Assignment 1

Ramzi Chariag

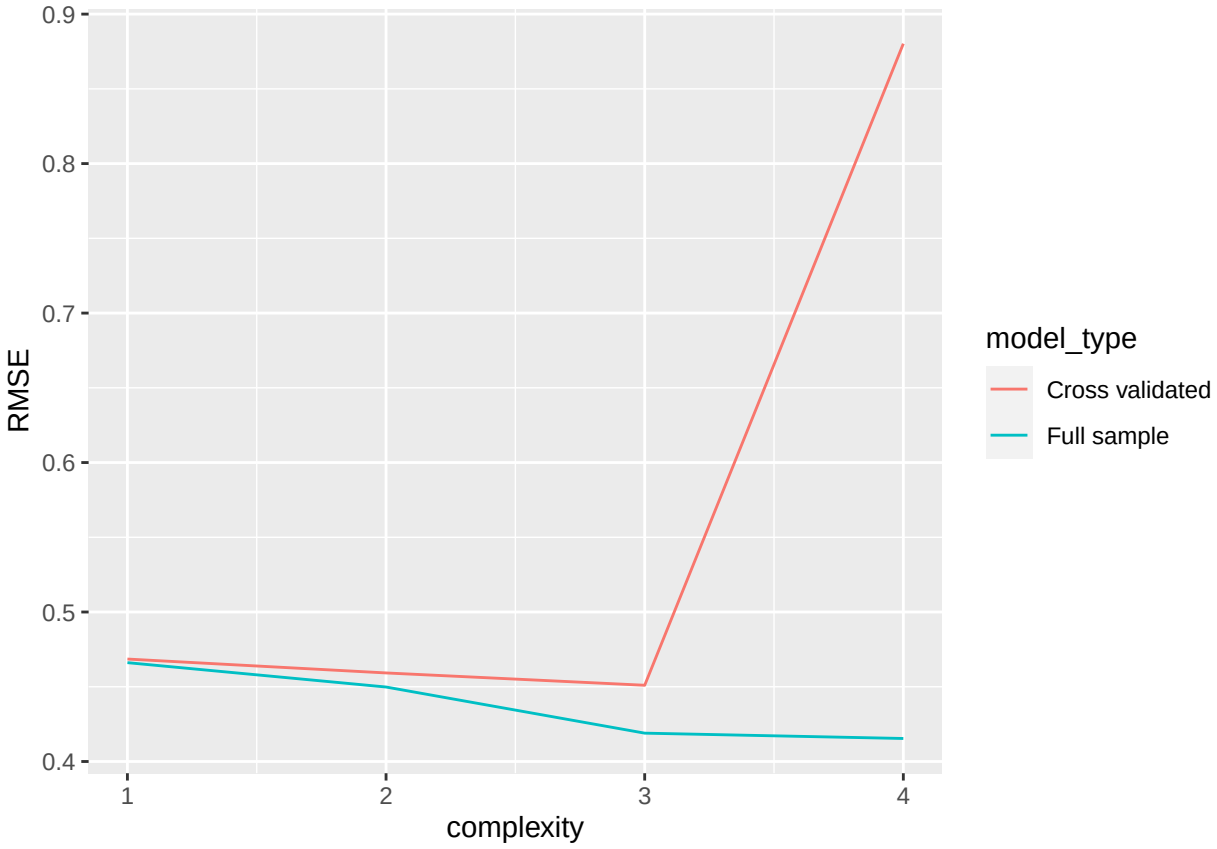
2023-01-21

In this assignment, I build four models to predict log wages using a large number of socio-demographic and economic variables from the CPS database. First, I create wage per hour by dividing earnings per week by the number of hours worked per week. I use the variable grade92 as a proxy for education. I also square education, since education might have a non-linear effect on earnings, at least theoretically it does. The same principle applies to the number of hours worked, for which I also generate a squared and cubic terms. I then turn all useful categorical variables into dummies making sure I leave at least one out. On some of them I left more than one out when the number of observations in the said category is negligibly low (e.g. below 100 out of the whole CPS sample).Race for example is divided into many categories, I aggregated them into white, black, asian and other (being the omitted category). I divided marital status into three categories: married, used to be married (this includes widows, separated and divorced) and never married. Information about the various variables is available in the description of the variables available on the CPS website. I restrict my analysis to legal jobs using their codes to extract them from the initial data. I also use the specific job as a categorical variable from which I generate dummies. I also get rid of variables that do not carry any useful information for this exercise since the very beginning: id, weight and timing of the survey.

Predictors in the first model are what anyone would associate naively with earnings. I used uhours (hours usually worked per week), age, education, married, black and female. In the second model, I added a couple polynomial terms and other more detailed variables. the variables added are non_USborn, lawyer, private_profit (which is a dummy for working at private forr profit firm), emp_absent (which is a dummy for being officially employed but absent/on leave from work), ownchild, chldpres (underage children present in the household), education squared, number of children squared, nionised, white, asian, clerk (another legal job dummy like lawyer), naturalized (dummy for being a naturalized citizen of the US), native_born_abroad, federal (dummy for working for the federal government, the omitted category here working for state government). In model 3, I add additional polynomial terms and state fixed effects. In the 4th model, I add 5^{th} degree polynomial terms to continuous variables. Model 4 is over-fitted. This results in a lower RMSE in the linear regression case, and a higher one in the cross-validated case, as demonstrated by the plot below. Initially complexity makes the model better at predicting, but at some point, it starts becoming less and less accurate. The different RMSEs and BIC are reported in the rable below:

Model	1	2	3	4
BIC	1697	1715	1937	1937
RMSE	0.466	0.450	0.419	0.415
CV.RMSE	0.469	0.459	0.451	0.880

BIC does not change between model 3 and 4. This is due to the fact that the increase in the number of variables between models 3 and 4 is not large enough to induce a change in BIC. However, RMSE for the cross validated models picks model 3. Theoretically, model 3 is the one that makes the most sense. Model 4 is forcibly over-fitted just to show how the cross-validated RMSE would increase if we add too many predictors to the linear model. I tried adding intercation terms with the state dummies, but the number of variables blew up and was above the number of observations in the restricted sample. In the whole sample that, such a model could be deployed to demonstrate that things could get even worse.




```
## Saving 6.5 x 4.5 in image
```