Introduction to Econometrics II: Recitation 9*

Seung-hun Lee†

April 8th, 2020

1 Quantile Regression

The usual linear regression, which has the form

$$y = X\beta + u$$

where $E(Xu) = 0$ (or more restrictively, $E(u|X) = 0$) is aimed at capturing the conditional mean of y given X at a certain value. However, there is no reason to restrict our attention to just a conditional mean. In fact, we an do more. For instance, what is the conditional median? What about the top 10%, or 1%? **Quantile regression** aims to capture different values of β depending on the location of the conditional distribution.

Rigorously speaking, the quantile regression seeks to estimate the conditional quantile

$$q_\tau(y|X) = X\beta_\tau$$

where $\tau \in [0, 1]$ is the percentile of our choice satisfying $F_{y|X}(X\beta_\tau|X) = \tau$. To characterize the conditional quantile function, we need to use the **check function**. Since we have

$$F_{y|X}(X\beta_\tau|X) = \Pr(y \leq X\beta_\tau|X) = \tau$$

we can write as

$$\tau - \Pr(y \leq X\beta_\tau|X) = 0$$

*This is based on the lecture notes of Professors Jushan Bai and Bernard Salanie. I was also greatly helped by previous recitation notes from Paul Sungwook Koh and Dong Woo Hahm. The remaining errors are mine.

†Contact me at sl4436@columbia.edu if you spot any errors or have suggestions on improving this note.

Since $\Pr(y \leq X\beta_\tau|X)$ is equal to $E[1(y - X\beta_\tau \leq 0)] = E[1(u \leq 0)]$, we can obtain the condition

$$E[\tau - 1(y - X\beta_\tau \leq 0)|X] = 0$$

and using the law of iterated expectations, this also implies $E[(\tau - 1(y - X\beta_\tau \leq 0))X] = 0$. The check function can be defined as

$$\rho_\tau(u) = u(\tau - 1(u \leq 0))$$

To see what this is, I will talk about two specific cases

- Median: Let $\tau = 1/2$. Then the check function becomes

$$\rho_{1/2}(u) = \begin{cases} -\frac{1}{2}u & (u \leq 0) \\ \frac{1}{2}u & (u > 0) \end{cases} = \frac{1}{2}|u| = \frac{1}{2}|y - X\beta_{1/2}|$$

  This becomes equivalent to solving the least absolute deviation problem.

- $\tau = 1/3$: Then the check function becomes

$$\rho_{1/3}(u) = \begin{cases} -\frac{2}{3}u & (u \leq 0) \\ \frac{1}{3}u & (u > 0) \end{cases}$$

  which has a kink at $u = 0$ and is asymmetric.

For all values of τ the check function has a kink at $u = 0$. However, the check function is convex and can be shown to be a lipschitz function.

Theorem 1.1 (Convex Functions are Lipschitz Functions?)**.** *Let f be a convex function and K be a closed, bounded set contained in the relative interior of the domain of f. Then f is Lipschitz continuous on K.*

Proof. Suppose that f is convex on $[a, b]$. Select c, d s.t. $a < c < d < b$ and let $t_1 \in [d, b]$, $t_2 \in [a, c]$ and $x_1, x_2 \in [c, d]$. Since f is convex we can write

$$\frac{f(c) - f(t_2)}{c - t_2} \leq \frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq \frac{f(t_1) - f(d)}{t_1 - d}$$

Then, let $L = \max\left[\left|\frac{f(c)-f(t_2)}{c-t_2}\right|, \left|\frac{f(t_1)-f(d)}{t_1-d}\right|\right]$, we then have

$$|f(x_2) - f(x_1)| \leq L|x_2 - x_1|$$

□

The minimization problem that should be solved in quantile regression is

$$\min_{\beta} E[\rho_\tau(y - X\beta_\tau)|X]$$

which can be written as

$$E[\rho_\tau(y - X_i\beta_\tau)|X] = (\tau - 1)\int_{-\infty}^{a}(y-a)f_{Y|X}(y|x)dy + \tau\int_{a}^{\infty}(y-a)f_{Y|X}(y|x)dy$$

where $a = X\beta_\tau$. Take the first order condition w.r.t. a to get

$$\begin{aligned} -(\tau-1)\int_{-\infty}^{a} f_{Y|X}(y|x)dy - \tau\int_{a}^{\infty} f_{Y|X}(y|x)dy = 0 \\ \iff -\tau + F(a|X) = 0 \end{aligned}$$

Thus, the β_τ that solves

$$X\beta_\tau = a = F^{-1}_{Y|X}(\tau|X)$$

is the β_τ that we are looking for. We can also solve for

$$\min_{\beta} E[\rho_\tau(y - X\beta_\tau)X]$$

or in a sample analogue

$$\min_{\beta} \frac{1}{n}\sum_{i=1}^{n}\rho_\tau(y_i - X_i\beta_\tau)X_i$$

to find the quantile estimator β_τ. For suitable conditions, the resulting estimator is CAN.

Example 1.1 (Levin, 2001)**.** *As part of the education production function literature, the effect of class size on various outcomes for the student is controversial (Lazear, 2001). This paper, while controlling for a large number of observable characteristics and endogeneity in class size variable,*

estimates education production function using a quantile regression. In doing so, it attempts to analyze the conventional claims that reducing class size improves learning directly via and increased allocation of teaching per student. While it does not find statistically significant impact of class size once peer effect is controlled for, peer effect has a positive and significant effect for those in the lower portion of the achievement distribution for math and language scores.

Example 1.2 (Autor, Houseman, & Kerr 2017[a])**.** *Using the Detroit's welfare-to-work program, this paper studies the the effect of two government employment programs - direct hire assistance and temporary-help job placements - on distribution of participant's earnings over a 7-quarter period. The paper finds that for the low-tail of the earnings distribution, neither programs are effective. The direct-hire increases earnings for the high-tail but temporary-help placements negatively affect the distribution for the same group. Autor and Houseman (2010) have studied the same program 7 years ago without using the quantile approach and find that on average, the same program lead to earnings gain. The key takeaway is that by using quantile regression, you can unmask the effect at a different quantile that you cannot find out through conditional expectations.*

[a] Technically, this paper uses Chernozhukov-Hansen Instumental Variables Quantile Regression, which is not part of our curriculum yet. You may learn this in Microeconometrics course next year.

2 Nonparametric Regression Models

Consider a setting where we observe the data (y_i, x_i) which is i.i.d. and drawn from a DGP $P_0(y|x)$. Unlike in the parametric setup, we are interested in backing out the whole or part of the data generating process $P_0(y|x)$ **without any modeling assumptions**. In essence, we are interested in an estimation problem involving an unknown function. This approach is called a **nonparametric** approach. We normally use nonparametric approach to conduct a diagnostic checking of an estimated parametric model, to conveniently display key features of the dataset, and to conduct an inference under very weak assumptions.

For discrete-valued Y and X, this is relatively easy. We can back out the data generating process of interest

$$\hat{P}(y \in A | x \in B) = \frac{n^{-1} \sum_{i=1}^{n} 1(y_i \in A, x_i \in B)}{n^{-1} \sum_{i=1}^{n} 1(x_i \in B)}$$

Provided that $P_0(x \in B) \neq 0$, then the above estimator converges to the true $P_0(y|x)$ as $n \to \infty$.

When we are working with continuous variables to get an estimate of $P_0(Y \leq y|x)$, one way we could approach is to define A in the above example as $A \equiv (-\infty, y], B \equiv [x - \epsilon, x + \epsilon]$

and use $\lim_{\epsilon \to 0} \hat{P}(A|B)$ to back out the DGP. This can be problematic when we have too many dimensions of X (and by too many, the number required for this to happen is not that large). Therefore, the method to conduct nonparametric estimation in this context is kernel density function approach.

2.1 Kernel Density Estimation

Consider the problem of estimating the probability density function $f(x)$ of a random scalar variable X at $X = x$. We are putting a minimal number of restriction on what $f(x)$ could be. Let $\{X_1, ..., X_n\}$ be a random sample of X. If f is a smooth function, we can approximate f by

$$f(x) \simeq \frac{\int_{x-h}^{x+h} f(u)du}{2h}$$

for a small h (which will later be our bandwith). The numerator on the right hand side can be estimated using a sample analogue of the following form

$$\frac{1}{n} \sum_{i=1}^{n} 1\{x - h \leq X_i \leq x + h\}$$

Combining the two expressions, we can approximate $f(x)$ with

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

where $K(u) = \frac{1}{2} 1(|u| \leq 1)$ and $K_h = \frac{1}{h} K\left(\frac{\cdot}{h}\right)$. In general, $K(\cdot)$ is a **kernel density estimator**. Specifically for this example, I have used a uniform kernel.

As for the type of kernel density estimators, there can be many. The requirements are that K is a real-to-real function, positive and smooth and integrates to 1 on its support. For instance, you can use a Gaussian Kernel, $K = \phi$ or an Epanechinkov kernel, defined as $K(u) = \frac{3}{4} \max\{1 - u^2, 0\}$. However, the selection of K does not matter in most cases.

2.2 Choice of the bandwidth h

While we may have some freedom in choosing K, same cannot be said for the bandwidth h. This is because our choice of h directly affects the trade-off between the bias and the variance. To see this.

$$\begin{aligned}
E[\hat{f}(x)] &= E\left[\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\right] \\
&= \frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{x-s}{h}\right) f(s)ds \\
&= -\frac{1}{h}\int_{-\infty}^{\infty} K(t)f(x-ht)(-hdt) \ (\because \frac{x-s}{h} = t \text{ transformation}) \\
&= \int_{-\infty}^{\infty} K(t)f(x-ht)dt \\
&= \int_{-\infty}^{\infty} K(t)\left[f(x) - f'(x)ht + \frac{f''(x)h^2t^2}{2} + o(h^2)\right]dt \\
&= f(x) - 0 + \frac{1}{2}\int_{-\infty}^{\infty} K(t)h^2t^2f''(x)dt + o(h^2)
\end{aligned}$$

Where $\int_{-\infty}^{\infty} K(t)dt = 1$ justifies $f(x)$ and $\int_{-\infty}^{\infty} tK(t)dt = 0$ (since kernel is symmetric around 0) justifies second term in the last line being 0.[1] As such, the bias is captured by

$$E[\hat{f}(x)] - f(x) = \frac{1}{2}\int_{-\infty}^{\infty} K(t)h^2t^2f''(x)dt$$

Therefore, smaller bandwidth h reduces the bias.

However, for variance, we have

$$\begin{aligned}
Var[\hat{f}(x)] &= E[\hat{f}^2(x)] - (E[\hat{f}(x)])^2 \\
&= E\left[\frac{1}{n^2h^2}\left(\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\right)^2\right] - (E[\hat{f}(x)])^2 \\
&= E\left[\frac{1}{n^2h^2}\left(\sum_{i=1}^{n} K^2\left(\frac{x-X_i}{h}\right) + 2\sum_{i<j} K\left(\frac{x-X_i}{h}\right)K\left(\frac{x-X_j}{h}\right)\right)\right] - (E[\hat{f}(x)])^2 \\
&= \frac{1}{nh^2}\int_{-\infty}^{\infty} K^2\left(\frac{x-s}{h}\right)f(s)ds + \frac{n(n-1)}{n^2h^2}\left(\int_{-\infty}^{\infty} K\left(\frac{x-s}{h}\right)f(s)ds\right)^2 - \frac{1}{h^2}\left(\int_{-\infty}^{\infty} K\left(\frac{x-s}{h}\right)f(s)ds\right)^2
\end{aligned}$$

Then, we use the Taylor expansion and the variable transformation $\frac{x-s}{h} = t$ on each of the terms. As a result, we can show that the first term can be written

$$\frac{1}{nh^2}\int_{-\infty}^{\infty} K^2\left(\frac{x-s}{h}\right)f(s)ds = \frac{1}{nh}\int_{-\infty}^{\infty} K^2(t)f(x-ht)dt \simeq \frac{1}{nh}\int_{-\infty}^{\infty} K^2(t)f(x)dt$$

[1] In the lecture slides, $O(h^2) = \frac{f''(x)h^2t^2}{2} + o(h^2)$

so that $\frac{1}{nh}\int_{-\infty}^{\infty} K^2(t)f(x)dt$ is the leading term. The other two terms will be cancelled out as $n \to \infty$, as

$$\left(\frac{n(n-1)}{n^2h^2} - \frac{1}{h^2}\right)\left(\int_{-\infty}^{\infty} K\left(\frac{x-s}{h}\right)f(s)ds\right)^2 = -\frac{1}{nh^2}\left(\int_{-\infty}^{\infty} K\left(\frac{x-s}{h}\right)f(s)ds\right)^2$$

Since the first term can be written as $O\left(\frac{1}{nh}\right)$, smaller h increases the variance.

So what determines the optimal h? The answer is the h that minimizes the loss function, or asymptotic mean integrated squared error (AMISE), defined as

$$\int E(\hat{f}(x) - f(x))^2 dx$$

where the bias-variance decomposition of MSE kicks in as follows

$$\begin{aligned}
E[(\hat{f}(x) - f(x))^2] &= E[(\hat{f}(x) - E(\hat{f}(x)) + E(\hat{f}(x)) - f(x))^2] \\
&= E[(\hat{f}(x) - E(\hat{f}(x)))^2 + (E(\hat{f}(x)) - f(x))^2 + 2(\hat{f}(x) - E(\hat{f}(x)))(E(\hat{f}(x)) - f(x))] \\
&= E[(\hat{f}(x) - E(\hat{f}(x)))^2] + E[(E(\hat{f}(x)) - f(x))^2] \\
&= \text{Variance} + \text{Bias}^2
\end{aligned}$$

From the discussion about the bias and variance above, we can write

$$\text{Variance} + \text{Bias}^2 = \frac{1}{nh}\int_{-\infty}^{\infty} K^2(t)f(x)dt + \frac{h^4}{4}(f''(x))^2\left(\int_{-\infty}^{\infty} K(t)t^2dt\right)^2$$

Thus, the final version of AMISE can be written as

$$\int(\text{Variance} + \text{Bias}^2)dx = \frac{1}{nh}\int_{-\infty}^{\infty} K^2(t)dt + \frac{h^4}{4}\int_{-\infty}^{\infty}(f''(x))^2\left(\int_{-\infty}^{\infty} K(t)t^2dt\right)^2 dx$$

For a shorthand notation, we can write $A = \frac{1}{4}\int_{-\infty}^{\infty}(f''(x))^2\left(\int_{-\infty}^{\infty} K(t)t^2dt\right)^2 dx$, $B = \int_{-\infty}^{\infty} K^2(t)dt$ to get

$$AMISE = Ah^4 + \frac{B}{nh}$$

Then, the minimization problem becomes $\min_h AMISE$. Therefore, we find h satisfying

$$4Ah^3 - Bn^{-1}h^{-2} = 0 \iff h^5 = \frac{B}{4An} \iff h = \left(\frac{B}{4An}\right)^{1/5}$$

In this framework, the bias and standard errors are both in $n^{-2/5}$ and AMISE will be in $n^{-4/5}$.

Therefore, we may not have a CAN estimator at $n^{-1/2}$, which is what is normally the case. Even bigger problem arises from $\int f''(x)dx$ term in A, as we are not sure of $f(x)$ to begin with, we do not know what $f''(x)$ would be. There are several ways to deal with this.

- **Silverman's Rule of Thumb**: If K is a normal kernel and $f(x)$ is normal with some variance σ^2, Silverman's rule of thumb implies the benchmark

$$h = 1.06\sigma n^{-1/5}$$

  where σ can be replaced with sample standard deviation s. For a more robust version, we can use

$$h = 0.9 \min\{s, IQ/1.34\}n^{-1/5}$$

  where IQ is an interquartile distance.

- **Cross-validation**: This method is similar to the original method in the sense that we are selecting some h that minimizes certain criterion function, $CV(h)$. In this case, we are using a leave-one-out estimator. This is usually calculated as

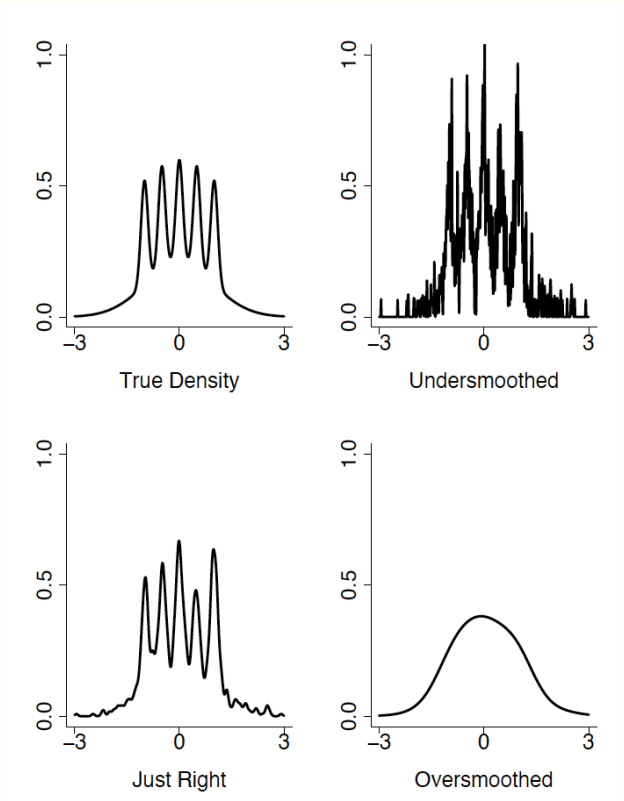
$$\hat{f}_{-i}(x) = \frac{1}{nh}\sum_{j\neq i} K\left(\frac{x - x_j}{h}\right)$$

  for every observation i. This is computationally burdensome and the resulting h converges very slowly ($n^{-1/10}$).

- **Local Bandwidth**: You may care only about $f(x)$ at a given point. The key idea is to make h larger in a low density area, as low density could imply larger spread in the distribution of $f(x)$. Making h larger would decrease variance. In a wiggly region, it is better to take h smaller. The distribution can take many values and bias can be problematic. Reducing h would minimize worries about biases.

Comment 2.1 (So why worry about bandwidth to begin with?)**.** *If h is smaller than the optimal bandwidth length, we say that there is an **undersmoothing**. If otherwise, we say **oversmoothing** is present. When there is an undersmoothing, there are too much more detail than necessary and in some areas, leading to computational difficulty. When there is an oversmoothing, we risk masking some important detail about the distribution of $f(x)$. We may end up with a normal distribution even if it really is not.*

Below is a figure that I borrowed from a statistics lecture note from CMU.[a]



[a] https://www.stat.cmu.edu/ larry/=sml/densityestimation.pdf

2.3 Curse of Dimensionality

Now assume that x is not necessarily a scalar, but of dimension d. Then we can use a d-dimensional kernel K and estimate $f(x)$ with

$$\hat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

where K can be a d-product of a uni-dimensional kernels. In addition, we are not necessarily confined to using a same bandwith for all d kernels. For instance, if you know that the variance of x_1 is larger than x_2 you may want higher h_1 relative to h_2 to avoid variance of the kernel density estimator for the first covariate from being too large. We may want to sphericize if the kernel estimators are correlated.

A bigger concern has to do with the computation cost from using a large d. This is what **curse of dimensionality** refers to. If we return to calculating $E[\hat{f}(x)^2]$ term in the variance

component of AMISE,

$$\begin{aligned}
E[\hat{f}(x)^2] &= E\left[\frac{1}{n^2h^{2d}}\left(\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\right)^2\right] \\
&= E\left[\frac{1}{n^2h^{2d}}\left(\sum_{i=1}^{n} K^2\left(\frac{x-X_i}{h}\right) + 2\sum_{i<j} K\left(\frac{x-X_i}{h}\right)K\left(\frac{x-X_j}{h}\right)\right)\right] \\
&= \frac{1}{nh^{2d}}\int_{-\infty}^{\infty} K^2\left(\frac{x-s}{h}\right)f(s)ds + \frac{n(n-1)}{n^2h^{2d}}\left(\int_{-\infty}^{\infty} K\left(\frac{x-s}{h}\right)f(s)ds\right)^2
\end{aligned}$$

Again, the leading term is

$$\frac{1}{nh^{2d}}\int_{-\infty}^{\infty} K^2\left(\frac{x-s}{h}\right)f(s)ds \simeq \frac{1}{nh^d}\int K^2(t)f(x-ht)dt$$

since s is actually d-dimensional. Thus, the variance is now $O\left(\frac{1}{nh^d}\right)$. What this means is that $AMISE = Ah^4 + \frac{B}{nh^d}$ and the optimal h is solved as

$$h = \left(\frac{Bd}{4An}\right)^{1/(4+d)}$$

Thus, h will be in $n^{-\frac{1}{4+d}}$ and convergence occurs in $n^{-\frac{2}{4+d}}$ - at an even slower rate that when $d = 1$, which was already slower than parametric convergence rate. AMISE would be in order $n^{-\frac{4}{4+d}}$

2.4 Nadaraya-Watson Estimation

Given the data (y_i, x_i), we are attempting to capture $E[g(y,x)|x] = m(x)$ for some $g(y,x)$. For instance, we can attempt to estimate the conditional expectation by letting $g(y,x) = y$. Note that the conditional expected value can be written as

$$\begin{aligned}
E[y|x] &= \int y f_{Y|X}(y|x)dy \\
&= \int y\frac{f_{Y,X}(y,x)}{f_X(x)}dy = \frac{\int y f_{Y,X}(y,x)dy}{\int f_{Y,X}(y,x)dy}
\end{aligned}$$

We can obtain the nonparametric estimator for the conditional expected value by replacing $f_{Y,X}$ with its kernel estimator. For simplicity, I will assume that the dimension of both y and

x is 1. The numerator becomes

$$\begin{aligned}\int y\hat{f}(y,x)dy &= \int y\frac{1}{nh^2}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)K\left(\frac{y-Y_i}{h}\right)dy \\ &= \frac{1}{nh^2}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)\int yK\left(\frac{y-Y_i}{h}\right)dy \\ &= \frac{1}{nh^2}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)\int (Y_i+sh)K(s)\,(hds)\ (\because s=\frac{y-Y_i}{h}) \\ &= \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)Y_i\ (\because \int K(s)ds=1, \int sK(s)ds=0)\end{aligned}$$

The denominator can be written as $\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)$. Thus, the estimator for the conditional expectation becomes

$$\frac{\sum_{i=1}^{n}Y_iK\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)}$$

Effectively we are putting weight $\frac{K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)}$ on each observation Y_i. This is also a **local constant estimation** in the sense that when solving the following minimization problem

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

The first order condition on a yields the following results

$$\sum_{i=1}^{n}Y_iK\left(\frac{x-X_i}{h}\right)=a\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)\implies a=\frac{\sum_{i=1}^{n}Y_iK\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)}$$

As for the optimal h, it is possible to use cross-validation method and minimizing AMISE. This method can be unreliable if $f(x)\to 0$.

We can also do something more general. For instance, a **local linear estimation**. This is when we regress $g(y,x)$ on a constant (a) and a linear term ($b(x-X_i)$). In mathematical expression, we solve

$$\min_{a,b}\frac{1}{nh}\sum_{i=1}^{n}(Y_i-a-b(x-X_i))^2K\left(\frac{x-X_i}{h}\right)$$

and obtain that $\hat{a}=\hat{g}$ and $\hat{b}$ is an estimate of $\frac{\partial g(x)}{\partial x}$. If it happens that the true functional

form is linear, then this estimate does not produce a bias. In addition, local linear estimation performs better than local constant estimation in the boundaries of the support for X.

We can do even more with **local polynomial estiation** by regressing $g(y, x)$ on a constant, $x - X_i$, $(x - X_i)^2$ and so on.

2.5 Semi-nonparametrics

- Flexible approach: Suppose that we are sure that $f(x)$ can be characterized by $f_{m,\sigma}$ where m, σ indexes some properties of the density function f. Then, by Weierstrass approximation theorem, we can choose a family of positive functions which increases in complexity $P_\theta^1, P_\theta^2, ...$ and maximize over the loglikelihood

$$\sum_{i=1}^{n} \log f_{m,\sigma}(X_i) P_\theta^M(X_i)$$

> **Theorem 2.1** (Weierstrass Approximation Theorem)**.** *If f is a continuous real-valued function on $[a, b]$ and if any $\epsilon > 0$ is given, then there exists a polynomial P on $[a, b]$ s.t.*
>
> $$|f(x) - P(x)| < \epsilon$$
>
> *for all $x \in [a, b]$*

- Mixture of normals: Suppose that $Y|X$ is drawn from the two distributions
  - $N_1(m_1(x, \theta), \sigma_1^2(x, \theta))$ with probability $q_1(x, \theta)$
  - $N_2(m_2(x, \theta).\sigma_2^2(x, \theta))$ with probability $q_2(x, \theta)$

  Then, we apply a maximum likelihood of the following form

$$\min_\theta \sum_i \sum_k q_k(x_i, \theta)[(y_i - m_k(x_i, \theta))' \sigma_k(x_i, \theta)^{-1}(y_i - m_k(x_i, \theta)) + \log\det \sigma_k(x_i, \theta)]$$

  As with other MLE estimators, there could be more than one local maxima. Furthermore, the optimal number of k is difficult to determine in practice.

- Series estimator: Let $\{P_k(x_i)|k = 1, 2...\}$ be the orthonormal basis for a smooth function.

By orthonormal, it means that

$$\int P_k(x)^2 dx = 1, \int P_k(x)P_m(x) = 0 \ (k \neq m)$$

These could be polynomials of degree k, sine functions and so on. What we do here is to run a linear regression that has the following form

$$y_i = \sum_{k=1}^{M} P_k(x_i)\theta_k + \epsilon_i$$

The $\sum_{k=1}^{M} P_k(x_i)\theta_k$ part is a series approximation to $g(x)$. However, depending on the number of M that we choose, the curse of dimensionality can kick in.