DA3 Exercise1

Peter Kaiser

2022 01 23

Introduction

This is a report about building models to predict the hourly wage of the driving jobs with multiple predictor variables in the cps data set. In the end we horse race the models to see which is the best one. It checks the BIC and the RMSE on the whole data set and the RMSE with 4 fold cross validation.

Data cleaning, Feature engineering

- Target Variable (Y): Hourly wage
- Predictor variables (X): Education (High school graduate as base), Race (White as base), Age, Sex (Male as base), Marital Status (Married as base), Union Status (Not in union as base)
- For more information see in Appendix

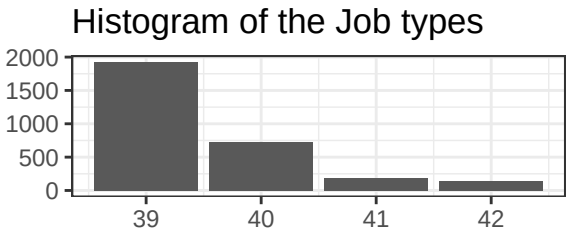
Models

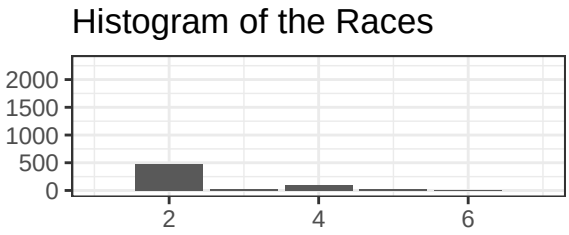
- Model 1: age, age squared
- Model 2: age, age squared, sex
- Model 3: age, age squared, sex, union, marital status
- Model 4: age, age squared, sex, union, marital status, race, education
- Model 5: age, age squared, sex, union, marital status, race, education, interaction age and female, interaction with age and union

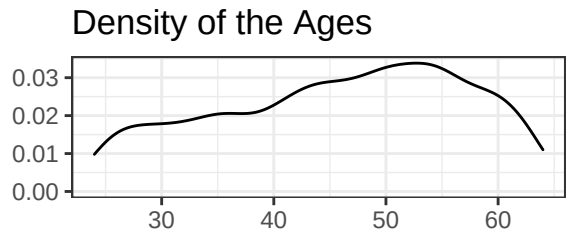
Appendix

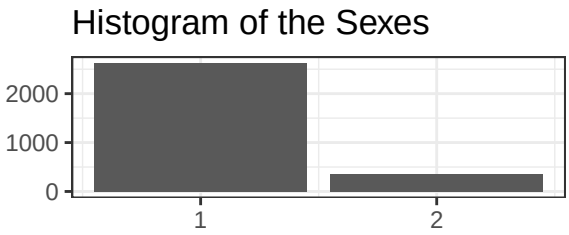
Distribution of the hourly wages

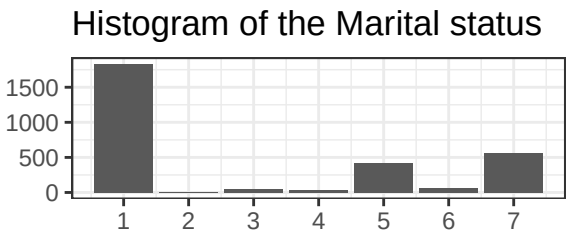
The distribution is normal with right long tail

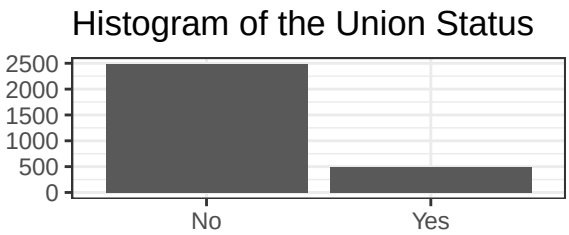

Histogram of the Job types
2000
1500
1000
500
0
39
40
41
42

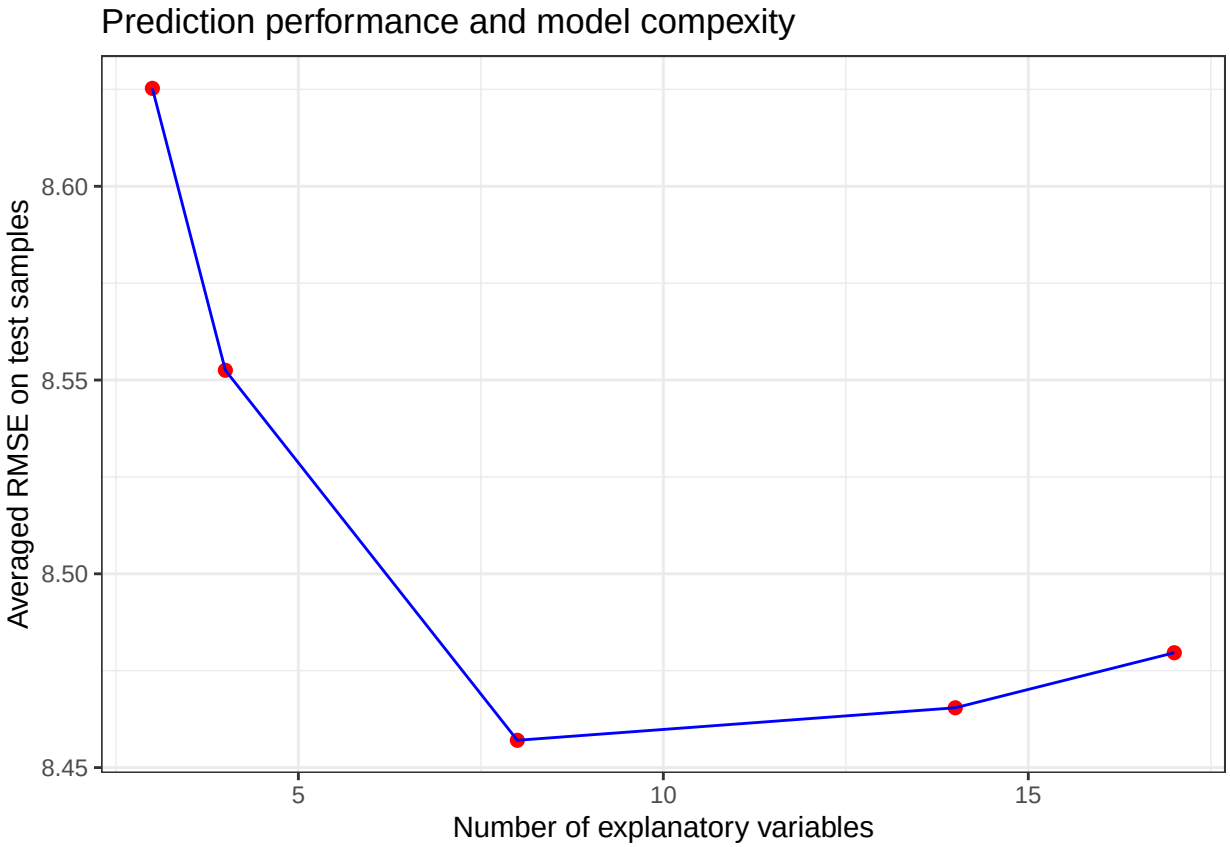

Histogram of the Races
2000
1500
1000
500
0
2
4
6


Density of the Ages
0.03
0.02
0.01
0.00
30
40
50
60


Histogram of the Sexes
2000
1000
0
1
2


Histogram of the Marital status
1500
1000
500
0
1
2
3
4
5
6
7


Histogram of the Union Status
2500
2000
1500
1000
500
0
No
Yes



Resample	RMSE	RMSE.1	RMSE.2	RMSE.3	RMSE.4
Fold1	8.590490	8.474870	8.338192	8.293919	8.297041
Fold2	8.926210	8.814151	8.682432	8.744360	8.745257
Fold3	8.309046	8.223868	8.162299	8.162094	8.186240
Fold4	8.664158	8.685491	8.634376	8.647591	8.676497
Average	8.625277	8.552521	8.457033	8.465423	8.479631

With 5-fold cross validation Model 3 has the best RMSE

Evaluation of the models using all the sample

	reg1	reg2	reg3	reg4	reg5
Dependent Var.:	earnho	earnho	earnho	earnho	earnho
Intercept	1.915	1.676	3.992	3.407	-2.449
	(2.392)	(2.371)	(2.504)	(2.563)	(9.631)
age	0.6876***	0.7182***	0.6146***	0.6460***	1.095
	(0.1157)	(0.1149)	(0.1172)	(0.1187)	(0.7073)
age squared	-0.0069***	-0.0072***	-0.0063***	-0.0067***	-0.0173
	(0.0013)	(0.0013)	(0.0013)	(0.0013)	(0.0167)
female		-3.639***	-4.000***	-4.073***	-4.134*
		(0.4378)	(0.4510)	(0.4575)	(1.917)
unionized			3.532***	3.543***	0.4839
			(0.4268)	(0.4212)	(1.890)
divorced			0.4361	0.3766	0.3872
			(0.4939)	(0.4937)	(0.4928)
never married			-0.9001*	-0.7635	-0.7563
			(0.4349)	(0.4357)	(0.4352)
other marital status			-0.1950	-0.1103	-0.1091
			(0.5845)	(0.5855)	(0.5852)
black				-0.8400*	-0.8445*
				(0.3966)	(0.3972)
asian				-1.219	-1.175
				(1.022)	(1.023)
other race				-1.991*	-2.073*
				(0.8207)	(0.8215)
college drop-out				0.5206	0.5151
				(0.3849)	(0.3854)
occupational degree				-1.838***	-1.849***
				(0.5385)	(0.5381)
academic degree				2.000*	2.019*
				(0.8355)	(0.8368)
age cubed					7.87e-5
.					(0.0001)
age x female					0.0012
.					(0.0428)
age x unionized					0.0645
.					(0.0396)
S.E. type	Hete.-rob.	Hete.-rob.	Hete.-rob.	Hete.-rob.	Het.-rob.
AIC	21,297.8	21,243.4	21,173.7	21,157.6	21,160.7
BIC	21,315.8	21,267.4	21,221.7	21,241.6	21,262.7
RMSE	8.6150	8.5339	8.4234	8.3837	8.3797
R2	0.01795	0.03636	0.06115	0.06997	0.07085
Observations	2,980	2,980	2,980	2,980	2,980
No. Variables	2	3	7	13	16

According to BIC, Model 3 is the best. According to RMSE Model 5 is the best.