Available online at www.sciencedirect.com
SCIENCE DIRECT
JOURNAL OF Econometrics
Journal of Econometrics 136 (2007) 189–211
www.elsevier.com/locate/jeconom

A simple ordered data estimator for inverse density weighted expectations

Arthur Lewbel[a,*], Susanne M. Schennach[b]

[a]*Department of Economics, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467, USA*
[b]*University of Chicago, USA*

Accepted 19 August 2005
Available online 19 October 2005

Abstract

We consider estimation of means of functions that are scaled by an unknown density, or equivalently, integrals of conditional expectations. The "ordered data" estimator we provide is root n consistent, asymptotically normal, and is numerically extremely simple, involving little more than ordering the data and summing the results. No sample-size-dependent smoothing is required. A similarly simple estimator is provided for the limiting variance. The proofs include new limiting distribution results for functions of nearest-neighbor spacings. Potential applications include endogenous binary choice, willingness to pay, selection, and treatment models.

JEL classification: C14; C21; C25; J24

Keywords: Semiparametric; Conditional expectation; Density estimation; Binary choice; Binomial response

1. Introduction

Given an iid dataset $(x_1, y_1), \ldots, (x_n, y_n)$, on a bivariate random variable (x, y), we propose to estimate an expectation of the inverse density weighted form

$$\theta = \mathrm{E}\left[\frac{y}{f(x)}\right], \tag{1}$$

*Corresponding author. Tel.: +1 617 552 3678; fax: +1 617 552 2308.
E-mail address: lewbel@bc.edu (A. Lewbel).
URL: http://www2.bc.edu/~lewbel/.

doi:10.1016/j.jeconom.2005.08.005

where $f(x)$ is the unknown density of the continuously distributed x. The ''ordered data'' estimator we provide possesses the rather surprising property of achieving root n consistency and asymptotic normality without requiring sample-size-dependent smoothing. It also offers the advantage of being numerically extremely simple, requiring little more than ordering the data and summing the results, thereby avoiding issues regarding the selection of smoothers such as bandwidths, kernels, etc. A similarly simple estimator is provided for the limiting variance.

Inverse density weighted estimation applies generically to the estimation of definite integrals of conditional expectations. Suppose

$$\theta = \int_{x \in \mathcal{X}} \mathrm{E}(w|x)\,\mathrm{d}x \tag{2}$$

for some random variable w, where $\mathcal{X} \subset \mathrm{supp}(x)$. Then θ can be rewritten as $\theta = \mathrm{E}[y/f(x)]$ with $y = wI(x \in \mathcal{X})$, and where I is the indicator function that equals one if its argument is true and zero otherwise.

A number of existing semiparametric estimators make use of inverse density weighted expectations either directly or indirectly, via their relationship with integrated conditional expectations. Examples include density weighted least squares (Newey and Ruud, 1994), average derivative estimation (Härdle and Stoker, 1989), estimators of willingness-to-pay models and general estimators of moments from binomial data (Lewbel, 1997; McFadden, 1999; Lewbel et al., 2005), semiparametric estimators of consumer surplus (Hausman and Newey, 1995; Newey, 1997), some discrete choice, sample selection, and other latent variable model estimators (Lewbel, 1998, 2000, 2002), entropy measures of dependence (Hong and White, 2005) and semiparametric functional tests (Hall and Yatchew, 2005).

Let $(y_{[i]}, x_{[i]})$ denote the ith observation when the data are sorted in increasing order of x, so $x_{[i]}$ is the ith order statistic and $y_{[i]}$ is the concomitant statistic to $x_{[i]}$. Let (y_i, x_i) denote the ith observation when the data are left unsorted. Let $F(x)$ denote the unknown distribution function of x. We show that the numerically trivial ''ordered data'' estimator

$$\hat{\theta} = \sum_{i=1}^{n-1} (y_{[i+1]} + y_{[i]})(x_{[i+1]} - x_{[i]})/2 \tag{3}$$

is root n consistent and asymptotically normal. Specifically, $n^{1/2}(\widehat{\theta} - \theta) \xrightarrow{\mathrm{d}} \mathrm{N}(0, 3\sigma^2/2)$, where

$$\sigma^2 = \mathrm{E}\left[\frac{\mathrm{Var}(y|x)}{f^2(x)}\right] \tag{4}$$

and a consistent estimator of σ^2 is the simple expression

$$\hat{\sigma}^2 = \frac{n}{4} \sum_{i=1}^{n-1} (y_{[i+1]} - y_{[i]})^2 (x_{[i+1]} - x_{[i]})^2. \tag{5}$$

For some intuition for this estimator, let x be a point that lies between $x_{[i]}$ and $x_{[i+1]}$ for some i, and let y be a corresponding point that lies between $y_{[i]}$ and $y_{[i+1]}$. Then $y \approx (y_{[i+1]} + y_{[i]})/2$ and

$$\frac{1}{f(x)} = \left(\frac{\mathrm{d}F(x)}{\mathrm{d}x}\right)^{-1} \approx \frac{x_{[i+1]} - x_{[i]}}{F(x_{[i+1]}) - F(x_{[i]})} \approx \frac{x_{[i+1]} - x_{[i]}}{1/n}, \tag{6}$$