Control Function Methods in Applied Econometrics

Jeffrey M. Wooldridge

ABSTRACT

This paper provides an overview of control function (CF) methods for solving the problem of endogenous explanatory variables (EEVs) in linear and nonlinear models. CF methods often can be justified in situations where "plug-in" approaches are known to produce inconsistent estimators of parameters and partial effects. Usually, CF approaches require fewer assumptions than maximum likelihood, and CF methods are computationally simpler. The recent focus on estimating average partial effects, along with theoretical results on nonparametric identification, suggests some simple, flexible parametric CF strategies. The CF approach for handling discrete EEVs in nonlinear models is more controversial but approximate solutions are available.

I. Introduction

The term "control function" has been part of the econometrics lexicon for several decades, but it has been used inconsistently, and its usage has evolved. In early work — notably, Barnow, Cain, and Goldberger (1981) (hereafter, BCG) — a control function is a variable that, when added to a regression, renders a policy variable appropriately exogenous. From the BCG perspective, multiple regression that includes the policy variable and one or more control functions provides consistent estimation of the causal effect of a policy intervention. Cameron and Trivedi (2005, p. 37) endorses this definition of a control function (CF), and, based on the usage in BCG, what Wooldridge (2010, Section 4.3.2) defines as a proxy variable would be considered a CF. As one example, a standardized intelligence test score, such as IQ score, can be considered a CF if conditioning on it appropriately controls for unobserved cognitive ability, thereby enabling consistent estimation of the causal effect of schooling in a standard wage equation.

Jeffrey M. Wooldridge is a University Distinguished professor of economics at Michigan State University.
He thanks four anonymous referees and the editors for very helpful comments on two earlier drafts. Data used in this article are available from the author from November 2015 through October 2018.
[Submitted April 2013; accepted February 2014]
ISSN 0022-166X E-ISSN 1548-8004 © 2015 by the Board of Regents of the University of Wisconsin System

In the application motivating BCG, variables measuring socioeconomic status (SES) are control functions if participation in a program—such as Head Start—is essentially determined by the SES variables. Goldberger (1972, reprinted 2008) was an important contribution studying the problem of whether controlling for observables could solve self-selection into program participation although it did not use the phrase "control function." Therefore, in early usage, the notion of a control function was closely tied to assumptions of "ignorable" or "unconfounded" treatment assignment that are prevalent today: Conditional on observed covariates, the key policy variables are appropriately exogenous. (For an overview, see Imbens and Wooldridge 2009.)

Heckman and Robb (1985), in the context of program evaluation with longitudinal data, also describes a control function as a variable that, when conditioned on, makes an intervention exogenous in a regression equation. It explicitly recognizes that CFs might depend on unknown parameters and that to operationalize a CF procedure the parameters must be estimated in a first stage. One example is a lagged residual in a program evaluation equation using longitudinal data.

For the most part, the modern usage of "control function" maintains the spirit of the earlier definitions but with an important defining feature: Constructing a valid CF relies on the availability of one or more instrumental variables. I take this perspective in the current paper: The control function approach to estimation is inherently an instrumental variables method. More precisely, the equation of interest—for brevity called the "structural equation"—contains at least one explanatory variable that is endogenous, or suspected of being so, in the sense that it is correlated with unobservables in the equation. Further, I have excluded exogenous variables from the structural equation that explain variation in the endogenous explanatory variables (EEVs for short). The exogenous variation induced by excluded instrumental variables provides separate variation in the residuals (or generalized residuals) obtained from a reduced form, and these residuals serve as the control functions. By adding appropriate control functions, which are usually estimated in a first stage, the EEVs become appropriately exogenous in a second-stage estimating equation. The purpose of this review is to show how this general description of the CF approach can be applied to various linear and nonlinear models.

In evaluating the scope of an estimation method, it is important to understand how it works in familiar settings, including cases when it is not necessarily needed. Consequently, in Section II, I discuss the control function approach applied to linear models with constant coefficients. Such models are still the workhorse in applied econometrics, and simple IV methods, such as two-stage least squares (2SLS), are usually sufficient for estimation. Nevertheless, even when the CF approach is identical to a 2SLS estimator, the CF perspective has a couple of attractive features. First, the CF approach produces a simple Hausman (1978) test that compares OLS and 2SLS, and the test is easily made robust to heteroskedasticity and cluster correlation (including serial correlation in a panel data setting) of unknown form. Second, the CF approach parsimoniously handles fairly complicated models that are nonlinear in endogenous explanatory variables.

In Section III, I turn to random coefficient models, where the partial effects of the endogenous explanatory variables can vary across individual units in unobserved ways. Estimating such models has fallen somewhat out of favor in empirical research yet they (implicitly) play a role in the recent program evaluation literature, where treatment effects are assumed to be heterogeneous. In the last decade or so, the focus in the treatment effects literature has been on quantities that are identified using stan-

dard IV methods under general assumptions, with the "local average treatment effect" (LATE) introduced in Imbens and Angrist (1994) being the most popular.

I argue in Section III that the control function approach is a useful complement to standard IV methods for a couple of reasons. First, we might hope to estimate treatment effects for identifiable populations or subpopulations, and the CF approach allows us to do that under certain assumptions. Second, the CF approach allows us to study the nature of self-selection. In particular, if we think units self-select into treatment when the treatment is likely to be beneficial, then we should be able to test that proposition. A classic example is the endogenous switching regression model, as in Heckman (1976), which is often applied to earnings equations under two different regimes (such as belonging to a union or not).

I discuss estimation of nonlinear models in Section IV, where the CF approach is particularly appealing compared with other approaches such as "plug-in" methods or joint maximum likelihood. Important contributions are Rivers and Vuong (1988), which developed a two-step CF method for estimation of a probit model with a continuous EEV, and Smith and Blundell (1986), which essentially did the same for the Tobit model. In these early applications of CF methods to nonlinear models, the focus was on parameter estimation. Many of the recent advances in CF methods demonstrate that average partial effects, or average causal effects, are identified quite generally. Wooldridge (2010) uses these results extensively and, in pioneering work, Blundell and Powell (2003) shows that the concept of a control function can be applied in nonparametric and semiparametric contexts. For discrete EEVs, I also summarize the CF methods recently proposed by Terza, Basu, and Rathouz (2008) and Wooldridge (2014). These methods are more controversial because they rely on nonstandard parametric assumptions.

To illustrate several of the CF methods, I present applications to three data sets. One data set allows estimation of a log wage equation allowing for education to be endogenous. Such equations can be estimated assuming a constant return to education, as in Section II, or the return to education can be individual-specific, as in Section III. A second application is to a math test score equation, where the EEV of interest is a binary indicator of attending a Catholic high school. Again, one can assume constant coefficients or allow the effect of attending a Catholic high school to depend on unobserved characteristics. As we will see in Section III, there is strong evidence for individual-specific heterogeneity in the effects of attending a Catholic high school.

To illustrate nonlinear models in Section IV, I use a data set on married women's labor force participation, where the variable measuring other sources of income is treated as endogenous. I show how to estimate the simple Rivers-Vuong model and also show how the CF approach can be made much more flexible with almost no additional computation. Finally, I briefly consider a binary response model for graduating from high school with attending a Catholic high school as the endogenous explanatory variable. The data sets and Stata® code used for all models estimated in the paper are available on request from the author.

II. Models Linear in Constant Coefficients

I begin with a standard linear model with constant coefficients for several reasons. The first is to show that a very common estimation method, two stage least squares (2SLS), can be derived using the control function approach. Second, the

control function (CF) approach leads to robust, regression-based Hausman tests of whether the suspected EEVs are actually endogenous. Third, the basic 2SLS approach can be contrasted with CF approaches that put structure on the reduced forms of the endogenous explanatory variables. CF approaches that use more information can improve precision of the estimates but are generally less robust.

I consider a setting where y_1 is a scalar response variable, y_2 is the endogenous explanatory variable (also a scalar for simplicity), and $\mathbf{z}$ is the $1 \times L$ vector of exogenous variables, which I assume contains unity to allow for a nonzero intercept. The "structural" equation in the population is

$$(1) \quad y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \gamma_1 y_2 + u_1,$$

where $\mathbf{z}_1$ includes unity and is a $1 \times L_1$ subvector of $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2)$. The sense in which $\mathbf{z}$ is exogenous is given by the L orthogonality (zero covariance) conditions

$$(2) \quad E(z_j' u_1) = 0, \quad j = 1, 2, ..., K.$$

The assumptions in Equation 2 hold if I make the stronger assumption $E(u_1|\mathbf{z}) = 0$, which is sometimes preferred if Equation 1 is supposed to be a structural equation—but I first derive the CF approach under the same assumption employed by 2SLS, which is Equation 2.

I make the standard assumption that the elements of $\mathbf{z}$ are not perfectly collinear. In addition, I assume the rank condition for identification holds. In the context of Model 1, the rank condition is most easily stated in terms of the linear reduced form for y_2. If I write

$$(3) \quad y_2 = \mathbf{z}\boldsymbol{\pi}_2 + v_2 = \mathbf{z}_1\boldsymbol{\pi}_{21} + \mathbf{z}_2\boldsymbol{\pi}_{22} + v_2$$

$$(4) \quad E(\mathbf{z}'v_2) = \mathbf{0},$$

then the rank condition holds if and only if $\boldsymbol{\pi}_{22} \neq \mathbf{0}$. This is just the usual requirement that there be at least one exogenous variable that is omitted from Equation 1 that is partially correlated with y_2. As is now widely appreciated, given a random sample, one should estimate the reduced from in Equation 3 and be able to reject the null $H_0 : \boldsymbol{\pi}_{22} = \mathbf{0}$ at a suitably small significance level.

A leading example of the above setup is when y_1 is the logarithm of hourly earnings, y_2 is a measure of schooling, and $\mathbf{z}_1$ includes other determinants of wages that are assumed to be exogenous (such as workforce experience). Many instrumental variables have been proposed for schooling in the literature, ranging from parents' education to quarter of birth; Card (2001) includes a survey of some of the more convincing efforts.

As a second example, suppose y_1 is performance on a standardized test and y_2 is a binary indicator of attending a Catholic high school, a problem studied by Altonji, Elder, and Taber (2005), among others. In thinking about the scope of the model in Equation 1, it is important to understand that it allows for y_2 to be continuous or discrete (or some mixture). The linear reduced form in Equation 3 under Condition 4 can always be specified regardless of the nature of y_2. I need provide no structural interpretation of this equation. I simply need $\mathbf{z}_2$ to be correlated with y_2 after partialling out $\mathbf{z}_1$.

The CF approach based on Equations 1 through 4 proceeds by noting that correlation between the structural error, u_1, and the reduced form error, v_2, can be captured using a linear relationship:

(5) $u_1 = \rho_1 v_2 + e_1$

(6) $E(v_2 e_1) = 0,$

where $\rho_1 = E(v_2 u_1) / E(v_2^2)$ is the population regression coefficient. Because u_1 and v_2 are uncorrelated with $\mathbf{z}$, it follows that e_1 is also uncorrelated with $\mathbf{z}$, and then e_1 must also be uncorrelated with y_2. Therefore, I obtain a valid estimating equation by plugging Equation 5 into the structural equation to get

(7) $y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \gamma_1 y_2 + \rho_1 v_2 + e_1.$

In the CF approach, one views v_2 as an explanatory variable in Equation 7. By including it, one obtains a new error term, e_1, that is uncorrelated with all other righthand-side variables, including y_2. In effect, including v_2 in the equation "controls for" the endogeneity of y_2. One can think of v_2 as proxying for the factors in u_1 that are correlated with y_2.

A. Control Function Procedure: Linear Reduced Form

If one could observe v_2 along with the other variables, Equation 7 immediately suggests a way to estimate $\boldsymbol{\delta}_1$, γ_1, and ρ_1: Run the OLS regression of y_{i1} on $\mathbf{z}_{i1}$, y_{i2}, and v_{i2} using a random sample of size N. The only problem is that one does not observe v_2. Nevertheless, from Equation 3, one can write $v_2 = y_2 - \mathbf{z}\boldsymbol{\pi}_2$ and, because data is collected on y_2 and $\mathbf{z}$, one can consistently estimate $\boldsymbol{\pi}_2$ by OLS. This leads to the following two-step control function procedure:

1. Run the OLS regression of the EEV, y_{i2}, on all exogenous variables, $\mathbf{z}_i$,

(8) y_{i2} on $\mathbf{z}_i$, $\quad i = 1, \ldots, N$

and obtain the OLS residuals, $\hat{v}_{i2}$.

2. Run the OLS regression

(9) y_{i1} on $\mathbf{z}_{i1}$, y_{i2}, $\hat{v}_{i2}$, $\quad i = 1, \ldots, N$

to obtain $\hat{\boldsymbol{\delta}}_1$, $\hat{\gamma}_1$, and $\hat{\rho}_1$.

It has been known since at least Hausman (1978) that this CF method produces coefficients on $\mathbf{z}_{i1}$ and y_{i2} that are numerically identical to the 2SLS estimates. Therefore, one might wonder what all the fuss is about. It is true that, in this particular setting, the CF approach does not lead to a new estimator. In fact, obtaining proper standard errors from the regression in Equation 9 is made difficult by the first-stage estimation of $\boldsymbol{\pi}_2$. Nevertheless, compared with the 2SLS approach, the inclusion of $\hat{v}_{i2}$ serves a valuable purpose: It produces a heteroskedasticity-robust Hausman test of the null hypothesis $H_0 : \rho_1 = 0$, which means y_2 is actually exogenous. The traditional form of the Hausman test that directly compares OLS and 2SLS is substantially harder to make robust to heteroskedasticity.

The importance of the identification requirement that $\boldsymbol{\pi}_{22} \neq \mathbf{0}$ can be seen by studying Equations 3 and 7. If $\boldsymbol{\pi}_{22} = \mathbf{0}$, then v_2 is a linear function of y_2 and $\mathbf{z}_1$, which means v_2 is collinear in Equation 7. The presence of $\mathbf{z}_2$ that is partially correlated with y_2 ensures v_2 has variation separate from $(\mathbf{z}_1, y_2)$. If there are no variables $\mathbf{z}_2$ then the CF regression in Equation 9 suffers from perfect multicollinearity in the sample, and estimates of all parameters cannot be produced. These same observations apply to more complicated CF procedures covered later.

I illustrate the CF approach using two data sets, one a wage data set used to estimate

Table 1
Estimates of the log(wage) Equation

Explanatory Variable	1 OLS	2 2SLS	3 2SLS	4 CF	5 CF
educ	0.0747	0.157	0.161	0.153	0.151
	(0.0036)	(0.052)	(0.054)	(0.048)	(0.048)
exper	0.0848	0.119	0.120	0.116	0.115
	(0.0068)	(0.023)	(0.024)	(0.021)	(0.021)
$exper^2$	–0.0023	–0.0024	–0.0024	–0.0022	–0.0022
	(0.0003)	(0.0004)	(0.0005)	(0.0003)	(0.0003)
black	–0.119	–0.123	–0.121	–0.107	–0.105
	(0.018)	(0.051)	(.062)	(0.048)	(0.048)
$black \cdot (educ - \overline{educ})$	—	—	–0.0008	0.018	0.019
			(0.0408)	(0.006)	(0.006)
$\hat{v}_2$	—	–0.082	—	–0.082	–0.106
		(0.048)		(0.048	(0.050)
$\hat{v}_2 \cdot educ$	—	—	—	—	0.0019
					(0.0010)
intercept	4.62	3.24	3.17	3.31	3.33
	(0.07)	(0.88)	(0.91)	(0.81)	(0.81)
Observations	3,010	3,010	3,010	3,010	3,010

Notes: (i) Each equation contains dummy variables for living in an SMSA and living in the South. In addition, they include regional dummies for where the man was living in 1966 and an indicator of whether the man lived in an SMSA in 1966.
(ii) Standard errors for OLS and 2SLS are robust to heteroskedasticity.
(iii) In Column 2, the 2SLS estimates are equivalent to the CF estimates.
(iv) The standard errors for the CF estimates in Columns 4 and 5 are based on 1,000 bootstrap replications.

the return to schooling in Card (1995) and the other a subset of the data on student performance and Catholic school attendance from Altonji, Elder, and Taber (2005) (hereafter, AET). In both cases, the authors provide detailed discussions about the exogeneity of the instruments, and AET casts doubt on the exogeneity of a commonly used distance instrument. Nevertheless, I proceed as if the instruments are exogenous.

I begin with a standard wage equation

$$lwage = \mathbf{z}_1\boldsymbol{\delta}_1 + \gamma_1 educ + u_1,$$

where *lwage* is the log of wage and $\mathbf{z}_1$ contains exogenous variables and a constant. Years of schooling, *educ*, can be correlated with u_1 for many reasons, such as omitted ability and measurement error. Rather than estimate γ_1 by OLS, I can try to find one or more instrumental variables for *educ*. Card (1995) uses two dummy variables indicating whether there is a two-year college (*nearc*2) or four-year college (*nearc*4) in the local labor market at age 16. Details of the data are described in Card (1995).

Table 1 reports several estimates. The first column contains the OLS estimates with the controls used in Card (1995). The return to a year of schooling is estimated to be

Table 2
Estimates of the math12 *Equation*

Explanatory Variable	1 OLS	2 2SLS	3 CF	4 2SLS	5 CF	6 CF
cathhs	1.49	2.36	1.59	2.06	2.30	–0.95
	(0.39)	(1.25)	(1.07)	(1.63)	(1.19)	(1.75)
motheduc	0.714	0.713	0.714	0.620	0.714	0.709
	(0.062)	(0.062)	(0.062)	(0.077)	(0.064)	(0.062)
fatheduc	0.893	0.887	0.893	0.908	0.886	0.876
	(0.056)	(0.057)	(0.057)	(0.071)	(0.058)	(0.058)
lfaminc	1.84	1.82	1.84	1.87	1.90	1.86
	(0.14)	(0.14)	(0.14)	(0.18)	(0.15)	(0.15)
$cathhs \cdot (motheduc - \overline{motheduc})$	—	—	—	1.61	–0.077	–0.085
				(0.73)	(0.262)	(0.262)
$cathhs \cdot (fatheduc - \overline{fatheduc})$	—	—	—	–0.198	0.089	0.184
				(0.684)	(0.235)	(0.238)
$cathhs \cdot (lfaminc - \overline{lfaminc})$	—	—	—	–0.688	–1.10	–0.691
				(2.082)	(0.61)	(0.634)
$\hat{r}_2$	—	—	–0.061	—	–0.290	–1.52
			(0.594)		(0.632)	(0.80)
$\hat{r}_2 \cdot cathhs$	—	—	—	—	—	3.31
						(1.31)
intercept	11.20	11.45	11.23	11.87	10.72	11.18
	(1.25)	(1.29)	(1.28)	(1.62)	(1.37)	(1.38)
Observations	7,444	7,444	7,444	7,444	7,444	7,444

Notes: (i) Standard errors for OLS and 2SLS are robust to heteroskedasticity.
(ii) The standard errors for the CF estimates are based on 1,000 bootstrap replications.
(iii) Using the estimates from Column 6, the average treatment effect on the treated is 3.99 ($t = 2.96$) and the average treatment effect on the untreated is –1.27 ($t = -0.73$).

0.075 ($t = 20.48$). Column 2 contains the 2SLS estimates reported in control function form with the reduced form residual included. The heteroskedasticity-robust *t*statistic on $\hat{v}_2$ is –1.72, which is a marginal rejection of the null that education is exogenous — even though the 2SLS point estimate for the return to education (0.157, $t = 3.00$) is much higher than the OLS estimate.

For the AET application, I estimate the model

$$math12 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 cathhs + u_1,$$

where $\mathbf{z}_1$ includes an intercept, mother's education, father's education, and the log of family income. The instruments for *cathhs*, which is a binary indicator for attending a Catholic high school, is distance from the nearest Catholic high school divided into five bins. Thus, four distance dummies are used as IVs for *cathhs*. The OLS and 2SLS estimates are given in Columns 1 and 2 of Table 2. The OLS estimate on *cathhs* is about 1.49 ($t = 3.84$), or about 0.16 standard deviations in the test score. The 2SLS estimate is 2.36 ($t = 1.90$). However, the heteroskedasticity-robust test statistic on the

control function (not reported in the table) is only $t = -0.75$, so the OLS and 2SLS estimates are not statistically different.

B. Exploiting a Binary EEV

The test score example raises an interesting question because the EEV, *cathhs*, is binary. The standard IV approach treats all EEVs the same: The structural equation is supplemented with the linear reduced form given in Equations 3 and 4. An alternative is to recognize the binary nature of y_2 and replace its linear reduced form with a binary response model. The two equations are then

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \gamma_1 y_2 + u_1 \tag{10}$$

$$y_2 = 1[\mathbf{z}\boldsymbol{\delta}_2 + e_2 > 0], \tag{11}$$

where $1[\cdot]$ is the binary indicator function. With these equations, one would assume that (u_1, e_2) is independent of $\mathbf{z}$, which is already much stronger than the zero correlation assumptions used by the previous CF (2SLS) estimator. If it is assumed that u_1 is linearly related to e_2 and that

$$e_2 \sim Normal(0,1), \tag{12}$$

then one can derive an alternative control function method. An implication of Equations 11 and 12 is that y_2 follows a probit model:

$$P(y_2 = 1|\mathbf{z}) = \Phi(\mathbf{z}\boldsymbol{\delta}_2), \tag{13}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Nothing was assumed of the sort to apply 2SLS to Equation 1.

A thorough understanding of the pros and cons of different CF approaches requires one to understand that the model for y_2 in Equations 11 and 12 is entirely compatible with the linear reduced form defined by Equations 3 and 4. The usual 2SLS approach assumes nothing about the distribution of y_2 given $\mathbf{z}$. By contrast, Equations 11 and 12 completely characterize the distribution of y_2 given $\mathbf{z}$.

C. Control Function Procedure: Probit Reduced Form

When one specifies a full distribution for y_2, the CF approach is based on the conditional expectation $E(y_1|\mathbf{z}, y_2)$. This is a subtle difference with the 2SLS approach, which is based on zero correlation assumptions only. It is well known—see, for example, Wooldridge (2010, Section 21.4.2)—that under the previous assumptions,

$$E(y_1|\mathbf{z}, y_2) = \mathbf{z}_1\boldsymbol{\delta}_1 + \gamma_1 y_2 + \eta_1[y_2\lambda(\mathbf{z}\boldsymbol{\delta}_2) - (1 - y_2)\lambda(-\mathbf{z}\boldsymbol{\delta}_2)], \tag{14}$$

where $\lambda(\cdot) = \phi(\cdot) / \Phi(\cdot)$ is the well-known inverse Mills ratio. The function

$$r(y_2, \mathbf{z}\boldsymbol{\delta}_2) \equiv y_2\lambda(\mathbf{z}\boldsymbol{\delta}_2) - (1 - y_2)\lambda(-\mathbf{z}\boldsymbol{\delta}_2) \tag{15}$$

is sometimes called a "generalized error" because it has a mean of zero conditional on $\mathbf{z}$.

1. Estimate the probit model in Equation 13. Obtain the "generalized residuals"

$$\hat{r}_{i2} \equiv y_{i2}\lambda(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2) - (1 - y_{i2})\lambda(-\mathbf{z}_i\hat{\boldsymbol{\delta}}_2), \quad i = 1, \ldots, N. \tag{16}$$

2. Run the OLS regression

$$y_{i1} \text{ on } \mathbf{z}_{i1}, y_{i2}, \hat{r}_{i2}, \quad i = 1, \ldots, N \tag{17}$$

to consistently estimate $\boldsymbol{\delta}_1$, γ_1, and η_1.

As with the first CF approach—the one that produces 2SLS—a simple test of the null hypothesis that y_2 is exogenous is obtained as the (heteroskedasticity-robust) t statistic on $\hat{r}_{i2}$.

The CF approach from Regression 17 is the same one computed by the "treatreg" command in Stata® using its two-step option. It exploits the binary nature of y_2 but not without cost. For one, it is generally inconsistent if the probit model for y_2 is misspecified. This is in contrast to the usual 2SLS estimator—equivalently, the CF estimator from Equation 9. The robustness of the 2SLS estimator compared with the estimator from Equation 17 is perhaps counterintuitive and has generated some confusion among empirical researchers. The key is that 2SLS does not use any distributional assumptions in the reduced form whereas the expression in Equation 14 does. If the probit model for y_2 is correctly specified, then the CF procedure in Equation 17 and 2SLS should give estimates that differ only due to sampling error.

Column 3 in Table 2 contains the CF estimates obtained from Equation 17 for the *math*12 equation. The *cathhs* coefficient is 1.59, which is close to the OLS estimate. This is expected because the t statistic on $\hat{r}_{i2}$ is only –0.10. If the coefficient on the generalized residual were statistically significant, one should adjust the standard errors for the two-step estimation. The bootstrap can be used if analytical methods are not readily available. Given the three estimates so far—OLS, 2SLS (the CF estimates from Equation 9), and the CF estimates from Equation 17—there is no reason to reject the OLS estimate. This is not the case when one turns to a richer set of models.

D. Models Nonlinear in the EEV

One benefit of the CF approaches in Equations 9 and 17 is that they are easily adapted to handle more complicated models. As one important example, consider a model where y_2 interacts with the exogenous variables (and appears on its own because $\mathbf{z}_1$ includes an intercept):

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + y_2\mathbf{z}_1\boldsymbol{\gamma}_1 + u_1. \tag{18}$$

If y_2 is continuous, then one can use the regression in Equation 9 where $y_{i2}\mathbf{z}_{i1}$ replaces y_{i2}, which means y_{i2} appears on its own and interacted with exogenous variables. If y_{i2} is binary, then one uses Regression 17, where y_{i2} appears by itself and interacted with exogenous variables. The t statistic on either $\hat{v}_{i2}$ or $\hat{r}_{i2}$, perhaps made robust to heteroskedasticity, is still a valid test of the null that y_2 is exogenous. Because Equation 18 contains only a single EEV, a one degree-of-freedom test is appealing.

A standard IV approach to estimating Equation 18 would require choosing IVs for the L_1 terms in $y_2\mathbf{z}_1$. For example, I can add interactions of elements in the excluded exogenous variables, $\mathbf{z}_2$, with $\mathbf{z}_1$. When y_2 is binary, there are other natural choices for IVs, as discussed in Wooldridge (2010, Section 21.3): Use the interactions $\Phi(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)\mathbf{z}_{i1}$,

where $\Phi(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$ are the probit fitted values. This IV estimator has a theoretical advantage over the CF estimator, at least if one assumes the linear model with constant coefficients is the correct specification: The IV estimator is generally consistent even if the probit model is misspecified. Thus, one can exploit the binary nature of y_2 but still obtain an estimator that does not require a correctly specified model for $D(y_2|\mathbf{z})$, the distribution of y_2 given $\mathbf{z}$. However, the CF approach offers a parsimonious way to account for endogeneity of y_2 even if it interacts with many exogenous variables. It seems likely that it is more efficient quite generally, but this possibility seems not to have been systematically investigated.

In Columns 3 and 4 of Table 1, I include an interaction between the race indicator, *black*, and *educ*, where I first center *educ* about its mean (roughly 13.3) before creating the interaction. Column 3 contains the 2SLS estimates where *nearc*2, *nearc*4, *black* · *nearc*2, and *black* · *nearc*4 are used as instruments for *educ* and *black* · (*educ* – *educ*). The coefficient on the latter is –0.0008, which is small and has a very wide 95 percent confidence interval. Column 4 contains the CF approach, and now the coefficient on the interaction term is positive and practically large, 0.018, and statistically significant with $t = 2.84$. The return to education is estimated to be about 1.8 percentage points higher for black men. Further, the earnings gap between black and nonblack men shrinks at high levels of education. The picture given by the CF estimates is different from the much less precise 2SLS estimates.

In the test score equation, I interact *cathhs* with *motheduc*, *fatheduc*, and *lincome*. Column 4 in Table 2 contains the estimates where the fitted probit probabilities and interactions are used as IVs, while Column 5 contains the CF estimates from adding the generalized residual. The estimates are notably different. The CF estimate of the average effect of *cathhs* is 2.30 ($t = 1.94$) and the interaction terms are all small and insignificant (although the interaction with *lincome* has $t = -1.79$). By contrast, the average effect estimated by 2SLS is insignificant but there appears to be a large, statistically significant interaction with mother's education. I cannot reconcile the difference in these estimates without allowing the treatment effect of *cathhs* to depend on unobservables, which I do in the next section.

Before ending this section, it is useful to summarize the key points of how the CF approach compares with other common approaches.

1. In the basic linear model with constant coefficients, where the EEV appears linearly, and where I use linear reduced forms, the CF approach is the same as 2SLS. The CF approach provides a simple, robust test of the null hypothesis that y_2 is exogenous.

2. When I exploit special features of the EEV y_2—for example, recognize that it is a binary variable—the CF approach uses generalized residuals. The CF approach is likely more efficient than 2SLS because it exploits the binary nature of y_2 but, in terms of consistency, the CF approach is usually less robust than IV approaches.

3. In models with multiple, nonlinear functions of EEVs, the CF approach parsimoniously handles endogeneity and provides simple exogeneity tests. For general nonlinearities, inserting fitted values for EEVs is generally inconsistent, even under strong assumptions. The IV approach, where nonlinear functions of exogenous variables are specified as instruments, is the most robust in terms of consistency, but in a model such as Equation 18 it treats any function of the EEVs as a separate endogenous variable; therefore, it can be quite inefficient relative to the more parsimonious CF approach.

III. Correlated Random Coefficient Models

The setup of the previous section allows the endogenous explanatory variable or variables to appear linearly or nonlinearly and to interact with observed covariates. This may be sufficient for some applications, but one may also want to allow the effect of y_2 to depend on unobservables. One might think, for example, that the return to schooling or the causal effect of attending a Catholic high school vary across individuals in ways that cannot be observed fully. When one allows random coefficients to be correlated with some explanatory variables, such as amount of school or choice of school, one obtains a "correlated random coefficient" (CRC) model, a label adopted by Heckman and Vytlacil (1998) and discussed in the context of the return to schooling by Card (2001). In the treatment effects literature, CRC models allow for heterogeneous treatment effects combined with self-selection into treatment—provided that there are suitable instrumental variables for treatment assignment.

Consider the problem of estimating a wage equation with an individual-specific return to schooling. For a random draw i,

$$(19) \quad lwage_i = \mathbf{z}_{i1}\boldsymbol{\delta}_1 + g_{i1}educ_i + u_{i1},$$

where g_{i1} is the individual-specific return to schooling. Now there are two sources of unobserved heterogeneity and both g_{i1} and u_{i1} might be correlated with $educ_i$. In fact, due to self-selection, one might expect the amount of education, $educ_i$, to be positively correlated with g_{i1}: people for whom the return to schooling is higher will choose, on average, to obtain more education.

Certainly, one cannot expect to estimate g_{i1} for each i. Instead, I focus on the average return to schooling in the population, $\gamma_1 = E(g_{i1})$. Then I can write $g_{i1} = \gamma_1 + v_{i1}$ where $E(v_{i1}) = 0$. Plugging into Equation 19 gives

$$(20) \quad lwage_i = \mathbf{z}_{i1}\boldsymbol{\delta}_1 + \gamma_1 educ_i + v_{i1}educ_i + u_{i1}.$$

If I apply the usual 2SLS estimator to Equation 20, then the error term is implicitly $e_{i1} = v_{i1}educ_i + u_{i1}$. As discussed in Wooldridge (2003), the 2SLS estimator is generally inconsistent for γ_1, although it is consistent if one assumes, in addition to the standard exogeneity requirements

$$(21) \quad E(u_{i1}|\mathbf{z}_i) = 0, E(v_{i1}|\mathbf{z}_i) = 0,$$

a constant conditional covariance assumption:

$$(22) \quad Cov(educ_i, v_{i1}|\mathbf{z}_i) = Cov(educ_i, v_{i1}).$$

Notice that Condition 22 allows arbitrary correlation between $educ_i$ and the random return to education, g_{i1}. But the conditional covariance cannot depend on the exogenous variables. Card (2001) discusses situations where this assumption is likely to fail in simple models of schooling decisions.

A control function approach is based on similar assumptions but has the added advantage of allowing estimation of a relationship between the level of education and the return to education. The method is due to Garen (1984), although one can relax the normality assumptions it uses. To describe the approach generally, let y_{i1} and y_{i2} be the endogenous variables, as before, and assume that

(23) $y_{i2} = \mathbf{z}_i\boldsymbol{\pi}_2 + v_{i2}, E(v_{i2}|\mathbf{z}_i) = 0.$

Assume also that both sources of unobservables, u_{i1} and v_{i1}, are linearly related to v_{i2}:

(24) $E(u_{i1}|v_{i2}) = \eta_1 v_{i2}, \; E(v_{i1}|v_{i2}) = \psi_1 v_{i2}$

and that all unobservables are independent of $\mathbf{z}_i$. The estimating equation is

(25) $E(y_{i1}|\mathbf{z}_i, y_{i2}) = E(y_{i1}|\mathbf{z}_{i1}, y_{i2}, v_{i2}) = \mathbf{z}_{i1}\boldsymbol{\delta}_1 + \gamma_1 y_{i2} + \eta_1 v_{i2} + \psi_1 v_{i2} y_{i2}.$

Equation 25 leads to the following simple CF approach. As before, estimate Equation 23 by OLS and obtain the residuals, $\hat{v}_{i2}$. Second, run the OLS regression

(26) y_{i1} on $\mathbf{z}_{i1}, y_{i2}, \hat{v}_{i2}, \hat{v}_{i2}y_{i2}, \quad i = 1, ..., N.$

Compared with the constant-coefficient case, I have added the interaction term $\hat{v}_{i2}y_{i2}$. Without the interaction, I know that Regression 26 produces the 2SLS estimates of $\boldsymbol{\delta}_1$ and γ_1. The interaction term accounts for the random coefficient on y_{i2}. It is of interest to test for statistical significance of the interaction term, but one must be careful: If the coefficient on v_{i2} is different from zero, the usual t statistic on $y_{i2}\hat{v}_{i2}$ is not valid because of the first-stage estimation. It is simple to bootstrap the two-step procedure to obtain valid standard errors for all of the coefficients. Conveniently, a test of joint significance of $(\hat{v}_{i2}, \hat{v}_{i2}y_{i2})$ is valid without adjusting the standard errors. The joint test is a test of the null hypothesis that y_2 is exogenous.

Given the results on 2SLS by Wooldridge (2003) described earlier, it is possible that the coefficient on $\hat{v}_{i2}y_{i2}$, $\hat{\psi}_1$, is large and statistically significant but the estimate of γ_1 is similar to the 2SLS estimate. Even if the two procedures give similar estimates of the average effect, $\hat{\psi}_1$ is of some interest because one can write

(27) $E(g_{i1}|v_{i2}) = \gamma_1 + \psi_1 v_{i2}.$

Even though I cannot estimate g_{i1}, I can estimate its expected value given the reduced form error, v_{i2}, which necessarily has a zero mean. In the return-to-schooling example, I might expect $\psi_1 > 0$ because, as v_{i2} increases, the person has more education than is predicted by the exogenous variables, $\mathbf{z}_i$. A positive ψ_1 is consistent with a selection story: conditional on $\mathbf{z}_i$, people obtain more education if their return to schooling is higher. One can estimate the righthand side of Equation 27 as $\hat{\gamma}_1 + \hat{\psi}_1\hat{v}_{i2}$ for each i and, if desired, study how these estimates vary across i. The average of the individual partial effects in the sample is, mechanically, $\hat{\gamma}_1$.

As with the simpler CF method from Section II, Regression 26 easily extends to allow any nonlinear functions of $(\mathbf{z}_{i1}, y_{i2})$, including quadratics and interactions. I estimate the wage equation using the Card (1995) data by including the interaction $black_i \cdot (educ_i - \overline{educ})$ along with $\hat{v}_{i2}$ and $\hat{v}_{i2} \cdot educ_i$; the results are in Column 5 of Table 1. The estimates on the $educ_i$, $black_i$, and $black_i \cdot (educ_i - \overline{educ})$ are similar to the CF estimates without the interaction term $\hat{v}_{i2} \cdot educ_i$, even though the latter is marginally significant ($t = 1.84$), revealing a certain robustness of the simpler CF approach. (Jointly, $\hat{v}_{i2}$ and $\hat{v}_{i2} \cdot educ_i$ are significant with p–value = 0.042). From Equation 27, the positive coefficient on $\hat{v}_{i2} \cdot educ_i$ implies that those with higher-than-predicted education have, on average, higher returns to schooling, thereby providing some evidence for self-selection into schooling.

Using the CF approach, I do not have to stop at interaction terms between observed variables or even just one random coefficient. A very general correlated random effects

analysis is obtained by choosing a $1 \times K_1$ set of regressors, $\mathbf{x}_{i1}$, to be any function of $(\mathbf{z}_{i1}, y_{i2})$, say $\mathbf{g}_1(\mathbf{z}_{i1}, y_{i2})$. This can include, in addition to $\mathbf{z}_{i1}$ and y_{i2}, terms such as y_{i2}^2 and $\mathbf{z}_{i1}y_{i2}$, or even higher order polynomials and interactions. If one separates out an intercept and allow all K_1 elements of $\mathbf{x}_{i1}$ to have random slopes $\mathbf{b}_{i1} = \boldsymbol{\beta}_1 + \mathbf{v}_{i1}$, one can write

$$(28) \quad y_{i1} = \alpha_1 + \mathbf{x}_{i1}\boldsymbol{\beta}_1 + u_{i1} + \mathbf{x}_{i1}\mathbf{v}_{i1},$$

where both u_{i1} and $\mathbf{x}_{i1}\mathbf{v}_{i1}$ are unobserved. After obtaining the reduced form residuals $\hat{v}_{i2}$ from OLS of y_{i2} on $\mathbf{z}_i$, the CF regression is simply

$$(29) \quad y_{i1} \text{ on } 1, \ \mathbf{x}_{i1}, \hat{v}_{i2}, \mathbf{x}_{i1}\hat{v}_{i2}.$$

So, I add $\hat{v}_{i2}$ and also interact all or some elements of $\mathbf{x}_{i1}$ with $\hat{v}_{i2}$. As before, it is simple to use the nonparametric bootstrap, where both estimation steps are included, to obtain valid inference. If $\hat{\boldsymbol{\psi}}_1$ is the K_1 vector of OLS coefficients on $\mathbf{x}_{i1}\hat{v}_{i2}$, one can estimate $E(\mathbf{b}_{i1}|v_{i2})$ as $\hat{\boldsymbol{\beta}}_1 + \hat{v}_{i2}\hat{\boldsymbol{\psi}}_1$ and possibly provide economic interpretations for the signs and magnitudes of the elements of $\hat{\boldsymbol{\psi}}_1$.

Even more flexibility is obtained by allowing $E(\mathbf{v}_{i1}|v_{i2})$ to be a nonlinear function in v_{i2}, such as $E(\mathbf{v}_{i1}|v_{i2}) = v_{i2}\boldsymbol{\psi}_1 + (v_{i2}^2 - \tau_2^2)\boldsymbol{\eta}_1$, where $\tau_2^2 = E(v_{i2}^2)$. Then the terms $\hat{v}_{i2}^2$ and $\mathbf{x}_{i1} \cdot (\hat{v}_{i2}^2 - \hat{\tau}_2^2)$, where $\hat{\tau}_2^2$ is the usual OLS variance estimate from the first stage, get added to Equation 29. It is evident that these extensions of Garen's (1984) CF approach allow significant flexibility in correlated random coefficient models.

The CF approach can also be used to estimate the random coefficient model when y_2 is binary. The typical endogenous switching model is

$$(30) \quad y_{i1} = \alpha_1 + \mathbf{z}_{i1}\boldsymbol{\delta}_1 + \gamma_1 y_{i2} + y_{i2}\mathbf{z}_{i1}\boldsymbol{\theta}_1 + u_{i1} + y_{i2}v_{i1}$$

and I combine this with the probit model for y_2, given in Equations 11 and 12, with all unobservables independent of $\mathbf{z}_i$. After obtaining the generalized residuals in Equation 16, the CF regression is

$$(31) \quad y_{i1} \text{ on } 1, \ \mathbf{z}_{i1}, y_{i2}, y_{i2} \cdot (\mathbf{z}_{i1} - \overline{\mathbf{z}}_1), \hat{r}_{i2}, y_{i2}\hat{r}_{i2},$$

where, again, centering $\mathbf{z}_{i1}$ about the sample averages ensures that the coefficient on y_{i2} is the average effect.

The estimates of the switching regression model for the test score data are given in Column 6 of Table 2. These estimates provide a very different picture than either the 2SLS estimates or the CF estimates that ignore the random coefficient on $cathhs_i$. First, the two terms $\hat{r}_{i2}$ and $cathhs_i \cdot \hat{r}_{i2}$ are jointly significant using a heteroskedasticity-robust test with p–value = 0.022. By contrast, when $\hat{r}_{i2}$ is included by itself, its t statistic is only –0.46. The coefficient on $cathhs_i \cdot \hat{r}_{i2}$ is very large, 3.31 with $t = 2.53$, providing evidence that the treatment effect of attending a Catholic high school depends strongly on unobserved heterogeneity. Even more importantly, the average treatment effect in the population is now negative and not statistically different from zero: $\hat{\gamma}_1 = -0.95$, $(t = -0.58)$.

How can one reconcile the estimated average treatment effect in Column 6 of Table 2 with the 2SLS estimate, which, in the model with interactions, is 2.37 ($t = 1.90$)? As is now well known from the work of Imbens and Angrist (1994), the 2SLS estimator can be given a LATE interpretation. Because the instruments are functions of distance to the nearest school, the interpretation is (somewhat loosely) as follows: The 2SLS

estimate is the average treatment effect for those who are induced to attend a Catholic high school because they live near a Catholic high school. This subpopulation can be very different from the overall population, where the effect estimated by the CF approach is not statistically different from zero.

One can shed further light on the difference between the 2SLS and CF estimates by computing the average treatment effect on the treated (ATT) and the average treatment effect on the untreated (ATU); see Imbens and Wooldridge (2009). The simplest way of obtaining these quantities is to estimate separate equations for the control ($y_{i2} = 0$) and treated ($y_{i2} = 1$) groups, in each case by regressing y_{i1} on 1, $\mathbf{z}_{i1}$, $\hat{r}_{i2}$. Then, fitted values from each regression are obtained for *all* observations i, say $\hat{y}_{i1}^{(0)}$ and $\hat{y}_{i1}^{(1)}$, respectively. Then

$$(32) \quad \widehat{ATT} = N_1^{-1}\sum_{i=1}^{N} y_{i2}[\hat{y}_{i1}^{(1)} - \hat{y}_{i1}^{(0)}],$$

which is simply the average in the difference of fitted values over the $y_{i2} = 1$ observations. (See, for example, Heckman, Tobias, and Vytlacil 2003.) Similarly, $\widehat{ATU}$ is the average of $\hat{y}_{i1}^{(1)} - \hat{y}_{i1}^{(0)}$ over the $y_{i2} = 0$ observations. Using the full endogenous switching specification, the estimated ATT (based on 452 students) is about 3.99 ($t = 2.96$). By contrast, the estimated ATU (based on 6,992 students) is about –1.27 ($t = –0.73$). The large difference is another way to illustrate the self-selection into attending a Catholic school: Those who would benefit based on factors unobserved to us are much more likely to attend a Catholic high school. The usual 2SLS estimation of a linear model is necessarily silent on such selection issues because it only estimates the LATE.

The CF regression in Equation 31 can be made even more general to allow random coefficients on some or all of the exogenous variables as well as on the interaction terms. If one takes the vector of explanatory variables to be $\mathbf{x}_{i1} = (\mathbf{z}_{i1}, y_{i2}, \mathbf{z}_{i1}y_{i2})$ and allow randomness in all coefficients $\mathbf{b}_{i1}$, then the CF regression (across all observations) becomes

$$(33) \quad y_{i1} \text{ on } 1,\ \mathbf{z}_{i1},\ y_{i2}, y_{i2} \cdot (\mathbf{z}_{i1} - \overline{\mathbf{z}}_1),\ \hat{r}_{i2}, y_{i2}\hat{r}_{i2},\ \hat{r}_{i2} \cdot (\mathbf{z}_{i1} - \overline{\mathbf{z}}_1),\ y_{i2} \cdot \hat{r}_{i2} \cdot (\mathbf{z}_{i1} - \overline{\mathbf{z}}_1).$$

The coefficient on y_{i2} in this regression is consistent for the average treatment effect. Alternatively, one can run separate regressions for the control and treated groups, where the regressions have the form y_{i1} on $1, \mathbf{z}_{i1}, \hat{r}_{i2}$, and $\hat{r}_{i2}\mathbf{z}_{i1}$. The estimated ATT is still obtained as in Equation 32, but the fitted values are obtained by adding the terms $\hat{r}_{i2}\mathbf{z}_{i1}$ to the separate regressions. As usual, bootstrapping is an attractive way to obtain valid standard errors. In the Catholic high school example, the expanded regression gives the following estimates (not reported in Table 2): $\widehat{ATT} = 3.59 (t = 2.62)$, $\widehat{ATU} = 0.063$ ($t = 0.03$), and $\widehat{ATE} = 0.28$ ($t = 0.14$). Thus, the picture is similar to the switching regression model with constant coefficients: The average treatment effect in the entire population is essentially zero, with a large average treatment effect for the relatively small treated subpopulation.

The CF estimator obtained from Regression 33 allows substantial heterogeneity across individuals — much more than is allowed in typical applications. Its main limitation is that it is based on a linear model for y_1 under zero conditional mean assumptions for the unobservables. Such models are likely to be good approximations when

y_1 is the log of wage or a test score, but linearity is harder to justify if y_1 is discrete or its range is otherwise restricted. I now turn to nonlinear models for y_1.

IV. Nonlinear Models

Control function methods have long been employed for particular nonlinear models, especially probit and Tobit, when the endogenous explanatory variables are continuous. Thanks largely to the work of Blundell and Powell (2003, 2004), the scope of such applications is now much broader. Wooldridge (2005) and Petrin and Train (2010) give several examples of where CF methods can be applied with continuous EEVs. Here, I cover some simple examples that illustrate the flexibility of the CF approach.

A. Continuous EEVs

Probably the leading example of a nonlinear model with continuous EEVs is the probit model, as analyzed in Rivers and Vuong (1988). With a single EEV y_2, the model can be written as

$$y_1 = 1[\mathbf{z}_1\boldsymbol{\delta}_1 + \gamma_1 y_2 + u_1 \geq 0] \tag{34}$$

$$y_2 = \mathbf{z}\boldsymbol{\delta}_2 + v_2, \tag{35}$$

where (u_1, v_2) is bivariate normal with mean zero, $Var(u_1) = 1$, and independent of $\mathbf{z}$. Here, both $\mathbf{z}$ and $\mathbf{z}_1$ include constants with $\mathbf{z}_1$, a strict subset of $\mathbf{z}$. In most cases, the parameters of interest are constant insofar as they index partial effects. As discussed in Wooldridge (2010, Section 15.7.2), the average partial effects are obtained by taking derivatives or changes of

$$E_{u_{i1}}\{1[\mathbf{z}_1\boldsymbol{\delta}_1 + \gamma_1 y_2 + u_{i1} \geq 0]\} = \Phi(\mathbf{z}_1\boldsymbol{\delta}_1 + \gamma_1 y_2), \tag{36}$$

where the notation $E_{u_{i1}}\{\cdot\}$ indicates averaging out the unobservables and treating $(\mathbf{z}_1, y_2)$ as fixed arguments. Equation 36 is an example of what Blundell and Powell (2003) calls an "average structural function," or ASF. In defining the ASF, the observables are taken as fixed arguments and the unobservables are averaged out. Under the assumptions given, the parameters in Equations 34 and 35 and those in the bivariate normal distribution can be estimated using joint MLE, and so the ASF can be estimated as $\Phi(\mathbf{z}_1\hat{\boldsymbol{\delta}}_1 + \hat{\gamma}_1 y_2)$.

For the purposes of the current paper, a control function approach is attractive. The CF approach is based on the following conditional probability; see Wooldridge (2010, Section 15.7.2):

$$P(y_1 = 1|\mathbf{z}, y_2) = P(y_1 = 1|\mathbf{z}_1, y_2, v_2) = \Phi(\mathbf{z}_1\boldsymbol{\delta}_{\eta 1} + \gamma_{\eta 1} y_2 + \rho_{\eta 1} v_2), \tag{37}$$

where $E(u_1|v_2) = \rho_1 v_2$, the η subscript denotes division by $(1 - \rho_1^2\tau_2^2)^{1/2}$, and $\tau_2^2 = Var(v_2)$. The expression in Equation 37 leads to a simple two-step CF estimator for estimating the scaled coefficients. First, the residuals, $\hat{v}_{i2}$, are obtained from the OLS regression of y_{i2} on $\mathbf{z}_i$. Then, the scaled coefficients are consistently estimated from a probit of y_{i1}

on $\mathbf{z}_{i1}, y_{i2}, \hat{v}_{i2}$. The null hypothesis that y_2 is exogenous is easily tested using the usual t statistic on $\hat{v}_{i2}$.

The CF approach appears to have the drawback that it does not estimate the parameters $\boldsymbol{\delta}_1$ and $\boldsymbol{\gamma}_1$ appearing in Equation 36. Fortunately, it turns out that the ASF is easily estimated using the scaled parameters identified by Equation 37. As discussed in Wooldridge (2010, Section 15.7.2), the ASF can be obtained as

$$(38) \quad ASF(\mathbf{z}_1, y_2) = E_{v_{i2}}[\Phi(\mathbf{z}_1\boldsymbol{\delta}_{\eta 1} + \gamma_{\eta 1}y_2 + \rho_{\eta 1}v_{i2})];$$

that is, one averages the control function, v_{i2} out of the conditional probability $P(y_1 = 1|\mathbf{z}_1, y_2, v_2)$. It follows that a consistent estimator of the ASF is

$$(39) \quad \widehat{ASF}(\mathbf{z}_1, y_2) = N^{-1}\sum_{i=1}^{N}\Phi(\mathbf{z}_1\hat{\boldsymbol{\delta}}_{\eta 1} + \hat{\gamma}_{\eta 1}y_2 + \hat{\rho}_{\eta 1}\hat{v}_{i2}),$$

and then I use derivatives or changes with respect to the elements of $(\mathbf{z}_1, y_2)$. After partial effects have been obtained, further averaging can be used, or one can average the partial effects across $(\mathbf{z}_{i1}, y_{i2}, \hat{v}_{i2})$ to obtain a single average partial effect (as is done by the "margins" command in Stata®).

Flexible extensions of the Rivers-Vuong approach can be obtained using the general results of Blundell and Powell (2003, 2004, hereafter, BP), which at its most general level is fully nonparametric. BP assumes a structural model of the form

$$(40) \quad y_1 = g_1(\mathbf{z}_1, y_2, \mathbf{u}_1)$$

for a vector of unobservables $\mathbf{u}_1$ where, for simplicity, y_2 is a scalar. The object of interest in BP is the ASF, defined generally as

$$(41) \quad ASF(\mathbf{z}_1, y_2) \equiv E_{\mathbf{u}_{i1}}[g_1(\mathbf{z}_1, y_2, \mathbf{u}_{i1})];$$

again, the notation means that the unobservables $\mathbf{u}_1$ are averaged out in the population and $\mathbf{z}_1$ and y_2 are fixed values. The ASF can be differentiated with respect to $(\mathbf{z}_1, y_2)$, or discrete differences can be calculated, to obtain average partial effects. Therefore, if one can consistently estimate the ASF, then one can get not only directions of effects but also magnitudes. As is now well known, parameters in nonlinear models often do not deliver magnitudes of partial effects.

A key representation assumed by BP is

$$(42) \quad y_2 = g_2(\mathbf{z}) + v_2,$$

where $(\mathbf{u}_1, v_2)$ is independent of $\mathbf{z}$ [and $E(v_2) = 0$ so that $E(y_2|\mathbf{z}) = g_2(\mathbf{z})$]. It is important to understand that independence between v_2 and $\mathbf{z}$ effectively limits the scope of the BP approach to continuous EEVs. If y_2 is discrete, or its range is restricted in some substantive way, v_2 in Equation 42 cannot be independent of $\mathbf{z}$. Together, Equations 40 and 42 are said to form a "triangular system" because the equation for y_2 does not have y_1 as an explanatory variable. Therefore, if y_1 and y_2 are simultaneously determined, then assuming Equation 42 can be restrictive.

When Equation 42 holds and $(\mathbf{u}_1, v_2)$ is independent of $\mathbf{z}$, the conditional distribution of the unobservables $\mathbf{u}_1$ in the structural function depends on $(\mathbf{z}, y_2)$ only through the reduced-form error, v_2:

$$D(\mathbf{u}_1|\mathbf{z}, y_2) = D(\mathbf{u}_1|\mathbf{z}, v_2) = D(\mathbf{u}_1|v_2). \quad (43)$$

As shown by BP, the ASF can be obtained by using v_2 as a proxy for $\mathbf{u}_1$, in the following sense. First, define the conditional expectation

$$h_1(\mathbf{z}_1, y_2, v_2) \equiv E(y_1|\mathbf{z}_1, y_2, v_2). \quad (44)$$

Then the key result is

$$ASF(\mathbf{z}_1, y_2,) = E_{v_{i2}}[h_1(\mathbf{z}_1, y_2, v_{i2})]. \quad (45)$$

The result in Equation 45 is critical to the CF approach, and it generalizes the probit case in Expression 38. It means that, for obtaining the ASF, it suffices to obtain $E(y_1|\mathbf{z}_1, y_2, v_2)$ and then average out across the population distribution of v_2. For identification purposes, I effectively observe the v_{i2} because $v_{i2} = y_{i2} - g_2(\mathbf{z}_i)$, and $g_2(\cdot)$ is generally identified by $E(y_2|\mathbf{z}) = g_2(\mathbf{z})$.

Let $\hat{g}_2(\cdot)$ be a consistent estimator of $g_2(\cdot)$ and define the reduced-form residuals as

$$\hat{v}_{i2} = y_{i2} - \hat{g}_2(\mathbf{z}_i). \quad (46)$$

A consistent estimator of the ASF, under weak regularity conditions, is

$$\widehat{ASF}(\mathbf{z}_1, y_2) = N^{-1}\sum_{i=1}^{N}\hat{h}_1(\mathbf{z}_1, y_2, \hat{v}_{i2}). \quad (47)$$

Consistent estimates of partial effects are obtained by taking derivatives or changes with respect to the elements in $(\mathbf{z}_1, y_2)$.

Wooldridge (2005) showed that the same analysis goes through if the deterministic equation in Equation 40 is replaced with a conditional mean specification,

$$E(y_1|\mathbf{z}, y_2, \mathbf{u}_1) = E(y_1|\mathbf{z}_1, y_2, \mathbf{u}_1) = g_1(\mathbf{z}_1, y_2, \mathbf{u}_1). \quad (48)$$

Stating the structural model as in Equation 48 allows for some cases that fall outside the BP framework, such as when y_1 is a fractional response or a count response.

A powerful implication of the BP work is that, provided one is interested in the average structural function for y_1 and one can specify a reduced form for y_2 with an additive, independent error, one need not start with a structural model at all. For example, when y_1 is binary case, the parameters in the structural Equation 34 are interesting insofar as they provide directions of effects and enter into the average partial effects. But the scaled coefficients in Equation 37 do just as nicely for getting directions of effects, ratios of coefficients, and average partial effects. In other words, one could start with the probit model in Equation 37 and learn everything desired, including magnitudes of the effects. The insight obtained from the probit model carries over to general situations. By focusing on $E(y_1|\mathbf{z}_1, y_2, v_2)$, I can achieve considerable flexibility even within a parametric framework. Of course, I need at least one exogenous variable that causes variation in y_2 not explained by $\mathbf{z}_1$, and I need to get suitable estimates of v_2.

As an example of how liberating the focus on the APEs can be, consider again the binary response model. Let $\mathbf{x}_1$ be any function of the exogenous and endogenous variables and let v_2 be the error in a reduced form for y_2, probably linear in parameters. Then one can jump directly to specifying flexible models for $P(y_1 = 1|\mathbf{z}_1, y_2, v_2)$, such as

$$P(y_1 = 1|\mathbf{z}_1, y_2, v_2) = \Phi(\mathbf{x}_1\boldsymbol{\beta}_1 + \rho_1 v_2 + \eta_1 v_2^2 + \mathbf{x}_1 v_2 \boldsymbol{\psi}_1). \quad (49)$$

It would be difficult, if not impossible, to derive Equation 49 from an underlying structural equation of the form $y_1 = g_1(\mathbf{z}_1, y_2, u_1)$. Instead, I am skipping the step of specifying a structural model and proceeding directly to estimating Equation 49. A two-step CF method is straightforward. First, obtain the reduced form residuals $\hat{v}_{i2}$ from an initial (flexible) OLS regression. Then, estimate the parameters in Equation 49 using probit of y_{i1} on $\mathbf{x}_{i1}$, $\hat{v}_{i2}$, $\hat{v}_{i2}^2$, $\mathbf{x}_{i1}\hat{v}_{i2}$. Testing the null hypothesis of exogeneity is the same as testing that the last three terms are jointly insignificant. Importantly, there is no need to worry that the coefficients might be scaled versions of underlying structural parameters because the parameters estimated are precisely those that can be used to estimate the ASF:

$$\widehat{ASF}(\mathbf{z}_1, y_2) = N^{-1}\sum_{i=1}^{N}\Phi(\mathbf{x}_1\hat{\boldsymbol{\beta}}_1 + \hat{\rho}_1\hat{v}_{i2} + \hat{\eta}_1\hat{v}_{i2}^2 + \mathbf{x}_1\hat{v}_{i2}\hat{\boldsymbol{\psi}}_1). \tag{50}$$

As before, $\mathbf{x}_1$ is a fixed argument and the averaging out is over the control function, $\hat{v}_{i2}$. With large sample sizes, one can be even more flexible, including higher order polynomials or other transformations in $\hat{v}_{i2}$.

If $\mathbf{x}_1$ includes nonlinear functions of $(\mathbf{z}_1, y_2)$, such as y_2^2 or interactions $\mathbf{z}_1 y_2$, methods where first-stage fitted values are inserted for y_2 do not consistently estimate anything interesting—either parameters or average partial effects. The CF approach has a distinct advantage: If one thinks Equation 49 provides a good approximation to $P(y_1 = 1|\mathbf{z}_1, y_2, v_2)$, then Equation 50 will deliver reliable estimates of the average partial effects.

As an application, consider estimating a binary response model of married women's labor force participation ($y_1 = inlf$). The data, on 5,634 married women, come from the May 1991 Current Population Survey. The EEV is other sources of income, $y_2 = nwifeinc$. I use husband's education (*huseduc*) as an instrument for *nwifeinc*. Other controls are education, experience (as a quadratic), and a dummy variable for having a child under the age of six. The first-stage t statistic on *huseduc* is 18.39; not surprisingly, husband's education is a good predictor of other sources of income.

Table 3 contains estimates of various models, starting with linear probability models estimated by OLS and 2SLS. The OLS coefficient on *nwifeinc* is about –0.0033 (t = –14.14), which implies that another $10,000 in other sources of income reduces the labor force participation probability by 0.033. The IV estimate is substantially smaller in magnitude, –0.0014, and not statistically different from zero (t = –1.42). Columns 3 and 4 contain the estimates for a probit model and the Rivers-Vuong control function approach, respectively. The average partial effect when *nwifeinc* is treated as exogenous is about –0.0033 (t = –14.21), the same as the OLS estimate of the linear probability model to four decimal places. The APE from the CF approach is –0.0015 (t = –1.60), which is very similar to the linear IV estimate. In the probit CF method, the first-stage residual has t = –1.93 and so there is marginal evidence of endogeneity.

Column 5 allows more flexibility by including a squared term in *nwifeinc* and an interaction between having a young child, *kidlt6*, and *nwifeinc*. Like Column 4, the estimates in Column 5 employ only a linear function in $\hat{v}_{i2}$. The square and interaction are both statistically significant, and the CF is now slightly more significant. The coefficients are especially difficult to interpret because of the nonlinearity in the model (including the probit functional form). Using the derivative, the APE of *nwifeinc* is estimated to be –0.00097 (t = –1.00).

Table 3
Estimates of the inlf *Equation*

Explanatory Variable	1 Linear OLS	2 Linear 2SLS	3 Probit MLE	4 Probit CF	5 Probit CF	6 Probit CF
nwifeinc	–0.0033	–0.0014	–0.0091	–0.0042	–0.0001	–0.0025
	(0.0002)	(0.0010)	(0.0010)	(0.0026)	(0.0028)	(0.0028)
educ	0.0350	0.0300	0.100	0.087	0.088	0.090
	(0.0026)	(0.0035)	(0.008)	(0.010)	(0.010)	(0.010)
exper	0.0033	0.00067	0.0080	0.0011	0.0007	0.0003
	(0.0024)	(0.00273)	(0.0072)	(0.0081)	(0.0081)	(0.0081)
$exper^2$	–0.00023	–0.00019	–0.00062	–0.00051	–0.00049	–0.00047
	(0.00005)	(0.00006)	(0.00006)	(0.00017)	(0.00017)	(0.00017)
kidlt6	–0.180	–0.183	–0.520	–0.512	–0.339	–0.346
	(0.016)	(0.016)	(0.045)	(0.046)	(0.064)	(0.064)
$(nwifeinc - \overline{nwifeinc})^2$	—	—	—	—	–0.000046	–0.00029
					(0.000018)	(0.00012)
$kidlt6 \cdot nwifeinc$	—	—	—	—	–0.0061	–0.0058
					(0.0016)	(0.0016)
$\hat{v}_2$	—	—	—	–0.0052	–0.0055	–0.0273
				(0.0027)	(0.0027)	(0.0072)
$\hat{v}_2^2$	—	—	—	—	—	–0.00052
						(0.00013)

$\hat{v}_2 \cdot nwifeinc$	—	—	—	—	—	0.00075
						(0.00024)
intercept	0.333	0.371	–0.494	–0.390	–0.474	–0.422
	(0.045)	(0.048)	(0.130)	(0.140)	(0.141)	(0.143)
$\widehat{APE}_{nwifeinc}$	–0.0033	–0.0014	–0.0033	–0.0015	–0.00097	–0.0015
	(0.0002)	(0.0010)	(0.0002)	(0.0009)	(0.00010)	(0.0010)
Observations	5,634	5,634	5,634	5,634	5,634	5,634

Notes: (i) Standard errors for OLS and 2SLS are robust to heteroskedasticity.
(ii) The standard errors for the CF estimates are based on 1,000 bootstrap replications.
(iii) $\widehat{APE}_{nwifeinc}$ is the estimated derivative-based average partial effect of *nwifeinc*, where the individual APEs are averaged across the entire sample.

Table 4
Average Partial Effects of nwifeinc *at Different Quartiles*

	1 Probit CF	2 Probit CF	3 Probit CF
No young children			
25th percentile	–0.00143	0.00026	0.00163
	(0.00087)	(0.00105)	(0.00129)
50th percentile	–0.00146	–0.00014	–0.00068
	(0.00091)	(0.00098)	(0.00095)
75th percentile	–0.00149	–0.00065	–0.00367
	(0.00095)	(0.00096)	(0.00158)
At least one young child			
25th percentile	–0.00157	–0.00197	–0.00067
	(0.00099)	(0.00120)	(0.00129)
50th percentile	–0.00156	–0.00240	–0.00295
	(0.00099)	(0.00115)	(0.00098)
75th percentile	–0.00155	–0.00291	–0.00535
	(0.00097)	(0.00109)	(0.00124)

Notes: (i) Column 1 is for the probit estimates reported in Column 4 of Table 3, Column 2 corresponds to Column 5 in Table 3, and Column 3 corresponds to Column 6 in Table 3.
(ii) All standard errors are obtained from 1,000 bootstrap replications.

One of the benefits of using a nonlinear model is that it allows the effects of the explanatory variables to change in a parsimonious way. Table 4 provides estimates of average partial effects for *nwifeinc*, evaluated at the median as well as the first and third quartiles. I also consider the APEs with and without a young child. All of the other variables are averaged out. The picture is now different than that for the simple model; those APEs are reported in Column 1 of the table. When *nwifeinc* appears linearly in the probit model, its APE is essentially flat across the six combinations of (*kidlt*6, *nwifeinc*). By contrast, in Column 2, the APEs vary substantially across different settings of the two covariates. The effect of *nwifeinc* is essentially zero at the three income settings for women without a young child although the point estimates show the effect increases in magnitude as income increases. For women with a young child, the effect is marginally significant at the lowest quartile, –0.0020 ($t = -1.65$), and is largest at the 75th percentile, –0.0029 ($t = -2.67$).

Finally, Column 6 in Table 3 contains estimated parameters of a model that adds a quadratic in the CF, $\hat{v}_{i2}$, along with an interaction between $\hat{v}_{i2}$ and $nwifeinc_i$. Now the three terms that depend on the CF are jointly very significant, with *p*–value equal to zero to four decimal places. Plus, each term is individually very significant, suggesting that the earlier models suffer from functional form misspecification. As often happens in comparing a variety of models, the estimated APE across all observations is very similar to the simpler models, including the linear model estimated by IV: –0.0015 ($t = -1.50$). But the pattern of APEs at different (*kidlt*6, *nwifeinc*) pairs differs. Column 3 in Table 4 contains the APEs. Now *nwifecinc* has a negative, statistically significant

effect at the highest quartile among women without a small child: –0.0037 ($t = -2.32$). Among women with a child, there is no income effect at the lowest quartile but a fairly large effect, –0.0054 ($t = -4.31$), at the highest quartile.

There is no guarantee that even the last model captures all of the important nonlinearities, but the example shows that accounting for the nonlinearities is potentially important. With large sample sizes, one can try interactions among all variables — including the control function — and quadratics in the continuous variables (including the control function). Two-step estimation is simple and the bootstrap efficiently computes standard errors of the coefficients and the average partial effects.

The BP setup, and therefore convenient parametric approximations, extends easily to the case of a vector of continuous EEVs, say $\mathbf{y}_2$, provided there are sufficient instruments. An example is Petrin and Train (2010), which studies multinomial consumer choice models with a vector of endogenous price variables. Rather than start with, say, a multinomial or nested logit model that depends on unobserved taste heterogeneity that can be correlated with price, Petrin and Train proposes estimating such models for $D(y_1|\mathbf{z}_1, \mathbf{y}_2, \mathbf{v}_2)$, where $\mathbf{v}_2$ is the vector of reduced for errors in $\mathbf{y}_2 = \mathbf{\Pi}_2\mathbf{z} + \mathbf{v}_2$. When $\mathbf{v}_2$ is replaced with reduced-form residuals $\hat{\mathbf{v}}_{i2}$ — obtained from OLS regressions using prices or log prices — the CF methods are computationally simple even for many choice alternatives. The standard approach, where the distribution of the heterogeneity is modeled and then integrated out, is much more complicated. Petrin and Train provides evidence that the CF approach works well.

B. Discrete EEVs

The major impediment to extending the BP framework to allow discrete EEVs is that the average structural function is nonparametrically unidentified even under fairly strong independence assumptions; see Chesher (2003). Consequently, parametric CF approaches when y_2 is discrete generally require the parametric assumptions to hold in order to achieve identification. By contrast, the parametric models discussed in the previous subsection are offered as flexible approximations to an analysis that, in principle, could be fully nonparametric.

The traditional approach to estimating nonlinear models with discrete y_2 is not a CF approach. Instead, maximum likelihood — or, in some cases, quasi-MLE (see Wooldridge 2014 for some recent examples) — is by far the leading method. One occasionally sees plug-in methods used but these are generally inconsistent. In this subsection, I discuss how two-step CF methods can be used in place of MLE approaches under a different set of parametric assumptions. The CF approach is somewhat controversial in this case because the assumptions under which it produces consistent partial effects are nonstandard.

To illustrate the issues, suppose that y_1 is binary and is generated by Equation 34. Now, y_2 is also binary and follows a linear index model:

$$y_2 = 1[\mathbf{z}\boldsymbol{\delta}_2 + v_2 \geq 0]. \quad (51)$$

As example, I could model a binary outcome, such as graduating from high school (y_1), as a function of attending a Catholic high school (y_2). Usually the parameters in Equations 34 and 51 are estimated jointly by MLE under the assumption that (u_1, v_2) is independent of $\mathbf{z}$ with a bivariate normal distribution, where u_1 and v_2 are both

standard normal. This model is sometimes called a "bivariate probit" model, where y_2 appears in the equation for y_1 but Equation 51 is taken to be a reduced form probit equation. The ASF, $\Phi(\mathbf{z}_1\boldsymbol{\delta}_1 + \gamma_1 y_2)$, is easily estimated given the MLEs of $\boldsymbol{\delta}_1$ and γ_1. A plug-in approach that replaces y_{i2} with probit fitted values, $\Phi(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$, in the second-stage probit inconsistently estimates both the parameters and the average partial effects.

Under the standard bivariate probit assumptions, there is no known CF method that consistently estimates the parameters. Nevertheless, as shown by Wooldridge (2014), an optimal test of the null hypothesis that y_2 is exogenous is obtained as the usual MLE t statistic on the generalized residual $\hat{r}_{i2} = y_{i2}\lambda(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2) - (1 - y_{i2})\lambda(-\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$. Therefore, if one knew $\boldsymbol{\delta}_2$, rather than having to estimate it, one would estimate the probit model

$$P(y_{i1} = 1|\mathbf{z}_{i1}, y_{i2}, r_{i2}) = \Phi(\mathbf{z}_{i1}\boldsymbol{\delta}_1 + \gamma_1 y_{i2} + \rho_1 r_{i2}) \quad (52)$$

and test $H_0 : \rho_1 = 0$. To operationalize the test, replace r_{i2} with $\hat{r}_{i2}$.

An intriguing possibility is that including $\hat{r}_{i2}$ in the second-stage probit along with $(\mathbf{z}_{i1}, y_{i2})$ might provide an accurate correction for "small" amounts of endogeneity, where smallness is measured by the size of ρ_1. Terza, Basu, and Rathouz (2008) (TBR) was the first to propose adding residuals to standard models — such as probit — to solve the endogeneity problem for discrete y_2. Rather than the generalized residual $\hat{r}_{i2}$, TBR uses the residual $\hat{e}_{i2} = y_{i2} - \Phi(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$, but the motivation is the same. As noted by Wooldridge (2014), in order to use Equation 52 to consistently estimate the average partial effects, one needs to add the assumption that r_2 acts as a kind of sufficient statistic for capturing the endogeneity of y_2. One can state the condition by recalling that $y_1 = 1[\mathbf{z}_1\boldsymbol{\delta}_1 + \gamma_1 y_2 + u_1 \geq 0]$. Then, assume that u_1 depends on $(\mathbf{z}, y_2)$ only through r_2 in the conditional distribution sense:

$$D(u_1|\mathbf{z}, y_2) = D(u_1|r_2). \quad (53)$$

When Equations 52 and 53 are combined, the average structural function can be consistently estimated, just as in the BP case, by averaging out the generalized residuals:

$$\widehat{ASF}(\mathbf{z}_1, y_2) = N^{-1}\sum_{i=1}^{N}\Phi(\mathbf{z}_1\hat{\boldsymbol{\delta}}_1 + \hat{\gamma}_1 y_2 + \hat{\rho}_1\hat{r}_{i2}). \quad (54)$$

In using Equation 52 as an estimating equation, I still require that $\mathbf{z}_i$ has at least one element with nonzero coefficient in $\boldsymbol{\delta}_2$ that is excluded from $\mathbf{z}_{i1}$. This ensures that r_{i2} has variation that is not determined entirely by $(\mathbf{z}_{i1}, y_{i2})$. As with any CF method, it is better to have more independent variation in r_{i2}. Because r_{i2} depends on $\mathbf{z}_i$ in a nonlinear way, technically I could get by with $\mathbf{z}_i = \mathbf{z}_{i1}$. However, as in other contexts, I should not achieve identification off of nonlinearities. That is, if a linear version of the model is not identified, then I should not proceed with a nonlinear model. For further discussion, see Wooldridge (2010, Section 9.5).

It is important to understand that the CF approach and the bivariate probit approach use the same probit reduced form for y_2 but use different assumptions about the conditional distribution $D(y_1|\mathbf{z}, y_2)$. The bivariate probit approach requires an extra integration that leads to a fairly complicated log likelihood function; see, for example, Wooldridge (2010, Section 15.7.3). By contrast, Assumption 52 leads to a straightforward two-step method. While the CF assumptions are nonstandard, they are no more or less general than the bivariate probit assumptions. Because Equation 52 is a valid

approximation for ρ_1 "near" zero, the simple CF method might provide good estimates of the ASF fairly generally.

Using the data in AET, but with only 5,979 students due to missing data on the binary response $y_1 = hsgrad$, the probit model in Equation 52 can be estimated by inserting the same generalized residuals used for the linear *math*12 equation. The exogenous variables are exactly as before, with the distance dummies playing the role of instruments. The coefficient from the second stage probit on $\hat{r}_{i2}$ is 0.626 ($t = 3.15$), suggesting a strong form of self-selection into attending a Catholic high school. The average partial effect of *cathhs* using the two-step CF approach is actually negative, –0.082, with *p*–value above 0.25. Thus, in this simple model, there is no evidence that attending a Catholic high school has a positive causal effect on graduating from high school. When the generalized residuals are dropped so that *cathhs* is treated as exogenous, the APE is 0.047 ($t = 4.92$), suggesting a nontrivial positive and very statistically significant effect. A complete set of estimates is available on request.

As in the case with a continuous EEV, I can use flexible parametric models to allow general interactive effects inside the probit function. For example,

$$(55) \quad P(y_{i1} = 1|\mathbf{z}_i, y_{i2}, r_{i2}) = \Phi(\mathbf{x}_{i1}\boldsymbol{\beta}_1 + \rho_1 r_{i2} + \mathbf{x}_{i1} r_{i2}\boldsymbol{\psi}_1),$$

where $\mathbf{x}_1$ is a general function of $(\mathbf{z}_1, y_2)$ and includes an intercept. I can use a standard Wald test of $H_0 : \rho_1 = 0, \boldsymbol{\psi}_1 = \mathbf{0}$ after replacing r_{i2} with its generalized residuals from the first-stage probit. The average structural function is estimated as in Equation 50 with $\hat{r}_{i2}$ replacing $\hat{v}_{i2}$.

If one embraces the flexibility of the control function approach when combined with sensible parametric functional forms, problems that can be computationally demanding using traditional approaches become much easier. For example, in the binary response model, there might be a continuous EEV, say y_2, and a binary EEV, say y_3. One can include functions of the OLS residuals from the reduced form for y_2 and the generalized residuals from the reduced form probit model for y_3 in a second-stage probit model for y_1. These functions might include quadratics, cubics, and various interactions among the OLS residuals, generalized residuals, and observed covariates.

V. Concluding Remarks

This survey of control function methods has focused on cross-sectional applications where the average partial effects on a mean response function are of primary interest — hence my focus on the average structural function. But one need not focus on the mean. For example, Imbens and Newey (2009) defines the notion of a "quantile structural function" and derives control function methods under monotonicity. It is important to understand that such a change of focus often restricts the amount of heterogeneity that one may have in a model, especially when one approaches the problem from a nonparametric perspective.

Control function methods are also very useful in panel data applications where one must account for unobserved heterogeneity as well as endogeneity. Papke and Wooldridge (2008) shows how the CF approach can be combined with the Chamberlain-Mundlak device for handling time-constant heterogeneity. Two-step estimation methods, where the first stage is a linear reduced form, are computationally

simple and are consistent and asymptotically normal in the presence of serial correlation of unknown form. Altonji and Matzkin (2005) considers nonparametric identification of panel data models and endogeneity in a very general setting.

References

Altonji, Joseph, Todd Elder, and Christopher Taber. 2005. "An Evaluation of Instrumental Variable Strategies for Estimating the Effects of Catholic Schooling." *Journal of Human Resources* 40(4):791–821.

Altonji, Joseph, and Rosa Matzkin. 2005. "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors." *Econometrica* 73(4):1053–102.

Barnow, Burt, Glen Cain, and Arthur Goldberger. 1981. "Selection on Observables." *Evaluation Studies Review Annual* 5(1):43–59.

Blundell, Richard, and James Powell. 2003. "Endogeneity in Nonparametric and Semiparametric Regression Models." In *Advances in Economics and Econonometrics: Theory and Applications*, Eighth World Congress, Volume 2, ed. Mathias Dewatripont, Lars Hansen, and Stephen Turnovsky, 312–57. Cambridge: Cambridge University Press.

——. 2004. "Endogeneity in Semiparametric Binary Response Models." *Review of Economic Studies* 71(3):655–79.

Cameron, Colin, and Pravin Trivedi. 2005. *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

Card, David. 2001. "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems." *Econometrica* 69(5):1127–60.

Chesher, Andrew. 2003. "Identification in Nonseparable Models." *Econometrica* 71(5):1405–41.

Garen, John. 1984. "The Returns to Schooling: A Selectivity Bias Approach with a Continuous Choice Variable." *Econometrica* 52(5):1199–218.

Goldberger, Arthur. 2008. "Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations." In *Advances in Econometrics*, Volume 21, ed. Daniel Millimet, Jeffrey Smith, and Edward Vytlacil, 1–31. Amsterdam: Elsevier.

Hausman, Jerry. 1978. "Specification Tests in Econometrics." *Econometrica* 46(6):1251–71.

Heckman, James. 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models." *Annals of Economic and Social Measurement* 5(4):475–92.

Heckman, James, and Richard Robb. 1985. "Alternative Methods for Evaluating the Impact of Interventions: An Overview." *Journal of Econometrics* 30(1–2):239–67.

Heckman, James, and Edward Vytlacil. 1998. "Instrumental Variables Methods for the Correlated Random Coefficient Model: Estimating the Average Rate of Return to Schooling When the Return Is Correlated with Schooling." *Journal of Human Resources* 33(4):974–87.

Heckman, James, Justin Tobias, and Edward Vytlacil. 2003. "Simple Estimators for Treatment Parameters in a Latent-Variable Framework." *Review of Economics and Statistics* 85(3): 748–55.

Imbens, Guido, and Joshua Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62(2):467–75.

Imbens, Guido, and Whitney Newey. 2009. "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity." *Econometrica* 77(5):1481–512.

Imbens, Guido, and Jeffrey Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47(1):5–86.

Papke, Leslie, and Jeffrey Wooldridge. 2008. "Panel Data Methods for Fractional Response Variables with an Application to Test Pass Rates." *Journal of Econometrics* 145(1–2): 121–33.

Petrin, Amil, and Kenneth Train. 2010. "A Control Function Approach to Endogeneity in Consumer Choice Models." *Journal of Marketing Research* 47(1):3–13.
Rivers, Douglas, and Quang Vuong. 1988. "Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models." *Journal of Econometrics* 39(3):347–66.
Smith, Richard, and Richard Blundell. 1986. "An Exogeneity Test for a Simultaneous Equation Tobit Model with an Application to Labor Supply." *Econometrica* 54(3):679–85.
Terza, Joseph, Anirban Basu, and Paul Rathouz. 2008. "Two-Stage Residual Inclusion Estimation: Addressing Endogeneity in Health Econometric Modeling." *Journal of Health Economics* 27(3):531–43.
Wooldridge, Jeffrey. 2003. "Further Results on Instrumental Variables Estimation of Average Treatment Effects in the Correlated Random Coefficient Model." *Economics Letters* 79(2):185–91.
——. 2005. "Unobserved Heterogeneity and Estimation of Average Partial Effects." In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, ed. Donald Andrews and James Stock, 27–55. Cambridge: Cambridge University Press.
——. 2010. *Econometric Analysis of Cross Section and Panel Data*, 2nd edition. Cambridge: MIT Press.
——. 2014. "Quasi-Maximum Likelihood Estimation and Testing for Nonlinear Models with Endogenous Explanatory Variables." *Journal of Econometrics* 182(1):226–34.

Copyright of Journal of Human Resources is the property of University of Wisconsin Press and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.