DA3 - Assignment 1

Oszkar Egervari

1/26/2022

Introduction

In this report I showcase four models, that predict the earnings per hour of Educational Administrators (occupational code 0230). The models have different levels of complexity, model 1 being the most simple, while model 4 contains the most predictor variables. The aim of this report is to find the model with the best performance based on RMSE in full sample, cross-validated RMSE and BIC in full sample.

Data

The data used for the report is the cps-earnings dataset. As mentioned above, I chose the Educational Administrators (occupational code 0230) occupation for the prediction.

The number of observations is 966 for all of our key variables.

[H]

Table 1: Evaluation of the models

	Model 1	Model 2	Model 3	Model 4
Dependent Var.:	w	w	w	w
(Intercept)	-7.651 (5.789)	-4.683 (5.773)	-116.5*** (9.936)	-115.6*** (9.964)
age	1.277*** (0.2884)	1.271*** (0.2821)	0.7511** (0.2743)	0.7627* (0.3009)
agesq	-0.0090** (0.0034)	-0.0090** (0.0033)	-0.0038 (0.0032)	-0.0041 (0.0035)
female		-4.001*** (0.9474)	-2.429** (0.9277)	-2.451** (0.9358)
grade92			2.839*** (0.2132)	2.818*** (0.2116)
unionmember				2.980* (1.224)
ownchild				-0.3632 (0.3872)
S.E. type	Heteroskedas.-rob.	Heteroskedas.-rob.	Heteroskeda.-rob.	Heteroskeda.-rob.
BIC	7,912.2	7,901.8	7,791.1	7,797.9
RMSE	14.379	14.251	13.409	13.361
Observations	966	966	966	966
No. Variables	2	3	4	6

As seen on table 1, Model 1 has only age and age square as predictor variables. I chose the square of age to model the potential non-linearity. For Model 2 I added the female binary variable, then the level of education, and finally for Model 4 I added the union member binary variable and the number of children.

Interpreting the coefficients, we can see that age is positively related to hourly wage in a concave fashion. The female binary variable is negative, meaning females are expected to earn less, meanwhile the union member binary variable and educational level variable are positive, so both union members and people with higher levels of education are expected to earn more.

Model 3 has the lowest BIC (7,791.1), closely followed by Model 4 (7,804.7), while Model 4 has a slightly lower RMSE (13.361) than Model 3 (13.409) as for RMSE in the full sample.

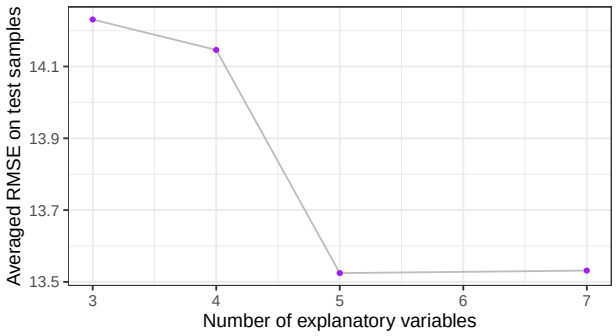
[H]

Table 2: Hourly wage models estimated and evaluated using 5-fold cross-validation and RMSE

Resample	Model1	Model2	Model3	Model4
Fold1	13.25895	13.11637	12.73828	12.64800
Fold2	14.02906	13.82337	12.89075	12.76970
Fold3	13.82035	13.61453	13.19488	13.27631
Fold4	15.69906	15.87379	15.13612	15.26790
Fold5	15.14692	14.93508	13.41458	13.32521
Average	14.23092	14.14618	13.52444	13.53150

On table 2 we can see the cross-validated RMSE for the four models. Here Model 3 (13.52444) has a slightly lower RMSE than Model 4 (13.54168). All these values are close, in which case it is advised to choose the more simple model, as it may help us avoid overfitting the live data. In this case that would mean choosing Model 3. The relationship between the increasing number of prediction variables and lower RMSE is displayed on Figure 1.

Figure 1: Prediction performance and model compexity



Conclusion

By increasing the number of the predictor variables, lower BIC and RMSE were achieved to a certain point, Model 3 and 4 produced similar scores. In which case it is advised to choose the less complex model, that way the chance of overfitting on the live data is lower.