Social Data Science: Machine Learning & Econometrics

Exercise class 5

March 20, 2020

Todays quick warmup

Q: Write a function that adds an arbitrary n numbers together. Your function should obey the signature

```
add(1,1)
>>> 2
add(1,1,1,1)
>>> 4
add(1,1,1,1,1, 1,1,1,1,1, 1,1,1,1,1,
    1,1,1,1,1, 1,1,1,1,1, 1,1,1,1,1)
>>> 30
```

Bonus: write a function that greets people with custom greetings, i.e.

```
greet(hello='kristian')
>>> 'hello kristian!'
greet(hello='kristian', hi='peter')
>>> 'hello kristian and hi peter!'
```

Last lecture in a nutshell

How can we move forward when there are *a lot* of covariate/instrument candidates?

- We need methods to select a subset among many candidates.
- This method must not induce biases in our estimates. (Which easily happens because of *omitted variable bias*).
- Having many candidates is more common than you might think:
  - Maybe nonlinearities matter; better include polynomials $x_1^2, x_2^2, x_3^2, x_1x_2, x_1x_3, x_2x_3$. n grows rapidly with polynomial degree!
  - What about rates of change? Might they also matter? $\Delta x_1, \Delta x_2, \Delta x_3$, Acceleration? $\Delta^2 x_1, \Delta^2 x_2, \Delta^2 x_3$.
  - Data in general are becomming more abundant $\rightarrow$ more likely to have many potential controls/instruments.

Last lecture in a nutshell

- Idea: use LASSO to select variables.
- Problem: LASSO biases variables downwards.
- Problem: LASSO selects variables for prediction, these might not be the "right" variables for inference.

Solution: depends a bit on the problem (Regression/IV etc.), but intuition is o.k.

- Use Lasso to select relevant variables. Use OLS to reestimate parameter values (Post-Lasso).
- Do this in a way that makes the problem insensitive to small mistakes in the lasso step (i.e. failing to include relevant variables). This is problem specific.
- In IV: first stage is a prediction problem, leaving out very weak instruments is o.k.