Final Exam Format

- Two hours (10 minutes reading time).
- Closed book, approved calculators permitted.
- Four questions, each worth 20-30 marks.
- A mixture of analytical and practical (approx 50:50), sometimes both within the same question.
- Analytical component will be in the same spirit as the assignments.
- Practical component will involve describing how one implements specific methods and interpreting STATA output.
  - Knowledge of STATA code/syntax will not be examined.

ECON6300/7320/8300
Advanced Microeconometrics
Nonparametric Methods

Christiern Rose [1]

[1]University of Queensland

Outline:

- Motivation
- Nonparametric Density Estimation

- Nonparametric Regression
- Semiparametric Regression

Motivation:

- The underlying decision problem is to choose

$$m(X) := \arg\min_{g \in \mathcal{G}} E[(Y - g(X))^2] \underbrace{=}_{\text{turns out}} E[Y|X]$$

where $\mathcal{G}$ is the space of functions defined on the support of X.

- If we know $m(X)$ up to a finite dimensional parameter $\theta \in \Theta \subset \mathbb{R}^K$, we may substantially restrict $\mathcal{G}$.
- If we knew $m(X) = g(X, \theta)$, we may consider the semi-parametric model

$$\mathcal{G}_\theta := \{g(X, \theta); \theta \in \Theta\} \subset \mathcal{G}$$

- A widely used example: $g(X, \theta) = X\theta$.

Motivation:

- The prior information, $m(X) \in \mathcal{G}_\theta$, improves efficiency, if it's correct.
- But, if we use $m(X) \in \mathcal{G}_\theta$ when it is wrong, our analysis may be invalid
- Often, economic theory is silent about the functional form of $m(X)$
- We wish to have a method that is robust to a functional form assumption
- For this reason, we study nonparametric analysis
- We do density estimation first and then regression

Density Estimation:

- Let $Y_1, \ldots, Y_n$ be a random sample from an unknown distribution.
- We wish to learn the distribution of Y_i from the random sample.
- We may assume $Y_1, \ldots, Y_n \overset{iid}{\sim} f(\cdot|\theta)$ with unknown $\theta \in \Theta$.
- For example, the normal model has $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_{++}$.
- In most cases, we do not know the correct parametric family, and our model may be likely misspecified and it can be very different from the true model.
- We want our analysis to be robust to any misspecification.

Density Estimation: Histogram

- We start with a histogram.
- Let y_i be the realisation of the random variable Y_i in the sample.

- To draw a histogram, we choose an increasing sequence of equidistant points $\{a_0, a_1, \ldots, a_J\}$ such that $a_0 < \min\{y_i\}_{i=1}^N$ and $a_J > \max\{y_i\}_{i=1}^N$.

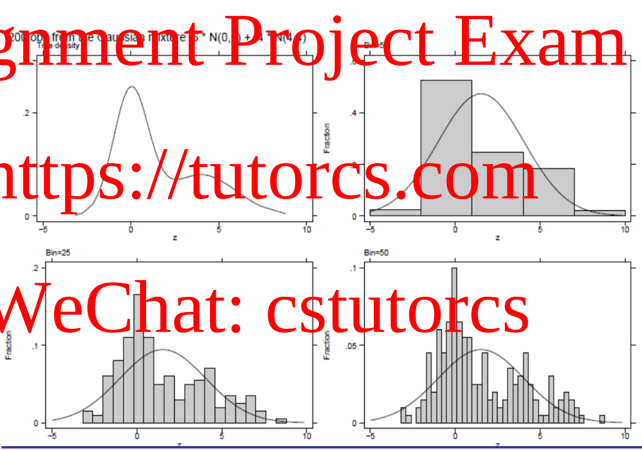
- The histogram, $\{H_1, \ldots, H_J\}$, counts the number of observations in each sub-interval $(a_{j-1}, a_j]$, i.e.,

$$H_j := \sum_{i=1}^{N} \mathbb{1}(y_i \in (a_{j-1}, a_j])$$

- By normalising $\{H_j\}_{j=1}^J$, we may have a naive density estimate.

Density Estimation: Histogram

Figure 1: Histogram estimates of density.



Density Estimation: Histogram

- The density estimate based on $\{H\}_{j=1}^{J}$ provides useful information on the data distribution
- small $J \rightarrow$ uninformative, but large $J \rightarrow$ noisy (too bumpy)
- J has to be appropriately chosen (depending on N).
- A good choice of J should increase as N grows
- The histogram estimate is a step function, but the true density may not be a step function.
- This density estimate cannot be used for any problem where a derivative of density plays a critical role, e.g., auction models.
- We wish to have a 'smooth' density estimate.

Kernel Density Estimation

- A kernel $K(\cdot)$ is a density with mean zero.
- The density $\phi(\cdot)$ of $\mathcal{N}(0,1)$ can be used as a kernel (Gaussian Kernel). The density of $\mathcal{N}(y, h^2)$ can be written as

$$f_{y,h}(z) = \frac{1}{h}\phi\left(\frac{z-y}{h}\right)$$

- Similarly, when a kernel is shifted by y_i and rescaled by h, we have

$$K_{y_i,h}(z) := \frac{1}{h}K\left(\frac{z-y_i}{h}\right)$$

- A kernel density estimate (KDE) at a point y is:

$$\hat{f}(y) := \frac{1}{N}\sum_{i=1}^{N}\frac{1}{h}K\left(\frac{y-y_i}{h}\right)$$

- When we have a bell-shape kernel, the KDE puts high weights on data near the point y and small weights on data far from y.

Kernel Density Estimation

- We choose a kernel $K(\cdot)$ and bandwidth (smoothing parameter) h.
- Some examples of kernels:

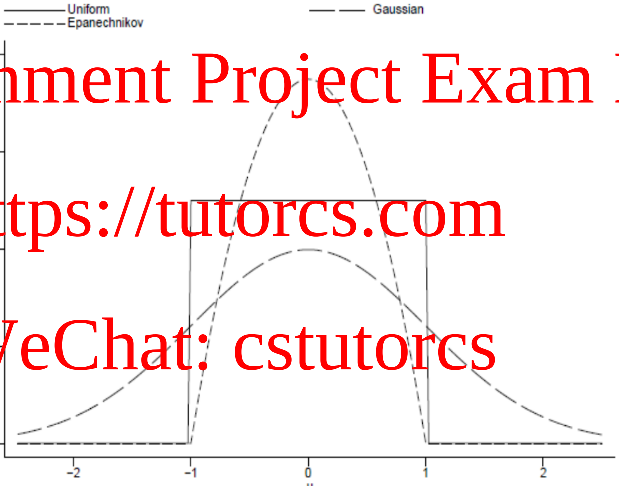
$$\text{Uniform Kernel: } K(u) = \frac{1}{2}\mathbb{1}(|u| < 1)$$

$$\text{Triangular Kernel: } K(u) = (1 - |u|)\mathbb{1}(|u| < 1)$$

$$\text{Epanechnikov Kernel: } K(u) = \frac{3}{4}(1 - z^2)\mathbb{1}(|u| < 1)$$

$$\text{Gaussian Kernel: } K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right)\mathbb{1}(u \in \mathbb{R})$$

Kernel Density Estimation



Kernel Density Estimation

- Different kernel weights differently the sample points.
- Uniform kernel puts equal weights within the support $[y - h, y + h]$
- Triangular kernel puts weight (inversely) proportional to the distance from y in $[y - h, y + h]$.
- Epanechnikov Kernel is bell-shape on the support $[y - h, y + h]$.
- Gaussian Kernel is bell-shape on the entire $\mathbb{R}$.
- A smooth kernel like Epanechnikov or Gaussian is often preferred.

Kernel Density Estimation

- We generate a random sample of size 100 drawn from the $\mathcal{N}(0, 25^2)$ and fit the date by KDEs with different kernels (all with $h = 12.5$)

Kernel Density: Different Kernels

Kernel Density Estimation

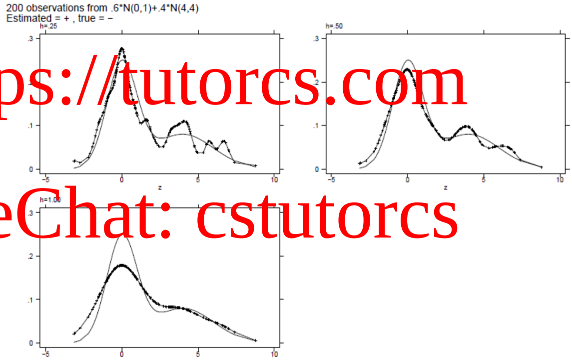
- Choice of h is much more important than choice of $K(\cdot)$
- large $h \rightarrow$ uninformative, but small $h \rightarrow$ noisy (too bumpy)

Figure 2: Uniform kernel density estimates.

Kernel Density Estimation

- The trade off is common for all kernels
- A good choice of h decreases as N increases

Figure [illegible]: Gaussian kernel density estimates.



Kernel Density Estimation

- The bias of the Kernel density estimator is

$$Bias(y) = E\left(\hat{f}(y)\right) - f(y) = \frac{1}{2}h^2 f''(y) \int z^2 K(z)\,dz$$

  which depends bandwidth, the curvature of the true density and the kernel.
- The bias is **increasing** in h, and disappears asymptotically if $h \to 0$ as $N \to \infty$.

- The variance is

$$V\left(\hat{f}(y)\right) \approx \frac{1}{Nh} f(y) \int K(z)^2\,dz$$

  which depends on the sample size, bandwidth, the true density and the kernel.
- The variance is **decreasing** in h, and disappears asymptotically if $Nh \to \infty$ as $N \to \infty$. That is that $h \to 0$ at a slower rate than the $N \to \infty$.

Kernel Density Estimation

- Provided that both the bias and variance terms disappear asymptotically (i.e. for appropriate h) the kernel estimator is **pointwise consistent**

$$\hat{f}(y) \xrightarrow{p} f(y)$$

  at a point y
- To obtain **uniform consistency** we need that

$$\sup_y |\hat{f}(y) - f(y)| \xrightarrow{p} 0$$

  which can be shown to occur if $Nh/\ln N \to \infty$. This requires a larger h than pointwise consistency.

- The limit distribution is:

$$\sqrt{Nh}\left(\hat{f}(y) - f(y) - Bias(y)\right) \xrightarrow{d} \mathcal{N}\left(0, f(y)\int K(z)^2 dz\right)$$

Kernel Density Estimation

- Optimal choice of bandwidth?
- We frequently represent our loss using the mean squared error (MSE). For a parametric model,

$$MSE(\hat{\theta}) := E[(\hat{\theta} - \theta)^2] = V(\hat{\theta}) + Bias(\hat{\theta})^2$$

- The nonparametric estimate $\hat{f}(y)$ is a function of y. So is $MSE(\hat{f}(y))$

- Thus, we use the overall MSE after integrating y out. The mean integrated squared error (MISE) is given as

$$MISE[\hat{f}(y)] := E\left(\int \left[\hat{f}(y) - f(y)\right]^2 dy\right) = \int V[\hat{f}(y)]dy + \int Bias[\hat{f}(y)]^2 dy$$

- The optimal bandwidth h^* minimises the MISE by balancing the variance and bias. But, h^* is infeasible because $f(y)$ is unknown.

Kernel Density Estimation

- Let s be the sample standard deviation.
- Silverman (1986) proposed the rule-of-thumb

$$\widehat{h}^* := 1.059 s N^{-1/5}$$

which is optimal when the reference distribution is $\mathcal{N}(\mu, \sigma^2)$

Kernel Density Estimation

- Silverman's rule may result in an inaccurate estimate if the true distribution is not close to the normal, in which case

$$\widetilde{h}^* := 1.3643\delta N^{-1/5} \min\left\{ s, \frac{\hat{q}_{.75} - \hat{q}_{.25}}{1.349} \right\}$$

  is often used, where $\hat{q}_{.75}$ and $\hat{q}_{.25}$ are sample percentiles and δ is a constant depending on the kernel.
- Values of δ are: 1.3510 (Uniform), 1.7188 (Epanechnikov), 2.0362 (Quartic/biweight), 2.3122 (Triweight), 0.7764 (Normal).

- The optimal bandwidth converges slowly to zero.
- There is no commonly accepted way to choose h and prior beliefs on smoothness are often exploited in practice.

Kernel Density Estimation: joint density

- Suppose we observe a random sample $(y_i, x_i)_{i=1}^{N}$ drawn from an unknown bivariate distribution.
- Then, a natural extension of the univariate KDE would be

$$\widehat{f}(x, y) = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{1}{h}K\left(\frac{x - x_i}{h}\right)\right\}\left\{\frac{1}{b}K\left(\frac{y - y_i}{b}\right)\right\}$$

- Similarly, we may obtain a high dimensional KDE.
- However, the required N rapidly increases $\Rightarrow$ curse of dimensionality.
- A density with dimensionality greater than 3 or 4 should be parametrised for any realistic sample size.

Kernel Density Estimation: conditional density

- The conditional density can be estimated by

$$\begin{aligned}
\widehat{f}(y|x) &= \frac{\widehat{f}(x,y)}{\widehat{f}(x)} \\
&= \frac{\frac{1}{N}\sum_{i=1}^{N}\left\{\frac{1}{h}K\left(\frac{x-x_i}{h}\right)\right\}\left\{\frac{1}{b}K\left(\frac{y-y_i}{b}\right)\right\}}{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{h}K\left(\frac{x-x_i}{h}\right)} \\
&= \sum_{i=1}^{N}\omega_{i,h}(x)\left\{\frac{1}{b}K\left(\frac{y-y_i}{b}\right)\right\}
\end{aligned}$$

where

$$\omega_{i,h}(x) := \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^{N}K\left(\frac{x-x_i}{h}\right)}$$

Kernel Density Estimation: conditional expectation

- The conditional expectation can be nonparametrically estimated by

$$
\begin{aligned}
\widehat{E}[Y|X=x] &:= \int y \widehat{f}(y|x) dy \\
&= \int y \sum_{i=1}^{N} \omega_{i,h}(x) \left\{ \frac{1}{b} K\left(\frac{y-y_i}{b} \right) \right\} dy \\
&= \sum_{i=1}^{N} \omega_{i,h}(x) \underbrace{\int y \left\{ \frac{1}{b} K\left(\frac{y-y_i}{b} \right) \right\} dy}_{=\text{a density with mean } y_i} \\
&= \sum_{i=1}^{N} \omega_{i,h}(x) y_i
\end{aligned}
$$

Nonparametric Regressions

- Now, recall that we want to estimate the regression function

$$m(X) = \arg\min_{g \in \mathcal{G}} E[(Y - g(X))^2] = E[Y|X]$$

where $\mathcal{G}$ is the space of functions defined on the support of X.

- We've just learned that a nonparametric regression is given as

$$\widehat{m}(x) = \sum_{i=1}^{N} \omega_{i,h}(x) y_i$$

where

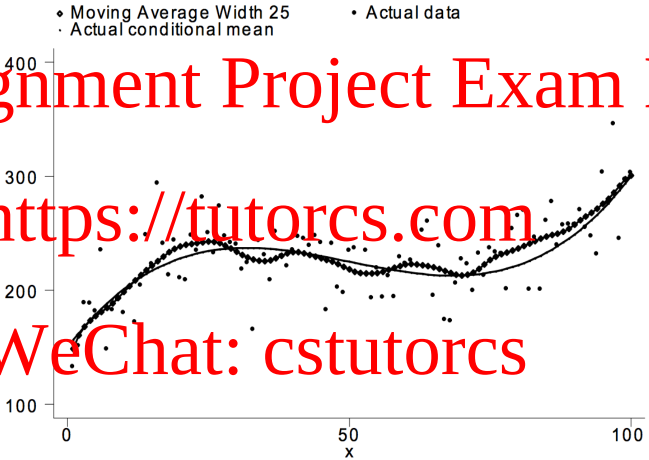
$$\omega_{i,h}(x) = \frac{K\left(\frac{x - x_i}{h}\right)}{\sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right)}$$

- This estimator is called Nadaraya & Watson estimator.

Nonparametric Regressions

- N-W estimator is a weighted average of y_i because $\sum_{i=1}^{N} \omega_{i,h}(x) = 1$.
- Especially, it puts large (small) weights on observations (y_i, x_i) with x_i close to (far from) the given point x.
- N-W is a special case of local weighted average (LOWESS) estimator.
- In a class of LOWESS estimators, there are many options to choose the weights $\omega_{i,h}(x)$.
- For example, a LOWESS estimate at x is the average of y_is whose associated x_is are included in the $h \times 100\%$ closest to the given x, where $h \in (0, 1)$. Note: Stata default for $h = 0.8$.
- Also, LOWESS estimate at x is the average of y_is whose associated x_is are included in the h nearest to the given x, where $h \in \mathbb{N}$.

Nonparametric Regressions



Kernel Density Estimates with Bandwidth 25

Nonparametric Regressions

- Just like KDE, $h \uparrow \Rightarrow$ Variance $\downarrow$ but Bias $\uparrow$. There is a trade-off. The optimal bandwidth h^* should balance the variance and bias. But, there is no commonly accepted rule for choosing the bandwidth.
- Moreover, x_i should be low dimensional: curse of dimensionality.
- When the regressor is high-dimensional, we may consider a semi-parametric specification such as

$$m(x, z; \beta) = x'\beta + \lambda(z)$$

which has both parametric part $x'\beta$ and nonparametric part $\lambda(z)$.

Series Estimation

- Series estimation is widely used for nonparametric analysis.
- Let $\phi_1(\cdot), \phi_2(\cdot), \ldots$ be a sequence of (basis) functions such that

$$\sum_{j=1}^{\infty} \theta_j \phi_j(\cdot)$$

approximates any function on the same domain with some $\theta_1, \theta_2, \ldots$.

- Then, if $\phi_1(\cdot), \phi_2(\cdot), \ldots$ are densities, by restricting θ_j to be positive and sum to one, we have a fully flexible density specification with infinite dimensional θ.
- If $\phi_1(\cdot), \phi_2(\cdot), \ldots$ are not densities, by normalising

$$\exp\left(\sum_{j=1}^{\infty} \theta_j \phi_j(\cdot)\right)$$

we may approximate any density.

Series Estimation

- In practice, we cannot have infinitely many components. So, we often use a finite number K of components.
- K determines the smoothness of the estimate like h of KDE.

- If K is too small, the density estimate would not be informative.
  If K is too large, the density estimate would be too noisy (bumpy).

- The optimal number of components K should increase as the sample size increases.

Series Estimation

- How to choose K?
- Frequentist:
  - often use BIC or AIC, or some data driven method such as cross-validation, or even use prior information informally.
  - There is no universally accepted rule for choosing K.
  - In any case, computation of parameter estimates is for each given fixed K, but inference is nonparametric (slow-convergence).
- Bayesian:
  - regard K as a latent variable (parameter), put a prior on K with a full support $\mathbb{N}$, and obtain the posterior of K as well as other parameters using an MCMC method.
  - Do not choose an arbitrary K as its distribution is determined.

Series Estimation

- There are many functions that can be used as basis functions $\{\phi_j(\cdot)\}$
- Examples include
  - Polynomials: Legendre polynomials, Bernstein polynomials, etc.
  - Splines: piecewise linear splines, *B*-splines, etc.
  - Densities: Beta densities (Bernstein polynomials), normal densities, Gamma densities, etc.

Series Estimation: BPD

- For example, the Bernstein polynomial density of order K is given as

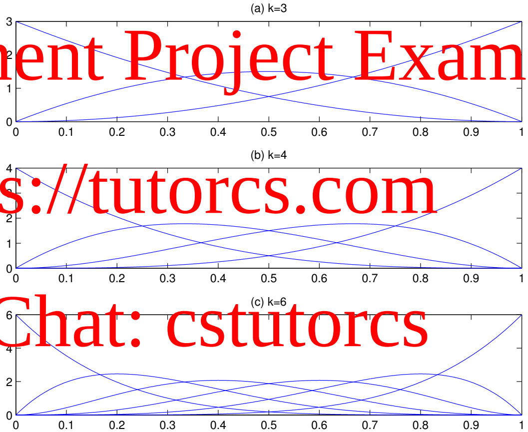
$$f(y|\theta_1, \ldots, \theta_K) := \sum_{j=1}^{K} \theta_j \mathrm{Beta}(j, K - j + 1)$$

where θ_j are all positive and sum to 1 and $\mathrm{Beta}(a, b)$ denotes the Beta density with parameters a and b, i.e., its mean is $ab/(a + b)$.

- When $K \to \infty$, the BPD approximates any absolutely continuous density on $[0, 1]$; see Petrone (1999) for Bayesian nonparametric method using BPD.

Series Estimation: BPD

- The Basis functions are plotted below



- The BPD is a histogram smoothing; see Petrone (1999)

Series Estimation: normal mixture

- Another example: normal densities can be used as basis functions.

$$f(y|\{\pi_j, \mu_j, \sigma_j^2\}) = \sum_{j=1}^{\infty} \pi_j \left\{ \frac{1}{\sigma_j} \phi\left(\frac{y - \mu_j}{\sigma_j} \right) \right\}$$

where $\phi(\cdot)$ is the PDF of $\mathcal{N}(0, 1)$.

- The normal mixture approximates any absolutely continuous density.