Logistics

Quick note: My name is pronounced as “ling thô”

- **Timing**
  - 10 lectures
  - In-class exam on April 24, 2018 (second half only)

- **Sections**
  - Materials: The course is new this year, and we’ll be learning together as we go along. We have limited time in sections compared to lectures, and therefore we will focus on concepts and the problem sets.
  - Section notes summarize the basic concepts, but in-person time will focus more on clarification and examples that may be outside of these notes.
  - The course is not set in stone, so section format will change. If you find any mistakes in the notes, definitely share and discuss with me and the class.
  - Your feedback: Any comments that you’d like to share on this course are very valuable. Please let Elie and me know, and letting us know earlier will give us time to adjust to make the course better for you.

- **Problem sets**
  - The grader will pick up the problem sets from my mailbox by end of the due date.
  - Group work: up to 3 students per group, though it’s also a good opportunity to practice for the exam so make sure you understand all the questions.
1 Preliminaries: Probability Review

1.1 Random Variable

**Definition** (Random variable). A random variable is a function from a sample space into the real number \( \mathbb{R} \).

**Example.** Experiment: Toss a fair coin 3 times

- Sample space: \( s \equiv \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\} \)
- Random variable: \( X(s) \) defined as the number of heads obtained

The definition above can be extended to a *random vector* which is a mapping from the sample space into the \( n \)-dimensional Euclidean space \( \mathbb{R}^n \).

1.2 Convergence of Random Variables

We consider the convergence of a sequence of random variables \( X_1, X_2, \ldots, X_n \) as the sample size \( n \) approaches infinity. The following definitions generally go from a weaker concept (easier to obtain) to a stronger one (which will imply the one weaker than it).

**Definition** (Convergence in distribution: \( X_n \xrightarrow{d} X \)). The sequence \( \{X_n\} \) of random variables converges in distribution to the random variable \( X \) if \( \lim_{n \to \infty} F_n(x) = F(x) \) for every \( x \in \mathbb{R} \) at which \( F \) is continuous. Here \( F_n \) and \( F \) are the cumulative distribution functions of random variables \( X_n \) and \( X \), respectively.

**Definition** (Convergence in probability: \( X_n \xrightarrow{p} X \)). The sequence \( \{X_n\} \) of random variables converges in distribution to the random variable \( X \) if for all \( \epsilon > 0 \) we have \( \lim_{n \to \infty} \Pr(|X_n - X| > \epsilon) = 0 \).

**Definition** (Convergence almost surely: \( X_n \xrightarrow{a.s.} X \)). The sequence \( \{X_n\} \) of random variables converges almost surely (or almost everywhere, or with probability 1, or strongly) towards \( X \) means that \( \Pr(\lim_{n \to \infty} X_n = X) = 1 \).

**Definition** (Convergence in \( r \)-th mean: \( X_n \xrightarrow{L^r} X \)). Given \( r \geq 1 \), the sequence \( \{X_n\} \) of random variables converges in the \( r \)-th mean (or in the \( L^r \)-norm) towards the random variable \( X \) if the \( r \)-th absolute moments \( \mathbb{E}[|X_n|^r] \) and \( \mathbb{E}[|X|^r] \) of \( X_n \) and \( X \) exist and \( \lim_{n \to \infty} \mathbb{E}[|X_n - X|^r] = 0 \).

- When \( r = 1 \), we say that \( X_n \) converges in mean to \( X \).
- When \( r = 2 \), we say that \( X_n \) converges in mean square to \( X \).
- Convergence in \( r \)-th mean implies convergence in \( s \)-th mean if \( r > s \geq 1 \).
• Convergence in $r$-th mean implies convergences in probability if $r \geq 1$.

**Note:** $a.s \Rightarrow p \Rightarrow d$ and mean/mean square $\Rightarrow p \Rightarrow d$. For a constant $c$, $X_n \xrightarrow{d} c \iff X_n \xrightarrow{p} c$.

**Definition** (Order in probability notation). Direct counterparts of big $O$ and small $O$ for sequence of real numbers:

- Small $O$ (convergence in probability): $X_n = o_p(b_n)$ if $x_n b_n \neq c_0$.
- Big $O$ (stochastic boundedness): $X_n = O_p(b_n)$ if there exists $M_\varepsilon < \infty$ and a finite $N_\varepsilon > 0$ such that $Pr\left(\left|\frac{X_n}{b_n}\right| > M_\varepsilon\right) < \varepsilon, \forall n > N_\varepsilon$.

If a sequence is $o_p(1)$ then it is $O_p(1)$ but the reverse does not hold.

**Definition** (Uniform convergence in probability). A sequence of functions $\hat{Q}_n(\theta)$ converges uniformly in probability to $Q_0(\theta)$ means $\sup_{\theta \in \Theta} \left|\hat{Q}_n(\theta) - Q_0(\theta)\right| \xrightarrow{p} 0$.

### 1.3 Some Useful Results

**Theorem** (Weak Law of Large Numbers). Let $X_1, \ldots, X_n$ be iid random variables drawn from the distribution of $X$ with $E[X] = \mu < \infty$. Let $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$. Then $\bar{X}_n \xrightarrow{p} \mu$.

This theorem extends naturally to random vectors by applying the theorem component-wise.

**Theorem** (Central Limit Theorem). Let $X_1, \ldots, X_n$ be iid random variables drawn from the distribution of $X$ with $E[X^2] < \infty$. Let $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$. Then $\sqrt{n}\left(\bar{X}_n - E[X]\right) \xrightarrow{d} N(0, \text{Var}[X])$.

**Question:** Do we also need $E[X] < \infty$?

There is a natural extension to random variables, when $\text{Var}[X]$ is defined as the covariance matrix amongst the components of the vector.

**Theorem** (Continuity Theorem). Let $X_1, \ldots, X_n$ be $\mathbb{R}^\ell$ random vectors, and $X_n \xrightarrow{p} c$ for some constant $c \in \mathbb{R}^\ell$. Let $h: \mathbb{R}^\ell \rightarrow \mathbb{R}^m$ be a function that is continuous at $c$, then $h(X_n) \xrightarrow{p} h(c)$.

In words: The probability limit of a continuous function is the value of that function evaluated at the probability limit.

**Theorem** (Generalized Slutsky’s Theorem). If the random vectors $X_n$ and $Y_n$ satisfy $X_n \xrightarrow{p} c$ and $Y_n \xrightarrow{d} Y$ then for a continuous function $g(\cdot)$, $g(X_n, Y_n) \xrightarrow{d} g(c, Y)$.

**Theorem** (Continuous Mapping Theorem). Let $X_1, \ldots, X_n$ be $\mathbb{R}^\ell$ random vectors, and $X_n \xrightarrow{d} X$ for some random vector $X$. Let $h: \mathbb{R}^\ell \rightarrow \mathbb{R}^m$ be a continuous function, then $h(X_n) \xrightarrow{d} h(X)$.
In words: Convergence in distribution, like convergence in probability, interchanges with continuous functions, as long as they are continuous everywhere (can be relaxed to hold only on the support of $X$).

**Theorem** (Delta Method). Suppose $v(n)(\hat{\theta}_n - \theta_0) \xrightarrow{d} X$, where $\hat{\theta}_n$, $\theta_0$, $X$ are random vectors, and $v(n)$ is a scalar function of $n$ that goes to $\infty$ as $n$ goes to infinity. If $h(\cdot)$ is continuously differentiable at $\theta_0$, then $v(n)(h(\hat{\theta}_n) - h(\theta_0)) \xrightarrow{d} \frac{dh(\theta_0)}{d\theta} X$.

**Question:** When might you have seen the Delta Method “in practice”?

## 2 Introduction of Extremum Estimators

### 2.1 Definition

An estimator $\hat{\theta}$ is an extremum estimator if there is an objective function $\hat{Q}_n(\theta)$ such that $\hat{\theta}$ maximizes $\hat{Q}_n(\theta)$ subject to $\theta \in \Theta$.

**Question:** What about minimizing instead of maximizing?

Note that here we suppress the dependence of the estimator $\hat{\theta}$ and the objective function $\hat{Q}_n(\theta)$ on the data: they vary depending on the sample available and are not population objects.

A special case: An M-estimator is an extremum estimator when the objective function $\hat{Q}_n(\theta)$ is of the form $\frac{1}{n} \sum_{i=1}^{n} g_i(\theta)$.

### 2.2 Examples

**Example** (Maximum likelihood).

- Setting: Data $z_i$ (can be a combination of outcome $y_i$ and covariates $x_i$) that are iid and with pdf $f(z \mid \theta)$ with true parameter $\theta_0$.
- $\hat{Q}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log f(z_i \mid \theta)$

**Question:** Why are we taking the log?

**Example** (Nonlinear least squares).

- Setting: Data $z_i = (y_i, x_i)$ with conditional mean $E[y \mid x] = h(x, \theta_0)$.
- $\hat{Q}_n(\theta) = -\frac{1}{n} [y_i - h(x_i, \theta)]^2$

**Question:** Nonlinear in what?
Example (General method of moments).

- Setting: Data $z_i$ such that we know that population moments $g(z, \theta)$ of the data satisfy: $E[g(z, \theta_0)] = 0$.

- We want to convert this into an extremum estimator. A GMM estimator is one that minimizes a squared Euclidean distance of sample moments from their population counterpart of zero, with a weight matrix $\hat{W}$ that is a positive semi-definite matrix.

  $$\hat{Q}_n(\theta) = -\left[\frac{1}{n} \sum_{i=1}^{n} g(z_i, \theta)\right]' \hat{W} \left[\frac{1}{n} \sum_{i=1}^{n} g(z_i, \theta)\right]$$

Question: Why is there a “hat” on the weighting matrix $W$?

A typical role of the weighting matrix $\hat{W}$ is to balance between information and noise: We may want to weigh down “noisy” moments while still taking them into account and using their information. Such a weighting matrix has the form of an inverse variance-covariance matrix.

Question: Why are we converting an exact condition into an extremum one?

Question: Can we do the opposite?

Example. Classical minimum distance

- Setting: The data can be reduced to some “reduced-form” parameters $\pi$ and there is a mapping from the “structural” parameters $\theta$ to $\pi$: $\pi_0 = h(\theta_0)$.

  $$\hat{Q}_n(\theta) = -[\hat{\pi} - h(\theta)]' \hat{W} [\hat{\pi} - h(\theta)]$$

In practice, the choice of the estimator depends on the setting and should be motivated by the theory.

(We will continue on extremum estimators next time. Please bring this section note with you to section 2.)
1 Consistency and Asymptotic Normality

“Large sample distribution theory is the cornerstone of statistical inference for econometric models. The limiting distribution of a statistic gives approximate distributional results that are often straightforward to derive, even in complicated econometric models. These distributions are useful for approximate inference, including constructing approximate confidence intervals and test statistics. Also, the location and dispersion of the limiting distribution provides criteria for choosing between different estimators.” (Newey McFadden 1994)

**Definition** (Consistency). A consistent estimator \( \hat{\theta} \) is one that converges in probability to the true value \( \theta_0 \), i.e. \( \hat{\theta} \xrightarrow{p} \theta_0 \), as the sample size \( n \) goes to infinity, for all possible true values.

This is a reasonable property, even if it is not very strong, and therefore it is often desirable for an estimator to be considered adequate.

**Definition** (Asymptotic normality). An asymptotically normal estimator \( \hat{\theta} \) one where there is an increasing function \( v(n) \) such that the distribution function of \( v(n)(\hat{\theta} - \theta_0) \) converges to the Gaussian distribution function with mean zero and variance \( V \), i.e. \( v(n)(\hat{\theta} - \theta_0) \xrightarrow{d} N(O,V) \). The variance \( V \) of the limiting distribution is referred to as the asymptotic variance of \( \hat{\theta} \). We generally refer to the case where \( v(n) = \sqrt{n} \), which is also called the \( \sqrt{n} \)-consistent case.

If we establish asymptotic normality and we have a consistent estimator \( \hat{V} \) of the asymptotic variance \( V \), we can then construct approximate confidence intervals for hypothesis testing. Comparing the asymptotic variances of different estimators can also be a criterion of choosing among them.

2 Consistency of Extremum Estimators

**Theorem** (The basic consistency theorem—Newey McFadden Theorem 2.1). *If there is a function \( Q_0(\theta) \) such that:

1. Identification: \( Q_0(\theta) \) is uniquely maximized at \( \theta_0 \)
2. Boundedness of the parameter set: \( \Theta \) is compact
3. Regularity condition (continuity): \( Q_0(\theta) \) is continuous
4. Regularity condition (uniform convergence): \( \hat{Q}_n(\theta) \) converges uniformly in probability to \( Q_0(\theta) \)

then \( \hat{\theta} \xrightarrow{p} \theta \).

Note that it is not necessary to assume that \( \hat{\theta} \) actually maximizes the objective function \( \hat{Q}_n \). We can instead replace it with \( \hat{Q}_n(\hat{\theta}) \geq \sup_{\theta \in \Theta} \hat{Q}_n(\theta) + o_p(1) \). Some of the assumptions can also be
weakened, but usually it is sufficient to check the four conditions above to prove consistency of an extremum estimator.

**Note 1:** The first step to show consistency is to find the function $Q_0(\theta)$. This involves finding the probability limit of $\hat{Q}_n(\theta)$ and the Law of Large Numbers can often be useful.

**Question:** Why is the first step finding the probability limit of $\hat{Q}_n(\theta)$?

**Note 2:** The Law of Large Numbers requires that the expected value has to be finite. It is therefore sometimes useful to normalize the objective function to $\hat{Q}_n(\theta) \neq \hat{Q}_n(\theta_0)$ rather than using $\hat{Q}_n(\theta)$ directly.

Condition (2) is often by assumption. We will then focus on the other three assumptions.

### 2.1 Identification condition

How is condition (1) on identification related to identification in the normal sense? What is identification in the normal sense?

A problem is identified if the distribution of the data at the true parameter is different than that at any other possible parameter value. Otherwise, we will not have any way to choose between the “true” parameter and an incorrect one. This is a matter of the setup of the problem, and in practice, the variation in the data. Do we have specify enough conditions and moments to pin down the parameters?

A problem has to be identified for condition (1) to be satisfied. This is however not a sufficient condition.

Throughout, we define $X \neq Y$ for two random variables $X$ and $Y$ if $\Pr(\{X \neq Y\}) > 0$.

Summary of the identification conditions of some standard extremum estimators:

- **Maximum likelihood estimator:** In this case, identification in the normal sense implies the identification condition (1), i.e., if $\theta \neq \theta_0$ then $f(z | \theta) \neq f(z | \theta_0)$. This is called the Information inequality lemma (Newey McFadden Lemma 2.2).

- **Nonlinear least squares:** If $\theta \neq \theta_0$, then $h(x, \theta) \neq h(x, \theta_0)$ (“conditional mean” identification). In the special case of linear least squares, when $h(x, \theta) = x'\theta$, conditional mean identification holds if $\mathbb{E}[xx']$ is not singular.

- **Generalized method of moments:** If $\hat{W} \overset{p}{\to} W$, $W$ is positive semidefinite and $W\mathbb{E}[g(z, \theta)] \neq 0$ for all $\theta \neq \theta_0$ then $Q_0(\theta) = -\mathbb{E}[g(z, \theta)]'W\mathbb{E}[g(z, \theta)]$ has a unique maximum at $\theta_0$. For nonsingular $W$ this reduces to $\mathbb{E}[g(z, \theta_0)] \neq 0$ if $\theta \neq \theta_0$. A necessary condition is that we have at least as many moments as we have parameters.

For non-linear function $g$ of $\theta$, this may not be trivial. This is one of the reasons to avoid formulating $\hat{\theta}$ as the solution to a first order condition when analyzing consistency (e.g.,
interpreting MLE as a GMM estimator with $g(z, \theta) = \nabla_{\theta} \log f(z | \theta)$. In this case, people often turn to local identification: A sufficient condition for a unique solution of $W \mathbb{E}[g(z, \theta)] = 0$ in a (small enough) neighborhood of $\theta_0$ is that $W \mathbb{E}[\nabla_{\theta} g(z, \theta)]$ has full column rank (Rothenberg 1971).

**Question:** What about the linear case? Think about Linear IV?

- **Classical minimum distance:** Very similar to GMM: If $\hat{\pi} \not \to \pi_0$ and $\hat{W} \not \to W$, $W$ positive semidefinite, then $Q_0(\theta) = -[\pi_0 - h(\theta)]'W[\pi_0 - h(\theta)]$, and condition (1) is satisfied when $h(\theta_0) = \pi_0$ and $h(\theta) - h(\theta_0)$ is not in the null space of $W$ if $\theta \neq \theta_0$ (which reduces to $h(\theta) \neq h(\theta_0)$ if $W$ is nonsingular).

## 2.2 Uniform convergence and continuity conditions

The following theorem (also Newey McFadden Lemma 2.4) is useful for estimators that depend on sample averages (M-estimators), such as MLE, NLS, and GMM.

**Theorem** (Uniform law of large numbers—Newey McFadden Lemma 2.4). Suppose we have the following conditions

1. The data are iid
2. $\Theta$ is compact
3. $m(z_i, \theta)$ is continuous at each $\theta \in \Theta$ with probability 1
4. There is $d(z)$ with $\|m(z, \theta)\| \leq d(z)$ for all $\theta \in \Theta$ and $\mathbb{E}[d(z)] < \infty$.

Then $\mathbb{E}[m(z, \theta)]$ is continuous and $\sup_{\theta \in \Theta} \left\| \frac{1}{n} \sum_{i=1}^{n} m(z_i, \theta) - \mathbb{E}[m(z, \theta)] \right\| \to 0$.

The two conditions imposed on $m(z, \theta)$ are both primitive: Continuity condition of $m(z, \theta)$ can often be verified by inspection. And the condition on the dominating function $d(z)$ only needs to have a bounded mean. In class, instead of using a bounded function $d(z)$, condition 4 is stated using the supremum as $\mathbb{E}[\sup_{\theta \in \Theta} \|m(z, \theta)\| < \infty]$. This result gives you both continuity and uniform convergence.

Outside of M-estimators with sample averages, there are other results that you can appeal to in order to show uniform convergence. Many of them utilize another concept, stochastic equicontinuity.

**Definition.** The sequence of random variables $\{H_n(\theta)\}$ is stochastically equicontinuous if for all $\epsilon > 0$, there exists $\delta > 0$ such that

$$
\lim_{n \to \infty} \Pr \left[ \sup_{\theta} \left( \sup_{\theta' \in B(\theta, \delta)} |H_n(\theta') - H_n(\theta)| \right) > \epsilon \right] < \epsilon
$$
**Theorem** (Stochastic equicontinuity and uniform convergence—Newey McFadden Lemma 2.8). Suppose we have the following conditions

1. $\Theta$ is compact
2. $Q_0(\theta)$ is continuous.

Then $\sup_{\theta \in \Theta} |\hat{Q}_n(\theta) - Q_0(\theta)| \overset{p}{\to} 0$ if and only if $\hat{Q}_n(\theta) \overset{p}{\to} Q_0(\theta)$ for all $\theta \in \Theta$ and $\hat{Q}_n(\theta)$ is stochastically equicontinuous.

In words: If the two conditions are satisfied, then uniform convergence is equivalent to having both pointwise convergence and stochastic equicontinuity. The proof can be found in Newey 1991 “Uniform Convergence in Probability and Stochastic Equicontinuity”.

Finally, another result will provide even more primitive conditions for stochastic equicontinuity.

**Theorem** (Primitives for uniform convergence—Newey McFadden Lemma 2.9). Suppose the following conditions are satisfied:

1. $\Theta$ is compact
2. $Q_0(\theta)$ is continuous
3. $\hat{Q}_n(\theta) \overset{p}{\to} Q_0(\theta)$ for all $\theta \in \Theta$
4. There is $\alpha > 0$ and $\hat{B}_n = O_p(1)$ such that for all $\theta', \theta \in \Theta$, $|\hat{Q}_n(\theta') - \hat{Q}_n(\theta)| \leq \hat{B}_n \|\theta' - \theta\|^\alpha$

Then $\sup_{\theta \in \Theta} |\hat{Q}_n(\theta) - Q_0(\theta)| \overset{p}{\to} 0$

From the previous theorem, it is then enough to show that condition 4 implies stochastic equicontinuity.
1 Asymptotic Normality of Extremum Estimators

All of these results have the starting point of assuming consistency.

Question: Why is this our starting point?

1.1 Smooth objective function

There are two basic results to establish asymptotic normality for well-behaved objective functions. The first, for M-estimators, requires a twice continuous differentiability in a neighborhood of $\theta_0$, whereas the second, for minimum distance estimators, only once continuous differentiability.

**Theorem** (Asymptotic normality 1—Newey McFadden Theorem 3.1). Suppose $\hat{\theta} \xrightarrow{P} \theta_0$ and the following 4 conditions are satisfied:

1. $\theta_0$ is in the interior of $\Theta$
2. $\hat{Q}_n(\theta)$ is twice continuously differentiable in a neighborhood $\Theta_0$ of $\theta_0$
3. $\sqrt{n}\nabla_{\theta} \hat{Q}_n(\theta_0) \xrightarrow{d} N(0,\Omega)$
4. There is $H(\theta)$ continuous at $\theta_0$ such that $\sup_{\theta \in \Theta_0} \|\nabla_{\theta} \hat{Q}_n(\theta) - H(\theta)\| \xrightarrow{P} 0$ and $H(\theta_0)$ is not singular.

Then $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0,H^{-1}\Omega H^{-1})$.

**Question**: Notice that in this last formula, the variance will “blow up” when $H$ is near-singular. What is the intuition here?

**Theorem** (Asymptotic normality 2—Newey McFadden Theorem 3.2). Suppose $\hat{Q}_n(\theta) = -\hat{g}_n(\theta)'\hat{W}\hat{g}_n(\theta)$ where $\hat{W} \xrightarrow{P} W$ and $W$ positive semidefinite. Suppose also that $\hat{\theta} \xrightarrow{P} \theta_0$ and the following conditions are satisfied:

1. $\theta_0$ is in the interior of $\Theta$
2. $\hat{g}_n(\theta)$ is continuously differentiable in a neighborhood $\Theta_0$ of $\theta_0$
3. $\sqrt{n}\hat{g}_n(\theta_0) \xrightarrow{d} N(0,\Omega)$
4. There is $G(\theta)$ continuous at $\theta_0$ such that $\sup_{\theta \in \Theta_0} \|\nabla_{\theta} \hat{g}_n(\theta) - G(\theta)\| \xrightarrow{P} 0$ and $G'WG$ is not singular with $G = G(\theta_0)$.

Then $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0,(G'WG)^{-1}G'W\Omega W G(G'WG)^{-1})$. 
We can think of $\hat{g}_n(\theta)$ as the first order condition for $\hat{Q}_n(\theta)$, and that’s why we only need once differentiability here instead of twice as above.

For both of these theorems, condition 4 can be satisfied by applying the uniform law of large numbers theorem in section 2: It says that you want uniform convergence around a small neighborhood of the true $\theta_0$. The nonsingularity part of this same condition will be satisfied with local identification.

**Question:** Why?

### 1.2 Non-smooth objective function

Smoothness of the objective function is not always available. A common case is when we measure distance using absolute difference. Given the insight that we often only need smoothness of the limit of the objective function around the true value rather than of the objective function itself, we can obtain asymptotic normality in various non-smooth cases. In particular, the limiting objective functions are often expectations which are smoother than their sample counterparts.

We can appeal to the following results, the first one applying to general extremum estimator, and the second one to minimum distance estimators.

**Theorem** (Asymptotic normality 3—Newey McFadden Theorem 7.1). Suppose that $\hat{Q}_n(\theta) \geq \sup_{\theta \in \Theta} \hat{Q}_n(\theta) - o_p\left(\frac{1}{n}\right)$, $\hat{\theta} \xrightarrow{p} \theta_0$ and the following conditions are satisfied:

1. $Q_0(\theta)$ is maximized on $\Theta$ at $\theta_0$
2. $\theta_0$ is an interior point of $\Theta$
3. $Q_0(\theta)$ is twice differentiable at $\theta_0$ with nonsingular second derivative $H$
4. $\sqrt{n} \hat{D}_n \xrightarrow{d} N(0, \Omega)$ where $\hat{D}_n = \nabla_\theta \hat{Q}_n(\theta_0)$
5. For any $\delta_n \to 0, \sup_{||\theta - \theta_n|| \leq \delta_n} \left| \hat{R}_n(\theta)/[1 + \sqrt{n}||\theta - \theta_0||]\right| \xrightarrow{p} 0$ where

$$\hat{R}_n(\theta) = \sqrt{n}\left[\hat{Q}_n(\theta) - \hat{Q}_n(\theta_0) - \hat{D}_n(\theta - \theta_0) - (Q_0(\theta) - Q_0(\theta_0))\right]/||\theta - \theta_0||.$$

Then $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, H^{-1}\Omega H^{-1})$

Notice the difference between condition 3 in this theorem and condition 2 in Newey McFadden Theorem 3.1 above: we have replaced the differentiability of $\hat{Q}_n(\theta)$ which is not smooth with the differentiability of $Q_0(\theta)$ which is usually smoother. (I will upload some simulation code that shows that the expectation of a non-smooth function can be very smooth based on the objective function in problem set 1 question 2.)

Notice also that we write a gradient for $\hat{Q}_n(\theta_0)$ in condition 4 even though it is not smooth. Read this as an “approximate derivative,” which is exactly true with probability 1 (but not everywhere).
Again, you can find an example by looking at the derivation for the Least Absolute Deviation estimator, which is also a quantile regression.

The condition that stands out as completely different from before is condition 5, which we can call “stochastic differentiability” but in this class we can just assume it. (Notice how $\hat{R}_n(\theta)$ has a similar format to the definition of a derivative as $\theta \overset{p}{\to} \theta_0$). Andrews (1997) in the same Handbook discusses ways to verify this condition.

Note that $Q_0$ and condition 1 arise naturally from consistency as in Newey-McFadden Theorem 2.1 and condition 2 can just be assumed.

**Question:** Do you have any guess why we’re explicitly not writing $\hat{\Theta}$ as the arg max of $Q_n(\Theta)$ here?

**Theorem** (Asymptotic normality 4—Newey McFadden Theorem 7.2). Suppose that $\hat{g}_n(\hat{\theta})^\top \hat{W} \hat{g}_n(\hat{\theta}) \leq \inf_{\theta \in \Theta} g_n(\theta)^\top \hat{W} g_n(\theta) + o_p\left(\frac{1}{n}\right)$, $\hat{\theta} \overset{p}{\to} \theta_0$, $\hat{W} \overset{p}{\to} W$, $W$ is positive semidefinite and the following conditions are satisfied:

1. There is $g_0(\theta)$ such that $g_0(\theta_0) = 0$
2. and $g_0(\theta)$ is differentiable at $\theta_0$ with derivative $G$ such that $G'WG$ is nonsingular
3. $\theta_0$ is an interior point of $\Theta$
4. $\sqrt{n} \hat{g}_n(\theta_0) \overset{d}{\to} N(0, \Sigma)$
5. For any $\delta_n \to 0$, $\sup_{\|\theta - \theta_0\| \leq \delta_n} \|g_n(\theta) - \hat{g}_n(\theta) - g_0(\theta)\| / [1 + \sqrt{n}\|\theta - \theta_0\|] \overset{p}{\to} 0$.

Then $\sqrt{n} (\hat{\theta} - \theta_0) \overset{d}{\to} N\left(0, (G'WG)^{-1}G'W\SigmaWG(G'WG)^{-1}\right)$

Similar to above, $g_0(\theta)$ can be thought of as the limit of $\hat{g}_n(\theta)$ and condition 5 can be assumed.

## 2 Two-step Estimators as Joint GMM

The class of GMM estimators is sufficiently general to include two-step estimators where moment functions from the first step and the second step can be “stacked” to form a vector of moment conditions. Newey McFadden Theorem 3.4 can then be applied to specify regularity conditions for asymptotic normality. Previous results can also be used to show consistency, which is an assumption for the asymptotic normality results.

For example, we may be interested in an estimator $\hat{\Theta}$ that satisfies some moment condition: $\frac{1}{n} \sum_{i=1}^{n} g(z_i, \Theta, \hat{\gamma}) = 0$. The estimator $\hat{\Theta}$ itself can be a vector, and $g(\cdot)$ has the same dimension as $\Theta$. Here we are treating $\hat{\gamma}$ as a known object. The reason that there is a “hat” is because this object itself was an estimator from a previous step procedure using the same dataset. If we do not take into account the fact that $\hat{\gamma}$ is not an exact population-level estimate, then simply deriving the
asymptotic distribution of \( \hat{\theta} \) will give us the wrong asymptotic variance, invalidating our hypothesis tests.

Now, if \( \hat{\gamma} \) itself was estimated as satisfying a moment condition of the same form \( \frac{1}{n} \sum_{i=1}^{n} m(z_i, \gamma) = 0 \), then what we can do instead is to use a “stacked” estimator and form the joint moment condition \( \frac{1}{n} \sum_{i=1}^{n} \tilde{g}(z_i, \hat{\theta}, \hat{\gamma}) \) with \( \tilde{g}(z, \theta, \gamma) = \left[ m(z, \gamma), g(z, \theta, \gamma) \right] \). The two-step estimator can be viewed as a GMM estimator.

In general, if you do the two steps sequentially and independently, you will underestimate the variance in the second step by taking the first-step estimate as a constant (for example, with Two Stage Least Squares). This is not always the case. If the first-step estimator has no first-order impact on the asymptotic variance of the second-step estimator, then the asymptotic variance of the second-step estimator is the same as if we knew the true value of the first-step estimand. In such situations, it’s often said that \( \hat{\theta} \) is oracle-efficient. This occurs when \( G_\gamma \equiv \Delta_\gamma \mathbb{E}[g(z_i, \theta_0, \gamma_0)] = 0 \), i.e., the expectation of the derivative of the second-stage moment condition with respect to the first-stage parameter is 0 when evaluated at the true parameters. Another way of seeing it is that small changes in \( \gamma_0 \) have negligible first order impact on the root \( \beta_0 \) of the second-step moment. This is something you can check directly, and will automatically drop out even if you do not check in advance and work instead with the stacked GMM.

**Note:** Check that in problem set 1, part 3 (d), even if we use a feasible estimator, the feasible estimator will have the same asymptotic variance as the infeasible one. The case when we can safely adjust the objective function by an estimated variance from the first stage is fairly general.

## 3 Quantile Regression

Instead of the mean, we can define the median. We can also generalize to quantiles, which help describe the entire distribution of a random variable, not just the center.

We define the \( u \)-th quantile of \( y \) as \( q_u \) which satisfies \( \Pr(y \leq q_u) \geq u \) and \( \Pr(y \geq q_u) \geq 1 - u \).

**Question:** Why are quantiles also more robust to outliers than averages?

In linear least squares, the conditional average of \( y \) given \( x \) is linear: \( \mathbb{E}[y \mid x] = x'\beta \).

**Question:** This is because the mean square error has a unique minimum at the conditional mean. Can you prove it? (Hint: Use the law of iterated expectation.)

Suppose instead that the \( \theta \)-quantile of the distribution of \( y \) given \( x \) is linear: \( Q_u(y \mid x) = x'\beta_u \). Then \( \beta_u = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \rho_u(y_i - x_i'\beta) \), where \( \rho_u(\epsilon) = (u - 1_{\{\epsilon \leq 0\}})\epsilon \). This is again an M-estimator. But the difficulty arises because it is not smooth (because of the indicator function).

Like mean regression, quantile regression can be expressed using an error term: \( y = x'\beta_u + \epsilon_u \) where \( Q_u(\epsilon_u \mid x) = 0 \) or equivalently \( \Pr(\epsilon_u \leq 0 \mid x) = u \).
This is identified if $E[f_{u|x}(0 \mid x)xx'] > 0$, or equivalently if $E[xx']$ is invertible and the $u$-th quantile of $y$ given $x$ is unique at every $x$. When $y$ is discrete, we are in trouble because of this uniqueness condition not being satisfied.

**Note:** This condition looks very similar to the identification condition of a linear least squares. Note that there is no uniqueness issue when it comes to the mean.

Under regularity conditions (applying Newey McFadden theorem 7.1 for asymptotic normality), the quantile regression estimator is consistent and asymptotically normal: $\sqrt{n}(\hat{\beta}_u - \beta_u) \overset{d}{\rightarrow} N(0, H^{-1}\Omega H^{-1})$, where $H = E[f_{u|x}(0)xx']$ and $\Omega = u(1-u)E[xx']$ and $f_{u|x}(\cdot)$ is the density of $\epsilon_u$ conditional on $x$.

**Note:** There is a reason why $H$ being invertible is the same condition as identification above. Recall from above that $H$ being invertible is a local identification condition. Here, local identification is sufficient. (See the proof in class for the median case.)

The challenge in computation arises because $H$ contains a conditional density, which we will have to estimate. This is on top of the computational challenge of finding the optimal $\hat{\beta}_u$ to minimize the non-smooth objective function itself.
1 The Paths to Statistical Inference

“Statistical inference is the process of using data analysis to deduce properties of an underlying probability distribution. Inferential statistical analysis infers properties of a population, for example by testing hypotheses and deriving estimates. It is assumed that the observed data set is sampled from a larger population.

Inferential statistics can be contrasted with descriptive statistics. Descriptive statistics is solely concerned with properties of the observed data, and it does not rest on the assumption that the data come from a larger population.” (Wikipedia)

In the ideal situation, we hope to obtain the exact distribution of our estimator in order to do statistical inference.

Let the data be a random sample of size $n$ denoted by $\{Z_1, \ldots, Z_n\}$. Let the true data generating process be represented by a probability distribution with cumulative distribution $F_0$ with parameters $\theta_0$: $F_0(z, \theta_0) \equiv \Pr(Z \leq z)$. Let $T_n = T_n(Z_1, \ldots, Z_n)$ be a test statistic, for example $T_n = \sqrt{n}(\hat{\theta}_n(Z_1, \ldots, Z_n) - \theta_0)$. Let $G_n(t, F_0) \equiv \Pr(T_n \leq t)$ denote the exact, finite-sample CDF of $T_n$ (depending on $n$). This object is often hard to obtain (recall Section 1).

Assuming we do have it, here is how we would do hypothesis testing with the example test statistic above. For a desired two-sided equal-tailed confidence interval with confidence level $1 - \alpha$ ($\alpha \in (0, 1)$), let:

$$a = \hat{\theta}_n - G_n^{-1}(1 - \alpha/2, F_0)$$
$$b = \hat{\theta}_n - G_n^{-1}(\alpha/2, F_0)$$

**Note:** We are dropping the scale $\sqrt{n}$ of the test statistic here because we are dealing with an exact distribution with a fixed $n$.

Then the exact confidence interval we want is:

$$\Pr(a \leq \theta_0 \leq b) = \Pr(\hat{\theta}_n - b \leq \hat{\theta}_n - \theta_0 \leq \hat{\theta}_n - a)$$
$$= G_n(\hat{\theta}_n - a, F_0) - G_n(\hat{\theta}_n - b)$$
$$= (1 - \alpha/2) - \alpha/2$$
$$= 1 - \alpha$$

Generally we do not know $\theta_0$, and we also cannot obtain $a$ and $b$ without knowing what $G_n(\cdot, F_0)$ is.

What we have been doing so far with large-sample approximation is to let $T_n \rightarrow T$ and find $G_n(t, F_0) \rightarrow G_\infty(t, F_0)$, where $G_\infty(\cdot, F_0)$ is the CDF of the limiting random vector $T$, the distribution of which depends on $F_0$. There are times when this may be challenging:
Some statistics have complicated limiting distributions and it can be hard to compute quantiles or moments of $G_{\infty}(\cdot, F_0)$. Even for the extremum estimators we considered with established results, many have various derivative terms involved. When there are no nice analytical form for these terms, computation becomes a challenge.

Large-$n$ approximations can work poorly with small $n$, and it’s not always easy to figure out how “large” $n$ needs to be for the approximation to be appropriate (depending on the unknown true distribution $F_0$).

Alternatively, we can approximate $G_n(t, F_0)$ with $G_n(t, F_n)$. $F_0$ is an unknown object, but we will approximate it using the data and use $F_n$ instead. As $n \neq \infty$, we also approach $G_{\infty}(t, F_0)$. This is the bootstrap. How good it is depends on how well we approximate $F_0$.

2 The Bootstrap

We will go over the procedure to compute the bootstrap in two steps. First, assume we know $F_0$, how do we go about approximating the exact sample distribution $G_n(\cdot, F_0)$ of $T_n$?

Monte Carlo simulation:

1. For each bootstrap $b = 1, \ldots, B$:
   
   (a) Draw $n$ data points from $F_0$ to get the sampled dataset $Z_1, \ldots, Z_n$.
   
   (b) Compute the statistic $T_{N,b}$ on the simulated dataset. For example, for $T_n = \sqrt{n}(\hat{\theta}_n - \theta_0)$, compute the estimator $\hat{\theta}_b$ on the simulated dataset, and set $T_{N,b} = \sqrt{n}(\hat{\theta}_b - \theta_0)$.

2. Approximate $G_n(\cdot, F_0)$ by the empirical distribution function (EDF) of $T_{N,b}$ across the $B$ simulated datasets: $G_n(t, F_0) \approx \frac{1}{B} \sum_{b=1}^{B} 1\{T_{N,b} \leq t\}$, $t \in \mathbb{R}$. This approximation becomes exact in the limit as we increase the computational effort $B \to \infty$ (the Glivenko-Cantelli theorem).

**Question:** Why do we draw $n$ data points? (This is important.)

Now, we don’t know $F_0$. Assume that we know the parametric form of $F_0$ and therefore can approximate $F_0$ with $\hat{F} = F(\cdot, \hat{\theta})$. This is the **parametric bootstrap**.

1. For each bootstrap $b = 1, \ldots, B$:
   
   (a) Draw $n$ data points from $\hat{F}$ to get the sampled dataset $Z_1, \ldots, Z_n$.
   
   (b) Compute the statistic $T_{N,b}$ on the simulated dataset. For example, for $T_n = \sqrt{n}(\hat{\theta}_n - \theta_0)$, compute the estimator $\hat{\theta}_b$ on the simulated dataset, and set $T_{N,b} = \sqrt{n}(\hat{\theta}_b - \hat{\theta})$. (Note that we don’t have $\theta_0$ here. A requirement is then that $\hat{\theta}$ is consistent for $\theta_0$.)
2. Approximate $G_n(\cdot,F_0)$ by the empirical distribution function (EDF) of $T_{n,b}^*$ across the $B$ simulated datasets.

**Note:** In the bootstrap world, it is $\hat{\theta}$ that governs the $\hat{F}$.

If we don’t have any information about $F_0$, if the data are iid, we can still approximate it using the EDF from the dataset: $\hat{F}(z) = \frac{1}{n} \sum_{i=1}^{n} 1\{Z_i \leq z\}$. This is the **nonparametric bootstrap**.

1. For each bootstrap $b = 1, \ldots, B$:
   
   (a) Draw $n$ data points $Z_1^*, \ldots, Z_n^*$ by randomly sampling from $Z_1, \ldots, Z_n$ with replacement.
   
   (b) Compute the statistic $T_{N,b}^*$ on the simulated dataset. For example, for $T_n = \sqrt{n} (\hat{\theta}_n - \theta_0)$, compute the estimator $\hat{\theta}_b^*$ on the simulated dataset, and set $T_{n,b}^* = \sqrt{n} (\hat{\theta}_b^* - \hat{\theta})$. (Note that we don’t have $\theta_0$ here. A requirement is then that $\hat{\theta}$ is consistent for $\theta_0$.)

2. Approximate $G_n(\cdot,F_0)$ by the empirical distribution function (EDF) of $T_{n,b}^*$ across the $B$ simulated datasets.

Most often, what is called “the bootstrap” is the nonparametric bootstrap.

The hope with bootstrapping is that we can approximate $F_0$ well, and that our approximation takes into account skewness, kurtosis, nonlinearities, things that the CLT and the Delta Method “iron” out in large sample asymptotic approximation.

Note also that we have always assumed that the number of simulations $B$ is large enough. This is a matter of computational time and power, and it gets easier over time. A rule of thumb to choose $B$ is to observe when the EDF of the statistics of interest are stable. If the $T_n$ is more “well-behaved,” we usually require a smaller $B$, i.e., if it is more sensitive to the tail behavior of the data (extreme quantiles, etc.) we often need a higher $B$.

## 3 When Does Bootstrap Work?

First, a formal statement:

**Theorem** (Bootstrap consistency—Horowitz Theorem 2.1). Let $\rho(\cdot, \cdot)$ be a metric (distance) on the space of probability distributions for the data $z$. Let $G_\infty(t,F_0)$ denote the asymptotic CDF of the statistic $T_n$. Assume, for any $\epsilon > 0$:

1. $\Pr(\rho(\hat{F}, F_0) > \epsilon) \to 0$ as $n \to \infty$

2. $G_\infty(t,F)$ is continuous in $t$, for each possible distribution $F$ of $z$
3. For any \( t \) and any sequence \( \{F_k\} \) of distributions for \( z \) satisfying \( \rho(F_k, F_0) \to 0 \) as \( k \to \infty \), we have \( G_n(t, F_n) \overset{d}{\to} G_{\infty}(t, F_0) \) as \( n \to \infty \).

Then for any \( \epsilon > 0 \), 
\[
\Pr\left( \sup_{t \in \mathbb{R}} |G_n(t, \hat{F}) - G_{\infty}(t, F_0)| > \epsilon \right) \to 0 \text{ as } n \to \infty.
\]

Because we mostly work with asymptotically normal or \( \chi^2 \) statistics (those from the first part of the class for example), condition 2 is satisfied. We still need to be concerned about conditions 1 and 3.

Condition 1 means that approximation \( \hat{F} \) of \( F_0 \) should be good asymptotically. For iid data and \( \hat{F} \) being the EDF, this holds by Glivenko-Cantelli. For a parametric bootstrap, this often holds if the model is correctly specified, and the data distribution \( F(\cdot, \theta_0) \) does not change discontinuously with \( \theta_0 \).

Condition 3 means that the CDF of \( T_n \) cannot change discontinuously with the data distribution \( F_0 \), in large samples.

**Best practice summary:**

- The bootstrap should mimic the dependence/independence in the data sampling procedure. Important for two-step estimators and clustered/panel data.

- The bootstrap is usually valid for smooth functions of sample moments, for asymptotically linear estimators, and for tests based on these.

**Worry about bootstrap consistency if:**

- The limiting distribution of \( T_n \) is not normal or \( \chi^2 \).

- The statistic is not a smooth function of the data.

- The limiting distribution of \( T_n \) changes functional form or convergence rate when the true parameter \( \theta_0 \) equals a specific point in the parameter space (e.g., 0 or a boundary point), and you cannot exclude this parameter value a priori.

- The distribution of the statistic is sensitive to “extreme” features of the data distribution (tail probabilities, high-order moments, etc.).

An example is given in Abadie and Imbens (ECMA 2008) “On the Failure of the Bootstrap for Matching Estimators”: “Here we show that the bootstrap is in general not valid, even in the simple case with a single continuous covariate when the estimator is root-N consistent and asymptotically normally distributed with zero asymptotic bias. Due to the extreme non-smoothness of nearest neighbor matching, the standard conditions for the bootstrap are not satisfied, leading the bootstrap variance to diverge from the actual variance.”
4 When is the Bootstrap “Better”?

The bootstrap works when it produces a consistent estimate. As an alternative to deriving the (first-order) asymptotic distribution, the bootstrap can do even better and with “higher-order” asymptotic refinement. In these cases, the bootstrap has a better approximation accuracy, and therefore a more accurate size for a test and a more correct coverage of a confidence interval. The bootstrap will be strictly preferred for these cases. But when and how?

First, some definitions related to the discussion in the first part:

4.1 Pivotal statistic

Usually $G_n(t, F)$ is a different function of $t$ for different distributions $F$. The statistic $T_n$ is called pivotal if $G_n(\cdot, F)$ does not depend on $F$. This is often not achievable without strong distributional assumptions.

For example, assume $Z \sim \mathcal{N}(\mu_0, \sigma_0^2)$ with both parameters unknown. Given iid data $Z_1, \ldots, Z_n$, the t-statistic for the null hypothesis $H_0 : \mu_0 = \mu_*$ is $t_n = \frac{\bar{Z} - \mu_*}{\bar{\sigma}/\sqrt{n}}$ where $\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Z_i - \bar{Z})^2$. This statistic has a Student-t distribution with $n - 1$ degree of freedom under the null hypothesis, and it is pivotal (for any $n$).

If a statistic is pivotal, then we do not need to approximate $F_0$ (it does not matter then!), and we achieve the exact confidence interval.

4.2 Asymptotically pivotal statistic

A statistic is asymptotically pivotal if $G_\infty(\cdot, F)$, the distribution of the limiting test statistic $T$ does not depend on $F$ (for example, if $T \sim \mathcal{N}(0, 1)$ or $\chi^2$).

Digression: Econometric parameter estimators usually are not asymptotically pivotal, but many are asymptotically normally distributed. If an estimator is asymptotically normally distributed, then its asymptotic distribution depends on at most two unknown parameters, the mean and the variance, that can often be estimated without great difficulty. The normal distribution with the estimated mean and variance can then be used to approximate the unknown $G_n(\cdot, F_0)$ if $n$ is sufficiently large. This is the case for the first part of this class when we derive asymptotic normal distributions of the extremum estimators.

4.3 Asymptotic refinement of the bootstrap

The bootstrap produces a more accurate approximation to the distribution of a statistic than the usual asymptotic approximation we derived if the statistic is asymptotically pivotal. The sensitivity of the
sampling distribution of \( T_n \) to changes in \( F_0 \) is negligible in large samples, hence, the bootstrap’s reliance on an approximation of \( F_0 \) is not as bad.

This is not the case with standard errors, but the bootstrap is still used often because computationally it can be simpler. If we want use the standard error to achieve testing or confidence interval however, it is better to apply the bootstrap directly to those instead of using the standard error and making another approximation using those.

**Heuristics for how we get refinements:**

An **Edgeworth expansion** to the distribution of a statistic provides a higher-order expansion of the CDF with terms in power of \( 1/\sqrt{n} \).

**Asymptotically pivotal case:**

Suppose that \( \sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \sigma_0^2) \), then \( \frac{\hat{\theta} - \theta_0}{\sigma_0/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1) \) is asymptotically pivotal if \( \hat{\sigma} \xrightarrow{p} \sigma_0 \).

Under some conditions, it can be shown that:

\[
\Pr\left( \frac{\hat{\theta} - \theta_0}{\sigma_0/\sqrt{n}} \leq t \right) = \Phi(t) + \phi(t) \frac{p_1(t, F_0)}{\sqrt{n}} + O(1/n)
\]

where \( \Phi(\cdot) \) and \( \phi(\cdot) \) are the standard normal CDF and PDF and \( p_1 \) is a continuous function. And the approximation error of the normal approximation

\[
\Pr\left( \frac{\hat{\theta} - \theta_0}{\sigma_0/\sqrt{n}} \leq t \right) - \Phi(t)
\]

is of order \( 1/\sqrt{n} \).

The bootstrap distribution on the other hand can be shown to be

\[
\Pr^*\left( \frac{\hat{\theta}^* - \theta_0}{\sigma_0/\sqrt{n}} \leq t \right) = \Phi(t) + \phi(t) \frac{p_1(t, F_n)}{\sqrt{n}} + O(1/n).
\]

The bootstrap approximation is then

\[
\Pr\left( \frac{\hat{\theta} - \theta_0}{\sigma_0/\sqrt{n}} \leq t \right) - \Pr^*\left( \frac{\hat{\theta}^* - \theta_0}{\sigma_0/\sqrt{n}} \leq t \right) = \frac{\phi(t)}{\sqrt{n}} (p_1(t, F_0) - p_1(t, F_n)) + O(1/n)
\]

Because \( F_n \) converges to \( F_0 \) with \( n \to \infty \), the order of the first term is smaller than \( 1/\sqrt{n} \). The rate of refinement depends on how fast \( F_n \) approaches \( F_0 \). If this rate is \( 1/\sqrt{n} \), then the bootstrap approximation if of order \( 1/n \).

**Non-pivotal case:**
For a non-pivotal statistic, for example: $T_n = \sqrt{n}(\hat{\theta} - \theta_0) \overset{d}{\rightarrow} \mathcal{N}(0, \sigma_0^2)$, then it can be shown that

$$\Pr(\sqrt{n}(\hat{\theta} - \theta_0)) = \Pr\left( \frac{\hat{\theta} - \theta_0}{\sigma_0/\sqrt{n}} \leq \frac{t}{\sigma_0} \right) + O(1/\sqrt{n})$$

$$\Pr*(\sqrt{n}(\hat{\theta}^* - \hat{\theta})) = \Pr\left( \frac{\hat{\theta}^* - \hat{\theta}}{\hat{\sigma}/\sqrt{n}} \leq \frac{t}{\hat{\sigma}} \right) + O(1/\sqrt{n})$$

The approximation error in this case is

$$\Pr(\sqrt{n}(\hat{\theta} - \theta_0)) - \Pr*(\sqrt{n}(\hat{\theta}^* - \hat{\theta})) = \Phi\left( \frac{t}{\sigma_0} \right) - \Phi\left( \frac{t}{\hat{\sigma}} \right) + O(1/\sqrt{n})$$

Because the difference in the first 2 terms is typically of order $1/\sqrt{n}$, the approximation error here is still of order $1/\sqrt{n}$ and there is no improvement over the normal approximation.
Nonparametric Methods
Ec 2140

Linh T. Tô
Harvard University

2 May 2018

With thanks to Alberto Abadie, Whitney Newey, and Mikkel Plagborg-Møller.
Nonparametric density estimation

We want to estimate the density $f(x)$ of random variable $X$ from i.i.d. data $X_1, \ldots, X_N$. Direct way to visualize distribution.

If $X$ is a discrete variable taking values in a countable set $\mathcal{X}$, the analogy principle suggests that we should estimate $f(x)$ (the probability mass function) by

$$
\hat{f}(x) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(X_i = x), \quad x \in \mathcal{X}.
$$

These are just the empirical frequencies of the points $\mathcal{X}$ in the support of $X$. Always have $\hat{f}(x) \in [0, 1]$ and $\sum_{x \in \mathcal{X}} \hat{f}(x) = 1$.

Asymptotic normality follows from the CLT: For any $x_0 \in \mathcal{X}$

$$
\sqrt{N}(\hat{f}(x_0) - f(x_0)) \xrightarrow{d} N(0, f(x_0)(1 - f(x_0)))
$$
Nonparametric density estimation

There are many situations where we are interested in estimating the distribution of a *continuous* variable $X$:

- GDP per capita across countries,
- wage rates for US workers,
- valuations from a second-price sealed-bid auction.

The analogy principle seems to fail us: The empirical distribution is always discrete, so there is no direct way to compute a corresponding PDF. (What is the density at a point $x$ that is not one of our data points?)

Solution 1: **Histogram**. This relies on *binning*, i.e., splitting the continuous support of $X$ up into finitely many bins.

Solution 2: Fit a parametric model (e.g., normal) to the data and use the implied density.
Nonparametric density estimation

Histogram and best-fitting normal density
Nonparametric density estimation

Neither solution does everything we might want:

- The parametric model could be misspecified.
- The histogram is a nonparametric method, but it throws away information by binning the data. How to choose the bins?

Kernel density estimators are similar to histograms, but:

- Provide a smooth depiction of the distribution of the data. Output actual densities $\hat{f}(x)$ (wrt. Lebesgue measure).
- Formal theory of how to pick the **bandwidth**, the analog of the bin size.
Kernel density estimation

Suppose that $X$ is univariate and continuously distributed with CDF $F$ and density $f$ that are unknown to us.

The density at $x_0$ is defined as

$$f(x_0) = \lim_{h \to 0} \frac{F(x_0 + h) - F(x_0 - h)}{2h}.$$ 

This suggests estimating $f(x_0)$ with

$$\hat{f}(x_0) = \frac{1}{2h} \left[ \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(X_i \leq x_0 + h) - \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(X_i \leq x_0 - h) \right]$$

$$= \frac{1}{2h} \left( \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(x_0 - h < X_i \leq x_0 + h) \right),$$

for $h$ small. Notice the similarity with histograms: it’s like a running bin centered around $x_0$. 
Kernel density estimation

Let

\[ K(z) = \mathbb{1}(|z| \leq 1)/2. \]

Then

\[ \frac{1}{2} \mathbb{1}(x_0 - h \leq X_i \leq x_0 + h) = K\left(\frac{X_i - x_0}{h}\right), \]

and (if we ignore the probability 0 event \( X_i = x_0 - h \)):

\[ \hat{f}(x_0) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{X_i - x_0}{h}\right). \]

The normalization is so that the function \( K(\cdot) \) is independent of \( h \). Intuitively, \( 2 \sum_{i=1}^{N} K((X_i - x_0)/h) \) is the number of observations in \([x_0 - h, x_0 + h]\). We divide this number by \( 2Nh \): the number of observations that we would expect in \([x_0 - h, x_0 + h]\) if \( X \) had a uniform density.
Kernel density estimation

The function $K$ is called the **kernel function**, the constant $h$ is called the **bandwidth**.

The kernel function and the bandwidth together are used to measure proximity of any observation $X_i$ to $x_0$:

$$K\left(\frac{X_i - x_0}{h}\right) = \begin{cases} \frac{1}{2} & \text{if } |X_i - x_0| \leq h, \\ 0 & \text{if } |X_i - x_0| > h. \end{cases}$$

The particular kernel function (called Boxcar or Uniform) used so far assigns the same weight to every observation in the $[x_0 - h, x_0 + h]$ band, and zero weight to observations outside that band.

The idea of kernel density estimation can be extended to smooth weighting schemes by adopting a smooth kernel. Thus, can give higher weight to observations $X_i$ that are closer to $x_0$. 
Kernel density estimation

\[ K(z) = 1\{ |z| \leq 1 \}/2 \]

\[ K(z) = (1 - |z|)1\{ |z| \leq 1 \} \]

\[ K(z) = (2\pi)^{-1/2}e^{-z^2/2} \]

\[ K(z) = (3/4)(1 - z^2)1\{ |z| \leq 1 \} \]
Kernel density estimation

A kernel is a weighting scheme for observations close to \( x_0 \).

The definition of the estimator remains the same:

\[
\hat{f}(x_0) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{X_i - x_0}{h}\right).
\]

The following properties hold for the usual kernels:

- \( K(z) \) is positive, symmetric around zero, bounded, and has thin tails (\( K(z) \) goes to zero fast as \( z \to \pm \infty \)). Implies \( \int z^r K(z) dz = 0 \) for any positive odd integer \( r \).
- \( \int K(z) dz = 1 \).
- \( \int z^2 K(z) dz < \infty \).

That is, common kernels are symmetric density functions with mean zero and finite variance. If \( K(\cdot) \) is a density function then \( \hat{f}(\cdot) \) is as well: \( \hat{f}(\cdot) \geq 0 \) and \( \int \hat{f}(x) \, dx = 1 \).
Kernel density estimation

An observation at a distance $z$ from $x_0$ is given weight equal to

$$K_h(z) = K(z/h)/h.$$  

The bandwidth $h$ controls the spread of the kernel.

- If $h$ is small, only observations close to $x_0$ get large weights.
- If $h$ is large, observations distant from $x_0$ may get large weights.

$K_h(z)$ with a small and a large bandwidth
Kernel density estimation: Twin Peaks?

kernel density estimator

log GDP per capita

density

kernel: normal, bandwidth = 0.5
Kernel density estimation: Twin Peaks?

Kernel density estimator

log GDP per capita

density

kernel: normal, bandwidth = 0.4
Kernel density estimation: Twin Peaks?

kernel density estimator

log GDP per capita

density

kernel: normal, bandwidth = 0.3
Kernel density estimation: Twin Peaks?

Kernel density estimator

log GDP per capita

density

kernel: normal, bandwidth = 0.2
Kernel density estimation: Twin Peaks?

kernel density estimator

log GDP per capita

kernel: normal, bandwidth = 0.1
Kernel density estimation: Twin Peaks?

kernel density estimator

log GDP per capita

kernel: normal, bandwidth = 0.2
Kernel density estimation: Twin Peaks?

Kernel density estimator

log GDP per capita

density

kernel: normal, bandwidth = 0.3
Kernel density estimation: Twin Peaks?

Kernel density estimator

log GDP per capita

density

kernel: normal, bandwidth = 0.4
Kernel density estimation: Twin Peaks?

Kernel density estimator

log GDP per capita

density

kernel: normal, bandwidth = 0.5
Kernel density estimation: Twin Peaks?

kernel density estimator

log GDP per capita

density

kernel: normal, bandwidth = 0.6
Kernel density estimation: Twin Peaks?

kernel density estimator

log GDP per capita

density

kernel: normal, bandwidth = 0.7
Kernel density estimation: Twin Peaks?

kernel density estimator

density

log GDP per capita

kernel: normal, bandwidth = 0.8
Kernel density estimation: Twin Peaks?

Kernel density estimator

log GDP per capita

density

kernel: normal, bandwidth = 0.9
Kernel density estimation: Twin Peaks?

kernel density estimator

kernel: normal, bandwidth = 1
Kernel density estimation: Alternative interpretation

Another derivation of kernel density estimators yields additional insight.

The analogy principle says we should base our estimate $\hat{f}(x_0)$ on the empirical distribution function (EDF)

$$\hat{F}(x) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(X_i \leq x).$$

As discussed, this is hard to do for continuous variables. But suppose we smooth out the empirical distribution by adding a little continuously distributed, independent random noise:

$$\tilde{X} = \hat{X} + hU.$$

Here $\hat{X}$ has the (discrete) empirical distribution $\hat{F}$, $h > 0$ is a small scalar, and $U$ is ctsly distributed with CDF $F_U(u) = \int_{-\infty}^{u} K(v) \, dv$. 
Kernel density estimation: Alternative interpretation

The CDF of the smoothed-out random variable \( \tilde{X} \) is

\[
F_{\tilde{X}}(x) = \Pr_{\hat{X},U}(\tilde{X} \leq x) = \mathbb{E}_{\hat{X}} \left[ \Pr_U(\hat{X} + hU \leq x \mid \hat{X}) \right] \\
= \mathbb{E}_{\hat{X}} \left[ \Pr_U(U \leq \frac{x - \hat{X}}{h} \mid \hat{X}) \right] = \mathbb{E}_{\hat{X}} \left[ F_U\left(\frac{x - \hat{X}}{h}\right)\right] \\
= \frac{1}{N} \sum_{i=1}^{N} F_U\left(\frac{x - X_i}{h}\right).
\]

This is a cts distribution. To find its density at \( x_0 \), differentiate:

\[
f_{\tilde{X}}(x_0) = F'_{\tilde{X}}(x_0) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x_0 - X_i}{h}\right).
\]

If \( K(\cdot) \) is symmetric, this is the kernel density estimator from above. So we can think of this estimator as arising from applying a version of the analogy principle. The kernel function is the PDF of the smoothing noise \( U \), and the bandwidth scales the noise.
Kernel density estimation: Bias/variance

How do we choose the bandwidth $h$?

- **Large** $h$: $\hat{f}(x_0)$ smooths over many data points far from $x_0$. Low variance $\text{var}(\hat{f}(x_0))$ but high bias $|\mathbb{E}[\hat{f}(x_0)] - f(x_0)|$.

- **Small** $h$: $\hat{f}(x_0)$ smooths over only the few closest data points to $x_0$. High variance, low bias.

We face a bias/variance tradeoff.

In fact, econometricians always face bias/variance tradeoffs, e.g.:

- Use full dataset (low var), or subsample with most accurately measured variables (low bias).
- Use tightly parametrized model (low var), or flexible/high-dimensional model (low bias).
- Ignore endogeneity (low var), or use an IV (low bias).

In the kernel smoothing problem, this fundamental tradeoff is phrased in terms of a single tuning parameter $h$. 

Picking too large a bandwidth would obscure the economically important peak in the 1979 density around the minimum wage.
Kernel density estimation: Bias/variance

\[ \mathbb{E}[\hat{f}(x_0)] = \mathbb{E}\left[ \frac{1}{Nh} \sum_{i=1}^{N} K\left( \frac{X_i - x_0}{h} \right) \right] = \mathbb{E}\left[ \frac{1}{h} K\left( \frac{X - x_0}{h} \right) \right] \]

\[ = \int \frac{1}{h} K\left( \frac{x - x_0}{h} \right) f(x) dx. \]

Using the change of variable \( z = (x - x_0)/h \) (so that \( dx = h \, dz \)):

\[ \mathbb{E}[\hat{f}(x_0)] = \int K(z) f(x_0 + hz) \, dz \]

Using a Taylor expansion:

\[ f(x_0 + hz) = f(x_0) + f'(x_0) hz + \frac{1}{2} f''(x_0)(hz)^2 \]

\[ + \frac{1}{6} f'''(x_0)(hz)^3 + O(h^4). \]

Reminder: A function \( g(h) \) is \( O(h^r) \) if \( g(h) \) is bounded by a constant times \( h^r \) (as \( h \to 0 \)).
Kernel density estimation: Bias/variance

Because \( \int K(z)dz = 1 \) (density function), \( \int zK(z)dz = 0 \), and \( \int z^3K(z)dz = 0 \) (symmetry):

\[
E \left[ \hat{f}(x_0) \right] = f(x_0) + \frac{h^2}{2} f''(x_0) \int z^2 K(z)dz + O(h^4).
\]

Therefore, the bias

\[
E \left[ \hat{f}(x_0) \right] - f(x_0) = \frac{h^2}{2} f''(x_0) \int z^2 K(z)dz + O(h^4),
\]

is of order \( O(h^2) \).

The higher the curvature \( f''(x_0) \) of the true density is at \( x_0 \), the higher the bias.

Intuition: High curvature means that \( f(x) \) changes rapidly and nonlinearly around \( x = x_0 \). Hard to estimate well in a given sample.
Kernel density estimation: Bias/variance

\[
\text{var}(\hat{f}(x_0)) = \text{var} \left( \frac{1}{Nh} \sum_{i=1}^{N} K \left( \frac{X_i - x_0}{h} \right) \right) \\
= \frac{1}{N} \text{var} \left( \frac{1}{h} K \left( \frac{X - x_0}{h} \right) \right) \\
= \frac{1}{N} \left( \mathbb{E} \left[ \left( \frac{1}{h} K \left( \frac{X - x_0}{h} \right) \right)^2 \right] - \left( \mathbb{E} \left[ \frac{1}{h} K \left( \frac{X - x_0}{h} \right) \right] \right)^2 \right) \\
= f(x_0) + O(h^2)
\]

Using a change of variable as before, we obtain:

\[
\text{var}(\hat{f}(x_0)) = \frac{1}{N} \left( \int \frac{1}{h} K(z)^2 f(x_0 + hz) dz - \left( f(x_0) + O(h^2) \right)^2 \right).
\]
Using a first order Taylor expansion

\[ f(x_0 + hz) = f(x_0) + O(h), \]

we obtain

\[
\text{var} \left( \hat{f}(x_0) \right) = \frac{1}{Nh} f(x_0) \int K(z)^2 dz + O \left( \frac{1}{N} \right) \\
- \frac{1}{N} \left( f(x_0) + O(h^2) \right)^2.
\]

The last term is \( O(1/N) \), so the variance is of order \( O \left( \frac{1}{Nh} \right) \):

\[
\text{var} \left( \hat{f}(x_0) \right) = \frac{1}{Nh} f(x_0) \int K(z)^2 dz + O \left( \frac{1}{N} \right).
\]
Kernel density estimation: Bias/variance

As our intuition told us, bias is increasing in $h$, while variance is decreasing in $h$.

If $N \to \infty$, $h \to 0$, and $Nh \to \infty$, we obtain

$$
\mathbb{E}\left[\hat{f}(x_0)\right] = f(x_0) + \frac{h^2}{2} f''(x_0) \int z^2 K(z)dz + O(h^4) \to f(x_0),
$$

$$
\text{var}\left(\hat{f}(x_0)\right) = \frac{1}{Nh} f(x_0) \int K(z)^2 dz + O\left(\frac{1}{N}\right) \to 0,
$$

and thus,

$$
\hat{f}(x_0) \overset{p}{\to} f(x_0).
$$

But this doesn’t tell us how to choose $h$ in practice.
Kernel density estimation: Choosing the bandwidth

The **mean square error** of an estimator is equal to the square of the bias plus the variance.

\[
MSE(\hat{f}(x_0)) = \mathbb{E}\left[ (\hat{f}(x_0) - f(x_0))^2 \right] \\
= \left( \mathbb{E}[\hat{f}(x_0)] - f(x_0) \right)^2 + \text{var}(\hat{f}(x_0)).
\]

A popular way to choose \( h \) is to minimize the **integrated mean square error**:

\[
IMSE(\hat{f}) = \int MSE(\hat{f}(x)) \, dx \\
= \frac{h^4}{4} \int (f''(x))^2 \, dx \left( \int z^2 K(z) \, dz \right)^2 + \frac{1}{Nh} \int K(z)^2 \, dz
\]

(plus smaller order terms).
Kernel density estimation: Choosing the bandwidth

Minimizing the IMSE with respect to $h$ we obtain:

$$h^* = \left( \frac{\int K(z)^2 \, dz}{\int (f''(x))^2 \, dx \left( \int z^2 K(z) \, dz \right)^2} \right)^{1/5} \frac{1}{N^{1/5}}.$$ 

The optimal bandwidth decreases at a rate $N^{1/5}$. The optimal MSE is then of order $N^{-4/5}$.

The result depends on $f''$, which is unknown to us. A simple rule of thumb is obtained assuming that $f$ is normal. This is called the normal reference rule. If $K(z)$ is also normal, then

$$h^* = \frac{1.059 \, \sigma}{N^{1/5}},$$

where $\sigma^2 = \text{var}(X)$, which can be estimated in the data.
Kernel density estimation: Choosing the bandwidth

Estimate of world income distribution (across countries) with bandwidth given by the normal reference rule.

kernel: normal, bandwidth = 0.68325676
Kernel density estimation: Choosing the bandwidth

The reference rule can be adapted to other kernels, changing the values of $\int K(z)^2 dz$ and $\int z^2 K(z) dz$ in the formula of $h^*$. 

For example:

- **Triangular**: $K(z) = (1 - |z|) \mathbb{1}(|z| \leq 1)$, then 
  \[ h^* = \frac{2.576 \sigma}{N^{1/5}}. \]

- **Epanechnikov**: $K(z) = (3/4)(1 - z^2) \mathbb{1}(|z| \leq 1)$, then 
  \[ h^* = \frac{2.345 \sigma}{N^{1/5}}. \]

$\hat{f}(x)$ is typically fairly insensitive to the choice of kernel, as long as the optimal bandwidth is used for each kernel. Theoretically, the Epanechnikov kernel minimizes MSE.
Kernel density estimation: Choosing the bandwidth

Kernel density estimates using the normal reference rule.

![Graph showing kernel density estimates using different kernels: Normal, Triangular, and Epanechnikov.](image)
Kernel density estimation: Choosing the bandwidth

Other proposals to choose $h$:

- **Eyeballing**.
- **Silverman’s rule of thumb**. The normal reference rule may oversmooth bimodal distributions. For a normal kernel, Silverman proposes to reduce the factor 1.059 to 0.90 and to use the minimum of two estimators of the standard deviation:

$$
\hat{h}^* = \frac{0.90 \min\{\hat{\sigma}, \frac{IQR}{1.349}\}}{N^{1/5}}.
$$

where $\hat{\sigma}$ is the sample standard deviation and $\hat{IQR}$ is the sample interquartile range (notice that $\sigma = IQR/1.349$ for a normal distribution with variance $\sigma^2$).
Kernel density estimation: Choosing the bandwidth

Estimate of world income distribution (across countries) with bandwidth given by the Silverman’s rule of thumb.
Other proposals to choose $h$ (cont.):

- **Plug-in methods.** Estimate $f''$ using a pilot bandwidth and plug-in $\int (\hat{f}''(x))^2 dx$ in the formula for $h^*$. Difficult because a (different!) optimal bandwidth is needed to estimate $f''$.

- **Cross validation.** An approximately unbiased estimate of the integral squared error $\int (\hat{f}(x) - f(x))^2 dx$, up to a constant, is

\[
CV(h) = \frac{1}{N^2 h} \sum_{i=1}^{N} \sum_{j=1}^{N} K^{(2)} \left( \frac{X_i - X_j}{h} \right) - \frac{2}{N} \sum_{i=1}^{N} \hat{f}_{-i}(X_i),
\]

where $K^{(2)}(z) = \int K(a)K(z-a) da$ and $\hat{f}_{-i}$ is the kernel density estimator that excludes observation $i$ from the sample. Let $h_{CV}$ be the bandwidth that minimizes $CV(h)$. It can be shown that, under some conditions:

\[
\frac{h_{CV}}{h^*} \xrightarrow{p} 1.
\]

However, this convergence is very slow.
Multivariate density estimation

Kernel density estimator for $X \in \mathbb{R}^k$:

$$\hat{f}(x_0) = \frac{1}{Nh^k} \sum_{i=1}^{N} K\left(\frac{X_i - x_0}{h}\right).$$

$K(z)$ typically is a **product kernel**, that is a product of $k$ univariate kernels

$$K(z) = K_1(z_1)K_2(z_2)\cdots K_k(z_k),$$

where $z = (z_1, \ldots, z_k)'$. In this case, if $N \to \infty$, $h \to 0$, and $Nh^k \to \infty$, then

$$\hat{f}(x_0) \overset{p}{\to} f(x_0).$$

If in addition $Nh^{k+4} \to 0$, then

$$\sqrt{Nh^k}(\hat{f}(x_0) - f(x_0)) \overset{d}{\to} N\left(0, f(x_0) \int K(z)^2 dz\right).$$
For case of $k = 1$, the best possible rate for the IMSE is $N^{-4/5}$.

Doing the same calculations for a general value of $k$, it can be shown that the best possible rate for the IMSE is $N^{-4/(k+4)}$.

The best rate deteriorates when $k$ is large.

This is a manifestation of the **curse of dimensionality**: The higher the dimension of the $X$ space, the less likely is it that one of the $N - 1$ other data points will be close (in every dimension) to $x_0$. This greatly complicates nonparametric estimation.
Multivariate density estimation: Curse of dimensionality

The curse of dimensionality is illustrated in the following table in Silverman (1986).

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>k</td>
<td>N</td>
<td></td>
<td>k</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>-----</td>
<td>---</td>
<td>---</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td></td>
<td>6</td>
<td>2790</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>19</td>
<td></td>
<td>7</td>
<td>10700</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>67</td>
<td></td>
<td>8</td>
<td>43700</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>223</td>
<td></td>
<td>9</td>
<td>187000</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>768</td>
<td></td>
<td>10</td>
<td>842000</td>
<td></td>
</tr>
</tbody>
</table>

The table considers the case of a normal density and a normal kernel and reports the number of observations required to attain:

\[
\mathbb{E} \left[ \frac{(\hat{f}(0) - f(0))^2}{f(0)^2} \right] < 0.1.
\]
Higher-order kernels

From the bias formula, if we changed the kernel so that

$$\int z^2 K(z) dz = 0,$$

then the order of the bias would be $O(h^4)$ instead of $O(h^2)$.

A kernel of this type is called kernel of order four because its first non-zero moment is the fourth moment. Kernels of order four or larger are called higher-order kernels. Notice that higher-order kernels must take on negative values (unintuitive).

In fact, provided that higher order derivatives of $f$ exist, we could arbitrarily reduce the order of the bias by adopting a kernel of high order, without affecting the order of the variance.

Unfortunately, this is a theoretical property that does not work well in practice. For the sample sizes used typically in practice, higher-order kernels inflate the variance substantially.
Nonparametric regression

Nonparametric regression methods seek to estimate the population regression function

\[ m(x) = \mathbb{E}[Y \mid X = x], \quad x \in \mathbb{R}^k, \]

without taking a strong stand on the functional form of \( m(x) \).

We want the data to influence the estimate in a flexible way, so we don’t miss important features of \( m(x) \).

In a finite sample, we cannot hope to estimate arbitrarily wiggly/discontinuous functions \( m(x) \). But we can hope to do well when \( m(x) \) is smooth.

We will discuss three methods:

2. Kernel regression.
3. Local linear regression.
Linear regression

OLS: Assume linearity and minimize sum of squared residuals.

But regression function may not be linear. Linearity *may* be okay for prediction but can obscure important economic associations.
Bin scatter

If $X$ is a scalar, the most obvious nonparametric regression strategy is binning.

1. Group the data points $X_1, \ldots, X_N$ into a finite number $S$ of bins, like in histogram estimation.
2. Compute the average outcome $Y$ in each bin.
3. Plot the average outcomes against the midpoint of each bin.

Sometimes called **bin scatter**. Corresponds to regressing $Y$ on a collection of $S$ indicator functions of the form $\mathbb{1}(X_i \in [a_j, a_{j+1}))$, where $-\infty = a_0 < a_1 < \cdots < a_S = \infty$.

This is a visually appealing first pass. But:

- Throws away information by binning (unless $X$ is discrete).
- Doesn’t yield a useful estimate of $m(x)$ for every $x$.
- Not obvious how to pick the bins (deciles/percentiles/etc.).
Kernel regression (Nadaraya-Watson)

Suppose we want to estimate the regression function

\[ m(x_0) = \mathbb{E}[Y \mid X = x_0]. \]

A kernel regression is a weighted average

\[ \hat{m}(x_0) = \sum_{i=1}^{N} w_i Y_i, \]

with weights

\[ w_i = \frac{K\left(\frac{X_i - x_0}{h}\right)}{\sum_{j=1}^{N} K\left(\frac{X_j - x_0}{h}\right)}. \]

Observations close to \( x_0 \) get large weights and observations distant from \( x_0 \) get small weights.
Kernel regression (Nadaraya-Watson)

Kernel regression, bandwidth = 0.80
Kernel regression (Nadaraya-Watson)

Kernel regression, bandwidth = 0.80

log Gini index
3.2 3.4 3.6 3.8 4 4.2

log GDP per capita
4 6 8 10 12

Kernel weight
0 0.5 1
Kernel regression (Nadaraya-Watson)

Kernel regression, bandwidth = 0.80
Kernel regression (Nadaraya-Watson)

Kernel regression, bandwidth = 0.80

log Gini index

log GDP per capita

Kernel weight
Kernel regression (Nadaraya-Watson)

The bandwidth $h$ is a **smoothing parameter**:

$\Rightarrow$ Large $h$ makes regression smooth

$\Rightarrow$ Small $h$ makes regression wiggly
School infrastructure expenditures and school attendance in Peru (Paxson & Schady)

FIGURE 7. Nonparametric Regressions of Change in District School Attendance Rate and Per Capita FONCODES Education Expenditures in 1992-95 on (Modified) FONCODES Index.

Source: Authors’ calculations based on 1993 census data, the 1996 INE-I household survey data, and FONCODES data.

Yields a coefficient of 0.0151, which implies that a one-standard-deviation (9.6) increase in per capita FONCODES spending on school infrastructure is associated with a gain in the attendance rate of 14.5 percentage points.

Another interesting feature of figure 7 is that the relationships it shows are nonlinear and nonmonotonic. Districts with FONCODES index values between 22 and 26 had greater attendance gains and higher school infrastructure spending than did poorer districts with index values between 26 and 28. Both measures increased again for even poorer districts with index values greater than 28. These nonlinear patterns are not the result of few observations at very high values of the FONCODES index. About 15 percent of districts (accounting for 14 percent of children) had FONCODES index values between 22 and 26, 6.3 percent of districts (4.7 percent of children) had index values between 26 and 28, and 9.6 percent of districts (8.4 percent of children) had index values greater than 28. Furthermore, these nonlinearities remain even when a wide range of bandwidths is used for the nonparametric regressions. Although these nonlinearities are striking, it is not clear what drives them. The districts with index values between 26 and 28 were allocated more FONCODES funds than the wealthier This content downloaded on Sat, 19 Jan 2013 22:22:28 PM All use subject to JSTOR Terms and Conditions
Kernel regression: Choosing the bandwidth

- **Eyeballing.**

- **Plug-in:** Define the **mean square error** as:

\[
MSE(h) = \int (\hat{m}(x) - m(x))^2 f(x) dx.
\]

A MSE-minimizing bandwidth sequence is given by:

\[
h^* = c N^{-1/(k+4)},
\]

where the constant $c$ depends on $K(z)$, $m(x)$, and $f(x)$. Estimation of $c$ by plug-in is possible but cumbersome.
Kernel regression: Choosing the bandwidth

- **Cross validation**: Let

\[
CV(h) = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \hat{m}_{-i}(X_i))^2,
\]

where \( \hat{m}_{-i}(X_i) \) is the leave-\( i \)-out kernel regression estimator of \( m(X_i) \) (with bandwidth \( h \)). Let \( h_{CV} \) be the bandwidth sequence that minimizes \( CV(h) \).

It can be shown that:

\[
\frac{MSE(h_{CV})}{\min_h MSE(h)} \xrightarrow{p} 1.
\]
Kernel regression: Boundary bias

Consider $x_0$ at the boundary of the support of $X$.
Kernel regression: Boundary bias

Consider \( x_0 \) at the boundary of the support of \( X \).
Kernel regression: Boundary bias

Consider $x_0$ at the boundary of the support of $X$. 

![Graph showing kernel regression with boundary bias](image)

- True regression value
- Kernel regression estimate

Large bias near boundary because all observations that are close to the boundary have regression values smaller than the regression value at $x_0$. 
Kernel regression: Boundary bias

Consider $x_0$ at the boundary of the support of $X$. 

![Diagram showing true regression value and kernel regression estimate](image)
Consider $x_0$ at the boundary of the support of $X$.

$\Rightarrow$ Large bias near boundary because all observations that are close to the boundary have regression values smaller than the regression value at $x_0$. 
Series regression

Fit a polynomial of order $p$: Minimize

$$
\sum_{i=1}^{N} \left( Y_i - b_0 - b_1 X_i - b_2 X_i^2 - \cdots - b_p X_i^p \right)^2,
$$

$$
\hat{m}(x_0) = \hat{b}_0 + \hat{b}_1 x_0 + \cdots + \hat{b}_p x_0^p.
$$
Local linear regression

A popular halfway point between kernel regression and linear regression is **local linear regression**.

Fit a linear regression *locally* around each point \( x_0 \) of interest: At \( X = x_0 \), minimize the Weighted Least Squares criterion

\[
\sum_{i=1}^{N} K \left( \frac{X_i - x_0}{h} \right) (Y_i - b_0 - b_1 X_i)^2,
\]

and compute the fitted value \( \hat{m}(x_0) = \hat{b}_0 + \hat{b}_1 x_0 \). Do this separately for each \( x_0 \) (with different \( \hat{b}_0, \hat{b}_1 \) at each \( x_0 \)).

Only requires Weighted Least Squares calculations.

This can potentially remove boundary bias if the true regression function \( m(x_0) \) is approximately linear close to the boundary.
Local linear regression

Local linear regression, bandwidth = 0.80

Local linear regression, bandwidth = 0.40
Local polynomial regression

Local polynomial regression extends this idea to polynomials:

$$\sum_{i=1}^{N} K\left(\frac{X_i - x_0}{h}\right) (Y_i - b_0 - b_1 X_i - \cdots - b_p X_i^p)^2.$$ 

Kernel regression can be viewed as a particular case of local polynomial regression that uses $p = 0$ (that is, the regression includes only an intercept).

(Series regression can be viewed as a particular case of local polynomial regression that uses a constant kernel.)

Local polynomial regression can be computed as the intercept value $\hat{b}_0$ obtained from minimizing the WLS criterion

$$\sum_{i=1}^{N} K\left(\frac{X_i - x_0}{h}\right) (Y_i - b_0 - b_1 (X_i - x_0) - \cdots - b_p (X_i - x_0)^p)^2.$$ 

The $\nu$-th derivative of the regression function at $x_0$ can be estimated by $\nu! \hat{b}_\nu$. 
Polynomial order and bandwidth selection

Local linear regression (and series regression) have lower boundary bias than kernel regression.

(For series regression, the order $p$ of the series is typically guided by cross-validation.)

For local polynomial regression, the order of the polynomial is typically chosen to be $\nu + 1$, where $\nu$ is the order of the derivative of interest ($\nu = 0$ if we aim to estimate $\mathbb{E}[Y \mid X = x_0]$).

There are plug-in methods to choose bandwidth for local polynomial regression. Cross-validation can also be applied.