Journal of Economic Perspectives—Volume 15, Number 4—Fall 2001—Pages 11–28

Nonparametric Density and Regression Estimation

John DiNardo and Justin L. Tobias

Even a cursory look at the empirical literature in most fields of economics reveals that a majority of applications use simple parametric approaches such as ordinary least squares regression or two-stage least squares accompanied by simple descriptive statistics. The use of such methods has persisted despite the development of more general *nonparametric* techniques in the recent (and perhaps not-so-recent) statistics and econometrics literatures.

At least two possible explanations for this come to mind. First, given the challenges—or lack of—provided by economic theories with empirical content, the parametric toolkit is more than sufficient. Where serious first-order problems in nonexperimental inference exist, they are in the inadequacy of the research design and data, not in the limitations of the parametric approach. Second, the predominant use of parametric approaches may reflect the lack of sufficient computational power or the difficulty of computation with off-the-shelf statistical software. Given the recent advances in computing power and software (as well as the development of the necessary theoretical foundation), only the first point remains an open question. The purpose of this article is to make the case that nonparametric techniques need not be limited to use by econometricians.

Our discussion is divided into two parts. In the first part, we focus on "density estimation"—estimation of the entire distribution of a variable or set of variables. In the second part, we discuss nonparametric regression, which concerns estimation of regression functions without the straightjacket of a specific functional form.

■ *John DiNardo is Professor of Economics and Public Policy, University of Michigan, Ann Arbor, Michigan. While writing this paper, he was on leave at the University of California, Berkeley, California. Justin Tobias is Assistant Professor of Economics, University of California at Irvine, Irvine, California. Their e-mail addresses are ⟨jdinardo@umich.edu⟩ and ⟨jtobias@uci.edu⟩, respectively.*

Why Nonparametrics? An Application

Consider the observation that that wage "inequality" among women grew from 1979 to 1989. One frequently employed metric of this inequality is the standard deviation of log wages, which increased 25 percent from 0.41 in 1979 to 0.50 in 1989.[1]

One way to get a handle on the economic importance of this change is to consider what such a change means when the logarithm of wages is distributed normally. We draw the functions under this assumption in the top half of Figure 1. In principle, the two *probability density functions* describe everything we need to know about the wages of women in 1979 and 1989. For example, the area under the curve between two different levels of the wage gives you the probability that a randomly chosen woman will have a wage between the two values. Under the assumption that log wages are distributed normally, the mean and standard deviation—the two *parameters* of the normal distribution—are all we need to describe everything about the log wage distributions.

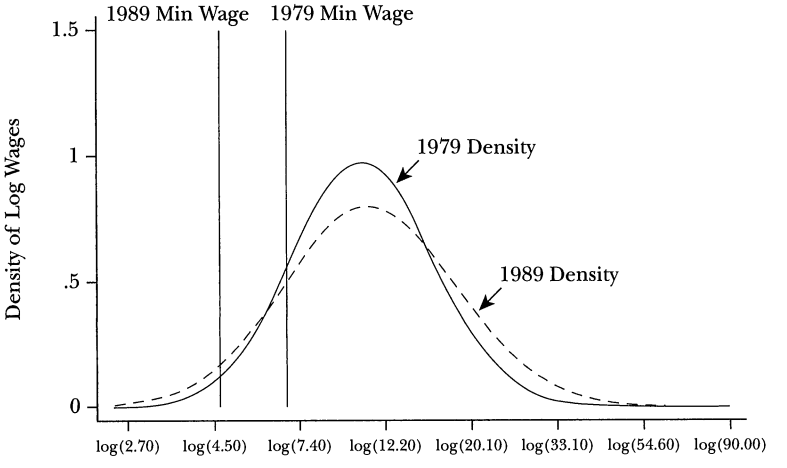
Parametric restrictions are powerful and important. With just two numbers, we can, in principle, answer any question about the probability density function of log wages (or wages) *if our assumption of normality is correct.* However, if the distribution isn't normal (or from a certain class of distributions), then two parameters aren't enough. Indeed, when the key of assumption of the parametric model (in this case, normality) is wrong, our answers can be seriously misleading. Consider this case: given the visual display in the top half of Figure 1, we might speculate on the causes of this increased inequality: international trade? organizational change? computers and skill-biased technical change? Judging from the variances of the parametric densities, it would be hard to say.

One of the likely culprits, however, does manifest itself when we consider a *nonparametric* estimate of the density—that is, we no longer *assume* a specific function like the normal distribution and instead let the data completely tell us what the distribution looks like. This nonparametric estimate tells a specific story. To make the point clearly, we have superimposed a line that is at the value of the legal minimum wage in real terms in 1979 and 1989. As the large "spike" makes evident, in 1979, the modal wage (the most common hourly wage) is the minimum wage! By 1989, however, the minimum wage had fallen so dramatically in real terms that it was no longer binding for most women. Evidently, the approximate truncation induced by the binding minimum wage in 1979 helps to explain the growing "inequality" among women from 1979 to 1989.

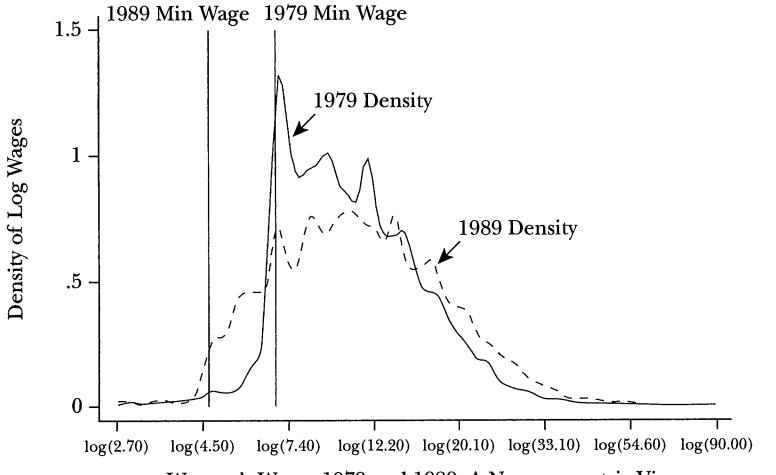
This example highlights a possible problem with *parametric* techniques—when

[1] The source for these numbers is the Monthly Outgoing Rotation Group Files of the Current Population Survey. The curious reader might wonder why we are referring to the natural logarithm of wages. One frequently cited reason is that log of wages is "roughly normally distributed." Indeed the log of wages is "more normal" than the level of wages, but as we will shortly see, not "normal enough"!

Figure 1
The Minimum Wage and Wage Inequality



Women's Wages 1979 and 1989: A Parametric View
(2000 Constant Dollars)



Women's Wages 1979 and 1989: A Nonparametric View
(2000 Constant Dollars)

the assumptions of the parametric approach are violated, the answers can be quite misleading.

Nonparametric Density Estimation

Most readers are already familiar with a nonparametric density estimator of sorts—the venerable histogram, a bar chart that shows the proportion of observations at different values. Indeed, for many applications, it is more than adequate.

A nice historical illustration of the histogram in practice can be found in the work of Louise-Adolphe Bertillon, as discussed in Stigler (1986). In 1863, Bertillon published an analysis of the heights of 9,002 French conscripts from an area called Doubs. Bertillon created a histogram of the data, and Figure 2 is generated from his analysis. To generate the histogram, Bertillon first took his original data (which had been recorded to the nearest centimeter) and converted them to inches. After creating one cell for all observations with heights less than 4′10″, Bertillon grouped the data into 1-inch bins: those with heights from 4′10″ to 4′11″, from 4′11″ to 5′ and so on, up to a bin that contained the relative frequency of conscripts with heights ranging from 5′10″ to 5′11″.

What captured Bertillon's fancy in gazing at this data was the appearance of two humps, or "modes," in his histogram, instead of the single-humped "normal" probability density functions that seemed to characterize so much other data. The two humps reappeared even when he took subsets of the data—those conscripted between 1851 and 1855 or just those men conscripted between 1856 and 1860, for example. Given such a robust finding, it is no surprise that social science was quick to come up with a "theory." The inhabitants of Doubs came from two different "races"—the Celts and Burgundians. The Celtic race tended to be tall, healthy and "true to their word," whereas the feckless Burgundian race was composed of shorter persons neither "so robust, nor temperate, nor obliging." The bimodality arose, it was alleged, because there were really *two* underlying distributions with different means, not one.

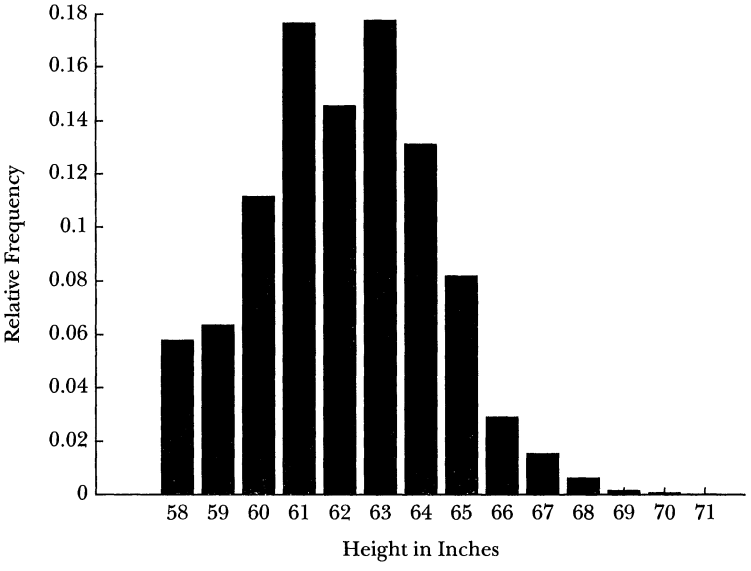
Twenty-six years (!) later, Livi (1896) noted that, among other things, even if the underlying data had one hump, Bertillon's method of constructing the histogram would have produced two humps anyway. Livi noted that the "dip" that appears in Bertillon's histogram at the bin containing the relative frequency of conscripts with heights between 5′1″ and 5′2″ contains only *two* centimeter classes, while the bins to the left and right, from 5′0″ to 5′1″ and from 5′2″ to 5′3″, contained *three* such centimeter classes.

Bertillon's method for estimating the distribution of conscript's heights can be regarded as a simpler form of a popular modern estimator of the probability density function: the kernel density estimator. The key distinction between the kernel and histogram estimators is that a histogram separates the data into distinct *nonoverlapping* "bins." In a histogram, the height of each bar—the proportion of observations falling in a bin—can be viewed as an estimate of the probability density function at the horizontal midpoint of the bin. For instance, the height of the bin that contains observations between 58 inches and 59 inches can be viewed as an estimate of the probability density function at 58.5 inches by observing that it is also the proportion of observations x_i such that $(58.5 - \frac{1}{2} < x_i \leq 58.5 + \frac{1}{2})$.

Using more helpful notation, Bertillon's estimator of the probability density function at x_0 can be written in the following way:

$$pdf(x_0) = \frac{1}{N} \sum_{i=1}^{N} I\left(x_0 - \frac{1}{2} < x_i \leq x_0 + \frac{1}{2}\right),$$

Figure 2
Doubts About the Doubs: Bertillon's Histogram of the Heights of 9,002 Conscripts from Doubs

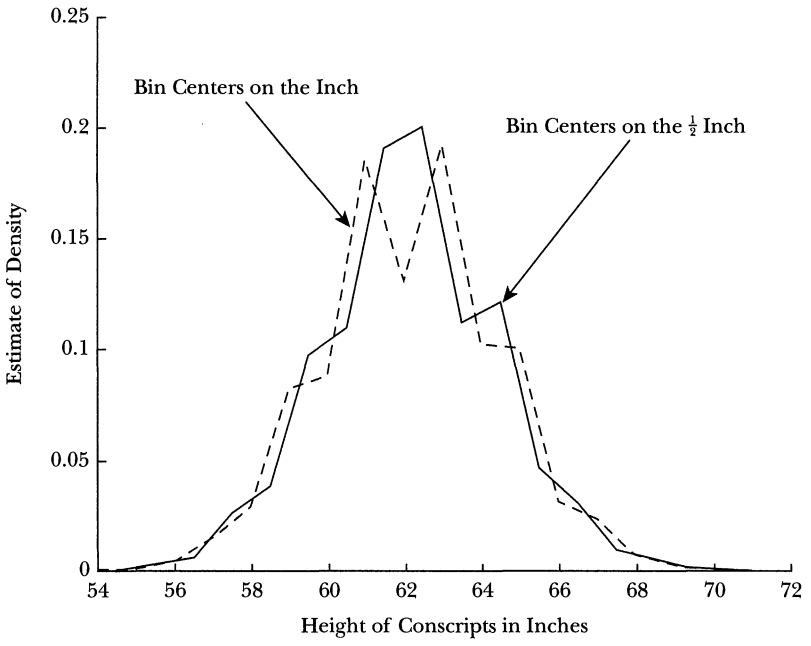


where the function $I(\cdot)$ is equal to one if the statement is true and zero otherwise, and x_0 represents the bin midpoint. Once it is written this way, the problem with the histogram becomes clearer. *The shape of the histogram can potentially be influenced by where you place the bin centers.* Moreover, with a histogram, choosing the width of the bins and the location of the first bin also determines the choice of bin centers. In Bertillon's case, he "chose" to estimate his probability density function at the points 58.5 inches, 59.5 inches, 60.5 inches, . . . Suppose he had estimated his probability density function at 58 inches, 59 inches, 60 inches, . . . instead? Would it matter? Figure 3 (with simulated data generated to approximate Bertillon's original data) illustrates the potential problem. Note that consistent with our viewing the histogram as an estimate of the probability density function for the center of the bins, we have dropped the usual bars and have labeled the axis suitably. As in Bertillon's case, the choice of bin centers clearly matters: with our simulated data, a possibly more precise conversion from centimeters to inches, and bin centers at the $\frac{1}{2}$ inch, the middle hump disappears.

The modern kernel density estimator differs from Bertillon's histogram estimator as described in the previous equation in two key ways. First, in a typical kernel density estimator, the bins are allowed to "overlap." This severs the link between bin size and bin centers that characterizes the histogram and is one reason to prefer the more sophisticated kernel density estimator.

Second, kernel density estimators typically place diminishing weight on data points as they move farther away from our point x_0, while the histogram

Figure 3
The Sensitivity of the Histogram to Alternate Bin Centers



assigns equal weight to all points falling in the bin. On this latter point, note that the indicator function in the earlier equation simply serves to count the number of data points lying in the bin, while potentially one could assign differing weights to points falling in the bin depending on their "closeness" to x_0. Before formally describing the relationship between these two estimators, we first define two terms that are key ingredients in a kernel density estimator: the bandwidth and the kernel.

What is a bandwidth? In a histogram, the bandwidth is the width of the bin divided by two. The bandwidth tells you how far to look to the left and to the right of x_0 when computing the probability density function at x_0. A bandwidth of $\frac{1}{2}$ inch, for example, would typically say that to estimate the probability density function at x_0, you would consider all points that are within $\frac{1}{2}$ inch of x_0. Bertillon's 1-inch bins imply a $\frac{1}{2}$ inch bandwidth. More generally, one can think of a bandwidth as simply a parameter used to determine the size of the "neighborhood" around x_0—large bandwidths define large neighborhoods, and small bandwidths define small ones. In the histogram, points falling outside the neighborhood receive zero weight, while those falling in the neighborhood receive constant weight.

What is a kernel? It is merely a smoothing or weight-assigning function. Bertillon's kernel (the kernel used in any histogram) is today called a rectangular kernel—so-called because it treats all points in a bin the same. In modern

notation, we would write the general kernel probability density estimator function as follows:

$$pdf(x_0) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right),$$

where the function $K(\cdot)$ is called the kernel, and the symbol h denotes the bandwidth. Bertillon's histogram estimator at x_0 can be seen as a particular (rectangular) kernel density estimator where

$$K(x) = \begin{cases} \frac{1}{2} & \text{if } |x| < 1 \\ 0 & \text{otherwise.} \end{cases}$$

A little manipulation shows that this is virtually identical to our previous expression for Bertillon's estimator when h (the bandwidth or bin size) is $\frac{1}{2}$.

In fact, this particular choice of K has been termed "the naïve estimator" (Silverman, 1986, p. 12), and we can improve on it a bit. The first thing to notice is that the rectangular kernel merely assigns the same weight—$\frac{1}{2}$—to all points in a bin no matter how far they are away from the center of the bin. A more intuitive approach, however, is generally to let the weight decline the further the observation is from x_0. Indeed, the most commonly used kernel functions have this property.

Some of these common kernel functions (where we have included the rectangular kernel used in the histogram for reference) are displayed in Table 1. A typical kernel density estimator proceeds by using the formula for the general kernel density where the function $K(\cdot)$ is replaced by one of the functions in this table. To illustrate how this works, consider using the normal kernel with a bandwidth of $\frac{1}{2}$ inch instead of the rectangular kernel that Bertillon used. Instead of the Bertillon estimator, then, we would have

$$pdf(x_0) = \frac{1}{N(\frac{1}{2})} \sum_{i=1}^{N} \frac{1}{\sqrt{2\pi}} \exp\left(-0.5 \cdot \left(\frac{x_i - x_0}{\frac{1}{2}}\right)^2\right).$$

If we were estimating the probability density function of conscripts at 57 inches in height, this function—quite sensibly—assigns the most weight to individuals whose height is exactly 57 inches. To see this, note that the value of this kernel is maximized when $x_i = x_0$. Because the "support" of this kernel is the entire real line, we use all the data to estimate the probability density function at $x_0 = 57$. Note, however, that as a practical matter, the weight we assign far away observations with a normal kernel is quite small. With a bandwidth of $\frac{1}{2}$, an observation exactly at 57 inches would get a weight almost 3,000 times as great as an observation at 55 inches. Different kernels merely change the relative weights. Using the same

Table 1
The Rectangular Kernel and Some Popular Alternatives

Kernel	$K(x)$	*Support*
Rectangular	$\frac{1}{2}$	$\lvert x\rvert \leq 1$
Epanechnikov	$\frac{3}{4\sqrt{5}}\left(1 - \frac{1}{5}x^2\right)$	$\lvert x\rvert \leq \sqrt{5}$
Biweight	$\frac{15}{16}(1 - x^2)^2$	$\lvert x\rvert \leq 1$
Triangular	$1 - \lvert x\rvert$	$\lvert x\rvert \leq 1$
Normal or "Gaussian"	$\frac{1}{\sqrt{2\pi}} e^{-.5x^2}$	$-\infty < x < \infty$

bandwidth, and estimating the probability density function at 57 inches, an observation at slightly less than 55 inches—that is, more than two inches away from the bin center—would get zero weight with a biweight or triangular kernel, but get a small weight with the Epanechnikov kernel in Table 1.

Also note that the particular kernel employed imposes nothing on the shape of the probability density function we estimate. For example, using a normal kernel doesn't restrict the resulting probability density function estimate to looking normal. One property of all these kernels is that the resulting density estimates will typically have a less "jagged" appearance than a histogram. Moreover, since kernel density estimates are often "inputs" to other econometric procedures that require computing derivatives, nonrectangular kernels are clearly useful.

As the above suggests, the choice of kernel appears to be relatively unimportant in practice; that is, choosing any of the kernel functions listed in Table 1 will typically produce a similar-looking estimate of the probability density function.

More important is the choice of bandwidth. Choice of bandwidth essentially involves a tradeoff between bias (misreporting the shape) and variance (lack of precision) of the estimates. Intuitively, the larger the bandwidth, the "smoother" the resulting estimates (lower variance), but we may have oversmoothed the true density and thus obtained a biased estimate of that density. Note that this is a problem for histograms as well; for example, if Bertillon had chosen very wide 5-inch bins, he would also have somewhat mischaracterized the data, though he probably wouldn't have mischaracterized the Burgundians!

Despite a huge literature, however, bandwidth choice remains largely an art. In practice, if a visual display of the data is all that is needed, the eyeball method—start with a rule-of-thumb value and then assess the sensitivity of the resulting estimate to somewhat smaller or somewhat larger bandwidths until the data do not appear either oversmoothed or undersmoothed—produces a perfectly adequate bandwidth. One can, of course, choose the bandwidth "optimally" using some criterion

for evaluating the tradeoff between bias and variance. Silverman (1986) has suggested a frequently used rule-of-thumb bandwidth—$\tilde{h}_n = 0.9(\min\{\hat{\sigma}, IQR/1.34\})N^{-1/5}$, where IQR is the interquartile range (the difference between the 75th and 25th percentile) and $\hat{\sigma}$ is the sample standard deviation of x—which is often a good starting point whether or not the underlying data are normally distributed. Like all desirable bandwidth selection procedures, this bandwidth gets smaller as N—the number of observations—increases, but does not go to zero "too fast."

The kernel density estimator generalizes easily to two (or more) dimensions. One of the earliest uses of bivariate kernel density estimation was Sir Francis Galton, a contemporary of Bertillon. Interestingly, both were interested in providing methods for the police to help identify criminals; Galton's fingerprint classification system eventually came to dominate the once-popular system devised by Bertillon, which comprised a large set of bodily measurements. Galton's (1886) focus was the relationship between the height of adult children and their "midparent" (the average of the mother's height times 1.08 and the father's height), and he collected data on 928 offspring of 205 parents. Galton observed that if a parent grew to be taller than average, the adult offspring's height would tend to be closer to the mean, and hence smaller than the parent. The same was true in reverse. Galton coined the word "regression" to explain this phenomenon. The term "regression" then evolved to describe the ordinary least squares technique that was frequently used to analyze this relationship. An abridged version of Galton's tabular display of the data appears as Table 2.[2] This data could readily be displayed as a three-dimensional histograph, a sort of checkerboard pattern, with squares rising up from each midparent height–child height combination to represent how many times that outcome occurred.

After explaining how he produced the cross-tabulation of the heights of midparents and their offspring, Galton (1886) goes on to say

> I found it hard at first to catch the full significance of the entries in the table, which had curious relations that were very interesting to investigate. They came out distinctly when I *smoothed* the entries in the table, by writing at each intersection of a horizontal column with a vertical one, the sum of the entries in the four adjacent squares and using these to work upon. I then noticed that lines drawn through entries of the same value formed a series of concentric and similar ellipses.

To illustrate Galton's calculation, consider the point (67 inches, 63.7 inches)

[2] The data in this table are from Galton (1886), "Regression Towards Mediocrity in Hereditary Stature." In what follows, however, we use Galton's actual data, which (although they yield similar results) are difficult to reconcile with the table in the article. Our thanks to the University of College London Rare Manuscripts Collection for their permission to use the data. Thanks to Jo Blanden, London School of Economics, for her help in tracking down the data, as well as to Steve Machin for his assistance in making this possible.

Table 2
Galton's 1886 Data

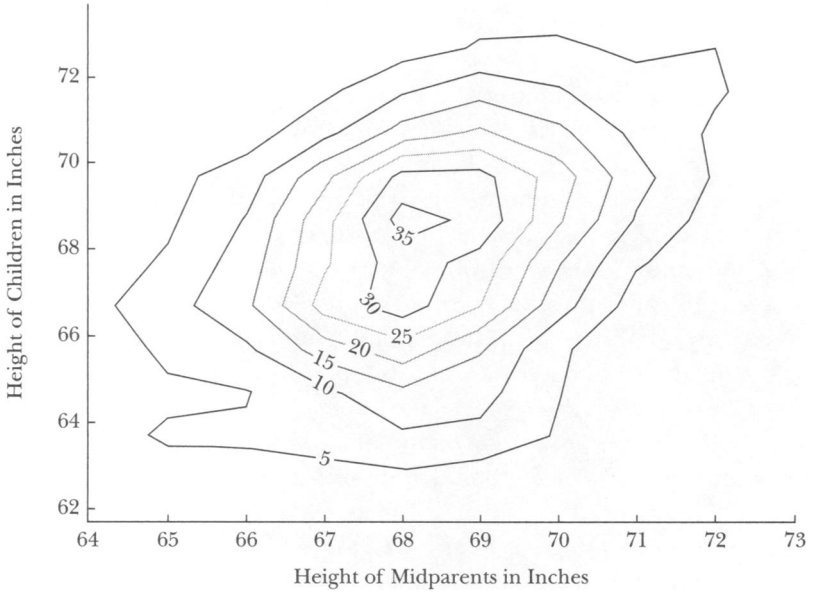
Height of Adult Children (inches)	*Height of Midparent (inches)*										
	Below	*64.5*	*65.5*	*66.5*	*67.5*	*68.5*	*69.5*	*70.5*	*71.5*	*72.5*	*Above*
Below	1	1	1	0	0	1	0	1	0	0	0
62.2	0	1	0	3	3	0	0	0	0	0	0
63.2	2	4	9	3	5	7	1	1	0	0	0
64.2	4	4	5	5	14	11	16	0	0	0	0
65.2	1	1	7	2	15	16	4	1	1	0	0
66.2	2	5	11	17	36	25	17	1	3	0	0
67.2	2	5	11	17	38	31	27	3	4	0	0
68.2	1	0	7	14	28	34	20	12	3	1	0
69.2	1	2	7	13	38	48	33	18	5	2	0
70.2	0	0	5	4	19	21	25	14	10	1	0
71.2	0	0	2	0	11	18	20	7	4	2	0
72.2	0	0	1	0	4	4	11	4	9	7	1
73.2	0	0	0	0	0	3	4	3	2	2	3
Above	0	0	0	0	0	0	5	3	2	4	0

(a 67-inch midparent and a 63.7-inch child), which corresponds to the vertices in Table 2 between midparent heights 66.5 inches and 67.5 inches and child heights 63.2 inches and 64.2 inches. The average of these points is $\frac{1}{4}(3 + 5 + 5 + 14)$ and becomes the smoothed value for (67 inches, 63.7 inches). With the minor additional step of converting his estimates from the number of persons at given heights to the percentage of persons at given heights, Galton's smoother—like the histogram—can be shown to be a simple kernel density estimator using a rectangular kernel! Today, Galton's ellipses are called "contour" plots and are often a useful way to display an inherently three-dimensional figure. Figure 4 displays the contour plots generated by Galton's estimator.

The illustration of the bivariate case is also instructive for illuminating the "curse of dimensionality." Going from one dimension to two greatly exacerbates the difficulties of computation and presentation. Had Galton tried a much smaller bandwidth, many of the resulting cells would have been empty—only by having information on many more individuals would a smaller bandwidth be appropriate.[3] For this reason and others, it is often more useful to step away from considering the entire distribution of all variables and instead to consider the simpler relationship between an explanatory variable (or variables) and the expected value of y (the dependent variable). This is the case of regression analysis, to which we now turn.

[3] For a particularly nice and accessible explication and application of kernel density estimation to real economic problems, including the bivariate case, see Deaton (1997). A nice, earlier discussion at a slightly more technical level can be found in Deaton (1995).

Figure 4
Galton's Original Smoother–Contour Plots: Contours after Galton Smoother



Nonparametric Regression

A standard ordinary least squares regression imposes an assumption of linearity. This framework can be adapted, with the use of logarithms and quadratics, for example, to allow for some curvature, but in all of these cases, the researcher brings a particular underlying functional form to the data. In nonparametric regression, the data is given flexibility to characterize its own shape. This section is devoted to describing a procedure for estimating regression functions nonparametrically, providing applications of the procedure, and illustrating how this flexible approach can provide remarkably accurate estimates even when the underlying regression function is quite nonlinear.

Local Linear Regression

There is a huge literature on flexible methods for estimating regression functions. Here, we focus on *local linear regression* (for example, Fan, 1992; Fan and Gijbels, 1996).[4] Our goal is to describe a method for estimating the function m in the regression equation

[4] A well-known cousin of local linear regression is *local constant regression*, as discussed, for example, in Nadaraya (1964) and Watson (1964). Local linear regression estimates offer several improvements over local constant regression, including improved performance at the boundaries and a reduction in bias. In general, there is a preference for odd-order polynomial fits: see Ruppert and Wand (1994) and Fan and Gijbels (1996) for further discussion.

$$y_i = m(x_i) + \varepsilon_i,$$

where x_i, our independent variable, is assumed to be a scalar for simplicity. Clearly, this generalizes the standard linear regression model, where it is assumed that $m(x_i) = x_i\beta$.

As with most nonparametric regression methods, this approach provides a method for obtaining *pointwise* estimates. That is, the researcher chooses an arbitrary point, say x_0, and then uses the local linear regression method to obtain an estimate of the regression function at that point. This procedure is then repeated for a variety of different choices of the arbitrary point x_0, and, in this way, an estimate of the entire function can be obtained.

The local linear regression approach proceeds by running several "local" regressions. Start by picking an arbitrary point x_0 and a bandwidth that controls the distance (or neighborhood size) around the point x_0. Then carry out a weighted least squares regression, where the data points farther away from the arbitrary starting point x_0 receive less weight than data points closer to x_0. As with density estimation, the weighting of the points is done through a kernel function, and, as a consequence, this procedure is often called *kernel regression.*[5]

Just as in the nonparametric density case, the choice of which kernel to use for a given bandwidth—which amounts to the choice of how to diminish the weight of more distant points—has a relatively small effect on the outcome of a nonparametric regression. However, the choice of bandwidth has a larger effect on results. Moreover, it remains true that bandwidths that are too large produce "oversmoothed" estimates, while bandwidths that are too small generally produce estimates that appear excessively bumpy or that are undersmoothed.

To provide an application of nonparametric regression, consider how two independent variables, cognitive ability and education, affect our dependent variable, log hourly wages. We obtain our wage, ability and education data from the National Longitudinal Survey of Youth (NLSY). Our measure of cognitive ability is a test score constructed from the Armed Services Vocational Aptitude Battery (ASVAB) tests given to the NLSY participants. This ability measure is standardized to have mean zero and unit variance. The education variable used is the highest grade completed by the respondent in 1990, and our dependent variable represents the log of 1990 hourly wages.

Since the education variable is discrete, we proceed by estimating separate nonparametric regressions for five education groups: from nine to twelve years of education; completed twelve years; completed more than twelve years but less than 16 years; completed 16 years; and completed more than 16 years. For the ability variable, we create a grid of 20 evenly spaced points denoting those individuals lying 0.3 standard deviations above the mean of the ability distribution to those lying

[5] Formally, the local linear regression estimator of $m(x_0)$ is given as the α_0^*, which minimizes the weighted least squares problem: $min_{\alpha_0,\alpha_1} \sum_{i=1}^{n} [(y_i - \alpha_0 - \alpha_1(x_i - x_0))^2 K((x_i - x_0)/h_n)]$, with K denoting the kernel and h_n the bandwidth.

1.5 standard deviations above the mean. Five nonparametric regressions are then carried out that describe the relationship between ability and (the log of) hourly wages within each of the five education groups. These five estimates are then pieced together to plot the conditional expectation of log wages over the ability-education space.

Results of this analysis are provided in Figure 5. The first figure is obtained using local linear regression and an optimally chosen smoothing parameter within each education cell.[6] To illustrate sensitivity to bandwidth choice, the right-hand panel of Figure 5 chooses a small bandwidth so that the surface is quite bumpy (undersmoothed) relative to the left-hand panel.

The left-hand panel of Figure 5 reveals some interesting results. First, wages are generally increasing across the education cells at all points in the ability distribution. The noticeable exception is that those in the highest education group (more than 16 years of education) have lower earnings than those in the second-highest education group (exactly 16 years of education) at the left tail of the ability distribution. This result may be attributable to the fact that the regression function is very poorly estimated at the left-tail of the ability distribution for the highest education group, since we observe very few highly educated individuals with low values of the ability index.

Second, for the highest education groups, the log wage surface is generally increasing in measured ability. Note, however, that for the lowest two education groups, the ability-log wage relationship actually decreases as we move into the right tail of the ability distribution. For the highest two education groups, the relationship of ability to log wages is increasing over the full range of ability, and the slope tends to be greatest at the right tail of the ability distribution. This suggests increasing returns to ability for those who are highly educated, while this relationship does not appear to exist for those with less education. Further, this analysis suggests some evidence that returns to schooling are concentrated among the more able, since the gain in log wages from moving to the next education group is generally highest at the right tail of the ability distribution.[7]

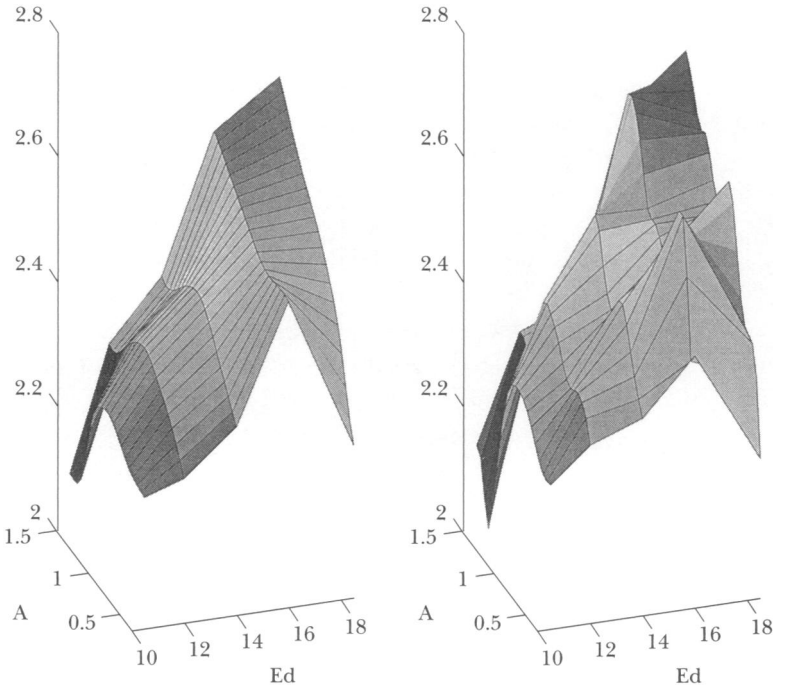
Extension to the Partially Linear Model

While nonparametric regression is flexible in recovering the true shape of the regression curve, it has certain limitations. The most obvious limitation is that it is

[6] An initial optimal bandwidth was obtained using the pilot bandwidth selector of Fan and Gijbels (1996, section 4.2). The results in the figure are slightly oversmoothed relative to the bandwidth obtained from this procedure.

[7] A similar result is reported in Blackburn and Neumark (1993) using a fully parametric approach. Such a result should, however, be interpreted with caution. Heckman and Vytlacil (2001) find that returns to schooling have risen over time for those in the highest ability quartile, and they point out that sorting by ability into higher education creates problems for identifying the return to schooling over the full range of ability. Tobias (2000) uses partially linear regression techniques to examine these issues and obtains pointwise standard errors associated with the estimates of the regression function. As in Heckman and Vytlacil (2001), his results suggest that we are not able to estimate returns to schooling accurately for the most or least able without extrapolating based on particular functional forms.

Figure 5
Nonparametric Estimates of the Effect of Ability (A) and Education (Ed) on Log Wages: Two Different Bandwidths



difficult to include many variables. One problem is computational: each additional variable increases the number of calculations that need to be done in a geometric way, which can become unmanageable. An even more fundamental problem is the *curse of dimensionality*—the rate at which the nonparametric estimator "collapses" around the true value of the regression function depends in a dramatic way on the number of independent variables included in the problem.

To combat these problems, a potentially useful approach in many situations is to remain nonparametric about a certain key variable of interest, but take a parametric stance about the other variables. Specifically, we might specify a model of the form

$$y_i = z_i\beta + m(x_i) + \varepsilon_i,$$

where z contains all of the other variables we would like to control for. This model is called the *partially linear* or *semilinear* regression model. Estimators of this model are often called semiparametric, since they provide both parametric estimates (of β) and nonparametric estimates (of m).

A procedure for estimating this model can be given as follows:

1) Sort the data by ascending values of the x variable (the variable you wish to treat in nonparametric fashion). Calculate first differences of all the sorted data.

2) Compute an ordinary least squares estimate of the β parameters using the differences of the z and y variables. That is, run a regression of the differenced y's on the differenced z's.

3) Now adjust the original dependent variable y by "purging" it of the effects of the z variables that enter in a linear fashion. This can be done by taking each value of y and subtracting from it the sum of all the coefficients from the previous regression in step 2 multiplied by the values of z for the given observation (that is, of the form $y - z\hat{\beta}$).

4) Finally, perform a local linear regression using the new "purged" dependent variable from step 3 and the independent variable x to obtain an estimate of m.

This procedure is based on the simplest form of differencing, and it is important to note that Yatchew (1997) suggests the use of higher order (that is, longer) differencing and optimal differencing weights to increase the efficiency of this estimator. He suggests replacing step 2 with a regression of a higher-order optimally differenced y variable on a similarly higher-order optimally differenced z variable.[8]

To illustrate how the partially linear model is estimated, and to show further the power of the nonparametric approach, consider the following experiment. We generate 300 data points from a regression model that includes both parametric and nonparametric components.[9] The true nonparametric function and the scatterplot of points generated from it are presented in the upper left-hand corner of Figure 6, labeled 6A.[10]

From Figure 6A, we immediately see that attempts to capture this relationship with a pure linear regression would be hopeless. However, the second panel, Figure 6B, goes even further and shows that the function m is not well estimated even after including a quadratic and a quartic polynomial in x.

The third panel, Figure 6C, presents two local linear regression estimates using different choices of the bandwidth. When we choose a bandwidth equal to 10, the result is oversmoothed, and when we set the bandwidth to 0.06, the result is undersmoothed. However, the undersmoothed estimate still appears to do a reasonable job at reproducing the overall shape of true regression function.

The final panel, Figure 6D, uses an optimal choice of smoothing parameter.

[8] Such a procedure does not significantly affect the computational complexity of the problem. See Speckman (1988), Robinson (1988) and Green and Silverman (1994) for alternate approaches.

[9] Specifically, we generate 300 observations from the model $y = 2z_1 + z_2 + m(x) + \varepsilon$, where $m(x) = 0.3exp[-4(x+1)^2] + 0.7exp[-16(x-1)^2]$ and $\varepsilon \sim N(0, 0.01I_n)$. We generate x uniformly on the interval $[-2, 2]$, $z_{1i} \sim N(0.5x_i, 1)$, and z_2 independently from a standardized student-t distribution with four degrees of freedom. Both the coefficients on z_1 and z_2 and the function m are estimated in this example.

[10] To focus on estimation of the nonparametric portion, we present the scatterplot of the $y_i - 2z_{1i} - z_{2i}$ values against x_i.

Figure 6
Partially Linear and Ordinary Least Squares Estimates of Regression Function $m(x)$ Using Generated Data

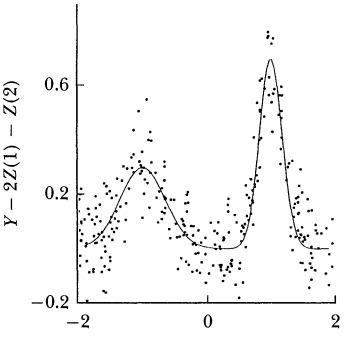


Fig. 6A: True Regression Curve and Data

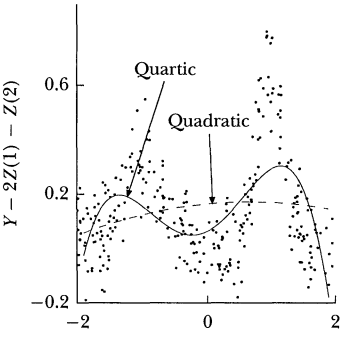


Fig. 6B: Quadratic and Quartic OLS Estimates

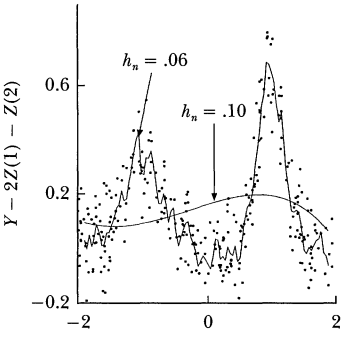


Fig. 6C: Local Linear Estimates

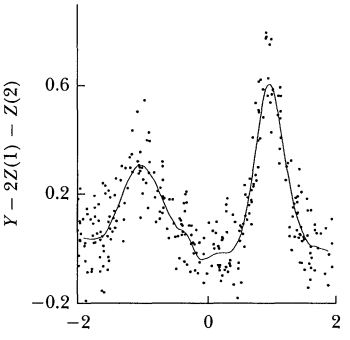


Fig. 6D: Local Linear Estimate − Optimal h_n

To obtain this figure, the dependent variable was first "purged" of the linear variables through a process of higher-order differencing, and a new "adjusted" dependent variable was created. Finally, a local linear regression was performed. Clearly, this estimated function captures the shape of the true regression curve in the first graph remarkably closely, and certainly it does a better job than an ordinary least squares regression even after including a variety of powers of the x variable.

Conclusion

The reader may be curious about drawbacks of these methods for the applied empirical researcher. In our estimation, the biggest limitations for the applied researcher at the moment are two: the lack of a commonly accepted method to choose an appropriate bandwidth and the lack of a simple way to compute reliable standard errors.

The root cause of the first problem is that typical expressions for an "optimal" bandwidth involve unknown properties of the function we are trying to estimate. Since the true shape is never known in practice (why use these techniques if you already know the true shape of the function!), econometricians have devised numerous bandwidth selector techniques. We have discussed some of these selector techniques here, including Silverman's rule-of-thumb bandwidth selector for density estimation, as well as pilot bandwidth selection for regression problems (for example, Fan and Gijbels, 1996, section 4.2). A variety of other bandwidth selection rules are available.

In this paper, we did not discuss the second limitation, and the calculation of pointwise standard errors associated with nonparametric density and regression estimates remains a source of uncertainty for the empirical practitioner. Several techniques are often used, including bootstrapping (Härdle, 1990, section 4.2; Simonoff, 1996; Horowitz, 2001, section 4.2), as well as other, more "direct" methods (in the regression context, see, for example, Fan and Gijbels, 1996, sections 4.3–4.6).

The good news is that the procedures described here are pretty easy to implement—even if one isn't using the estimates as part of a more complicated statistical procedure. Further, these techniques often prove to be very useful in visualizing and describing the data, and thus they can help to avoid drawing some potentially foolish inferences.

For more on nonparametric density estimation, see Silverman (1986). One application of nonparametric density estimation can be found in DiNardo, Fortin and Lemieux (1996). For more on nonparametric regression, see the survey articles by Blundell and Duncan (1998) and Yatchew (1998) or the texts by Härdle (1990), Fan and Gijbels (1996) and Simonoff (1996). Applications of such methods have appeared in Hausman and Newey (1995), Subramanian and Deaton (1996), Chevalier and Ellison (1997), Heckman et al. (1998) and Yatchew (2000), among others.

■ *We would like to thank Steve Machin and the Rare Manuscripts Collection of the University College London Library, and most especially Jo Blandon for helping us to obtain Sir Francis Galton's original Record of Family Faculties data. We would also like to thank Sarah Senesky for her helpful comments and suggestions. We would also like to thank the editors for comments that led to a radical revision of the scope of the paper.*

References

Bertillon, Louise Adolphe. 1876. "Moyenne," in *Dictionnaire Encyclopédique des Sciences Médicales.* Volume 10, Second Edition. Paris: Masson & Asselin, pp. 296–324.

Blackburn, McKinley and David Neumark. 1993. "Omitted-Ability Bias and the Increase in the Return to Schooling." *Journal of Labor Economics.* July, 11:3, pp. 521–44.

Blundell, Richard and Alan Duncan. 1998. "Kernel Regression in Empirical Microeconomics." *Journal of Human Resources.* Winter, 33:1, pp. 62–87.

Chevalier, Judith and Glenn Ellison. 1997. "Risk Taking by Mutual Funds as a Response to Incentives." *Journal of Political Economy.* December, 105:6, pp. 1167–200.

Deaton, Angus. 1995. "Data and Econometric Tools for Development Analysis," in *Handbook of Development Economics, Volume 3A.* Hollis Chenery and T.N. Srinivasan, eds. New York: North-Holland, pp. 1785–882.

Deaton, Angus. 1997. *The Analysis of Household Surveys: A Microeconomic Approach to Development Policy.* Baltimore: Johns Hopkins University Press.

DiNardo, John, Nicole Fortin and Thomas Lemieux. 1996. "Labor Market Institutions and the Distribution of Wages, 1973–1993: A Semi-Parametric Approach." *Econometrica.* September, 64:5, pp. 1001–045.

Fan, Jianqing. 1992. "Design-Adaptive Nonparametric Regression." *Journal of the American Statistical Association.* December, 87:420, pp. 998–1004.

Fan, Jianqing and Irene Gijbels. 1996. *Local Polynomial Modeling and its Applications.* London: Chapman and Hall.

Galton, Sir Francis. 1886. "Regression Towards Mediocrity in Hereditary Stature." *Journal of the Anthropological Institute of Great Britain and Ireland.* 15, pp. 246–63.

Green, P. J. and B. W. Silverman. 1994. *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach.* London: Chapman and Hall.

Härdle, Wolfgang. 1990. *Applied Nonparametric Regression.* Cambridge, Mass.: Cambridge University Press.

Hausman, Jerry and Whitney Newey. 1995. "Nonparametric Estimation of Exact Consumer Surplus and Deadweight Loss." *Econometrica.* November, 63:6, pp. 1445–476.

Heckman, James and Edward Vytlacil. 2001. "Identifying the Role of Cognitive Ability in Explaining the Level of and Change in the Return to Schooling." *Review of Economics and Statistics.* February, 83:1, pp. 1–12.

Heckman, James, H. Ichimura, J. Smith and P. Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica.* September, 66:5, pp. 1017–098.

Horowitz, Joel. 2001. "The Bootstrap," in *Handbook of Econometrics, Volume 5.* E.E. Leamer and J.J. Heckman, eds. Amsterdam: Elsevier Science Publishers, forthcoming.

Livi, Ridolfo. 1896. *Antropometria Militare.* Rome: Presso il Giornale Medico del Regio Esercito.

Nadaraya, E. A. 1964. "On Estimating Regression." *Theory of Probability and its Applications.* 10, pp. 186–90.

Robinson, Peter. 1988. "Root-N Consistent Semiparametric Regression." *Econometrica.* July, 56:4, pp. 931–54.

Ruppert, D. and M. P. Wand. 1994. "Multivariate Locally Weighted Least Squares Regression." *Annals of Statistics.* 22:3, pp. 1346–370.

Silverman, B. W. 1986. *Density Estimation for Statistics and Data Analysis.* London: Chapman and Hall.

Simonoff, Jeffrey S. 1996. *Smoothing Methods in Statistics.* New York: Springer-Verlag.

Speckman, Paul L. 1988. "Kernel Smoothing in Partial Linear Models." *Journal of the Royal Statistical Society, Series B.* 50:3, pp. 413–36.

Stigler, Stephen M. 1986. *The History of Statistics: The Measurement of Uncertainty Before 1900.* Cambridge, Mass.: Belknap Press of Harvard University Press.

Subramanian, Shankar and Angus Deaton. 1996. "The Demand for Food and Calories." *Journal of Political Economy.* February, 104:1, pp. 133–62.

Tobias, Justin L. 2000. "Are Returns to Schooling Concentrated Among the Most Able? A Semiparametric Analysis of the Ability-Earnings Relationships." University of California, Irvine, Department of Economics Working Paper.

Watson, Geoffrey S. 1964. "Smooth Regression Analysis." *Sankhya, Series A.* 26, pp. 359–72.

Yatchew, Adonis. 1997. "An Elementary Estimator of the Partial Linear Model." *Economics Letters.* December, 57:2, pp. 135–43.

Yatchew, Adonis. 1998. "Nonparametric Regression Techniques in Economics." *Journal of Economic Literature.* June, 36:2, pp. 669–721.

Yatchew, Adonis. 2000. "Scale Economies in Electricity Distribution: A Semiparametric Analysis." *Journal of Applied Econometrics.* March/April, 15:2, pp. 187–210.

[Footnotes]

[2] **Regression Towards Mediocrity in Hereditary Stature.**
Francis Galton
The Journal of the Anthropological Institute of Great Britain and Ireland, Vol. 15. (1886), pp. 246-263.
Stable URL:
http://links.jstor.org/sici?sici=0959-5295%281886%2915%3C246%3ARTMIHS%3E2.0.CO%3B2-T

[4] **Multivariate Locally Weighted Least Squares Regression**
D. Ruppert; M. P. Wand
The Annals of Statistics, Vol. 22, No. 3. (Sep., 1994), pp. 1346-1370.
Stable URL:
http://links.jstor.org/sici?sici=0090-5364%28199409%2922%3A3%3C1346%3AMLWLSR%3E2.0.CO%3B2-U

[7] **Omitted-Ability Bias and the Increase in the Return to Schooling**
McKinley L. Blackburn; David Neumark
Journal of Labor Economics, Vol. 11, No. 3. (Jul., 1993), pp. 521-544.
Stable URL:
http://links.jstor.org/sici?sici=0734-306X%28199307%2911%3A3%3C521%3AOBATII%3E2.0.CO%3B2-S

[7] **Identifying the Role of Cognitive Ability in Explaining the Level of and Change in the Return of Schooling**

James Heckman; Edward Vytlacil

The Review of Economics and Statistics, Vol. 83, No. 1. (Feb., 2001), pp. 1-12.

Stable URL:

http://links.jstor.org/sici?sici=0034-6535%28200102%2983%3A1%3C1%3AITROCA%3E2.0.CO%3B2-O

[7] **Identifying the Role of Cognitive Ability in Explaining the Level of and Change in the Return of Schooling**

James Heckman; Edward Vytlacil

The Review of Economics and Statistics, Vol. 83, No. 1. (Feb., 2001), pp. 1-12.

Stable URL:

http://links.jstor.org/sici?sici=0034-6535%28200102%2983%3A1%3C1%3AITROCA%3E2.0.CO%3B2-O

[8] **Root-N-Consistent Semiparametric Regression**

P. M. Robinson

Econometrica, Vol. 56, No. 4. (Jul., 1988), pp. 931-954.

Stable URL:

http://links.jstor.org/sici?sici=0012-9682%28198807%2956%3A4%3C931%3ARSR%3E2.0.CO%3B2-3

References

Omitted-Ability Bias and the Increase in the Return to Schooling

McKinley L. Blackburn; David Neumark

Journal of Labor Economics, Vol. 11, No. 3. (Jul., 1993), pp. 521-544.

Stable URL:

http://links.jstor.org/sici?sici=0734-306X%28199307%2911%3A3%3C521%3AOBATII%3E2.0.CO%3B2-S

Kernel Regression in Empirical Microeconomics

Richard Blundell; Alan Duncan

The Journal of Human Resources, Vol. 33, No. 1. (Winter, 1998), pp. 62-87.

Stable URL:

http://links.jstor.org/sici?sici=0022-166X%28199824%2933%3A1%3C62%3AKRIEM%3E2.0.CO%3B2-8

LINKED CITATIONS

Risk Taking by Mutual Funds as a Response to Incentives
Judith Chevalier; Glenn Ellison
The Journal of Political Economy, Vol. 105, No. 6. (Dec., 1997), pp. 1167-1200.
Stable URL:
http://links.jstor.org/sici?sici=0022-3808%28199712%29105%3A6%3C1167%3ARTBMFA%3E2.0.CO%3B2-X

Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach
John DiNardo; Nicole M. Fortin; Thomas Lemieux
Econometrica, Vol. 64, No. 5. (Sep., 1996), pp. 1001-1044.
Stable URL:
http://links.jstor.org/sici?sici=0012-9682%28199609%2964%3A5%3C1001%3ALMIATD%3E2.0.CO%3B2-K

Design-adaptive Nonparametric Regression
Jianqing Fan
Journal of the American Statistical Association, Vol. 87, No. 420. (Dec., 1992), pp. 998-1004.
Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28199212%2987%3A420%3C998%3ADNR%3E2.0.CO%3B2-0

Regression Towards Mediocrity in Hereditary Stature.
Francis Galton
The Journal of the Anthropological Institute of Great Britain and Ireland, Vol. 15. (1886), pp. 246-263.
Stable URL:
http://links.jstor.org/sici?sici=0959-5295%281886%2915%3C246%3ARTMIHS%3E2.0.CO%3B2-T

Nonparametric Estimation of Exact Consumers Surplus and Deadweight Loss
Jerry A. Hausman; Whitney K. Newey
Econometrica, Vol. 63, No. 6. (Nov., 1995), pp. 1445-1476.
Stable URL:
http://links.jstor.org/sici?sici=0012-9682%28199511%2963%3A6%3C1445%3ANEOECS%3E2.0.CO%3B2-J

Identifying the Role of Cognitive Ability in Explaining the Level of and Change in the Return of Schooling
James Heckman; Edward Vytlacil
The Review of Economics and Statistics, Vol. 83, No. 1. (Feb., 2001), pp. 1-12.
Stable URL:
http://links.jstor.org/sici?sici=0034-6535%28200102%2983%3A1%3C1%3AITROCA%3E2.0.CO%3B2-O

LINKED CITATIONS

Characterizing Selection Bias Using Experimental Data
James Heckman; Hidehiko Ichimura; Jeffrey Smith; Petra Todd
Econometrica, Vol. 66, No. 5. (Sep., 1998), pp. 1017-1098.
Stable URL:
http://links.jstor.org/sici?sici=0012-9682%28199809%2966%3A5%3C1017%3ACSBUED%3E2.0.CO%3B2-U

Root-N-Consistent Semiparametric Regression
P. M. Robinson
Econometrica, Vol. 56, No. 4. (Jul., 1988), pp. 931-954.
Stable URL:
http://links.jstor.org/sici?sici=0012-9682%28198807%2956%3A4%3C931%3ARSR%3E2.0.CO%3B2-3

Multivariate Locally Weighted Least Squares Regression
D. Ruppert; M. P. Wand
The Annals of Statistics, Vol. 22, No. 3. (Sep., 1994), pp. 1346-1370.
Stable URL:
http://links.jstor.org/sici?sici=0090-5364%28199409%2922%3A3%3C1346%3AMLWLSR%3E2.0.CO%3B2-U

The Demand for Food and Calories
Shankar Subramanian; Angus Deaton
The Journal of Political Economy, Vol. 104, No. 1. (Feb., 1996), pp. 133-162.
Stable URL:
http://links.jstor.org/sici?sici=0022-3808%28199602%29104%3A1%3C133%3ATDFFAC%3E2.0.CO%3B2-S

Nonparametric Regression Techniques in Economics
Adonis Yatchew
Journal of Economic Literature, Vol. 36, No. 2. (Jun., 1998), pp. 669-721.
Stable URL:
http://links.jstor.org/sici?sici=0022-0515%28199806%2936%3A2%3C669%3ANRTIE%3E2.0.CO%3B2-J

Scale Economies in Electricity Distribution: A Semiparametric Analysis
A. Yatchew
Journal of Applied Econometrics, Vol. 15, No. 2. (Mar. - Apr., 2000), pp. 187-210.
Stable URL:
http://links.jstor.org/sici?sici=0883-7252%28200003%2F04%2915%3A2%3C187%3ASEIEDA%3E2.0.CO%3B2-X