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 $f(y \mid x)$.

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, $\theta(x) = E(t(Y,x) \mid X=x)$, is the target of the econometric investigation.

Examples of transformations of interest are

(1) $t(y,x) \equiv y$, in which case $\theta(x) = E(Y \mid X=x)$ is the conditional expectation of Y given x, or the *regression function* of Y on x;

(2) $t(y,x) = yy'$, in which case $\theta(x) = E(YY' \mid X=x)$ is the array of second conditional moments, and this function combined with the first example, $E(YY' \mid X=x) - \{E(Y \mid X=x)\}\{E(Y \mid X=x)\}'$ is the conditional variance; and

(3) $t(y,x) = 1_A(y)$, 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 $E(\varepsilon \mid x) = 0$.

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.

Where can an econometrician go wrong in setting out to analyze the generalized regression relationship *t*(y,x) = θ(x) + ε? First, there is nothing in the formulation of this model *per se* that assures that θ(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.

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.

The only property that is guaranteed for the disturbances ε when $\theta(x)$ is correctly specified is the conditional first moment condition $E(\varepsilon \mid x) = 0$. 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 canbe used to estimate 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.

Example: If F is a CDF with density f on $\mathbb{R}$, then $G(u) = F(\Phi^{-1}(u))$ is a CDF on (0,1) with density $g(u) = f(\Phi^{-1}(u))/\phi(u)$.

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:

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

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

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

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*,**

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

Then, the MSE is bounded by

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

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),

$$\text{MISE} = E\int(\hat{g}(x) - g(x))^2dx$$

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

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

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

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

$$\leq K/n + L^2/3K^2.$$

The RHS of this expression is minimized at $K = (2L^2n/3)^{1/3}$, with MISE $\leq (3L/2n)^{2/3}$. 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 $K = 20$.