ECON675: Assignment 2

Anirudh Yadav

October 10, 2018

Contents

1 Question 1: Kernel Density Estimation

1.1 Density derivatives

I follow the derivation in Hansen's notes. We are interested in estimating

$$f^{(s)}(x) = \frac{d^s}{dx^s} f(x).$$

The natural estimator is

$$\hat{f}^{(s)}(x) = \frac{d^s}{dx^s} \hat{f}(x)$$

Now, we know that $\hat{f}(x) = \frac{1}{nh} \sum_i K\left(\frac{X_i - x}{h}\right)$. Thus,

$$\begin{aligned}
\hat{f}^{(1)}(x) &= \frac{-1}{nh^2} \sum_{i=1}^{n} K^{(1)}\left(\frac{X_i - x}{h}\right), \\
\hat{f}^{(2)}(x) &= \frac{1}{nh^3} \sum_{i=1}^{n} K^{(2)}\left(\frac{X_i - x}{h}\right), \\
&\vdots \\
\hat{f}^{(s)}(x) &= \frac{(-1)^s}{nh^{1+s}} \sum_{i=1}^{n} K^{(s)}\left(\frac{X_i - x}{h}\right).
\end{aligned}$$

Now,

$$\begin{aligned}
\mathbb{E}[\hat{f}^{(s)}(x)] &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\frac{(-1)^s}{h^{1+s}} K^{(s)}\left(\frac{X_i - x}{h}\right)\right] \\
&= \mathbb{E}\left[\frac{(-1)^s}{h^{1+s}} K^{(s)}\left(\frac{X_i - x}{h}\right)\right], \text{ since } X_i \text{ are iid.} \\
&= \int_{-\infty}^{\infty} \frac{(-1)^s}{h^{1+s}} K^{(s)}\left(\frac{z - x}{h}\right) f(z) dz
\end{aligned}$$

Next, we want to use integration by parts: $\int u dv = uv - \int v du$. Define

$$dv = \frac{(-1)^s}{h^s} \frac{1}{h} K^{(s)}\left(\frac{z - x}{h}\right) \implies v = \frac{(-1)^s}{h^s} K^{(s-1)}\left(\frac{z - x}{h}\right)$$

And

$$u = f(z) \implies du = f^{(1)}(z).$$

Thus,

$$\begin{aligned}
\mathbb{E}[\hat{f}^{(s)}(x)] &= \left[\frac{(-1)^s}{h^s} K^{(s-1)}\left(\frac{z - x}{h}\right) f^{(1)}(z)\right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \frac{(-1)^s}{h^s} K^{(s-1)}\left(\frac{z - x}{h}\right) f^{(1)}(z) dz. \\
&= -\int_{-\infty}^{\infty} \frac{(-1)^s}{h^s} K^{(s-1)}\left(\frac{z - x}{h}\right) f^{(1)}(z) dz
\end{aligned}$$

Repeating this s times give

$$\begin{aligned}\mathbb{E}[\hat{f}^{(s)}(x)] &= (-1)^s \int_{-\infty}^{\infty} \frac{(-1)^s}{h} K\left(\frac{z-x}{h}\right) f^{(s)}(z)dz \\ &= \int_{-\infty}^{\infty} \frac{1}{h} K\left(\frac{z-x}{h}\right) f^{(s)}(z)dz\end{aligned}$$

Next, use the following change of variables: $u = \frac{z-x}{h}$, which implies $z = x + hu \implies dz = hdu$. Thus,

$$\mathbb{E}[\hat{f}^{(s)}(x)] = \int_{-\infty}^{\infty} K(u) f^{(s)}(x+hu)du \tag{1}$$

The next step is to take a Taylor expansion of $f^{(s)}(x+hu)$ around $x+hu = x$, which is valid if $h \to 0$. We get

$$f^{(s)}(x+hu) = f^{(s)}(x) + f^{(s+1)}(x)hu + \frac{1}{2}f^{(s+2)}(x)h^2u^2 + ... + \frac{1}{P!}f^{(s+P)}(x)h^Pu^P + o(h^P).$$

Substituting this expression back into (1), integrating over each term, and using the fact that $\int_{-\infty}^{\infty} K(u)du = 1$ and the notation

$$\mu_\ell(K) = u^\ell K(u)$$

gives

$$\mathbb{E}[\hat{f}^{(s)}(x)] = f^{(s)}(x) + f^{(s+1)}(x)h\mu_1(K) + \frac{1}{2}f^{(s+2)}(x)h^2\mu_2(K) + ... + \frac{1}{P!}f^{(s+P)}(x)h^P\mu_P(K) + o(h^P).$$

Finally, noting that since K is a P-order kernel, $\mu_\ell(K) = 0$ for all $\ell < P$, gives the desired result

$$\mathbb{E}[\hat{f}^{(s)}(x)] = f^{(s)}(x) + \frac{1}{P!}f^{(s+P)}(x)h^P\mu_P(K) + o(h^P). \tag{2}$$

Next we consider the variance of the derivative estimator.

$$\begin{aligned}\mathbb{V}[\hat{f}^{(s)}(x)] &= \mathbb{V}\left[\frac{(-1)^s}{nh^{1+s}} \sum_{i=1}^{n} K^{(s)}\left(\frac{X_i - x}{h}\right)\right] \\ &= \frac{1}{nh^{2+2s}} \mathbb{V}\left[K^{(s)}\left(\frac{X_i - x}{h}\right)\right],\end{aligned}$$

since $\{X_i\}$ are iid there are no covariance terms and each term has the same variance. Continuing,

$$\begin{aligned}\mathbb{V}[\hat{f}^{(s)}(x)] &= \frac{1}{nh^{2+2s}} \left\{\mathbb{E}\left[K^{(s)}\left(\frac{X_i - x}{h}\right)^2\right] - \mathbb{E}\left[K^{(s)}\left(\frac{X_i - x}{h}\right)\right]^2\right\} \\ &= \frac{1}{nh^{2+2s}} \mathbb{E}\left[K^{(s)}\left(\frac{X_i - x}{h}\right)^2\right] - \frac{1}{n}\mathbb{E}\left[\frac{1}{h^{1+s}}K^{(s)}\left(\frac{X_i - x}{h}\right)\right]^2\end{aligned} \tag{3}$$

Now, from above we know that

$$\begin{aligned}\mathbb{E}\left[\frac{1}{h^{1+s}}K^{(s)}\left(\frac{X_i-x}{h}\right)\right] &= f^{(s)}(x)+\frac{1}{P!}f^{(s+P)}(x)h^P\mu_P(K)+o(h^P)\\ &= f^{(s)}(x)+o(1)\end{aligned}$$

since the remainder goes to zero as $h\to 0$. Thus, the second term in (3) is $O(\frac{1}{n})$; i.e. the same order as $1/n$. Furthermore $O(\frac{1}{n})$ is of smaller order than $O(\frac{1}{nh^{1+2s}})$ since $h\to 0$ and $n\to\infty$. Accordingly, we can write

$$\mathbb{V}[\hat{f}^{(s)}(x)] = \frac{1}{nh^{2+2s}}\mathbb{E}\left[K^{(s)}\left(\frac{X_i-x}{h}\right)^2\right]+o\left(\frac{1}{nh^{1+2s}}\right),$$

Thus,

$$\mathbb{V}[\hat{f}^{(s)}(x)] = \frac{1}{nh^{1+2s}}\int_{-\infty}^{\infty}\frac{1}{h}K^{(s)}\left(\frac{z-x}{h}\right)^2 f(z)dz+o\left(\frac{1}{nh^{1+2s}}\right)$$

Again we use the change of variables $u=\frac{z-x}{h}$ so that

$$\mathbb{V}[\hat{f}^{(s)}(x)] = \frac{1}{nh^{1+2s}}\int_{-\infty}^{\infty}K^{(s)}(u)^2 f(x+hu)du+o\left(\frac{1}{nh^{1+2s}}\right)$$

With the usual Taylor expansion of $f(x+hu)$ we can write

$$\begin{aligned}\mathbb{V}[\hat{f}^{(s)}(x)] &= \frac{1}{nh^{1+2s}}\int_{-\infty}^{\infty}K^{(s)}(u)^2(f(x)+O(h))du+o\left(\frac{1}{nh^{1+2s}}\right)\\ &= \frac{f(x)}{nh^{1+2s}}\int_{-\infty}^{\infty}K^{(s)}(u)^2du+o\left(\frac{1}{nh^{1+2s}}\right)\\ &= \frac{1}{nh^{1+2s}}f(x)\vartheta_s(K)+o\left(\frac{1}{nh^{1+2s}}\right),\end{aligned}$$

where $\vartheta_s(K)=\int_{-\infty}^{\infty}K^{(s)}(u)^2du$ as required.

1.2 Optimal bandwidth

We have

$$\begin{aligned}\text{AIMSE}[h] &= \int_{-\infty}^{\infty}\left[\left(h^P\mu_P(K)\cdot\frac{f^{(P+s)}(x)}{P!}\right)^2+\frac{1}{nh^{1+2s}}\vartheta_s(K)f(x)\right]dx\\ &= h^{2P}\left(\frac{\mu_P(K)}{P!}\right)^2\vartheta_{s+P}(f)+\frac{1}{nh^{1+2s}}\vartheta_s(K),\end{aligned}$$

since $f(x)$ integrates to 1 and where $\vartheta_{s+P}(f)=\int(f^{(P+s)}(x))^2dx$. Thus,

$$\begin{aligned}\frac{d}{dh}\text{AIMSE}[h] = 2Ph^{2P-1}\left(\frac{\mu_P(K)}{P!}\right)^2\vartheta_{s+P}(f)-(1+2s)\frac{1}{nh^{2+2s}}\vartheta_s(K)=0\\ \implies 2Ph^{1+2P+2s}\left(\frac{\mu_P(K)}{P!}\right)^2\vartheta_{s+P}(f)=(1+2s)\frac{1}{n}\vartheta_s(K),\end{aligned}$$

which gives the optimal bandwidth

$$h^*=\left[\frac{1+2s}{2Pn}\left(\frac{P!}{\mu_P(K)}\right)^2\frac{\vartheta_s(K)}{\vartheta_{s+P}(f)}\right]^{\frac{1}{1+2P+2s}}.$$

A fully data-driven method for estimating h^* is cross-validation. This procedure attempts to directly estimate the mean-squared error, and then choose the bandwidth which minimizes this estimate. From the lecture notes the cross-validation bandwidth is the value h which minimizes the criteria

$$\hat{h}_{CV}=\arg\min_h CV(h)=\frac{1}{n^2h}\sum_{i=1}^n\sum_{j=1}^n(K*K)\left(\frac{X_i-X_j}{h}\right)-\frac{2}{n}\sum_{i=1}^n\hat{f}_{(i)}(X_i)$$

where $\hat{f}_{(i)}(x_i)$ is the density estimate computed without observation X_i.

1.3 Monte Carlo experiment

(a) First, we want to compute the theoretically optimal bandwidth for $s=0$, $n=1000$, using the Epanechnikov kernel ($P=2$), with the following Gaussian DGP:

$$x_i\sim 0.5\mathcal{N}(-1.5,-1.5)+0.5\mathcal{N}(1,1)$$

From Table 1 in Hansen's notes, $\mu_2(K)=1/5$ and $\vartheta(K)=3/5$ for the Epanechnikov kernel. Thus, the only other ingredient we need is $\vartheta_2(f)=\int[f^{(2)}(x)]^2dx$ for the above DGP. Note that the second derivative of the normal density with mean μ and variance σ^2 is

$$\phi^{(2)}_{\mu,\sigma^2}(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{\sigma^2}\right)\left[\left(\frac{x-\mu}{\sigma^2}\right)^2-\frac{1}{\sigma^2}\right]$$

Since differentiation is a linear operation, we have

$$\vartheta_2(f) = \int_{-\infty}^{\infty} [0.5 \times \phi^{(2)}_{-1.5,1.5}(x) + 0.5 \times \phi^{(2)}_{1,1}(x)]^2 dx \approx 0.0388.$$

Finally, we get the theoretically optimal bandwidth

$$h^* = \left[\frac{1}{2 \times 2 \times 1000} \left(\frac{2!}{1/5} \right)^2 \frac{3/5}{\vartheta_2(f)} \right]^{\frac{1}{1+2\times 2}} \approx 0.827.$$

(b) I plot the IMSE estimates for the full-sample and leave-one-out sample below (see Appendix for the code).

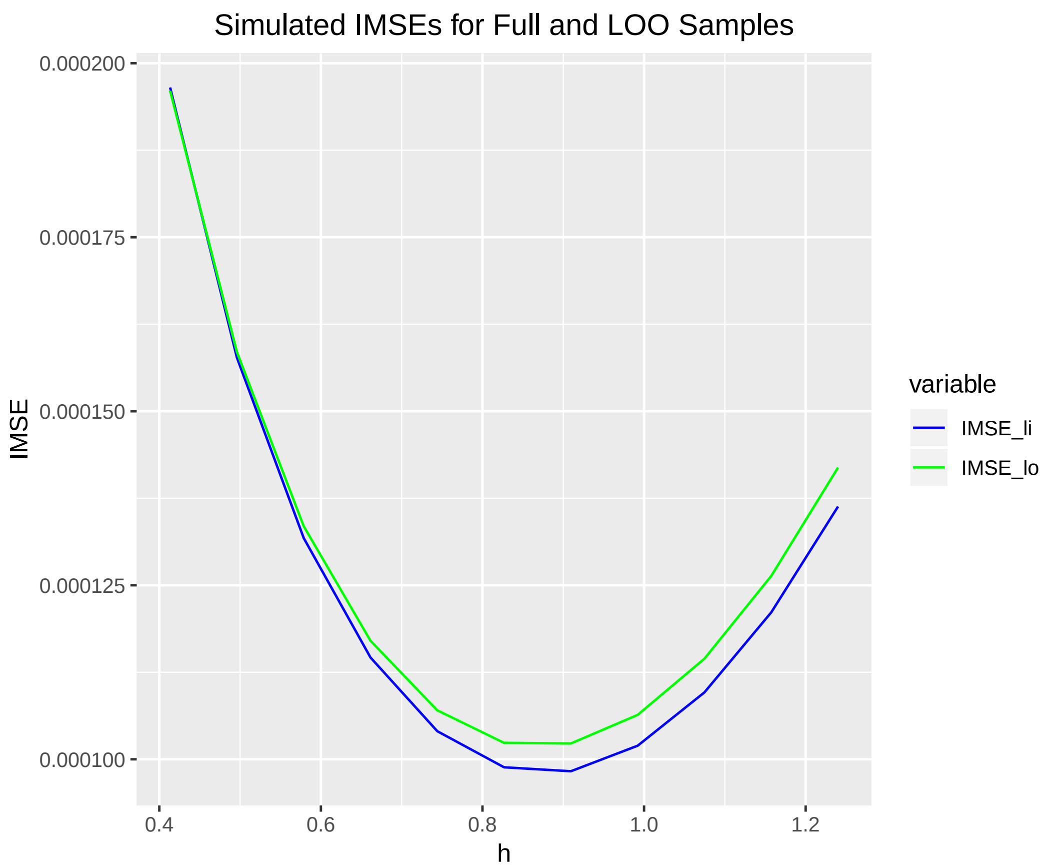

Figure 1: Estimated IMSE for $M = 1000$ simulations.

(c) Somewhat strangely, I find that $h_{\widehat{IMSE},LI} = h_{\widehat{IMSE},LO} = 1.1 \times h^*$. I suppose as we increase M, the estimates should converge to h^*.

(d) I get the following rule-of-thumb bandwidth

$$\bar{h}_{\texttt{AIMSE}} = \frac{1}{M} \sum_{i=1}^{M} \hat{h}_{\texttt{AIMSE},m} \approx 0.985,$$

which is about $1.2 \times h^*$.

2 Linear smoothers, cross-validation and series

2.1 Local polynomial and series estimation as linear smoothers

We are interested in estimating the regression function $e(x) = \mathbb{E}[y_i|x_i = x]$. The idea of local polynomial regression is to approximate $e(x)$ locally by a polynomial of degree p, and estimate this local approximation by weighted least squares. For each x we solve

$$\hat{\boldsymbol{\beta}}(x) = \arg\min_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} [y_i - \beta_0 - \beta_1(x_i - x) - \beta_2(x_i - x)^2 - ... - \beta_p(x_i - x)^p]^2 K\left(\frac{x_i - x}{h}\right).$$

where

$$\hat{e}(x) = \hat{\beta}_0$$

Note that this is motivated by a Taylor expansion of the true regression function $e(x_i)$ around x. And note that the kernel is just a 'smooth' way of weighting observations that are close to the evaluation point x.

More compactly, we write

$$\hat{\boldsymbol{\beta}}_{\text{LP}}(x) = \arg\min_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} [y_i - \boldsymbol{r}_p(x_i - x)'\boldsymbol{\beta}]^2 K\left(\frac{x_i - x}{h}\right)$$

where $\boldsymbol{r}_p(u) = (1, u, u^2, ..., u^p)'$.

From the lecture notes, we know that

$$\hat{\boldsymbol{\beta}}_{\text{LP}}(x) = (\boldsymbol{R}_p' \boldsymbol{W} \boldsymbol{R}_p)^{-1} \boldsymbol{R}_p' \boldsymbol{W} \boldsymbol{y}$$

where

$$\boldsymbol{R}_p = \begin{bmatrix} 1 & (x_1 - x) & (x_1 - x)^2 & \dots & (x_1 - x)^p \\ 1 & (x_2 - x) & (x_2 - x)^2 & \dots & (x_2 - x)^p \\ \vdots & \vdots & \dots & \ddots & \vdots \\ 1 & (x_n - x) & (x_n - x)^2 & \dots & (x_n - x)^p \end{bmatrix}$$

and $\boldsymbol{W} = \text{diag}\left(K\left(\frac{x_1 - x}{h}\right), K\left(\frac{x_2 - x}{h}\right), ..., K\left(\frac{x_n - x}{h}\right)\right)$.

Then

$$\begin{aligned} \hat{\boldsymbol{e}}(x) &= \boldsymbol{e}_1' \hat{\boldsymbol{\beta}}_{\text{LP}}(x) \\ &= \boldsymbol{e}_1' (\boldsymbol{R}_p' \boldsymbol{W} \boldsymbol{R}_p)^{-1} \boldsymbol{R}_p' \boldsymbol{W} \boldsymbol{y} \end{aligned}$$

where $\boldsymbol{e}_1$ the first standard basis vector of length $(1 + p)$ (i.e. it has a 1 in the first entry and zeros in the remaining p entries). I think in summation form we can write

$$\hat{\boldsymbol{e}}(x) = \boldsymbol{e}_1' (\sum_{i=1}^{n} \boldsymbol{r}_p(x_i - x) \boldsymbol{r}_p(x_i - x)' w_i)^{-1} (\sum_{i=1}^{n} \boldsymbol{r}_p(x_i - x) w_i y_i)$$

where $w_i = K\left(\frac{x_i - x}{h}\right)$.

Next we consider series estimation of the regression function $e(x)$. A series approximation to $e(x)$ is a global approximation, unlike the local polynomial regression. A series approximation that uses a polynomial basis (c.f. splines) takes the form

$$\hat{\boldsymbol{\beta}}_{\texttt{Series}} = \arg\min_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} (y_i - \boldsymbol{r}_p(x_i)'\boldsymbol{\beta})^2$$

where $\boldsymbol{r}_p(x_i) = (1, x_i, x_i^2, ..., x_i^p)$. And

$$\hat{e}(x) = \boldsymbol{r}_p(x)'\hat{\boldsymbol{\beta}}_{\texttt{Series}}$$

Accordingly, we have

$$\hat{\boldsymbol{\beta}}_{\texttt{Series}} = \left(\boldsymbol{R}_p'\boldsymbol{R}_p\right)^{-1} \boldsymbol{R}_p\boldsymbol{y}$$

where

$$\boldsymbol{R}_p = \begin{bmatrix} 1 & (x_1) & (x_1)^2 & \dots & (x_1)^p \\ 1 & (x_2) & (x_2)^2 & \dots & (x_2)^p \\ \vdots & \vdots & \dots & \ddots & \vdots \\ 1 & (x_n) & (x_n)^2 & \dots & (x_n)^p \end{bmatrix}$$

And,

$$\hat{e}(x) = \boldsymbol{r}_p(x)'\left(\boldsymbol{R}_p'\boldsymbol{R}_p\right)^{-1} \boldsymbol{R}_p\boldsymbol{y},$$

which is of the linear smoother form. In summation form

$$\hat{e}(x) = \boldsymbol{r}_p(x)'(\sum_{i=1}^{n} \boldsymbol{r}_p(x_i)\boldsymbol{r}_p(x_i)')^{-1}(\sum_{i=1}^{n} \boldsymbol{r}_p(x_i)y_i).$$

2.2 Cross validation

The idea of cross-validation is to choose the tuning parameter (e.g. bandwidth, etc.) that minimizes the mean squared leave-one-out error

$$\hat{c} = \arg\min_{c} \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{e}_{(i)}(x_i; c))^2$$

where $\hat{e}_{(i)}(x_i)$ is the estimator of the regression function that "leaves out" x_i.

From the above results we know that both the local polynomial and series estimators can be written as

$$\hat{\boldsymbol{e}}(x) = \boldsymbol{S}\boldsymbol{y}$$

where $\boldsymbol{S}$ is the 'smoothing' matrix. Note that for local polynomial and series estimators the smoothing matrix is constant preserving in the sense $\boldsymbol{S1} = \boldsymbol{1}$. That is, the rows of $\boldsymbol{S}$ sum to one. In leave-one-out cross validation, we want to use the same smoother with the i-th row and column deleted; we also want this to be an $(n-1) \times (n-1)$ smoother matrix. Accordingly, we must renormalize the rows to sum to one. Let w_{ij} denote the elements of $\boldsymbol{S}$. When we delete the i-th column, then the i-th row now sums to $1 - w_{ii}$. So, we divide by $1 - w_{ii}$ to renormalize. Accordingly, the leave-one-out estimator is

$$\hat{e}_{(i)}(x_i) = \frac{1}{1 - w_{ii}} \sum_{j=1, j \neq i}^{n} w_{ij} y_i$$

And note that the full-sample estimator is just

$$\hat{e}(x_i) = \sum_{j=1}^{n} w_{ij} y_i.$$

From the above expression we get

$$\begin{aligned}
\hat{e}_{(i)}(x_i)(1 - w_{ii}) &= \sum_{j=1, j \neq i}^{n} w_{ij} y_i \\
\hat{e}_{(i)}(x_i) &= \sum_{j=1, j \neq i}^{n} w_{ij} y_i + w_{ii} \hat{e}_{(i)}(x_i) \\
&= \sum_{j=1}^{n} w_{ij} y_i + w_{ii} \hat{e}_{(i)}(x_i) - w_{ii} y_i \\
&= \hat{e}(x_i) + w_{ii} \hat{e}_{(i)}(x_i) - w_{ii} y_i \\
\implies y_i - \hat{e}_{(i)}(x_i) &= y_i - \hat{e}(x_i) - w_{ii} \hat{e}_{(i)}(x_i) + w_{ii} y_i \\
&= y_i - \hat{e}(x_i) + w_{ii}(y_i - \hat{e}_{(i)}(x_i)) \\
\therefore y_i - \hat{e}_{(i)}(x_i) &= \frac{1}{1 - w_{ii}} (y_i - \hat{e}(x_i)),
\end{aligned}$$

which gives the desired result.

2.3 Asymptotic distribution

First note that we have iid data. Also note that we must have $\sum_{i=1}^{n} w_{n,i}(x_i) = 1$. To ease notation, denote $\mathbb{E}[\cdot|x_1, x_2, ..., x_n; x]$ as $\mathbb{E}[\cdot|x]$. Then

$$\begin{aligned}
\mathbb{E}[\hat{e}(x)|x] &= \mathbb{E}[\sum_{i=1}^{n} w_{n,i}(x_i) y_i | x] \\
&= \sum_{i=1}^{n} \mathbb{E}[w_{n,i}(x_i) y_i | x] \\
&= \sum_{i=1}^{n} w_{n,i}(x_i) \mathbb{E}[y_i | x] \\
&= \mathbb{E}[y_i | x].
\end{aligned}$$

Thus, so long as $\hat{e}(x)$ has a finite second moment we can use the classical CLT to get asymptotic normality. Now,

$$\begin{aligned}\mathbb{V}[\hat{e}(x)|x] &= \mathbb{V}[\sum_{i=1}^{n} w_{n,i}(x) y_i | x] \\ &= \sum_{i=1}^{n} \mathbb{V}[w_{n,i}(x) y_i | x] \\ &= \mathbb{V}[y_i|x] \sum_{i=1}^{n} w_{n,i}(x)^2\end{aligned}$$

Then we get the consistent variance estimator

$$\hat{V}(x) = \hat{\sigma}^2 \sum_{i=1}^{n} w_{n,i}(x)^2$$

where $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \hat{e}(x_i))^2$

2.4 Confidence interval

The pointwise asymptotically valid 95% CI for $e(x)$ is

$$CI_{95}(x) = [\hat{e}(x) - 1.96\sqrt{\hat{V}(x)}, \hat{e}(x) + 1.96 \cdot \sqrt{\hat{V}(x)}].$$

This is clearly different to a confidence band that is uniformly valid over all x. Uniform confidence bands would be specified as

$$\sup_{x \in \chi} \left| \frac{\hat{e}(x) - e(x)}{\sqrt{\hat{\mathbb{V}}(x)}} \right| \leq q_{1-\alpha/2},$$

which is clearly a harder problem than the pointwise intervals.

3 Semiparametric semi-linear model

We have the partially linear model

$$y_i = t_i\theta_0 + g_0(\boldsymbol{x_i}) + \epsilon_i, \tag{4}$$

with the usual heteroskedasticity assumptions for the error.

3.1

From Li-Racine 7.1.1 (p 222) we know that for θ_0 to be identifiable, t_i must not contain a constant (since t_i is a treatment dummy, this is clearly satisfied) or any deterministic functions of $\boldsymbol{x}_i$. Now, somehow we need to show

$$\mathbb{E}[(t_i - h_0(\boldsymbol{x}_i))(y_i - t_i\theta_0)] = 0$$

Then we have

$$\begin{aligned}
\mathbb{E}[y_i(t_i - h_0(\boldsymbol{x}_i)) - t_i\theta_0(t_i - h_0(\boldsymbol{x}_i))] &= 0 \\
\mathbb{E}[t_i\theta_0(t_i - h_0(\boldsymbol{x}_i))] &= \mathbb{E}[y_i(t_i - h_0(\boldsymbol{x}_i))] \\
\therefore \theta_0 &= \mathbb{E}[t_i(t_i - h_0(\boldsymbol{x}_i))]^{-1}\mathbb{E}[y_i(t_i - h_0(\boldsymbol{x}_i))].
\end{aligned}$$

The IV interpretation is that we are using $t_i - h_0(\boldsymbol{x}_i)$ as an instrument for t_i.