

Estimating Marginal Treatment Effects
in Heterogeneous Populations

Robert Moffitt
Johns Hopkins University

December, 2006
Revised, July, 2007

The author would like to thank Marc Chan for research assistance and Lorraine Dearden for generous help in obtaining and using the data. Comments from Joshua Angrist, James Heckman, Guido Imbens, Tiemen Woutersen, and participants at workshops at the BLS, Johns Hopkins, SITE, Western Ontario, and Wharton are appreciated, as well as comments from Thierry Magnac and other participants at the Conference on Econometric Evaluation of Public Policies: Methods and Applications,” Paris, December, 2005 on an earlier closely related paper. Steffen Reinhold also corrected two errors. Research support from the National Institute of Child Health and Human Development is gratefully acknowledged.

welfls0_v3.wpd
7/15/07

Abstract

This paper proposes a nonparametric method of estimating marginal treatment effects in heterogeneous populations. Building upon an insight of Heckman and Vytlačil, the conventional treatment effects model with heterogeneous effects is shown to imply that outcomes are a nonlinear function of participation probabilities. The degree of this nonlinearity, and hence the shape of the marginal response curve, can be estimated with series methods (e.g., power series or splines). An illustration is provided for the returns to higher education in the U.K, indicating that marginal returns to higher education fall as the proportion of the population with higher education rises, thus providing evidence of heterogeneity in returns.

The possible existence of individual heterogeneity in the effect of a treatment on outcomes in a population has been a focus of much work in the causal effects literature. In economics, heterogeneity in the effect of a binary endogenous regressor was introduced in the literature on switching regression models by Quandt (1972), Heckman (1978), and Lee (1979), while in the statistics literature the causal model of potential outcomes of Rubin (1974) also allowed full heterogeneity in treatment effects. This heterogeneity was reformulated as a random coefficient by Heckman and Robb (1985) and by Björklund and Moffitt (1987). The latter paper also introduced the concept of the marginal treatment effect (termed the ‘marginal gain’) in the context of a multivariate-normal switching regression model and showed that the model was observationally equivalent to the Lee switching regression model. Imbens and Angrist (1994) showed that the treatment effect in a heterogeneous population across two points in the distribution, termed the Local Average Treatment Effect (LATE), could be nonparametrically estimated with instrumental variables (IV) and Angrist et al. (1996) elaborated and clarified this method. Heckman and Vytlačil (1999, 2005) have clarified the distinctions between the marginal treatment effect (MTE), the LATE, and other treatment effects of interest.

In this paper, we build upon a remark by Heckman and Vytlačil (2005, p.691) that the treatment effects model with heterogeneous effects of a binary treatment implies that outcomes are simply a nonlinear function of participation probabilities. A model is set up in this paper which demonstrates that point in a slightly reformulated random coefficients model which makes minimal identifying assumptions for the identification of the nonlinearity. A simple series

estimation method is proposed to nonparametrically estimate the shape of the outcome-participation-probability relationship, and hence marginal returns to treatment, which can be implemented with widely-available software packages.

An empirical illustration is provided for the effect of a binary higher education indicator on earnings in the UK using the data from a study by Blundell et al. (2005). The literature on the effect of education on earnings has seen the largest number of discussions of heterogeneity in the return, a concept discussed in the Woytinsky Lecture of Becker (1975) and in Mincer (1974). Surveys of the empirical literature by Card (1999, 2001) have emphasized the impact of possible heterogeneity in the return on the interpretation of the estimates in that literature (see also Lang (1993)). The large majority of these estimates use only a binary instrument and hence only one piece of the marginal return function can be nonparametrically identified, whereas in this paper a wider portion of the return function is estimated because multiple, multi-valued instruments are used. Carneiro et al. (2003a, 2003b) have also used a wider range of instruments, together with other identifying assumptions, and have been able to estimate the full range of returns to education. Oreopoulos (2006) has examined heterogeneity in returns to education by comparing LATE estimates based on compulsory schooling laws between two countries which have different fractions of the population affected by the laws, which implicitly uses a three-valued instrument rather than a binary one.

The next section lays out the model and estimation method, and the subsequent section provides the illustration. A summary appears at the end.

I. Estimation of the Heterogeneous Effects Model

The model presented here is adapted from those in the treatment effects literature referenced in the Introduction. Let y_i be an outcome variable for individual i , D_i a dummy variable signifying participation in the program, and Z_i an instrumental variable with a multinomial distribution. An unrestricted model, assuming no other covariates, can be written as

$$y_i = \beta_i + \alpha_i D_i \quad (1)$$

$$D_i^* = k(Z_i, \delta_i) \quad (2)$$

$$D_i = 1(D_i^* \geq 0) \quad (3)$$

where β_i and α_i are scalar random parameters and δ_i is a vector of random parameters. All parameters are allowed to be individual-specific and to have some unrestricted joint distribution $f(\beta_i, \alpha_i, \delta_i)$; thus a separate model (1)-(3) exists for each individual. The function k is likewise unrestricted and hence the model for D_i can be saturated in Z_i , though restrictions on δ_i will be necessary for interpretation (see below). Eqn (1) is to be interpreted as a description of potential outcomes, not just a description of how y_i varies with D_i in any particular population; hence α_i is the object of interest.¹ What can be estimated, however, is only the mean of α_i on some subpopulation(s).

If we condition (1) on D_i , we obtain $E(y_i | D_i) = E(\beta_i | D_i) + E(\alpha_i | D_i) D_i$, which illustrates one conditional mean of interest. But to see which of the classes of objects can be

¹ In the language and notation of potential outcomes, Y_{0i} ($=\beta_i$) is the value of the outcome if individual i does not participate, Y_{1i} ($=\beta_i+\alpha_i$) is the value of the outcome if individual i does participate, and $\alpha_i=Y_{1i}-Y_{0i}$ is the program impact for individual i .

identified, we work instead with the reduced form by conditioning (1)-(3) on Z_i :

$$E(y_i | Z_i=z) = E(\beta_i | Z_i=z) + E(\alpha_i | D_i=1, Z_i=z) \text{Prob}(D_i=1 | Z_i=z) \quad (4)$$

$$E(D_i | Z_i=z) = \text{Prob} [k(z, \delta_i) \geq 0] \quad (5)$$

We make the following minimal identifying assumptions:

$$A1. E(\beta_i | Z_i=z) = \beta \quad (6)$$

$$A2. E(\alpha_i | D_i=1, Z_i=z) = g[E(D_i | Z_i=z)] \quad (7)$$

Assumptions A1 and A2 are mean independence assumptions needed for Z_i to be a valid exclusion restriction. They state that the mean of the random intercept and the conditional mean of the random coefficient, respectively, are, in the former case, independent of Z_i and, in the latter case, dependent only on the fraction treated and not directly on Z_i . The existing literature often assumes, instead of A2, that all potential outcomes are independent of treatment status, which, together with the assumption that Z_i enters the D_i equation, implies that the full distribution of α_i in the treated population is dependent on Z_i only through the probability of participation. A2 is a slightly weaker condition which assumes that α_i in the treated population is only mean independent of Z_i conditional on the participation probability, and states this assumption as a primitive rather than deriving it from other assumptions.²

² In most applications, full independence may hold in any case. But there may be applications where the variation in the participation rate induced by the instrument is located only in one part of the alpha distribution, and one may have more confidence in the similarity of that

To interpret the estimates of marginal treatment effects estimated below as the mean α_i of those who change participation, we also need a “monotonicity” assumption originally formulated by Imbens and Angrist (1994):

$$A3. D_i(Z_i=z) - D_i(Z_i=z') \text{ is zero or the same sign for all } i \text{ for any distinct values } z \text{ and } z' \quad (8)$$

where $D_i(Z_i=z)$ is the function described in (2)-(3). This assumption constitutes a restriction on the distribution of δ_i (see also Heckman and Vytlacil, 2005, for a discussion).

With these assumptions, and letting $F(Z_i)=E(D_i | Z_i)$, (4) and (5) can be rewritten as

$$y_i = \beta + g[F(Z_i)] F(Z_i) + \epsilon_i \quad (9)$$

$$D_i = F(Z_i) + v_i \quad (10)$$

where F is a proper probability function and where $E(\epsilon_i | gF) = E(v_i | F) = 0$ by construction.

No other restriction on the distribution of ϵ_i or v_i is made. The implication of response heterogeneity can be seen in (9) to be that the effect of program participation (F) on y varies with the level of participation because g is a function of F , thus inducing an inherent nonlinearity of y in F , a feature of heterogeneous treatment effects models noted by Heckman and Vytlacil (2005, p.691) and also discussed in Heckman et al. (2006). A homogeneous-effects model is one in which g is a constant.

Nonparametric identification of the parameters of (9) and (10) is straightforward given

part of the distribution across values of the instrument than in other parts of the alpha distribution.

that D_i is binary and Z_i has a multinomial distribution. $F(Z_i)$ is identified at each point $Z_i=z$ from the population mean of D_i at that z . The elements of the function g that can be identified depend on the support of $F(Z_i)$ and, as has been emphasized in the literature and originally emphasized by Imbens and Angrist (1994), not all elements can be identified if the support of Z_i in the sample does not generate full support of F from 0 to 1. For two or more points in the support of F , the local average treatment effect between two participation fractions F_j and $F_{j'}$ is the discrete slope of the y function between those points, $\Delta y/\Delta F = [F_j g(F_j) - F_{j'} g(F_{j'})] / (F_j - F_{j'})$. The marginal treatment effect at some point F_j is instead the continuous derivative, $\partial y/\partial F = g'(F_j)F_j + g(F_j)$, which must be obtained by some smoothing method given the multinomial assumption on Z_i . If the support of F contains the value 0, $g(F_j)$, the effect of the treatment on the treated, is likewise identified at all other points in the support of F .³ If $F=1$ as well as $F=0$ is contained in the support, the average treatment effect, $g(1)$, is therefore also identified.

Nonparametric estimation of the g function will be conducted here by series estimation methods rather than with kernel methods. Series estimation methods, whether by power functions or spline functions, are easily implemented in conventional regression packages because they merely involved adding additional regressors to a linear model. Here, (9) simply becomes a linear regression model with regressors that are nonlinear in $F(Z)$. However, rather than using IV, which would require instrumenting each of the regressors in the series and programming the two-step standard errors, (9) and (10) will be estimated jointly by nonlinear least squares (NLS) and the standard errors will be computed by the conventional robust formula

³ The effect of the treatment on the treated as defined here is conditional on z ; however, by integrating z out, the effect unconditional on z can be obtained.

which is simpler. NLS is also available in most regression packages and hence should also be relatively easy to implement by practitioners.⁴ The estimation problem is

$$\text{Min}_{\theta_1, \theta_2} \sum_i w_{1i} [y_i - \beta - g[F(k(Z_i, \theta_2)); \theta_1] F(k(Z_i, \theta_2))]^2 + \sum_i w_{2i} [D_i - F(k(Z_i, \theta_2))]^2 \quad (11)$$

where θ_1 and θ_2 are the parameter vectors of the series expansions of functions g and k , respectively, and w_{1i} and w_{2i} are weights to improve efficiency.⁵

Adding a vector of exogenous observables X_i , the model becomes:

$$y_i = \alpha_i D_i + h_i(X_i) \quad (12)$$

$$D_i^* = k(Z_i, X_i, \delta_i) \quad (13)$$

$$D_i = 1(D_i^* \geq 0) \quad (14)$$

We assume

$$\text{B1. } E[h_i(X_i) | X_i=x, Z_i=z] = h(x) \quad (15)$$

$$\text{B2. } E(\alpha_i | D_i=1, X_i=x, Z_i=z) = g[E(D_i | X_i=x, Z_i=z), x] \quad (16)$$

$$\text{B3. } D_i(Z_i=z, X_i=x) - D_i(Z_i=z', X_i=x) \text{ is zero or the same sign for all } i \text{ for any distinct values } z \text{ and } z' \quad (17)$$

Then, conditioning (12)-(14) on X_i and Z_i , we have:

⁴ As will be seen below, we will be parametric on F , assuming it to be normal. This is a nonrestrictive assumption if F is saturated in Z . Without saturation and with parametric restrictions, this model differs from the linear probability model because it forces the conditional mean of Z to follow a normal c.d.f. shape with respect to Z instead of a linear shape. Estimating a nonparametric regression of D on Z would nest the two forms.

⁵ In the application below, cross-equation weights will be used as well.

$$y_i = g[F(Z_i, X_i), X_i] F(Z_i, X_i) + h(X_i) + \epsilon_i \quad (18)$$

$$D_i = F(Z_i, X_i) + v_i \quad (19)$$

where, again, the errors are mean-independent of the regressors by construction. The two equations could be estimated jointly with nonlinear least squares.

Nonparametric methods could, in this case, be used to estimate the unknown functions g, F , and h . However, in our empirical application below, this is not attempted, for index functions will be instead used for all three functions except for the form in which F affects g :

$$y_i = X_i \beta + [X_i \lambda + g(F(X_i \delta + Z_i \eta))] F(X_i \delta + Z_i \eta) + \epsilon_i \quad (20)$$

$$D_i = F(X_i \delta + Z_i \eta) + v_i \quad (21)$$

with an appropriate redefinition of the function g , and where F is taken as the normal c.d.f. We will test for nonlinearities in g by approximating it with series methods, as noted above. Note that, even with its linear index restrictions, this model allows an interaction of X with the effect of treatment on y as long as λ is nonzero, which is a departure from most IV practice.⁶ Note as well that the parametric nature of the model will allow estimation of the entire distribution of g , since both power functions and splines can be extrapolated beyond the range of $F(Z)$ in the data. However, it will be important to note that these estimates are the result of extrapolation and that the estimates of g within the range of F in the data are presumably more reliable.

⁶ Blundell et al. (2005), however, have an extensive discussion of interactions of X with treatment in the IV model. Note that the vector $X_i \lambda$ excludes a constant term.

Denote u_i as a 2×1 column vector of residuals for individual i , the first being the residual in (20) and the second being the residual in (21). Then the unknown parameters in both equations, which we denote θ , are estimated by the criterion

$$\hat{\theta} = \underset{\theta}{\text{Arg Min}} \sum_{i=1}^n u_i' \hat{\Omega}^{-1} u_i \quad (22)$$

where $\hat{\Omega}$ is a 2×2 matrix of variances of the residuals obtained from an unweighted first-stage estimation. The covariance matrix of $\hat{\theta}$ is calculated as

$$\text{Var}(\hat{\theta}) = \left(\sum_{i=1}^n M_i' \hat{\Omega}^{-1} M_i \right)^{-1} \left(\sum_{i=1}^n M_i' \hat{\Omega}^{-1} u_i u_i' \hat{\Omega}^{-1} M_i \right) \left(\sum_{i=1}^n M_i' \hat{\Omega}^{-1} M_i \right)^{-1} \quad (23)$$

where M_i is the $2 \times K$ matrix of derivatives of the two conditional mean functions w.r.t. the K parameters in the model.

II. An Empirical Illustration

Preliminaries. The empirical illustration presented here will be for the case where the effect of higher education on future earnings is the object of interest, focusing as well (as in much of the literature) on the effect of a discrete change in education from less-than-college to college-or-more. The education-earnings literature is the literature where heterogeneity in returns has been discussed most heavily, as noted in the Introduction. As for whether the MTE for the return to college should be expected to rise or fall as a larger fraction of individuals go to college, this depends, as always, on the nature of the instrument and what portion of the ability distribution is

swept out by its variation. The usual practice in the literature is to seek instruments which proxy, or are correlated with, costs of schooling. In this case, the Becker Woytinsky Lecture model implies that the MTE will fall if costs fall and participation expands as the lower-return individuals are drawn into any given level of schooling. Therefore, that will be the prior for the empirical exercise conducted here.⁷

It is also worth noting that the empirical literature to date has generally found OLS estimates of the return to be below IV estimates, where the latter are interpretable as LATE or, in continuous terms, as the MTE (Card, 1999, 2001). One possible explanation of this result (see Card as well as Angrist and Krueger (1999, pp. 1324-1325)) is that an instrument may affect different individuals in the population in different ways and may affect those with high MTE values disproportionately. The same result applies in the model in (1)-(3) above because that model allows unobserved heterogeneity in δ_i . This is formally shown in Appendix A. However, it is also shown there that if this the correct explanation for the finding that OLS are greater than IV estimates, the MTE nevertheless cannot be larger than OLS in the neighborhood of $F=0$ or $F=1$, and that OLS must be smaller than the effect of the treatment on the treated, whose estimate requires an instrument generating values of $F=0$. Therefore, a test of this explanation for the MTE-OLS difference is available if the instruments provide variation in those ranges of F . We will illustrate this in the application.

⁷ It should be noted that the relationship of interest here is how the MTE changes as the fraction of the population with a fixed level of schooling increases. The usual Becker-Woytinsky diagram which portrays returns vs the level of schooling must be analyzed with a vertical line drawn at the fixed level of schooling. A shift in the marginal cost curve then has the effects just noted. This is somewhat different than the question of whether the LATE falls at successively higher levels of schooling, which Card (1999, p.1854) tentatively found to be the case.

Application. For our application, we use the data employed in Blundell et al. (2005), who estimated the effect of higher education on earnings in the UK in 1993.⁸ The data set consisted of information on 3,639 males whose earnings were observed at age 33 in 1991, and whose families had been interviewed periodically since birth to collect child and family background information. The regressor of interest was a dummy variable indicating whether the individual had undertaken some form of higher education, and a set of other socioeconomic characteristics were available for use as control variables. The OLS estimate of the effect of higher education on the log of the hourly wage was .287. The authors obtained IV estimates with three variables used as instruments: (1) a dummy variable for whether the parents reported an adverse financial shock at either age 11 or age 16 of the child, (2) a dummy variable for whether the child's teacher ranked the parent's "interest in education" high or low when the child was 7, and (3) the number of older siblings of the child (the total number of siblings was used as a control variable in the wage regression). The authors argued that these variables could be excludable from the wage equation, and noted that they have high F-statistics in the first-stage regression. In this paper, we do not question the credibility of the instruments but take their validity as a maintained assumption in order to illustrate the estimation method, which is our main interest. Blundell et al. found IV estimates of the return to higher education to fall in a very wide range (.05, 1.17) for the three different instruments, and made a priori arguments for why different instruments should have different effects, depending on their correlation with unobserved returns and costs in the

⁸ The author would like to thank Lorraine Dearden for providing the data and explaining the variables and samples.

population.⁹

Here we use the same data as Blundell et al. and estimate a slightly condensed model with fewer X variables, excluding those with coefficients of low significance and condensing categories (e.g., region) where coefficient differences are of low significance. The means of the variables in the data set are shown in Appendix B, along with the OLS regressions, which generate an estimate of the effect of higher education of .287 (robust s.e.=.02), identical to that of Blundell et al. We then estimate our models using all three instruments (Z). The literature has noted that different instruments may sweep out different portions of the return distribution and hence may have different MTEs (Imbens and Angrist, 1994; Card, 1999, 2001; Heckman and Vytlacil, 2005; see also Blundell et al. for a discussion focused on these three instruments), in which case the MTE estimates from a model which includes all instruments must be interpreted as weighted averages of the MTEs in those different populations. However, different instruments may also simply sweep out different ranges of the F distribution, and this will also generate different estimates of the MTE when the instruments are used separately if heterogeneity exists and hence the MTE is a function of F. The method used here assumes each Z to sweep out the same portion of the return distribution at the same F but allows each Z to operate in a different portion of the F distribution, which will generate a different value of the MTE for each Z for this reason alone.

Table 1 shows the estimates of the treatment effects not allowing the effect of participation to vary with X (i.e., assuming $\lambda=0$). The g function (=effect of the treatment on the

⁹ In an earlier version of their paper, Blundell et al. (2001) used all three instruments together.

treated) is estimated with both linear splines and polynomials:

$$g(F) = \gamma_0 + \sum_{j=1}^J \gamma_j \text{Max}(0, F - \pi_j) \quad (24)$$

$$g(F) = \gamma_0 + \sum_{j=1}^J \gamma_j F^j \quad (25)$$

where J is the number of terms in the series and where the π_j are preset knots, in this case taken to be quartile points of the estimated F distribution. Linear splines with preset knots have the advantage of allowing one to see slopes directly off the estimates in different regions rather than having to generate them from a polynomial and of allowing γ to have zero regions, but have the disadvantage of generating discontinuous derivatives (=the MTE) at knot points and requiring, at least in the simple method used here, a priori determination of the knots.¹⁰

Column (1) shows estimates of a model with just a constant in (24)-(25), equivalent to the homogeneous-effects model. The estimate of .34 is slightly above the OLS estimate, consistent with much of the literature (estimates of the other parameters in the model are shown in Table B1).

Figure 1 shows a histogram of predicted participation rates from the estimated first-stage equation and indicates a concentration of probabilities in the lower ranges of F and with sizable fractions of the data at higher probabilities as well, although the distribution becomes thin above .70. However, most of this variation is generated by variation in X , and the relevant issue for

¹⁰ There are many more sophisticated spline methods which address some of these features, such as methods which allow estimation of the knot points and which allow derivatives to be continuous at knot points (e.g., de Boor, 2001).

this model is instead the incremental effect of the instruments on these probabilities. The coefficients on the instruments are generally significant (see Table B1) and have an F-statistic of 18 in the nonlinear form of the first-stage equation estimated here and an F-statistic of 13 if a linear first-stage equation is estimated.¹¹ Table 2 shows a box plot of the incremental effect of the instruments on the spread of predicted F, where the “baseline” F is obtained by setting the values of the instruments equal to their means (but allowing X to vary in the sample) and the “actual” F is obtained by allowing both Z and X to vary. The instruments provide quite a bit of additional variation in the middle ranges of the probabilities (e.g., .30 to .70) but very little additional variation at both low and high values of F. This is an important result because it demonstrates that, despite the concentration of the overall predicted probabilities in the region around F=0, the instruments have very little power in that region. They have more power in the higher regions, but there is also relatively little data in the higher regions. Thus the region where both there is a reasonable fraction of the data and where the instruments have their most relevance is in the relatively narrow region of approximately (.30, .60). These remarks also suggest that, for models with effect heterogeneity, instruments can be strong in some regions of F but weak in other regions, a feature not generally noted in the weak instruments literature.¹²

¹¹ Linear IV (i.e., a two-step procedure with the linear probability model for the first stage) yields a treatment-effect estimate of .71 (s.e.=.16). This suggests that the linearity assumption in the first stage is important and possibly restrictive.

¹² The particular functional form of the incremental effects of the instruments shown in Figure 2 is, to some extent, driven by the normal distribution, which necessarily implies a smaller incremental effect of any regressor at high regions and low regions of F. However, this must necessarily also hold in a more nonparametric model qualitatively. It is worth noting that a linear probability model for the first stage would generate the same incremental effects on F at all points in the F distribution, suggesting another limitation of such a model for the purposes of this paper.

The rest of the columns in Table 1 show the degree of nonlinearity with respect to F using splines and polynomials. Column (2) allows the g function to vary linearly with F and indicates that the treatment effect declines as F rises and more of the population is engaged in higher education. This is therefore consistent with the prior. Column (3) adds a spline knot at the 50th percentile point of the predicted F distribution, showing that the t-statistics on the nonlinear F parameters decline markedly and the parameters reach implausible magnitudes in some ranges. Column (4) adds two further splines showing parameters that, while retaining significance at conventional levels, reach further implausible magnitudes. Column (5) shows the effect of adding one additional polynomial term, a quadratic in F (which implies that log wages are cubic in F) and shows no significant evidence of higher nonlinearity. Taken as a whole, these estimates do not provide evidence of any reliably-estimated nonlinearities beyond the first order, although there are hints in the spline results of some convexity in the function.¹³

The rapid decline in the stability of the estimates as additional nonlinearities are introduced arises from two related sources. The more important is the already-noted weakness of the instruments in high and low ranges of F; instruments which have little or no effect on F in those regions should be expected to generate unstable and implausible values. Figures 3 and 4 plot the g function (treatment on the treated) and the MTE (derivative of the equation w.r.t. F), respectively, for columns (1), (2), (3), and (5) of Table 1 (note that the effect of the treatment on the treated is identified because $F=0$ is in the support of the data). In the F region $[\.30, \.60]$, the three models allowing nonlinearities, including the polynomial, are reasonably close to one

¹³ A cross-validation statistic could be used to more formally choose the degree of nonlinearity but is left for future work.

another. Further, in Figure 4, these three models also show definite evidence of declining MTE in that range.¹⁴ However, the functions diverge dramatically at both higher and lower values of F, precisely where the instruments are very weak.

A second, related factor is that the instruments, while generating more than a single variation in predicted F that is allowed with a binary instrument, nevertheless generate only a limited set of values. Two of the three instruments are binary and the third (number of older siblings) is concentrated in only three values (0, 1, and 2). Thus the number of discrete points of support in the incremental predicted F distribution is still modest. Adding parameters to the model by introducing spline and polynomial terms necessarily requires a sufficient range of instruments to support estimation of those parameters, and that range may still not be sufficient with these instruments. In estimates not reported here, interactions between the three instruments and nonlinearities in the third instrument were added to the first-stage equation to generate a greater range of incremental F contributions, but those additional interactions and nonlinearities were extremely weak. The F statistic for five instruments falls to 9 and a more extensive set of interactions leading to fifteen instruments yields an F statistic of 4. Tests of interactions of the initial three instruments with the variables in the X vector leads to F values of 2 or 3. The instruments in these data are therefore too weak to obtain more variation in predicted probabilities and therefore a wider range of probabilities over which to estimate nonlinear treatment effects.

Table 2 allows interactions with treatment and the variables in the X vector ($\lambda \neq 0$). The first three columns, showing results for two of the spline models and the polynomial model,

¹⁴ However, the confidence intervals for these estimates (not shown) heavily overlap.

show that the nonlinear treatment effects are rendered insignificant or much less significant in the spline models but slightly more significant in the polynomial specification. At least for the two spline specifications, this suggests that the unobservable heterogeneity in return found in the Table 1 results may be masking heterogeneity in the effects by observables. However, as can be seen by an inspection of the results, the interaction coefficients for the large majority of the seventeen variables have large standard errors. Restricting the interactions to the five variables that are significant at conventional levels restores the spline-model nonlinear effects to significance. Thus estimates of the effect of unobservables on estimates of the return are quite sensitive to whether and which interactions are allowed, suggesting that a more formal determination of which interactions should be included in the model is needed. The insignificance of most of the interaction terms may also be related, once again, to weaknesses in the instruments in generating sufficient incremental effects on the F distribution for different values of X. Further work is needed on these issues.

Finally, note that the OLS estimates are smaller than the model estimates over virtually all of the distribution of F (see Figure 4). However, the MTE is always less than the TT for all models as well, and, further, OLS is usually less than the TT. Thus there is little support for the explanation for the OLS-IV difference noted in prior work and described in Appendix A, at least for these instruments.¹⁵

¹⁵ A qualification is necessary because the Appendix proof pertains to local OLS estimates in the same range of F as the IV model estimates. The OLS estimates obtained here, however, are global. Also, note that for the MTE to be greater than the TT requires that $g'(F) > 0$ at some point over the range of F, and this is not the case for values of F in the lower 60 percent of the distribution.

III. Summary and Conclusions

We have proposed a method of estimating the shape of the marginal return function in the treatment-effects model when heterogeneous returns are present and have applied the method to the data from a prior study of the effect of higher education on earnings of men in the UK. The application shows significant effects of heterogeneity, indicating that marginal returns to higher education fall as the proportion of the population with higher education rises. This direction of effect is consistent with the Becker Woytinsky Lecture model. However, the instruments used are weak in some ranges of the F distribution and hence these findings apply to only a limited range of the participation-rate spectrum. There is also some uncertainty regarding the relative contributions of observable and unobservable heterogeneity to this finding. These topics suggest further work on more formal methods of addressing these issues.

Appendix A

Relationship of MTE to OLS and Interpretation of IV Estimates

As noted by Card (1999, 2001), heterogeneity in the effect of an instrument on may lead to IV-based LATE or MTE estimates that exceed OLS estimates. This effect operates in the model in (1)-(3) through the δ_i . A reformulated model for the education case is:

$$y_i = \bar{\beta} + \alpha_i D_i + \epsilon_i \quad (\text{A1})$$

$$D_i^* = \alpha_i - c_i + v_i \quad (\text{A2})$$

$$D_i = 1(D_i^* \geq 0) \quad (\text{A3})$$

where $\beta_i = \bar{\beta} + \epsilon_i$ and where the education choice equation is assumed to be based on the earnings return minus costs (c_i) plus other unobserved determinants (v_i), an equation which drops out of the standard theory. Let $c_i = \delta_i Z_i$ where Z_i measures observed costs or a proxy for it (the instrument) and where $\delta_i > 0$ is a measure of the responsiveness of an individual to a change in costs; hence

$$D_i^* = \alpha_i - \delta_i Z_i + v_i \quad (\text{A4})$$

Those with greater values of δ_i have a lower probability of $D_i=1$, hence lower schooling levels.

The expectation of the OLS estimate of the effect of D_i on y_i , familiar from the literature, is:

$$\begin{aligned}
\alpha_{OLS} &= E(y_i | D_i=1) - E(y_i | D_i=0) \\
&= E(\alpha_i | D_i=1) + [E(\epsilon_i | D_i=1) - E(\epsilon_i | D_i=0)]
\end{aligned}
\tag{A5}$$

where the first term is the effect of the treatment on the treated (TT). Although the second term (in brackets) could be negative if those who attend college would have had lower earnings than those who did not attend college if they also did not, this is unlikely. Let us therefore assume that the second term is non-negative. This implies that OLS will be equal to or greater than the TT. OLS can be guaranteed to be above the MTE as well if the TT exceeds the MTE, but not otherwise; hence the question is the relationship between the TT and the MTE.

The TT conditional on Z_i is:

$$E(\alpha_i | D_i=1, Z_i) = E(\alpha_i | u_i > 0, Z_i) \tag{A6}$$

where $u_i = \alpha_i - \delta_i Z_i + v_i$. This mean will be above $\bar{\alpha}$ if positive sorting on α_i takes place, and this will depend on the sign of the covariance of α_i and u_i . That covariance will be positive unless α_i is sufficiently positively correlated with δ_i or negatively correlated with v_i , but the former is implausible because costs are generally thought to be negative correlated with α_i and there is no reason to expect any particular sign for the correlation of α_i and v_i . For present purposes let us assume that α_i is uncorrelated with δ_i and v_i . Then positive sorting on α_i conditional on Z_i results. Integrating (A6) over the distribution of Z_i guarantees that the unconditional-on- Z TT is also positively sorted on α_i .

Given this, it might be assumed that the MTE must be less than TT since the MTE is

equal to $E(\alpha_i | u_i=0)$, which is the minimum of the TT distribution. But the question instead is what values of MTE are swept out by a change in Z_i . To determine this we must calculate the MTE conditional on δ_i and then integrate over it. The MTE in general is $\partial E(y)/\partial F$ where F is $\text{Prob}(D=1)$. Conditional on δ_i , and recalling that $E(y_i | Z_i, \delta_i) = \bar{\beta} + E(\alpha_i | D_i=1, Z_i, \delta_i)F(Z_i, \delta_i)$, the MTE generated by a change in Z_i is therefore

$$\begin{aligned}
\text{MTE}(Z_i, \delta_i) &= \frac{\partial E(y_i | Z_i, \delta_i) / \partial Z_i}{\partial F(Z_i, \delta_i) / \partial Z_i} & (A7) \\
&= \frac{[\partial E(\alpha_i | D_i=1, Z_i, \delta_i) / \partial Z_i] F(Z_i, \delta_i) + E(\alpha_i | D_i=1, Z_i, \delta_i) [\partial F(Z_i, \delta_i) / \partial Z_i]}{\partial F(Z_i, \delta_i) / \partial Z_i} \\
&= \Delta_i(Z_i, \delta_i) + E(\alpha_i | D_i=1, Z_i, \delta_i)
\end{aligned}$$

where

$$\Delta_i(Z_i, \delta_i) = \frac{[\partial E(\alpha_i | D_i=1, Z_i, \delta_i) / \partial Z_i] F(Z_i, \delta_i)}{\partial F(Z_i, \delta_i) / \partial Z_i} \quad (A8)$$

is the difference between the MTE and TT conditional on δ_i . With positive sorting on α_i , $\Delta_i < 0$ and hence the MTE is less than the TT again (conditional on δ_i). How the value of Δ_i varies with δ_i depends on the shape of the distribution of α_i and the rest of the parameters of the problem, so let us assume for simplicity that it is independent of δ_i and hence $\Delta_i(Z_i, \delta_i) = \Delta_i(Z_i)$.

Then the unconditional-on- δ_i MTE is

$$\begin{aligned}
\text{MTE}(Z_i) &= \frac{\int [\partial E(y_i | Z_i, \delta_i) / \partial Z_i] dG(\delta_i)}{\int [\partial F(Z_i, \delta_i) / \partial Z_i] dG(\delta_i)} \\
&= \Delta_i(Z_i) / dF_T(Z_i) + \int E(\alpha_i | D_i=1, Z_i, \delta_i) p(Z_i, \delta_i) dG(\delta_i)
\end{aligned} \tag{A9}$$

where G is the c.d.f. of δ_i , $dF_T(Z_i) = \int [\partial F(Z_i, \delta_i) / \partial Z_i] dG(\delta_i)$ is the total change in the fraction with $D_i=1$, and

$$p(Z_i, \delta_i) = \frac{\partial F(Z_i, \delta_i) / \partial Z_i}{\int [\partial F(Z_i, \delta_i) / \partial Z_i] dG(\delta_i)} \tag{A10}$$

is the fraction of the change in the fraction with $D_i=1$ arising from each δ_i subpopulation. Since the unconditional-on- δ_i TT is

$$E(\alpha_i | D_i=1, Z_i) = \int E(\alpha_i | D_i=1, Z_i) dG(\delta_i) \tag{A11}$$

it is clear that the second term in (A9) can be greater than this TT if $p(Z_i, \delta_i)$ is positively related to the conditional-on- δ_i TT. But that is the case in this problem. This concludes the demonstration that the MTE can be greater than the TT, and hence that OLS may be smaller than the MTE.

Despite this result, it must still be the case that the TT is greater than OLS, which follows from (A5) and its associated assumptions. In addition, the MTE must be equal to the TT at $F=0$ (the α_i of the first person to participate constitutes both the MTE and the TT) and the MTE must be less than the TT as F approaches 1, for the TT for each δ_i approaches the same number and

hence the second term in (A9) approaches the unconditional-on- δ_1 TT. If the TT can be estimated, a test of this hypothesis for the reason for an MTE greater than OLS can also be tested.

References

- Angrist, J.; ; G. Imbens; and D. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." Journal of the American Statistical Association 91 (June): 444-472.
- Angrist, J. and A. Krueger. 1999. "Empirical Strategies in Labor Economics." In Handbook of Labor Economics, Vol. 3A, eds. O. Ashenfelter and D. Card. Amsterdam: North-Holland.
- Becker, G. 1975. Human Capital. 2nd Ed. New York and London: Columbia University Press.
- Björklund, A. and R. Moffitt. 1987. "The Estimation of Wage and Welfare Gains in Self-Selection Models." Review of Economics and Statistics 69: 42-49.
- Blundell, R.; L. Dearden; and B. Sianesi. 2001. "Evaluating the Returns to Education: Models, Methods and Results." Paper presented at the Meetings of the Royal Statistical Society.
- Blundell, R.; L. Dearden; and B. Sianesi. 2005. "Evaluating the Effect of Education on Earnings: Models, Methods and Results from the National Child Development Survey." Journal of the Royal Statistical Society A, Part 3: 473-512.
- Card, D. 1999. "The Causal Effect of Education on Earnings." In Handbook of Labor Economics, Vol. 3A, eds. O. Ashenfelter and D. Card. Amsterdam: North-Holland.
- Card, D. 2001. "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems." Econometrica 69 (September): 1127-1160.
- Carneiro, Pedro; James J. Heckman; and Edward Vytlacil. 2003a. "Understanding What Instrumental Variables Estimate: Estimating Marginal and Average Returns to Education." Mimeo.
- Carneiro, P. ; J. Heckman; and E. Vytlacil. 2003b. "Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice." International Economic Review 44 (May): 361-422.
- de Boor, C. 2001. A Practical Guide to Splines. New York: Springer.
- Heckman, J. 1978. "Dummy Endogenous Variables in a Simultaneous Equation System." Econometrica 46: 931-960.
- Heckman, J. and R. Robb. 1985. "Alternative Methods for Evaluating the Impact of Interventions." In Longitudinal Analysis of Labor Market Data, eds J. Heckman and B. Singer. Cambridge University Press..

Heckman, J.; S. Urzua; and E. Vytlacil. 2006. "Understanding Instrumental Variables in Models with Essential Heterogeneity." REStat 88 (August): 389-432.

Heckman, J. and E. Vytlacil. 1999. "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects." Proceedings of the National Academy of Sciences 96 (April): 4730-4734.

Heckman, J. and E. Vytlacil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." Econometrica 73 (May): 669-738.

Imbens, G. and J. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." Econometrica 62: 467-76.

Lang, K. 1993. "Ability Bias, Discount Rate Bias, and the Return to Education." Mimeographed. Boston: Boston University.

Lee, L.F. 1979. "Identification and Estimation in Binary Choice Models with Limited (Censored) Dependent Variables." Econometrica 47: 977-996.

Mincer, J. 1974. Schooling, Experience, and Earnings. New York and London: Columbia University Press.

Oreopoulos, P. 2006. "Estimating Average and Local Average Treatment Effects of Education When Compulsory Schooling Laws Really Matter." American Economic Review 96 (March): 152-175.

Quandt, R. 1972. "A New Approach to Estimating Switching Regressions." Journal of the American Statistical Association 67 (1972): 306-310.

Rubin, D. 1974. "Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies." Journal of Educational Psychology 66: 688-701.

Table 1

Gamma Parameter Estimates Assuming $\lambda=0$

	(1)	(2)	(3)	(4)	(5)
Constant	.34 (.09)	.89 (.16)	1.56 (.53)	7.17 (1.71)	.90 (.33)
F	--	-.62 (.15)	-2.89 (1.71)	-48.55 (13.58)	-.65 (.76)
Max(0,F-F(.25))	--	--	--	43.13 (12.86)	--
Max(0,F-F(.50))	--	--	2.21 (1.64)	3.16 (1.78)	--
Max(0,F-F(.75))	--	--	--	1.29 (.64)	--
F ²					.37 (.60)

Notes:

1. Standard errors in parentheses.
2. Parameter estimates for the full model including β , δ , and η are shown in Table B2 for Column (1).
3. Percentile points for splines: $F(.25)=.10$, $F(.50)=.24$, $F(.75)=.43$

Table 2

Gamma and Lambda Parameter Estimates Assuming $\lambda \neq 0$

	(1)	(2)	(3)	(4)	(5)
<u>Gamma</u>					
Constant	.85 (.32)	1.40 (.59)	.85 (.50)	.82 (.18)	1.62 (.55)
F	-.00 (.44)	-2.44 (2.24)	-1.61 (1.02)	-.63 (.16)	-3.32 (1.73)
Max(0,F-F(.50))	--	2.13 (1.90)	--	--	2.59 (1.66)
F ²	--	--	1.61 (.79)	--	--
<u>Lambda</u>					
Public School	.03 (.19)	.07 (.19)	.03 (.23)	--	--
Other School	.47 (.28)	.45 (.28)	.47 (.29)	.39 (.25)	.38 (.25)
Math Ability at age 7	-.01 (.03)	-.00 (.03)	-.01 (.04)	--	--
Verbal Ability at age 7	-.04 (.04)	-.03 (.05)	-.04 (.05)	--	--
Verbal Ability at age 7 Missing	.08 (.23)	.16 (.25)	.08 (.25)	--	--
Math Ability at age 11	-.01 (.05)	.01 (.05)	-.01 (.05)	--	--
Verbal Ability at age 11	-.06 (.05)	-.05 (.05)	-.06 (.05)	--	--

Table 2 (continued)

	(1)	(2)	(3)	(4)	(5)
Verbal Ability at age 11 Missing	-.22 (.23)	-.11 (.25)	-.22 (.26)	--	--
Father's Education	-.02 (.02)	-.01 (.02)	-.02 (.03)	--	--
Father's Education Missing	-.10 (.24)	-.03 (.25)	-.10 (.27)	--	--
Mother Employed in 1974	.01 (.07)	.01 (.07)	.01 (.07)	--	--
No. of Siblings	-.04 (.02)	-.04 (.02)	-.04 (.02)	-.05 (.02)	-.06 (.02)
Father Unskilled Manual in 1974	.49 (.41)	.49 (.40)	.49 (.35)	--	--
Father Occupation Missing	.11 (.26)	.16 (.27)	.11 (.29)	--	--
Region Group 1	.28 (.09)	.28 (.09)	.28 (.100)	.18 (.08)	.18 (.08)
Region Group 2	.27 (.11)	.27 (.11)	.27 (.12)	.17 (.10)	.17 (.10)
Region Group 3	.33 (.11)	.33 (.11)	.33 (.12)	.23 (.11)	.23 (.10)

Notes:

1. Standard errors in parentheses.
2. Parameter estimates for β , δ , and η are not shown.
3. Percentile points for splines: $F(.25)=.10$, $F(.50)=.24$, $F(.75)=.43$

Table B1

Means of the Variables in the Data Set

Log wage	2.04
D (=1 if higher education)	.28
<u>X</u>	
Public School	.05
Other School	.02
Math Ability at age 7	2.72
Verbal Ability at age 7	2.55
Verbal Ability at age 7 missing	.11
Math Ability at age 11	2.41
Verbal Ability at age 11	2.34
Verbal Ability at age 11 missing	.19
Father's Education	7.27
Father's Education missing	.28
Mother Employed in 1974	.51
No. of Siblings	1.69
Father Unskilled Manual in 1974	.03
Father Occupation Missing	.11
Region Group 1	.47
Region Group 2	.13
Region Group 3	.15

Table B1 (continued)

<u>Z</u>	
Adverse Financial Shock	.16
Parental Interest	.39
No. Older Siblings	.82

Notes:

N=3,639

Region Group 1: North Western, North, East and W. Riding, North Midlands, South Western, Midlands

Region Group 2: Eastern, Southern

Region Group 3: Wales, Scotland

London and Southeast omitted

Table B2

Full Estimates for OLS and Basic Nonlinear Least Squares Specifications

	OLS	Nonlinear Least Squares
Higher Education	.287 (.015)	.340 (.091)
β		
Public School	.121 (.032)	.110 (.037)
Other School	-.104 (.056)	-.100 (.056)
Math Ability at age 7	.028 (.006)	.027 (.006)
Verbal Ability at age 7	.012 (.006)	.009 (.007)
Verbal Ability at age 7 missing	.192 (.034)	.139 (.036)
Math Ability at age 11	.028 (.006)	.014 (.009)
Verbal Ability at age 11	.033 (.008)	.031 (.008)
Verbal Ability at age 11 missing	.174 (.031)	.110 (.034)
Father's Education	.012 (.004)	.010 (.006)
Father's Education missing	.104 (.047)	.085 (.057)
Mother Employed in 1974	.035 (.015)	.036 (.015)

Table B2 (continued)

	OLS	Nonlinear Least Squares
No. of Siblings	-.009 (.004)	-.008 (.004)
Father Unskilled Manual in 1974	-.093 (.032)	-.091 (.033)
Father Occupation Missing	-.133 (.031)	-.048 (.061)
Region Group 1	-.192 (.020)	-.192 (.020)
Region Group 2	-.106 (.026)	-.107 (.026)
Region Group 3	-.242 (.024)	-.242 (.024)
Constant	1.716 (.051)	1.75 (.072)
<u>δ</u>		
Public School	--	.584 (.116)
Other School	--	-.306 (.234)
Math Ability at age 7	--	.103 (.023)
Verbal Ability at age 7	--	.162 (.026)
Verbal Ability at age 7 missing	--	1.034 (.126)

Table B2 (continued)

	OLS	Nonlinear Least Squares
Math Ability at age 11	--	.209 (.033)
Verbal Ability at age 11	--	.104 (.034)
Verbal Ability at age 11 missing	--	1.102 (.122)
Father's Education	--	.112 (.016)
Father's Education missing	--	1.060 (.19)
Mother Employed in 1974	--	-.094 (.065)
No. of Siblings	--	.008 (.027)
Father Unskilled Manual in 1974	--	-.113 (.192)
Father Occupation Missing	--	1.003 (.209)
Region Group 1	--	.005 (.078)
Region Group 2	--	.084 (.097)
Region Group 3	--	-.001 (.095)
Constant	--	-3.691 (.226)

Table B2 (continued)

	OLS	Nonlinear Least Squares
η		
Adverse Financial Shock	--	-.394 (.095)
Parental Interest	--	.272 (.056)
No. Older Siblings	--	-.054 (.034)

Notes:

Standard errors in parentheses

Nonlinear Least Squares corresponds to Table 1, Column (1)

Figure 1: Histogram of Predicted Participation Rates

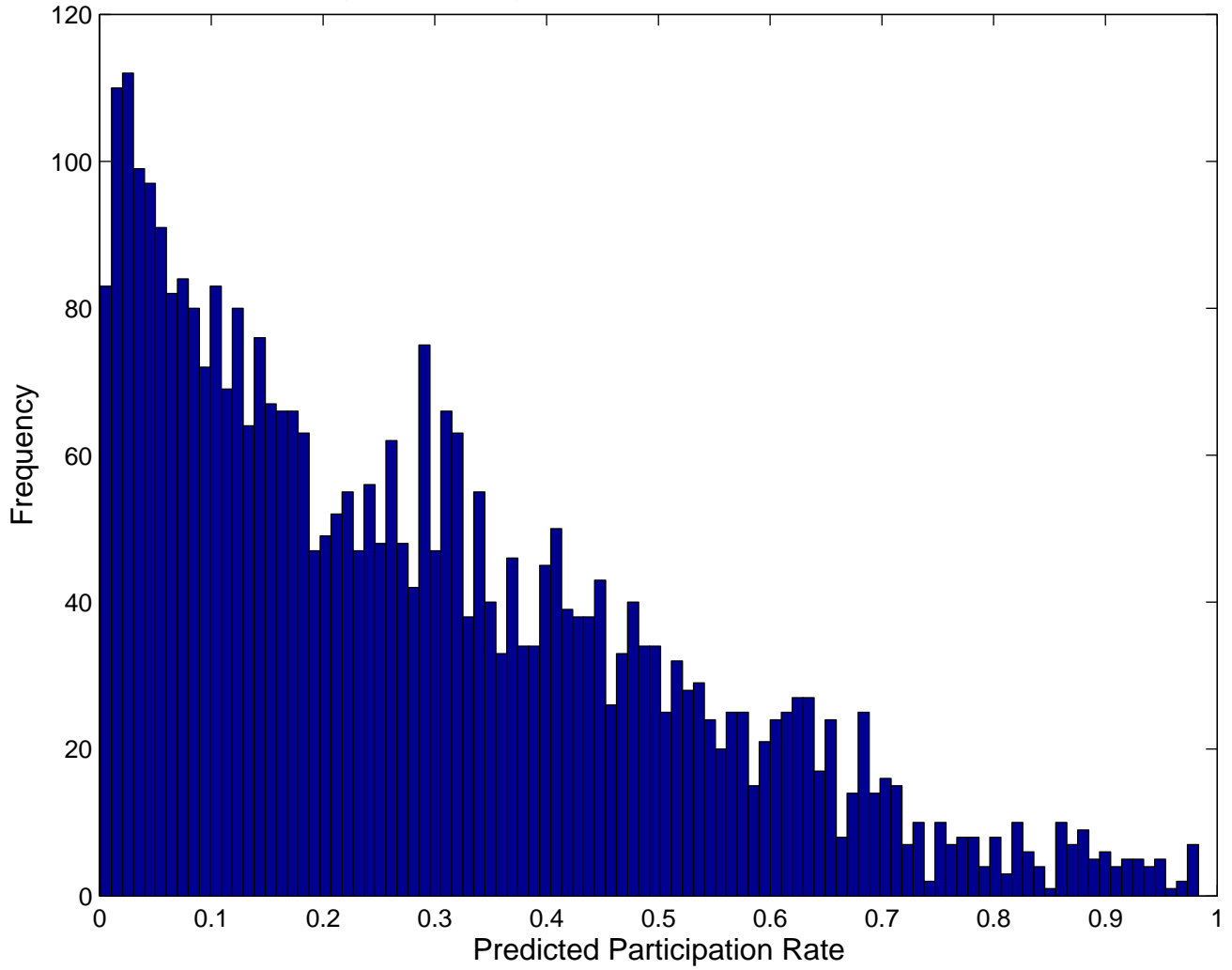
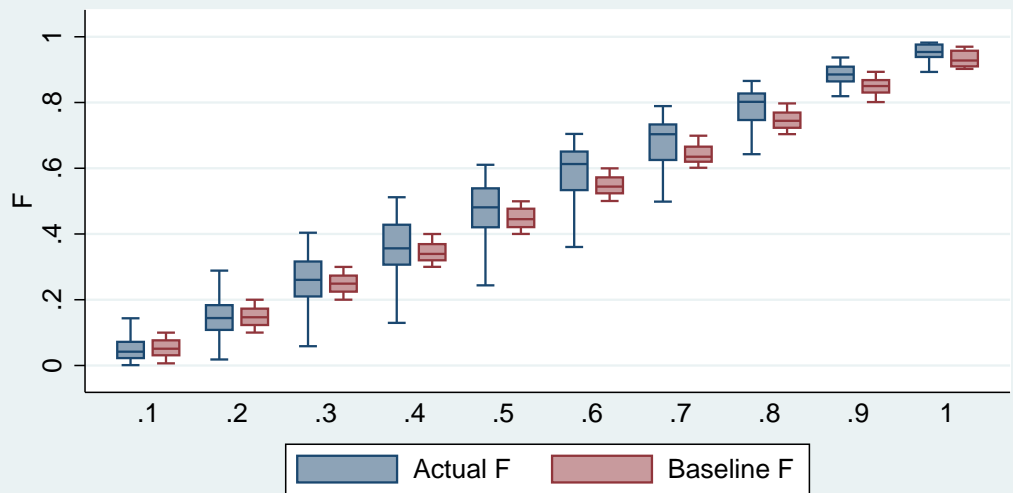


Figure 2: Baseline and Actual F Distribution at Deciles of Baseline F



Baseline F is the predicted probability holding the Z vector at its mean.
 Actual F is the predicted probability allow the Z vector to vary.
 Horizontal axis represents decile ranges of Baseline F.
 The upper and lower points of the rectangles are 75th and 25th percentile points of the distribution, respectively, and the horizontal lines inside the recentangles are medians.
 Upper and lower tick marks above and below the rectangles are upper and lower ranges, respectively.

Figure 3: Value of the Effect of the Treatment on the Treated for Different Models

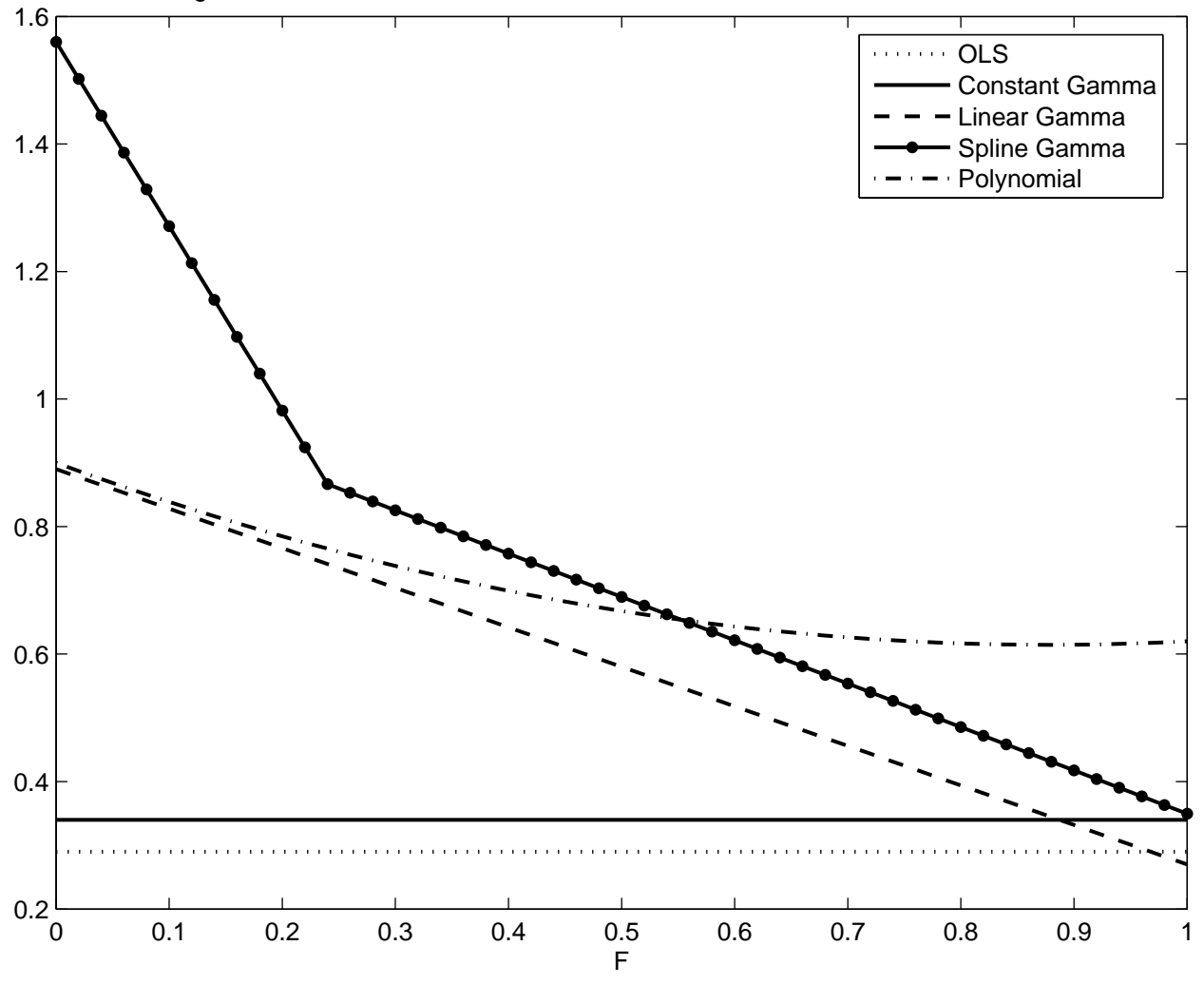


Figure 4: MTE for Different Models

