**Assignment 1**

Dilnovoz Abdurazzakova

For the purpose of this assignment, I choose Economists and 4 types of finance-specialized occupations (Accountants and auditors, Appraisers and assessors of real estate, Budget analysts, Credit analysts and Financial analysts) to predict the hourly wages of the respondents based on my own interest. Analyses are based on USA workers, cross sectional dataset taken from <https://osf.io/g8p9j/>.

Firstly, the respondents of this specific occupation groups are selected, and relevant predictor variables are selected for the analyses. As an individual/home related factors, I used the age of the respondent, gender, marital status, race and the number of children they have. Furthermore, I added education level and what type of sector they are working for which are directly related with their labor level outcomes. I have re-categorized relevant unbalanced variables including marital status, race and education, so they are more balance among respondents. To illustrate, race of the respondents are categorized into white black and other categories. And it showed that hourly wages of respondents whose races are either white or other are on average relatively higher than black respondents’. My regression estimations further show that the age of the respondents are nonlinearly associated with the earnings they get, however, marital status and education level of respondents are not differently associated with wages between females and males. I have tried both log and dollar amount of hourly earnings. According to adjusted R squared statistics, I continued my analyses using dollar amount of wages. Additionally, in my sample the wages of the respondents are already relatively normally distributed as maybe due to these occupations are quite similar to each other. I first estimated individual/household related factors separately, and education and working for private sector separately. And in my last estimation, I added both group of variables in one model. As we can see from the BIC, Full sample RMSE and cross-validated RMSE, the last model including all variables are most preferable consistently. However, I should note that increasing the number of variables are not the main factor in the prediction analyses as we can see in the last graph of RMSE, even though the number of individual/household variables are more than education and working for private sector model, the prediction get relatively worse off at first. And once nonlinearity of the age variable are introduced and education, working for private sectors are introduced, RMSE decreased noticeably!