{-1} W^T G]^{-1}$, where G is defined as before.

Note that in the case where the model is overidentified, one can test the

overidentifying restrictions using the minimized value of the criterion function (4A) or (4B). This value is distributed as χ^2 with degrees of freedom equal to the number of overidentifying restrictions (i.e., the number of extra instruments). (See Davidson and MacKinnon, 1993, pp. 614ff.) If these restrictions are valid, the value of the criterion function should be small; large values lead one to reject the overidentifying restrictions. This would call into question the assumptions underlying those restrictions as well as the parameter estimates that depend on them.

GMM Estimation of the Logit Model

A simple non-linear model of considerable interest is logistic regression, where

$$\Pr(y_t = 1) = \frac{e^{X_t\beta}}{(1 + e^{X_t\beta})} \quad \text{or} \quad (1 + e^{-X_t\beta})^{-1.} \tag{5}$$

Y equals 1 if an individual t experiences the event (e.g., finishes high school) or has the characteristic of interest (e.g., receives welfare) and 0, otherwise. In this case, the estimator-defining equations are orthogonality conditions between x and the residuals

$$\frac{1}{N}\sum_{t=1}^{N} x_{t,k}(y_t - (1 + e^{-X_t\tilde{\beta}})^{-1}) = 0, \tag{6}$$

where $k = 1$ to K. (X and x are defined as above.) Solving these equations for the parameters of interest (either directly or by defining the appropriate criterion function) produces GMM estimates ($\hat{\beta}$) of the model parameters. It is worth noting that (6) represents the first-order conditions for maximum-likelihood estimation. It is apparent that in this case MLE is a special case of GMM. (See Greene, 1993.)

The intuition behind (6) is similar to that for the linear model: by assumption, the prediction error should not be correlated with regressors. If one or more regressors are endogenous, the estimator-defining equations are again incorrect.

Instrumental Variables Estimation for Logistic Regression

T. Amemiya (1985) demonstrates that in the case of endogenous regressors, non-linear IV estimates are obtained in the same manner as they are in the linear case: the troublesome regressors are simply replaced in the estimator-defining equations by appropriate instruments. In the case of logistic regression, this produces

$$\frac{1}{N}\sum_{t=1}^{N} w_{t,k}(y_t - (1 + e^{-X_t\tilde{\beta}})^{-1}) = 0, \tag{7}$$

where $k = 1$ to K, where K equals the number of explanatory variables. When there is one instrument per variable, the model is just identified, and one can solve

these equations for the K elements of $\tilde{\beta}$. In the overidentified case, one can form a criterion function like (4A) where the residual is defined as $y_t - (1 + e^{-X_t\tilde{\beta}})^{-1}$. The resulting estimates have been labeled non-linear two-stage least squares estimates (Amemiya, 1985).

3. AN ILLUSTRATIVE EXAMPLE

As an example, we consider the effect of the neighborhood dropout rate on the likelihood that an individual completes high school (Foster and McLanahan, 1995). For a variety of reasons (Jencks and Mayer, 1989), one would expect individuals growing up in neighborhoods with a high dropout rate to be more likely to drop out themselves. This could involve peer effects, or the neighborhood dropout rate could reflect conditions in the neighborhood (such as opportunities to earn money through criminal activities) that influence whether an individual finishes high school. Whether neighborhood effects are peer effects or the effects of the neighborhood conditions, the key question is whether moving an individual to a neighborhood where the dropout rate is lower would lower his or her chance of dropping out.

The model of interest is one where

$$\Pr(y_t = 1) = (1 + e^{-X_t\beta - N_t\gamma})^{-1} \tag{8}$$

y_t is a dichotomous variable indicating whether the individual drops out (i.e., fails to finish high school by age 20). X refers to a vector of family characteristics and β to the corresponding regression coefficients. N is the neighborhood dropout rate, and γ determines the effect of the neighborhood dropout rate on an individual's likelihood of dropping out. (In the earlier equations, N was part of the X matrix.) Family characteristics are included to improve the explanatory power of the model and to prevent the effect of the neighborhood dropout rate from being confounded with neighborhood patterns in family characteristics (like family structure). β and γ could be estimated using logistic regression. For reasons discussed above, however, it seems likely that estimates of the latter will be inconsistent because neighborhood conditions are endogenous—i.e., an individual's chance of finishing high school is jointly determined with the type of neighborhood in which he or she lives.

We examine the relationship between an individual's chance of completing high school and the neighborhood in which he or she grows up using information on a sample of young men and women drawn from the 1987 wave of the Panel Study of Income Dynamics (PSID). The PSID is a sample of 5000 families first interviewed in 1968 and re-interviewed each year thereafter. The PSID collects information annually on the economic and social characteristics of sample households and on members of these households (Economic Behavior Program, 1984). Our analyses are based on a sample of 667 girls and 655 boys from panel families who were between the ages of 1 and 5 in 1968. (The sub-sample of individuals meeting the age restriction only totaled 684 boys and 687 girls. 20 girls and 29 boys lacked needed information on schooling or family background

TABLE 1
Means of Background Characteristics
All Cases (n = 1322)

Variable	Mean	Std. deviation
Individual finished high school by age 20	0.78	0.41
Race (1 = white; 0 = other)	0.51	0.50
Characteristics of household child lived in at age 16		
Female-headed (0 = male; 1 = female head)	0.28	0.45
Head high-school dropout (1 = yes; 0 = no)	0.53	0.50
Head attended college (1 = yes; 0 = no)	0.20	0.40
Neighborhood dropout rate	14.56	9.97
Characteristics of labor market area		
Dropout rate	13.28	3.81
Rate of public assistance	8.38	3.11
Poverty rate	12.62	4.84
% Families with 1979 income >$30,000	24.72	8.84
% Workers in		
Executive occupations	10.08	2.36
Professional occupations	14.89	2.98
Manufacturing industries	22.92	8.56
Service industries	15.15	2.49

Source. 1968–1987 tapes of the Panel Study of Income Dynamics.

and were excluded from the analyses.) These individuals turned 20 between 1983 and 1987, and we examined whether sample members had finished high school (or received a GED) at that point. Seventy-eight percent of the sample had finished high school by age 20. Our analysis attempts to predict which of these individuals failed to finish high school, using the information on family background and the neighborhood dropout rate.

Sample members were age 16 between the years of 1979 and 1983. At that time, an array of family characteristics were recorded including race, family structure (specifically whether the household was headed by woman) and several measures of economic status (family income, welfare receipt as well as household head's employment status and education). Table 1 presents sample means for the variables used in the analysis. One can see that at age 16 more than one-fourth (28%) of the sample lived in a family with a female head, and more than half of (53%) the sample lived in a family where the head of the household did not finish high school. Just over half of the sample was white, reflecting the overrepresentation of the poor and African Americans in the PSID's original sampling frame.

Information on the census tract and local labor market in which the individual

and his or her family lived when the former was age 16 was obtained using information from the 1980 census and a special address file supplied by the Institute for Social Research.[7] Neighborhood conditions are represented using tract characteristics when possible. In non-tracted areas, we used information on Minor Civil Divisions (MCD), the ''primary political/administrative subdivisions of counties (or county-equivalents),'' in place of census tract information. (Most MCD's are called ''townships.'' (Adams, 1992, p. 26).) The average tract in which a PSID sample member lives has 5020 persons and 1832 households.

In the discussion that follows, city-level characteristics are actually characteristics of the ''local labor market area.'' The local labor market area was created by the Institute for Social Research and is defined as ''one or more counties with close economic ties defined by patterns of commuting to work'' (Adams, 1992, p. 15). In practice, the local labor market area is the metropolitan statistical area for metropolitan counties and the county or state employment area for non-metropolitan areas depending on whether fewer or greater than 20% of the residents commute to work outside of the county, respectively.

For each neighborhood, we used census data to calculate a neighborhood (tract/MCD) dropout rate which is defined as *the proportion of individuals ages 16 to 19 who were not in school and had not completed high school.* Census data are also used to calculate a variety of characteristics for each local labor market area: the dropout rate, the poverty rate among the non-elderly and the rate of public assistance (percentage of families receiving AFDC, General Assistance, or Aid to the Aged, Blind, or Disabled). Census data also provided information on the proportion of workers who are employed in professional occupations, in executive occupations, in manufacturing, and in the service sector as well as the proportion of families who have incomes above $30,000.[8]

Because neighborhood conditions are likely determined jointly with an individual's likelihood of finishing high school, conventional logistic estimates of the effect of neighborhood conditions on high school completion are likely to be inconsistent. What might serve as an instrument for the neighborhood dropout rate? One prominent source of instruments is variables that influence the problematic regressor but do not influence the outcome directly. Thinking along these

[7] These geocode files are part of the sensitive data files of the Panel Study of Income Dynamics and were obtained under special contractual arrangements designed to protect the anonymity of respondents. These data are not available from the authors. Persons interested in obtaining PSID sensitive data files should contact Panel Study of Income Dynamics, Box 1248, Ann Arbor, MI, 48106-1248.

[8] These characteristics are also available for the neighborhood as well. This analysis focuses on the neighborhood dropout rate for several reasons. First, if neighborhood effects capture peer effects, then the neighborhood dropout rate seems most appropriate. If neighborhood effects capture the impact of neighborhood institutions, then the neighborhood dropout rate would appear to be the best measure of these institutions as they influence high school completion. Second, empirical analyses (described in Foster and McLanahan (1995)) reveal that the dropout rate is the neighborhood characteristic that best predicts high school completion. Those analyses also reveal that collinearity problems arise if more than one neighborhood characteristic is included in the analysis.

lines, we use characteristics of the local labor market (or city) as instruments. These are the characteristics discussed above: (1) the dropout rate for the labor market area; (2) the proportion of workers who are employed in professional occupations; (3) the proportion employed in executive occupations; (4) the proportion employed in manufacturing; (5) the proportion employed in the service sector; (6) the proportion of all families who have incomes above $30,000; (7) the poverty rate; and (8) the proportion of families receiving public assistance.

The choice of these characteristics as instruments rests on the assumption that these characteristics do not *directly* influence a young person's chance of finishing high school, *except insofar as they influence neighborhood conditions.* Stated another way, given whether an individual lives in a 'good' or 'bad' neighborhood, his or her chance of finishing high school does not depend on whether the larger labor market area is prospering or failing. Living in a city where labor market conditions are deteriorating, however, makes it more likely a given individual will live in a neighborhood where the dropout rate is high.

In the next section, we use this assumption to obtain instrumental variables estimates of the effect of the neighborhood dropout rate.[9] We then test the validity of this assumption.

Results

Table 2 presents coefficient estimates for boys and girls (the top and bottom panel, respectively). Each panel presents three sets of estimates: conventional logit estimates, what we refer to as ''pseudo-2SLS'' estimates and GMM-IVE estimates. The family characteristics described above were used as additional regressors in all analyses. (The complete analyses are available from the author.)

The first row in each panel presents estimates from standard logistic regression. (Note that these analyses are unweighted as are all that follow. Results from weighted analyses are very similar.) These estimates suggest that neighborhood conditions strongly influence a girl's likelihood of not completing high school. This effect is statistically significant ($p < .01$), and the implied impact on a young woman's probability of dropping out is quite large: a ten percentage point change in the neighborhood dropout rate implies a 4.79 percentage point increase in an her likelihood of dropping out. In contrast, the effect for boys is very small in both statistical and practical terms.[10]

For comparison purposes, row 2 in each panel presents estimates labeled ''pseudo-2SLS'' because they were obtained in a manner analogous to 2SLS

[9] Another way to take advantage of this information would be to jointly estimate models of neighborhood conditions and high school completion. This was done in Evans, Oates, and Schwab (1992). These maximum likelihood estimates require additional assumptions (in particular, normality of the neighborhood dropout rate) that are not required by the approach taken here.

[10] While not presented here, GMM estimates of the logit model were also calculated. The point estimates are identical to those for the maximum likelihood estimates presented in Table 2.

TABLE 2
Alternative Estimates of the Effect of the Neighborhood Dropout Rate on an Individual's Chance of *Not* Completing High School by Age 20

		Beta	Standard error	Sig.
A. Boys	Logit	0.0099	0.0103	0.34
	Pseudo-2SLS[a]	0.0267	0.0292	0.36
	GMM-IVE	0.0379	0.0256	0.14
B. Girls	Logit	0.0400	0.0108	**
	Pseudo-2SLS	0.0388	0.0294	0.19
	GMM-IVE	0.0422	0.0247	*

Source. 1968–1987 tapes of the Panel Study of Income Dynamics.

Note. All estimates were obtained using the education, race and sex of the household head as additional controls. The coefficients on those regressors were of a sign and magnitude consistent with prior research. The full results are available from the author.

[a] Pseudo-2SLS refers to estimates obtained by first regressing the neighborhood dropout rate on the other controls and on conditions in the local labor market. This regression is used to generate a predicted value of the neighborhood dropout rate. That prediction is used as a regressor in the second stage logit analysis of the likelihood that a specific individual finishes high school.

** $p < .05$.

* $.05 < p < .10$. The actual p value is reported in other cases.

estimates. In particular, the neighborhood dropout rate first was regressed on the other controls and on conditions in the local labor market. This resulting coefficient estimates were used to generate a predicted neighborhood dropout rate.[11] That prediction was used as a regressor in the second-stage logit analysis of whether a specific individual finishes high school. The resulting estimate for girls is fairly similar to the conventional logit estimate. The standard error is much larger, and so the 95% confidence interval for the estimate now includes zero. The magnitude of the point estimate for boys is now much closer to that for girls. The estimated effect, however, is still not statistically significant.

Row 3 of each panel provides GMM-IV estimates of the neighborhood effect. For girls, these estimates are virtually identical to the conventional logit estimates. One can test the hypothesis that the two estimates are the same using the Hausman test.[12] The test statistic has a standard normal distribution, and for girls,

[11] These results are available from the author. The estimated coefficients generally were as expected. Race and parental education were significant predictors of the neighborhood dropout rate as was the city-level dropout rate, the proportion of families with income above $30,000 (girls only), the proportion of workers in executive occupations (boys only) and the proportion of workers in manufacturing (boys only).

[12] The Hausman test allows one to compare two estimators, one of which is consistent whether the null is true or false. The other estimator is consistent only under the null but is efficient in that case (Davidson and MacKinnon, 1993, p. 239). If the two estimators produce estimates that differ statistically, one can reject the null. In that case, the estimates produced by the estimator that is

has a value of .10 ($p = .92$). This suggests that any simultaneity bias has a negligible effect on estimates of the neighborhood effect for girls.

The GMM-IVE estimates for boys are very similar to those for girls. They are substantially larger than the conventional logit estimates.[13] Because the GMM-IVE standard errors are larger, however, the hypothesis that neighborhood conditions do not influence boys cannot be rejected ($p = .14$). Similarly, the hypothesis that the conventional logit and the GMM-IVE are the same can be rejected only at $p = .23$. (The value of the test statistic for the Hausman test is 1.20.)

These results suggest that the neighborhood in which one lives does influence one's likelihood of completing high school. The point estimates for both boys and girls are similar and suggest that moving a young person to a neighborhood where the dropout rate is 10 percentage points lower would reduce his or her chance of dropping out by 4 percentage points or more. While the effect for boys is not statistically significant, it is much larger than the impact measured in conventional logit estimates.

Of course, these estimates depend on the presumed structure of the model—i.e., on the assumption that the city-level characteristics do not directly influence the likelihood that an individual drops out. Is this model reasonable? We can gauge the validity of the model to some extent by determining whether the overidentifying restrictions fit the data. Such a test depends on the value of the criterion function corresponding to (7); it is distributed χ^2 with 7 degrees of freedom (the number of overidentifying restrictions or extra instruments). For girls, the value of the criterion function is 8.73. This is not statistically significant ($p = .27$) and suggests that *for girls* the hypothesized over-identifying restrictions cannot be rejected. For boys, however, the validity of the overidentifying restrictions is questionable. The value of the criterion function is 11.27, a value that is nearly significant at conventional levels ($p = .13$). These results suggest that the estimates presented here for boys should be interpreted with caution.

Why might the model structure be invalid for boys? Other research suggests that boys' schooling decisions are based on the overall economic health of an area (Clark and Wolfe, 1992). If that is the case, the city-level characteristics belong in the equation predicting high school completion, and therefore treating them as instruments is a mistake. For boys, therefore, it appears that we are back where we

---

consistent whether the null is true or not is to be preferred. In the case at hand, the null hypothesis is that the neighborhood dropout rate is uncorrelated with the error term. If that is indeed true, the ordinary logistic estimates are efficient. If such a correlation does exist, however, only the IV estimates are consistent. In the case at hand, the hypothesis that the two sets of estimates are the same—that the neighborhood dropout rate is not correlated with the error term—cannot be rejected.

[13] Why are the instrumental variable estimates for boys larger than conventional estimates? One explanation is that the parents of boys likely to drop out move their families to other neighborhoods to improve their son's chance of finishing high school. To know whether this is in fact the case, one might jointly model residential mobility among parents and the high school completion of their children.

started: we fear that the OLS estimates are inconsistent but have no means of assessing whether this is the case nor of correcting the estimates if we find that they are. (It is worth noting that we also estimated the just-identified model where only the labor market dropout rate is used as an instrument. In this model, all of the other labor market characteristics are included as regressors. None of these variables have a statistically significant impact on the likelihood that a young man drops out. The estimated impact of the neighborhood dropout rate in this model is very similar to the GMM-IVE estimate in Table 2.)

A Final Caveat

Are these results misleading? The discussion above seems to suggest that IV estimates are better than the standard logistic estimates, even though—at least for girls—the numbers are fairly similar. The similarity between the two seems to suggest that simultaneity is not particularly severe. An alternative explanation for the similarity is that the instruments may be particularly effective in predicting the neighborhood dropout rate. In short, the IV estimates may resemble the GMM estimates because there is no difference between predicted and actual neighborhood conditions. In that case, both sets of estimates are biased in finite samples, no matter how well IVE performs in infinitely large samples. (See pages 222ff of Davidson and MacKinnon, 1993.)

We considered this possibility by examining how well the instruments (and other explanatory variables) predict the neighborhood dropout rate. r^2 for this equation is .28. (For boys, the figure is .23.) Thus there appears to be a real difference between predicted and actual neighborhood conditions. While we cannot say for sure, it does not appear that the standard logit and GMM-IVE estimates are similar (for girls) because we have been especially effective in predicting the neighborhood dropout rate.

SUMMARY

Estimates of the effect of a regressor are not consistent when that regressor is simultaneously determined with the outcome of interest. This paper outlines the use of the Generalized Method of Moments to obtain instrumental variables estimates for logistic regression. This methodology is then illustrated with an analysis of neighborhood effects. The results from that example suggest that neighborhood conditions do influence an individual's likelihood of finishing high school. The estimated effect is statistically significant for girls but not boys. A test of the structure of the model suggests that the assumptions underlying instrumental variables estimation in this application may be untenable, at least for boys.

Instrumental Variables for Logistic Regression (April 2, 1997)

APPENDIX 1

Sample LIMDEP Program for Generating GMM-IVE Estimates

```
   Comments are included in italics.
open; output=c:\work\gmm-ive\gmmive.out$                    Opens output file
Read;                                                       Reads data
   file=c:\work\gmm-ive\gmmive.wk1;                         Data stored as lotus file
        format=wks;
        names $
skip $                                          Tells LIMDEP to use listwise deletion.
namelist;                       Limdep convention for referring to groups of variables.
   y1=ido;                                                  Dependent variable
   x1=one,pardo,psomecol,raceh68,sexh16,ndrpout;            Explanatory variables
   x2=one,pardo,psomecol,raceh68,sexh16                     Variables that will serve
                                                            as their own instruments.
   instr=lpoccex,lpoccpr,lpmanu,lpserv,lpfam30k,lpubass,lppoor,ldrpout $ Instruments for neighbor-
                                                            hood dropout rate.
title; Newton logit $
logit; lhs=ido; rhs=x1; marginal $
matrix; logitB=b $              Coefficient estimates are stored in matrix for use
                                as starting values for estimation below.
title; GMM (NO IVE; Heter Correction) $
nlsq; lhs=ido;
      labels=constant,b_do,b_col, b_race,b_sexh,b_ndo;      Specifies the parameters.
      start=logitB;
      fcn=Lgp(Dot[x1]);         Lgp is the logit function.
                                Dot takes the dot product of the variables in x1 with
                                the parameters specified in the label statement.
       pds=0 $                  pds=0 employs the heteroskedasticity-consistent
                                criterion function.
title; GMM-IVE $
nlsq; lhs=ido;
      labels=constant,b_do,b_col,b_race,b_sexh,b_ndo;
      start=B;
      fcn=Lgp(Dot[x1]);
      inst=x2, instr;
      pds=0 $
```

REFERENCES

Adams, T. K. (1991). ''Documentation for 1968–1985 PSID-geocode match files.'' Mimeo. Economic Behavior Program, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI.

Amemiya, T. (1974). ''The non-linear two-stage least-squares estimator,'' Journal of Econometrics **2,** 105–110.

Amemiya, Y. (1990). ''Two-stage instrumental variable estimators for the nonlinear errors-in-variables model,'' Journal of Econometrics **44,** 311–332.

Angrist, J. D., and Imbens, G. W. (1995). ''Two-stage least squares estimation of average causal effects in models with variable treatment intensity,'' Journal of the American Statistical Association **90**(430), 431–442.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996a). ''Identification of causal effects using instrumental variables,'' Journal of the American Statistical Association **91**(434), 444–455.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996b). ''Rejoinder [to Comment by Heckman],'' Journal of the American Statistical Association **91**(434), 468–472.

Brooks-Gunn, J., Duncan, G., Klebanov, P., and Sealand, N. (1993). ''Do neighborhoods influence child and adolescent development?'' American Journal of Sociology, 353–395.

Carroll, R. J., and Stefanski, L. A. (1994). ''Measurement error, instrumental variables and corrections for attenuation with applications to meta-analyses,'' Statistics in Medicine **13,** 1265–1282.

Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995). Measurement Error in Nonlinear Models, Chapman and Hall, London.

Davidson, R., and MacKinnon, J. G. (1993). Estimation and Inference in Econometrics, Oxford Univ. Press, New York.

Economic Behavior Program (1984). User Guide to the Panel Study of Income Dynamics, Inter-university Consortium for Political and Social Research, Center for Political Studies, Institute for Social Research, University of Michigan, Ann Arbor, MI.

Econometric Software, Inc. (1995). LIMDEP: User's Manual and Reference Guide, Econometric Software, Bellport, NY.

Evans, W. N., Oates, W. E., and Schwab, R. M. (1992). ''Measuring peer group effects: A study of teenage behavior,'' Journal of Political Economy **100,** 966–91.

Foster, E. M., and McLanahan, S. (1996). ''A beginner's guide to instrumental variables,'' Psychological Methods **3**(1), 249–260.

Foster, E. M. (1996). ''Logistic regression and regressors measured with error,'' unpublished manuscript.

Fuller, W. A. (1987). Measurement Error Models, Wiley, New York.

Gallant, A. R. (1987). Nonlinear Statistical Models, Wiley, New York.

Gallant, A. R., and White, H. (1988). A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models, Basil Blackwell, Oxford.

Greene, W. H. (1993). Econometric Analysis, 2nd ed., Macmillan, New York.

Hansen, L. P. (1982). ''Large sample properties of generalized method of moments estimators,'' Econometrica **34,** 646–660.

Hansen, L. P. (1985). ''A method for calculating bounds on the asymptotic covariance matrices of generalized method of moments estimators,'' Journal of Econometrics **30,** 203–238.

Hansen, L. P., and Singleton, K. J. (1982). ''Generalized instrumental variables estimators of nonlinear rational expectations models,'' Econometrica **50,** 1269–1286.

Heckman, J. J. (1996). ''Comment [on 'identification of causal effects using instrumental variables'],'' Journal of the American Statistical Association **91**(434), 459–461.

Imbens, G. W., and Angrist, J. D. (1994). ''Identification and estimation of local average treatment effects,'' Econometrica **62**(2), 467–475.

Jencks, C., and Mayer, S. E. (1989). ''The social consequences of growing up in poor neighborhood,'' working paper, Center for Urban Affairs and Policy Research, Northwestern University.

Moffitt, R. (1996). ''Comment [on 'identification of causal effects using instrumental variables'],'' Journal of the American Statistical Association **91**(434), 462–465.

Phillips, P. C. B. (1983). ''Small sample theory in the simultaneous equations model,'' *in* Handbook of Econometrics (Z. Griliches and M. D. Intriligator, Eds.), Vol. 1, Chap. 8, North Holland, Amsterdam.

Rubin, D. (1974). ''Estimating causal effects of treatments in randomized and non-randomized studies,'' Journal of Educational Psychology **66,** 688–701.

Stefanski, L. A., and Buzas, J. S. (1995). ''Instrumental variables estimation in binary regression measurement error models,'' Journal of the American Statistical Association **90,** 541–550.

Taylor, W. E. (1983). ''On the relevance of finite sample distribution theory,'' Econometric Reviews **2,** 1–84.

White, H. (1980). ''A heteroskedasticity consistent covariance matrix and a direct test for heteroskedasticity,'' Econometrica **48,** 817–838.

Wonnacott, R. J., and Wonnacott, T. H. (1979). Econometrics, 2nd ed., Wiley, New York.