Data Science for Everyone

Week 11: Bootstrap and confidence intervals

Angela Lai

New York University

Outline

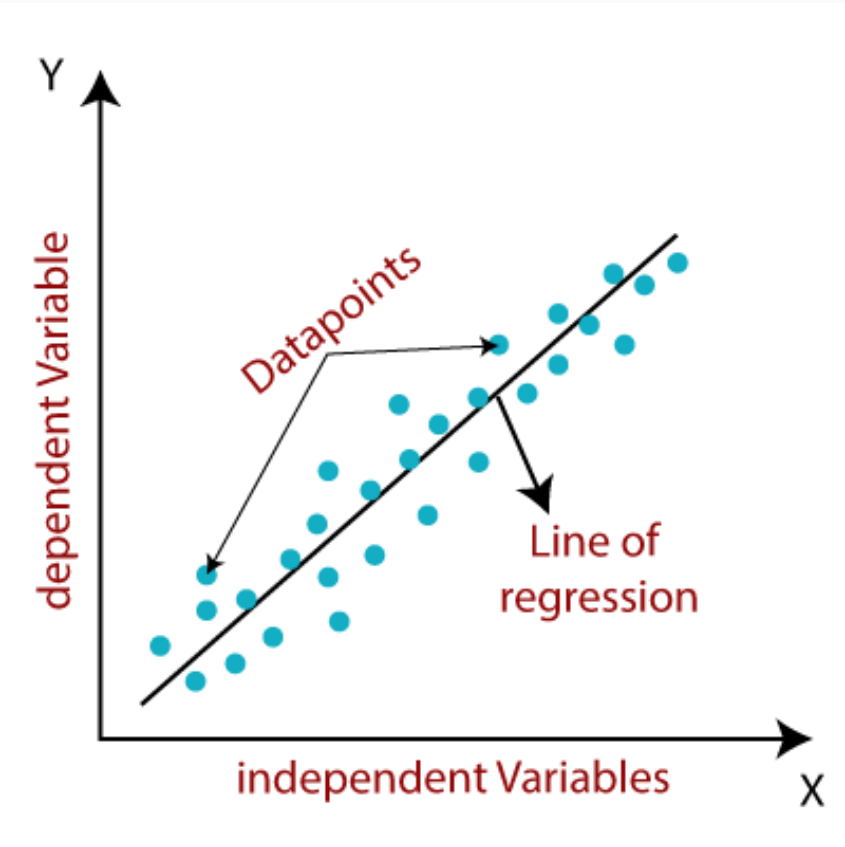
- Logistics
- Demo with notebook
- Questions?

Logistics

- Lab 8 out, due at 8 p.m. ET on April 22
- Homework 3/4 out, due at 8 p.m. ET on April 27
- Remember to get started on your project if you haven't already!

Logistics

- Note this project requirement: ”You may select any topic and use any dataset that you like as long as it’s publicly available and it contains **two continuous variables whose association you are interested in examining.**”
- The focus of the project analysis is simple linear regression



Review Questions: Bootstrapping & CIs

In a large random sample of U.S. households, the median annual income is $54,000. Researchers bootstrap this original sample 5,000 times and the sample median is recorded for each of the bootstrap samples. The middle 95% interval of these values is ($53,000, $55,000).

Review Questions: Bootstrapping & CIs

1. True or false: The interval ($53,000, $55,000) is an approximate bootstrap 95% confidence interval for the median income of all the households in the sample.
2. The percent of all U.S. households with annual incomes in the range ($53,000, $55,000)
   (i) is about 95%.
   (ii) is about 50%.
   (iii) cannot be approximated based on the information given.
3. If you calculate the mean of each of the 5,000 bootstrap samples and take the middle 95% interval of the 5,000 means, the center of the new interval will be
   (i) less than $54,000.
   (ii) about $54,000.
   (iii) more than $54,000.

Review Questions: Bootstrapping & CIs

1. True or false: The interval ($53,000, $55,000) is an approximate bootstrap 95% confidence interval for the median income of all the households in the sample.
2. The percent of all U.S. households with annual incomes in the range ($53,000, $55,000)
   (i) is about 95%.
   (ii) is about 50%.
   (iii) cannot be approximated based on the information given.
3. If you calculate the mean of each of the 5,000 bootstrap samples and take the middle 95% interval of the 5,000 means, the center of the new interval will be
   (i) less than $54,000.
   (ii) about $54,000.
   (iii) more than $54,000.

Review Questions: Bootstrapping & CIs

True/False: The researchers are estimating that the median household income in the city is between $53,000 and $55,000, but they could be wrong.

True/False: If the researchers had constructed an approximate 90% confidence interval based on the same boot-strap samples they used for the 95% interval, then both ends of their 90% confidence interval would have been inside the range $53,000 to $55,000.

Review Questions: Bootstrapping CIs

Let's say we're interested in the mean of a population and have a large random sample from it. We take 5000 bootstrap samples and calculate the sample mean for each one, obtaining a given 95% confidence interval for the population mean. Call this confidence interval c_1.

Which of these interpretations is correct?

- Our parameter of interest (the population mean) lies in c_1 with 95% probability.
- If we repeat this process n times and get a CI for the mean each time (obtaining $c_2, c_3, ..., c_n$) 95% of the time, the true value lies in c_1.
- If we repeat this process n times and get a CI for the mean each time, the true value lies in about 95 percent of the calculated CIs $(c_1, c_2, c_3, ..., c_n)$.

Review Questions: Bootstrapping and CIs

Let's say we're interested in the mean of a population and have a large random sample from it. We take 5000 bootstrap samples and calculate the sample mean for each one, obtaining a given 95% confidence interval for the mean. Call this confidence interval c_1.

Which of these interpretations is correct?

- Our parameter of interest (the population mean) lies in c_1 with 95% probability.
- If we repeat this process n times and get a CI for the mean each time (obtaining $c_2, c_3, ..., c_n$) 95% of the time, the true value lies in c_1.
- *If we repeat this process n times and get a CI for the mean each time, the true value lies in about 95 percent of the calculated CIs* $(c_1, c_2, c_3, ..., c_n)$.

Concept Review: Confidence Intervals

Key points:

- The "true value" of the mean ***does not change***! Our parameter—in this case, the population mean—is fixed.
- The calculated CI will be different each time we take our 5000 bootstrap samples to get the "distribution" of the mean.
- If we're given a defined CI, the parameter (which, again, is just a fixed number) either does or does not lie in that CI. There is no in-between.

Demo Time

Time for a demo using Will's notebook!

Questions?

Any questions?