Regression

by CHARLES F. MANSKI
University of Wisconsin–Madison

The writing of this expository article was supported by National Science Foundation grant SES-8805276. I am grateful to Arthur Goldberger, Peter Gottschalk, and Donald Hester for their useful comments.

I. Introduction

There was a time when a researcher reporting an estimate of a regression did not need to specify the kind of regression nor the type of estimate. The regression of a random variable $y$ on another such variable $x$ was understood to be the mean of $y$ conditional on $x$, considered as a function of $x$. The estimate was inevitably a least-squares fit of the sample data on $x$ to the data on $y$.

The term regression is used in a much broader sense today. A regression of $y$ on $x$ is any feature of the probability distribution of $y$ conditional on $x$, considered as a function of $x$. The feature of interest might be the mean, median, mode, or variance. Hence we speak of the mean regression, median regression, mode regression, variance regression, and so on.

Whatever the regression of interest, one can now select among many estimates. The selection made depends primarily on the prior information one is willing to assert. If little is known about the form of the regression, nonparametric estimators may be applied. If a parametric regression model has been specified, method-of-moments estimates may be used.¹

This article exposit the achievements and concerns of the rich modern literature on the estimation of regressions. To keep the discussion within bounds, I restrict attention to the case in which $y$ is scalar and focus on those regressions that provide best predictors of $y$ conditional on $x$. This class of regressions is easily described and covers the majority of applications.

Section II introduces conditional prediction problems. Section III examines the main parametric and nonparametric approaches to the estimation of best-predictor regressions in the context of random sampling. Section IV considers estimation from nonrandom samples. Section V makes concluding comments.

As we begin I must caution readers

¹The terms parametric and nonparametric are conventionally used to distinguish those problems in which the regression is known up to a finite-dimensional parameter from those in which the regression is known to be a member of some non-finite-dimensional space of functions of $x$. For example, the regression might be known to be a continuous function of $x$. Use of the term nonparametric to mean that the parameter space is a space of functions is an illogical but firmly entrenched semantic convention.
that the richness of modern regression theory can be unsettling. The old convention treating regression as synonymous with mean regression and estimation as synonymous with least-squares was arbitrary but simple. Recognizing that many regressions are potentially of interest and that many estimates have attractive properties, one now must make choices. Economic theory asserts that offering a person new options cannot decrease his welfare. Nevertheless, I have found that researchers often prefer the comfort of convention to the freedom of choice.

II. Conditional Prediction Problems

A conditional prediction problem presumes that one observes a realization of a random vector \( x \) and wishes to predict the realization of some random variable \( y \). A best predictor of \( y \) conditional on \( x \) is a prediction that minimizes expected loss with respect to a specified loss function. That is, a best predictor solves a minimization problem

\[
\min_p E[L(y - p)|x]. \tag{1}
\]

Here \( L(\cdot) \) denotes the loss function, \( p \) denotes a predictor for \( y \), and \( E[L(y - p)|x] \) denotes expected loss, conditional on \( x \), when \( p \) is used to predict \( y \). A "loss function" is any function \( L(\cdot) \) satisfying the condition

\[
0 < u < v \Rightarrow 0 = L(0) \leq L(u) \leq L(v) \leq L(\infty),
\]

where \( u \) and \( v \) denote values for the prediction error \( y - p \). With the loss function fixed, the value of the best predictor depends only on the probability distribution of \( y \) conditional on \( x \). Hence, considered as a function of \( x \), the best predictor is a regression.

Best predictors are the most extensively studied of all regressions. One reason is that conditional prediction problems often arise in practice. Another is that best predictors are useful summary statistics; they describe succinctly the manner in which the "location" of \( y \) varies with \( x \). There are, however, many interesting loss functions and correspondingly many ways to interpret formally the "location" of a random variable. Some leading examples follow. For proofs of and elaboration on the results stated in the remainder of Section II, see Thomas Ferguson (1967) and Manski (1985b, chs. 4 and 6).

Square Loss. Let \( L(\cdot) \) be the square loss function

\[
L(u) = u^2 \tag{2}
\]

Then it can be shown that the best predictor is the mean of \( y \) conditional on \( x \).

Absolute Loss. Let \( L(\cdot) \) be the absolute loss function

\[
L(u) = |u|. \tag{3}
\]

Then the best predictor is the median of \( y \) conditional on \( x \).

Step Loss. For a given positive constant \( \delta \), let \( L(\cdot, \delta) \) be the step loss function

\[
L(u) = 0 \text{ if } |u| < \delta, \quad 1 \text{ if } |u| \geq \delta. \tag{4}
\]

Then the best predictor is the midpoint of the interval of length \( 2\delta \) that has the highest probability of containing \( y \). As \( \delta \) approaches zero, the best predictor approaches the mode of \( y \) conditional on \( x \). For fixed \( \delta \), the best predictor may be referred to as the \( \delta \)-mode of \( y \) conditional on \( x \).

\[\] some auxiliary rule to pick out one solution for each value of \( x \).

Defining a best predictor to be one that minimizes expected loss is only a convention, but one that will be maintained here. The qualitative idea of a best predictor is retained if one minimizes some other location function of the loss distribution, say the median. This idea is developed in Charles F. Manski (1985a).
Mean Versus Median Versus δ-Mode Regression. The mean, median, and δ-mode express the location of a probability distribution in equally reasonable but distinct ways. In general, the mean, median, and δ-mode regressions of $y$ on $x$ differ from one another. The mean regression might be linear in $x$ and the median regression nonlinear, or both might be linear but with different slopes. It is even possible that one regression increases with $x$ while the other decreases.1

Asymmetric Absolute Loss Functions. The square, absolute, and step loss functions are all symmetric around zero. Sometimes one wishes to treat under- and overpredictions of $y$ asymmetrically. A leading case is asymmetric absolute loss. Here the loss function is

$$L(u) = \begin{cases} (1 - \alpha)|u| & \text{if } u \leq 0 \\ \alpha|u| & \text{if } u > 0, \end{cases}$$

(5)

where $\alpha$ is a specified constant in the interval $(0, 1)$. This loss function is asymmetric except in the special case $\alpha = .5$. The best predictor is the $\alpha$-quantile of $y$ conditional on $x$, that is, the smallest number $\tau$ such that $\text{Prob}(y \leq \tau|x) \geq \alpha$. The median is the $.5$-quantile. It is interesting to compare the best predictors under asymmetric absolute loss as $\alpha$ varies. As $\alpha$ increases, the loss function penalizes underpredictions of $y$ more heavily and overpredictions less heavily. The result is that the best predictor increases in value. Thus, if $\alpha_1 < \alpha_2$, the $\alpha_2$-quantile regression of $y$ on $x$ lies above the $\alpha_1$-quantile regression. The various quantile regressions need not, however, be parallel to one another or even have slopes of the same sign.

A. Choosing a Loss Function

In a prediction problem, the chosen loss function determines the regression of interest. The chosen loss function should reflect the actual losses associated with mispredictions. Equivalently, the loss function should reflect the way one wants to measure the location of $y$ conditional on $x$. The need for care in choosing a loss function lessens only if one has distributional information implying that distinct best predictors behave similarly.

Two cases stand out.

One may know that, for each value of $x$, the distribution of $y$ conditional on $x$ is symmetric. If so, the best predictors associated with all convex, symmetric loss functions coincide. It follows that, if one has somehow narrowed attention to convex, symmetric loss functions, any choice among these functions yields the same predictions for $y$. Square loss and absolute loss are convex but step loss is not.

One may alternatively know that, as $x$ varies, the distributions of $y$ conditional on $x$ retain the same shape, shifting only in location.4 If so, all best predictor regressions are parallel to one another. It

---

3 To see that the slopes can differ in sign consider the simple case in which $x$ can only take the value zero or one. Then the mean and median regressions of $y$ on $x$ can be written as the linear functions

$$E(y|x) = E(y|x = 0) + (E(y|x = 1) - E(y|x = 0))x$$

and

$$m(y|x) = m(y|x = 0) + (m(y|x = 1) - m(y|x = 0))x.$$  Here $m(y|x)$ denotes the median of $y$ conditional on $x$. The slopes of these two regressions are $E(y|x = 1) - E(y|x = 0)$ and $m(y|x = 1) - m(y|x = 0)$. The four quantities $E(y|x = 1), E(y|x = 0), m(y|x = 1)$, and $m(y|x = 0)$ can take any values. Hence the two slopes need not be equal nor even have the same sign. The two slopes are equal if the distributions of $y$ conditional on $x = 1$ and on $x = 0$ are both symmetric.

4 Equivalently, one may know that $y = g(x) + \varepsilon$, where $g(\cdot)$ is some function of $x$ and $\varepsilon$ is an unobserved random variable known to be distributed independently of $x$. Then a change in $x$ shifts the location of $y$, via $g(x)$, but leaves the shape of the distribution unchanged.

Any restriction on the distribution of $y$ conditional on $x$ can be restated as an equivalent restriction on the unobserved random variable $\varepsilon$. For example, saying that $g(\cdot)$ is the mean regression of $y$ on $x$ is equivalent to saying that $E(\varepsilon|x) = 0$. This article will continue to work directly with the distribution of $y$ conditional on $x$ rather than through restrictions framed in terms of unobserved variables.
follows that, if one wishes to learn the way the location of \( y \) varies with \( x \), it does not matter how one chooses to measure location.

B. Extrapolation

Given a population and a specification of \( y, x, \) and \( L(\cdot) \), suppose that one has determined the best predictor of \( y \) conditional on \( x \) under \( L(\cdot) \). Now alter the prediction problem in some respect. It is natural to ask whether extrapolation is possible. That is, does the known best predictor, or some transformation thereof, solve the revised prediction problem?

We have already encountered one extrapolation question, that raised by a change in the loss function. In general, knowing the best predictor under one loss function does not reveal the best predictor under other loss functions. But knowing the best predictor under one convex symmetric loss function does reveal the best predictor under other convex symmetric loss functions if the distributions of \( y \) conditional on \( x \) are symmetric.

Four additional extrapolation questions are considered here. These are (1) prediction off the support of \( x \); (2) prediction when the distribution of \( x \) changes; (3) prediction of functions of \( y \) conditional on \( x \); and (4) prediction of \( y \) conditional on functions of \( x \).

**Prediction off the Support of \( x \).** Perhaps the most familiar sense of extrapolation is prediction of \( y \) conditional on \( x \) taking a value that is logically possible but does not occur in the observed population. This kind of extrapolation is possible only if one has sufficient prior information on the form of the best predictor. The need for prior information arises because the distribution of \( y \) conditional on \( x \) is not well defined off the support of the distribution of \( x \). (A point \( \xi \) is said to be off the support of \( x \) if there is probability zero that \( x \) falls within a neighborhood of \( \xi \).)

Extrapolation off the support of \( x \) is com-

monly achieved by specifying a model for the best predictor. Suppose that the best predictor has the form \( f(x) \), where \( f(\cdot) \) is a known function defined on the entire logical domain of \( x \). Then \( y \) can be predicted off the support of \( x \), the best predictor at a point \( \xi \) being \( f(\xi) \).

**Prediction Following a Change in the Distribution of \( x \).** A common economic problem is to predict \( y \) following a policy or environmental change that alters the distribution of \( x \) but, for each value of \( x \), leaves unchanged the distribution of \( y \) conditional on \( x \). Regressions of \( y \) on \( x \), being features of the latter conditional distribution, are unaffected by such a change. Hence a best predictor of \( y \) ex ante remains a best predictor ex post.

For example, let \( y \) be household consumption and \( x \) be household after-tax income. Suppose that a change in the tax code takes place. Then the distribution of after-tax incomes is altered. If the tax code change does not affect consumption conditional on after-tax income, the best predictor of \( y \) conditional on \( x \) remains valid.

Using a best predictor to make forecasts following a policy change is inappropriate if the change alters the distribution of \( y \) conditional on \( x \). It is up to the economist to specify what effects a policy change may have and, consequently, whether extrapolation is possible. Concern with extrapolation may legitimately influence one's choice of predictor variables. Let \( x \) and \( z \) be two, possibly overlapping, vectors of predictors. Suppose that a given change in the environment alters the distribution of \( y \) conditional on \( x \) but not the distribution of \( y \) conditional on \( z \). Then one can extrapolate best predictors of \( y \) conditional on \( z \) but not best predictors conditional on \( x \).\(^5\)

\(^5\)A population property that is invariant with respect to a given policy or environmental change is often called a *structural property* with respect to that change. Here regressions of \( y \) on \( z \) are structural but regressions of \( y \) on \( x \) are not.
Prediction of a Function of y Conditional on x. Suppose that one wishes to predict \( g(y) \) conditional on \( x \), where \( g(\cdot) \) is some function of \( y \). It would be convenient if the best predictor of \( g(y) \) under a given loss function should be obtainable from the best predictor of \( y \) under that loss function. Unfortunately, this is so only in special cases. In general, knowing the best predictor of \( y \) does not reveal the best predictor of \( g(y) \).

To see the problem, consider prediction under square loss. The best predictor of \( y \) is \( E(y|x) \) and that of \( g(y) \) is \( E[g(y)|x] \). If \( g(\cdot) \) is linear in \( y \), then \( E[g(y)|x] = g(E(y|x)) \). But if \( g(\cdot) \) is nonlinear, \( E[g(y)|x] \) is not a function of \( E(y|x) \).

Perhaps the most salient case in which one can extrapolate from prediction of \( y \) to prediction of \( g(y) \) occurs when \( g(y) \) is increasing in \( y \) and \( L(\cdot) \) is \( \alpha \)-absolute loss. In that case, the best predictor of \( y \) is the \( \alpha \)-quantile of \( y \) conditional on \( x \), denoted \( q_\alpha([g(y)|x]) \), and that of \( g(y) \) is \( q_\alpha([g(y)|x]) \). It can be shown that, if \( g(\cdot) \) is increasing, then \( q_\alpha([g(y)|x]) = g(q_\alpha([y]|x)) \). See Manski (1988a, proposition 1) and Manski (1988b, section 6.2.6).

Prediction of \( y \) Conditional on a Function of \( x \). Suppose that one wishes to predict \( y \) conditional on \( h(x) \), where \( h(\cdot) \) is some function of \( x \). The feasibility of extrapolation depends on whether \( h(\cdot) \) is one to one or many to one.

If \( h(\cdot) \) is one to one, conditioning on \( x \) is equivalent to conditioning on \( h(x) \). Hence the best predictors conditional on \( x \) and on \( h(x) \) are the same. If \( h(\cdot) \) is many to one, conditioning on \( x \) and on \( h(x) \) are not equivalent. The best predictor of \( y \) conditional on \( h(x) \) is a feature of the distribution of \( y \) conditional on \( h(x) \). This distribution is the expectation of the distribution of \( y \) conditional on \( x \) with respect to the distribution of \( x \) conditional on \( h(x) \). It follows that the best predictor of \( y \) conditional on \( x \), which depends only on the distribution of \( y \) conditional on \( x \), cannot generally reveal the best predictor of \( y \) conditional on \( h(x) \). Knowledge of the distribution of \( x \) conditional on \( h(x) \) is needed as well.

III. Estimation of Best-Predictor Regressions

Suppose that a conditional prediction problem has been specified. Assume that one does not know the distribution of \( y \) conditional on \( x \) but does observe a sample of \( n \) realizations of \( (y, x) \). Then one cannot determine the best predictor but may be able to estimate it. The possibilities depend primarily on the sampling process and on what one knows a priori about the best predictor. In this section we assume random sampling and consider several informational settings.

A. Method-of-Moments Estimation of Parametric Models

Assume that the best predictor has the form \( f(x, b) \), where \( f(\cdot, \cdot) \) is a known function and where \( b \) is an unknown parameter in some parameter space \( B \). Assume that \( b \) is identified. Then estimation of \( b \) yields an estimate for the best predictor.

The method of moments is the prevalent approach. For each value of \( x \), \( b \) solves the problem

\[
\min_{\in B} \mathbb{E}[L(y - f(x, b))|x].
\]  

(6)

It follows that \( b \) minimizes the unconditional expected loss. That is, \( b \) solves

\[
\min_{\in B} \mathbb{E}[L(y - f(x, b))]
\]  

(7)

the expectation now being taken with respect to the joint distribution of \( (y, x) \). If

The parameter \( b \) is identified within the parameter space \( B \) if, for each \( x \in \mathcal{X} \) except \( x = b \), there exists a set \( \mathcal{X}_x \) of \( x \)-values such that \( P(X_x) > 0 \) and, for all \( x \in \mathcal{X}_x \), \( f(x, b) \neq f(x, c) \).
b is identified, it is the only solution to (7).

If the probability distribution of \((y, x)\) is unknown, one cannot solve (7). One can, however, replace expectations with sample averages and solve the sample analog of (7), namely

\[
\min_{b} \frac{1}{N} \sum_{i=1}^{N} L[y_i - f(x_i, c)]. \tag{8}
\]

A solution to (8) is termed a \textit{method-of-moments estimate} for \(b\). The phrase \textit{method of moments} dates back to Karl Pearson who, as early as 1894, advocated replacing population expectations by sample averages as a systematic approach to estimation.\(^7\)

Two familiar estimators emerge when \(L(\cdot)\) is square loss or absolute loss. Then we obtain the least-squares or least-absolute deviations estimators

\[
\min_{b} \frac{1}{N} \sum_{i=1}^{N} (y_i - f(x_i, c))^2 \tag{9}
\]

\[
\min_{b} \frac{1}{N} \sum_{i=1}^{N} |y_i - f(x_i, c)|. \tag{10}
\]

\textbf{Consistency of Method-of-Moments Estimates.} The fundamental results that make method-of-moments estimation work are "uniform laws of large numbers." These theorems show that, if \(L(\cdot)\) and \(f(\cdot, \cdot)\) satisfy weak regularity conditions and if the parameter space \(B\) is not too large, then as \(N\) increases the averages in (8) converge to the expectations in (7), uniformly over all \(c\) in \(B\). Hence the estimate minimizing (8) converges to the point minimizing (7), namely \(b\). In other words, the estimate is consistent.\(^8\)

The condition that the parameter space should not be "too large" is satisfied in the familiar case where \(B\) is finite dimensional and in some cases where \(B\) is not finite dimensional. But the method of moments breaks down if the parameter space is so large that it is always possible to fit the data perfectly. This occurs, for example, in the nonparametric setting to be discussed in Section III.B.

Suppose that, whatever the sample may be, the set of equations \(y_i = f(x_i, c), i = 1, \ldots, N\) always has a solution in \(B\). (The solution may vary with the sample.) Then the minimum average loss in (8) always equals zero even though the minimum expected loss in (7) is positive. As \(N\) increases, the averages in (8) do not converge uniformly to the expectations in (7).

The estimate minimizing (8) does not generally converge to \(b\). For further discussion of the behavior of the method of moments on different parameter spaces, see Manski (1988b, sections 3.3.3 and 4.3.1).

\textbf{Weighted Estimators.} The solution to (8) is "a" method-of-moments estimate, not "the" estimate. The reasoning leading to (8) yields a class of estimators, obtained by weighting the observations differentially. Because \(b\) solves (6) for each value of \(x\), it also solves

\[
\min_{b} \frac{1}{N} \sum_{i=1}^{N} L[y_i - f(x_i, c)]x, \tag{11}
\]

where \(w(\cdot)\) is any positive-valued function of \(x\) chosen by the researcher. Hence \(b\) solves the problem

\[
\min_{c} E[w(x)L[y - f(x, c)]], \tag{12}
\]

\(^7\) Instead of solving the sample analog of (7), one might consider solving the sample analog of (6) for each \(x\). This works if \(x\) is discrete but not if \(x\) is a continuous random variable. In the latter case, the sample almost surely contains \(N\) distinct values of \(x\), each occurring only once. Sample averages conditioning on these \(N\) values are not satisfactory estimates of the relevant population conditional expectations. See Manski (1988b, ch. 3) for further discussion.

\(^8\) Proofs of the asymptotic statistical properties of method-of-moments estimates can be found in such recent econometrics texts as Takeshi Amemiya (1985), A. Ronald Gallant (1987a), and Manski (1988b). The asymptotic properties of least absolute deviations estimation were not studied in depth until the late 1970s. See Roger Koenker and Gilbert Bassett (1978).
whose sample analog is

$$\min_{\alpha, \beta} \frac{1}{N} \sum_{i=1}^{N} w(x_i) L(y_i - f(x_i, \alpha))$$  \hspace{1cm} (13)$$

If the parametric model is correctly specified, the estimate minimizing (13) is consistent for any choice of the weighting function w(·). Different weighting functions, however, may yield estimates that differ numerically and that vary in precision.

**Estimators Implied by Distributional Restrictions.** Application of an estimator of the form (13) presumes that one knows the best predictor up to the parameter b but does not require any other knowledge of the distribution of y conditional on x. Additional method-of-moments estimators are available if one has distributional information implying that distinct best predictors coincide or behave similarly.

For example, suppose it is known that the distribution of y conditional on x is symmetric for all values of x. Then the best predictors derived from all convex, symmetric loss functions coincide. It follows that, if the specified loss function is convex and symmetric, b can be estimated by solving problem (13) for any convex, symmetric loss function. Each such estimate is consistent but the value of the estimate in any given sample and its precision may depend on the loss function used.

Alternatively, suppose it is known that, as x varies, the distributions of y conditional on x retain the same shape, shifting only in location. Then all best predictors have the form $$f(x, b_L) = g(x, \beta) + \alpha_L$$ where $$b_L = (\beta, \alpha_L)$$ and where \(\beta\) does not depend on L. It follows that the \(\beta\) part of \(b_L\) can be estimated by solving (13) for any loss function.

Finally, suppose it is known that the distributions of y conditional on x are members of a parametric family of distributions. Then the maximum likelihood method may be applied to estimate the unknown parameters. This yields an estimate for the conditional distributions and, therefore, for all regressions of y on x. Maximum likelihood is a method-of-moments estimator. See Manski (1988b, ch. 5).

Choosing an Estimator. The foregoing shows that a researcher who has specified a parametric best-predictor model can choose among many consistent method-of-moments estimates. How then should one choose an estimator?

Computation is obviously a consideration. Solution of the minimization problem (13) may be easy or difficult, depending on the loss function and the parametric model specified. The most benign computational problem is least-squares estimation of a linear model, as (13) has a simple, explicit solution. Not quite as simple but still straightforward is least absolute deviations estimation of a linear model; here (13) can be transformed into a linear programming problem (Peter Bloomfield and William L. Steiger 1983). Estimation of nonlinear models often presents more difficulty.

The classical statistical prescription is to select the most precise estimator, with precision measured by the concentration of an estimate's sampling distribution. The exact sampling distributions of method-of-moments estimates are generally compli-

---

9 The most precise estimate of the best predictor is not necessarily the same as the optimal prediction that can be made given the available sample data. Let $$p(x)$$ denote the unknown best predictor and let $$p_N(x)$$ denote an estimate thereof. The precision of $$p_N(x)$$ is defined by reference to its distribution if repeated, independent samples of size N were to be drawn. In contrast, the optimal prediction given the available data minimizes expected loss conditional on the information in the one sample that was actually drawn. Optimal prediction given the available data is not well defined in classic statistics but is well defined in Bayesian statistics. See Thomas Ferguson (1967).

Using an estimate $$p_N(x)$$ to predict y conditional on x makes sense if this estimate is consistent and the sample size N is large. Then $$p_N(x)$$ is likely to be close to the unknown $$p(x)$$ and hence a good predictor of y, in the sense of yielding close to minimal expected loss. But one cannot, in classical statistics, claim that any estimate $$p_N(x)$$ is the optimal prediction of y given the available sample data.
Manski: Regression

41

value \( c^* \) minimizes (12). But \( f(x, c^*) \) is not the best predictor of \( y \) conditional on \( x \).

B. Nonparametric Estimation of Continuous Regressions

Section III.A assumed a parametric model for the regression of interest. In practice, it is usually difficult to justify any model specification. Nevertheless, empirical researchers estimate parametric regression models routinely. How should their work be interpreted?

Most researchers seem to view parametric models not literally but as approximations to a true regression of unknown form. Empirical studies often report estimates for several models, justifying this practice as "sensitivity analysis." Some researchers apply a "model-selection" criterion in an attempt to choose one among several estimated models. Some researchers use "specification tests" to judge the performance of a given model against specified or unspecified alternatives. These activities are undertaken in the hope of finding an adequate approximation to a regression whose true form is unknown.

Attempts to approximate regressions by parametric models are inevitably ad hoc. It is preferable to confront directly the problem of estimating a regression of unknown form. For a long time it was believed that this problem does not have a formal solution but the recent, revolutionary development of methods for nonparametric estimation of regressions proves otherwise. A central achievement has been the development of easily computed "local" estimates usable when the regression of interest is known only to be continuous.\(^\text{11}\)

\(^{10}\) Efficiency must always be defined relative to the available prior information. It may be that one estimator is more precise than another in a given informational setting but that their ranking reverses in a different setting. Indeed an estimator that is more efficient in the presence of extensive prior information may become inconsistent if less information is available. The basic work on the efficiency of method-of-moments estimates is Gary Chamberlain (1987).

\(^{11}\) See B. Lo. S. Prakasa Rao (1933), Herman Bierens (1987), and Wolfgang Hardle (1980) for comprehensive treatments of the material in this section. The literature on nonparametric estimation of continuous regressions is one of several complementary literatures on nonparametric estimation of regressions. A separate body of work provides methods for estimation of regressions whose forms can be ex-
Basic Ideas. To introduce these estimates, let \( \xi \) denote a value of \( x \) and consider the problem of estimating a regression at \( \xi \). Suppose first that \( x \) equals \( \xi \) with positive probability. Then the regression at \( \xi \) is estimable even if no prior information is available. An intuitive procedure is to isolate the subsample of observations \( i \) for which \( x_i = \xi \) and act as if the empirical distribution of \( y \) within this subsample were the population distribution of \( y \) conditional on \( \xi \). For example, one might estimate \( E(y|\xi) \), the mean of \( y \) conditional on \( \xi \), by the sample average of \( y \) across the observations for which \( x_i = \xi \), namely

\[
\sum_{i=1}^{N} y_i 1(x_i = \xi) / \sum_{j=1}^{N} 1(x_j = \xi) \tag{14}
\]

(The indicator function \( 1(\cdot) \) takes the value one if the bracketed logical condition holds and zero otherwise.) Similarly one might estimate the median of \( y \) conditional on \( \xi \) by the sample median of \( y \) across these observations.

Now suppose that \( x \) equals \( \xi \) with probability zero but the support of \( x \) contains \( \xi \); that is, \( x \) is arbitrarily close to \( \xi \) with positive probability. This is the situation when \( x \) has a continuous distribution with positive density at \( \xi \). Estimation based on the subsample where \( x_i \) equals \( \xi \) does not work; with probability one the event \( x_i = \xi \) never occurs in the sample. On the other hand, one can isolate a subsample of observations for which \( x_i \) is "near" \( \xi \) and act as if the empirical distribution of \( y \) within this subsample were the population distribution of \( y \) conditional on \( \xi \). For example, the mean of \( y \) conditional on \( \xi \) can be estimated by a local average; that is, by the sample average of \( y \) across the observations for which \( x_i \) is near \( \xi \). The median conditional on \( \xi \) can be estimated similarly by a local median.

This approach to estimation of the regression at \( \xi \) works provided that (a) the regression is continuous at \( \xi \), (b) one tightens the criterion of nearness as the sample size increases, but (c) one does not tighten the criterion of nearness too rapidly. The need for these provisos can be seen by examining the estimate for the mean of \( y \) conditional on \( \xi \).

To form an estimate, one selects a metric \( \rho(\cdot, \cdot, \cdot) \) to measure the distance between \( \xi \) and a value of \( x \). Any reasonable metric will do; Euclidean distance is most commonly used. One also selects a positive "bandwidth" \( \delta_N \) to be the cutoff distance defining \( x \) to be near \( \xi \). This done, the estimate is

\[
\frac{\sum_{i=1}^{N} y_i 1(\rho(x_i, \xi) < \delta_N)}{\sum_{j=1}^{N} 1(\rho(x_j, \xi) < \delta_N)} \tag{15}
\]

Suppose that \( \delta_N \) is kept fixed at some value \( \delta \). Then as \( N \) increases, the strong law of large numbers implies that the estimate (15) converges to \( E[y|\rho(x, \xi) < \delta] \), the mean of \( y \) conditional on \( x \) being within \( \delta \) of \( \xi \). If \( E(y|x) \) is continuous at \( x = \xi \), then as \( \delta \) approaches zero, \( E(y|\rho(x, \xi) < \delta) \) approaches \( E(y|\xi) \). These two facts suggest that an estimate converging to \( E(y|\xi) \) can be obtained by adopting a bandwidth-selection rule that makes \( \delta_N \) approach zero as \( N \) increases. This works, provided that \( \delta_N \) does not approach zero too quickly. The rate of approach must be slower than \( 1/N^{1/K} \), where \( K \) is the dimension of the vector \( x \). The last proviso is needed to make the number of observations used to calcu-
late the estimate increase adequately with $N$.\footnote{To understand this, assume that $x$ has positive, continuous density at the point $\xi$. Continuity implies that the density of $x$ is approximately constant in a neighborhood of $\xi$. It follows that, for small $b_N$, the probability that a realization of $x$ is within $b_N$ of $\xi$ is approximately proportional to the volume of a neighborhood of radius $b_N$. This volume is proportional to $b_N^d$. Hence the number of observations with $x_i$ within $b_N$ of $\xi$ is approximately proportional to $N b_N^d$. Consistent estimation of the regression at $\xi$ requires that $N b_N^d$ should approach infinity as $N$ increases. Thus $b_N$ must approach zero at a rate slower than $1/N^{d/2}$.
}

Computation and Presentation of the Estimated Regression. I have often heard it asserted that nonparametric estimates are generically more difficult to compute than are parametric estimates. Consideration of the local average estimate (15) should lay this misperception to rest. To compute the estimate, one need only determine the observations for which $x_i$ is within $b_N$ of $\xi$ and then compute the average of $y_i$ for these observations. Computation of a local median is similarly straightforward. The truth is that local average estimates are generally easier to compute than are method-of-moments estimates of nonlinear parametric models. Whereas the former estimates have simple explicit forms, the latter can generally be determined only by numerical solution of the minimization problem (13).

It is true that local average estimates are more difficult to present than are parametric estimates. An estimate of a parametric regression model can be reported succinctly, through the parameter estimate. In contrast, a local estimate of a continuous regression must be reported by evaluating the estimate at particular values of $x_i$.

The usual practice is to estimate the regression at a set of closely spaced $x$-values and to plot the results. If $x$ is scalar, a single graph suffices to present accurately the estimated regression function. If $x$ is bivariate, the shape of the regression can be conveyed by the pseudo-three-dimensional graphing programs included in GAUSS and other statistical software programs. Greater accuracy can be achieved by plotting a series of cross-sections. In each cross-section, one component of $x$ is fixed and the regression is graphed as a function of the other component. If the dimension of $x$ is three or greater, cross-section plotting is a necessity.

Choosing an Estimator. Many local estimates of continuous regressions are available. Many metrics can be chosen to define the distance between $\xi$ and $x$. Many bandwidth-selection rules can be applied to determine how near $x_i$ should be to $\xi$ for observation $i$ to be used in estimating the regression at $\xi$.

Indeed the class of available methods can be enlarged by loosening the sharp line delineating observations used to estimate the regression at $\xi$ from those not used. The logic of local estimation continues to hold if all observations are used, those with $x_i$ near $\xi$ being given the most weight. For example, the mean of $y$ conditional on $\xi$ can be estimated by a local-weighted average

$$\frac{\sum_{i=1}^{N} y_i W_N(x_i - \xi)}{\sum_{i=1}^{N} W_N(x_i - \xi)},$$

(16)

where $W_N(x_i - \xi)$ is a non-negative weight that declines in value as the distance between $x_i$ and $\xi$ increases. In practice, unimodal symmetric probability density functions with mode zero are often used as weighting functions.

The major problem confronting an empirical researcher wishing to estimate a continuous regression is that the available statistical theory offers too little guidance on choosing a criterion for "nearness." Consider, for example, estimate (15) for the mean of $y$ conditional on $\xi$. Many metrics
and rules for choosing the bandwidth yield consistent estimates with limiting normal distributions. The available theory explains how to control the rate at which the bandwidth approaches zero so as to optimize the estimate's rate of convergence. But we know relatively little about how best to select the bandwidth given a sample of fixed size and composition.

The absence of theoretical guidance on setting the bandwidth, and more generally on defining nearness, leaves the empirical researcher with enormous discretion. This discretion gives applied nonparametric regression analysis a subjective flavor. As the selected bandwidth approaches zero, the estimated regression curve fits the data increasingly well but becomes increasingly jagged. As the bandwidth approaches infinity, the estimated curve becomes smoother and flatter. Thus the researcher can, through the choice of bandwidth, predetermine some properties of the estimate.

Attempting to reduce the degree of subjectivity in their analyses, researchers sometimes report multiple estimates computed using different nearness criteria. This practice recalls the sensitivity analyses of researchers working with parametric regression models. Many researchers seek to reduce subjectivity by using data-dependent, automated rules to choose the bandwidth.

The most prominent such rule is “cross-validation.” Here one fixes the bandwidth, estimates the regression on each of the \( N \) possible subsamples of size \( N - 1 \) and, in each case, uses the estimate to predict \( y \) conditional on \( x \) for the left-out observation. The cross-validated bandwidth is that yielding the best predictions of the left-out values of \( y \).

C. Estimation of Dimension-Reducing Models

Method-of-moments estimation of parametric models and nonparametric estimation of continuous regressions invoke very different assumptions. These two approaches to regression analysis have correspondingly distinct attractions.

One can usually be confident that the regression of interest is continuous. Hence one can usually trust nonparametric estimates to be consistent. On the other hand, these estimates are often imprecise in practice. Moreover, they cannot be extrapolated off the support of \( x \). Parametric modeling permits more precise estimation and makes extrapolation possible. The problem, of course, is that an assumed parametric model may be misspecified.

Recent work by statisticians and econometricians seeks to provide the empirical researcher with intermediate options, estimation approaches that assume more than continuity but less than a parametric model. Most notable has been the development of dimension-reducing approaches.

There is consensus among practitioners that nonparametric estimation of continuous regressions works well when the regressor vector \( x \) has low dimension. It is well known, however, that precision falls sharply as the dimension of \( x \) rises. The difficulty of estimating high-dimensional, continuous regressions has motivated interest in models that impose dimension-

---

\( \text{13 The rate of convergence depends on various factors but is necessarily slower than the } 1/\sqrt{N} \text{ rate attainable when the regression is known to belong to a finite-dimensional family of functions. This result applies to the problem of estimating the regression of } y \text{ on } x \text{ at a point } \xi. \text{ In some applications, one is interested in estimating an average across } x \text{ of some function of the regression. Recent work establishes that } 1/\sqrt{N} \text{ rate estimation of such averages is often possible. See, for example, James Stock (1989) and James Powell, James Stock, and Thomas Stoker (1989).} \)

\( \text{14 This phenomenon has a simple explanation. It was pointed out in Footnote 12 that, for small } \delta, \text{ the number of observations with } x_i \text{ within } \delta \text{ of } \xi \text{ is approximately proportional to } N^{-\delta^2}. \text{ This quantity is exponential in } K. \text{ Thus, for } \delta < 1, \text{ the number of observations with } x_i \text{ within } \delta \text{ of } \xi \text{ falls exponentially as } K \text{ rises.} \)
Manski: Regression

reducing restrictions but remain nonparametric or semiparametric. Three of the
two possibilities will be mentioned here. These are the "additive"
model, the "index" model, and the "semi-
additive" model.

The additive model assumes that the re-
gression has the form \( f_1(x^1) + f_2(x^2) + \ldots + f_K(x^K) \). Here \((x^1, x^2, \ldots, x^K)\)
are the components of \(x\) and \(f_1(\cdot), f_2(\cdot), \ldots, f_K(\cdot)\) are unknown functions of these
components. This model reduces the prob-
lem of estimating a regression function with
domain \(\mathbb{R}^K\) to the simpler problem of esti-
mating \(K\) functions, each with domain the
real line.\(^{16}\)

The index model assumes that the re-
gression has the form \( f[g(x, \beta)] \). Here \(\beta\)
is an unknown finite-dimensional parameter,
\(g(x, \beta)\) is a known function mapping
\((x, \beta)\) into the real line, and \(f(\cdot)\) is an
unknown function of the "index" \(g(x, \beta)\). This
model reduces the original nonparametric esti-
model reduces the estimation problem to
one of estimating a parameter \(\beta\) and a func-
tion \(f(\cdot)\). A method for estimating the semi-
additive model has been studied by Peter

IV. Estimation from Nonrandom Samples

Random sampling is a central idealiz-
ation in statistics, much as competitive be-
havior is in economics. Under random sam-
ping, the process generating observations of
\((y, x)\) is the same as the process defining
the population.

Actual sampling processes are often non-
random in some respect. A particular non-
random sampling process may be informa-
tive about some features of the population
but not about others. The central problem
in the study of estimation under nonran-
dom sampling is to determine what a given
sampling process reveals about the feature
of interest. In some cases a nonrandom sam-
ping process identifies the feature of
interest. Then estimation is usually
straightforward. In some cases the sam-
ping process does not identify the feature
of interest but reveals something about it.
Then estimation may be possible if one can
combine the sample data with suitable
prior information.

The available theory on the estimation
of regressions from nonrandom samples is
vast but fragmented. The brief discussion
here serves mainly to distinguish sampling
processes that identify regressions from
ones that are informative but not identify-
ing.

A. Regression-Preserving Sampling

We may call a sampling process regres-
sion-preserving if, for each observation \(i\),
the sampling distribution of \(y_i\) conditional
on \(x_i\) is the same as the population dis-
tribution of \(y_i\) conditional on \(x_i\). As a rule, re-
gression-preserving sampling identifies re-
gressions. Moreover, the estimation ap-
proaches discussed in Section III.B can be
applied without modification. Two leading cases follow.

Stationary Sampling. Suppose that the data \((y_i, x_i), i = 1, \ldots, N\) are drawn from a stationary stochastic process whose marginal distribution is \(P\). The term stationary means that, at each date \(i\), \(P\) is the marginal sampling distribution of \((y_i, x_i)\); that is, the distribution not conditioning on the realizations of \((y, x)\) at dates other than \(i\).

Stationary sampling does not provide a random sample from \(P\) except in the special case where the observations are serially independent. Nevertheless, dependent sampling versions of the law of large numbers show that, if the serial dependence is not too severe, the empirical distribution of the data converges to \(P\) as the sample size increases. This fulfills the basic requirement for successful estimation. The methods discussed in Section III continue to work. Lars Hansen (1982) studies method-of-moments estimation under stationary sampling. Bierens (1987) treats nonparametric estimation of continuous mean regressions.

Regressor-stratified Sampling. Suppose that, at each sampling point \(i\), a regressor value \(x_i\) is chosen by the researcher and then \(y_i\) is drawn at random conditional on \(x_i\). This sampling process is applied routinely in controlled experiments. More generally, suppose that a set \(X\) of regressor values is chosen by the researcher and then \((y_i, x_i)\) is drawn at random conditional on the event that \(x_i\) is in \(X_i\). This regressor-stratified sampling process is regression preserving. The estimation methods of Section III work.\(^{17}\)

\(^{17}\)The precision with which a regression may be estimated depends on the chosen strata \(X_i, i = 1, \ldots, N\) and on the available prior information. A large literature studies optimal stratification in the context of the standard linear model. Unfortunately, the lessons of this literature do not extend to other settings. In general, the precision attainable under alternative stratifications depends on how the conditional distribution of \(y\) varies with \(x\). In practice, these conditional distributions are not known. (If they

B. Some Informative but Not Identifying Sampling Processes

Many practically important sampling processes are informative about regressions but do not identify them. Little was known about such processes before the mid-1970s. Since then econometricians have learned much about the prior information necessary and sufficient to achieve identification and permit estimation. Three leading cases are described here.

Response-based Sampling. Discussion of regressor-stratified sampling leads one to consider the complementary sampling process wherein, at each sampling point \(i\), a stratum \(Y_i\) of \(y\)-values is chosen by the researcher and then \((y_i, x_i)\) is drawn at random conditional on the event that \(y_i\) is in \(Y_i\). This is variously referred to as response-based, choice-based, retrospective, or case-control sampling.

Response-based sampling is not regression preserving.\(^{18}\) This sampling process

were, the regression of interest would be known and the estimation problem would disappear. Hence classical statistics offers no general solution to the problem of selecting the best stratification.

\(^{18}\)To see this, consider the case in which \(x\) has a probability density \(p(x)\) and \(y\) is a binary random variable, taking the values 0 and 1. Let \(P(y = 1|x)\) denote the population probability that \(y = 1\), conditional on \(x\). Let \(Q(y = 1)\) denote the marginal probability that \(y = 1\). Let \(q(x|y)\) denote the density of \(x\) conditional on \(y\).

Assume that, in observations \(i = 1, \ldots, N\), the value \(y_i = 0\) is chosen by the researcher and then \(x\) is chosen at random from \(x_i\) on the event. In the remaining \(N_0 = N - N_1\) cases, \(y_i = 1\) is chosen and \(x\) is drawn conditional on this. Thus \(N_0\) realizations of \(x\) are drawn according to the density \(q(x|y = 0)\) and \(N_1\) according to the density \(q(x|y = 1)\). Let \(H_0 = N_0/N\) and \(H_1 = N_1/N\) be the sampling frequencies for the two values of \(y\).

The sampling probability that \(y = 1\), conditional on \(x\), can be shown to be

\[
P(y = 1|x, H_1, Q(y = 1))\]

\[
P(y = 1|x, H_1, Q(y = 1)) + P(y = 0|x, H_1, Q(y = 0))
\]

See Manski and Daniel McFadden (1981). Thus the sampling probability that \(y = 1\), conditional on \(x\), differs from the population conditional probability \(P(y = 1|x)\) except in the special case where \(H_1 = Q(y = 1)\).
yields information about regressions of $y$ on $x$ but is not identifying. Suppose that the union of the strata $Y_i$, $i = 1, \ldots, N$ is the complete support of $y$. If so, combining the sampling process with easily available prior information achieves identification. In particular, it is enough to know the marginal population probabilities that $y$ falls in each of the strata.\(^\text{19}\)

Considerably stronger prior information is needed if the union of the strata is a proper subset of $y$'s support. A leading case is "truncated sampling." Here $y$ is a continuous random variable with support the real line but the observations are drawn at random conditional on the event $y_i > \eta$, where $\eta$ is a given constant.\(^\text{20}\)

**Censored Sampling.** A censored-sampling process is one in which realizations of $(y, z)$ are incompletely observed. The censored-sampling process that has received the most attention by econometricians assumes that each member of the population is characterized by $(y, x)$ and by a binary indicator $z$. A random sample of $N$ realizations of $(y, z, x)$ is drawn. In each case $i$, the researcher observes $(z_i, x_i, y_i)$ but observes $y_i$ only if $z_i = 1$. Thus the realizations of $y$ are censored, $z$ being the censoring variable.\(^\text{21}\)

In the absence of prior information, this censored-sampling process does not identify regressions of $y$ on $x$.\(^\text{22}\) Identification can be achieved given prior information on the censoring process. The simplest case assumes that, conditional on $x$, $y$ is statistically independent of $z$. Then the censored-sampling distribution of $y$, which is the distribution of $y$ conditional on $x$ and on the event $z = 1$, is the same as the population distribution of $y$ conditional on $x$. Hence the censored-sampling process is regression preserving. The estimation methods of Section III apply directly.

The plausibility of the foregoing assumption in economic applications has been questioned sharply. The focus of much econometric work has been to derive identifying restrictions from parametric latent-variable models of the censoring process.\(^\text{23}\)

\(^\text{19}\) The literature emphasizes the case in which $y$ is a discrete random variable but this is not essential (see Manski and Steven R. Lerman 1977; Manski and McFadden 1981; and David A. Hsieh, Manski, and McFadden 1985). These articles develop various method-of-moments estimators, the computation of which is of the same order of difficulty as maximum likelihood estimation under random sampling. The weighted maximum likelihood estimator introduced in Manski and Lerman (1977) is particularly easy to compute and has been implemented in software packages such as LIMDEP.

\(^\text{20}\) That is, one samples only from the $y$-stratum ($n, \infty$). All of the literature on truncated sampling assumes parametric regression models. Jerry Hausman and David A. Wise (1977) study maximum likelihood estimation of linear models under the assumption that the distribution of $y$ conditional on $x$ is normal. Powell (1986b) develops method-of-moments estimates that may be applied when the distribution of $y$ conditional on $x$ is known to be symmetric for each value of $x$. Myoung-jai Lee (1989) develops method-of-moments estimates for 0-mode regressions. Whereas the first estimator has often been used in practice, the two more recent contributions have yet to be applied.

\(^\text{21}\) The study of this censoring process has been motivated largely by applications to the analysis of labor supply. There a typical problem is to estimate a wage regression when wages are observed only for those people who work. Thus $y$ is wage and $z$ indicates whether a person works or not. See Reuben Gronau (1974) for an early, lucid discussion of the problem.

\(^\text{22}\) Consider the case of mean regression. The objective is to learn

$$E(y|x) = E(y|x, z = 1)P(z = 1|x) + E(y|x, z = 0)P(z = 0|x).$$

The sampling process identifies the censored regression $E(y|x, z = 1)$ and the censoring probabilities $P(z|x)$. The sampling process does not identify $E(y|x, z = 0)$. Hence it does not identify $E(y|x)$. The identification problem is examined in depth in Manski (1989).

\(^\text{23}\) See James Heckman (1976) and G. S. Maddala (1983). The usual latent-variable model assumes that $y = f_i(x) + u_i$; that $E(u_i|x) = 0$, and that $z = 1[f_i(x) + u_2 > 0]$, where $[f_i^*, f_i^+]$ are functions of $x$ and $(u_1, u_2)$ are unobserved random variables. It follows that $E[y|x] = f_i[x]$ and $E[y|x, z = 1] = f_i(x) + E[u_1|x, f_i(x) + u_2 > 0]$. The sampling process identifies the censored regression $E[y|x, z = 1]$. The uncensored regression $E[y|x]$ is identified if sufficient prior information is available on the form of $[f_i^*, f_i^+]$ and on the distribution of $(u_1, u_2)$ conditional on $x$. Such information enables one to decompose $E[y|x]$.}
There has been substantial study of the self-censoring process wherein \( z = 1 \) if \( y > \eta \); here \( \eta \) is a given constant. Recently, it has been shown that the distribution function of \( y \) conditional on \( x \) can be bounded even if nothing is known about the censoring process (see Manski, 1989).

Contaminated Sampling. Throughout this article we have maintained the assumption that the data are from the population of interest. In practice, the sampling process may be contaminated. Among the many possible forms of contamination, two have drawn particular attention. These are "errors in variables" and "gross errors."

Errors in variables occur if the data are not realizations of \((y, x)\) but rather of \((y + \varepsilon, x + \delta)\), where \((\varepsilon, \delta)\) are random variables distributed independently of \((y, x)\). This kind of contamination is usually motivated by reference to inaccuracy in the instrument used to measure \((y, x)\). The study of errors in variables has a long history, most of it set in the context of random sampling and linear models for mean regression. Random sample data with errors in variables identify the mean regression of \( y + \varepsilon \) on \( x + \delta \) but do not identify the mean regression of \( y \) on \( x \). The latter regression is identified if sufficient information about the distribution of \((\varepsilon, \delta)\) is available.

Error \( z = 1 \) uniquely into its two parts, \( f_1(x) \) and \( E[z|x, f_2(x)] + v_2 > 0 \).

James Tobin (1958) and Amemiya (1973) study maximum likelihood estimation of linear models under the assumption that the distribution of \( y \) conditional on \( x \) is normal. Powell (1986a) develops method-of-moments estimates for quantile regressions. The maximum likelihood approach has been applied widely. Powell's censored least-absolute deviations estimator does not require one to impose the strong distributional assumptions needed for maximum likelihood estimation to be consistent. It is also relatively easy to compute. Nevertheless, this appealing method has not yet received from applied economists the attention it warrants.

The bound is extremely easy to compute. It suffices to compute local average estimates of two continuous regressions.

Identification can be achieved through knowledge of the "signal-to-noise" ratio, that is the ratio of the variances of \( x \) and \( \delta \). Steven Klepper and Ed-
Recently, regression theory has advanced extremely rapidly. Inevitably, it takes time for applied researchers to determine which advances are of practical importance and how to exploit them.

The present gap between theory and practice is, however, not entirely inevitable. A contributing factor, I believe, is the increasing isolation of econometric theorists from applied work. In the past, advances in econometrics were usually motivated by a desire to answer specific empirical questions. This symbiosis of theory and practice is less common today.

It has become typical for a theorist to set out an abstract estimation problem, propose a method solving that problem, work out the method's statistical properties, perform a modest Monte Carlo study, and then go on to the next theoretical problem. Econometric theorists today do not usually consider it their responsibility to apply their contributions nor to develop computer programs that would make their advances usable by others. I believe that the distancing of methodological research from its applied roots is unhealthy, both for empirical economics and for econometrics.

REFERENCES


