Junior Quantitative Analyst in Immigration Research

Technical Interview / Assignment

Dear Applicant,
This is the problem set for the technical interview. We would ask you to solve the problems using R. You are allowed to use any additional packages unless stated otherwise. Your submission should be sent to us in form of a ZIP archive named `firstname_lastname.zip`. The ZIP archive should include:

- A PDF file generated using LaTeX with a write-up of your answers to all questions. Please also include the `tex` file.
- R code files
- All generated figures & tables
- Original and generated data sets

Make sure that we only need to unzip the archive and change the working directory for the code to run on our machines. Please label all files clearly with your name.

Question 1: Job episodes

The data file `dem_dat.csv` contains hypothetical data on the arrival times of immigrants into a particular country. For each immigrant (each row in `dem_dat`), you will need to compute the total number of months that they had a job over the course of their first 3 years in the country. To do this, you will also make use of `job_dat.csv`.

Info on the data:

- `dem_dat.csv` contains observations for unique immigrants. `YM_in` is their arrival year-month in the country, and `YM_out` is the time in which they leave the country.
- `job_dat.csv` contains job 'episodes', with start and end year-months indicated. If an individual does not have any entries in `job_dat.csv`, then they have no job episodes.

What are the mean, median, and maximum number of months worked by the immigrants in their first 3 years in the country?

Question 2: Fertility and Labor Supply

This problem set examines the impact of fertility on various labor market variables: female labor supply, earnings etc. using a dataset from the US in 1980.

The analysis is based on Angrist and Evans (henceforth AE), 1998, "Children and Their Parents' Labor Supply: Evidence from Exogneous Variation in Family Size", *American Economic Review*, 1998, 88(3):450-477.

The dataset in `aedata.csv` consists of 254,654 households in the US in the 1980 with at least two children. You find a data description in the Appendix on page 3.

1. Using OLS, regress various labor market outcomes (female labor participation, female hours of work, total family income) on number of children (our measure of fertility) while controlling for age of the mother, age of mother at first birth, education of the mother and racial variables (i.e., `blackm`, `hispm`, `othracem`). Explain why these estimates might not reflect the causal effect of fertility on labor outcomes.

2. The treatment we now consider is the birth of a third child and its effect on these labor outcomes. The instrument we are going to use is whether the first two children have the same sex. Explain the reasoning for the choice of the instrument. Explain carefully what the treatment parameter is. (No estimation required.)

3. Assess the relevance of the instrument empirically. Explain what it means for an instrument to be relevant and what happens if this is not the case.

4. Estimate by 2SLS the effect of the treatment using the controls from 1. and compare to the OLS estimates. Briefly justify your choice of standard errors.

5. Obtain the same 2SLS point estimate as in the previous question, but implement the 2SLS estimator manually. That is, do not use *any* external R packages to answer this question.

Question 3: Machine learning

We use a tweaked (i.e., synthetically generated) version of the 1990 SIPP survey that has been widely used to study the effect of 401(k) eligibility and participation in the US. In short, 401(k) allows to deduct pension contributions from taxable income. 401(k) plans are provided by employers, only workers in firms offering plans are eligible for participation.

The main predictors in the data set are: age (`age`), income (`inc`), education (`educ`), family size (`fsize`), married (`marr`), indicator if partner has positive income (`twoearn`), DB pension plan (`db`), IRA participation (`pira`), home owner (`hown`), where DB and IRA are alternative pension plans.

Your task is to predict net total financial assets (`net_tfa`). Proceed as follows:

1. Load the data from `PVW_synth.csv`. Set the randomization seed at the beginning of your script to ensure that we can replicate your results. Randomly split the data in (approximately) two halves. The first half is the training sample, the other half is the validation sample.
2. For this prediction task, we consider lasso regression and random forests. Briefly explain these methods (no more than 150 words in total).
3. Fit these learners to the training sample using net financial assets as the outcome and using the predictors listed above. Briefly motivate your choice of (hyper)tuning parameters.
4. Assess the relative prediction performance of lasso and random forests using the validation sample. (Short statement.)

Question 4: White flight

In a recent working paper, Bayer et al. study if households are more likely to move away if a neighbor of different race moves into the same neighborhood. The study thus attempts to shed lights on the drivers of segregation. Explain why it is challenging to identify the causal effect of a new arrival in a neighborhood on the probability of other-race households to move away, and explain how Bayer et al. tackle this task. (Please do not use more than 250 words. Note that this question requires no estimation.)

Bayer, P., Casey, M. D., McCartney, W. B., Orellana-Li, J., & Zhang, C. S. (2022). Distinguishing Causes of Neighborhood Racial Change: A Nearest Neighbor Design (No. w30487). *National Bureau of Economic Research.*

Appendix

Question 2: Data description

Variable name	Description
kidcount	number of kids
morekids	had more than 2 kids
boys2	first two births boys
girls2	first two births girls
boy1st	first birth boy
boy2nd	second birth boy
samesex	first two kids are of same sex
multi2nd	=1 if 2nd birth is twin
agem1	age of mom
agefstm	age of mom at first birth
whitem	=1 if mom is white
blackm	=1 of black
hispm	=1 if hispanic
othracem	=1 if other race (white is ref)
workedm	mom worked last year
weeksm1	weeks worked mom
hourswm	hours of work per week, mom
incomem	moms labor income
faminc1	family income
nonmomi	income not generated by mom
educm	moms years of education