Problem Set 1

Report

I used data from Current Population Survey of US to predict wages. I chose occupations related to finance and financial sector. So namely I chose “Securities, commodities, and financial services sales agents”, “Financial clerks, all other”, “Financial managers”, “Accountants and auditors”, “Appraisers and assessors of real estate”, “Budget analysts”, “Credit analysts”, “Financial analysts”, “Personal financial advisors”, “Insurance underwriters”, “Financial examiners”, “Loan counselors and officers”, “Tax examiners, collectors, and revenue agents”, “Tax prepares”, “ Financial specialists, all other”. I went from 149k observations to well above 5k observations/entries/rows. I found that “ethnic” feature has 4k missing values, so I dropped it. The target is hourly wages imputed by me.

Overall, I had bad choice of feature so that I ended up with multicollinearity in the models. In forth model I added dummy variables for