Econometrics II
Prof. Brendan Kline

Homework on Bootstrap

Due on September 25 (in class)

This homework asks you to explore the bootstrap.

For questions that require the use of a computer, you should be sure to submit your code and output as part of your response. For some questions, the code/output will be the entirety of your response, whereas for some other questions, the code/output will only be part of your response, along with a written answer. There should be code/output that specifically addresses question 1, and then different code/output that specifically addresses question 2, etc., as appropriate. In other words, be clear about what part of the code/output addresses each question. Be sure to make your code/output legible! Also, be sure to make a clear distinction between the "code/output" part of your answer, and the normal written part of your answer.

I recommend using Stata for this project, although any statistical software package is acceptable. I do give explicit reference to certain Stata commands or help files. If you use something other than Stata, you are responsible for translating to that software.

Questions about the bootstrap with estimating an ordinary least squares regression:

We will explore how the bootstrap provides an approximation to the sampling distribution of the estimates from an ordinary linear regression.

First, we will numerically simulate the true sampling distribution.

1. Run the following code to add a Stata program called `genest` that generates a sample of $N = 50$ observations of the model

   $$y = -2 + 1.5x + \epsilon$$

   where $x \sim N(3, 4)$ and $\epsilon \sim N(0, 1)$. And then, `genest` estimates the model using ordinary least squares.

   ```
   program genest
   drop _all
   set obs 50
   gen eps = rnormal()
   gen x = 3 + 2*rnormal()
   gen y = -2 + 1.5*x + eps
   regress y x
   end
   ```

There is nothing to submit for this question.

2. Run the command `genest` to see what the program does.

   Include the output in your answers. And be sure to include other output in this homework.

3. Now run the command

   ```
   set seed 1
   simulate est = _b[x] se = _se[x], reps(200): genest
   summarize
   ```

   to do a Monte Carlo simulation. This command runs the `genest` command 200 times, and stores the results: the `est` variable stores the estimates of the slope coefficient and the `se` variable stores the estimates of the standard error associated with the estimate of the slope coefficient. The command `set seed 1` sets the "seed" for the pseudo-random number generator, so that the results are replicable.

   (a) What is the average of the estimates? What is the standard deviation of the estimates?

   (b) What is the average of the standard errors?

   (c) How does the standard deviation of the estimates compare to the average of the standard errors? (the answer is fairly straightforward, this is not a "trick question")

   Include the output in your answers, in addition to the answers to the questions.

Now, we will explore using the bootstrap to approximate the sampling distribution.

4. Now, run the following code that is similar to the code from question 1, *except that it computes the standard errors and confidence intervals using the bootstrap!*, due to the `vce(bootstrap)` option. Basically, this option instructs Stata to use the bootstrap rather than the standard asymptotic approximation.

   ```
   program genestboot
   drop _all
   set obs 50
   gen eps = rnormal()
   gen x = 3 + 2*rnormal()
   gen y = -2 + 1.5*x + eps
   regress y x, vce(bootstrap)
   end
   ```

   There is nothing to submit for this question.

5. Run the command `genestboot` to see what the program does.

   Include the output in your answers.

6. And now run the command

   ```
   set seed 1
   simulate estboot = _b[x] seboot = _se[x], reps(200): genestboot
   summarize
   ```

   to do a Monte Carlo simulation. This command runs the `genestboot` command 200 times, and stores the results: the `estboot` variable stores the estimates of the slope coefficient and the `seboot` variable stores the estimates of the standard error associated with the estimate of the slope coefficient (that comes from the bootstrap!).

   (a) What is the average of the estimates? How does it compare to the answer from question 3? Why does it have that relationship?

   (b) What is the standard deviation of the estimates? What is the average of the (bootstrap) standard errors?

   (c) So, how does the average of the (bootstrap) standard errors compare to the standard deviation of the estimates for the true sampling distribution?

   Include the output in your answers, in addition to the answers to the questions.

Questions about the bootstrap with estimating an instrumental variables regression:

Now we will explore how the bootstrap provides an approximation to the sampling distribution of the estimates from an instrumental variables regression.

First, we will numerically simulate the true sampling distribution.

7. Now write code for a program called `genestiv`, similar to that in question 1, to add a Stata program that generates a sample of $N = 50$ observations of the model:

   $$y = -2 + 1.5x + \epsilon_2$$

   where

   $$x = 3 + z + \epsilon_1,$$

   where $z \sim N(0, 3)$ and

   $$(\epsilon_1, \epsilon_2) \sim N\left(0, \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}\right)$$

   And then, `genestiv` estimates the model using instrumental variables, using z as the instrument for x. Hint: you might find the code from question 1 useful to base your answer on, and you might find the Stata commands `drawnorm` and `ivreg y (x = z)` (or a similar command `ivregress 2sls y (x = z)`) useful. You can use the Stata commands `help drawnorm` and `help ivreg` and `help ivregress` to get information about these commands. There are usually examples of using commands at the bottom of the help files in Stata, which are particularly useful. `ivreg` is technically an out-of-date command in more recent editions of Stata, but you can still use it although the help file might not exist anymore.

   Submit your code in your answers!

8. Run the new command `genestiv` to see what the program does.

   Include the output in your answers.

9. Now run the command

   ```
   set seed 1
   simulate estiv = _b[x] seiv = _se[x], reps(200): genestiv
   summarize
   ```

   to do a Monte Carlo simulation.

   (a) What is the average of the estimates? What is the standard deviation of the estimates?

   (b) What is the average of the standard errors?

   (c) How does the standard deviation of the estimates compare to the average of the standard errors?

   Include the output in your answers, in addition to the answers to the questions.

Now, we will explore using the bootstrap to approximate the sampling distribution.

10. Now, write code for a program called `genestivboot` that is similar to the code from question 7, *except that it computes the standard errors and confidence intervals using the bootstrap!*, due to the `vce(bootstrap)` option.

    Submit your code in your answers!

11. Run the command `genestivboot` to see what the program does.

    Include the output in your answers.

12. And now run the command

    ```
    set seed 1
    simulate estivboot = _b[x] seivboot = _se[x], reps(200): genestivboot
    summarize
    ```

    to do a Monte Carlo simulation.

    (a) What is the average of the estimates? How does it compare to the answer from question 9? Why does it have that relationship?

    (b) What is the standard deviation of the estimates? What is the average of the (bootstrap) standard errors?

    (c) So, how does the average of the (bootstrap) standard errors compare to the standard deviation of the estimates for the true sampling distribution?

    Include the output in your answers, in addition to the answers to the questions.