DA3_HW1_Gyebnar_Daniel

Gyebnar Daniel

2022 01 23

Introduction

This project aims to evaluate different models that predict the hourly wages of Computer and Information System managers. I will use the RSME and the BIC metrics on the full sample and on a 4-fold cross validation sample. I have used the CPS Annual Earnings from 2014 dataset, for full time employees. This resulted in 691 observations.

Exploratory analysis

I have examined each variable's distribution and conditional mean of each explanatory variable. Based on these results, I have selected my final variables for the models, and transformed them:

- I transformed the categorical variables into factors, e.g. grade92, marital status, owning a child, ect.
- I have aggregated the values of multiple categorical variables, where many categories had only a few observations but very similar meaning, conditional mean and SD. E.g. I simplified marital status variable from 8 different values to only 3 (Never married, Married, Used to be married), and educational level to below BsC, BsC, MsC, above MsC.

Variable and model selection

I have utilized all potential variables that had an underlying meaning with conditional means on y. I have selected the following four models:

- Model 1: w ~ grade92 + age
- Model 2: w ~ grade92 + age + sex + prcitshp + white + ownchild
- Model 3: w ~ grade92 + age + sex + prcitshp + white + ownchild + class + (age)x(grade92) + (age)x(sex)
- Model 4: w ~ grade92 + age + sex + prcitshp + white + ownchild + class + marital + (age)x(grade92) + (age)x(sex) + (age)x(prcitshp) + (sex)x(ownchild) + (white)x(grade92) + age2 + (age2)x(grade92) + (age2)x(sex) + (age2)x(ownchild) + (age2)*(prcitshp)

Model 1 is the simplest, capturing experience and education level. Model 2 captures additional social parameters such as sex, citizenship (referring to place of birth), race and having a child Model 3 includes interactions for the most important variables, age, educational level and sex. Model 4 includes all potential interactions and age as a quadratic term

Models	RMSE_full_sample	BIC_full_sample	RMSE_k_fold_avg
Model1	13.72668	5613.593	13.75631
Model2	13.30785	5616.536	13.41762
Model3	13.19799	5657.385	13.42640
Model4	12.88207	5741.588	13.57844

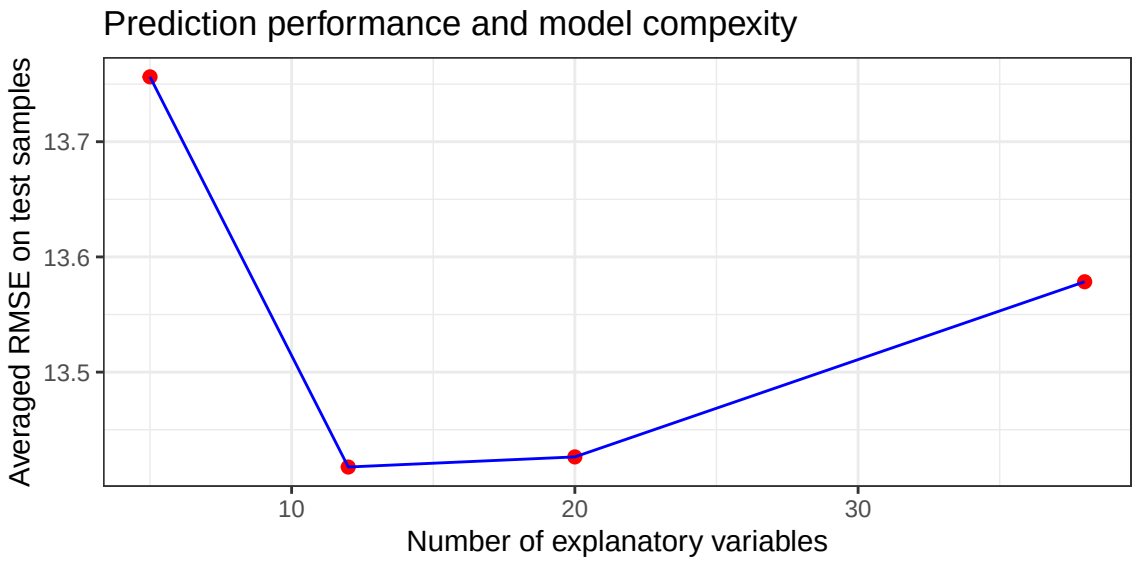
Results

Based on evaluating the full sample, Model 2 is the best choice: On the full sample, the RMSE decreases from 13.73 (Model 1) to 12.88 (MOdel 4) as complexity increases, while the BIC is the smallest for Model 2, meaning that Model 3 and Model 4 would likely overfit the live set.

Evaluating with 4 fold cross validation, the RMSE is also the lowest for Model 2.

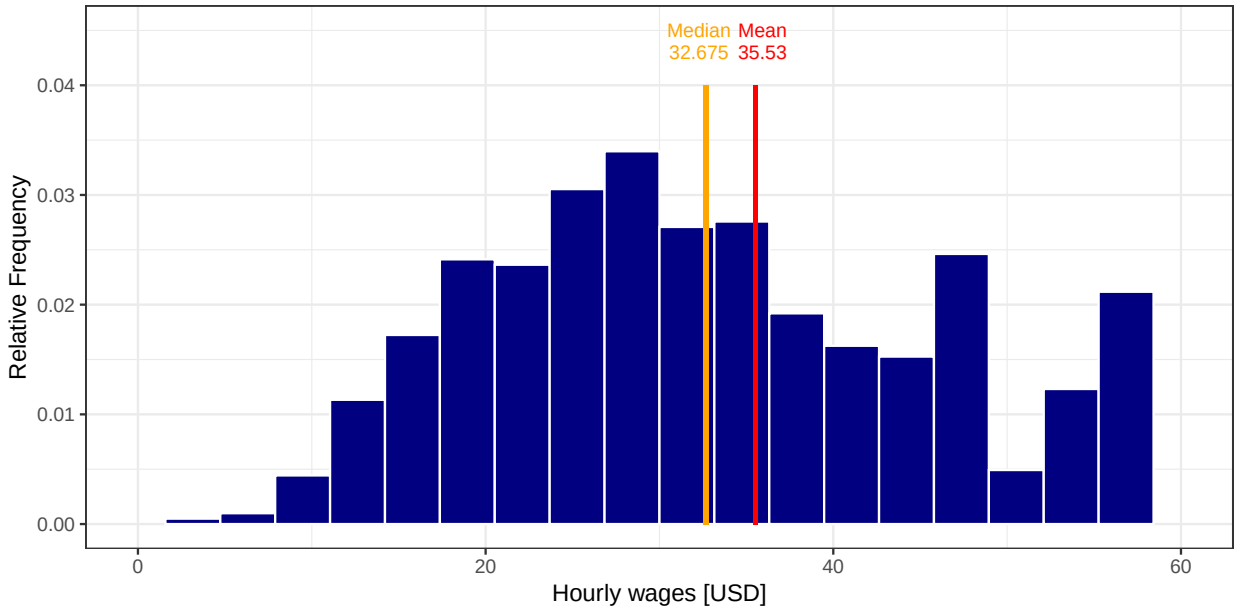
Visaulizing number of variables and RMSE from the 4 fold cross validation, Model 2 achieves the lower RMSE with the lowest complexity.

Based on the shape of the graph, it could be possible that with 4-5 more variables, RMSE could be further decreased.

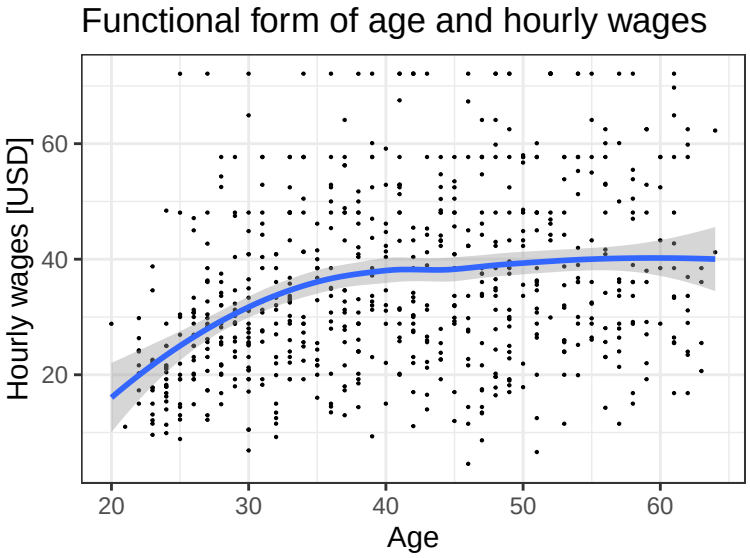


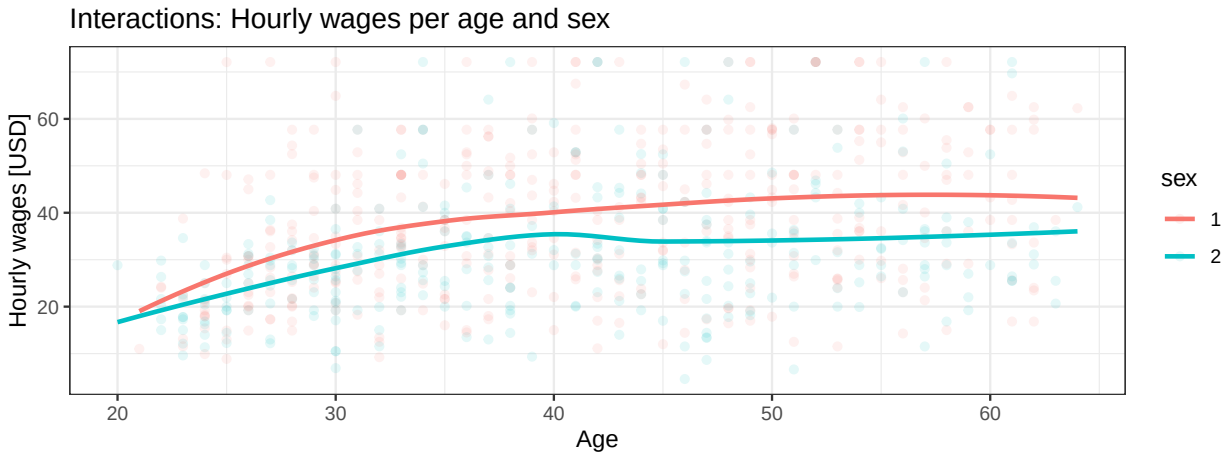
Appendix

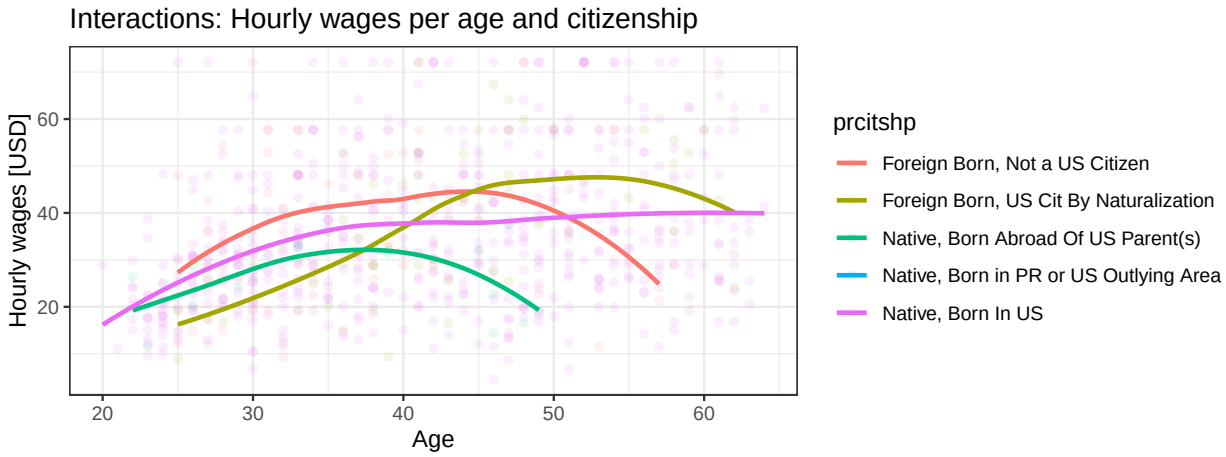
Distribution of hourly wage

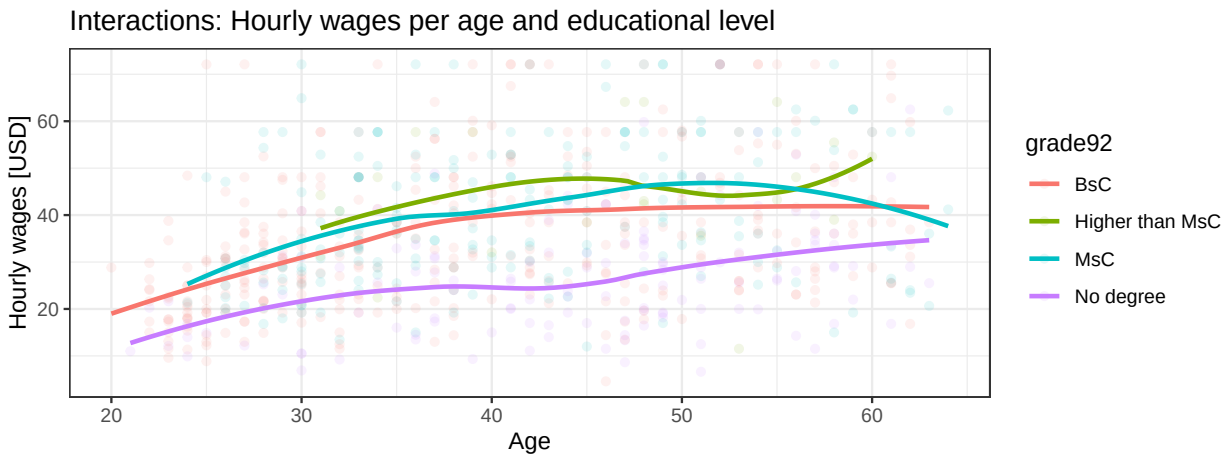


Functional form for continous variable




Interactions: Hourly wages per age and sex
Hourly wages [USD]
60
40
20
20
30
40
50
60
Age
sex
1
2


Interactions: Hourly wages per age and citizenship
Hourly wages [USD]
60
40
20
20
30
40
50
60
Age
prcitshp
Foreign Born, Not a US Citizen
Foreign Born, US Cit By Naturalization
Native, Born Abroad Of US Parent(s)
Native, Born in PR or US Outlying Area
Native, Born In US


Interactions: Hourly wages per age and educational level
Hourly wages [USD]
60
40
20
20
30
40
50
60
Age
grade92
BsC
Higher than MsC
MsC
No degree