Prediction Competition

Description: The goal of this competition is to predict annual wage and salary incomes for individuals. Participation is ***entirely voluntary***.

Data: a sample extract from ACS 2019 at IPUMS

Training Sample: use the dataset called "**train_sample.dta**" to develop a prediction model for **incwage.**

train_sample.dta

```
  obs:       318,074
 vars:            13
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
sex             byte    %8.0g      SEX        sex
age             byte    %8.0g      AGE        age
marst           byte    %8.0g      MARST      marital status
race            byte    %8.0g      RACE       race [general version]
bpl             int     %8.0g      BPL        birthplace [general version]
language        byte    %8.0g      LANGUAGE   language spoken [general version]
educ            byte    %8.0g      EDUC       educational attainment [general version]
degfield        byte    %8.0g      DEGFIELD   field of degree [general version]
occ             int     %8.0g                 occupation
wkswork1        byte    %8.0g                 weeks worked last year
uhrswork        byte    %8.0g      UHRSWORK   usual hours worked per week
incwage         long    %12.0g                wage and salary income
vetstat         byte    %8.0g      VETSTAT    veteran status [general version]
-------------------------------------------------------------------------------
```

After you develop a prediction model, predict **incwage** (in integer values) for each **case_id** in the (out-of-sample) test sample called "**test_sample.dta**". In this test sample, all other variables are given except **incwage.**

Rule:

Size of a competition team: minimum 1; maximum 4

Submission deadline: January 17, 2022 before 11:59pm US EST

Submission materials:

- a file containing predictions of **incwage** (in integer values) for each **case_id**
- a written report of the methods and empirical results
- a zipped file of all replication files

Competition Criteria:

- Prediction performance measured by mean square prediction errors, which will be computed after taking the log of the annual wages.
- A quality of the report
- Completeness of replication files

Prizes: A prize will be given to the top performing competition team. More prizes will be awarded if there are multiple award worthy teams.

Tips:

- Use the log wages as the dependent variable for estimation but predict the wages in integer value.
- Be careful how to construct covariates and be explicit in your report.
- Use high-dimensional regression models learned in class and try other models if you can.
- It might improve prediction performance if you take an average of predictions from individual models.