CHAPTER 7. ROBUST METHODS IN ECONOMETRICS

1. THE PARAMETERS OF ECONOMETRICS

Econometrics deals with complex multivariate relationships and employs non-experimental or "field" data that are influenced by many factors. Occasionally econometricians have data from *designed experiments* in which treatments are randomized, and/or other factors are held constant, to assure that there can be no confounding of the measured effects of treatments. Almost as good are "*natural experiments*", also called "*quasi-experiments*", in field data where a factor of direct interest (or an instrument correlated with a factor of interest) has clearly operated in a manner that is independent of confounding effects. The scientific value of such quasi-experiments is high, and econometricians should actively seek designed or natural experiments that can illuminate economic issues. That said, there remain important problems in economic theory and policy for which experimental data are not available within the time frame in which answers are needed. It is imperative that econometricians deal with these problems using the best tools available, rather than reverting to an orthodoxy that they are too "messy" for econometric treatment.

Econometricians must make educated guesses about the structure of the data generation processes in non-experimental data. The studies that result rely on these structural assumptions can be misleading if the assumptions are not realistic. This has important implications for the conduct of econometric analysis. First, it is desirable to have large data sets in which the "signal" contained in systematic relationships is strong relative to the "statistical noise". Second, it is important to "proof" econometric models, testing the plausibility of the specification both internally and against other data and other studies, and avoiding complex or highly parametric formulations whose plausibility is difficult to check. Fourth, it is desirable to use statistical methods that are "robust" in the sense that they do not force conclusions that are inconsistent with the data, or rely too heavily on small parts of the data.

Most of classical econometric analysis, from linear regression models to maximum likelihood estimation of non-linear models, lays out the assumptions under which the procedures will produce good statistical results, and simply assumes that these postulates can be checked and will be checked by users. To some extent, the development of diagnostic and specification tests provides the capacity to make these checks, and good econometric studies use these tests. However, some basic assumptions are difficult to check, and they are too often accepted in econometric studies without serious examination. Fortunately, in many economic applications, particularly using linear models, the analysis is more robust than the assumptions, and sensibly interpreted will provide useful results even if some assumptions fail. Further, there are often relatively simple estimation alternatives that provide some protection against failures, such as use of instrumental variables or heteroskedasticity-consistent standard errors. New developments in econometrics expand the menu of procedures that provide protection against failures of classical assumptions. This chapter introduces three areas in which "robust" methods are available: the use of nonparametric and semiparametric methods, the use of simulation methods and "indirect inference", and the use of bootstrap methods.

Econometrics first developed from classical parametric statistics, with attention focused on linear systems. This was the only practical alternative in an era when computation was difficult and data limited. Linear parametric models remain the most useful tool of the applied econometrician. However, the assumption of known parametric functional forms and distributions interposes an untidy veil between econometric analysis and the propositions of economic theory, which are mostly abstract without specific dimensional or functional restrictions. Buoyed by good data and computers, contemporary econometricians have begun to attack problems which are not *a priori* parametric. One major line of attack is to use general nonparametric estimation methods to avoid distributional assumptions. The second, closer to classical methods, is to use flexible forms to approximate unknown functions, and specification tests to search for parsimonious representations. The added dimension in a modern rendition of the second approach is explicit recognition of the statistical consequences of adding terms and parameters as sample sizes grow.

Many problems of econometric inference can be cast into some version of the following setup: There is a random vector $(Y,X) \in \mathbb{R}^k \times \mathbb{R}^m$ such that X has a (unknown) density $g(x)$ and almost surely Y has a (unknown) conditional density $f(y|x)$. There is a known transformation $t(y,x)$ from $\mathbb{R}^k \times \mathbb{R}^m$ into the real line $\mathbb{R}$, and the conditional expectation of this transformation, $\theta(x) = \mathbf{E}(t(Y,x)|X=x)$, is the target of the econometric investigation. Examples of transformations of interest are (1) $t(y,x) \equiv y$, in which case $\theta(x) = \mathbf{E}(Y|X=x)$ is the conditional expectation of Y given x, or the *regression function* of Y on x; (2) $t(y,x) = yy'$, in which case $\theta(x) = \mathbf{E}(YY'|X=x)$ is the array of second conditional moments, and this function combined with the first example, $\mathbf{E}(YY'|X=x) - \{\mathbf{E}(Y|X=x)\}\{\mathbf{E}(Y|X=x)\}'$ is the conditional variance; and (3) $t(y,x) = \mathbf{1}_A(y)$, the indicator function of the set A, in which case $\theta(x)$ is the conditional probability of the event A, given $X = x$. Examples of economic applications are Y a vector of consumer demands, and x the vector of income and prices; or Y a vector of firm net outputs and x a vector of levels of fixed inputs and prices of variable inputs.

Define the disturbance $\varepsilon = \varepsilon(y,x) \equiv t(y,x) - \theta(x)$. Then the setup above can be summarized as a *generalized regression model*,

$$t(y,x) = \theta(x) + \varepsilon,$$

where $\mathbf{E}(\varepsilon|x) = 0$. Econometric problems fitting this setup can be classified as *fully parametric*; *semiparametric*; or *nonparametric*. The model is fully parametric if the function θ and the distribution of the disturbance ε are both known *a priori* to be in finite-parameter families. The model is nonparametric if both θ and ε have unknown functional forms, except possibly for shape and regularity properties such as concavity or continuous differentiability. The model is semiparametric if it contains a finite parameter vector, typically of primary interest, but parts of θ and/or the distribution of ε are not restricted to finite-parameter families. This is a rather broad definition of semiparametric, which includes for example linear regression under Gauss-Markov conditions where the distribution of the disturbances is not restricted to a parametric family, and only the first two moments are parametric. Some econometricians prefer to reserve the term semiparametric for situations where the problem can be characterized as one with a finite-dimensional parameter vector that is the target of the analysis and an infinite-dimensional

vector of nuisance parameters (which might, for example, determine an unknown function), for it is in this case that non-classical statistical methods are needed.

Where can an econometrician go wrong in setting out to analyze the generalized regression relationship $t(y,x) = \theta(x) + \varepsilon$? First, there is nothing in the formulation of this model *per se* that assures that $\theta(x)$ has any causal or invariance properties that allow it to be used to predict the distribution of values of $t(y,x)$ if the distribution of x shifts. Put another way, the model will by definition be descriptive of the conditional mean in the current population, but not necessarily predictive under policy changes that alter the distribution of x. Because econometricians are often interested in conditional relationships for purposes of prediction or analysis of policy scenarios, this is potentially a severe limitation. The prescription for "robust" causal inference is to use statistical methods and tests that can avoid or detect joint or "wrong-way" causality (e.g., instrumental variables, Granger invariance tests in time series, exogeneity tests); avoid claiming causal inferences where confounding of effects is possible; and avoid predictions that require substantial extrapolation from the data. Second, when $\theta(x)$ is approximated by a parametric family, there will be a specification error if the parametric family fails to contain $\theta(x)$. Specification errors are particularly likely if the parametric family leaves out variables or variable interactions that appear in the true conditional expectation. Third, the only property that is guaranteed for the disturbances ε when $\theta(x)$ is correctly specified is the conditional first moment condition $\mathbf{E}(\varepsilon \mid x) = 0$. There is no guarantee that the conditional distribution of ε given x is independent of x, or for that matter that the variance of ε is homoskedastic. In addition, there is no guarantee that the distribution of ε has thin enough tails so that higher moments exist, or are sufficiently well behaved so that estimates are not unduly (and unstably) influenced by a small number of high influence observations. In these circumstances, statistical methods that assume well-behaved disturbances can be misleading, and better results may be obtained using methods that bound the influence of tail information. At minimum, it is often worth providing estimates of estimator dispersion that are consistent in the presence of various likely problems with the disturbances.

In statistics, there is a fairly clear division between *nonparametric statistics*, which worries about the specification of $\theta(x)$ or about tests of the qualitative relationship between x and *t*, and *robust statistics*, which worries about the properties of ε. In econometrics, both problems appear, usually together, and it is useful to refer to the treatment of both problems in economic applications as *robust econometrics*.

Despite the leading place of fully parametric models in classical statistics, elementary nonparametric and semiparametric methods are used widely without fanfare. Histograms are nonparametric estimators of densities. Contingency tables for data grouped into cells are one approach to estimating a regression function nonparametrically. Linear regression models, or any estimators that rely on a finite list of moment conditions, can be interpreted as semiparametric, since they do not require complete specification of the underlying distribution function.

2. HOW TO CONSTRUCT A HISTOGRAM

One of the simplest examples of a nonparametric problem is that of estimating an unknown univariate unconditional density $g(x)$, given a random sample of observations x_i for $i = 1,...,n$. Assume, by transformation if necessary, that the support of *g* is the unit interval. An elementary

method of approximating g is to form a histogram: First partition the unit interval into K segments of length 1/K, so that segment k is $(c_{k-1},c_k]$ with $c_k = k/K$ for k = 0,...,K. Then estimate g within a segment by the share of the observations falling in this segment, divided by segment length. If you take relatively few segments, then the observation counts in each segment are large, and the variance of the sample share in a segment will be relatively small. On the other hand, if the underlying density is not constant in the segment, then this segment average is a biased estimate of the density at a point. This bias is larger when the segment is longer. Segment length can be varied to balance variance against bias. As sample size rises, the number of segments can be increased so that the contributions of variance and bias remain balanced.

Suppose the density g has the following smoothness property:

$$|g(x') - g(x)| \leq L|x' - x|,$$

where L is a positive constant. Then the function is said to satisfy a *Lipschitz condition*. If g is continuously differentiable, then this property will be satisfied. Let n_k be the number of observations from the sample that fall in segment k. Then, the histogram estimator of g at a specified argument x is

$$\hat{g}(x) = Kn_k/n \quad \text{for } x \in (c_{k-1},c_k].$$

Compute the variance and bias of this estimator. First, the probability that an observation falls in segment k is the segment mean of g, $p_k = K\cdot \int_{c_{k-1}}^{c_k} g(x)dx$. Then, n_k has a binomial distribution with probability p_k/K, so that it has mean np_k/K and variance $n(p_k/K)(1 - p_k/K)$. Therefore, for $x_o \in (c_{k-1},c_k]$, $\hat{g}(x_o)$ has mean p_k and variance $(K/n)p_k(1 - p_k/K)$. The bias is $B_{nK}(x) = p_k - g(x)$. The *mean square error* of the estimator equals its variance plus the square of its bias, or

$$MSE(x) = (K/n)p_k(1 - p_k/K) + (p_k - g(x))^2.$$

A criterion for choosing K is to minimize the mean square error. Looking more closely at the bias, note that by the theorem of the mean, there is some argument z_k in the segment $(c_{k-1},c_k]$ such that $p_k/K = \int_{c_{k-1}}^{c_k} g(x)dx = g(z_k) \int_{c_{k-1}}^{c_k} dx = g(z_k)/K$. Then, using the Lipschitz property of g,

$$|p_k - g(x)| = |g(z_k) - g(x)| \leq L|z_k - x| \leq L/K,$$

Then, the MSE is bounded by

$$MSE(x) \leq (K/n)p_k(1 - p_k/K) + L^2/K^2.$$

Approximate the term $p_k(1 - p_k/K)$ in this expression by g(x), and then minimize the RHS in K. The (approximate) minimand is $K = (2L^2n/g(x))^{1/3}$, and the value of MSE at this minimand is approximately $(Lg(x)/2n)^{2/3}$. Of course, to actually do this calculation, you have a belling-the-cat

problem that you need to know $g(x)$. However, there are some important qualitative features of the solution. First, the optimal K goes up in proportion to the cube root of sample size, and MSE declines proportionately to $n^{-2/3}$. Compare this with the formula for the variance of parametric estimators such as regression slope coefficients, which are proportional to 1/n. Then, the histogram estimator is *consistent* for g, since the mean square error goes to zero. However, the cost of not being able to confine g to a parametric family is that the rate of convergence is lower than in parametric cases. Note that when L is smaller, so that g is less variable with x, K is smaller.

If you are interested in estimating the entire function g, rather than the value of g at a specified point x, then you might take as a criterion the Mean Integrated Square Error (MISE),

$$\text{MISE} = \mathbf{E}\int(\hat{g}(x) - g(x))^2dx = \sum_{k=1}^{K} \int_{c_{k-1}}^{c_k} \mathbf{E}(\hat{g}(x) - p_k + p_k - g(x))^2dx$$

$$= \sum_{k=1}^{K} \mathbf{E}(Kn_k/n - p_k)^2/K + \sum_{k=1}^{K} \int_{c_{k-1}}^{c_k} (p_k - g(x))^2dx$$

$$= \sum_{k=1}^{K} (1/n)p_k(1 - p_k/K) + \sum_{k=1}^{K} \int_{c_{k-1}}^{c_k} (g(z_k) - g(x))^2dx$$

$$\le K/n + \sum_{k=1}^{K} \int_{c_{k-1}}^{c_k} L^2\cdot(z_k - x)^2dx \le K/n + L^2/3K^2.$$

The RHS of this expression is minimized at $K = (2L^2n/3)^{1/3}$, with MISE $\le (3L/2n)^{2/3}$. Both minimizing MSE at a specified x and minimizing MISE imply that the number of histogram cells K grows at the rate $n^{1/3}$. When $g(x) < 3$, the optimal K for the MISE criterion will be smaller than the optimal K for the MSE criterion; this happens because the MISE criterion is concerned with average bias and the MSE criterion is concerned with bias at a point. One practical way to circumvent the belling-the-cat problem is to work out the value of K for a standard distribution; this will often give satisfactory results for a wide range of actual distributions. For example, the triangular density $g(x) = 2x$ on $0 \le x \le 1$ has L = 2 and gives $K = 2(n/3)^{1/3}$. Thus, a sample of size n = 81 implies K = 6, while a sample of size n = 3000 gives K = 20.

3. KERNEL ESTIMATION OF A MULTIVARIATE DENSITY

One drawback of the histogram estimator is that it is estimating a continuous density by a step function, and the constancy of this estimate within a cell and the steps between cells contribute to bias. There would seem to be an advantage to using an estimator that mimics the smoothness that you know (believe?) is in the true density. This section describes the commonly used *kernel method* for estimating a multivariate density.

Suppose one is interested in estimating an unknown density $g(x)$ for $x = (x_1,...,x_m)$ in the domain $[0,1]^m$. Suppose that g is not known to be in a parametric family, but is known to be strictly

positive on the interior of $[0,1]^m$ and is known to have the following smoothness property: g is continuously differentiable up to order p (where $p \geq 0$), and the order p derivatives satisfy a Lipschitz condition. Some notation is needed to make this precise. Let $\mathbf{r} = (r_1,...,r_m)$ denote a vector of non-negative integers, and $|\mathbf{r}| = \sum r_j$. Let

$$g^{\mathbf{r}}(x) = \frac{\partial^{|\mathbf{r}|} g(x)}{\partial x_1^{r_1} \cdot \cdot \cdot \partial x_m^{r_m}}$$

denote the mixed partial derivative of g of order $|\mathbf{r}|$ with respect to the arguments in $\mathbf{r}$. The assumption is that $g^{\mathbf{r}}(x)$ exists and is continuous for all $\mathbf{r}$ satisfying $|\mathbf{r}| \leq p$, and that there exists a constant L such that $|g^{\mathbf{r}}(x) - g^{\mathbf{r}}(y)| \leq L|x - y|$ for any $\mathbf{r}$ satisfying $|\mathbf{r}| = p$. In applications, the most common cases considered are $p = 0$, where one is assuming g continuous and not too variable (e.g., Lipschitz), and $p = 2$, where one is assuming g twice continuously differentiable. Define $z^{\mathbf{r}} = z_1^{r_1}...z_m^{r_m}$. A function g that satisfies the smoothness condition above has a Taylor's expansion (in h) that satisfies

$$g(x - hz) = \sum_{q=0}^{p} \frac{(-h)^q}{q!} \sum_{|r|=q} g^{\mathbf{r}}(x) \cdot z^{\mathbf{r}} + \lambda \cdot \frac{h^{p+1}}{p!} \sum_{|r|=p} |g^{\mathbf{r}}(x) \cdot z^{\mathbf{r}}| \cdot L|z|$$

for some scalar $\lambda \in (-1,1)$.

Exercise 1. Verify that for $m = 1$, these smoothness conditions reduce to the requirement that g be p-times continuously differentiable, with $d^p g(x)/dx^p$ satisfying a Lipschitz condition, so the Taylor's expansion is a textbook expansion in derivatives up to order p.

Exercise 2. Show that in the case $p = 0$, the expansion reduces to $g(x - hz) = g(x) + \lambda h \cdot L|z|$.

Suppose you have a random sample x_i for $i = 1,...,n$ drawn from the density $g(x)$. In applications, it is almost always desirable to first do a linear transformation of the data so that the components of x are orthogonal in the sample, with variances that are the same for each component. Hereafter, assume that the x's you are working with have this property. Suppose that you estimate g using a kernel estimator,

$$\hat{g}(\mathbf{x}) = \frac{1}{nh^m} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) .$$

The function $K(\mathbf{z})$ is the *kernel*, and the scalar h is the *bandwidth*. The kernel K is a function on $(-\infty,+\infty)^m$ with the properties that $\int K(z)dz = 1$, and for some integer s with $0 \leq s \leq p$, $\int z^{\mathbf{r}} \cdot K(z)dz = 0$ for $|\mathbf{r}| \leq s$ and $\int z^{\mathbf{r}} \cdot K(z)dz = k_{\mathbf{r}}$ for $|\mathbf{r}| = s+1$, where the $k_{\mathbf{r}}$ are constants that are finite and not all zero. In words, K is a "density-like" function which integrates to one, but which is not necessarily always non-negative. All the moments of this function up through order s vanish, and moments of order s + 1 exist and some do not vanish. This is called a *kernel of order* s. In applications, you will encounter mostly first-order kernels satisfying $\int z_i K(z)dz = 0$ and $\int z_i^2 K(z)dz > 0$; these are usually

constructed as non-negative densities that are symmetric about zero. Higher-order kernels, for s > 1, will be used to take advantage of problems where *g* is known to be differentiable to higher order than two. Higher order kernels will necessarily sometimes be negative.

An example of a first-order kernel is $K(z) = (2\pi)^{-m/2} \cdot \exp[-z'z/2]$, a Gaussian kernel formed by the product of univariate standard normal densities. Forming products of univariate kernels in this fashion is a convenient way to build up multivariate kernels. Another example of a multivariate kernel is the multivariate Epanechnikov kernel, $K(z) = (\frac{1}{2})c_m \cdot (m+2) \cdot (1 - z'z) \cdot \mathbf{1}(z'z < 1)$, where c_m is the volume of a unit sphere in $\mathbb{R}^m$, which can be calculated recursively using the formulas $c_1 = 2$, $c_2 = \pi$, and $c_n = c_{n-2} \cdot n/(n-1)$ for $n > 2$. An example of a second-order kernel derived from a first-order kernel K is

$$K^*(\mathbf{z}) = [K(\mathbf{z}) - \gamma^3 K(\gamma \mathbf{z})]/(1-\gamma^2),$$

where γ is a scalar in (0,1). (If K is symmetric about zero, then K* is actually a third-order kernel.) Kernels to any order can be built up recursively as linear combinations of lower order kernels.

Mean and Variance of the Kernel Estimator

The mean of the kernel estimator is

$$\mathbf{E}\hat{g}(\mathbf{x}) = \frac{1}{nh^m} \sum_{i=1}^{n} \mathbf{E}\, K\left(\frac{x - x_i}{h}\right) = \frac{1}{h^m} \int K\left(\frac{x - y}{h}\right) \cdot g(y)dy .$$

Using the fact that the observation x_i are independent, the variance of the kernel estimator is

$$\mathbf{V}\hat{g}(x) \equiv \mathbf{E}[\hat{g}(x) - \mathbf{E}\hat{g}(x)]^2 = \frac{1}{n^2 h^{2m}} \sum_{i=1}^{n} \left\{ \boldsymbol{E}\, K\left(\frac{x - x_i}{h}\right)^2 - \left[\boldsymbol{E}\, K\left(\frac{x - x_i}{h}\right)\right]^2 \right\}$$

$$= \frac{1}{nh^{2m}} \left\{ \int K\left(\frac{x - y}{h}\right)^2 g(y)dy - \left[\int K\left(\frac{x - y}{h}\right) \cdot g(y)dy\right]^2 \right\} .$$

Consistency, Bias, and Mean Square Error

Require $h \to 0$ and $n \cdot h^{2m} \to +\infty$. Then, $\mathbf{E}\hat{g}(x) \to g(x)$ and $\mathbf{V}\hat{g}(x) \to 0$, so that $\hat{g}(x)$ converges to $g(x)$ in mean square error, and is hence consistent. Note that for m large, these conditions require that h fall quite slowly as n rises. This is called the *curse of dimensionality*.

Next approximate the bias and variance of the estimator when h is small. Assume that the order of the kernel s is less than or equal to the degree of differentiability p. Introduce the change of variables $y = x - hz$ in the expressions for the mean and variance of $\hat{g}(x)$, and then use the Taylor's expansion for $g(x - hz)$ up to order s, to obtain

$$\mathbf{E}\hat{g}(x) = \frac{1}{h^m} \int K\left(\frac{x - y}{h}\right) \cdot g(y)dy = \int K(z) \cdot g(x - hz)dz$$

$$= g(x) + \sum_{q=0}^{p} \frac{(-h)^q}{q!} g^r(x) \int K(z)\cdot z^r dz + \lambda\cdot \frac{h^{s+1}}{s!} \sum_{|r|=s} g^r(x)\cdot \int K(z)\cdot z^r\cdot L|z|dz$$

$$= g(x) + \lambda'\cdot L\cdot \frac{h^{s+1}}{s!} \sum_{|r|=s} |g^r(x)|\cdot C_r ,$$

where $C_r = \int |K(z)\cdot z^r|\cdot|z|dz$ is a positive constant determined by the kernel, and λ' is a scalar in (-1,1). Then,

$$\text{Bias}(x) = \lambda'\cdot L\cdot \frac{h^{s+1}}{s!} \sum_{|r|=s} |g^r(x)|\cdot C_r .$$

From this formula, one sees that the magnitude of the bias shrinks at the rate h^{s+1}, where s is the order of the kernel, as long as $s \le p$. Thus, when one knows that *g* has a high degree of differentiability, one can use a higher order kernel and control bias more tightly. The reason this works is that when *g* is very smooth, you can in effect estimate and remove bias components that change smoothly with x; e.g., bias terms that are linear in deviations from the target x. However, if one uses a low order kernel, the bias is determined by the order of the kernel, and is not reduced even if the function *g* is very smooth. At the other extreme, the bias is of order h^{p+1} for any kernel of order $s \ge p$, since the Taylor's series cannot be extended beyond the order of differentiability of *g*, so nothing is gained on the bias side by going to a kernel of order $s > p$. For example, if $p = 0$, so that one knows only that *g* is Lipschitz, then one cannot reduce the order of bias by using a symmetric kernel.

Next consider the variance. Making the change of variables $y = x - hz$,

$$\mathbf{V}\hat{g}(x) = \mathbf{E}[\hat{g}(x) - \mathbf{E}\hat{g}(x)]^2 = \frac{1}{nh^m} \int K(z)^2\cdot g(x - hz)dz - \frac{1}{n}\left(\int K(z)\cdot g(x - hz)dz\right)^2$$

$$= \frac{g(x)}{nh^m} \int K(z)^2 dz + \frac{D}{n\cdot h^{m-1}} ,$$

where D is a constant that depends on K and *g*. As $h \to 0$, the first term in the variance will dominate. Then, the mean square error of the estimator $\hat{g}$ at x is bounded by

$$\text{MSE}(x) = \text{Bias}(x)^2 + \mathbf{V}\hat{g}(x) = L^2\cdot \frac{h^{2(s+1)}}{(s!)^2} \left(\sum_{|r|=s} |g^r(x)|\cdot C_r\right)^2 + \frac{g(x)}{nh^m} \int K(z)^2 dz + \text{HOT},$$

where HOT stands for "Higher Order Terms". The *mean integrated square error* (MISE) is then

$$\text{MISE} = \int \text{MSE}(x)dx = L^2\cdot \frac{h^{2(s+1)}}{(s!)^2}\cdot A + \frac{1}{nh^m}\int K(z)^2 dz + \text{HOT},$$

where

$$A = \int \left(\sum_{|r|=s} |g^r(x)| \cdot C_r \right)^2 dx .$$

The optimal bandwidth h minimizes MISE:

$$h_{opt} = \left(\frac{m(s!)^2}{2(s+1)n \cdot A \cdot L^2} \int K(z)^2 dz \right)^{\frac{1}{m+2(s+1)}} .$$

Then, the bandwidth falls with n, at a slower rate the higher the dimension m or the higher the order of the kernel s. Intuitively, this is because when m is high, there are more dimensions where data can "hide", so the sample is less dense and one has to look more widely to find sufficient neighboring points. Also, when the order of the kernel s high, more distant points can be used without adding too much to bias because the function is smooth enough so that leading bias terms can be taken out. Increasing the order of derivatives typically increases A and/or L, and this also shrinks bandwidth. In an applied problem, direct application of the formula for h_{opt} is impractical because it depends on functions of *g* that one does not know.

Substituting the optimal bandwidth in MISE yields

$$MISE(h_{opt}) = n^{\frac{2(s+1)}{m+2(s+1)}} \cdot \left\{ \frac{2(s+1)AL^2}{m(s!)^2} \right\}^{\frac{m}{m+2(s+1)}} \cdot \left\{ \int K(z)^2 dz \right\}^{\frac{2(s+1)}{m+2(s+1)}} \cdot \frac{m+2(s+1)}{2(s+1)} .$$

Note first that MISE will always fall more slowly than 1/n. This is due to the nonparametric nature of the problem, which implies in effect that only local data is available to estimate the density at each point. Chuck Stone has shown that the rate above is not particular to kernel estimation, but is a best rate that can be obtained by any estimation method. Second, the higher the dimension m, the lower the rate at which MISE falls with sample size, the *curse of dimensionality*. If the problem is very smooth, and one exploits this by using a higher-order kernel, one can offset some of the curse of dimensionality. In the limiting case, as s –→ +∞, the rate approaches the limiting 1/n rate. However, other terms in MISE also change when one goes to higher order kernels. In particular, $\int K(z)^2 d\mathbf{z}$ will increase for higher order kernels, and the constant A will typically increase rapidly because higher order derivatives are less smooth than lower order ones.

Least-Squares Cross-Validation

The idea behind cross-validation is to formulate a version of the MISE criterion that can be estimated from the data alone. Then, the bandwidth that minimizes this empirical criterion is close to the optimal bandwidth. The MISE criterion can be written

$$MISE = \mathbf{E} \int [\hat{g}(x) - g(x)]^2 dx = \mathbf{E} \int \hat{g}(x)^2 dx - 2 \cdot \mathbf{E} \int \hat{g}(x) \cdot g(x) dx + \int g(x)^2 dx .$$

The approach is to obtain unbiased estimators of the terms involving $\hat{g}(x)$, and then to choose h iteratively to minimize this estimated criterion. Consider first the term $\mathbf{E} \int \hat{g}(x)^2 dx$. This

expression can be estimated using the kernel estimator $\hat{g}$. To get a convenient computational formula, first define $K^{(2)}(z) = \int K(w - z)\cdot K(w)dw$. This is a convolution that defines a new kernel starting from K, and is an expression that can often be determined analytically. When K is a probability density, $K^{(2)}$ has a simple interpretation: if $\mathbf{W}_1$ and $\mathbf{W}_2$ are independent random vectors with density K, then the density of $\mathbf{Z} = \mathbf{W}_1 - \mathbf{W}_2$ is $K^{(2)}$. For example, if K is a product of univariate standard normal densities, then $K^{(2)}$ is a product of univariate normal densities with mean 0 and variance 2. Using the definition of $K^{(2)}$, and making the transformation of variables $w = (x - x_i)/h$,

$$\int \hat{g}(x)^2 dx = \frac{1}{n^2 h^{2m}} \sum_{i=1}^{n} \int K\left(\frac{x - x_i}{h}\right) \cdot K\left(\frac{x - x_j}{h}\right) \cdot dx$$

$$= \frac{1}{nh^m} \sum_{i=1}^{n} \sum_{j=1}^{n} K^{(2)}\left(\frac{x_j - x_i}{h}\right) .$$

This statistic converges to its expectation as $n \to +\infty$.

Next consider the term $\int \hat{g}(x)\cdot g(x)dx = \frac{1}{n^2 h^{2m}} \sum_{i=1}^{n} \int K\left(\frac{x - x_i}{h}\right) g(x)dx$. Replace the unknown $g(x)$ in the expression $\int K\left(\frac{x - x_i}{h}\right) g(x)dx$ by the empirical density from the sample, excluding $\mathbf{x}_i$; this puts probability 1/(n-1) at each data point x_j for $j \neq i$. This gives an estimator $\frac{1}{nh^m} \cdot \frac{1}{n-1} \sum_{i=1}^{n} \sum_{j \neq i} K\left(\frac{x_j - x_i}{h}\right)$ for $\int \hat{g}(x)\cdot g(x)dx$.

Exercise 3. Show that $\int \hat{g}(x)\cdot g(x)dx$ and the estimator for it given above have the same expectation.

Putting together the estimators for the first two terms in the MISE, one obtains the empirical criterion

$$\mathrm{MISE}'(h) = \frac{1}{n^2 h^m} \sum_{i=1}^{n} \sum_{j=1}^{n} K^{(2)}\left(\frac{x_j - x_i}{h}\right) - \frac{2}{nh^m} \cdot \frac{1}{n-1} \sum_{i=1}^{n} \sum_{j \neq i} K\left(\frac{x_j - x_i}{h}\right)$$

$$= \frac{1}{n^2 h^m} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[K^{(2)}\left(\frac{x_j - x_i}{h}\right) - \frac{2n}{n-1} K\left(\frac{x_j - x_i}{h}\right)\right] + \frac{2K(0)}{(n-1)h^m} .$$

For application, use a nonlinear search algorithm to minimize this expression in h. The minimand h_{lsxv} is the optimal bandwidth estimated by the cross-validation method. An important theoretical

result due to Chuck Stone is that if g is bounded, then $\text{MISE}(h_{opt})/\text{MISE}(h_{lsxv}) \to 1$ as $n \to +\infty$, so that asymptotically one can do as well using the bandwidth obtained by minimizing the empirical criterion $\text{MISE}'(h)$ as one can do using the optimal bandwidth.

4. NONPARAMETRIC REGRESSION

Now consider the general problem of estimating $\theta(x)$ in the regression model $t_i = \theta(x_i) + \varepsilon_i$, where x_i is of dimension m, $t_i = t(y_i,x_i)$ is a known transformation, θ is an unknown function, ε_i is a disturbance satisfying $\mathbf{E}(\varepsilon_i | x_i) = 0$, but otherwise not restricted, and (y_i,x_i) for $i = 1,...,n$ is a random sample. This is the general setup from the introduction. Consider *locally weighted* estimators of the form

$$T_n(x) = \sum_{i=1}^{n} w_{ni}(x;x_1,...,x_n)t(y_i,x_i),$$

where the w_{ni} are scalars that put the most weight on observations with x_i near x. The weights do not have to be non-negative, but their sum has to approach one as $n \to +\infty$. Here are some examples of nonparametric estimation methods that are of this form, and their associated weight functions:

1. *Kernel Estimation*: Suppose K is a *kernel* function from $\mathbb{R}^m$ into $\mathbb{R}$, and h is a *bandwidth*. The function K will be large near zero, and will go to zero at arguments far away from zero; common examples for $m = 1$ are the *uniform* kernel, $K(v) = \mathbf{1}_{[-1,+1]}(v)$; the *normal* kernel $K(v) = \varphi(v)$, where φ is the standard normal density; the *triangular* kernel $K(v) = \text{Max}\{1-|v|,0\}$; and the *Epanechnikov* kernel $K(v) = (3/4)(1-v^2)\mathbf{1}_{[-1,+1]}(v)$, which turns out to have an efficiency property. The local weights are

$$w_{ni}(x;x_1,...,x_n) = \frac{1}{h_n^{\,m}} K\left(\frac{x-x_i}{h_n}\right) \Big/ \sum_{j=1}^{n} \frac{1}{h_n^{\,m}} K\left(\frac{x-x_j}{h_n}\right) ,$$

where the bandwidth h_n shrinks with sample size. The kernel estimator of $\theta(x)$ is

$$T_n(x) = \frac{\frac{1}{nh^{\,m}} \sum_{i=1}^{n} t(y_i,x_i)K\left(\frac{x-x_i}{h_n}\right)}{\frac{1}{nh^{\,m}} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h_n}\right)} .$$

The denominator of this expression can be interpreted as an estimator of $g(x)$, and the numerator as an estimator of $g(x)\mathbf{E}_{y|x}t(y,x) = g(x)\theta(x)$. The kernel function K is typically defined so that $\int K(v)dv$

= 1, and is taken to be symmetric so that $\int vK(v)dv = 0$. If θ is known to be a smooth function, with Lipschitz derivatives of order p, then there turns out to be an advantage (in large enough samples) to using a *higher-order* kernel that satisfies $\int v^j k(v)dv = 0$ for j = 1,...,p.

2. *Nearest Neighbor Estimator*. For the given x, order the observations $(y_{(i)},x_{(i)})$ so that $|x - x_{(1)}| \leq |x - x_{(2)}| \leq ... \leq |x - x_{(n)}|$. To simplify discussion, rule out ties. Define a sequence of scalars $w_{n,(i)}$ that sum to one, and define

$$T_n(x) = \sum_{i=1}^{n} w_{n,(i)} t(y_{(i)},x_{(i)}) .$$

If $w_{n,(i)} = 0$ for i > r, this is termed a *r-nearest neighbor* estimator. Examples of weights are uniform, $w_{n,(i)} = 1/r$ for $i \leq r$ and zero otherwise, and triangular, $w_{n,(i)} = 2(r-i+1)/r(r+1)$. If θ is known to be a smooth function with Lipschitz derivatives of order p, then it is advantageous to run a *local regression*, in which $t(y_i,x_i)$ is regressed on all points of the form $\prod_{h=1}^{m} x_{ih}^{p_h}$ with $\sum_{h=1}^{m} p_h \leq p$, with weights $w_{n,(i)}$, and the fitted value of this regression at x is the estimator of $\theta(x)$. This extension reduces bias by taking into account the fact that a smooth function must vary regularly in its arguments, allowing larger neighborhoods so that variance as well as bias can be reduced.

Uniform nearest neighbor and uniform kernel estimators have the following relationship: If the bandwidth in a uniform kernel estimator is chosen as a function of the data, a *variable kernel* method, so that exactly r observations fall in the interval where the kernel is positive, then this estimator is a uniform nearest neighbor estimator.

3. *Other Nonparametric Methods*. There are several widely used nonparametric estimation methods other than locally weighted estimators. First, the function $\theta(x)$ may be approximated by sums of standard functions, such as polynomials, with the number of terms in the sums growing with sample size. A traditional form of these series approximations is the use of *Fourier* or *Laplace* approximations, or other series of *orthogonal polynomials*. These series are truncated at some point, depending on the sample size, the dimension of the problem, and the smoothness assumed on $\theta(x)$. Once this is done, the problem is effectively parametric, and ordinary regression methods can be used. (Judicious choice of the series so that the terms are orthogonal results in computational simplifications, as you do not have to invert very large matrices.) This approach to nonparametric regression is called, awkwardly, *semi-nonparametric estimation*. The traditional econometric practice of adding variables to regression models as sample sizes grow, and using some criterion based on t-statistics to determine how many variables to keep in, can be interpreted as a version of this approach to estimation. What nonparametric econometrics adds is a mechanism for choosing the number of terms in an "optimal" way, and an analysis that determines the statistical properties of the result.

More recently it has become common to use a functional approximation approach with functions whose determination is more local; popular functional forms are *splines*, *neural nets*, and

wavelets. This approach is called the *method of sieves*. Loosely speaking, splines are piecewise polynomials, neural nets are nested logistic functions, and wavelets are piecewise trigonometric functions. Another approach to nonparametric estimation is *penalized maximum likelihood*, in which the log likelihood of the sample, written in terms of the infinite-dimensional unknown function, is augmented with a penalty function that controls the "roughness" of the solution.

All the nonparametric estimation methods listed above will be consistent, in the sense that the mean square error MSE(x) of $T_n(x)$ at a given point x converges to zero, with asymptotically normal distributions (although not at a root-n rate) under suitable regularity conditions and choices of estimation tuning features such as bandwidth. Further, the conditions on the underlying problem needed to get this result are essentially the same for all the methods. An important result, due to Chuck Stone, is that given sample size, the dimensionality of a problem, and the smoothness that can be assumed for the regression function, there is a maximum rate at which MSE(x) can decline. Any of the estimation methods listed above can achieve this maximum rate. Thus, at least in terms of asymptotic properties, one method is as good as the next. In practical sample sizes, there are no general results favoring one method over another. Kernel methods are usually the easiest to compute at a point, but become computationally burdensome when an estimator is needed for many points. Nearest neighbor estimators require large sorts, which are time-consuming. The method of sieves involves more computational overhead, but has the advantage of being "global" so that once the coefficients of the series expansion have been estimated, it is easy to produce forecasts for different points. The method of sieves is currently the most fashionable approach, particularly using neural net or wavelet forms which have been spectacularly successful in recovering some complex test functions. On the whole, nonparametric methods in finite samples place a considerable burden on the econometrician to decide whether nonlinearities in nonparametric estimators are true features of the data generation process, or are the result of "over-fitting" the data.

Consistency:

As in the case of the histogram estimator of a density, good large sample properties of a locally weighted estimator are obtained by giving sufficient weight to nearby points to control variance, while down-weighting distant points to control bias. As sample size increases, distant observations will be down-weighted more strongly, since there will be enough observations close by to control the variance. The following theorem, adapted from C. Stone (1977), gives sufficient conditions for consistency of a locally weighted estimator.

Theorem 1. Assume (i) g(x) has a convex compact support $\mathbf{B} \subseteq \mathbb{R}^m$; (ii) $\theta(x)$ satisfies a Lipschitz property $|\theta(x') - \theta(x)| \leq L|x' - x|$ for all $x', x \in B$; (iii) the conditional variance of $t(y,x)$ given x, denoted $\Omega(x)$, satisfies $\Omega_o \leq \Omega(x) \leq \Omega_1$, where Ω_o and Ω_1 are finite positive definite matrices; (iv) a random sample i = 1,..,n is observed; and (v) as $n \to +\infty$ the local weights w_{ni} satisfy

$$\text{(a)} \quad E_{\{x_i\}} \sum_{i=1}^{n} w_{ni}(x;x_1,...,x_n)^2 \to 0$$

(b) $E_{\{x_i\}} \sum_{i=1}^{n} w_{ni}(x;x_1,...,x_n) - 1 \to 0$

(c) $E_{\{x_i\}} \sum_{i=1}^{n} |w_{ni}(x;x_1,...,x_n)| \cdot |x - x_i| \to 0$.

Then $T_n(x) - \theta(x)$ converges to zero in mean square.

Proof: The bias of the estimator is

$$B_n(x) = E_{\{x_i\}} \sum_{i=1}^{n} w_{ni}(x;x_1,...,x_n)[\theta(x_i) - \theta(x)] + \theta(x) \left\{ E_{\{x_i\}} \sum_{i=1}^{n} w_{ni}(x;x_1,...,x_n) - 1 \right\} ,$$

so that assumption (v), (b) and (c) imply

$$|B_n(x)| \le L \cdot E_{\{x_i\}} \sum_{i=1}^{n} |w_{ni}(x;x_1,...,x_n)| \cdot |x_i - x| + \theta(x) \left\{ E_{\{x_i\}} \sum_{i=1}^{n} w_{ni}(x;x_1,...,x_n) - 1 \right\} \to 0.$$

The variance of the estimator is, by assumption (v), (a),

$$V_n(x) = E_{\{x_i\}} \sum_{i=1}^{n} w_{ni}(x;x_1,...,x_n)^2 \Omega(x_i) \le \Omega_1 \; E_{\{x_i\}} \sum_{i=1}^{n} w_{ni}(x;x_1,...,x_n)^2.$$

Then, $MSE = V_n(x) + B_n(x)^2 \to 0$, completing the proof. ■

It is useful to work out conditions on nearest neighbor and kernel estimators that satisfy the sufficient conditions in Theorem 1. First, consider a uniform nearest neighbor estimator, with r_n points included in the neighborhood at sample size n. Then, $w_{n(i)} = 1/r_n$ for the points in the neighborhood. The LHS of condition (v), (b) in Theorem 1 equals $1/r_n$, so the condition is satisfied if $r_n \to +\infty$. Next, we show that a sufficient condition for (v), (c) in Theorem 1 is $r_n/n \to 0$. Let $N_t(x)$ denote a neighborhood of x of radius t. For any $\lambda > 0$, define τ_n such that $g(N_{\tau_n}) = (1+\lambda)r_n/n$, and note that $r_n/n \to 0$ and $x \in \mathbf{B}$ implies $\tau_n \to 0$. Let $\mathbf{R}_n$ denote the (random) number of observations in the neighborhood N_{τ_n}; then $\mathbf{E}\mathbf{R}_n = n\, g(N_{\tau_n}) = (1+\lambda)r_n$ and $\mathrm{Var}(\mathbf{R}_n) = n\, g(N_{\tau_n})\, [1 - g(N_{\tau_n})] \le$

$(1+\lambda)r_n$. Let $\mathbf{T}_n$ denote the (random) radius of the neighborhood that contains exactly r_n of the observations x_i. Then

$$P(\mathbf{T}_n > \tau_n) = P(\mathbf{R}_n < r_n) = P(\mathbf{R}_n - \mathbf{E}\mathbf{R}_n < r_n - (1+\lambda)r_n) = P(R_n - \mathbf{E}R_n < -\lambda r_n)$$
$$\leq \mathrm{Var}(\mathbf{R}_n)/\lambda^2 r_n^{\,2} \leq (1+\lambda)/\lambda^2 r_n,$$

with the first inequality obtained by applying Chebyshev's inequality to the sum of the independent random indicators for the events $x_i \in N_{\tau_n}$; these indicators sum to $\mathbf{R}_n$. From this result, and a bound $|x - x'| \leq M$ for $x, x' \in B$ implied by the compactness of B,

$$\mathbf{E}\mathbf{T}_n \leq \tau_n \cdot P(\mathbf{T}_n \leq \tau_n) + M \cdot P(\mathbf{T}_n > \tau_n) \leq \tau_n + M(1+\lambda)/\lambda^2 r_n \to 0.$$

Then,

$$E_{\{x_i\}} \sum_{i=1}^{n} |w_{ni}(x;x_1,\ldots,x_n)| \cdot |x - x_i| \leq \mathbf{E}\mathbf{T}_n \to 0,$$

establishing that (v), (c) in Theorem 1 holds. The kernel estimator of $\theta(x)$ is

$$T_n(x) = \frac{\frac{1}{n \cdot h^m} \sum_{i=1}^{n} t(y_i,x_i) \cdot K\left(\frac{x - x_i}{h}\right)}{\frac{1}{n \cdot h^m} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)} .$$

Note that this estimator is of the generic form $T_n(x) = \sum_{i=1}^{n} w_{in} t(y_i,x_i)$, where the w_i are weights that sum to one. Because the kernel $K\left(\frac{x - x_i}{h}\right)$ is small unless x_i is near x, the weights w_i will be concentrated on points with x_i near x. Then, this estimator corresponds to intuition on how a non-parametric estimator can be constructed. You will recognize the denominator in the formula for $T_n(x)$ is simply a kernel estimator of $g(x)$. The numerator is an estimator of $\int t(y,x) \cdot f(y|x) dy \cdot g(x)$. Then, $T_n(x)$ can be interpreted as an estimator of $\int t(y,x) \cdot f(y|x) dy = [\int t(y,x) \cdot f(y|x) dy \cdot g(x)]/g(x)$.

Now suppose that $\theta(x)$ and $g(x)$ are continuously differentiable to order p, with Lipschitz order p derivatives, and that the kernel is of order $s \leq p$. Also assume that $\sigma^2(x)$ is finite and Lipschitz in x. As in the case of density estimation, require that $h \to 0$ and $nh^m \to +\infty$ as $n \to +\infty$. This will ensure that the numerator of $T_n(x)$ converges in mean square error to $\theta(x) \cdot g(x)$ and that the denominator converges in mean square error to $g(x)$, so that the ratio is a consistent estimator of $\theta(x)$.

Arguments similar to those for density estimation are used to establish further statistical properties of $T_n(x)$. Treat the numerator and the denominator separately. The denominator is the

earlier density estimator, where we found that the bias satisfied $\text{Bias}_{\text{denom}}(x) = C \cdot h^{s+1}$, where C is a constant. Make a Taylor's expansion of the function $q(x - hz) \equiv \theta(x - hz) \cdot g(x - hz)$ to order s:

$$q(\mathbf{x} - h\mathbf{z}) = \sum_{j=0}^{s} \frac{(-h)^j}{j!} \sum_{|r|=j} q^r(x) \cdot z^r + \lambda \cdot \frac{h^{s+1}}{s!} \sum_{|r|=s} |q^r(x) \cdot z^r| \cdot L' |z| .$$

Then, the numerator satisfies

$$\mathbf{E} \frac{1}{nh^m} \sum_{i=1}^{n} t(y_i, x_i) \cdot K\left(\frac{x - x_i}{h}\right) = \int g(x-hz) \cdot \theta(x-hz) \cdot K(z) dz = g(x) \cdot \theta(x) - \lambda'' \cdot A' \cdot h^{s+1},$$

where A' is a constant that depends on the order s derivatives of t, and on the Lipschitz constant L'. Then, $\text{Bias}_{\text{numer}}(x) = \lambda'' \cdot A' \cdot h^{s+1}$.

The variance of the denominator, from the previous analysis, is $\frac{g(x)}{nh^m} \int K(z)^2 dz + \text{HOT}$.

An analogous argument applied to the numerator establishes that its variance is $\frac{\sigma^2(x) \cdot g(x)}{n \cdot h^m} \int K(z)^2 dz + \text{HOT}$. The covariance of the numerator and denominator is zero.

Consider a ratio α_n/β_n of random variables α_n and β_n that have finite second moments, satisfy $\alpha_n \to_p \alpha_0$ and $\beta_n \to_p \beta_0$ as $n \to +\infty$, and have β_n uniformly bounded and bounded away from zero. Then, $\mathbf{E}\alpha_n \to \alpha_0$, $\mathbf{E}\beta_n \to \beta_0$, and the ratio can be rewritten

$$\frac{\alpha_n}{\beta_n} - \frac{\alpha_0}{\beta_0} = \frac{\frac{\alpha_n - \boldsymbol{E}\alpha_n}{\boldsymbol{E}\beta_n} - \frac{\alpha_0}{\beta_0} \cdot \frac{\beta_n - \boldsymbol{E}\beta_n}{\boldsymbol{E}\beta_n} + \frac{\boldsymbol{E}\alpha_n - \alpha_0}{\boldsymbol{E}\beta_n} - \frac{\alpha_0}{\beta_0} \cdot \frac{\boldsymbol{E}\beta_n - \beta_0}{\boldsymbol{E}\beta_n}}{1 + \frac{\beta_n - \boldsymbol{E}\beta_n}{\boldsymbol{E}\beta_n}} .$$

The expectation of the square of this expression is the mean square error of α_n/β_n. For n large, the denominator is almost always very close to one, and is rarely close to zero. The expectation of the square of the numerator can be written

$$\frac{\boldsymbol{V}\alpha_n}{\beta_0^2} + \left(\frac{\alpha_0}{\beta_0}\right)^2 \cdot \frac{\boldsymbol{V}\beta_n}{\beta_0^2} - \frac{2\alpha_0}{\beta_0} \cdot \frac{\boldsymbol{cov}(\alpha_n, \beta_n)}{\beta_0^2} + \left(\frac{bias_\alpha}{\beta_0} - \frac{\alpha_0 \cdot bias_\beta}{{\beta_0}^2}\right)^2$$

Applying this formula to the numerator and denominator of $T_n(x)$, substituting the expressions just derived for variances and biases, the mean square error in $T_n(\mathbf{x})$ is

$$\text{MSE}(x) = \frac{\sigma^2(x)}{n \cdot h^m \cdot g(x)} \int K(z)^2 dz + \frac{\theta(x)^2}{n \cdot h^m \cdot g(x)} \int K(z)^2 dz + h^{2(s+1)} \cdot \frac{C}{g(x)^2} ,$$

where C is a constant depending on order s derivatives, Lipschitz constants, and K. The h_{opt} that minimizes MSE(x), or the integral MISE of MSE(x) over a domain where $g(x)$ is bounded positive,

is proportional to $n^{-1/(m+2(s+1))}$, and the mean square error criterion is proportional to $n^{-2(s+1)/(m+2(s+1))}$, just as in the case of density estimation. Again, the precision of the estimator falls when dimensionality m rises, and high-dimension problems require immense sample sizes to achieve accurate estimators. A high degree of smoothness, exploited using high-order kernels can offset some of the negative impacts of dimensionality, but can never get mean square error to fall at a 1/n rate. As in the case of density estimation, a least squares cross-validation procedure can be used to determine an approximately optimal bandwidth in applications. W. Hardle and O. Linton (1994) give the formulas.

Optimal Rates

The number of observations included in a nearest neighbor estimator, or the bandwidth in a kernel estimator, can vary over considerable ranges and still produce consistent estimators. However, there are typically optimal values for these design parameters that minimize mean square error. These values depend on the properties of the function being estimated, but their qualitative properties are of interest. These notes mentioned earlier the result of Stone that there will be a best rate at which MSE(x) declines, for *any* nonparametric method, and that all the standard methods can achieve this rate. This best rate of decline turns out to be very slow when the dimension m of x is large. This is called the *curse of dimensionality*, and is a consequence of the fact that when dimensionality is high, data are more sparse. (This proposition can be made precise by considering the statistical problem of the expected radius of the largest sphere that can be circumscribed around a data point without encountering any other data points. For a given sample size, this expected radius rises with dimension m at a rate that corresponds to the curse of dimensionality.)

I will give a rough outline of an argument that determines the optimal bandwidth for kernel estimation in the case that $\theta(x)$ is Lipschitz, and after that a rough outline of an argument that determines the optimal number of neighbors for nearest neighbor estimation. These arguments draw heavily from the demonstrations following the proof of Theorem 1, and parallel the arguments for consistent kernel estimation of a multivariate density given earlier.

Kernel Estimation: From the earlier analysis, the variance of the estimator is approximately proportional to $K(0)/g(x)nh^m$, and the bias is approximately proportional to h. Then, the first-order-condition for minimization of variance plus squared bias is $h_n = D/n^{1/(m+2)}$ for a constant D, and the corresponding MSE declines at rate $n^{-2/(2+m)}$. For m = 1, this is the same $n^{-2/3}$ rate that was achieved by the optimal histogram estimator of a Lipschitz density.

Nearest Neighbor Estimation: From the earlier analysis, if there are r observations in the neighborhood, with $r \to +\infty$ and $r/n \to 0$, then the estimator is a (weighted) average of r observations, so that its variance is approximately D_0/r, where D_0 is a constant that does not depend on r. The volume of a sphere of radius t in $\mathbb{R}^m$ is $C_m t^m$, where C_m is a constant depending only on m. Then, for $g(x) > 0$, the radius τ_n of a neighborhood that is expected to contain $(1+\lambda)r$ points satisfies $(1+\lambda)r/n = g(N_{\tau_n}) \approx g(x)C_m\tau_n^m$ and the random radius $\mathbf{T}_n$ of a neighborhood that contains exactly r points satisfies $\mathbf{E}\mathbf{T}_n \leq \tau_n + D_1/r \approx D_2(r/n)^{1/m} + D_1/r$ for some constant D_2. Suppose for the moment that we omit the D_1 term. Then, the first-order-condition for minimizing the sum of variance and squared

bias is $D_0/r_n = (D_2/m)r_n \cdot n^{-2/m}$, which implies that the optimal r_n is proportional to $n^{2/(2+m)}$. Substituting this into the formula for the bias shows that at this rate the D_1 term becomes negligible relative to the D_2 term, justifying its omission. Finally, when r_n is proportional to $n^{2/(2+m)}$, the MSE declines at the rate $n^{-2/(2+m)}$.

The common rate $n^{-2/(2+m)}$ at which MSE declines for the "best" nearest neighbor and kernel estimators of a Lipschitz nonparametric regression is in fact the maximum rate found by Stone for a problem of m dimensions with Lipschitz θ that has no further known smoothness properties. Hence the rates above for the number of neighbors and for bandwidth are also "best". Note that for m even moderately large, the rate of decline of MSE is agonizingly slow. When m = 8 for example, to reduce MSE by a factor of 10, it is necessary to increase sample size by a factor of 100,000. This is the curse of dimensionality in action. The only way to circumvent this problem is to assume (and justify the assumption) that θ is differentiable to high order, and use this in constructing the nonparametric estimator, or to assume that θ depends only on low-dimensional interactions of the variables, e.g., θ is a sum of functions of the variables taken two at a time.

Asymptotic Normality

Returning to the general family of locally weighted estimators, we look for conditions, in addition to those guaranteeing consistency, that are sufficient to establish that the nonparametric estimator is asymptotically normal. The following theorem gives a general result; the added conditions are (iv) and in (vi), strengthened conditions (b) and (c), and new conditions (d)-(f):

Theorem 2. Assume (i) g(x) has a convex compact support $B \subseteq \mathbb{R}^m$; (ii) θ(x) satisfies a Lipschitz property $|\theta(x') - \theta(x)| \le L|x' - x|$ for all $x', x \in B$; (iii) the conditional variance of $t(y,x)$ given x, denoted Ω(x), satisfies $\Omega_o \le \Omega(x) \le \Omega_1$, where Ω_o and Ω_1 are finite positive definite matrices; (iv) $\mathbf{E}_{y|x}|t(y,x) - \theta(x)|^3 \le A|\Omega(x)|^{3/2}$ for some constant A; (v) a random sample i = 1,..,n is observed; and (vi) as $n \to +\infty$ the local weights w_{ni} satisfy

$$\text{(a)} \quad \sum_{i=1}^{n} E_{\{x_i\}} \; w_{ni}^2(x;x_1,...,x_n) \to 0$$

$$\text{(b)} \quad \left(\sum_{i=1}^{n} \boldsymbol{E}_{\{x_i\}} w_{ni}^2(x;x_1,...,x_n)\Omega(x_i) \right)^{-1/2} \left\{ \boldsymbol{E}_{\{x_i\}} \sum_{i=1}^{n} w_{ni}(x;x_1,...,x_n) - 1 \right\} \to 0$$

$$\text{(c)} \quad \left(\sum_{i=1}^{n} \boldsymbol{E}_{\{x_i\}} w_{ni}^2(x;x_1,...,x_n)\Omega(x_i) \right)^{-1/2} E_{\{x_i\}} \sum_{i=1}^{n} |w_{ni}(x;x_1,...,x_n)| \cdot |x - x_i| \to 0$$

(d) $$\frac{\boldsymbol{E}_{\{x_i\}} \sum_{i=1}^{n} |w_{ni}(x;x_1,\ldots,x_n)|^3}{\left\{\boldsymbol{E}_{\{x_i\}} \sum_{i=1}^{n} w_{ni}(x;x_1,\ldots,x_n)^2\right\}^{3/2}} \rightarrow 0$$

(e) $$\frac{\left\{\boldsymbol{E}_{\{x_i\}} \sum_{i=1}^{n} |w_{ni}(x;x_1,\ldots,x_n)|.|x - x_i|\right\}^2}{\boldsymbol{E}_{\{x_i\}} \sum_{i=1}^{n} w_{ni}(x;x_1,\ldots,x_n)^2 |\Omega(x_i)|} \rightarrow 0$$

(f) $$\frac{\left\{\boldsymbol{E}_{\{x_i\}} \sum_{i=1}^{n} w_{ni}(x;x_1,\ldots,x_n) - 1\right\}^2}{\boldsymbol{E}_{\{x_i\}} \sum_{i=1}^{n} w_{ni}(x;x_1,\ldots,x_n)^2 |\Omega(x_i)|} \rightarrow 0$$

Then $\left\{\boldsymbol{E}_{\{x_i} \sum_{i=1}^{n} w_{ni}(x;x_1,\ldots,x_n)^2 \Omega(x_i)\right\}^{-1/2}$ $\{T_n(x) - \theta(x)\}$ converges in distribution to $N(0,\mathbf{I})$.

Proof: We make use of the following central limit theorem, which is a corollary of the Lindeberg-Feller theorem for triangular arrays; see Serfling (1980, 1.9.3, Corollary, p. 32): *For each n, let* ζ_{ni} *for* $i \le n$ *be independent random variables with mean zero, finite variances* σ_{ni}^2*, and for some* $\nu > 2$, $\left(\sum_{i=1}^{n} \boldsymbol{E}|\zeta_{ni}|^{\nu}\right) / \left(\sum_{i=1}^{n} \sigma_{ni}^2\right)^{\nu/2} \rightarrow 0$. *Then,* $\left(\sum_{i=1}^{n} \zeta_{ni}\right) / \left(\sum_{i=1}^{n} \sigma_{ni}^2\right)^{1/2} \rightarrow_d N(0,\mathbf{I})$.

Assume that $T_n(x)$ is a scalar, or else consider a fixed linear combination of components. Define $\zeta_{ni} = w_{ni}[t(y_i,x_i) - \theta(x_i)]$; then for each n, the ζ_{ni} are independent with finite variances $\sigma_{ni}^2 = w_{ni}^2 \Omega(x_i)$. Hypotheses (iv) and (vi), (d) imply

$$\left(\sum_{i=1}^{n} \boldsymbol{E}_{\{x_i\}} |w_{ni}|^3 |t(Y,x_i) - \theta(x_i)|^3\right) / \left(\sum_{i=1}^{n} \sigma_{ni}^2\right)^{3/2}$$

$$\le A \left(\sum_{i=1}^{n} \boldsymbol{E}_{\{x_i\}} |w_{ni}|^3 |\Omega(x_i)|^{3/2}\right) / \left(\sum_{i=1}^{n} \boldsymbol{E}_{\{x_i\}} w_{ni}^2 \Omega(x_i)\right)^{3/2}$$

$$\leq A(|\Omega_1|/|\Omega_o|) \left(\sum_{i=1}^{n} \boldsymbol{E}_{\{x_i\}} |w_{ni}|^3 \right) \Big/ \left(\sum_{i=1}^{n} \boldsymbol{E}_{\{x_i\}} w_{ni}^2 \right)^{3/2} \to 0.$$

Finally, consider the scaled bias term

$$\left(\sum_{i=1}^{n} \boldsymbol{E}_{\{x_i\}} w_{ni}^2 \Omega(x_i) \right)^{-1/2} \quad [\ \boldsymbol{E}_{\{x_i\}}\ T_n(x) - \theta(x)]$$

$$= \left(\sum_{i=1}^{n} \boldsymbol{E}_{\{x_i\}} w_{ni}^2 \Omega(x_i) \right)^{-1/2} \quad \left\{ \sum_{i=1}^{n} \boldsymbol{E}_{\{x_i\}} w_{ni} [\theta(x_i) - \theta(x)] + \theta(x) [\sum_{i=1}^{n} \boldsymbol{E}_{\{x_i\}} w_{ni} - 1] \right\} .$$

This converges to zero by (vi), (e) and (f). Then, the limiting distribution has mean zero. ■

Consider the "best" kernel and nearest neighbor estimators. The assumptions on these estimators made in the discussion of consistency and best rates, along with assumptions (i)-(v) in Theorem (ii), are sufficient to establish (vi), (a)-(d). These in turn are sufficient to establish consistency and asymptotic normality, but possibly with a non-zero mean. A device introduced by Herman Bierens allows one to get this asymptotic mean to zero while preserving the "best" rate. I will explain the trick for a nearest neighbor estimator. Suppose $r_n = Dn^{2/(2+m)}$ and $r_n' = 2^m r_n$ are two cutoff numbers for nearest neighbor estimation, both growing at the "best" rate, where D is some constant. Let $T_n(x)$ and $T_n'(x)$ be the corresponding estimators. Since $r_n' > r_n$, the estimator $T_n'(x)$ will have a larger bias and a smaller variance than $T_n(x)$. Now consider an estimator $T^*(x) = 2T_n(x) - T_n'(x)$. This estimator is also a locally weighted estimator, with weights that are the {2,-1} linear combination of the weights for the two original estimators. It is easy to check that these weights satisfy the same properties in Theorems 1 and 2 as do the original weights, so that $T^*(x)$ is consistent for $\theta(x)$. These combined weights increase at the "best" rate $n^{1/(2+m)}$, so that $T^*(x)$ is again a "best" estimator. Recall from the discussion of optimal rates that except for terms that are negligible in large samples, the bias for a nearest neighbor estimator with $r = Cn^{2/(2+m)}$ points is proportional to $(r/n)^{1/m} = C^{1/m} n^{-1/(2+m)}$. For $T_n(x)$, C = D, while for $T_n'(x)$, $C = 2^m D$. Therefore, except for higher-order terms, the bias in $T^*(x)$ is proportional to $2D^{1/m} n^{-1/(2+m)} - (2^m D)^{1/m} n^{-1/(2+m)} = 0$. Then, there is a "best" nearest neighbor estimator that is asymptotically normal with mean zero. The weights for the estimator $T^*(x)$ can be interpreted as "higher order" weights that remove more bias; note that these weights are sometimes negative. This trick has reduced bias, at the expense of increasing variance, since the variance of $T^*(x)$ is greater than that of $T_n(x)$, while leaving the "best" rate unchanged. A similar device works for kernel estimators, using a higher-order kernel that is a linear combination of two kernels whose bandwidths differ by a multiplicative constant.

Exercise 4. Find the appropriate constants for a second-order kernel that removes asymptotic bias from the estimator so that its asymptotic distribution is centered at zero.

.

5. SEMIPARAMETRIC ANALYSIS

Semiparametric methods provide estimates of finite parameter vectors without requiring that the complete data generation process be assumed in a finite-dimensional family. By avoiding bias from incorrect specification, such estimators gain robustness, although usually at the cost of decreased precision. The most familiar semiparametric method in econometrics is ordinary least squares, which estimates the parameters of a linear regression model without requiring that the distribution of the disturbances be in a finite-parameter family. The recent literature in econometric theory has extended semiparametric methods to a variety of nonlinear models. Four overlapping major areas are models for censored duration data (e.g., employment duration); limited dependent variable (partial observability) models for discrete or censored data (e.g., employment status or employment hours); models for data with (natural or intentional) endogenous sample selection (e.g., wage determination among self-selected workers, or case-control sampling); and models for additive non-parametric effects. The following table summarizes some applications.

Model	**Applications**
Regression and Single Index Models for Censored Duration Data: $Y\|x \cong Y\|x'\beta$	Employment Duration, Innovation Lags, Mobility
Limited Dependent Variable Models (E.g., Discrete response or censored response) $Y^* = x'\beta - \varepsilon$, $\varepsilon\|x \sim F(\cdot)$, *observability transformation* $Y = \Psi(Y^*)$ E.g., Discrete: $Y = \text{sgn}(Y^*)$, Censored: $Y = \text{Min}(Y^c, Y^*)$	Discrete: Employment Status, Brand Choice Censored: Employment Hours, Expenditure Levels
Endogenous Sample Selection $Y = x'\beta - \varepsilon$, $\varepsilon\|x \sim f(\cdot)$, $x \sim g(\cdot)$, Natural: (Y,x) observed iff $Y > 0$ Intentional: (Y,x) sampled iff $Y > 0$ $P(Y,x\|\text{Obs}) = \dfrac{f(Y-x'\beta)g(x)\mathbf{1}(Y>0)}{\int_{z=-\infty}^{+\infty}\int_{y=0}^{+\infty} f(y-z'\beta)g(z)dydz}$ $P(Y\|x,\text{Obs}) = f(Y-x'\beta)/\int_{y=0}^{+\infty} f(y-x'\beta)dy$	Natural: Self-selected Workers, Self-selected Homeowners Intentional: Case-Control Sample Designs
Additive Non-Parametric Effects: $Y = x'\beta + H(z) + \varepsilon$	Robust policy analysis

In most cases, the primary focus of semiparametric analysis is estimation of coefficients of covariates that index the location of the distribution of a dependent variable; then, the unknown distribution is a (infinite-dimensional) nuisance parameter. There are also applications where some functional of the unknown distribution, such as the expectation of the dependent variable conditioned on covariates, is of primary interest. The final objective may be point estimates or confidence

intervals for the objects of interest, or hypothesis tests involving these parameters. Usually, it is important to have measures of precision for the estimates of interest, including convergence rates, asymptotic distributions, and bootstrap or other indicators of finite-sample precision and accuracy of asymptotic approximations.

These notes will not survey the full range of semiparametric models in econometrics, or develop the properties of semiparametric estimator except for illustrative cases. A good survey of the foundations of semiparametric analysis can be found in Powell (1994). These notes will instead survey two areas of application. The first is the analysis of censored employment duration data, perhaps the leading case of applied semiparametric work. The second is the analysis of data on stated willingness-to-pay for natural resources.

Censored Employment Duration

The main focus of the literature on employment duration has been the effect of covariates such as sex, race, age, and education on the hazard of leaving a job. Data on employment duration is typically censored because employment spells start before a panel study is initiated (and the start date may not be recovered accurately using retrospective questions) and/or continue past the end of the panel study, or because of attrition from the panel. In this chapter, we consider only right-censoring before the end of a spell. Parametric analysis of the duration problem has typically used exponential or Weibull survival curves, or the Cox proportional hazards model, which qualifies as one semiparametric formulation.

Horowitz and Newmann (1987) make perhaps the first empirical application of semiparametric censored regression methods to data on employment duration. To provide some context for the economic application, consider the hazards that may lead to termination of a spell of employment. First, termination may be initiated either by the employee (quits), or the employer (layoffs, separations). The quit decision of an employee is presumably influenced by nonpecuniary job features (e.g., safety, variety, and work rules), wage opportunity cost, and worker characteristics such as education, race, and loyalty. The termination decision of the firm is influenced by the expected productivity of the worker, net of wages. The worker's job-specific human capital influences both wage opportunity cost and expected productivity. Wage opportunity cost is also influenced by expected unemployment insurance benefits and duration of unemployment. Macroeconomic and product cycles influence expected productivity. Several aspects of this description are important for modeling employment duration:

1. Quits and separations are competing risks, with overlapping but not identical covariates. Structural estimates of duration must distinguish these two hazards. Data on whether employment spells end in quits would greatly aid identification and estimation of the separate hazards.
2. Important covariates such as the level of macroeconomic activity and job-specific human capital vary in elapsed or chronological time, so a structural model must accommodate time-varying covariates. To do this is fairly easy in discrete time using heterogeneous Markov models, and quite difficult in continuous time.

3. Unobserved variables such as worker loyalty are heterogeneous in the population and are selected by survival. Thus, it is necessary for structural modeling of duration to determine the distribution of these unobservables. The presence of unobserved heterogeneity also selects the subpopulation that start employment spells during the interval of observation. The subpopulation starting employment spells near the beginning of the observation interval will be less loyal on average than all workers. Those whose first observed employment spell start comes near the end of the observation period will be more loyal on average if the panel is long enough.
4. In a structural model of employment duration, the hazard must depend solely on the history of economic variables, and not directly on elapsed time. Thus, models that postulate a reduced-form "baseline" hazard are removing variation that must have a structural source. From the standpoint of structural estimation of the economic determinants of duration, emphasis on the effect of covariates with the baseline hazard treated as a nuisance parameter is misplaced.
5. Economic theory provides neither a tight specification of functional forms or the distributions of unobservables; the assumption that observables enter in a parametric additive combination must be justified as an approximation. Consequently, analyses that assume observables appear in an exact additive combination within unknown transformations or distributions in fact assume too much on the structure of the additive combination, and perhaps too little on the unknown transformations, which may be approximable to comparable accuracy using flexible finite-parameter families.

The duration data generation process can be characterized by a *survival curve* $q(t|x)$ stating the proportion of a population with spells starting at time zero who survive at elapsed time t, given an observed covariate process $x(\cdot)$. If there are unobserved covariates ξ distributed in the initial population with density $\nu(\cdot|x,0)$, and the "structural" survival curve is $q(t|x,\xi)$, then the data generation process satisfies

$$q(t|x) = \int_{-\infty}^{+\infty} q(t|x,\xi)\cdot\nu(\xi|x,0)d\xi. \tag{1}$$

The density of the unobserved covariates, conditioned on survival, is modified over time by selection, satisfying

$$\nu(\xi|x,t) = \nu(\xi|x,0)q(t|x,\xi)/q(t|x). \tag{2}$$

The survival curve can also be described by the *hazard rate*,

$$h(t|x,\xi) = -\nabla_t \mathrm{Ln}(q(t|x,\xi)). \tag{3}$$

The *average hazard rate* in the surviving population is

$$h^*(t|x) = -\nabla_t \mathrm{Ln}(q(t|x)) \tag{4}$$

$$= \left(\int_{-\infty}^{+\infty} h(t|x,\xi) q(t|x,\xi) \nu(\xi|x,0) d\xi \right) \quad / q(t|x) = \quad \int_{-\infty}^{+\infty} \; h(t|x,\xi) \nu(\xi|x,t) d\xi \; .$$

Equation (3) can be inverted to obtain

(5) $$q(t|x,\xi) = \exp \left(-\int_0^t h(s|x,\xi) ds \right) \quad \equiv \exp \left(-\Lambda(t|x,\xi) \right) \; ;$$

with $\Lambda(t|x,\xi)$ termed the *integrated hazard*. The mean duration of completed spells is

(6) $$\mathbf{E}(t|x,\xi) = - \int_0^{\infty} \; t \, \nabla_t \, q(t|x,\xi) dt = \int_0^{\infty} \; q(t|x,\xi) dt,$$

with the second formula obtained using integration by parts.

When the observation interval is finite, some spells are *interrupted* or *right-censored*; the survivor function defined up to the censoring point continues to characterize the data generation process. The mean duration of all spells whether ended naturally (at t) or by censoring (at t^c) is

(7) $$\mathbf{E}(\text{Min}(t,t^c)) = - \int_0^{t^c} \; t \, \nabla_t q(t|x,\xi) dt + t^c q(t^c|x.\xi) = \int_0^{t^c} \; q(t|x,\xi) dt.$$

Analogous formulas hold for the average hazard rate.

With sample attrition, the censoring time becomes a random variable, with an associated censoring survivor function $r(t^c|x,\xi)$. Then the probability that a spell is observed to extend to t is $q(t|x,\xi)r(t|x,\xi)$; the combined hazard rate for termination of an observed spell either naturally or by censoring is $h(t|x,\xi) - r'(t|x,\xi)/r(t|x,\xi)$; for a spell ending at time t, the probability that it is censored is $h(t|x,\xi)/(h(t|x,\xi) - r'(t|x,\xi)/r(t|x,\xi))$; and the mean duration of observed spells is

$$\int_0^{\infty} \; q(t|x,\xi) r(t|x,\xi) dt.$$

An example of a parametric duration model when x is time-invariant is the *Weibull* model, which specifies

(8) $$q(t|x) = \exp(-t^{\alpha} e^{-x'\beta}),$$

with α a positive parameter, β a vector of parameters, and x a vector of covariates. The associated hazard rate is

(9) $$h(t|x) = \alpha t^{\alpha-1} e^{-x'\beta}$$

and the mean duration of completed spells is

(10) $$\mathbf{E}(t|x) = e^{x'\beta/\alpha}\Gamma(1+1/\alpha),$$

where Γ is the gamma function. When $\alpha = 1$, this simplifies to the *exponential* duration model.

There are three strategies for statistical inference of censored duration data:

1. The fully parametric approach, with $q(t|x)$, or in the case of unobserved heterogeneity $q(t|x,\xi)$ and $\nu(\xi|x,0)$, assumed to be in a finite-parameter family.[1]

2. The fully nonparametric approach, in which $q(t|x)$ is estimated without parametric restrictions, using for example a Kaplan-Meier estimator.[2]

3. The single-index semiparametric approach, in which $q(t|x)$ depends on x through a scalar function $V(x,\beta)$ that is known up to a finite parameter vector β, but $q(t|v)$ is not confined to a parametric family. In the case of unobserved heterogeneity, either $q(t|v,\xi)$ or $\nu(\xi|v,t)$ may be nonparametric (but not both, without further restrictions, due to identification requirements).[3]

We survey some of the alternative semiparametric problems that have been discussed in the literature. Let x be a vector of covariates, assumed now to be *time-invariant*. Let β be a vector of unknown parameters, $V(x,\beta) \equiv x'\beta$ be a single index function known up to β, and $q(t|x'\beta)$ the survivor function. Let T^* be the random variable denoting completed duration, and T^c the censoring time, so observed duration is $T = \text{Min}(T^*,T^c)$. Four alternative models for T are

---

[1]Typical examples are a Weibull or log-normal distribution for $q(t|x)$, or an exponential distribution for $q(t|x,\xi)$ combined with a gamma distribution for ξ. The parameters of the distribution can be estimated by maximum likelihood.

[2]The classical Kaplan-Meier estimator is formulated for duration data without covariates. Suppose that in a data set spells starting at a common time 0 are observed to end (naturally or by censoring) at times $t_1 < ... < t_J$. Let n_j denote the number that end naturally at time t_j, and let m_j denote the number that are censored at this time. The total number "at risk" at time t_j is $N_j = \sum_{i=j}^{J} (n_i+m_i)$. The Kaplan-Meier estimate of the hazard rate at t_j is $h^*(t_j) = n_j/N_j$. A corresponding estimate of the survival function is $q^*(t_j) = (1-h^*(t_j))q^*(t_{j-1})$, or $q^*(t_j) = \prod_{i=1}^{j} (1-n_j/N_j)$. In the presence of categorical covariates, the Kaplan-Meier estimator obviously applies cell-by-cell for each configuration of the covariates. Using the nearest neighbor idea from non-parametric regression, the Kaplan-Meier estimator can be adapted to the general case of non-categorical covariates. In the case of unobserved heterogeneity, it is not possible in general to identify the structural survivor functions and the density of the unobserved covariates when both are non-parametric. Heckman and Singer (1984) establish this result, and also establish semiparametric methods for estimation of a parametric structural survivor function $q(t|x,\xi,\beta)$ in the presence of a non-parametric heterogeneity density $\nu(\xi|x,0)$.

[3]Other semiparametric approaches include multiple-index models and methods that parameterize quantiles without fully parameterizing the distribution.

1. *Regression model*: $\text{Ln } T^* = x'\beta + \varepsilon$, with $\varepsilon | x$ distributed with an unknown density $f(\varepsilon)$ with zero mean. The density f is often assumed symmetric and homoskedastic. This model yields the survivor function

(11) $$q(t|x'\beta) = 1 - F(\text{Ln } t - x'\beta),$$

where F is the cumulative distribution function of f. The associated hazard rate is

(12) $$h(t|x'\beta) = f(\text{Ln } t - x'\beta)/t[1 - F(\text{Ln } t - x'\beta)].$$

A generalized version of this model allows ε to be heteroskedastic, with variance depending on the index $x'\beta$, or more generally on some other function of x. The *censored regression model* is simply

(13) $$\text{Ln } T = \text{Min}(\text{Ln } T^c, x'\beta+\varepsilon);$$

it has the property in the case of non-stochastic censoring that

(14) $$\mathbf{E}(\text{Ln } T|x) = \int [1 - F(y-x'\beta)]dy$$

is an increasing function of $x'\beta$.

2. *Transformation* (Generalized Box-Cox) *model*: Suppose G is an unknown monotone increasing transformation from $(0,+\infty)$ onto the real line, and assume

(15) $$G(T^*) = x'\beta + \varepsilon,$$

with $\varepsilon | x$ distributed with a known or unknown density $f(\varepsilon)$. The associated survivor function is

(16) $$q(t|x'\beta) = 1 - F(G(t) - x'\beta),$$

and the associated hazard rate is

(17) $$h(t|x'\beta) = G'(t)f(G(t) - x'\beta)/[1 - F(G(t) - x'\beta)].$$

Again, the model can be generalized to allow heteroskedasticity depending on $x'\beta$.

3. *Projection Pursuit* (single index) *regression*: Suppose H is a unknown transformation from the real line into the real line. Assume

(18) $$\text{Ln } T^* = H(x'\beta) + \varepsilon,$$

with $\varepsilon | x$ distributed with a known or unknown density $f(\varepsilon)$. The associated survivor function is

(19) $$q(t|x'\beta) = 1 - F(\text{Ln } t - H(x'\beta)),$$

and hazard rate is

(20) $$h(t|x'\beta) = f(\text{Ln } t - H(x'\beta))/t[1 - F(\text{Ln } t - H(x'\beta))].$$

The error distribution is usually assumed homoskedastic, but some estimators for this model permit heteroskedasticity depending on $x'\beta$.

4. *Proportional Hazards model*: Let $h_o(t)$ be an unknown nonnegative "baseline hazard" function, and assume the covariates exert a proportional effect on the hazard, so that

(21) $$h(t|x) = h_o(t)\exp(-x'\beta).$$

Define the integrated baseline hazard

(22) $$\Lambda_o(t) = \int_0^t h_o(s)ds.$$

Then the survivor function is

(23) $$q(t|x'\beta) = \exp\left(-\Lambda_o(t)\ e^{-x'\beta}\right),$$

and

(24) $$\text{Ln } \Lambda_o(T^*) = x'\beta + \varepsilon,$$

where ε has the extreme value cumulative distribution function

(25) $$F(\varepsilon) = 1 - \exp\left(-e^{-\varepsilon}\right).$$

Other error distributions may result from a proportional hazards model with unobserved heterogeneity. For example, following Lancaster (1979), assume

(26) $$h(t|x,\xi) = h_o(t)\exp(-x'\beta)\xi,$$

with ξ having a gamma density, $\nu(\xi|x,0) = \xi^{\theta-1}e^{-\xi}/\Gamma(\theta)$. Then, applying the relation (1),

(27) $$q(t|x) = \left(1 + e^{\Lambda_o(t) - x'\beta}\right)^{-\theta},$$

which implies that (15) holds with ε having a generalized logistic distribution (or, e^{ε} having a Pareto distribution),

(28) $$F(\varepsilon) = 1 - (1+e^{\varepsilon})^{-\theta}.$$

The average hazard for (26),

(29) $$h^*(t|x) = \theta h_0(t) e^{\Lambda_o(t)} / \left(e^{\Lambda_o(t)} + e^{x'\beta} \right) ,$$

is no longer of the proportional hazards form. The conditional distribution of the unobserved covariates given survival $\nu(\xi|x,t)$ remains Gamma with parameter θ, but in the transformed variable

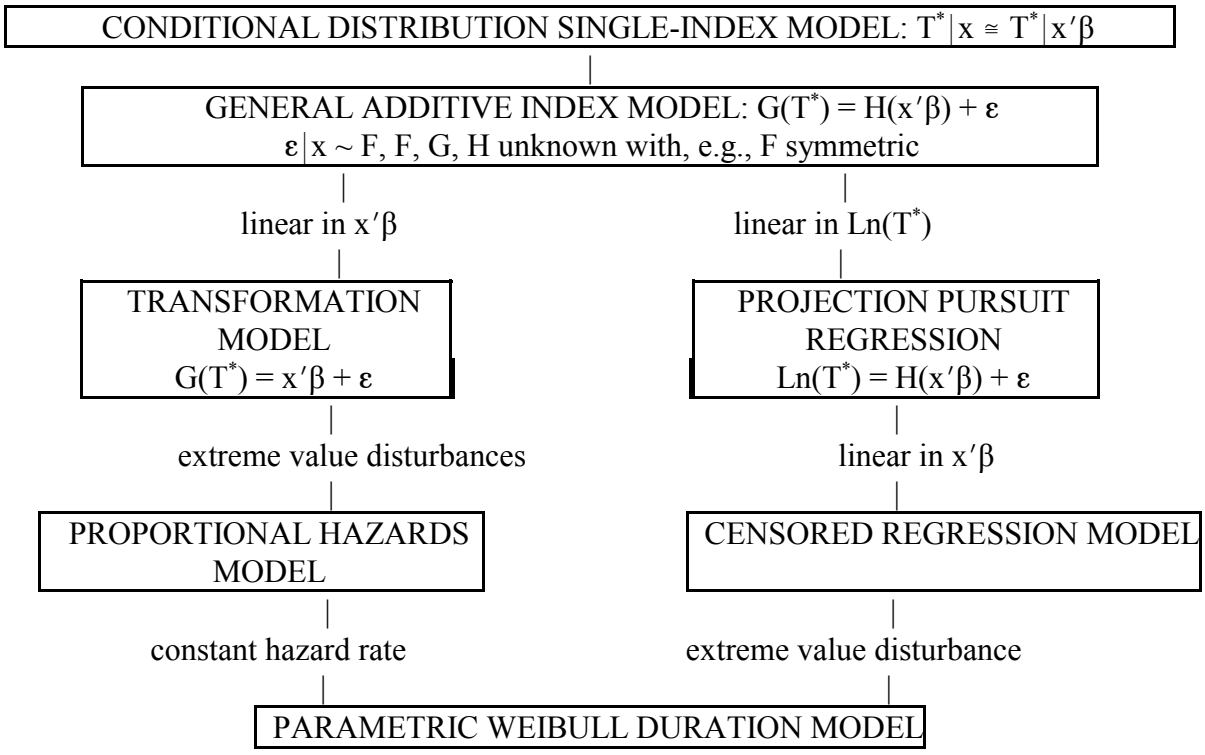
$$(1 + e^{\Lambda_o(t) - x'\beta})\xi \ .$$

The proportional hazards model (21) is a special case of the transformation model where the disturbance has the distribution (25). The proportional hazards model with heterogeneity (26) is also a specialization of the transformation model. When the baseline hazard varies with a power of t, $h_o(t) = \alpha t^{\alpha-1}$, (21) specializes to the parametric Weibull duration model, and also can be interpreted as a censored regression model with extreme value distributed disturbances.

FIGURE 1. SINGLE-INDEX MODELS

Observation Rules: $T = \text{Min}(T^c, T^*)$ for right-censored data
$T = \text{sgn}(\text{Ln}(T^*))$ for binomial discrete response data

(Specificity Increases as You Move Down the Table)



A common "generalized additive single-index" model in which the four models above are nested is

$$G(T^*) = H(x'\beta) + \varepsilon, \tag{30}$$

with ε distributed with cumulative distribution function F. The associated survivor function is

$$q(t|x'\beta) = 1 - F(G(T) - H(x'\beta)). \tag{31}$$

Figure 1 shows the logical relationship between these models. All the models are special cases of *single-index sufficiency* where the conditional distribution of the dependent variable depends on covariates x solely through the index $x'\beta$. The proportional hazards model and the censored regression model are logically distinct, except when both specialize to the common Weibull parametric model. Both are specializations of the transformation model. The censored regression model is a specialization of the projection pursuit regression model. The transformation model can be rewritten as a heteroskedastic projection pursuit model: If $G(T^*) = x'\beta + \varepsilon$ with G monotone increasing, then $\text{Ln } T^* = H(x'\beta) + \zeta$, where $H(x'\beta) = \mathbf{E}_\varepsilon \text{Ln } G^{-1}(x'\beta + \varepsilon)$, and ζ has the distribution function $F(G(\exp(\zeta + H(x'\beta)) - x'\beta)$, which in general is heteroskedastic.

The statistical issues that arise in application of these models are the (large sample and, potentially, small sample) distributional properties of estimators that are available under various assumptions, and the efficiency of alternative estimators. Most of the work to date has concentrated on finding computationally feasible estimators, establishing consistency and asymptotic normality, and establishing asymptotic efficiency bounds.

Horowitz and Newmann use two estimators for the censored regression model, a quantile estimator (Powell, 1986) and one-step semiparametric generalized least squares estimator (SGLS) (Horowitz, 1986). Other estimators that have been proposed for this problem include flexible parametric approximation of the cumulative distribution function (e.g., Duncan, 1986, who considers spline approximations--the "method of sieves"). Chamberlain (1986) and Cosslett (1987) have established for the censored regression problem the existence of a positive information bound on the parametric part. This suggests that the it is adequate to use relatively crude estimators of the nonparametric part in order to achieve |n asymptotically normal estimation of the parametric part. The Powell and Horowitz estimators have been shown |n asymptotically normal. Neither achieves the information bound for i.i.d. errors, and in general neither is efficient relative to the other.

Estimation of the proportional hazards model with an unknown baseline hazard function has been studied extensively; see Kaplan and Meier (1968), Cox (1972), Kalbfleisch and Prentice (1982), and Meyer (1990). A particularly useful "semiparametric" method for this model, applicable to the case where duration is measured in "weeks", is to flexibly parameterize the baseline hazard; Meyer (1990) shows this method is root-n asymptotically normal.

Estimators for the projection pursuit (single index) model have been proposed by Ichimura (1987), Ruud (1986), Stoker (1986), and Powell, Stock, and Stoker (1989). The Ichimura estimator chooses β to minimize the conditional variance of Ln T given $x'\beta$, using a kernel estimator of the conditional mean to form an estimate of the conditional variance. This estimator is consistent even if the disturbances are heterogeneous in the index function, so it can also be applied to the

transformation model. The Ichimura estimator is $n^{1/2}$ asymptotically normal, and has recently been argued to achieve the semiparametric information bound for the homoskedastic projection pursuit problem with normal disturbances. It is almost certainly not efficient for the transformation model. The Ruud and Stoker estimators rely on the fact that under suitable conditions the regression of Ln T on x is proportional to β; these are also |n asymptotically normal.

An estimator for the transformation model, applicable also to the proportional hazards model, is the maximum rank correlation method of Han (1987) and Doksum (1985).

Newey (1990) has established the asymptotic efficiency of some kernel and quantile estimators for the censored regression model when error distributions are symmetric. The status of these estimators under some other information conditions remains unresolved. A problem requiring further work is construction of reliable and practical covariance estimators for the semiparametric estimators. An interesting empirical question is whether the censored regression model or the proportional hazards models can be accepted as restrictions on the transformation model (and what are appropriate and practical test statistics)?

Stated Willingness-to-Pay for a Natural Resource

A method for eliciting *Willingness-to-Pay* (WTP) for natural resources is a *referendum contingent valuation* experiment: Survey respondents are asked if they are willing to pay an amount *b*, where *b* is a bid set by experimental design. Let *d* denote a dummy variable that is one for a "Yes" response, zero otherwise. A sample of n observations are collected on (*b*,*d*) pairs, plus covariates x characterizing the respondent. Suppose WTP is distributed in the population as $w = x\beta - \varepsilon$, where ε has a cumulative distribution function $G(\varepsilon)$ that is independent of x. Then, $\Pr(d=1 \mid x'\beta) = G(x\beta - b)$, or

$$d = G(x\beta - b) + \varepsilon, \tag{32}$$

Suppose β and the function G are unknown. The econometric problem is to estimate β and, if necessary, G, and use these to estimate a measure of location of the distribution of WTP, conditional on x or unconditional. This is an example of a projection-pursuit regression model.

Contingent valuation experiments are controversial because they are very sensitive to psychometric context effects, such as anchoring that leads respondents who are unsure about their preferences to take the offered bid as a cue to the "politically correct" range of values. Some subjects also appear to misrepresent their responses strategically, giving extreme values that they would not practically pay, but which express "protest" positions. These effects make WTP estimates imprecise, and their connection to welfare economics tenuous.

Why do contingent valuation experiments use the referendum elicitation format, rather than a format in which subjects would be asked to give an open-ended WTP response? One answer is that the open-ended format produces a much higher non- response rate, so that the referendum method reduces selection bias caused by non- response. Another is that psychologically the referendum and open-ended methods elicit quite different behaviors. Some argue that the referendum format is closer to the voting mechanisms used elsewhere to make social decisions, and there is a virtue in mimicking this mechanism for social decisions on natural resources.

One issue that enters the contingent valuation experimental design is the location of the bid levels b. Alternatives are to randomize b, or to choose b on a grid with a specified mesh. In practice, coarse meshes have been used, which limits the accuracy of semiparametric estimates. Let $h(b|x)$ be the density from which the bid level b is drawn, given x. Since this is chosen by experimental design, it is known to the analyst.

Econometric analysis of referendum WTP data can use the fact that (32) is a binary response model and a single-index model (that is heteroskedastic, but with the heteroskedasticity depending on the index). Then, available methods to estimate β are the Manski (1978) maximum score estimator, the Cosslett (1987) semiparametric maximum likelihood estimator, the Ichimura (1986) estimator that minimizes expected conditional variance, the Horowitz (1992) estimator that is a smoothed version of maximum score, and the Klein-Spady (1993) estimator. The key result for the binomial response model is that under some smoothness conditions, there are root-n consistent estimators β_n for β; i.e., $n^{1/2}(\beta_n - \beta)$ is asymptotically normal. A nonparametric estimator of G can be obtained jointly with the estimation of β, as in the Cosslett procedure, or by conventional kernel methods in a second step after the β estimate is plugged in to form the index; it can be estimated only at a nonparametric rate less than root-n.

One particularly simple estimator for the index parameters β has been proposed for this problem by Lewbel and McFadden (1997): Carry out a least squares regression on the model,

$$(33) \qquad (d_i - 1(b_i<0))/h(b_i|x_i) = x_i\beta + \zeta_i.$$

The authors show that the coefficients from this regression are consistent for β, and are asymptotically normal at a $n^{1/2}$ rate. The estimates are not particularly efficient, but their simplicity makes them an excellent starting point for analysis of model specification and construction of more efficient estimators.

Exercise 5. Prove that the estimator based on (33) is consistent. Apply a law of large numbers to conclude that

$$\frac{1}{n}\sum_{i=1}^{n} x_i'(d_i - 1(b_i<0))/h(b_i|x_i) \to_p \mathbf{E}_x\mathbf{E}_{b|x}\, x'(G(x\beta-b) - \mathbf{1}(b<0))/h(b|x).$$

Then apply integration by parts to conclude that

$$\mathbf{E}_{b|x}\, x(G(x\beta-b) - \mathbf{1}(b<0))/h(b|x) = x\cdot \int_{-\infty}^{0} (G(x\beta-b) - 1)\cdot db + x\cdot \int_{0}^{+\infty} G(x\beta-b)\cdot db$$

$$= x' \int_{-\infty}^{+\infty} b\cdot G(x\beta-b)\cdot db = x'x\beta.$$

From this conclude that the least squares coefficients converge to $(\mathbf{E}x'x)^{-1}(\mathbf{E}x'x\beta) = \beta$.

The authors also establish that the r-th moment of WTP, conditioned on $x = x_o$, can be estimated consistently at a root-n rate by

(34) $$M_r = (x_o\beta)^r + r \sum_{i=1}^{n} (b_i+(x_o-x_i)\beta)^{r-1} \cdot \frac{d_i - \mathbf{1}((x_i\beta>b_i)}{\sum_{j=1}^{n} h(b_i+(x_j-x_i)\beta|x_j)} .$$

The estimators (33) and (34) are good examples of statistical procedures for a semiparametric problem that are "robust" in the sense that they do not depend on parametric assumptions on the distribution of WTP, and provide an easily computed alternative to use of a kernel-type nonparametric estimator.

6. SIMULATION METHODS AND INDIRECT INFERENCE

Econometric theory has traditionally followed classical statistics in concentrating on problems that yielded analytic solutions. This explains the emphasis on the linear model, and on asymptotic approximations in situations where nonlinearities or other factors make exact sample analysis intractable. Increased computational power, and better understanding of the uses and limitations of numerical analysis, have greatly expanded the ability of econometricians to explore the characteristics of the methods they use under realistic conditions. The idea is straightforward. The economist can write down one or more trial data generation processes, perhaps after an initial round of econometric analysis, and use these data generation processes to generate simulated or virtual samples. If a comparison of a real sample with these virtual samples reveals inconsistencies, this is evidence that the trial data generation process is unrealistic. Conversely, if the econometrician has discovered the true data generation process, then the virtual samples generated from it should not differ systematically from the real sample. Computers and Monte Carlo simulation methods come in at the stages of drawing the virtual samples and comparing the real and virtual samples.

If the kinds of comparisons just described are done casually, without attention to statistical properties, they can mislead the analyst. Traditional calibration exercises in economics and other disciplines often suffer from this deficiency. However, it is possible to develop a statistical theory to support these comparisons, and use this theory to consistently identify the real data generation process, or good approximations to it. In various manifestations, this theory has been developed by Hendry, Mizon, and Richard under the name *encompassing*, by Gourieroux and Monfort under the name *indirect inference*, and by McFadden under the name *simulation-assisted inference*.

Consider two parametric families of data generation processes, H_f containing models $f(y|x,\alpha)$ for parameter vectors α in a set **A**, and H_g containing models $g(y|x,\beta)$ for parameter vectors β in a set **B**. Both of these families have the same dependent variable y, and are conditioned on the same explanatory variables x. It may be the case that one of these families is nested within the other; this is the situation in classical hypothesis testing where the null hypothesis (say H_g) is a subset of the universe (say H_f), and the true data generation process is a member of H_f and under the null a member of H_g. However, we will now consider more general situations where the two families are not necessarily nested, and the true data generation process may not be in either.

Example. The family H_f is the family of linear models $y = x\gamma + \varepsilon$, where x is a vector of explanatory variables and ε is a normal disturbance with variance σ^2. This family is parameterized by $\alpha' = (\gamma,\sigma^2)$. H_g is the family $y = z\delta + \eta$, where z is a vector of explanatory variables and η is a normal disturbance with variance λ^2, parmeterized by $\beta = (\delta,\lambda^2)$. The vectors x and z may have some variables in common, but in the most general case will each contain some distinct variables

so that neither is contained (nested) within the other. $y = x\alpha + \varepsilon$ and the family H_g of linear models $y = z\beta + \eta$, where x and z may have some variables in common, but also contain distinct variables corresponding to alternative theories of the determination of y. The families are said to be *non-nested* when neither can be written as a linearly restricted case of the other.

A proximity measure between densities is the Kullback-Leibler Information Criterion (KLIC),

$$K_{fg}(\alpha,\beta,x) = \int \log(f(y|x,\alpha)/g(y|x,\beta))\cdot f(y|x,\alpha)dy.$$

The KLIC is always non-negative, and is zero only if *f* and *g* coincide. This measure depends on exogenous variables x. We could alternately take its expectation with respect to x,

$$K_{fg}(\alpha,\beta) = \mathbf{E}_x K_{fg}(\alpha,\beta,x)$$

and approximate this expectation by a sample average

$$K_{fgn}(\alpha,\beta) = \frac{1}{n}\sum_{i=1}^{n} K_{fg}(\alpha,\beta,x_i).$$

For the model *g*, define the *pseudo-true value* $\beta_f(\alpha)$ to be the $\beta \in B$ that minimizes $K_{fg}(\alpha,\beta)$, and the *conditional pseudo-true value* $\beta_{fn}(\alpha)$ to be the $\beta \in B$ that minimizes $K_{fgn}(\alpha,\beta)$. Then, $g(y|x,\beta_f(\alpha))$ is the data generation process in the *g* family closest to $f(y|x,\alpha)$, and

$$J_f(\alpha,B) \equiv K_{fg}(\alpha,\beta_f(\alpha))$$

is the proximity of *f* and the $g \in H_g$ that is closest to *f*. In an earlier chapter, where $f(y|x,\alpha_o)$ was identified as the true data generation process, we called $g(y|x,\beta_f(\alpha_o))$ the *least misspecified* model in H_g. However, we will now consider more general situations where the *f* family may not contain the true data generation process.

Exercise 6. In the linear model example, Show that

$$\log(f/g) = 0.5\cdot\{\log(\lambda^2/\sigma^2) - (y-x\gamma)^2/\sigma^2 + (y-z\delta)^2/\lambda^2\}$$
$$= 0.5\cdot\{\log(\lambda^2/\sigma^2) - (y-x\gamma)^2(1/\sigma^2 - 1/\lambda^2) + 2(y-x\gamma)(x\gamma-z\delta)/\lambda^2 + (x\gamma-z\delta)^2/\lambda^2\},$$

and hence that $K_{fg}(\alpha) = 0.5\cdot\{\log(\lambda^2/\sigma^2) + \sigma^2/\lambda^2 - 1 + \mathbf{E}(x\gamma-z\delta)^2/\lambda^2\}$. The pseudo-true values in the model H_g are the values $\beta_f(\alpha)$ that minimize $K_{fg}(\alpha)$. Show that the pseudo-true value for δ is $(\mathbf{E}z'z)^{-1}(\mathbf{E}z'x)\gamma$ and the pseudo-true value for λ^2 is $\sigma^2 + \gamma'\{\mathbf{E}x'x - (\mathbf{E}x'z)(\mathbf{E}z'z)^{-1}(\mathbf{E}z'x)\}\gamma$. Show that the minimum distance from f to H_g is

$$J_f(\alpha) = 0.5\cdot\log(1 + \gamma'\{\mathbf{E}x'x - (\mathbf{E}x'z)(\mathbf{E}z'z)^{-1}(\mathbf{E}z'x)\}\gamma/\sigma^2).$$

The distance is zero if z can be written as a linear combination of the variables in x

A model $f(y|x,\alpha)$ is said to *encompass* the family g if f can account for, or explain, the results obtained with the g family. Operationally, this concept says the g family will fit similarly the observed sample data and virtual data generated by the model $f(y|x,\alpha)$. If we define

$$b_n = \text{argmax}_\beta \sum_{i=1}^{n} g(y_i|x_i,\beta);$$

to be the maximum likelihood estimate from the family H_g for the observed sample, and $f(y|x,\alpha)$ encompasses H_g, then b_n should converge to the pseudo-true value $\beta_f(\alpha)$. Conversely, if $b_n - \beta_f(\alpha)$ converges to a non-zero limit, $f(y|x,\alpha)$ fails to encompass H_g. This is the same as saying that as judged from the family H_g, samples generated by the model $f(y|x,\alpha)$ look like samples generated by the true data generation process.

Exercise 7. In the linear model example with n observations, write the models H_f and H_g as $y = X\gamma + \varepsilon$ and $y = Z\delta + \eta$ respectively. Show that the maximum likelihood estimates in the family H_g are $\delta_e = (Z'Z)^{-1}Z'y$ and $\lambda_e^2 = y'[I - Z(Z'Z)^{-1}Z']y/n$, and in the H_f family are $\gamma_e = (X'X)^{-1}X'y$ and $\sigma_e^2 = y'[I - X(X'X)^{-1}X']y/n$. Suppose the model $y = X\beta + \varepsilon$ with parameters α is true. Show that the differences of the maximum likelihood estimates in the H_g family and the corresponding pseudo-true values for this family, evaluated at α, converge in probability to zero.

If $f(y|x,\alpha_o)$ is the true data generation process, then by definition it encompasses any other family of models H_g. It is possible for a member of H_g to encompass the true data generation process $f(y|x,\alpha_o)$; this means that the member of g can generate data that looks like data drawn from $f(y|x,\alpha_o)$. This could obviously happen if H_g contains one or more models that are observationally equivalent to f, but could also occur if H_g contains models that are more "structural" than f so that they potentially can explain the same phenomena as f, and more.

In the theory of *tests of non-nested hypotheses*, the setup is to have two families of data generation processes, H_f and H_g, which are not nested, with the true data generation process assumed to be in one of the two families. Then, the family containing the true data generation process will encompass the other, but not vice versa (except in the unidentified case where there are models in either family that can mimic the true data generation process). Let a_n be the maximum likelihood estimator of α from the model $f(y|x,\alpha)$. Then $b_n - \beta_{fn}(a_n)$ converges to zero if and only f encompasses g, and $a_n - \beta_{gn}(b_n)$ converges to zero if and only if g encompases f. These observations form the basis for practical test statistics for non-nested hypotheses; see Pesaran (1987) and Gourieroux & Monfort (1994). These ideas also form the basis for an estimation method called *indirect inference*, or in a more general but less focused form, *method of simulated moments*: If the family H_f contains the true data generation process $f(y|x,\alpha_o)$, then this model encompasses g and one has $b_n - \beta_{fn}(\alpha_n)$ converging to zero if α_n converges to α_o, and with an assumption of identifiability, to a non-zero limit if α_n converges to something other than α_o. Then, choosing α_n to make $b_n - \beta_{fn}(\alpha_n)$ small will under some regularity conditions make these estimators consistent for α_o. The reason to consider these indirect estimates, rather than direct maximum likelihood estimates of α from the model $f(y|x,\alpha)$, is that the true model may be very complex or very difficult to work with computationally. For example, $f(y|x,\alpha)$ may involve a complex structural model, or may involve probabilities that require high-dimensional numerical integration to evaluate. Then, the indirect inference may utilize a

simpler family of models H_g that are easier to compute or more "robust". For example, g may be a reduced form model and indirect inference may involve choosing structural parameters so that their transformation to reduced form parameters gives the same values as direct least squares estimation of the reduced form. Or, indirect inference may utilize a select list of moment conditions that you are confident hold in the population. The reason simulation methods enter is that the practical way to calculate $\beta_{fn}(\alpha_n)$ is to use Monte Carlo methods to draw virtual samples from the data generation process $f(y|x,\alpha)$ for various trial α, and select α_n to minimize the distance between the estimator b_n from the observed sample and estimators $b_n(\alpha)$ obtained from a virtual sample from $f(y|x,\alpha)$ by estimating β by maximum likelihood estimation applied to this virtual sample. Because this process can also be interpreted as matching the "moments" b_n from the virtual sample with simulated "moments" $b_n(\alpha)$ from the simulated virtual sample by varying α, it is also called the *method of simulated moments*.

Encompassing is a limited concept when comparing the true data generation process with an alternative, since the true data generation process will encompass any alternative model. However, it becomes more general and more interesting under two circumstances: (1) the true data generation process may fail to lie in either H_f or H_g, or (2) the results from H_f and H_g are based on limited information, such as GMM estimates that rely on specific orthogonality conditions, rather that a full parametric specification of a data generation process. Then, encompassing can be a useful approach to model selection.

We will not attempt to provide any general introduction to simulation and Monte Carlo methods in these notes. However, there a few key concepts that are important enough to introduce at this stage. First consider the problem of drawing a virtual sample from the data generation process $f(y|x,\alpha)$ for a trial value of α. Consider the simplest case when y is one-dimensional. The corresponding CDF $U = F(Y|x,\alpha)$ has a uniform distribution, and a Monte Carlo draw of y for observation i is $y^* = F^{-1}(u_i|x_i,\alpha)$, where u_i is a draw from a uniform distribution. This is a practical method of drawing a realization of a random variable if F^{-1} can be determined analytically or efficiently evaluated numerically. When it is impractical to calculate F^{-1}, one may be able to use *Monte Carlo Markov Chain* (MCMC) methods. A *Metropolis-Hastings* (MH) *sampler* for $f(y|x,\alpha)$ is defined by a conditional density $q(y'|y,x)$ chosen by the analyst and kernel $w(y,y',x) = \text{Min}\{q(y'|y,x), f(y'|x,\alpha)\cdot q(y|y',x)/f(y|x,\alpha)\}$. This kernel is associated with a transition process in which y' is sampled from $q(y'|y,x)$, then the process moves to y' with probability $p(y,y',x)$, and otherwise stays at y, where $p(y,y',x) = \text{Min}\{1, q(y|y',x)\cdot f(y'|x,\alpha)/q(y'|y,x)\cdot f(y|x,\alpha)\}$. A simple choice for $q(y'|y,x)$ is a density $q(y')$ independent of y and x from which it is computationally easy to draw and which has the property that $f(y|x,\alpha)/q(y)$ is never too large, a key determinant of the efficiency of the sampling process. The MH sampler is a generalization of what are called *acceptance/rejection* methods.

The Metropolis-Hastings sampler starts from an arbitrary point, and proceeds recursively. Suppose at step t-1, the draw is y^{t-1} and $f_{t-1} = f(y^{t-1}|x,\alpha))$. Draw y' from the conditional density $q(\cdot|y^{t-1})$, and define $q_{t+} = q(y'|y^{t-1})$ and $q_{+t} = q(y^{t-1}|y')$, Calculate $\alpha(y^{t-1},y') = \text{Min}\{1, q_{+}f_t/q_{t+}f_{t-1}\}$. Draw a uniform [0,1] random number ζ. If $\zeta \le p(y^{t-1},y',x)$, set $y^t = y'$; otherwise, set $y^t = y^{t-1}$. Once it is "burned in", the sequence y^t behaves like a sample drawn from $f(\cdot|x,\alpha)$. Note that the terms in the sequence are not statistically independent. When one needs to form expectations with respect to $f(y|x,\alpha)$, these can be approximated by means over the y^t draws.

In indirect inference or method of simulated moments, one searches iteratively for parameter values that satisfy some criterion, such as minimizing the distance of $b_n - \beta_{fn}(\alpha)$ from zero, using simulation to approximate $\beta_{fn}(\alpha)$. It is important in doing this that the simulated value of $\beta_{fn}(\alpha)$, considered as a function of α, have a property called *stochastic equicontinuity*. Informally, this means that the simulator does not "chatter" as α varies. The way to accomplish this is to keep the Monte Carlo draws that drive the simulation fixed as α changes. For example, when a virtual sample from $f(y|x,\alpha)$ is drawn by the inverse method $y^* = F^{-1}(u|x,\alpha)$, keeping the uniformly distributed draws u fixed as α is varied does the job.

Further reading on simulation methods and indirect inference can be found in McFadden (1989), Gourieroux & Monfort (1994), and Hajivassiliou & Ruud (1994).

7. THE BOOTSTRAP

The idea fundamental to all of statistical inference is the principle that a statistical sample forms an *analogy* to the target population, and to estimate the results of an operation on the target population, one can complete the analogy by carrying out the same operation on the statistical sample. Thus, the sample mean is analogous to the population mean, and hence has decent statistical properties as an estimate of the population mean. Manski (1994) shows how this principle can guide the construction of estimators.

Extending the analogy principle, if one is interested in the relationship between a target population and a given sample drawn from this population, one could form an analogy by starting from the given sample, drawing subsamples from it, and forming analogous relationships between the original sample and the subsamples. When the subsamples are drawn with replacement and are the same size as the original sample, this is called the *bootstrap*.

To illustrate the operation of the bootstrap, suppose you have an estimate a_n of the parameter in a data generation process $f(y|x,\alpha)$, obtained from a sample of size n from the target population. You would like to know the variance of the estimator a_n. Note that this is a property of the relationship between the population and the sample that could in principle be determined by drawing repeated samples from the population, and estimating the variance of a_n from the repeated samples. The bootstrap idea is to start from the observed sample, draw repeated subsamples from it (with replacement), and complete the analogy by forming the estimator a^* for each subsample, and computing the sample variance of these estimators. The bootstrap process is computationally intensive, because it involves the subsampling process and the computation of a^*, repeated many times. Under very general regularity conditions, the analogy principle applies and the estimate of the variance of a_n formed in this way will have good statistical properties. Specifically, the bootstrap estimate of the variance of a_n will have the same properties as the first-order asymptotic approximation to the variance, without the effort of determining analytically and computing the asymptotic approximation. Further, the bootstrap estimator will under some conditions pick up higher order effects, so that it is a better finite sample approximation that the first-order asymptotic approximation. In particular, if the expression being studied has a limiting distribution that is independent of the parameters of the problem, as for example when one is interested in the finite sample distribution of the ratio of a parameter estimate to its standard error which has a limiting

T-distribution, the bootstrap will be more accurate for finite samples that the first-order asymptotic approximation. A statistic with the last property is called *pivotal*.

Bootstrap methods can often be used to estimate the distribution of statistics, for purposes of estimating moments or critical levels, in situations where asymptotic analysis is intractable or tedious. The bootstrap is itself one member of a broad class of techniques called *resampling methods*. There are various pitfalls to be avoided in application of resampling methods, and a variety of shortcuts and variants that can speed calculation or make them more accurate. For further reading, see Efron & Tibshirani (1993), Hall (1994), and Horowitz (1999).

REFERENCES

Chamberlain, G. (1986) "Asymptotic Efficiency in Semiparametric Models with Censoring" *J. of Econometrics* **29**, 189-218.

Chamberlain, G. (1992) "Efficiency Bounds for Nonparametric Regression" *Econometrica* **60**, 567-96.

Cosslett, S. (1987) "Efficiency Bounds for Distribution-Free Estimators of the Binary Choice and the Censored Regression Models" *Econometrica* **55**, 559-585.

Cox, D. (1972) "Regression Models and Life Tables" *Journal of the Royal Statistical Society* **B, 34**, 187-220.

Delgado, M. and P. Robinson (1992) "Nonparametric and Semiparametric Methods for Economic Research" *J. of Economic Surveys* **6**, 201-49.

Doksum, K. (1985) "An Extension of Partial Likelihood Methods for Proportional Hazard Models to General Transformation Models" U. of California, Berkeley working paper.

Duncan, G. (1986) "A Semiparametric censored Regression Estimator" *Journal of Econometrics* **29**, 5-34.

Gourieroux, C. and A, Monfort (1994) "Testing Non-Nested Hypotheses," in R. Engle and D. McFadden (eds) Handbook of Econometrics IV, North-Holland: Amsterdam, 2585- 2640.

Hajivassiliou, V. and P. Ruud (1994) "Classical Estimation Methods for LDV Models using Simulation," in R. Engle and D. McFadden (eds) Handbook of Econometrics IV, North-Holland: Amsterdam, 2384-2443.

Hall, P. (1994) "Methodology and Theory for the Bootstrap," in R. Engle and D. McFadden (eds) Handbook of Econometrics IV, North-Holland: Amsterdam, 2342- 2383.

Han, A. (1987) "Nonparametric Analysis of Generalized Regression Models: The Maximum Rank Correlation Estimator", *J. of Econometrics* **35**, 303-16.

Hardle, W. and O. Linton (1994) "Applied Nonparametric Methods," in R. Engle and D. McFadden (eds) Handbook of Econometrics IV, North-Holland: Amsterdam, 2297- 2341.

Heckman, J. and B. Singer (1984) "A method for minimizing the impact of distributional assumptions in econometric models for duration data" *Econometrica* **52**, 271-320.

Horowitz, J. (1986) "A Distribution-Free Least Squares Method for Censored Linear Regression Models" *J. of Econometrics* **29**, 59-84.

Horowitz, J. (1989) "Semiparametric M-Estimation of Censored Linear Regression Models" in G. Rhodes and T. Fomby (eds) *Nonparametric and Robust Inference*, Advances in Econometrics **7**, 45-83.

Horowitz, J. (1992) "A Smoothed Maximum Score Estimator for the Binary Response Model" *Econometrica* **60**, 505-31.

Horowitz, J. and G. Newmann (1987) "Semiparametric Estimation of Employment Duration Models" *Econometric Reviews* **6**, 5-40.

Horowitz, J. and G. Newmann (1989) "Computational and Statistical Efficiency of Semiparametric GLS Estimators" *Econometric Reviews* **8**, 223-25.

Ichimura, H. (1986) "Estimation of Single Index Models" PhD Dissertation, MIT.

Kalbfleisch, J. and R. Prentice (1980) *The Stochastic Analysis of Failure Time Data*, New York: Wiley.

Kaplan, E. and P. Meier (1958) "Nonparametric estimation from Incomplete Observations" *J. American Statistical Association* **53**, 487-491.

Klein, R. and R. Spady (1993) "An Efficient Semiparametric Estimator for Binary Response Models" *Econometrica* **61**, 387-422.

Lancaster, T. (1979) "Econometric methods for the duration of unemployment" *Econometrica* **47**, 141-165.
Manski, C. (1978) "Maximum Score Estimation of the Stochastic Utility Model of Choice" *Journal of Econometrics* **3**, 205-228.
Manski, C. (1994) "Analog Estimation of Econometric Models," in R. Engle and D. McFadden (eds) Handbook of Econometrics IV, North-Holland: Amsterdam, 2560- 2584.
McFadden, D. (1989) "A Method of Simulated Moments for Estimation of Multinomial Probit Models without Numerical Integration," *Econometrica*,.
Meyer, B. (1987) "Unemployment Insurance and Unemployment Spells" *Econometrica* **58**, 757-82.
Newey, W. (1990) "Semiparametric Efficiency Bounds" *J. of Applied Econometrics* **5**, 99- 135.
Newey, W. and D. McFadden (1994) "Large Sample Estimation and Hypothesis Testing," in R. Engle and D. McFadden (eds) Handbook of Econometrics IV, North-Holland: Amsterdam, 2113-2247.
Pesaran, M. (1987) "Global and Partial Non-Nested Hypotheses and Asymptotic Local Power," *Econometric Theory* 3, 69-79.
Powell, J. (1986) "Censored Regression Quantiles" *J. of Econometrics* **29**, 143-155.
Powell, J., J. Stock, and T. Stoker (1989) "Semiparametric estimation of weighted average derivatives" *Econometrica* **57**, 1403-30.
Powell, J. (1994) "Semiparametric Econometric Methods" *Handbook of Econometrics*, Vol. 4, R. Engle and D. McFadden (eds.), North Holland.
Ritov, Y. (1985) "Efficient and unbiased estimation in nonparametric linear regression with censored data" U. of California, Berkeley working paper.
Robinson, P. (1986) "Semiparametric Econometrics: A Survey" London School of Economics working paper.
Ruud, P. (1986) "Consistent Estimation of Limited Dependent Variable Models Despite Misspecification of Distribution" *J. of Econometrics* **29**, 157-187.
Serfling, R. (1980) *Approximation Theorems of Mathematical Statistics*. Wiley: New York.
Silverman, B. (1986) *Density Estimation*. Chapman and Hall: London.
Stoker, T. (1986) "Consistent Estimation of Scaled Coefficients" *Econometrica* **54**, 1461-1481.
Stone, C. (1977) *Annals of Statistics*.