Nonparametric Estimation

Jacob M. Montgomery

2018

Nonparametric estimation

Overview

- So far we have talked about a number of ways to make estimates about population parameters.

Overview

- So far we have talked about a number of ways to make estimates about population parameters.
- All of these methods are based on a "model" of the DGP.

Overview

- So far we have talked about a number of ways to make estimates about population parameters.
- All of these methods are based on a “model” of the DGP.
- Today, we will move away from this approach and instead try and learn from our data without assuming much of anything.
  1. Find some pointwise estimate
  2. Calculate it for the whole sample.
  3. Do something to the sample and re-calculate.
  4. Repeat and summarize.

How to think about all of this

- We generally have a lot of tools for generating point estimates.
- For example, we can estiamte μ for any population without making *any* assumptions about the DGP at all.

How to think about all of this

- We generally have a lot of tools for generating point estimates.
- For example, we can estiamte μ for any population without making *any* assumptions about the DGP at all.
- Where we get hung up is in calculating the standard errors.
    - Thinking broadly, *most* possible estimands do not have known asymptotic standard errors.
    - The best we can do in other circumstances is use the delta method (or higher-order Taylor approximations).

How to think about all of this

- We generally have a lot of tools for generating point estimates.
- For example, we can estiamte μ for any population without making *any* assumptions about the DGP at all.
- Where we get hung up is in calculating the standard errors.
  - Thinking broadly, *most* possible estimands do not have known asymptotic standard errors.
  - The best we can do in other circumstances is use the delta method (or higher-order Taylor approximations).
- (You will see a lot of estimands that can be "non-parametrically" identified. Always think about where their error bars came from.)
- But what we really need are more general ways to estimate standard errors.

Example: The Jackknife

- We have $X_1, \ldots, X_n$ that are iid samples from some unknown function or function space

$$X_i \sim F$$

- We calculate some statistic based on the data $\hat{\theta} = t(\mathbf{x})$.

Example: The Jackknife

- We have $X_1, \ldots, X_n$ that are iid samples from some unknown function or function space

$$X_i \sim F$$

- We calculate some statistic based on the data $\hat{\theta} = t(\mathbf{x})$. This is our estimate.
- Let $\mathbf{x}_{(i)}$ be the sample with the i^{th} observation removed.

$$x_{(i)} = (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)'$$

- Let $\hat{\theta}_{(i)} = t(\mathbf{x}_{(i)})$

- Then the jackknife estimate of the standard error of $\hat{\theta}$ is

$$\hat{se}_{jack} = \left[\frac{n-1}{n}\sum_{1}^{n}\left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}\right)^2\right]^{1/2}$$

- Where $\theta_{(\cdot)} = \sum_1^n \theta_{(i)}/n$

Discussion: Jackknife and the sample mean

- When $\bar{x} = \hat{\theta}$, what is the jackknife standard error?

Discussion: Jackknife and the sample mean

- When $\bar{x} = \hat{\theta}$, what is the jackknife standard error?
- $\hat{\theta}_{(i)} = (n\bar{x} - x_i)/(n-1)$ and $\hat{\theta}_{(\cdot)} = \bar{x}$
- $\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)} = (\bar{x} - x_i)/(n-1)$

Discussion: Jackknife and the sample mean

- When $\bar{x} = \hat{\theta}$, what is the jackknife standard error?
- $\hat{\theta}_{(i)} = (n\bar{x} - x_i)/(n-1)$ and $\hat{\theta}_{(\cdot)} = \bar{x}$
- $\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)} = (\bar{x} - x_i)/(n-1)$
- $\hat{se}_{jack} = \frac{s}{\sqrt{n}}$

Example: Correlation coefficient

- Imagine we have two iid random samples of the same size, $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$. The correlation between these two variables is:

$$\frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i^n (x_i - \bar{x})^2 \sum_i^n (y_i - \bar{y})^2}}$$

Example: Correlation coefficient

- Imagine we have two iid random samples of the same size, $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$. The correlation between these two variables is:
$$\frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i^n (x_i - \bar{x})^2 \sum_i^n (y_i - \bar{y})^2}}$$

- We often talk about the "correlation" between two variables, but the asymptotic standard errors provided in most textbooks (and the defaults in R) assume that both are normal.

Example: Correlation coefficient

- Imagine we have two iid random samples of the same size, $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$. The correlation between these two variables is:

$$\frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i^n (x_i - \bar{x})^2 \sum_i^n (y_i - \bar{y})^2}}$$

- We often talk about the "correlation" between two variables, but the asymptotic standard errors provided in most textbooks (and the defaults in R) assume that both are normal.
- What if we want to be more agnostic?

Example: Correlation coefficient

- Imagine we have two iid random samples of the same size, $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$. The correlation between these two variables is:

$$\frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i^n (x_i - \bar{x})^2 \sum_i^n (y_i - \bar{y})^2}}$$

- We often talk about the "correlation" between two variables, but the asymptotic standard errors provided in most textbooks (and the defaults in R) assume that both are normal.
- What if we want to be more agnostic?
  - Could approximate using the delta method? Gets pretty nasty.

Example: Correlation coefficient

- Imagine we have two iid random samples of the same size, $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$. The correlation between these two variables is:
$$\frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i^n (x_i - \bar{x})^2 \sum_i^n (y_i - \bar{y})^2}}$$
- We often talk about the "correlation" between two variables, but the asymptotic standard errors provided in most textbooks (and the defaults in R) assume that both are normal.
- What if we want to be more agnostic?
  - Could approximate using the delta method? Gets pretty nasty.
  - Could approximate using the jackknife

Example: Correlation coefficient

$$\frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i^n (x_i - \bar{x})^2 \sum_i^n (y_i - \bar{y})^2}}$$

- For a sample of

```
x=c(2, 4, 5, 3, 8, 10)
y=c(-2, 1, -1, 2, 5, 5)
```

- Plot the two variables
- Estimate the correlation (use the cor() function)
- Find the jackkknife standard error

Jackknife discussion

- This is a truly non-parametric method.

Jackknife discussion

- This is a truly non-parametric method.
- There is a hidden assumption here that $s(\mathbf{x})$ will not deviate wildly when sample size is $n-1$. Can be weird for things like medians (defined differently when sample size is odd or even).

Jackknife discussion

- This is a truly non-parametric method.
- There is a hidden assumption here that $s(\mathbf{x})$ will not deviate wildly when sample size is $n-1$. Can be weird for things like medians (defined differently when sample size is odd or even).
- Estimates of standard errors can be upwardly biased.
- Behaves poorly where local derivatives are eratic.

The bootstrap

- We are trying to understand how some statistic $\hat{\theta}$ varies around our estimate.

The bootstrap

- We are trying to understand how some statistic $\hat{\theta}$ varies around our estimate.
- From a frequentist perspective, we would like to re-sample from our distribution $\mathcal{F}$ and then re-calculate over and over and over again. But how?

The bootstrap

- We are trying to understand how some statistic $\hat{\theta}$ varies around our estimate.
- From a frequentist perspective, we would like to re-sample from our distribution $\mathcal{F}$ and then re-calculate over and over and over again. But how?
- What if we used the empirical distribution we have to estimate $\mathcal{F}$ itself and then sampled from that distribution?

Estimating an unknowing distribution

The **empirical distribution function** $\hat{\mathcal{F}}$ is a CDF that puts a mass of $\frac{1}{n}$ at each data point X_i.

$$\hat{\mathcal{F}}(x) = \frac{\sum_{i=1}^{n} I(X_i \leq x)}{n}$$

where $I(\cdot)$ is an indicator function.

Class exercise: Estimate $\mathcal{F}$ for the following dataset

```
x=c(2, 4, 5, 3, 8, 10, -2, 1, -1, 2, 5, 5)
```

Discussion of the empirical estimation of an unknown funciton

1. $$E(\hat{\mathcal{F}}(x)) = F(x)$$

Discussion of the empirical estimation of an unknown funciton

1. $$E(\hat{\mathcal{F}}(x)) = F(x)$$

2. $$Var(\hat{\mathcal{F}}(x)) = \frac{F(x)\,(1 - F(x))}{n}$$

Discussion of the empirical estimation of an unknown funciton

1. $$E(\hat{\mathcal{F}}(x)) = F(x)$$

2. $$Var(\hat{\mathcal{F}}(x)) = \frac{F(x)\,(1 - F(x))}{n}$$

3. $$MSE = \frac{F(x)\,(1 - F(x))}{n} \to 0,$$

4. $\hat{\mathcal{F}}(x)$ converges in probability to $F(X)$.

Moving from the estimate of the distribution to the bootstrapped standard error

- Now that we have an estimate of $\mathcal{F}$, we want to imagine "sampling" repeatedly from it.
- In some sense, this is the same "plug in" estimation method we have used before.

Moving from the estimate of the distribution to the bootstrapped standard error

- Now that we have an estimate of $\mathcal{F}$, we want to imagine "sampling" repeatedly from it.
- In some sense, this is the same "plug in" estimation method we have used before.
- The idea is that we sample from from $\hat{\mathcal{F}}$

Moving from the estimate of the distribution to the bootstrapped standard error

- Now that we have an estimate of $\mathcal{F}$, we want to imagine "sampling" repeatedly from it.
- In some sense, this is the same "plug in" estimation method we have used before.
- The idea is that we sample from from $\hat{\mathcal{F}}$
- We can then calculate our estimate over and over and over again and via this method get an approximation of the frequentist standard error without making any parametric assumptions.

How to calculate a bootstrapped standard error

- We have $X_1, \ldots, X_n$ that are iid samples from some unknown function or function space

$$X_i \sim F$$

.

How to calculate a bootstrapped standard error

- We have $X_1, \ldots, X_n$ that are iid samples from some unknown function or function space

$$X_i \sim F$$

.

- We calculate some statistic based on the data $\hat{\theta} = t(\mathbf{x})$.

How to calculate a bootstrapped standard error

- We have $X_1, \ldots, X_n$ that are iid samples from some unknown function or function space

$$X_i \sim F$$

.

- We calculate some statistic based on the data $\hat{\theta} = t(\mathbf{x})$. This is our estimate.

How to calculate a bootstrapped standard error

- We have $X_1, \dots, X_n$ that are iid samples from some unknown function or function space

$$X_i \sim F$$

.

- We calculate some statistic based on the data $\hat{\theta} = t(\mathbf{x})$. This is our estimate.
- Now we draw a *bootstrap sample* $\mathbf{x}^* = (x_1^*, x_2^*, \dots, x_n^*)$ where *each* x_i^* is drawn randomly with equal probability and **with replacement**.

How to calculate a bootstrapped standard error

- We have $X_1, \ldots, X_n$ that are iid samples from some unknown function or function space

$$X_i \sim F$$

.
- We calculate some statistic based on the data $\hat{\theta} = t(\mathbf{x})$. This is our estimate.
- Now we draw a *bootstrap sample* $\mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)$ where *each* x_i^* is drawn randomly with equal probability and **with replacement**. Why?

How to calculate a bootstrapped standard error

- We have $X_1, \ldots, X_n$ that are iid samples from some unknown function or function space
$$X_i \sim F$$
.
- We calculate some statistic based on the data $\hat{\theta} = t(\mathbf{x})$. This is our estimate.
- Now we draw a *bootstrap sample* $\mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)$ where *each* x_i^* is drawn randomly with equal probability and **with replacement**. Why?
- Calculate $\hat{\theta}^* = t(\mathbf{x}^*)$

How to calculate a bootstrapped standard error

- We have $X_1, \ldots, X_n$ that are iid samples from some unknown function or function space

$$X_i \sim F$$

.

- We calculate some statistic based on the data $\hat{\theta} = t(\mathbf{x})$. This is our estimate.
- Now we draw a *bootstrap sample* $\mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)$ where *each* x_i^* is drawn randomly with equal probability and **with replacement**. Why?
- Calculate $\hat{\theta}^* = t(\mathbf{x}^*)$
- Repeat

- Let B indicate the total number of bootstrap samples and b index each such that:

$$\hat{\theta}^{*(b)} = t(\mathbf{x}^{*(b)})$$

- Let B indicate the total number of bootstrap samples and b index each such that:

$$\hat{\theta}^{*(b)} = t(\mathbf{x}^{*(b)})$$

- For some large number B (B=500 will do for calculating standard errors)

$$\hat{se}_{boot} = \left[\frac{\sum_{b=1}^{B} (\hat{\theta}^{*(b)} - \hat{\theta}^{*\cdot})^2}{B-1} \right]^{1/2},$$

- Where

$$\hat{\theta}^{*\cdot} = \frac{\sum_{1}^{B} \hat{\theta}^{*(b)}}{B}$$

Example: Correlation coefficient

$$\frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i^n (x_i - \bar{x})^2 \sum_i^n (y_i - \bar{y})^2}}$$

- For a sample of

```
x=c(2, 4, 5, 3, 8, 10)
y=c(-2, 1, -1, 2, 5, 5)
```

- Estimate the correlation (use the cor() function)
- Find the non-parametric bootstrapped standard error

Discussion

- This is a completely non-parametric approach
- More dependable (but more computationally intensive) than the jackknife
- We could just as easily have calculated the absolute bias, the MSE or anything else.

Class exercise

- For a sample of

```r
x=c(2, 4, 5, 3, 8, 10, -2, 1, -1, 2, 5, 5)
```

1. Find the bootstrapp/jackknife SE for the median.
2. Estimate the absolute error $E(|\hat{\theta} - \theta|)$ for the median and the mean using the bootstrap.

Problems with re-sampling methods

- One of the assumptions we did *not* get rid of was the iid assumption.

Problems with re-sampling methods

- One of the assumptions we did *not* get rid of was the iid assumption.
- It is important realize that *everything* we have done in this session rests on the assumption that we can collect truly independent draws.

Problems with re-sampling methods

- One of the assumptions we did *not* get rid of was the iid assumption.
- It is important realize that *everything* we have done in this session rests on the assumption that we can collect truly independent draws.
- In many (most?) cases, our data do not meet this assumption.
- In these cases, the bootstrap must be used carefully or not at all.

Moving blocks bootstrap

- Suppose that $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ instead of being iid is a time series dataset.

Moving blocks bootstrap

- Suppose that $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ instead of being iid is a time series dataset.
- We can create a set of "blocks" that can be re-sampled

$$\{(x_1, x_2, x_3), (x_2, x_3, x_4), \ldots (x_{n-2}, x_{n-1}, x_n)\}$$

- The length of the block must be chosen so that the correlations of the items at the beginning of the block and the end are small.

Moving blocks bootstrap

- Suppose that $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ instead of being iid is a time series dataset.
- We can create a set of "blocks" that can be re-sampled

$$\{(x_1, x_2, x_3), (x_2, x_3, x_4), \ldots (x_{n-2}, x_{n-1}, x_n)\}$$

- The length of the block must be chosen so that the correlations of the items at the beginning of the block and the end are small.
- We can then construct a bootstrap sample by sampling from these blocks rather than from the "raw" sample.

Moving blocks bootstrap

- Suppose that $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ instead of being iid is a time series dataset.
- We can create a set of "blocks" that can be re-sampled

$$\{(x_1, x_2, x_3), (x_2, x_3, x_4), \ldots (x_{n-2}, x_{n-1}, x_n)\}$$

- The length of the block must be chosen so that the correlations of the items at the beginning of the block and the end are small.
- We can then construct a bootstrap sample by sampling from these blocks rather than from the "raw" sample.
- An even more complex scheme must be developed for time-series cross-sectional datsets, and insufficient work has been done in this area to recommend a "clean" solution.

Additional bootstrap schemes

- Bayesian bootstraps
- "Robust" estimation for quantities such as the trimmed mean to reduce the effect of out outliers
- See Efron and Hastie Chapter 10 for additional details

Example: Lowess curves

- Lowess draws a line through a scatterplot using kernal smoothing over local regions of the covariate space.

Example: Lowess curves

- Lowess draws a line through a scatterplot using kernal smoothing over local regions of the covariate space.
- For each possible value of $x = x_0$, we predict according to the formula:

$$\hat{f}(x_0) = \sum_{i=1}^{n} y_i K_\gamma(x_0, x_i)$$

Example: Lowess curves

- Lowess draws a line through a scatterplot using kernal smoothing over local regions of the covariate space.
- For each possible value of $x = x_0$, we predict according to the formula:

$$\hat{f}(x_0) = \sum_{i=1}^{n} y_i K_\gamma(x_0, x_i)$$

- In words this mean that we estimate the outcome to be similar to outcomes that are "close to" the observation in terms of the "Kernal" $K_\gamma(x_0, x_i)$
- The γ subscript indicates that we ignore some subset of observations that are too far away.

Example: Lowess curves

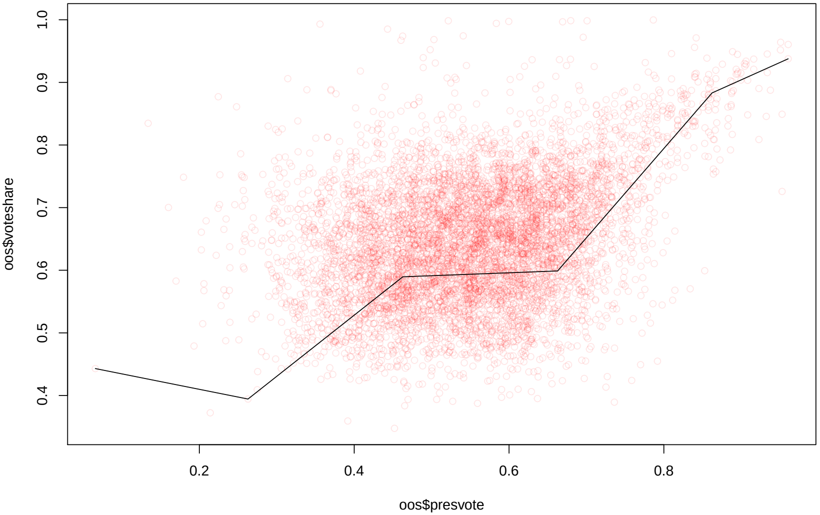
- Lowess draws a line through a scatterplot using kernal smoothing over local regions of the covariate space.
- For each possible value of $x = x_0$, we predict according to the formula:

$$\hat{f}(x_0) = \sum_{i=1}^{n} y_i K_\gamma(x_0, x_i)$$

- In words this mean that we estimate the outcome to be similar to outcomes that are "close to" the observation in terms of the "Kernal" $K_\gamma(x_0, x_i)$
- The γ subscript indicates that we ignore some subset of observations that are too far away.
- The tricubic kernal used by default in R is:

$$K_s(x_0, x_i) = \begin{cases} (1 - u_i^3)^3 \text{ if } u_i \leq 1 \\ 0 \text{ otherwise} \end{cases}$$

```r
oos<-read.csv(file =
url("https://jmontgomery.github.io/ProblemSets/incumbents.csv"))
myFit<-lowess(oos$voteshare~oos$presvote, delta=.2 )
plot(oos$presvote, oos$voteshare, col=rgb(1,0,0, alpha=.1))
points(myFit, type="l")
```



For problem set exercise

1. Use the bootstrap method (assuming iid) to estimate the 95% CI for this curve.
2. Add the CI to the plot.
3. Re-do this for several different values of delta (small and big).
4. What is driving this result?