Assignment 1

Ádám Vig (1903211)

October 4, 2021

In this assignment, I build a model to predict wages per hour in the Current Population Survey (CPS) done by the US government. This is the monthly household survey of employment and labor market, I had a cross section of one survey, and chose three related occupations: computer programmers (1010), software developers, applications and systems software (1020) and web developers (1030).

I started by looking at the documentation of variables in the dataset. My goal was to create the target variable and a set of variables I use for prediction. First, I removed all id variables and a variable used for representativeness. The target variable is *wage per hours*. I have two continuous variables: (1) *age* which I use as it is and additional I created a quadratic term of age and (2) education which was coded into categories. I created a *years in education* variable based on the codebook (I use this to create work experience later) and also a categorical *education* variable with 4 categories: no degree, BA degree, MA degree and Phd degree. My objective in aggregating was to create few but relevant categories as my sample size is only 2067, with loosing as few variation as possible.

Race is also divided into a lot of categories, I aggregated into white, black, asian and other. I divided *marital status* into three categories: married, used to be married (here goes widows, separated, divorced) and never married. I aggregated *number of children* to one category if it's greater or equal to three and left the other categories (no children, 1 children etc) as they were. I created a variable indicating a person is in a *union* and another indicating the person *born in the US* or not. I use *gender*, *employment status*, *working class* (govt, private sector) *occupation* as they are given. The *industry* variable is classified in to 4 level industry codes and even in this sample there are more than 160 different 4 level codes therefore I created a 2 level code indicating industry from the first two letters of the 4 level code and aggregated this industries that had less than 200 observations into one "other" category. I created work *experience* with *age – years in education – 6* but I did not used it in my estimation as it is closely similar to age (there is only a small variation in years in education). The remaining dataset has 2067 observations with 13 features and a target variable.

I used the following models estimated by OLS:

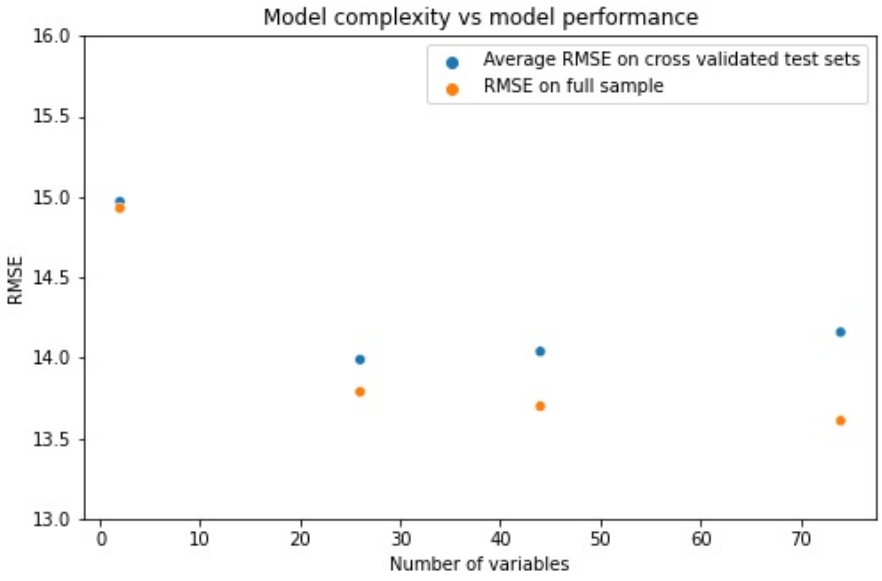
- Model 1 - *age, age*^{2} included
- Model 2 – all variables included

- Model 3 - all variables and potential interactions included
- Model 4 - all variables and all interactions of categorical and continuous variables included

My results are in the following table:

	Model1	Model2	Model3	Model4
BIC on full sample	17064.73	16918.31	17030.36	17230.74
RMSE on full sample	14.93	13.79	13.70	13.61
Fold1 rmse on test	14.93	13.77	13.72	13.86
Fold2 rmse on test	14.74	13.98	14.13	14.21
Fold3 rmse on test	14.99	14.34	14.52	14.55
Fold4 rmse on test	15.24	13.86	13.81	14.03
Average RMSE on test	14.97	13.99	14.04	14.16

Both BIC and the cross validated RMSE in the test set indicates that my best model is Model 2. Model 1 is clearly not enough, but it is only using age and age squared. Model 3 and Model 4 perform good in the whole sample, but both BIC and cross validation reveals that they are overfitted, and more likely to perform worse on live data, then Model 2. This relationship between model complexity and performance is demonstrated in the following figure. Finally, I note that there is no great variation on model performance, my best model is not much better than my worst and even the best model is not really a good one[1].



[1]See Python notebook for prediction intervals.