*Nonparametric and Semiparametric estimators*

1. Introduction
2. Nonparametric estimation;
3. Semiparametric estimators;

# Introduction

- A *parametric* estimator is used to estimate *finite-dimensional parameters* of a model where these parameter values uniquely identify the conditional probability distribution of the dependent variables (see chapter on Maximum Likelihood Estimation).
- A *nonparametric* estimator is used to estimate features of a model without making any assumptions on the population distribution. It is often (but not always) used to estimate **functions**, such as a probability density function or a conditional mean without assuming strong conditions such as normality. These are denoted in the literature *infinite dimensional parameters*.
- A *semiparametric* estimator is one used to estimate finite-dimensional parameters but these do not uniquely identify the conditional probability distribution of the dependent variables.
- Some authors define also *seminonparametric estimators* for the case in which a **semiparametric estimator** requires some form of nonparametric estimation.

**Remark:**

- Note that the same estimator can be seen as parametric, nonparametric and semiparametric depending on the context. Consider the linear regression model

$$y = x'\beta_0 + u.$$

and the Ordinary Least Squares Estimator (OLS):

  - OLS can be considered a parametric estimator if on assumes $u|x \sim N(0, \sigma^2)$ and obtain this estimator via Maximum Likelihood.
  - OLS can be considered a semiparametric estimator if we only assume that $E(u|x) = 0$ and consequently $E[y|x] = x'\beta_0$. Note that we do not assume normality of the error term, but assume a specific form for the mean of $y$ conditional on $x$.
  - It can also be seen as nonparametric if our objective is to estimate the best linear predictor coefficients. That is, we want to estimate the value of $\beta$ that minimizes $E\left[(y - x'\beta)^2\right]$. This is given by $\beta_0 = E[xx']^{-1} E[xy]$.

- In this chapter we focus on a particular type of nonparametric estimators denoted as *kernel estimators* (other types of nonparametric estimators are possible; eg. series estimators).
- Suppose $\{y_i\}_{i=1}^{n}$ are iid, where $y_i$ is a random variable with a continuous distribution.
- Our objective is to estimate the density function of $y, f\left(y\right)$.
- A crude estimator of the density of $y$ is the histogram.
- The shape of the histogram depends on the **bin width**, on the **number of bins** and on the chosen **origin**.
- Fixing these parameters, the density of $y$ at $y = y_0$ given by the histogram is

$$\hat{f}\left(y_0\right) = \frac{1}{n}\frac{\text{\# of observations in the same bin as } y_0}{\text{width of the bin containing } y_0}$$

- If we define $h$ as the bin width and let $y^*$ denote the midpoint of the interval containing $y_0$, the histogram estimate is given by

$$\hat{f}(y_0) = \frac{1}{nh} \sum_{j=1}^{n} \mathbf{1}\left( -\frac{1}{2} \leq \frac{y_j - y^*}{h} < \frac{1}{2} \right)$$

  where $\mathbf{1}(A) = 1$ if $A$ is true and $\mathbf{1}(A) = 0$ if $A$ is false
- This estimator is not continuous in $y_0$.

Another estimator of the density function can be motivated as follows.

- From the definition of a probability density, we have that

$$\begin{aligned} f(y_0) &= \lim_{h\to 0} \frac{F(y_0+h/2)-F(y_0-h/2)}{h} \\ &= \lim_{h\to 0} \frac{1}{h} \Pr(y_0-h/2<y<y_0+h/2) \\ &= \lim_{h\to 0} \frac{1}{h} \mathrm{E}\left[\mathbf{1}\left(y_0-\frac{h}{2}\le y<y_0+\frac{h}{2}\right)\right] \end{aligned}$$

  where $F(y)$ is the cumulative distribution function of $y$.

- Hence the estimator of $f(y_0)$ given by the analogy principle is the *naïve* density estimator:

$$\begin{aligned} \hat{f}(y_0) &= \frac{1}{nh}\sum_{j=1}^{n}\mathbf{1}\left(y_0-\frac{h}{2}\le y_j<y_0+\frac{h}{2}\right) \\ &= \frac{1}{nh}\sum_{j=1}^{n}\mathbf{1}\left(-\frac{1}{2}\le\frac{y_j-y_0}{h}<\frac{1}{2}\right) \end{aligned}$$

# Nonparametric Estimation

- It is quite easy to prove that this estimator is *consistent* in iid samples provided that $n \to \infty$, $nh \to \infty$ and $h \to 0$.
- Note that

$$E\left[\hat{f}\left(y_0\right)\right] = \frac{F\left(y_0 + \frac{h}{2}\right) - F\left(y_0 - \frac{h}{2}\right)}{h} \to f\left(y_0\right)$$

  as $h \to 0$
- Also

$$Var\left(\hat{f}\left(y_0\right)\right) =$$

$$= \frac{1}{nh} \frac{F\left(y_0 + \frac{h}{2}\right) - F\left(y_0 - \frac{h}{2}\right)}{h} \left[1 - \left(F\left(y_0 + \frac{h}{2}\right) - F\left(y_0 - \frac{h}{2}\right)\right)\right]$$

$$\leq \frac{1}{nh} \frac{F\left(y_0 + \frac{h}{2}\right) - F\left(y_0 - \frac{h}{2}\right)}{h} \to 0$$

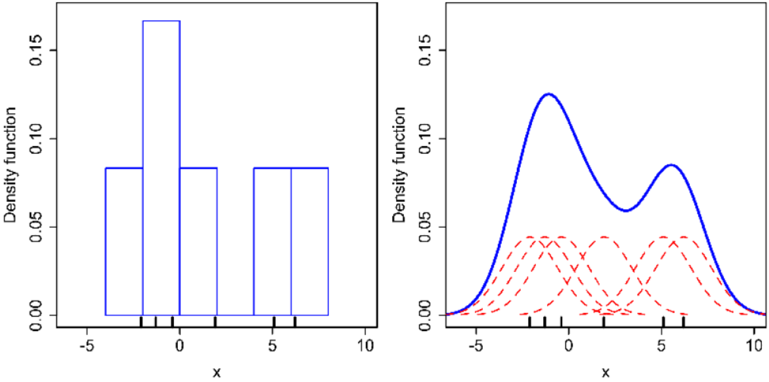
  provided that $nh \to \infty$.
- Consequently $\operatorname{plim} \hat{f}\left(y_0\right) = f\left(y_0\right)$.
- Like the histogram, this estimator is **not continuous in** $y_0$.

- A continuous and smooth estimator can be obtained by replacing the function $K[z] = \mathbf{1}(-1/2 \leq z < 1/2)$ by another *kernel* function. A *kernel* function is any function $K(\cdot)$ such that $\int_{\mathbb{R}} K(u)\, du = 1$ (Rosenblatt, 1956).
- Note that we do not require $K(\cdot) \geq 0$ (although this assumption ensures that the estimator is non-negative).
- Many kernels are available, but a popular choice is to use $K[z] = \phi(z)$ (the density function of the standard normal distribution).
- In general, for a bandwidth $h > 0$, the kernel density estimator is given by

$$\hat{f}(y_0) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{y_j - y_0}{h}\right).$$

- The choice of the kernel affects the bias $E\left[\hat{f}(y_0)\right] - f(y_0)$.
- As before consistency of this estimator requires $n \to \infty$, $nh \to \infty$ and $h \to 0$

# Nonparametric Estimation



- This can be generalized to $d$-dimensional data using a multivariate kernel function $K_d(\cdot)$ such that $\int_{\mathbb{R}^d} K_d(u)du = 1$.
- **Example:** $K_d(u) = \prod_{j=1}^{d} \phi\left(u_j\right)$, where $u = (u_1, ..., u_d)'$.
- To ensure **consistency**, among other things, we need to assume
  - $n \longrightarrow \infty$;
  - $h \longrightarrow 0$;
  - $nh^d \longrightarrow \infty$.
- In this case the variance of the estimator goes to zero as $nh^d \longrightarrow \infty$. Notice the "**curse of dimensionality**": the larger the value of $d$, the slower the rate of convergence.
- Under additional conditions it is also possible to ensure asymptotic *normality*.
- However, convergence is slower than rate $n^{-1/2}$ and the estimator is biased.
- The rate of convergence depends on $K_d(\cdot)$ but according to some authors this choice is relatively unimportant.
- The *bandwidth* $h$ affects the bias and variance of the estimator: as $h$ decreases, so does the bias; however, the variance increases.

# Nonparametric Estimation

- The choice of bandwidth is critical for the results.
- *Too large* a bandwidth, means over-smoothing and eliminates important features of the function being estimated.
- *Too small* a bandwidth, means under-smoothing and the important features of the function become obscured by noise.
- Choices usually implemented by:
  - Automatic rules;
  - Graphically analysing the effects of varying the bandwidth;
  - Minimizing selection criterion.
- Silverman's rule-of-thumb, for the Gaussian kernel, is a popular choice

$$h = \min\left\{\sigma, \frac{IQR}{1.34}\right\} \frac{0.9}{n^{1/5}}.$$

# Nonparametric Estimation

- The kernel density estimator can be used for *Kernel Regression*.
- Particularly in the **exploratory analysis** of the data, it may be convenient not to assume a functional form for the relation between two variables.
- Let $\{(y_i, x_i)\}_{i=1}^{n}$ be iid, where $y_i$ and $x_i$ are random variables with continuous distributions and suppose we want to estimate $\mathrm{E}(y_i|x_i) = m(x_i)$.
- Write this as

$$y_i = m(x_i) + \epsilon_i$$

where $\mathrm{E}(\epsilon_i|x_i) = 0$.

- If we were interested in estimating $m\left(x_j\right)$ at a particular value of $x$ and **if multiple observations** of $y$ were available for that particular value of $x$, we could use

$$\begin{aligned}\hat{m}\left(x_0\right) &= \frac{\sum_{j=1}^{n} \mathbf{1}\left(x_j = x_0\right) y_j}{\sum_{j=1}^{n} \mathbf{1}\left(x_j = x_0\right)} = \sum_{j=1}^{n} \frac{\mathbf{1}\left(x_j = x_0\right)}{\sum_{j=1}^{n} \mathbf{1}\left(x_j = x_0\right)} y_j \\ &= m\left(x_0\right) + \frac{\sum_{j=1}^{n} \mathbf{1}\left(x_j = x_0\right) \epsilon_j}{\sum_{j=1}^{n} \mathbf{1}\left(x_j = x_0\right)}\end{aligned}$$

  which, under mild assumptions, is consistent if $n \to \infty$, provided that $\Pr\left(x_j = x_0\right) > 0$.
- Consequently this estimator is nor valid if $x_j$ is a continuous random variable.

# Nonparametric Estimation

- For continuos random variables if $m(\cdot)$ is sufficiently smooth, we can estimate $m(x_0)$ using observations in a **neighbourhood** of $x_0$.
- A **smoothing estimator** of $m(x_0)$ can be written as

$$\hat{m}(x_0) = \sum_{j=1}^{n} \omega_h (x_0 - x_j)\, y_j$$

where the weights $\omega_h(x_0 - x_j)$ decrease with the distance between $x_j$ and $x_0$.

- Frequently, the weights are defined as

$$\omega_h (x_0 - x_j) = \frac{K\left(\frac{x_0 - x_j}{h}\right)}{\sum_{k=1}^{n} K\left(\frac{x_0 - x_k}{h}\right)}.$$

- This defines the **Nadaraya-Watson** estimator of $m(x_0)$.
- Under appropriate regularity conditions, $\hat{m}(x_0)$ is consistent when $\lim_{n\to\infty} h = 0$, but the **convergence is slow** (not $n^{1/2}$).

- Again, the properties of the estimator depend on the choice of $K(\cdot)$ and $h$.
  - According to some authors, the choice of $K(\cdot)$ is not that important, and often $K(\cdot) = \phi(\cdot)$.
  - The choice of $h$ is critical: If $h$ is too small, the estimates become too irregular; If $h$ is too large, the estimates approach $\bar{y}$.
- A popular method to choose $h$ is to use leave-one-out cross-validation, i.e.,

$$h_{CV} = \arg\min_{h} \frac{1}{n} \sum_{i=1}^{n} \delta(x_i) \left[ y_i - \textstyle\sum_{j \neq i} \omega_h (x_i - x_j) y_j \right]^2$$

where $\delta(x)$ is a trimming function to reduce boundary effects.

- This can be generalized for the case where we condition on a **vector of regressors**.
- However, for the multivariate case, not only do we have to face the "*Curse of Dimensionality*", but we also lose the ability to display the results *graphically*.

# Semiparametric estimation

- Most non-ML estimators are semiparametric in the sense that they do not require the full specification of the conditional probability distribution of the dependent variable.
- Later we will see examples of semiparametric estimators in the context of some specific microeconometric models.
- Here, we will look at a useful *semi(non)parametric* estimator: The partially linear model.
- Suppose that $\mathrm{E}(y_i|z,x) = z_i\beta + g(x_i)$, where $g(\cdot)$ is left unspecified, and

$$y_i = z_i\beta + g(x_i) + \epsilon_i, i = 1, \ldots, n.$$

- Robinson (1988) and Speckman (1988) have shown that, under certain conditions, it is possible to obtain $\sqrt{n}$-consistent estimators for $\beta$.
- Notice that we can write

$$\begin{aligned} \mathrm{E}(y_i|x) &= \mathrm{E}(z_i|x)\beta + g(x_i) \\ y_i - \mathrm{E}(y_i|x) &= [z_i - \mathrm{E}(z_i|x)]\,\beta + \epsilon \end{aligned}$$

# Semiparametric estimation

- If $\mathrm{E}(y_i|x)$ and $\mathrm{E}(z_i|x)$ where known, $\beta$ could be estimated by *OLS*.
- However, if $\mathrm{E}(y_i|x)$ and $\mathrm{E}(z_i|x)$ are replaced by *kernel estimators* that converge sufficiently quickly, their use will *not affect* the rate of convergence of the OLS estimator of $\beta$, which, using an obvious notation, can be estimated by (see Yatchew, 2003)

$$\hat{\beta} = \left[\left(Z - \widehat{\mathrm{E}(Z|x)}\right)' \left(Z - \widehat{\mathrm{E}(Z|x)}\right)\right]^{-1} \left(Z - \widehat{\mathrm{E}(Z|x)}\right)' \left(Y - \widehat{\mathrm{E}(Y|x)}\right),$$

- This estimator is consistent and asymptotically normal distributed.