

Working Paper 99-10
Department of Economics
University of North Carolina, Chapel Hill
<http://www.unc.edu/depts/econ>

Overfitting and Biases in Nonparametric Kernel Regressions Using Cross-Validated Bandwidths: A Cautionary Note

Thomas A. Mroz
Timothy H. Savage

September 1999

ABSTRACT

Using Monte Carlo experiments, we examine the performance of nonparametric kernel regression methods. With cross-validated bandwidths and normal kernels, our experiments find that the kernel estimator fails a simple test that should be passed by classically unbiased estimators. The implicit functional form of the nonparametric kernel regression, in finite, realistic sized samples, is actually quite restrictive; it fails to capture well either a classical OLS model or a simple discrete outcome model with multiple regressors. We also find that nonparametric regression techniques tend to overfit data for sample sizes common in microeconomic analyses, and we present two measures that help researchers to gauge the extent of the overfitting. Researchers should be exceptionally wary of interpreting estimates from the nonparametric regressions as conditional expected values in multivariate settings.

Thomas Mroz is a professor of economics at the University of North Carolina at Chapel Hill and a fellow of the Carolina Population Center. Timothy Savage is a Senior Associate at Charles River Associates, Inc. The authors thank Hyungtaik Ahn for his programs, data, and advice. Ron Gallant, two referees, and the associate editor also made important contributions to this paper.

INTRODUCTION

During the last decade, economists have begun to use a variety of semiparametric and nonparametric regression techniques in empirical analysis. This is due both to advances in theory and to the increases in computational speed that make such estimation feasible. This journal (Winter 1998) recently published papers from a symposium on the application of these techniques in microeconomic analyses. In that symposium, Blundell and Duncan (1998) present an overview of nonparametric regression theory and provide several applied examples. Yatchew (1998) also provides a recent review of the theory and discusses additional applied examples.

Unlike the favorable reviews of these techniques by Blundell and Duncan and Yatchew, our experiences with the nonparametric kernel regression estimators have been disappointing. We find that the nonparametric techniques fail to perform as well as might be expected if one were to rely on these favorable reviews. Many of the properties that researchers associate with ordinary least squares regression functions and residuals do not come close to being true for nonparametric kernel regression functions and residuals in realistically sized samples. For example, the residuals from the nonparametric regressions are often strongly correlated with the predicted values. This makes it difficult to interpret the nonparametric regression as a conditional expected value. We first encountered some of these problems in an analysis of the effect of union membership on wages using both semiparametric and nonparametric techniques (Mroz and Savage, 1997). The problems we encountered, however, are not limited to that specific example, as the Monte Carlo evidence we present here demonstrates.

We find, when the true underlying model is either a linear regression or a probit model with a linear index, that the nonparametric kernel regression with a normal kernel and a cross-validated bandwidth yields predicted values that possess quite undesirable small sample properties. In particular, we find that the implicit functional form of the kernel regression model in “small” samples fails to capture the standard interpretation of a predicted value as an expectation of the dependent variable conditional on the explanatory variables. At the same time these fitted values fit the data too closely, in the sense of having higher predictive power than is theoretically possible for a conditional expectation. These problems are mitigated somewhat with larger sample sizes, but even with 10,000 observations and only five regressors the bias and “overfitting” are substantial. Using standard cross-validated bandwidths with normal kernels, we think it unlikely that there will ever be a social science data set of sufficient size to permit an accurate analysis of any but the most simple multivariate phenomena.¹

Below we describe two real world examples and several Monte Carlo experiments that illustrate the problems we encountered when using these techniques in practice. The purpose this paper is not to provide a description of the advantages and disadvantages of various parametric, semiparametric, and nonparametric estimators in a variety of settings in a manner that could help one to choose among alternative approaches. Instead, we document and discuss the shortcomings when one uses a standard nonparametric procedure in applied settings that are very familiar to all empirical economists. Hastie and Tibrishani (1990, p. 35) suggest that a comprehensive Monte Carlo analysis of the finite-

¹ While Blundell and Duncan discuss general applications of these methods, their empirical examples focus only on univariate nonparametric functions. The magnitude of the problems we document is diminished with univariate models, but it can still be substantial. The issues we raise are quite evident when the dimensionality is increased even to two or three explanatory variables.

sample characteristics of these techniques “would be very useful.” Our study provides important evidence that these estimators need much further evaluation before researchers can use them confidently in practice.

SPECIFICATION OF THE NONPARAMETRIC ESTIMATOR

Before turning to the example that motivated this study, we briefly summarize the mechanics of nonparametric kernel regression.² Suppose a researcher has a sample of size N of data points (g_i, z_i) , where g_i is the outcome and z_i is a vector of covariates. The outcomes, g_i , may be either discrete or continuous and depend on some function of the covariates, z_i . One is interested in estimating nonparametrically the conditional mean function, $E(g|z)$. Using a bandwidth-dependent kernel K_h , the nonparametric predicted value for outcome g_i with covariates z_i is:

Equation 1

In practice, the most common kernel is the multivariate normal kernel.³ This kernel often takes the form:

² Both Blundell and Duncan and Yatchew provide a comprehensive discussion of kernel based nonparametric regression together with a thorough list of references.

³ See, for example, Ahn and Powell (1993).

Equation 2

where u_k is the k^{th} element of the vector u , and h is the bandwidth. Often one uses orthonormalized explanatory variables and a single bandwidth in the kernel. One could replace this multivariate kernel with a multivariate normal distribution function that allows for unequal “variances” and “correlated” u_k 's. Such modifications, however, are not necessary for deriving asymptotic properties. Asymptotic theory provides some guidance on the rate at which the bandwidth approaches zero in large samples in order to guarantee consistency. It is silent, however, on how to choose a bandwidth in finite samples or whether the explanatory variables should be orthonormalized prior to use.

A standard method for choosing the single bandwidth is cross validation. As noted by Blundell and Duncan, cross validation is “asymptotically optimal with respect to the mean squared error” of the prediction. To solve for the cross-validated bandwidth, define:

Equation 3

This is the predicted value of outcome g_i when the i^{th} observation is omitted from the nonparametric regression function and the bandwidth is h . Using the predicted values for all observations, define the mean squared error of the prediction (MSEP) as:

Equation 4

The cross-validated bandwidth is the h that minimizes $MSEP(h)$. Later in the paper we will refer to the predicted values using Equation 3 with the cross-validated bandwidth as the “cross-validated” predicted values.

AN ILLUSTRATIVE EXAMPLE

Estimating the return to union membership on wages occupies a large place in empirical microeconomics. Lewis (1986), for example, cites over 200 separate analyses of the union wage effect. Perhaps another 30 studies have been published since then.⁴ To measure this effect, researchers typically regress an hourly earnings measure on a dummy variable indicating union membership, together with demographic and human capital controls.

In more recent studies, it is standard to use some type of two-step method to correct for potential selection bias due to the nonrandom allocation of union/nonunion members. Often this correction takes the form of a first-step probit to estimate the probability of union membership and a second-step least squares regression using the first-step predicted probability of union membership rather than the observed dummy variable.⁵

Table 1 displays estimates of the union wage effect using OLS and the probit two-step method described above.⁶ It also displays an estimate using a nonparametric two-step method. With this latter method, the first step obtains a nonparametric predicted value of union membership. This involves finding the optimal bandwidth via cross validation and

⁴ See, for example, Robinson (1992).

⁵ The model we discuss here is quite simple, with union membership only changing the intercept of the log wage regression. This sort of specification is quite common in the literature. Both linear two-stage least squares and the Heckman two-step method are also customary. Our preferred specification for the wage function is a switching regression model (Mroz and Savage, 1997). Consequently the simple dummy shift estimates reported here should only be considered illustrative.

⁶ The data for these estimates are white, prime-age males (aged 30-50) from the 1990 March Supplement of the Current Population Survey with positive weeks worked and positive earnings. The sample size is 3,325. The dependent variable is log average hourly earnings. The other regressors in the log wage equation are quadratics in age and education, an age-education interaction, marital status, whether a man lived in a state with a right-to-work law, city and region dummies, and a dummy variable for a private-sector employment.

then generating the nonparametric predicted value. In the second step, this nonparametric predicted value rather than the dummy variable is used in the (log) wage regression.

Table 1: Estimates of the Union Wage Effect

Standard errors in parentheses

OLS*	Probit Two Step*	Nonparametri c Two Step*
0.146	0.060	0.400
(0.018)	(0.230)	(0.118)

*Standard error based upon a jackknife estimator.

The OLS estimate closely agrees with what is found in the literature when not controlling for the potential endogeneity of union membership, with wages being about 15% higher for union workers. After controlling for potential selection bias using two-step methods, we obtain much different point estimates. With the two-step probit,⁷ the point estimate falls to six percent.⁸ The standard error, however, rises by more than a factor of ten, to 0.23. At a standard 95 percent significance level, one could not reject either the hypothesis that unions cause wages to *decline* by a third or the hypothesis that unions increase wages by over 50 percent. Such imprecision is not uncommon in union wage studies, and this lack of precision has lead some prominent researchers to question attempts to control for the endogeneity of union membership (e.g., Lewis, 1986 and Freeman and Medoff, 1984).

A first stage kernel regression model for the indicator of union membership, by relaxing several strong assumptions, might improve the precision of the two step estimator. This more refined approach could predict better the dummy variable, which could lead to both less biased and more precise estimates of the union wage effect. This

⁷ Preliminary tests indicated that measures of the number of children in the household were not a significant determinant of wages, but they did appear to influence union membership significantly. Consequently we used these measures as the “identifying instruments” in the first stage.

⁸ This may be due to the positive selection mechanism discussed in Lewis (1986).

approach would also drop the reliance on the linear functional form and a homoscedastic normal distribution assumption inherent in the first stage probit model. One could consider the nonparametric two-step method to be a “gold standard” since it imposes no distributional assumptions in either step of the estimation procedure. Applying the nonparametric kernel regression instead of the probit model in the first stage yields a second stage union effect estimate indicating a large, positive return to union membership. Unions raise workers’ wages by about 40% with a t-statistic over 3.0. With these results, a researcher could reject, at the five percent level, the hypothesis that unions raise wages by as little as 15 percent. We were quite surprised that the nonparametric procedure in the first stage would yield such markedly different estimates and inferences about the impact of unions on wages from the first-step probit model. The dramatic differences in these two estimates that control for selection prompted us to examine the performance of the nonparametric method in a simple dummy-shift model.

MONTE CARLO EVIDENCE ON A SIMPLE DUMMY-SHIFT MODEL

Consider the following system of equations in which a latent index, η_i , influences a continuous outcome, c_i , through its discrete realization, d_i .⁹

Equation 5

Equation 6

⁹ For notational ease, the subscript i is omitted.

The disturbance terms have an error-components specification with a factor, v , and factor loadings, ρ_d and ρ_c :

Equation 7

Equation 8

where ε_d and ε_c are distributed as independent $N(0,1)$ random variables that are also independent of the random factor v . This error-components specification forces all correlation in the joint density of d and c to enter solely through the factor v .¹⁰

The factor v can be used to induce both error correlation and deviations from normality on the joint density. Let v follow a compound normal distribution such that

The implied distribution of v is both skewed and bimodal. In this case, d and c still have expectation zero, and d is statistically endogenous. The joint density of d and c is not normal, and the marginal distributions are convolutions of a normal random variable and a compound normal random variable. The probit two-step method is inconsistent because the first step, being based on the assumption of a normal distribution, is specified

¹⁰ See Mroz (1998) for a thorough discussion of this error-components specification. A bivariate normal distribution is a special case of this specification.

incorrectly. The nonparametric two-step estimator, however, should be able to control for both the endogeneity and the non-normality of disturbances.

To preserve the same form of the marginal distributions but with independence of the errors ϵ_{it} , one can replace the single common factor v in Equations 7 and 8 with two independent equation specific factors v_{it} . In this instance, with v_{it} independent of each other, the endogeneity problem vanishes, and OLS is consistent in the second step. The two-step probit method, however, continues to be inconsistent due to the remaining nonnormality of ϵ_{it} in the latent index. The estimator of δ using the kernel regression to predict the dummy variable in the first stage is, of course, consistent. We examine the performance of OLS and the two-step estimators (with a probit and a kernel regression used in the first stage to predict the dummy variable) with and without endogeneity of the dummy explanatory variable in Equation 6.

Tables 2 and 3 present Monte Carlo results using a data generating process that follows Equations 5 through 8 under the above assumptions. The true intercept, α , equals 0, and the true impact of the dummy variable, δ , is 1. We set $\rho_d = \rho_c = 0.25$ in the data generating process. These results are based upon a sample size of 1,000 and five z 's generating the latent index in Equation 5. These z 's are independent, Chi-squared random variables with 5 degrees of freedom; each z is normalized to have mean zero and variance one. Each table contains the mean and standard deviation of 1,000 replications from the bivariate regression of the outcome c on a constant and d in Equation 6. Table 2 presents results in which d is endogenous, while Table 3 presents results in which d is exogenous.

Note that the estimated standard errors for the means reported in these tables are the corresponding standard deviations divided by 32 ().

Table 2: Monte Carlo Results with Endogenous Dummy Explanatory Variable
Means and Standard Deviations of Estimates of Equation 6 (1000 Replications)
Truth: Intercept (α) = 0; Dummy (δ) = 1
Latent index is non-normal

	OLS	Probit Two Step	Nonparametric Two Step
Intercept	-0.286	-0.001	-0.210
tSt. dev.	0.066	0.077	0.084
Dummy	1.596	1.006	1.489
St. dev.	0.089	0.128	0.145

Table 3: Monte Carlo Results with Exogenous Dummy Variable
Means and Standard Deviations of Estimates of Equation 6 (1000 Replications)
Truth: Intercept (α) = 0, Dummy (δ) = 1
Latent index is non-normal

	OLS	Probit Two Step	Nonparametric Two Step
Intercept	0.001	0.003	-0.093
St. dev.	0.064	0.078	0.083
Dummy	0.996	0.994	1.236
st. dev.	0.092	0.128	0.144

As expected, the OLS estimates are quite biased in the presence of an endogenous dummy variable (Table 2). The mean estimated intercept (standard deviation) for OLS is -0.29 (0.07) and the mean dummy shift effect is 1.60 (0.09), indicating a 60% upward bias in the structural shift parameter. When d is exogenous (Table 3), of course, the bias in the OLS estimator vanishes. The mean estimates in this case are 0.00 (0.06) and 1.00 (0.09). The probit two-step estimator does quite well in both cases, showing little bias despite its inconsistency. These results indicate that the probit two-step estimator can control well for the endogeneity in this example. It is largely unaffected by its misspecification in the first step in this simple data generating process.¹¹

What is remarkable about these results, however, is the poor performance of the nonparametric two-step estimator. In the face of endogeneity, it has an upward bias of almost 50 percent, performing nearly as poorly as the OLS estimator.¹² The mean difference (standard deviation) of the structural shift estimates for OLS and the nonparametric procedure's estimates in the 1,000 replications is 0.11 (0.09). Using the standard deviations as measures of the standard errors obtained in practice, a researcher finding these mean estimates and testing for equality of the OLS and nonparametric two-

¹¹ These Monte Carlo results accord closely to those found in Mroz (1999).

¹² For each estimation technique, the same set of Monte Carlo data was used for a given replication. This allows direct comparison, replication by replication, of the point estimates.

step estimators would not be able to distinguish between them. In other words, the nonparametric two-step estimates would indicate that there is no problem of endogeneity when in fact there is a large problem.

Even when the dummy variable is exogenous, the nonparametric two-step estimator has a bias of nearly 25 percent. Using the standard deviations from the Monte Carlo experiments to approximate the standard errors of the difference between these estimators as above, if one obtained average estimates of the slope coefficients one would falsely reject the null hypothesis of exogeneity at conventional levels. These limited Monte Carlo results prompted us to be suspicious of the nonparametric estimate of the union wage effect in Table 1. They also prompted us to do a somewhat more thorough Monte Carlo analysis of the performance of nonparametric regression techniques.

EXPERIMENTAL DESIGN OF MONTE CARLO EXPERIMENTS

We consider two, standard, one equation econometric models with exogenous regressors. The first is a simple binary (or discrete) outcome model. In this case, the true data-generating process (DGP) we examine is a probit model with linear covariates and an independent, normally distributed error. The second is a classical linear regression model (OLS) also with linear covariates and an independent, normally distributed error. We focus on the performance of the nonparametric kernel regression models when predicting outcomes for these two DGPs. For the discrete outcome, the true DGP is:

Equation 9

For the continuous outcome, the DGP is:

Equation 10

The vector of covariates z is distributed as $N(0, I_k)$. The error terms are distributed as $N(0, 1)$ and are independent of the covariates. We define the Index R^2 for a particular model to be:¹³

Equation 11

The coefficients are set equal to some common value η , and we adjust η to achieve a particular value of the Index R^2 :

Equation 12

We examine two particular values of the Index R^2 , 0.25 and 0.75. The constant in the probit model, δ_{it} , is used to obtain a desired frequency of zeroes and ones. Here we choose it such that .¹⁴ In the continuous outcome model we set .

For each of these DGPs, the standard parametric estimator, probit or OLS, is known to be asymptotically efficient, in the sense of asymptotically being unbiased and achieving the Cramer-Rao lower bound.

¹³ This measure is exactly the definition of the R^2 in the linear regression model. It is also the R^2 from a regression of the latent index on the explanatory variables. See, Estrella (1998). In a regression of the dummy variables d on the true probabilities that $d = 1$, the resulting “ R^2 ” would be smaller than that for the latent index regression.

¹⁴ The unconditional expectation, $E(d)$, equals 0.280 for the lower R^2 and 0.368 for the higher R^2 .

A SIMPLE TEST

Elementary statistics implies that one can always write a regression function:

Equation 13

where, by definition, $\mu(z)$ is mean independent of z and hence uncorrelated with z ,

provided the expected value exists. Suppose one had an estimate of $\mu(z)$, denoted

$\hat{\mu}(z)$. By analogy, one could estimate the regression model:

Equation 14

If the conditional mean estimator is unbiased and fairly precise, it is reasonable to expect the estimates of α and β to be close to zero and one, respectively.¹⁵ Our simple test is to examine whether the predicted values from the nonparametric estimator satisfy these conditions.

To illustrate this simple test with real data in the context of discrete outcomes, consider Ahn and Powell's (1993) semiparametric estimation of Mroz's (1987) female labor supply model. As the first step in a two-step procedure to control for nonrandom selection, Ahn and Powell use a normal kernel and a cross-validated bandwidth with orthonormalized data to estimate the probability of labor force participation. Their first-step method is identical to that described above for discrete outcomes.¹⁶

¹⁵ Classical measurement error in constructing the conditional expected value, for example, results in an upward bias in α and a downward bias in β .

¹⁶ We replicate the results in Ahn and Powell when using the bandwidth they report. However, we find a slightly smaller optimal bandwidth (with a smaller MSE), and we use it here.

When we estimate Equation 14 via OLS using the predicted values from their first-step kernel regression as explanatory variables, we obtain:

where D is the dummy variable of labor force participation and \hat{y} is the predicted value of y from the kernel regression.¹⁷ Given the estimates of α and β and their standard errors, one would be hard-pressed to call these nonparametric kernel estimates conditional expected values. Clearly, these nonparametric estimates are difficult to interpret as standard regression functions or as typical predicted values.

The poor performance of these predicted values, however, is not due to a poor fit of the data. The log-likelihood value associated with the kernel regression is -362.6 , where the log-likelihood function is defined, as in a probit or logit model, to be:

This is a 35-point improvement in the log-likelihood value of the parametric probit specification used in Mroz, which is -398.2 . Using predicted values from a parametric probit as in Mroz (1987), our simple regression test yields:¹⁸

¹⁷ Jackknife standard errors.

¹⁸ Jackknife standard errors.

Note that the nonparametric estimator has an R^2 80 percent larger than that of the parametric estimator.

It is important to note that the centered R^2 measures reported here can inflate the goodness of fit measure for the kernel regression model. In particular, there is no guarantee of orthogonality between the kernel regression residuals and the predicted values in finite samples. The correlation only disappears in infinite sized samples. Thus, the explanatory power of the simple kernel regression might be overstated by this metric. The standard R^2 measure, however, may be an appropriate metric to consider if one plans to use the predicted probabilities to control for endogeneity or in other situations where there could be linear transformations of the predicted probabilities. The union wage model examined above with structural shifts is an example of a commonly occurring situation with precisely that property. Also, since the kernel regression adjusts perfectly for any mean shift in the dependent variable,¹⁹ it may not be unreasonable to focus on models that allow for an intercept when examining whether the kernel regression can provide reasonable finite sample estimates of conditional expected values. We never found instances where the mean of the kernel regression predicted values failed to provide a reasonable estimate of the unconditional mean of the dependent variable, but it almost always failed to describe well the variations about the mean.

MONTE CARLO TEST RESULTS

Clearly, the nonparametric approach fits the data “better” than the parametric approach. But does it fit the data too well? Also, does the inability of the kernel

¹⁹ Adding any constant μ to each outcome, g_j , in Equation 1 would shift each estimated expected value by μ exactly.

regression to capture well a conditional expected value seem to be peculiar to the two real world data sets we examined above? To examine these issues, we use Monte Carlo experiments. The explanatory variables and the errors are independent and identically distributed normal random variables in the DGPs described above. Given these specifications, we know precisely the theoretical maximum R^2 that could be obtained by conditioning only on the explanatory variables and the precise values of the conditional expected values.

Tables 4 through 7 contain the experimental results of our simple test. Tables 4 and 5 examine discrete outcomes, while Tables 6 and 7 examine continuous outcomes. We consider three different sample sizes (1,000, 5,000 and 10,000), denoted N , and five different values for the number of regressors (1, 2, 3, 4 and 5), denoted K . Tables 4 and 6 set the Index R^2 to 0.25, while Tables 5 and 7 set it to 0.75. It is important to note that, for the discrete DGP examined in Tables 4 and 5, the R^2 corresponding to the linear regression of the dummy variable on the true probabilities will be less than the Index R^2 . In particular, the R^2 when using the true predicted values from probit estimation is 0.148 for Table 4 (Index $R^2=0.25$) and 0.534 in Table 5 (Index $R^2=0.75$). In Tables 4 and 5, the expected frequency of ones is 0.25 when $z = 0$.²⁰

Table 4 reports the means and standard deviations from 100 replications of the DGP for the cross-validated bandwidth h and the means and standard deviations of α , β and R^2 in Equation 14. Also reported is a measure of the effective degrees of freedom (DF) which will be discussed in the next section. As expected, the optimal bandwidth

²⁰ We examined specifications in which the frequency of ones is 0.50 and found similar results. For the most part, only the intercepts α changed in Tables 4 and 5 as we varied $E(d|z=0)$.

decreases with N (*ceteris paribus*) and increases with K . In each of the 15 different specifications in Table 4, the mean of α is significantly less than zero and the mean of β significantly greater than one at conventional levels, the exact obverse of what one would expect if the only problem with the predicted values were measurement error. This is true even under the most ideal circumstances examined here: low dimensionality and large sample sizes. The performance of β is quite miserable under more realistic circumstances that are faced by applied researchers. When the Index $R^2 = 0.25$ and $K = 5$, its average value varies between 1.50 and 1.83 depending on the sample size. As can be seen in the R^2 column, these nonparametric predicted values fit the data better than is theoretically possible given the DGPs. This overfitting occurs even with only one or two explanatory variables. At higher dimensions, the R^2 fit of the nonparametric estimator can exceed the theoretical maximum by up to 50 percent. The results are qualitatively similar for the higher Index R^2 value used in Table 5.

This problem is not limited to discrete outcomes. Tables 6 and 7 contain similar information for the continuous outcome case. The results in these tables are qualitatively similar to those in Tables 4 and 5.²¹ In all instances, the average β is significantly greater than one.²² All of the R^2 values exceed the theoretical maximum. The supposed “fit” of the data increases substantially with additional regressors. Even at sample size 10,000, the nonparametric estimator “fits” the data by up to five percentage points greater than should be possible.

²¹ We set the intercept to zero for the continuous outcomes. This affects only the estimate of α in Table 2. The estimates of β and their standard deviations, as well as the cross-validated bandwidth and the R^2 are not affected by this linear transformation.

²² Note that the overall mean of the dependent variable is zero in Tables 6 and 7. The small estimates of α in these tables indicate that the kernel regression model fits the mean of the data quite well. Not surprisingly, restricting α to be zero had little impact on the estimates of β or the R^2 measures.

To contrast these test results with those from the parametric counterparts of probit and OLS, Table 8 contains selected Monte Carlo results when $R^2 = 0.25$. OLS, by definition, guarantees that α equals zero and β equals one. With these two parametric estimators, there is no indication of overfitting or inappropriate expected values. This is not surprising.

THE EFFECTIVE DEGREES OF FREEDOM

It is clear from the R^2 in Tables 4 through 7 that there seems to be a substantial amount of overfitting. By this, we mean that the nonparametric regression is able to explain more of the non-systematic variance than should be possible. Another metric we use to measure this overfitting is that provided by Hastie and Tibirishani (1990, pp. 52-55).²³ They discuss the effective degrees of freedom (DF) of nonparametric regression with respect to the amount of fitting that is done by the regression. The greater is the DF, the closer is the fit of the regression. Given the fundamental bias-variance tradeoff of nonparametric regression, the greater is the DF, the smaller is the bias but the greater is the variance.

The nonparametric estimator in Equation 1 can be expressed in matrix form as follows.

Equation 14

²³ Their discussion is in terms of the class of linear smoothing estimators, which includes OLS and kernel regression. The probit estimator is not a linear smoothing estimator. Nonetheless, in the following discussion we refer to its “effective degrees of freedom” as if it were an OLS estimator.

Equation 15

Examination of $S(h)$, the nonparametric projection matrix, shows that considerable weight can be placed on the own observation (the diagonal elements) relative to that

placed on other observations (the off-diagonal elements). This occurs when the value of $K_h(0)$ is large relative to the value of the kernel evaluated at the other points in the sample.

The trace of $S(h)$ is:

Equation 16

Following Hastie and Tibshirani, we define the effective DF of an estimator to be $\text{trace}[S(h)] - K$, where K is dimension of z . The trace of the OLS projection matrix, P , is always K with linearly independent regressors,²⁴ so its effective degrees of freedom is zero. The last column of Tables 4 through 7 displays the mean and standard deviation of the effective DF from the 100 Monte Carlo replications. One can loosely interpret the effective DF as the number of “additional regressors generated” by the kernel regression procedure (e.g., “powers” and “interactions” of the z).

The DF metric strongly mirrors the R^2 metric. It rises substantially as additional regressors are included in the model. In Table 4 at sample size 10,000, for example, as the number of regressors increases from one to five the average effective DF rises from 17 to 397. A fivefold increase in the number of regressors increases the effective DF increases by a *factor* of 23. For the higher Index R^2 in Table 5, the average effective DF rises from 27 to 1267, a *factor* of nearly 47, for sample size 10,000. Tables 6 and 7 contain the nonparametric results for continuous outcomes and so can be directly compared to results for OLS. Recall that for OLS, the effective DFs are always zero. At the lower R^2 in

²⁴ Or $K+1$ when there is an intercept as well as K explanatory variables z .

Table 6, the effective DF rises by a *factor* of nearly 33 as the number of regressors rise by a factor of five for a sample size of 10,000. For the higher R^2 in Table 7, they rise by a *factor* of nearly 62.

Another interpretation of the trace of the projection matrix further supports the notion that it is a measure of the amount of overfitting of the regression model. In the kernel regression model, the sum of each row of the projection matrix equals one by definition, so the $\text{trace}[S(h)]/N$ is a measure of the average of the weight put on a randomly chosen observation i 's dependent variable when calculating the predicted value for observation i . When an OLS regression model contains an intercept, each row of the projection matrix also sums to one. In this case the $\text{trace}[\quad]/N$ measures the average weight put on observation i 's observed outcome when constructing observation i 's predicted value, and it equals exactly K/N .

Looking at the rows for five explanatory variables in Tables 4 through 7 reveals the importance of each observation's actual observed outcome as a determinant of the predicted outcome for itself in the nonparametric kernel regression. For discrete outcomes, Index $R^2 = 0.25$, and sample size 1,000 (Table 4), we see that the predicted values on average are constructed using a weight of eight percent, or $(75.61+5)/1000$, on the "own" observation's observed outcome. There is a similar average weight on the "own" observation of almost 16% in the continuous outcome model with $R^2 = 0.75$, 10,000 observations, and five regressors (Table 7, $[(1575.25+5)/10000]$).

The kernel regression procedure, even with optimally chosen, cross-validated bandwidths, relies much more heavily on each observation's observed outcome than most parametric methods for calculating the expected value for that observation. This is exactly

what one would expect, of course, given that the kernel regression weights more heavily “nearby” observations in the explanatory variable space. The weights the normal kernel regression places on each observation’s own observed outcome, however, do seem extraordinarily large. This could severely impact the usefulness of the kernel regression procedures in controlling for endogeneity issues.

ADEQUACY OF THE KERNEL REGRESSION “FUNCTIONAL FORM”

Besides tending to rely excessively upon an observation’s observed outcome when constructing the predicted value of that outcome, the kernel regression procedure also fails to fit the data well in the simple models examined here. To display this failure of the approach, we examine a discrete regression model with 10,000 observations, five regressors, and an Index R^2 of 0.25. This data generating process corresponds directly to the results reported in the last row of Table 4. To facilitate the analysis we hold the set of explanatory variables fixed across twenty replications and examine the average predicted values across the 20 replications for each of the 10,000 sets of explanatory variables. We define the true kernel predicted value as \hat{y}_i , where \mathcal{Z} represents the set containing the explanatory variables for all observations, and z_i is the point at which we wish to evaluate this predicted value. Note that this expectation conditions on the DGP for the explanatory variables. We estimate this true predicted value by averaging the predicted values over the 20 replications for each set of explanatory variables, z_i , in the data.²⁵ This average can be quite noisy, and we also smooth these averages by taking

²⁵ Averaging the predictions across replications for each set of explanatory variables diminishes the dependence of the prediction on the “own” observed outcome in each replication.

moving averages using nearby observations when ordered by the true expected value of the discrete outcome for the DGP.

Figure 1(a) displays the average of the probit predicted probabilities and the nonparametric kernel regression predictions as a function of true probability for each of the 10,000 observations. We also display the “cross-validated” nonparametric predicted values. They are the \hat{p}_i in Equation 3, evaluated at the cross-validation selected bandwidth. The 45-degree line measures the theoretically ideal prediction; this line is almost indistinguishable from the line measuring the average probit probabilities across the 20 replications. Figure 1(b) uses a moving average of the average predicted probabilities displayed in Figure 1(a) to smooth the predicted probabilities.²⁶ Similar graphs for DGPs with continuous outcomes display substantively similar features.

These figures display the weakness of the nonparametric functional form that was suggested by the large slope coefficients in the simple test for a conditional expected value in Tables 4 through 7. Each of the average nonparametric prediction curves passes through the unconditional mean of the dependent variable about the unconditional mean of the true probability, but each curve has a much lower slope than the ideal value of one. There is no indication that the probit model fails to approximate well this theoretical ideal.

This functional form weakness of the non-parametric kernel regression is not due to the cross-validated bandwidth being a poor choice in finite samples. Table 9 contains information for the same 20 replications of the DGP just discussed evaluated at a variety

²⁶ The moving averages use equal weights for each of the 51 observations centered about each true probability. The vertical lines in these graphs indicate the points where it is not possible to construct a centered moving average because there are too few observations in the “tails” for the DGP.

of bandwidths. The bandwidths in this table are proportional to the bandwidths that were selected by the cross-validation approach (from one half to twice the cross-validated bandwidth). For each replication we evaluate the goodness of fit and the propensity to overfit at the indicated fraction of the cross-validated bandwidth. The row labeled 1.0, for example, contains information for the 20 replications (with “fixed” regressors) at the cross-validated bandwidth, and it corresponds to the last row of Table 4.

The slope measure reported in this table is the estimated regression coefficient obtained when the nonparametric prediction is regressed on the true probability (and an intercept). This is a different slope measure than that displayed in Tables 4 through 7. If the only shortcoming of the nonparametric predicted values were that they were noisy measures of the true predicted value, the slope coefficients in the second column of Table 9 should be close to one. At the cross-validated bandwidths (row 1.0) the average slope is about 0.77 with a small standard deviation across replications. Shrinking the bandwidth does make the slope increase towards 1, but the cost of this is an increase in the R^2 for the nonparametric model to well above the true maximum of 0.148. Note that in order to increase the slope to over 0.90, one must use almost one effective degree of freedom for every four or five observations in these samples of size 10,000, even though there are only five regressors used in the estimation. This set of results indicates that the implicit functional form of the nonparametric regression (with a normal kernel) might be too restrictive in realistic sized samples.

How Can Nonparametric Predictions Fit Too Well with a Poor Functional Form?

The normal kernel predictions fit too well precisely because the nonparametric procedure relies heavily on each observation's observed dependent variable in the construction of the fitted value for that observation. The effective degrees of freedom measures in Tables 4 through 7 suggested that this might indeed be a serious problem. To demonstrate the extent of this problem, Table 10 contains summary information on log-likelihood function values for the twenty replications used above in Table 9 and Figure 1. Each row of this table uses a different approach for defining the probabilities used to construct the log-likelihood value.²⁷ Note that only the probabilities used in rows 2), 9), and 10) could be calculated with a single real data set.

The first row of Table 10 indicates that the average likelihood function evaluated at the true probabilities is -5205 . Using the probit probabilities from each replication, the average log-likelihood increases by less than three points (row 2). The third row uses the average across the 20 replications of the probit predictions for each of the 10,000 sets of explanatory variables to define the predicted values at each of the 10,00 points, as in Figure 1 (a). We see that the average value of the likelihood function is nearly identical to that obtained using the true probabilities. Further smoothing of the probit estimated probabilities, by taking a moving average based upon the true probabilities (row 4) does little to impact the value of the likelihood function value.

Using the average of the kernel regression predictions to construct the probabilities (row 5, as in Figure 1(a)) reinforces the issue raised about the poor fit of the

²⁷ The log likelihood value for observation i is $[d_i \ln(P_{m,i}) + (1-d_i) \ln(1-P_{m,i})]$, where $P_{m,i}$ is the predicted probability of the event $\{d=1\}$ given observation i 's explanatory variables using method m to construct the conditional probability. The overall log-likelihood is the sum of these terms across individuals.

nonparametric regression model. The log likelihood value when one uses these kernel regression predicted probabilities is 36 points *lower* than that obtained with the true probabilities. One would expect the overall goodness of fit to improve or remain unchanged by reducing the noise in the predicted probabilities, but the results in row 6 indicate a somewhat large deterioration of the goodness of fit.

The averages of the “cross-validated” predictions from Equation 3 (row 7) also demonstrate a poor fit relative to the likelihood function value evaluated at the true probabilities. Using a moving average to smooth these “cross-validated” predictions (row 8, as in Figure 1(b)) slightly improves the value of the log-likelihood function value, in contrast to the decline observed for the smoothed, actual nonparametric predictions. This suggests that there might be much noise in these predictions even after averaging across the 20 replications.

The value of the likelihood function constructed from the actual predictions from the nonparametric model (Equation 1) is in the ninth row of Table 10, and the function value from the “cross-validated” predictions (Equation 3) is in row 10. The cross-validated predictions are clearly quite noisy, given the near thirty point improvement in the likelihood function value that can be obtained by averaging and smoothing these predictions (row 10 compared to rows 7 and 8).

The standard nonparametric predictions (row 9), as defined in Equation 1, increase the likelihood function value by more than 224 points over the probit model. These predictions also yield a 275-point increase over the likelihood function value from the smoothed and averaged kernel regression estimates. Merely allowing each observation to be included in the set of observations used to construct its own predicted value has a

substantial impact on the fit of the model. To put the increase in the log likelihood function into perspective, consider adding 400 regressors to the probit model and testing whether these 400 regressors could be excluded from the probit model. With an increase in the log likelihood of 224 one would just conclude, at a five percent significance level, that the additional 400 regressors significantly improve the fit of the model. The average effective degrees of freedom measure in Table 4 for this DGP implies a remarkably similar number, namely, 397.

When one averages predicted values from the nonparametric regressions for identical sets of explanatory variables across replications of the DGP, it is obvious that the kernel regression approach used with the data generating processes examined here does not fit the data very well. This failure to fit well does not appear to be due to a poor choice of bandwidth. Rather, on any given replication, the kernel regression predictions rely so heavily on each observation's dependent variable as a predictor of itself that the likelihood function value constructed from these predictions exceeds the known true value of the likelihood function value by absurd amounts. The kernel regression approach in any single data set will appear to provide an excellent fit, but the excellent fit is due solely to an over-reliance on each observation's observed outcome for predicting that outcome.

Further evidence on overfitting after controlling for biases due to the functional form comes from an examination of the DGP for the endogenous dummy shift model (Equation 6) used to construct Tables 2 and 3. Recall that there are five independent variables that influence the endogenous dummy variable but have no direct impact on the continuous outcome analyzed in these tables. Instead of using the prediction from the nonparametric regression directly as a regressor in the second stage, as was done for

Tables 2 and 3, consider using the nonparametric prediction as an instrumental variable when estimating Equation 6 by two stage least squares (TSLS). Because the nonparametric prediction is a function of the five instrumental variables, it should be a valid instrument for predicting the right hand side dummy variable in the “first stage” of the TSLS procedure. Additionally, this new procedure should eliminate the bias in the estimate of the dummy shift effect arising from the restrictive functional form of the nonparametric regression that was documented in the previous section.

Using 1,000 replications, the mean estimated structural shift effect (standard deviation; standard error of mean) from this three step estimation approach is 1.199 (0.117; 0.004) when the dummy shift is endogenous (as in Table 2), and 0.996 (0.115; 0.004) when the dummy shift is exogenous (as in Table 3). With an endogenous explanatory variable, the mean difference (standard deviation) between the two step probit estimator and this three-step estimator is -0.193 (0.051). Treating the parametric probit estimator as a special case of the nonparametric approach and without knowing the true DGP, one would most likely conclude incorrectly, in over 95% of the replications, that the probit estimator exhibits considerable bias, even though it actually has only an imperceptible bias.

The fact that this three-step estimator performs quite well when there are no endogeneity problems indicates that it is the endogeneity of the dummy variable, rather than any other feature of the DGP or the estimation approach, giving rise to the bias. By eliminating the bias due to the functional form of the kernel regression, the only remaining source of bias must arise from the dependence of the nonparametric predictions on the actual values of the endogenous dummy variables. Since the errors are uncorrelated

across observations, it must be the case that the heavy reliance of the prediction for each observation on the observed endogenous outcome for that observation gives rise to this additional bias. With this bias being nearly 20% of the true parameter value, we conclude that there is substantial overfitting in the first stage kernel regression.

Other Indications of Unreliability of the Normal Kernel, Nonparametric Regression

We carried out less complete Monte Carlo experiments for some minor variations of the data generating process. The most interesting results come from variations in the correlation structure of the explanatory variables, and how different levels of correlations interact with the choice of the basis used to span the space of the explanatory variables. It should be noted that variations in the correlation structure among the regressors can be generated by taking particular linear combinations of uncorrelated variables like those used above. Predicted values from both the OLS estimator and the probit estimator are, by definition, completely independent of such linear transformations of the explanatory variables.

The first set of additional experiments uses five regressors, equal and positively correlated explanatory variables, and identical coefficients on each explanatory variable. The DGPs for Tables 4 through 7 can be considered as special cases of this set of experiments, with the correlations of explanatory variables set to zero. Provided that the explanatory variables are not orthogonalized, at high levels of correlation (e.g., 0.8 or 0.9) the performance of the kernel regression model appears quite improved. Regressions of the observed dependent variable (for both the continuous and discrete outcomes) on the nonparametric predicted values yield slope coefficients much closer to 1 than those

reported in Tables 4 through 7. The reason for this improved performance of the nonparametric estimator appears related to the fact that most of the overall variation in highly and equi-correlated variables (with identical magnitude and signed effects) can be captured in a single “factor.” So with five highly correlated explanatory variables, the kernel regression can perform similarly to a kernel regression with only one or two explanatory variables. If one orthonormalizes the correlated explanatory variables, however, the slope coefficients appear quite similar to those reported in Tables 4 through 7. Nonparametric kernel regression appears quite sensitive to linear transformations of explanatory variables.

Next, we examine the consequences of different impacts for each of the five equi-correlated covariates on the dependent variable. In particular, we hold the R^2 constant but specify the regression coefficients so that the “factor” giving rise to the correlation among the regressors cancels out in the linear combinations, $z_i'\gamma$, in Equations 9 and 10. In these experiments, the predictions from the nonparametric regression model appear quite similar to those reported in Tables 4 through 7. Additionally, unlike the above case where each of the correlated covariates had the same impact, the performance of the nonparametric predictions does not improve or deteriorate when one orthonormalizes the data.

As a final set of experiments, we examine negative and equi-correlated regressors with each covariate having an identical impact on the outcome. We use a level of negative correlation close to that which would just guarantee a positive definite correlation matrix of the regressors. Without orthonormalization of the explanatory variables, the fit measures appear much worse than those reported in Tables 4 through 7. For example, with a continuous outcome, four regressors, $R^2 = 0.25$, 10,000 observations, and pairwise

correlation of -0.30 instead of 0.00 (see Table 6), the slope coefficient rises from 1.28 to 1.62, the R^2 increases from 0.28 to 0.35, and the effective degrees of freedom climbs from 325 to 1170. However, if one orthonormalizes the explanatory variables, then the fit statistics appear quite similar to those reported in Tables 4 through 7. These results indicate that the nonparametric kernel regression depends crucially on simple linear transformations of the explanatory variables. In turn, these variations across transformations depend critically on precise form of the true relationship.

CONCLUSIONS

Although the Monte Carlo results presented here are somewhat limited, they are striking. We use commonly practiced nonparametric methods with normal kernels and cross-validated bandwidths, and we examine data generating processes for two classic econometric models. The results indicate that these nonparametric methods fail a simple test that should be passed by accurate, unbiased estimators. This failure appears due to the nonparametric procedure having a restrictive functional form in finite samples. These methods, however, also overfit the data, often grossly so. This overfitting problem is related to the “curse of dimensionality,” and adding more explanatory variables exacerbates the problem. These are not minor deficiencies, and they have substantive impacts on the two real world data sets we examined. Without serious modifications and extensive Monte Carlo analyses, we believe that these methods will not be useful for even simple, multivariate analyses.

We have, of course, only compared the nonparametric approach to the “gold standard” estimation procedures for the data generating processes we examined.²⁸ This is clearly a shortcoming of this study, as there surely are situations where nonparametric regression will outperform naïve OLS and probit approaches. Nonetheless it is crucial for applied researchers to recognize that many of the desirable properties associated with standard parametric estimators do not carry over to the nonparametric estimators. Predicted values appear systematically “biased,” regression residuals can be correlated with predicted values, and linear transformations of the explanatory variables can have substantive impacts on the estimators. Predicted values are also “too good,” as the nonparametric predictions sometimes come close to behaving as if the observed dependent variable were its own expected value conditional on the exogenous variables.

These are not trivial problems. The fact that they appear in simple models like those examined here, regardless of how the nonparametric estimators compare to OLS and probit, should make researchers hesitant to rely confidently upon these approaches. In Newey, Powell and Vella’s (1999) recent paper on two-step nonparametric estimation, for example, the empirical example for their preferred model uses a standard OLS estimator in the first stage in conjunction with polynomials augmenting other linear regressors in the second stage. The nonparametric aspect of their analysis is a cross-validation approach to choosing the degrees for two separate polynomials. The study examines up to fifth order polynomials each (at most 10 degrees of freedom) to capture the nonlinear features of the model. They use these parametric procedures because of the

²⁸ This is not entirely correct. The first set of Monte Carlo experiments reported compare the nonparametric approach to an inconsistent parametric estimator. The experimental results in this instance indicate that the parametric estimator outperforms the nonparametric estimator.

risk of the “curse of dimensionality.” It is telling that a major paper on nonparametric procedures uses as its applied example what most applied researchers would consider a parametric approach in the estimation.

We recommend that applied researchers follow this example and make more use of polynomial expansions and spline functions. Alternatively, it might be useful to consider using semiparametric, single and multiple index models (Ichimura, 1993; Ichimura and Lee, 1991). The reductions in dimensionality that these restrictions provide might help overcome many of the problems encountered with nonparametric kernel regression. Kernels other than the normal kernel might improve dramatically the performance of these approaches we studied, and techniques such as locally linear regression (Fan, 1993) might provide reliable nonparametric estimators. What is clear from our paper is that researchers should not rely upon black box, nonparametric kernel regression approaches and interpret their estimated regression functions as if they had estimated a classical regression function. In finite samples nonparametric regression actually relies on a restrictive parametric form and, in the cases we examined, the poor functional form can be masked by the tendency of the procedure to fit data much better than is theoretically possible for a conditional expectation.

Table 4: Discrete Outcome Monte Carlo Results

Index $R^2 = 0.25$
True Dummy Variable Regression $R^2 = 0.148$
 $E(d|z=0) = 0.25$; $E(d) = 0.28$

N	K	h (St. Dev.)	α (St. Dev.)	β (St. Dev.)	R^2 (St. Dev.)	DF (St. Dev.)
1000	1	0.2881 (0.0813)	-0.0255 (0.0074)	1.0995 (0.0292)	0.1561 (0.0219)	8.78 (5.03)
1000	2	0.4674 (0.0680)	-0.0704 (0.0124)	1.2701 (0.0483)	0.1649 (0.0261)	20.21 (8.81)
1000	3	0.6048 (0.0505)	-0.1223 (0.0188)	1.4600 (0.0658)	0.1810 (0.0290)	33.47 (9.63)
1000	4	0.7000 (0.0530)	-0.1723 (0.0211)	1.6572 (0.0808)	0.2094 (0.0355)	52.97 (14.74)
1000	5	0.7815 (0.0521)	-0.2257 (0.0315)	1.8344 (0.1206)	0.2374 (0.0407)	75.61 (20.60)
5000	1	0.2111 (0.0460)	-0.0124 (0.0036)	1.0485 (0.0143)	0.1500 (0.0107)	13.52 (6.01)
5000	2	0.3660 (0.0253)	-0.0383 (0.0035)	1.1478 (0.0141)	0.1528 (0.0101)	39.97 (6.71)
5000	3	0.4691 (0.0229)	-0.0704 (0.0049)	1.2692 (0.0159)	0.1657 (0.0116)	90.51 (14.57)
5000	4	0.5655 (0.0240)	-0.1112 (0.0067)	1.4203 (0.0240)	0.1792 (0.0132)	153.23 (24.42)
5000	5	0.6397 (0.0204)	-0.1545 (0.0082)	1.5790 (0.0306)	0.2010 (0.0155)	240.41 (33.21)
10000	1	0.1818 (0.0425)	-0.0093 (0.0028)	1.0360 (0.0112)	0.1474 (0.0067)	16.94 (7.76)
10000	2	0.3205 (0.0205)	-0.0290 (0.0026)	1.1123 (0.0100)	0.1527 (0.0069)	57.86 (8.87)
10000	3	0.4276 (0.0184)	-0.0573 (0.0026)	1.2182 (0.0098)	0.1587 (0.0072)	131.90 (18.79)
10000	4	0.5137 (0.0168)	-0.0912 (0.0037)	1.3465 (0.0133)	0.1722 (0.0087)	247.01 (30.48)
10000	5	0.5888 (0.0153)	-0.1311 (0.0046)	1.4933 (0.0174)	0.1906 (0.0104)	396.74 (45.27)

Table 5: Discrete Outcome Monte Carlo Results

Index $R^2 = 0.75$
True Dummy Variable Regression $R^2 = 0.534$
 $E(d|z=0) = 0.25$; $E(d) = 0.37$

N	K	h (St. Dev.)	α (St. Dev.)	β (St. Dev.)	R^2 (St. Dev.)	DF (St. Dev.)
1000	1	0.1662 (0.0429)	-0.0091 (0.0027)	1.0282 (0.0085)	0.5346 (0.0251)	15.95 (8.32)
1000	2	0.2938 (0.0354)	-0.0278 (0.0043)	1.0858 (0.0121)	0.5495 (0.0268)	50.80 (15.45)
1000	3	0.4099 (0.0270)	-0.0540 (0.0056)	1.1646 (0.0153)	0.5630 (0.0258)	95.89 (16.69)
1000	4	0.4975 (0.0276)	-0.0772 (0.0073)	1.2340 (0.0210)	0.6030 (0.0310)	157.80 (25.42)
1000	5	0.5747 (0.0230)	-0.1005 (0.0100)	1.2940 (0.0267)	0.6351 (0.0312)	221.61 (27.10)
5000	1	0.1204 (0.0253)	-0.0044 (0.0014)	1.0140 (0.0044)	0.5354 (0.0124)	24.01 (7.13)
5000	2	0.2286 (0.0145)	-0.0159 (0.0015)	1.0500 (0.0046)	0.5405 (0.0095)	102.77 (12.57)
5000	3	0.3187 (0.0146)	-0.0328 (0.0021)	1.1008 (0.0063)	0.5548 (0.0129)	263.78 (31.39)
5000	4	0.4037 (0.0121)	-0.0543 (0.0024)	1.1631 (0.0067)	0.5724 (0.0132)	473.68 (41.43)
5000	5	0.4745 (0.0121)	-0.0745 (0.0036)	1.2231 (0.0097)	0.6021 (0.0151)	745.94 (64.41)
10000	1	0.1105 (0.0185)	-0.0036 (0.0010)	1.0113 (0.0031)	0.5336 (0.0081)	27.03 (6.85)
10000	2	0.2029 (0.0145)	-0.0124 (0.0013)	1.0389 (0.0042)	0.5389 (0.0084)	143.07 (20.65)
10000	3	0.2883 (0.0108)	-0.0268 (0.0013)	1.0826 (0.0042)	0.5494 (0.0094)	396.93 (36.40)
10000	4	0.3673 (0.0097)	-0.0454 (0.0017)	1.1372 (0.0046)	0.5650 (0.0093)	770.64 (64.60)
10000	5	0.4364 (0.0092)	-0.0649 (0.0019)	1.1941 (0.0058)	0.5923 (0.0108)	1266.76 (89.96)

Table 6: Continuous Outcome Monte Carlo Results

True Regression $R^2 = 0.25$						
N	K	h (St. Dev.)	α (St. Dev.)	β (St. Dev.)	R^2 (St. Dev.)	DF (St. Dev.)
1000	1	0.2801 (0.0553)	0.0005 (0.0036)	1.0885 (0.0244)	0.2575 (0.0270)	8.48 (2.90)
1000	2	0.4386 (0.0397)	0.0002 (0.0106)	1.2159 (0.0285)	0.2676 (0.0261)	22.29 (5.11)
1000	3	0.5457 (0.0356)	0.0034 (0.0141)	1.3509 (0.0338)	0.2954 (0.0247)	44.80 (8.54)
1000	4	0.6427 (0.0409)	-0.0015 (0.0234)	1.4857 (0.0577)	0.3246 (0.0379)	71.62 (16.47)
1000	5	0.7213 (0.0383)	0.0003 (0.0263)	1.6119 (0.0648)	0.3560 (0.0397)	101.42 (21.92)
5000	1	0.2019 (0.0309)	0.0000 (0.0010)	1.0439 (0.0097)	0.2528 (0.0120)	13.63 (4.17)
5000	2	0.3340 (0.0179)	0.0000 (0.0025)	1.1234 (0.0100)	0.2591 (0.0100)	48.37 (5.83)
5000	3	0.4336 (0.0209)	0.0001 (0.0049)	1.2212 (0.0149)	0.2729 (0.0131)	114.33 (15.66)
5000	4	0.5163 (0.0139)	0.0009 (0.0070)	1.3301 (0.0160)	0.2928 (0.0133)	211.85 (21.69)
5000	5	0.5882 (0.0166)	0.0003 (0.0090)	1.4426 (0.0206)	0.3202 (0.0155)	340.76 (36.88)
10000	1	0.1773 (0.0295)	0.0000 (0.0021)	1.0337 (0.0082)	0.2513 (0.0081)	16.55 (5.35)
10000	2	0.2959 (0.0142)	0.0000 (0.0013)	1.0958 (0.0072)	0.2559 (0.0080)	66.78 (7.36)
10000	3	0.3943 (0.0133)	0.0001 (0.0028)	1.1813 (0.0087)	0.2656 (0.0091)	168.45 (17.08)
10000	4	0.4770 (0.0122)	-0.0006 (0.0038)	1.2796 (0.0100)	0.2810 (0.0092)	324.56 (29.89)
10000	5	0.5471 (0.0105)	-0.0003 (0.0055)	1.3825 (0.0109)	0.3058 (0.0097)	540.11 (42.24)

Table 7: Continuous Outcome Monte Carlo Results

True Regression $R^2 = 0.75$						
N	K	h (St. Dev.)	α (St. Dev.)	β (St. Dev.)	R^2 (St. Dev.)	DF (St. Dev.)
1000	1	0.1753 (0.0303)	0.0001 (0.0021)	1.0315 (0.0091)	0.7539 (0.0135)	14.17 (3.15)
1000	2	0.2876 (0.0173)	0.0002 (0.0061)	1.0793 (0.0088)	0.7625 (0.0140)	51.22 (6.14)
1000	3	0.3783 (0.1670)	0.0018 (0.0083)	1.1250 (0.0118)	0.7855 (0.0132)	115.77 (12.27)
1000	4	0.4613 (0.0158)	-0.0003 (0.0116)	1.1617 (0.0146)	0.8084 (0.0156)	192.36 (17.24)
1000	5	0.5268 (0.0168)	0.0006 (0.0130)	1.1794 (0.0162)	0.8351 (0.0160)	282.41 (25.29)
5000	1	0.1282 (0.0186)	0.0003 (0.0006)	1.0169 (0.0042)	0.7512 (0.0068)	21.93 (4.73)
5000	2	0.2230 (0.0132)	-0.0005 (0.0016)	1.0501 (0.0060)	0.7559 (0.0061)	107.72 (10.43)
5000	3	0.3048 (0.0083)	0.0003 (0.0030)	1.0899 (0.0045)	0.7675 (0.0065)	293.21 (19.95)
5000	4	0.3758 (0.0066)	0.0005 (0.0040)	1.1265 (0.0048)	0.7859 (0.0061)	584.14 (30.76)
5000	5	0.4387 (0.0068)	-0.0004 (0.0048)	1.1546 (0.0062)	0.8090 (0.0069)	959.14 (44.31)
10000	1	0.1155 (0.0126)	0.0000 (0.0002)	1.0136 (0.0027)	0.7499 (0.0043)	25.26 (3.48)
10000	2	0.2003 (0.0076)	0.0001 (0.0009)	1.0406 (0.0028)	0.7542 (0.0047)	145.37 (10.16)
10000	3	0.2793 (0.0072)	0.0000 (0.0018)	1.0770 (0.0038)	0.7625 (0.0040)	427.76 (26.18)
10000	4	0.3458 (0.0059)	0.0001 (0.0025)	1.1123 (0.0040)	0.7778 (0.0047)	923.65 (47.10)
10000	5	0.4082 (0.0054)	0.0002 (0.0028)	1.1421 (0.0042)	0.7995 (0.0052)	1575.25 (66.29)

Table 8: Selected Results for Probit and OLS

Index $R^2 = 0.25$					
True Dummy Variable Regression $R^2 = 0.148$					
$E(d z=0) = 0.25$					
	N	K	α (St. Dev.)	β (St. Dev.)	R^2 (St. Dev.)
Probit					
	1000	1	0.0003 (0.0047)	0.9988 (0.0155)	0.1483 (0.0205)
	1000	5	0.0001 (0.0047)	1.0003 (0.0158)	0.1545 (0.0220)
	10000	1	0.0000 (0.0018)	1.0001 (0.0060)	0.1482 (0.0075)
	10000	5	0.0003 (0.0017)	0.9991 (0.0056)	0.1475 (0.0066)
OLS					
	1000	1	0.0000 (0, by def.)	1.0000 (0, by def.)	0.2467 (0.0212)
	1000	5	0.0000 (0, by def.)	1.0000 (0, by def.)	0.2518 (0.0247)
	10000	1	0.0000 (0, by def.)	1.0000 (0, by def.)	0.2493 (0.0070)
	10000	5	0.0000 (0, by def.)	1.0000 (0, by def.)	0.2502 (0.0067)

Table 9: Effect of Varying Bandwidth from that Chosen by Cross-Validation

20 replications with “Fixed” Explanatory Variables

Sample Size 10,000; Five Independent Regressors; Index $R^2 = 0.25$

Bandwidth as a Proportion of the Cross-Validated Bandwidth	Slope on True Probability when the Non-Parametric Prediction (using Fraction of Bandwidth) Regressed on Truth	R^2 from Regressions of the Observed Outcomes on the Non-Parametric Predictions at this Bandwidth	Effective Degrees of Freedom (DF) for Predicting the Observed Outcome at this Bandwidth
0.5	0.970 (0.028)	0.576 (0.026)	3824 (225)
0.6	0.937 (0.029)	0.431 (0.024)	2427 (176)
0.7	0.897 (0.029)	0.331 (0.020)	1535 (128)
0.8	0.855 (0.030)	0.264 (0.017)	976 (91)
0.9	0.812 (0.030)	0.220 (0.014)	625 (64)
1.0	0.768 (0.030)	0.191 (0.012)	403 (45)
1.1	0.727 (0.029)	0.167 (0.011)	263 (31)
1.2	0.684 (0.029)	0.155 (0.010)	173 (22)
1.3	0.645 (0.028)	0.143 (0.009)	115 (15)
1.4	0.607 (0.028)	0.134 (0.008)	78 (11)
1.5	0.571 (0.027)	0.126 (0.008)	54 (8)
1.6	0.528 (0.026)	0.119 (0.008)	38 (5)
1.7	0.506 (0.025)	0.113 (0.007)	27 (4)
1.8	0.476 (0.024)	0.108 (0.007)	19 (3)
1.9	0.448 (0.024)	0.103 (0.007)	14 (2)
2.0	0.422 (0.023)	0.098 (0.007)	10 (2)

Table 10: Likelihood Function Values for Various Measures of the Predicted Probabilities for Discrete Outcomes

20 replications with “Fixed” Explanatory Variables

Sample Size 10,000; Five Independent Regressors; Index $R^2 = 0.25$

Form of Predicted Probability	Average Log Likelihood in 20 Replications (standard deviation)
1) True Probabilities (constant across replications)	-5204.62 (49.22)
2) Probit Predicted Probabilities (probabilities vary across replications)	-5201.71 (39.97)
3) Average Probit Probabilities (Averaged across the 20 replications, as in Figure 1(a).; constant across replications)	-5204.50 (39.99)
4) Moving Average of Average Probit Probabilities (As in Figure 1(b); constant across replications)	-5204.59 (40.22)
5) Average Kernel Regression Predictions (Averaged, so constant across the 20 replications for each observation, as in Figure 1(a))	-5240.64 (35.71)
6) Moving Average of Average Kernel Regression Predictions (As in Figure 1(b); constant across replications)	-5253.15 (35.29)
7) Average “Cross-Validated” Kernel Regression Predictions (As in Equation 3, and Figure 1(a); constant across replications)	-5263.91 (35.28)
8) Moving Average of Average “Cross-Validated” Kernel Regression Predictions (As in Figure 1(b), constant across replications)	-5261.42 (35.01)
9) Kernel Regression Predictions (As in Equation 1; Probabilities vary across replications.)	-4977.38 (55.71)
10) “Cross-Validated” Kernel Regression Predictions (As in Equation 3; Probabilities vary across replications.)	-5290.72 (38.70)

REFERENCES

- Ahn, H. and Powell, J. 1993, "Semiparametric Estimation of Censored Models with a Nonparametric Selection Mechanism," *Journal of Econometrics*, 58(1), 3-29.
- Blundell, R. and Duncan, A. 1998, "Kernel Regression in Empirical Microeconomics," *Journal of Human Resources*, 21(1), 62-87.
- Estrella, A. 1998, "A New Measure of Fit for Equations with Dichotomous Dependent Variables," *Journal of Business and Economic Statistics*, 16(2), 198-205.
- Fan, J. 1993, "Local Linear Regression Smoothers and Their Minimax Efficiency," *Annals of Statistics*, 21, pp.196-216.
- Freeman, R. and Medoff, J., 1984. *What Do Unions Do?* New York: Basic Books.
- Hastie, T.J. and Tibrishani, R.J., 1990. *Generalized Additive Models*, London: Chapman and Hall Ltd.
- Ichimura, H. 1993, "Semiparametric Least-Squares (SLS) and Weighted SLS Estimation of Single-Index Models," *Journal of Econometrics*, 58: (1-2), pp. 71-120, July.
- Ichimura, H., and Lee, L.-F, 1991, "Semiparametric Least Squares Estimation of Multiple Index Models: Single Equation Estimation," in W. Barnett, J. Powell, and G. Tauchen, eds., *Nonparametric and Semiparametric Methods in Econometrics and Statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics*. International Symposia in Economic Theory and Econometrics Series, Cambridge; New York and Melbourne: Cambridge University Press, pp. 3-49.
- Lewis, H.G, 1986. *Union Relative Wage Effects: A Survey*, Chicago: University of Chicago Press.
- Mroz, T., 1987, "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions," *Econometrica*, 55(4), 765-799.
- Mroz, T., 1999, "Discrete Factor Approximations in Simultaneous Equation Models: Estimating the Impact of a Dummy Endogenous Variable on a Continuous Outcome," *Journal of Econometrics* (forthcoming).
- Mroz, T and Savage, T., 1997, "The Union Wage Gap in the Presence of Endogenous Union Membership," *UNC Working Paper Series*, December 1997.
- Newey W., Powell J., and Vella F., 1999, "Nonparametric estimation of triangular simultaneous equations models," *Econometrica*, 67: (3), pp. 565-603.

- Park, B. and Marron, J., 1990, "Comparison of Data-Driven Bandwidth Selectors," *Journal of the American Statistical Association*, 85(409), 66-73.
- Robinson, C., 1989, "The Joint Determination of Union Status and Union Wage Effects: Some Tests of Alternative Methods," *Journal of Political Economy*, 97, 639-667.
- Silverman, B., 1986, *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall Ltd.
- Yatchew, A., 1998, "Nonparametric Regression Techniques in Economics," *Journal of Economic Literature*, 36(2), 669-721.

