The Assignment 1

This is an individual assignment.

The dataset

Consider the cps-earnings dataset at https://osf.io/g8p9j/ (Cross section. N=149 316 individuals)

- Pick an occupation and filter data accordingly. You must all pick different occupations / occupation individually.
- Occupation codes are here: https://osf.io/57n9q/
- You may merge occupations as you see fit (ie all tax/insurance specialists, etc).

You can see some ideas working with this code here.

- https://github.com/gabors-data-analysis/da_case_studies/tree/master/ch09-gender-age-earnings
- https://github.com/gabors-data-analysis/da_case_studies/tree/master/ch10-gender-earnings-understand

Tasks

Build four predictive models using linear regression for earnings per hour.

1. Models: the target variable is earnings per hour, all others would be predictors.
2. Model 1 shall be the simplest, model 4 the more complex. It shall be OLS. You shall explain your choice of predictors.
3. Compare model performance of these models (a) RMSE in the full sample, (2) cross-validated RMSE and (c) BIC in the full sample.
4. Discuss the relationship between model complexity and performance. You may use visual aids.
5. You should submit your code in Github and a 1 page report in pdf on Moodle along with the final Jupyter notebook file
    1. Pay attention for markdown formatting in your Jupyter notebook.
    2. We encourage to use Quarto to write your 1 page report in pdf.

Work individually. But you may collaborate in your support group, check and comment (add issues) on each other code.

Hints re Git and commit

- Committing is a habit, and people may have different ways.
- Some people commit very frequently, others less so.
- We basically expect you to have a few commits, one per major parts of the exercise. The first commit will set up the folder/file for A1.
- Then you can commit, say data work, descriptive stats, graphics, and regressions. And then, commit your edits.
- Make sure the commit text is short but meaningful: Good: "adding graphs", "calculate RMSE", "edit typos". Bad: "update"

Grading

This assignment is worth 30 points.

- 5 points will be for Git use.
- 15 points will be technical aspects the analysis
- 10 points will be based on your report