Estimating Marginal Treatment Effects

in Heterogeneous Populations

Robert Moffitt
Johns Hopkins University

December, 2006
Revised, January, 2008

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, Matthew White, Tiemen Woutersen, and participants at workshops at several universities, research institutes, and conferences are appreciated, as well as comments from two anonymous referees, 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. The author also thanks Steffen Reinhold for correcting two errors in the Appendix. Research support from the National Institute of Child Health and Human Development is gratefully acknowledged.

welfls0_v4b.wpd
1/22/08

Abstract

This paper proposes a nonparametric method of estimating marginal treatment effects in heterogeneous populations. Building upon an insight of Heckman and Vytlacil, 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 such as 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 allows 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 Vytlacil (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 Vytlacil (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. (2003) and Aakvik et al. (2005) also obtained a wider range of estimates of the return function but partly because of parametric assumptions; however, Carneiro et al. (2006) used a wide range of instruments to nonparametrically estimate the full range of returns to education, similar to the exercise here. Oreopoulos (2006) 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

$$y_i = \beta_i + \alpha_i D_i \tag{1}$$

$$D_i^* = k(Z_i, \delta_i) \tag{2}$$

$$D_i = 1(D_i^* \geq 0) \tag{3}$$

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 and its distribution is the object of interest.[1] There are two sources of possible selection in the model: first, selection occurs if β_i covaries with δ_i (those with different unobserved participation propensities have different levels of y in the absence of the treatment) and, second, selection

---

[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.

occurs if α_i covaries with δ_i (those with different unobserved participation propensities have different unobserved 'gains' from the treatment).

If we condition (1) on D_i, we obtain $E(y_i \mid D_i) = E(\beta_i \mid D_i) + E(\alpha_i \mid D_i) D_i$, which illustrates one conditional mean of interest. But to see which of the classes of objects can be identified, we work instead with the reduced form by conditioning (1)-(3) on Z_i:

$$E(y_i \mid Z_i{=}z) = E(\beta_i \mid Z_i{=}z) + E(\alpha_i \mid D_i{=}1, Z_i{=}z) \, Prob(D_i{=}1 \mid Z_i{=}z) \tag{4}$$

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

We make the following minimal identifying assumptions:

A1. $E(\beta_i \mid Z_i{=}z) = \beta$ (6)

A2. $E(\alpha_i \mid D_i{=}1, Z_i{=}z) = g[E(D_i \mid Z_i{=}z)]$ (7)

Assumptions A1 and A2 are mean independence assumptions needed for Z_i to be a valid exclusion restriction. Eqn(6) states that the mean of the random intercept must be independent of Z_i (individuals must have the same level of y in the absence of treatment for all values of Z). Eqn(7) states that the mean 'gain' from the treatment among those who participate must depend on Z_i only through the fraction treated and not otherwise. If α_i covaries with δ_i, a change in Z_i will alter the types of individuals who participate and the mean of α_i among participants will change. For example, in the usual positive selection case, as participation in a treatment expands, those brought into the treatment have smaller positive α_i than those who have already participated, and the mean α_i among participants will fall. At different levels of Z_i, therefore, that mean will vary.

The existing literature usually assumes, instead of A2, that both potential outcomes are fully independent of Z_i and therefore that their difference, α_i, is also fully independent of Z_i; however, because Z_i enters the D_i equation, the distribution of α_i in the D_i=1 subpopulation is dependent on Z_i through the probability of participation in that case (assuming α_i covaries with δ_i), so (A2) holds. But A2 is a slightly weaker condition than full independence because it states that only the mean of α_i in the D_i=1 subpopulation need be independent of Z_i, conditional on the participation probability. This condition is stated as a primitive rather than deriving it from other assumptions.[2]

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. $D_i(Z_i=z) - D_i(Z_i=z')$ is zero or the same sign for all i for any distinct values z and z' (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 $F(Z_i)=E(D_i \mid Z_i)$, (4) and (5) can be rewritten as

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

$$y_i = \beta + g[F(Z_i)]\, F(Z_i) + \epsilon_i \tag{9}$$

$$D_i = F(Z_i) + \nu_i \tag{10}$$

where F is a proper probability function and where $E(\epsilon_i \mid gF) = E(\nu_i \mid F) = 0$ by construction. No other restriction on the distribution of ϵ_i or ν_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 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 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.[4] 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). Estimation of (10) is possible in several different ways. For example, (9) and (10) could be jointly estimated with nonlinear least squares, allowing for heteroskedasticity (particularly in (10)) and for a nonzero across-equation error covariance.[5] However, here, instead, the more traditional two-step method will be employed, using first-stage estimates based on probit estimation of F(Z) followed by second-stage estimation of (9) using predicted values of F as regressors. Robust standard errors correcting for estimation error in F and for the nonlinearity of F in (9) are obtained by applying formulas from Newey and McFadden (1994, eqn(6.11)).[6]

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

---

[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.

[4] Carneiro et al. (2006) add a vector of X variables and apply the partially-linear model to estimate g(F)F by kernel methods, for example.

[5] Earlier versions of this paper used this method.

[6] The normality restriction on F could be relaxed by applying a more nonparametric estimation procedure to the first stage. Note that the linear probability model would be inappropriate if it were to predict negative probabilities (in the application below, it does so), for it would not be sensible to provide estimates of g at negative values of F.

$$y_i = \alpha_i D_i + h_i(X_i) \tag{11}$$

$$D_i^* = k(Z_i, X_i, \delta_i) \tag{12}$$

$$D_i = 1(D_i^* \geq 0) \tag{13}$$

We assume

B1. $E[\, h_i(X_i) \mid X_i{=}x, Z_i{=}z) = h(x)$ (14)

B2. $E(\alpha_i \mid D_i{=}1, X_i{=}x, Z_i{=}z) = g[E(D_i \mid X_i{=}x, Z_i{=}z), x]$ (15)

B3. $D_i(Z_i{=}z, X_i{=}x) - D_i(Z_i{=}z', X_i{=}x)$ is zero or the same sign for all i for any distinct values z and z' (16)

Then, conditioning (11)-(13) on X_i and Z_i, we have:

$$y_i = g[F(Z_i,X_i), X_i]\, F(Z_i,X_i) + h(X_i) + \epsilon_i \tag{17}$$

$$D_i = F(Z_i,X_i) + \nu_i \tag{18}$$

where, again, the errors are mean-independent of the regressors by construction. Nonparametric methods could, in this case, be used to estimate the unknown functions g and h. However, in our empirical application below, this is not attempted. Instead, index functions will be used for all functions except g:

$$y_i = X_i\beta + [X_i\lambda + g(F(X_i\eta + Z_i\delta))]\, F(X_i\eta + Z_i\delta) + \epsilon_i \tag{19}$$

$$D_i = F(X_i\eta + Z_i\delta) + \nu_i \tag{20}$$

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.[7] 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.

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 to 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 how the conditional mean of α (usually interpreted as arising from unobserved ability) varies with that instrument. 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

---

[7] 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.

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.[8]

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, where it is demonstrated that, for the MTE to be greater than OLS, it is necessary that the MTE also be greater than the TT (effect of treatment on the treated). However, it is also shown there that OLS must nevertheless be greater than the TT and, in addition, the MTE be larger than the TT or OLS in the neighborhood of F=0 or F=1. Therefore, a test of this explanation for the MTE-OLS difference is available if the instruments provide variation in those ranges of F, which are also necessary to obtain an estimate of the TT. We will illustrate this in the application.

Application. For our application, we use the data employed in Blundell et al. (2005), who

---

[8] 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, which differs slightly from the standard textbook analysis. 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.

estimated the effect of higher education on earnings in the UK in 1993.[9] The data set consisted of information on 3,639 males whose earnings were observed at age 33 in 1993, 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 excluded 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 population.[10]

---

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

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

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. In principle, it would be possible to test whether the three instruments generate different estimates of the return to education at the same F if the supports of F generated by the instruments overlap, but this is not done here because the methodological goal is best served by maximizing the range of F and that is achieved by using all three instruments together. In practice, the results can be interpreted as weighted averages as discussed in the articles referenced above.

Table 1 shows the estimates of the treatment effects not allowing the effect of participation to vary with X (i.e., assuming λ=0). The g function (=effect of the treatment on the treated) is estimated with both linear splines and polynomials:

$$g(F) = \gamma_0 + \sum_{j=1}^{J} \gamma_j \mathrm{Max}(0, F-\pi_j) \tag{21}$$

$$g(F) = \gamma_0 + \sum_{j=1}^{J} \gamma_j F^j \tag{22}$$

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.[11]

Column (1) shows estimates of a model with just a constant in (21)-(22), equivalent to the homogeneous-effects model. The estimate of .33 is slightly above the OLS estimate, consistent with much of the literature (estimates of the other parameters in the model are shown in Table B2).

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

[11] 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).

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 this model is instead the incremental effect of the instruments on these probabilities. The coefficients on the instruments are generally significant (see Table B2) and have an F-statistic of 18 in a nonlinear least squares estimation of the first-stage equation and an F-statistic of 13 if a linear first-stage equation is estimated, within the rule-of-thumb ranges for small numbers of instruments (Stock and Yogo, 2005).[12] 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, 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 those regions. The region where there is both a reasonable fraction of the data and where the instruments have 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.[13]

---

[12] Almost 10 percent of predicted F values from the linear probability model are negative. As noted earlier, this makes it inappropriate for the purpose of this exercise.

[13] 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

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 standard errors on the nonlinear F parameters increase 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.[14]

The rapid decline in the stability of the estimates as additional nonlinearities are introduced could simply reflect the truth; that is, there are indeed no higher-order nonlinearities. In fact, the function g which is being estimated is equal to the conventional Heckman normal lambda term if the unobservables are multivariate normal, and that term is known to be

---

incremental effect of any regressor at high regions and low regions of F. However, this must necessarily also hold in a more nonparametric model, at least 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.

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

approximately linear in the probability of selection, at least in the middle range of probabilities. However, there are two other, related, sources of instability in the higher-order nonlinear terms. 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 log wage equation w.r.t. F), respectively, for columns (2), (3), and (5) of Table 1, along with OLS and the constant-effect estimate (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 another. Further, in Figure 4, these three models also show definite evidence of declining MTE in that range. However, the functions diverge much more at both higher and lower values of F, precisely where the instruments are very weak.[15]

A second, related factor is that the instruments, while generating more than the single variation in predicted F that is allowed with only 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

[15] To ascertain whether stronger instruments would affect the results, Monte Carlos were conducted assuming the coefficients on the three instruments were double and then triple what they are shown to be in Appendix Table B-2. All coefficients in the X vector were assumed to equal what they were estimated to be in that model, and 500 repetitions of multivariate normal errors were drawn with nonzero correlations to generate heterogeneity, for a sample size of 3639 and the same X and Z distribution as in the data. While the Monte Carlo estimates of gamma were, on average, the same regardless of the magnitude of the coefficients on the instruments, the standard errors and confidence intervals for gamma were dramatically lower when the coefficients were double or triple what they are here.

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.

On the central issue of whether the MTE is constant, the evidence from the three models with nonlinearities nevertheless provides strong evidence of nonconstancy and therefore of heterogeneous treatment effects in the population. Figure 5 shows a 95 percent confidence interval for the MTE in the most stably estimated model, that with a linearly declining MTE. Although the confidence intervals would allow a horizontal line in some regions, the intervals are narrow enough to make such horizontality very unlikely.

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, 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, shown in the fourth and fifth columns, 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, recall that the relationship between the MTE and the TT (=the g function) provides a test of whether the increase in the constant treatment effect when going from OLS to IV is arising from the differential effects of the instrument in ranges of F between 0 and 1. Specifically, if the MTE is greater than the TT in some range (it cannot be so at F=0 or F=1), it is possible for the MTE to also be greater than OLS.[16] However, all three nonlinear functions shown in Figures 3 and 4 have MTE values that lie below the TT values for all values of positive F. The TT is g(F) and the MTE is [g(F)+Fg'(F)], so the MTE must be below the TT so long as $g'(F)<0$. But $g'(F)<0$ holds for all the estimated nonlinear models. Thus, with the qualification that the TT estimates obtained here are based on weak instrument variation in the neighborhood

[16] The OLS estimate shown in Figure 3 is not a "local" OLS estimate, and therefore does not strictly conform to the proof in the Appendix, which compares a local OLS estimate to local MTE estimates. Therefore, the test here is based on the relationship between the TT and MTE, which have been locally estimated.

of F=0, there is little support for the explanation for the OLS-IV difference noted in prior work and described in Appendix A for these instruments and for these data.

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. Estimating a wide range of marginal treatment effects puts greater demands on the instruments than is the case for either a binary instrument or the average treatment effect obtained when estimating a single IV coefficient with multi-valued instruments. The results also reveal some uncertainty regarding the relative contributions of observables and unobservables to the heterogeneity that has been found. 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 choices may lead to IV-based LATE or MTE estimates that exceed OLS estimates. This effect operates in the model in (1)-(3) through the heterogeneous δ_i. A reformulated model for the education case is:

$$y_i = \bar{\beta} + \alpha_i D_i + \epsilon_i \tag{A1}$$

$$D_i^* = \alpha_i - c_i + \upsilon_i \tag{A2}$$

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

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 (υ_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

$$D_i^* = \alpha_i - \delta_i Z_i + \upsilon_i \tag{A4}$$

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

We demonstrate the following proposition.

Proposition A1. Let the model be (A1), (A4), and (A3). Define

$$\alpha_{OLS} = E(y_i \mid D_i{=}1) - E(y_i \mid D_i{=}0) \tag{A5}$$

$$\alpha_{TT} = E(\alpha_i \mid D_i{=}1) = \int E(\alpha_i \mid u_i{>}0, Z_i)\, dH(Z_i) \tag{A6}$$

$$\alpha_{MTE} = \int \alpha_{MTE}(Z_i)\, dH(Z_i) \tag{A7}$$

where $u_i=\alpha_i-\delta_i Z_i+\upsilon_i$, $H(Z_i)$ is the cdf of Z_i, $\alpha_{MTE}(Z_i)=\partial E(y_i \mid Z_i) / \partial F(Z_i)$ and $F(Z_i)=\text{Prob}(u_i{>}0 \mid Z_i)$. Assume that $E(\epsilon_i \mid Z_i) = 0$ and that positive sorting takes place, defined as:

$$E(\alpha_i \mid D_i{=}1,Z_i,\delta_i) > E(\alpha_i \mid Z_i,\delta_i) = E(\alpha_i) \tag{A8}$$

where a standard mean independence assumption is embodied in the second equality. Then (1) it is possible that $\alpha_{OLS}<\alpha_{MTE}$ over some ranges of Z_i but (2) this cannot be true in the neighborhood of $F(Z_i)=0$ and $F(Z_i)=1$.

The proposition is not obvious because positive sorting should imply that $\alpha_{OLS}>\alpha_{TT}>\alpha_{MTE}$, but the proposition states that this need not be the case in ranges of F between 0 and 1. The proof of the proposition is based on demonstrating that it is possible that $\alpha_{TT}<\alpha_{MTE}$, which makes $\alpha_{OLS}<\alpha_{MTE}$ possible.

From (A5) and (A1), we have

$$\alpha_{OLS} = E(\alpha_i \mid D_i{=}1) + [E(\epsilon_i \mid D_i{=}1) - E(\epsilon_i \mid D_i{=}0)] \tag{A9}$$

where the first term is the 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. If it is negative, it is obvious that α_{OLS} can be arbitrarily low. Therefore let us only consider the case where it is positive, implying that $\alpha_{OLS}>\alpha_{TT}$. It would appear that $\alpha_{TT}>\alpha_{MTE}$, for the TT conditional on Z_i is

$$E(\alpha_i \mid D_i=1, Z_i) = E(\alpha_i \mid u_i>0, Z_i) \tag{A10}$$

where $u_i=\alpha_i-\delta_i Z_i+\upsilon_i$. The assumption of positive sorting implies that this is greater than $E(\alpha_i \mid u_i=0, Z_i)$, which is the minimum of the TT distribution and constitutes one definition of the MTE (integrating (A10) over the distribution of Z_i guarantees that the unconditional-on-Z TT is also positively sorted). However, the question instead is what values of the 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. Recalling that $E(y_i \mid Z_i, \delta_i)=\bar{\beta}+E(\alpha_i \mid D_i=1, Z_i, \delta_i)F(Z_i,\delta_i)$, the MTE conditional on Z_i is

$$\begin{aligned}\alpha_{MTE}(Z_i) &= \frac{\int [\partial E(y_i \mid Z_i, \delta_i) / \partial Z_i] \; dG(\delta_i)}{\int [\partial F(Z_i, \delta_i) / \partial Z_i] \; dG(\delta_i)} \\ &= \{ \int [\, \partial E(\alpha_i \mid D_i=1,Z_i,\delta_i) / \partial Z_i] \, F(Z_i,\delta_i) \, dG(\delta_i) \} / dF_T(Z_i) \\ &\quad + \int E(\alpha_i \mid D_i=1,Z_i,\delta_i) \, p(Z_i,\delta_i) \, dG(\delta_i)\end{aligned} \tag{A11}$$

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

$$p(Z_i, \delta_i) = \frac{\partial F(Z_i, \delta_i) / \partial Z_i}{\int [\partial F(Z_i, \delta_i) / \partial Z_i] \; dG(\delta_i)} \tag{A12}$$

is the proportion of the change in the fraction with D_i=1 arising from each δ_i subpopulation. The first term in (A11) is negative since positive sorting implies that a rise (say) in F lowers the TT. However, the second term can be arbitrarily greater than the TT. The unconditional-on-δ_i TT is

$$E(\alpha_i \mid D_i{=}1, Z_i) = \int E(\alpha_i \mid D_i{=}1, Z_i, \delta_i) \, dG(\delta_i) \tag{A13}$$

which can be smaller than the second term in (A11) 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.

However, the MTE must equal 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 (A11) approaches the unconditional-on-δ_i TT. It must also be the case that OLS must be everywhere greater than or equal to the TT, at least if the second term in (A9) is non-negative.

References

Aakvik, A.; J. Heckman; and E. Vytlacil. 2005. "Estimating Treatment Effects for Discrete Outcomes When Responses to Treatment Vary: An Application to Norwegian Vocational Rehabilitation Programs." Journal of Econometrics 125: 15-51.

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, P. ; J. Heckman; and E. Vytlacil. 2003. "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.

Carneiro, P.; J. Heckman; and E. Vytlacil. 2006. "Estimating Marginal and Average Returns to Education." Mimeo.

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." Review of Economics and Statistics 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.

Newey, W. and D. McFadden. 1994. "Large Sample Estimation and Hypothesis Testing." Handbook of Econometrics, Volume IV, eds. R. Engle and D. McFadden. Amsterdam: Elsevier.

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.

Stock, J. and M. Yogo. 2005. "Testing for Weak Instruments in Linear IV Regression." In Identification and Inference for Econometric Models, eds. D.W.K. Andrews and J. Stock. Cambridge: Cambridge University Press.

Table 1

Gamma Parameter Estimates

	(1)	(2)	(3)	(4)	(5)
Constant	.33 (.10)	1.01 (.19)	1.67 (.61)	7.12 (1.77)	1.21 (.39)
F	--	-.73 (.17)	-2.91 (1.86)	-45.87 (13.13)	1.21 (.85)
Max(0,F-F(.25))	--	--	--	40.66 (12.31)	--
Max(0,F-F(.50))	--	--	2.09 (1.77)	2.83 (1.89)	--
Max(0,F-F(.75))	--	--	--	1.26 (.70)	--
F^2					.37 (.66)

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). All models constrain λ=0.
3. Percentile points for splines: F(.25)=.10, F(.50)=.24, F(.75)=.43

Table 2

Gamma and Lambda Parameter Estimates

	(1)	(2)	(3)	(4)	(5)
Gamma					
Constant	.87 (.36)	1.43 (.65)	1.63 (.55)	.95 (.20)	1.75 (.62)
F	.08 (.53)	-2.46 (2.52)	-1.73 (1.15)	-.76 (.18)	-3.35 (1.87)
Max(0,F-F(.50))	--	2.14 (2.08)	--	--	2.47 (1.78)
F^2	--	--	1.50 (.83)	--	--
Lambda					
Public School	-.14 (.19)	-.09 (.20)	-.27 (.21)	--	--
Other School	.47 (.28)	.44 (.27)	.44 (.27)	.41 (.26)	.40 (.26)
Math Ability at age 7	-.01 (.04)	.00 (.04)	-.01 (.04)	--	--
Verbal Ability at age 7	-.04 (.05)	-.02 (.05)	-.03 (.05)	--	--
Verbal Ability at age 7 Missing	.19 (.26)	.28 (.27)	.26 (.26)	--	--
Math Ability at age 11	.01 (.06)	.03 (.06)	.02 (.06)	--	--
Verbal Ability at age 11	-.08 (.05)	-.06 (.05)	-.07 (.05)	--	--

Table 2 (continued)

	(1)	(2)	(3)	(4)	(5)
Verbal Ability at age 11 Missing	-.13 (.26)	-.01 (.28)	-.03 (.27)	--	--
Father's Education	-.03 (.03)	-.02 (.03)	-.05 (.03)	--	--
Father's Education Missing	-.13 (.27)	-.04 (.27)	-.29 (.28)	--	--
Mother Employed in 1974	-.01 (.07)	-.02 (.07)	-.01 (.07)	--	--
No. of Siblings	-.03 (.02)	-.04 (.02)	-.03 (.02)	-.05 (.02)	-.06 (.02)
Father Unskilled Manual in 1974	.54 (.41)	.52 (.40)	.51 (.40)	--	--
Father Occupation Missing	-.03 (.29)	.06 (.29)	-.20 (.30)	--	--
Region Group 1	.24 (.09)	.24 (.09)	.24 (.09)	.16 (.08)	.15 (.08)
Region Group 2	.26 (.11)	.27 (.11)	.25 (.11)	.18 (.10)	.18 (.10)
Region Group 3	.35 (.12)	.34 (.12)	.36 (.12)	.24 (.11)	.24 (.11)

Notes:

1. Standard errors in parentheses.
2. Parameter estimates for β, δ, and η are not shown in Table B2 for Column (1).
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
X	
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)

Z	
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 2SLS Specifications

	OLS	2SLS
Higher Education	.287 (.015)	.326 (.102)
β		
Public School	.121 (.032)	.116 (.037)
Other School	-.104 (.056)	-.101 (.056)
Math Ability at age 7	.028 (.006)	.027 (.006)
Verbal Ability at age 7	.012 (.006)	.010 (.007)
Verbal Ability at age 7 missing	.192 (.034)	.144 (.037)
Math Ability at age 11	.028 (.006)	.015 (.009)
Verbal Ability at age 11	.033 (.008)	.031 (.009)
Verbal Ability at age 11 missing	.174 (.031)	.115 (.036)
Father's Education	.012 (.004)	.010 (.006)
Father's Education missing	.104 (.047)	.092 (.058)
Mother Employed in 1974	.035 (.015)	.035 (.015)

Table B2 (continued)

	OLS	2SLS
No. of Siblings	-.009 (.004)	-.008 (.004)
Father Unskilled Manual in 1974	-.093 (.032)	-.092 (.032)
Father Occupation Missing	-.133 (.031)	-.041 (.062)
Region Group 1	-.192 (.020)	-.192 (.020)
Region Group 2	-.106 (.026)	-.106 (.026)
Region Group 3	-.242 (.024)	-.239 (.024)
Constant	1.716 (.051)	1.74 (.074)
$\underline{\eta}$		
Public School	--	.467 (.105)
Other School	--	-.276 (.206)
Math Ability at age 7	--	.097 (.022)
Verbal Ability at age 7	--	.147 (.024)
Verbal Ability at age 7 missing	--	.953 (.117)

Table B2 (continued)

	OLS	2SLS
Math Ability at age 11	--	.194 (.031)
Verbal Ability at age 11	--	.121 (.033)
Verbal Ability at age 11 missing	--	1.056 (.112)
Father's Education	--	.104 (.015)
Father's Education missing	--	.962 (.175)
Mother Employed in 1974	--	-.064 (.060)
No. of Siblings	--	-.003 (.025)
Father Unskilled Manual in 1974	--	-.097 (.172)
Father Occupation Missing	--	.919 (.192)
Region Group 1	--	-.014 (.074)
Region Group 2	--	.057 (.093)
Region Group 3	--	-.083 (.091)
Constant	--	-3.485 (.197)

Table B2 (continued)

	OLS	2SLS
δ		
Adverse Financial Shock	--	-.300
		(.082)
Parental Interest	--	.241
		(.054)
No. Older Siblings	--	-.065
		(.032)

Notes:

Standard errors in parentheses
2SLS corresponds to Table 1, Column (1)

Figure 1: Histogram of Predicted Participation Rates, First Step Probit

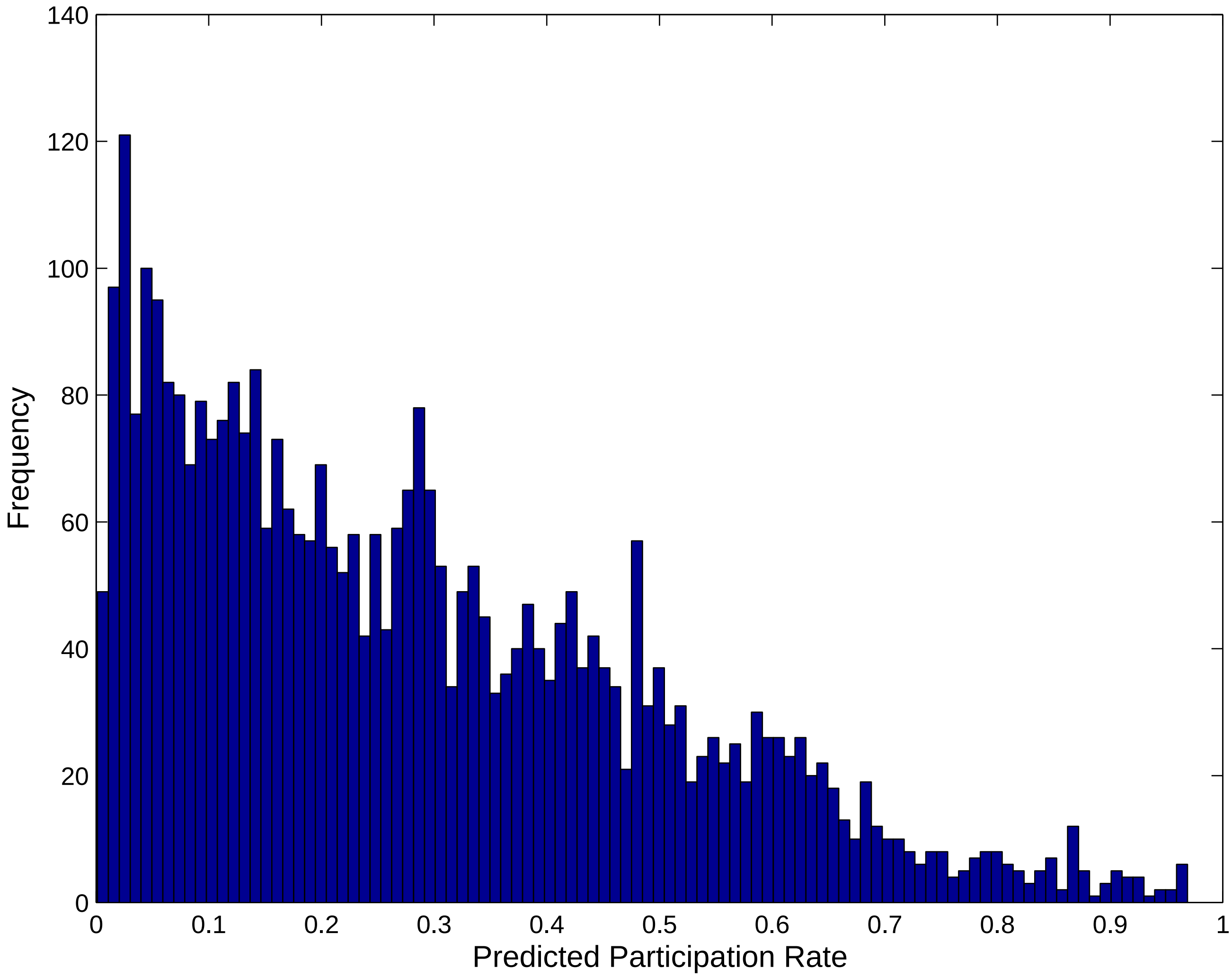
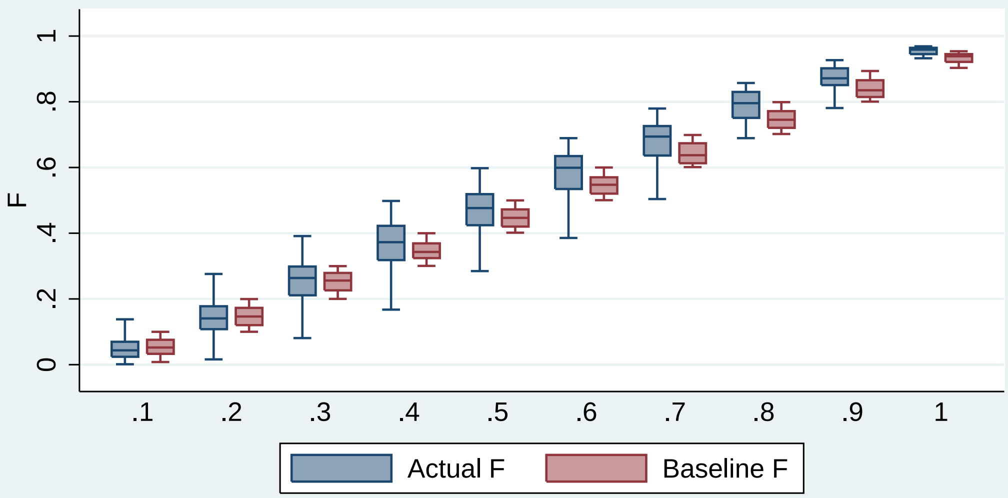


Figure 2: Baseline and Actual F Distribution
at Deciles of Baseline F



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

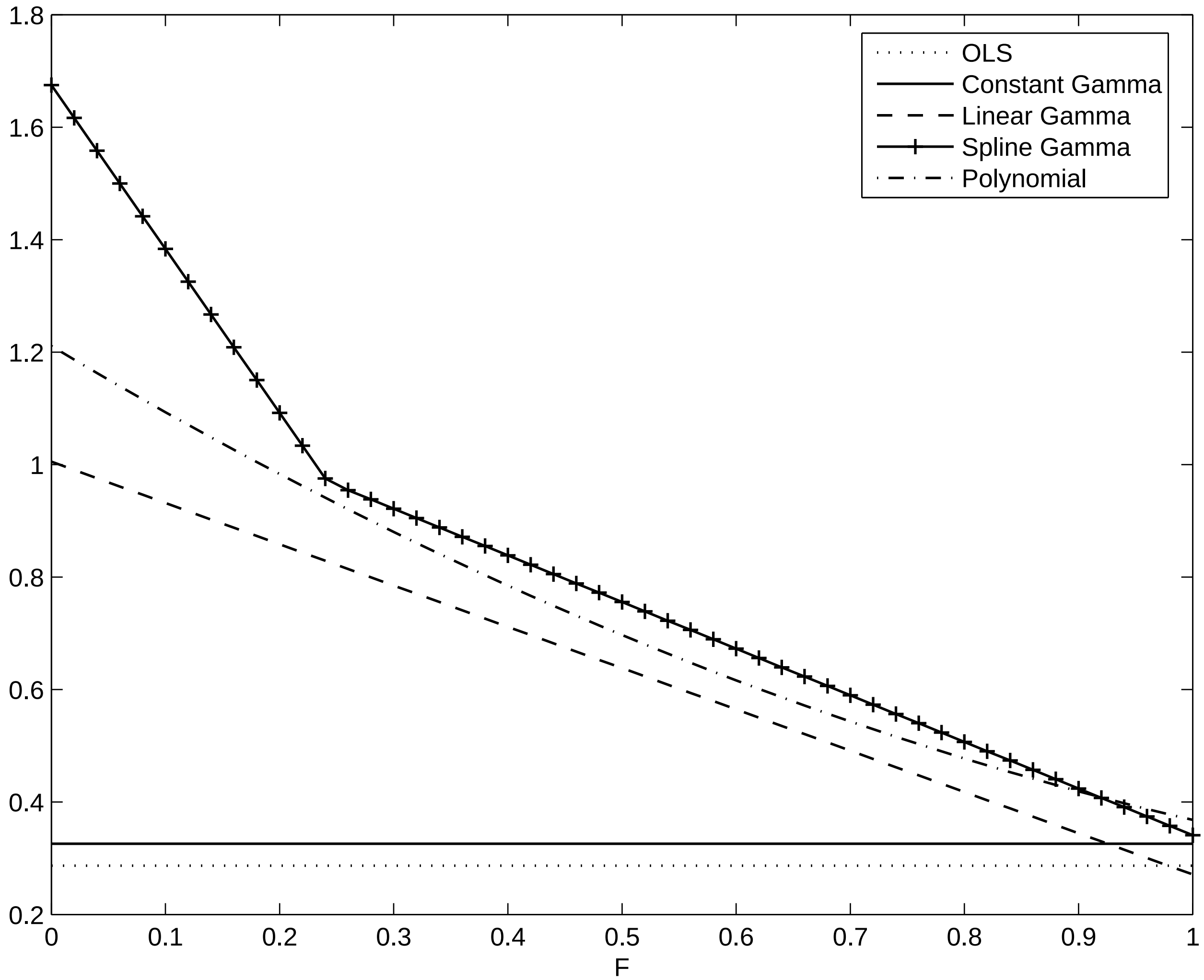


Figure 4: MTE for Different Models

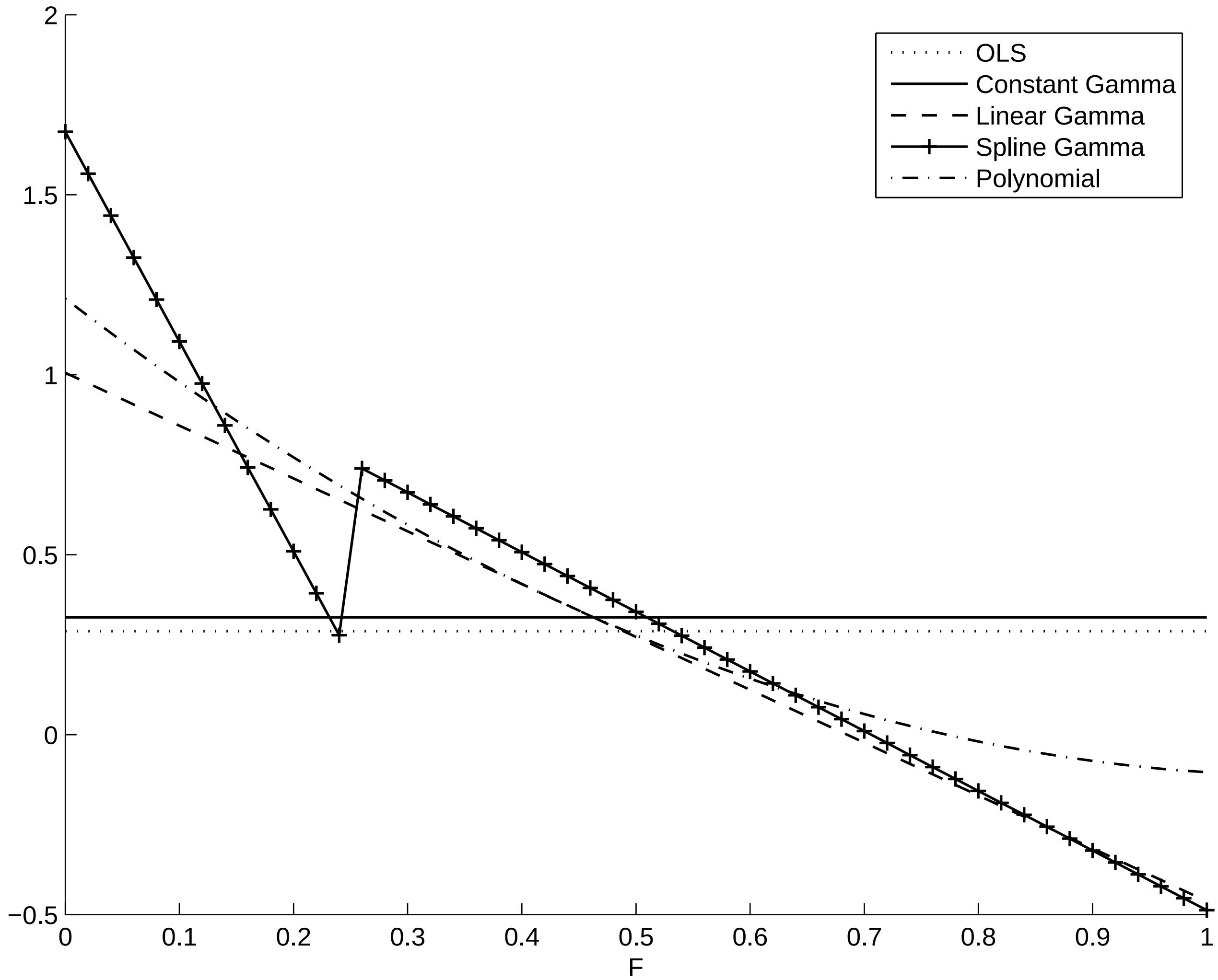


Figure 5: 95% C.I. for MTE of Constant Gamma Model