ApEc 8213: Econometric Analysis III -- Lecture #6

Nonparametric Density Estimation
Hansen, *Probability and Statistics for Economists* Ch. 17

I. Introduction

"Narrow-minded" linear models, that is linear models without higher ordered (e.g. squared) or interaction terms, are very restrictive. Adding higher ordered and interaction terms provides more flexibility. But even this may still be too restrictive (Horowitz, *Econometrica*, March 2011) .

Non-parametric methods let the data show the "shape" of the relationship between y and the x variables **without any parameters**. **Semi-parametric methods have some parameters** but in other "dimensions" are non-parametric. One example is a partially linear model with 3 x variables; the impact of two of them on y is linear but the impact of the third is completely non-parametric.

Both nonparametric and semi-parametric methods work by letting the data show the distribution of at least one variable without any distributional assumptions. This is **density estimation**. This lecture explains density estimation methods. Lecture 7 explains how to use these methods in many different kinds of regression models.

II. Histogram Density Estimation (17.2)

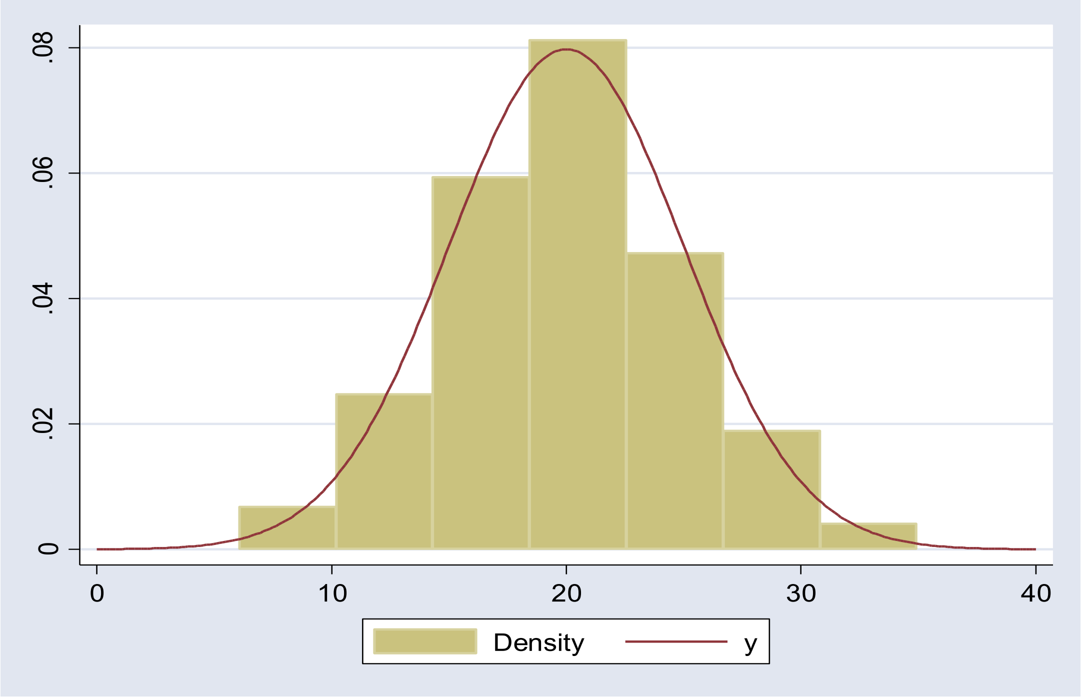
Suppose you have n observations for a variable X, and you want to estimate its density, denoted by $f(x)$. One way to do this is to use a **histogram**. Divide the range of X into B “bins” of width w and count the number in each bin. Let n_j be the number of observations in bin j. The **histogram** estimator of $\boldsymbol{f(x)}$ **for all $\boldsymbol{x}$ in bin $\boldsymbol{j}$** is:

$$\hat{f}(x) = \frac{n_j}{nw} \qquad (17.1)$$

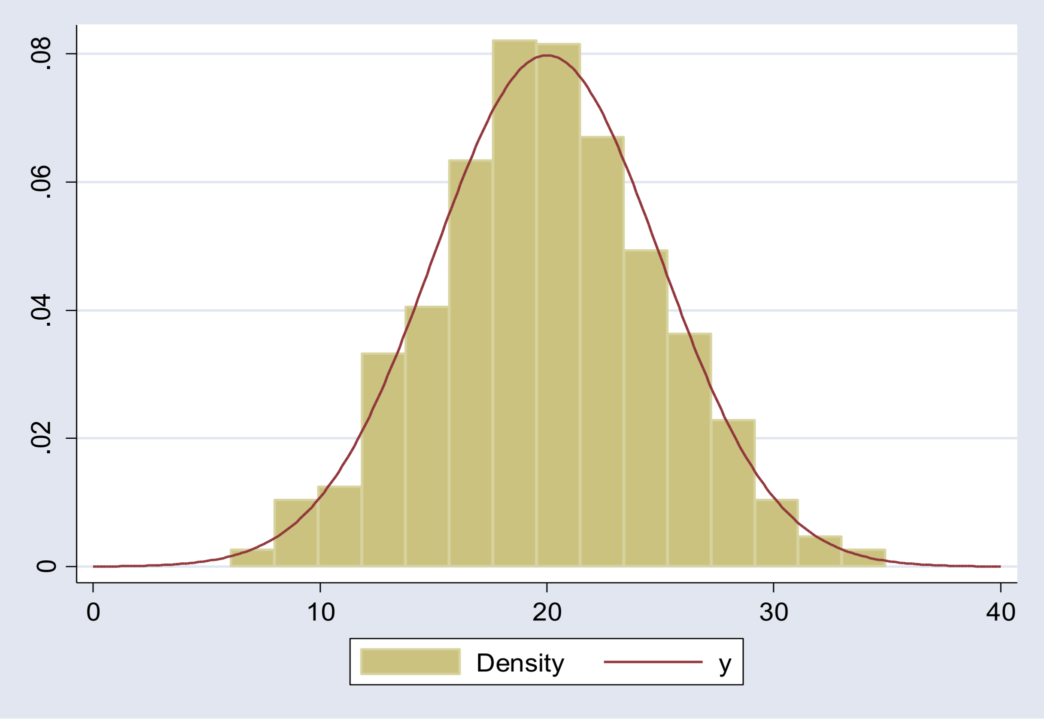
That is, you calculate the **percentage of observations in bin $\boldsymbol{j}$**, which is $\boldsymbol{n_j/n}$, and divide by w. You need to divide by w so that the “integral” of $\hat{f}(x)$, which is the total area under all the rectangles (for each rectangle (column) j, the rectangle’s area is $w \times n_j/(nw)$), equals 1: $\sum_{j=1}^{B} wn_j/nw = 1$.

Histograms are rather **clumsy** ways **to estimate a density**, especially a continuous density. The following graphs make this point. They also show how changing the number of bins can make the graphs “better” or “worse”, and how the “optimal” number of bins can vary depending on the size of your sample (n).

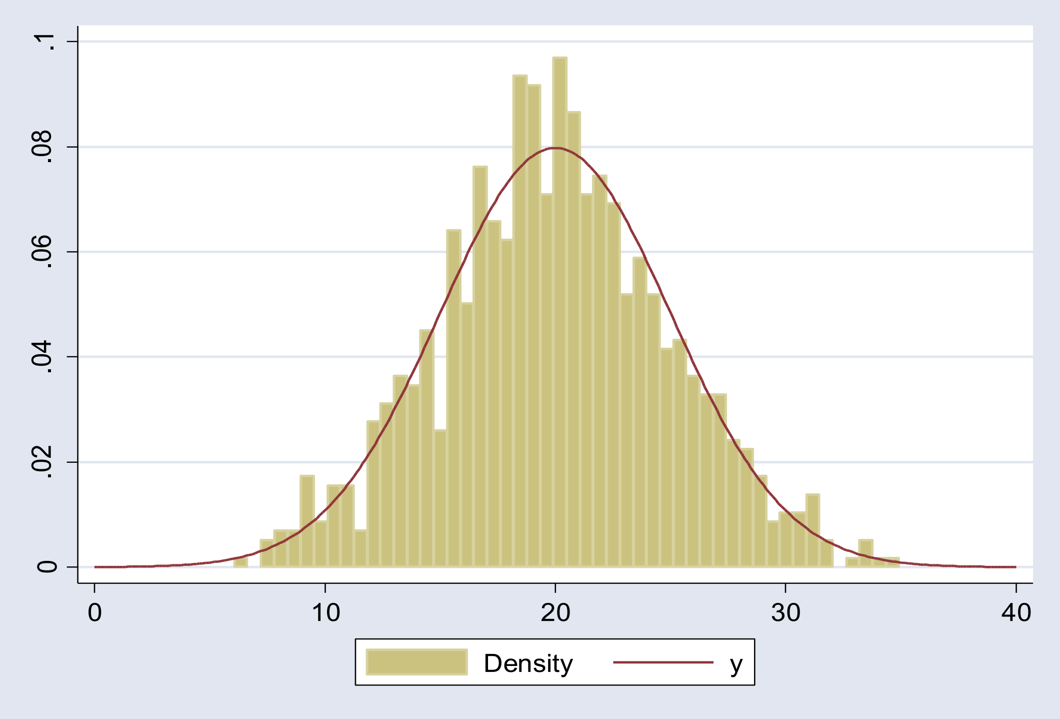
Graph 1: Histogram of a Normally Distributed Variable with Mean=20, SD=5 (1000 obs., 7 bins)



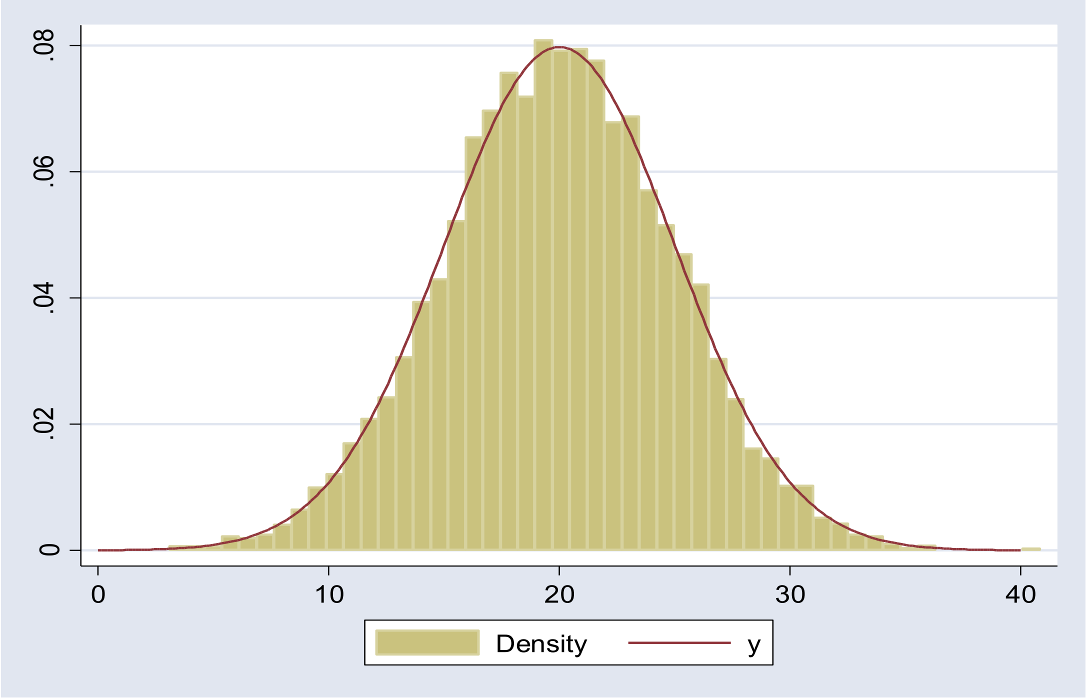
Graph 2: Histogram of a Normally Distributed Variable with Mean=20, SD=5 (1000 obs., 15 bins)



Graph 3: Histogram of a Normally Distributed Variable with Mean=20, SD=5 (1000 obs., 50 bins)



Graph 4: Histogram of a Normally Distributed Variable with Mean=20, SD=5 (10,000 obs., 50 bins)



III. Kernel Density Estimation (17.3)

Histograms have some limitations for estimating the density of X. **First**, within any given bin the density is the same, even if within the bin there are more observations in one part than in other parts. **Second**, for most points x in the density, $f(x)$, the observations used to calculate $f(x)$ are not symmetric around x. **Third**, as we will see in Lecture 9 it is useful to calculate the derivative of $f(x)$, and for histograms the derivatives are either 0 or undefined.

The **kernel density estimator** avoids these limitations:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \qquad \textbf{[W.O.B.]} \qquad (17.2)$$

where ***K*(*u*)** is a **weighting function** known as a **kernel function**, and $h > 0$ is a scalar called a **bandwidth**, which is analogous to the bin width of a histogram. For this to work, kernel functions must have the following properties:

Definition 17.1. A **kernel function** $K(u)$ must satisfy:

1. $0 \le K(u) \le \overline{K} < \infty$

2. $K(u) = K(-u)$ ($K(u)$ is symmetric around zero)

3. $\int_{-\infty}^{\infty} K(u)du = 1$

4. $\int_{-\infty}^{\infty} |u|^r K(u)du < \infty$ for all positive integers r

Hansen says that the fourth property "is not essential for most results but is a convenient simplification…".

It is **also convenient** to require the **kernel functions** to **have a variance of 1**, which is not restrictive:

Definition 17.2. A **normalized kernel function** satisfies:

$$\int_{-\infty}^{\infty} u^2 K(u)du = 1$$

In practice, only a few kernel functions are commonly used:

1. **Rectangular**: $$\begin{aligned} K(u) &= \frac{1}{2\sqrt{3}} \quad \text{if } |u| < \sqrt{3} \\ &= 0 \qquad\quad \text{if } |u| \geq \sqrt{3} \end{aligned}$$

For any given x this will give a histogram centered on x.

2. **Gaussian** (normal): $$K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$$

This has the disadvantage that you need to look at all of the observed values of x (there is no cut-off point).

3. **Epanechnikov**: $$\begin{aligned} K(u) &= \frac{3}{4\sqrt{5}} (1 - u^2/5) \text{ if } |u| < \sqrt{5} \\ &= 0 \qquad\qquad\qquad\quad \text{if } |u| \geq \sqrt{5} \end{aligned}$$

This kernel does have a cut-off point but it does not have a derivative at $|u| = \sqrt{5}$, which can be a disadvantage.

4. **Triangular**: $$K(u) = \frac{1}{\sqrt{6}}(1 - |u|/\sqrt{6}) \quad \text{if } |u| < \sqrt{6}$$
$$= 0 \quad \text{if } |u| \geq \sqrt{6}$$

This is similar to the rectangular kernel, except it gives greater weight to points that are closer to *x*.

5. **Biweight (Quartic)**: $$K(u) = \frac{15}{16\sqrt{7}}(1\text{-}u^2/7)^2 \quad \text{if } |u| < \sqrt{7}$$
$$= 0 \quad \text{if } |u| \geq \sqrt{7}$$

This kernel has a cut-off point (at $|u| = \sqrt{7}$), and it also has a derivative defined at $u = \sqrt{7}$.

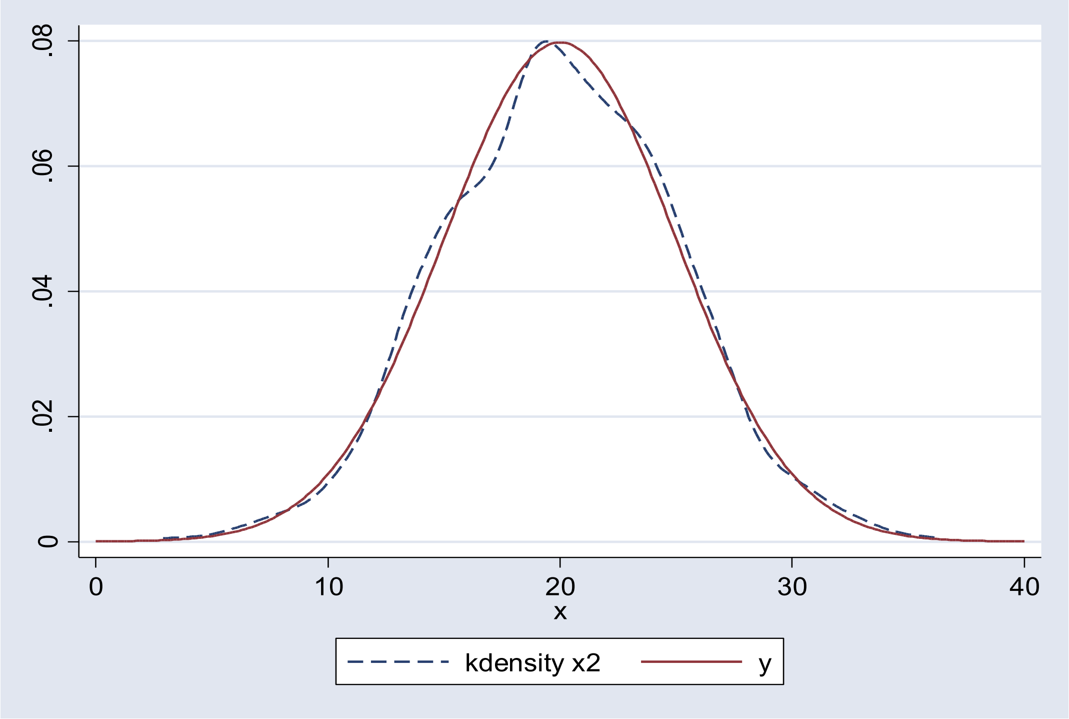
The **bandwidth** *h* is **very important**, so Hansen defines it:

Definition 17.3. A **bandwidth** (tuning) **parameter** $h > 0$ is a real number used to control the degree of smoothing for a nonparametric estimator.

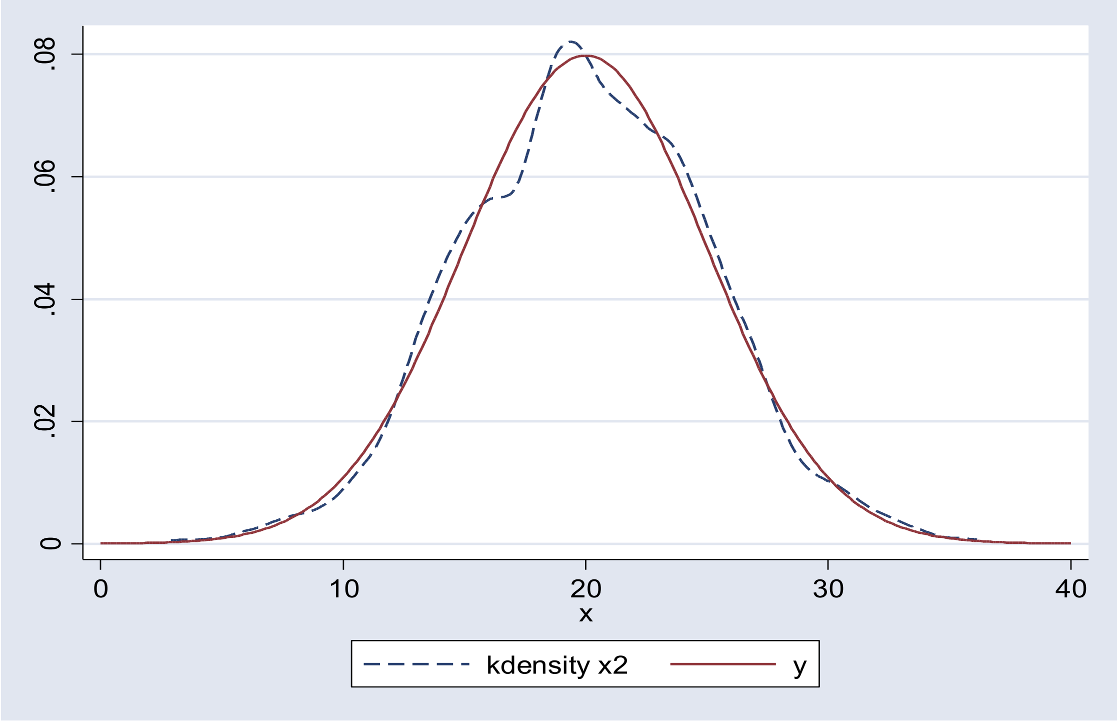
In general, **larger *h* gives "smoother" density estimates but**, as we will see, "**too smooth**" **can lead to bias**.

In practice, the choice of the kernel has little effect for drawing a parametric density, but the size of the bandwidth has a big effect, as seen in the following figures:

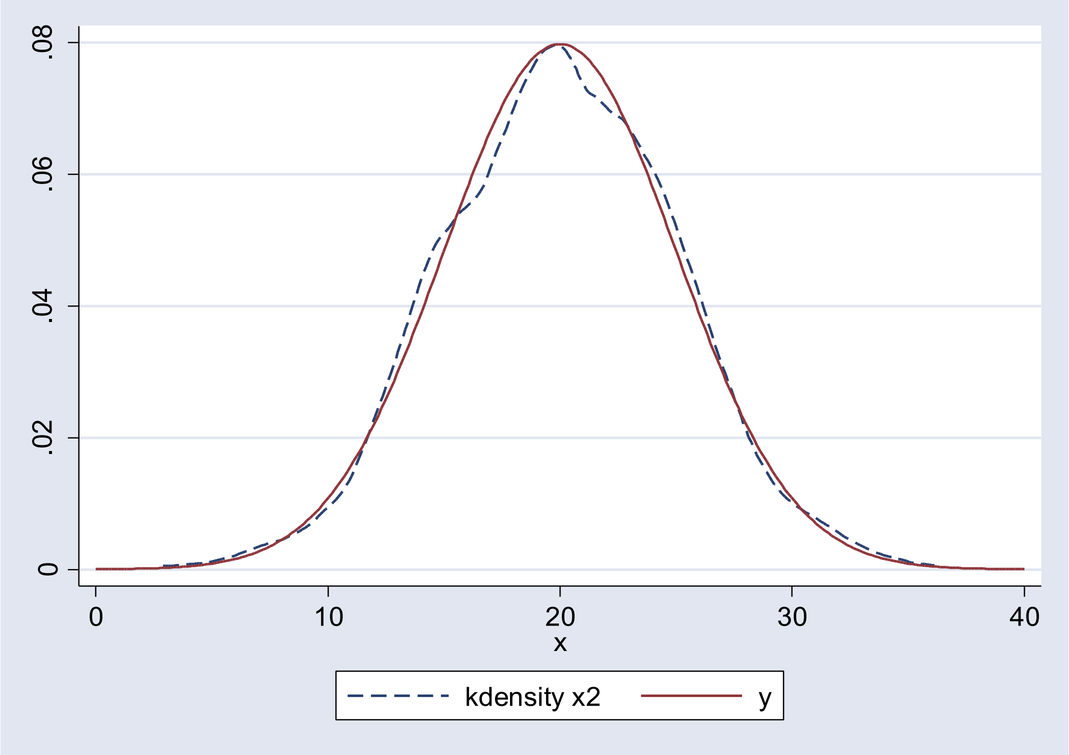
Graph 5: Density of a Normally Distributed Variable, Mean=20, SD=5 (1000 obs., Gaussian kernel, bw=1)



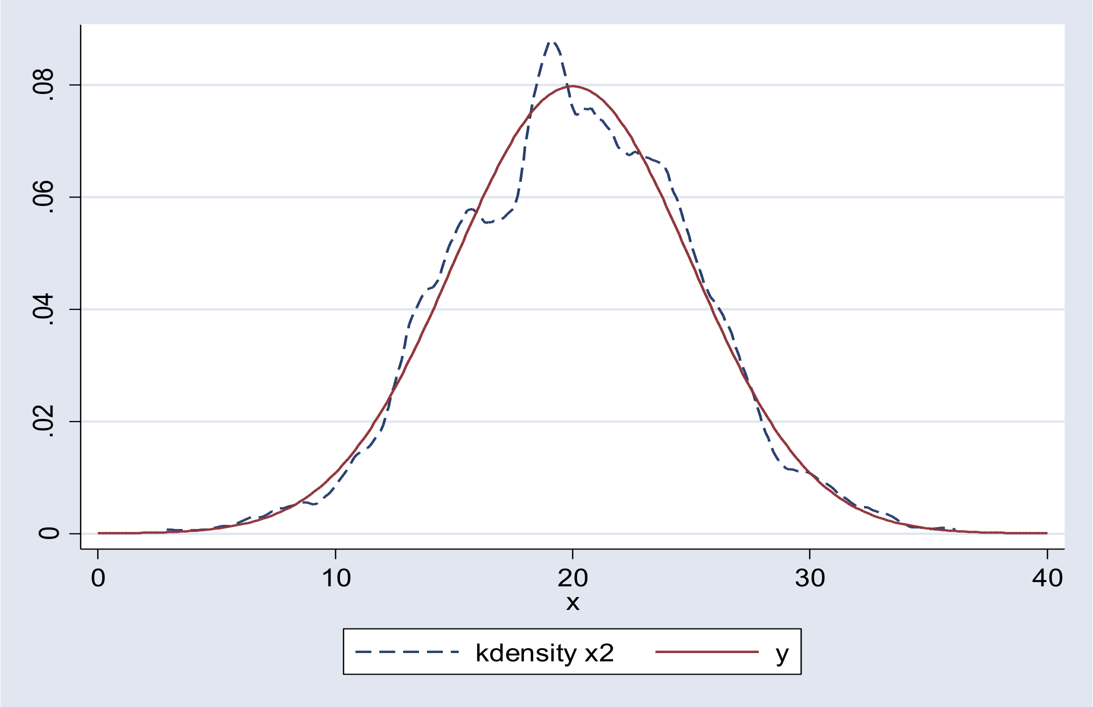
Graph 6: Density of a Normally Distributed Variable, Mean=20, SD=5 (1000 obs., Quartic kernel, bw=2)



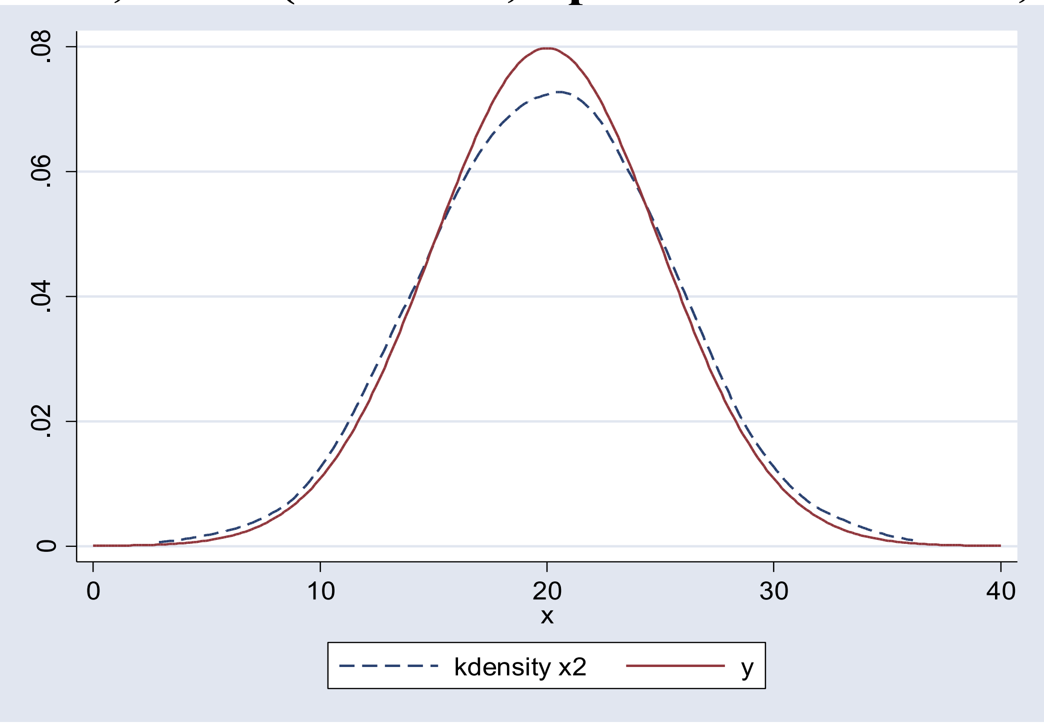
Graph 7: Density of a Normally Distributed Variable, Mean=20, SD=5 (1000 obs., Epanechnikov kernel, bw=1)



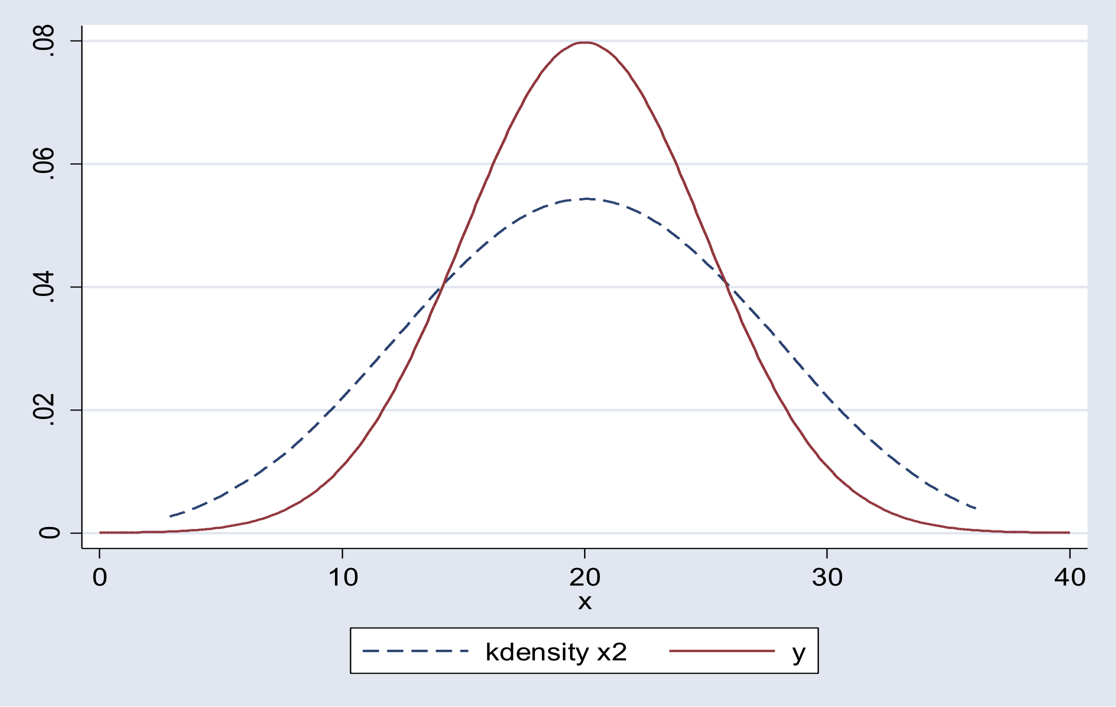
Graph 8: Density of a Normally Distributed Variable, Mean=20, SD=5 (1000 obs., Epanechnikov kernel, bw=0.5)



Graph 9: Density of a Normally Distributed Variable, Mean=20, SD=5 (1000 obs., Epanechnikov kernel, bw=2)



Graph 10: Density of a Normally Distributed Variable, Mean=20, SD=5 (1000 obs., Epanechn. kernel, bw=5)



A final property of using kernel functions is that they ensure that the estimated density function is a "proper" density function in that it integrates to 1:

$$\int_{-\infty}^{\infty} \hat{f}(x)dx = \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{\infty} K\left(\frac{X_i - x}{h}\right)dx = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} K(u)du = 1$$

IV. Bias and Variance of Density Estimators (17.4 - 17.6)

It turns out that, for any value of *x*, the **density estimate of *f*(*x*) is likely to be biased**, even though asymptotically **it is consistent**. However, the **extent of bias can be estimated**, which as we will see is useful for choosing the bandwidth (*h*). Since the density estimate in equation (17.2) for some *x* is an average of i.i.d observations, its expectation is:

$$\mathrm{E}[\hat{f}(x)] = \mathrm{E}[\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)] = \mathrm{E}[\frac{1}{h}K\left(\frac{X - x}{h}\right)]$$

Taking the expectation of a nonlinear function, ***K*( )**, of a random variable (*X*) **is tricky, but it can be done**:

$$\mathrm{E}[\frac{1}{h}K\left(\frac{X - x}{h}\right)] = \int_{-\infty}^{\infty} \frac{1}{h}K\left(\frac{v - x}{h}\right)f(v)dv$$

This can be simplified by defining $u = (v - x)/h$, which implies that $v = x + hu$ and $dv = hdu$. Then:

$$\mathrm{E}[\hat{f}(x)] = \mathrm{E}[\frac{1}{h}K\left(\frac{X-x}{h}\right)] = \int_{-\infty}^{\infty} K(u)f(x+hu)du$$

$$= f(x) + \int_{-\infty}^{\infty} K(u)[f(x+hu) - f(x)]du \qquad (17.3)$$

The last line simply add and subtracts $f(x)$, noting that $f(x)$ within the integral sign is a constant and $\int_{-\infty}^{\infty} K(u)du = 1$.

The **second term** in (17.3) **is the bias**. Notice that the bias goes to zero has h goes to zero. This can be expressed as:

$$\mathrm{E}[\hat{f}(x)] = f(x) + o(1)$$

where $o(1)$ is a term that goes to 0 as $n \rightarrow \infty$.

Hansen then derives the following (the derivations in Hansen's book are optional):

Theorem 17.1. If $f(x)$ is continuous in the neighborhood $\mathcal{N}$ around x, then as $h \rightarrow 0$:

$$\mathrm{E}[\hat{f}(x)] = f(x) + o(1) \qquad (17.4)$$

If $f''(x)$ is continuous in the neighborhood $\mathcal{N}$ around x, then as $h \rightarrow 0$,

$$\mathrm{E}[\hat{f}(x)] = f(x) + (1/2)f''(x)h^2 + o(h^2) \qquad (17.5)$$

The term ***o*(*h*^{2}) is much smaller than** the term ***o*(1)**. For any sequence x_n, $x_n = o(1)$ implies that $x_n \rightarrow 0$ as $n \rightarrow \infty$, but for another sequence x'_n, the property $x'_n = o(h^2)$ implies that $x'_n/h^2 \rightarrow 0$ as $n \rightarrow \infty$. Since $h \rightarrow 0$, x'_n must become very small for $x'_n/h^2 \rightarrow 0$ as $n \rightarrow \infty$.

Equation (17.5) is useful in two ways. **First**, it tells us that bias is positive if $f''(x) > 0$ and negative if $f''(x) < 0$. **Second**, if $f(x)$ "has a lot of curvature" in the sense that there are many parts of it with $f''(x)$ that is large in absolute value, the bias will, in general be worse.

The first property can be seen in Graphs 9 and 10 on page 10 (and can also be seen in Figure 17.4 in Hansen).:

Variance of Density Estimator

Since $\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$, and the n observations are i.i.d., **the variance of $\hat{f}(x)$ is the sum of the variances of these *n* terms**:

$$\mathrm{Var}[\hat{f}(x)] = \frac{1}{n^2h^2}\mathrm{Var}\left[\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)\right] = \frac{1}{nh^2}\mathrm{Var}\left[K\left(\frac{X - x}{h}\right)\right]$$

Note that in the last expression the only variable is X, so if the kernel $K(\ )$ is a nonlinear function, we will have to approximate the variance. This leads to:

Theorem 17.2. The exact variance of $\hat{\boldsymbol{f}}(x)$ is:

$$\text{Var}[\hat{f}(x)] = \frac{1}{nh^2}\text{Var}\left[K\left(\frac{X-x}{h}\right)\right] \qquad (17.6)$$

Also, **if $\boldsymbol{f(x)}$ is continuous** in the neighborhood $\mathcal{N}$ around x, then as $h \rightarrow 0$ and $nh \rightarrow \infty$

$$\text{Var}[\hat{f}(x)] = \frac{f(x)R_K}{nh} + o(\frac{1}{nh}) \qquad (17.7)$$

where $$R_K = \int_{-\infty}^{\infty}(K(u))^2 du \qquad (17.8)$$

R_K measures the "roughness" of the kernel function $K(u)$. It has been calculated for commonly used kernels (see Table 17.1 in *Hansen*). For example, $R_K = 0.289$ for the rectangular kernel, 0.268 for the Epanechnikov kernel, and 0.279 for the biweight (quartic) kernel, Since $nh \rightarrow \infty$, the term $o(1/nh)$ is much smaller than $o(1)$ and can be ignored.

In most applications, we assume that $f(x)$ is continuous, so we use (17.7). We can think if nh as the "effective" sample size. The variance decreases as both n and h increase.

Estimating the Variance and the Standard Errors

The formulas for $\text{Var}[\hat{f}(x)]$ in equations (17.6) and (17.7) provide **two different ways to estimate $\widehat{\textbf{Var}}[\hat{\boldsymbol{f}}(\boldsymbol{x})]$**.

For equation (17.6), rewrite $\frac{1}{nh^2}$Var[$K\left(\frac{X-x}{h}\right)$] as $\frac{1}{n}$Var[$\frac{1}{h}K\left(\frac{X-x}{h}\right)$]. The sample variance of $\frac{1}{h}K\left(\frac{X-x}{h}\right)$ is:

$$(\frac{1}{n-1})\sum_{i=1}^{n}[\frac{1}{h}K\left(\frac{X_i-x}{h}\right) - \frac{1}{n}\sum_{j=1}^{n}\frac{1}{h}K\left(\frac{X_j-x}{h}\right)]^2$$

$$= (\frac{1}{n-1})\sum_{i=1}^{n}[\frac{1}{h}K\left(\frac{X_i-x}{h}\right) - \hat{f}(x)]^2$$

$$= (\frac{1}{n-1})\sum_{i=1}^{n}[(\frac{1}{h}K\left(\frac{X_i-x}{h}\right))^2 - 2\frac{1}{h}K\left(\frac{X_i-x}{h}\right)\hat{f}(x) + (\hat{f}(x))^2]$$

$$= (\frac{1}{n-1})\{\sum_{i=1}^{n}[(\frac{1}{h}K\left(\frac{X_i-x}{h}\right))^2] - 2\hat{f}(x)\sum_{i=1}^{n}[\frac{1}{h}K\left(\frac{X_i-x}{h}\right)] + n(\hat{f}(x))^2\}$$

$$= (\frac{1}{n-1})\{\sum_{i=1}^{n}[(\frac{1}{h}K\left(\frac{X_i-x}{h}\right))^2] - 2\hat{f}(x)n\hat{f}(x) + n(\hat{f}(x))^2\}$$

$$= (\frac{1}{n-1})\{\sum_{i=1}^{n}[(\frac{1}{h}K\left(\frac{X_i-x}{h}\right))^2] - n(\hat{f}(x))^2\}$$

This is $\widehat{\text{Var}}$[$\frac{1}{h}K\left(\frac{X-x}{h}\right)$], which implies that:

$$\widehat{\text{Var}}[\hat{f}(x)] = \frac{1}{n}\widehat{\text{Var}}\ [\frac{1}{h}K\left(\frac{X-x}{h}\right)]$$

$$= (\frac{1}{n-1})\{\frac{1}{nh^2}[\sum_{i=1}^{n}(K\left(\frac{X_i-x}{h}\right))^2] - (\hat{f}(x))^2\}$$

For equation (17.7) one can replace $f(x)$ with $\hat{f}(x)$ and ignore $o(\frac{1}{nh})$ term:

$$\widehat{\text{Var}}[\hat{f}(x)] = \frac{\hat{f}(x)R_K}{nh} \qquad (17.9)$$

To calculate the **standard error**, just **take the square root** of either of these estimates of $\widehat{\text{Var}}[\hat{f}(x)]$.

V. IMSE of Density Estimator (17.7, 17.8)

Recall that bias decreases as h gets smaller, but a smaller h increases the variance. An intuitive **indicator of a good fit** for the estimator of a density function is the **integrated mean squared error (IMSE)**. To start, consider the mean squared error (MSE) of $\hat{f}(x)$ **for a particular value of *x***:

$$\text{MSE}(\hat{f}(x)) \equiv \text{E}[(\hat{f}(x) - f(x))^2]$$

In fact, this is the sum of $\text{Var}[\hat{f}(x)]$ and the square of the bias of $\hat{f}(x)$, so it **accounts for both variance and bias**:

$$\text{MSE}(\hat{f}(x)) \equiv \text{E}[(\hat{f}(x) - f(x))^2]$$

$$= \text{E}[(\hat{f}(x))^2] - 2f(x)\text{E}[\hat{f}(x)] + (f(x))^2$$

$$= \mathrm{E}[(\hat{f}(x))^2] - (\mathrm{E}[\hat{f}(x)])^2 + (\mathrm{E}[\hat{f}(x)])^2 - 2f(x)\mathrm{E}[\hat{f}(x)] + (f(x))^2$$

$$= \mathrm{E}[(\hat{f}(x) - \mathrm{E}[\hat{f}(x)])^2] + (\mathrm{E}[\hat{f}(x)] - f(\mathrm{x}))^2$$

$$= \mathrm{Var}[\hat{f}(x)] + [\mathrm{bias}(\hat{f}(x))]^2$$

This is for just one value of *x*, so to measure the "fit" of $\hat{f}(x)$ **over all *x*** we **integrate** over all values to get IMSE:

$$\mathrm{IMSE} = \int_{-\infty}^{\infty} \mathrm{E}[(\hat{f}(x) - f(x))^2]dx$$

Using Theorems (17.1) and (17.2) [**homework?**] we get:

$$\mathrm{IMSE} = \frac{1}{4}R(f'')h^4 + \frac{R_K}{nh} + o(h^4) + o(\frac{1}{nh})$$

where $R(f'') = \int_{-\infty}^{\infty}(f''(x))^2 dx$; this is called the **roughness** of the second derivative $f''(x)$. **Asymptotically, the last two terms** (stochastic order symbols) **drop out**, yielding:

$$\mathrm{AIMSE} = \frac{1}{4}R(f'')h^4 + \frac{R_K}{nh} \qquad (17.10)$$

This is called the **asymptotic integrated mean squared error**. It shows that $\hat{f}(x)$ is less accurate when $R(f'')$ is large, that is when there is a large amount of "curvature" in $f(x)$. Note that the **first term is increasing in *h*** while the **second term is decreasing in *h***; this is the **trade-off between bias** (decreases with smaller h) **and variance** (increases with smaller h).

To get the **optimal bandwidth** (the bandwidth that minimizes AIMSE), one can show that differentiating (17.10) with respect to h gives:

$$h_0 \text{ (optimal bandwidth)} = (\frac{R_K}{R(f'')})^{1/5} n^{-1/5} \qquad (17.11)$$

While **we can calculate R_K** based on the kernel function we are using, **we do not know $R(f'')$. However**, there are **two useful results** here:

1. The optimal bandwidth h_0 is related to n as $h_0 = c/n^{1/5}$ for some constant c, which means that **h_0 declines at a very small rate as n increases**.
2. AIMSE is proportional to $n^{-4/5}$ (via inserting optimal h_0 into equation (17.10)), which means that the **density estimator converges at a rate of $n^{-2/5}$**, which is a slightly slower rate than the standard $n^{-1/2}$ for almost all parametric estimators.

To summarize:

Theorem 17.3. If $f''(x)$ is uniformly continuous, then:

$$\text{IMSE} = \frac{1}{4} R(f'')h^4 + \frac{R_K}{nh} + o(h^4) + o(\frac{1}{nh})$$

and the (asymptotic) optimal bandwidth, h_0, is $(\frac{R_K}{R(f'')})^{1/5} n^{-1/5}$.

Optimal Kernel

Which kernel function should one use? Hansen discusses this in Section 17.8 of Chapter 17. **Mathematically, the Epanechnikov kernel is optimal** in that it **minimizes AIMSE relative to all other kernels**. However, **the efficiency gain** here is **rather small**; for example, AIMSE of the Gaussian kernel is only about 2% higher than the AIMSE of the Epanechnikov kernel.

On the other hand, the Gaussian kernel (and the quartic kernel) is differentiable for all values of u.

Hansen also mentions that the estimator $\boldsymbol{\hat{f}(x)}$ **is > 0** for all values of x **when the Gaussian kernel is used**, and that this could be an advantage if you need to take the inverse of the $\hat{f}(x)$ function. This is not true of the Epanechnikov kernel or the biweight kernel. But the Gaussian kernel can slow down computation of $\hat{f}(x)$ because you need to calculate over all values if x in your data for any given point x for which you are calculating $\hat{f}(x)$, which is not needed for the Epanechnikov or biweight kernels. On the other hand, computers are getting faster every year so this may not matter much unless you have a very large data set.

VI. Optimal Bandwidth (17.9, 17.10 and 17.11)

Recall that the optimal bandwidth formula in equation (17.11) depends on $f(x)$, or more specifically on the second derivative of $f(x)$. We do not know this, so how can we use the formula?

Silverman (1986) worked out the **exact formula if $f(x)$ follows a normal distribution** ("Silverman rule"):

$$h_r = \sigma C_K n^{-1/5} \qquad (17.12)$$

where σ is the standard deviation of X and:

$$C_K = \left(\frac{8\sqrt{\pi}R_K}{3}\right)^{1/5}$$

In fact, you can apply this to any kernel function, which have slightly different R_K (Hansen gives values of R_K and C_K in Table 17.1). **Strictly speaking**, Silverman's rule is **optimal only if X is normally distributed** (in which case C_K is approximately 1.059). **In practice, it seems to work well even if X is not normally distributed**. Silverman also noted that this formula may not work as well with bimodal and "thick-tailed" distributions, and for those he recommends that $C_K = 0.9$.

A few years later, **Sheather and Jones (1991)** proposed a way to calculate an optimal bandwidth by estimating $R(f'')$ nonparametrically. This involves using a "starting value" bandwidth and then updating until the algorithm converges to the optimal bandwidth. See Section 17.10 for details. This performs "quite well" according to Hansen, and Stata (and presumably other software) can implement this procedure.

So what should you do? In practice, try 2-3 bandwidths and compare your estimates. If your estimate is very smooth, you may be "oversmoothing" and so you have too large of a bandwidth. Try smaller bandwidths until you get "small bumps" that are probably "not real" but just reflect random variation, and go to a slightly larger bandwidth and stop there. See further discussion in Hansen, Section 17.11.

Sections 17.12 and 17.13 in Hansen provide some practical advice, including what different software does.

VI. Asymptotic Distribution (17.14)

Finally, note that $\hat{f}(x)$ is consistent and asymptotically normally distributed:

Theorem 17.6. If $f(x)$ is continuous in the neighborhood $\mathcal{N}$ around x, then as $h \to 0$ and $nh \to \infty$, $\hat{f}(x) \underset{p}{\to} f(x)$.

Theorem 17.7. If $f''(x)$ is continuous in the neighborhood $\mathcal{N}$ around x, then as $nh \to \infty$ such that $h = O(n^{-1/5})$:

$$\sqrt{nh}(\hat{f}(x) - f(x) - \frac{1}{2}f''(x)h^2) \underset{d}{\to} \mathrm{N}(0, f(x)R_K).$$

Three things to note about Theorem 17.7 are:

1. It **converges at a rate of $\sqrt{\boldsymbol{nh}}$ instead of the usual rate of $\sqrt{\boldsymbol{n}}$**. This is because at any given value of x the effective sample size nh rather than h.

2. **The term $\frac{1}{2}\boldsymbol{f''(x)h^2}$** is an explicit adjustment for bias. It is ***not asymptotically negligible*** and needs to be "acknowledged".

3. The extra condition that $h = O(n^{-1/5})$ requires $h \to 0$ at a rate of at least $n^{-1/5}$.

Finally, Hansen discusses in Section 17.15 how "undersmoothing" can reduce the bias pointed out in Theorem 17.7. This amounts to choosing h to be smaller than the optimal bandwidth. However, this smaller h increases the variance of $\hat{f}(x)$ and is also inefficient from the perspective of minimizing AIMSE. So Hansen does not recommend undersmoothing.