Simulation-based CIs for SLR

11/13/24

Housekeeping

- Office hours changed this week:
  - Wednesday (today!) 4-5pm
  - Friday: cancelled, moved to next week before midterm
- Coding practice due tonight

Recap

- Point estimates (b_0, b_1) also have variability as their specific values depend on a given set of data
- We saw how to use output from `lm()` to test hypotheses about and create confidence intervals for β_0 and β_1
  - Relies on LINE conditions being met
- Let's turn to simulation-based techniques (good refresher before midterm!)

Bootstrap CI for slope

`evals` data

First six observations:

course_id	prof_id	score	bty_avg
1	1	4.7	5
2	1	4.1	5
3	1	3.9	5
4	1	4.8	5
5	2	4.6	3
6	2	4.3	3

Recall our model:

$$\underbrace{\text{score}}_{y} = \beta_0 + \beta_1 \underbrace{\text{bty\_avg}}_{x} + \epsilon$$

- We can index to denote specific row/observation pairs (x_i, y_i)
  - e.g. $(x_1, y_1) = (5, 4.7)$

Bootstrapping

- Repeat B times:
  1. Sample with replacement from the original data, of the same sample size as the original data
  2. Calculate the quantity of interest using the resampled data
- In the case of SLR: what exactly should we be "resampling"? What is the quantity/quantities of interest?

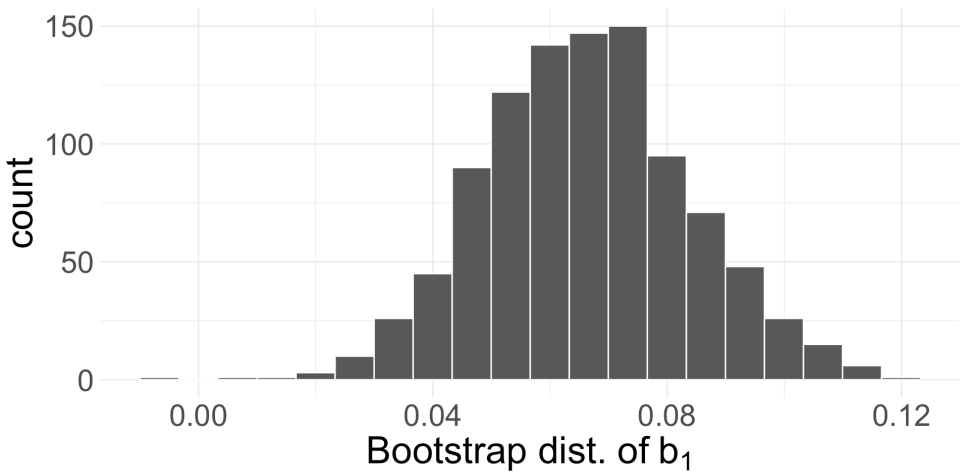
Bootstrapping for SLR

1. For a given observation i, we need to keep (x_i, y_i) together
    - Want to keep pairs of `score` and `bty_avg` together, but different pairs may be re-sampled
    - We will re-sample with replacement *row-by-row*
2. For each re-sampled dataset, fit a linear regression model and record b_1

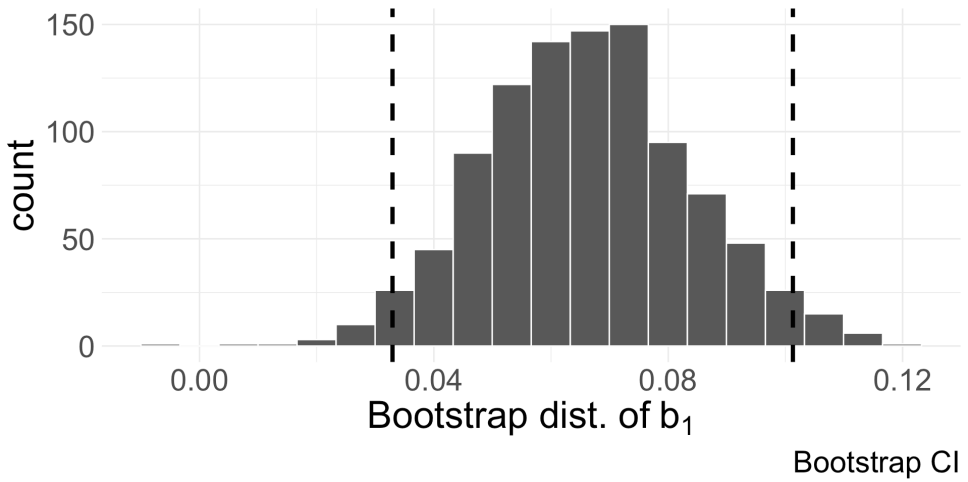
- This yields bootstrap distribution of estimated slope coefficients!

Live code for bootstrapped slope

Bootstrap distribution of b_1:



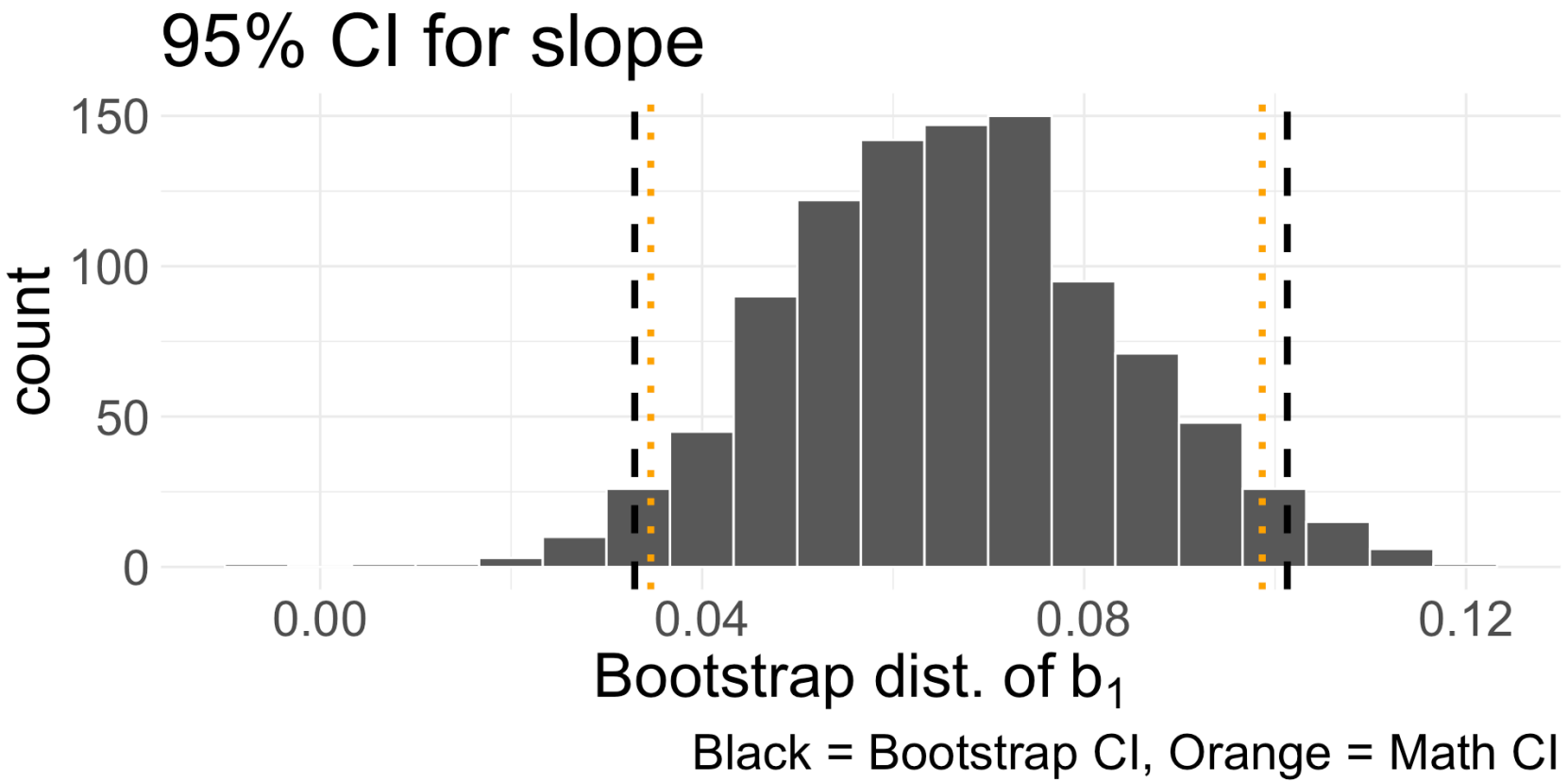
95% bootstrap CI for β_1: (0.033, 0.101):



Confidence intervals

term	estimate	std.error	statistic	p.value
(Intercept)	3.880338	0.0761430	50.961213	0.00e+00
bty_avg	0.066637	0.0162912	4.090382	5.08e-05

Compare to our 95% CI for β_1 using mathematical model: (0.035, 0.099)



Looking towards testing

Recall our hypotheses for the slope: $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$

How might we use simulation to test these hypotheses? (i.e. how can we simulate "null world"?)