D_An3_hw1

Introduction

This analysis aims to predict the possible hourly wage of lawyers and what kind of factors affect by how much this amount, like age, education level, sex, race, and many others. For the sake of the analysis, the cps_data dataset was used that includes information on many thousands of people.

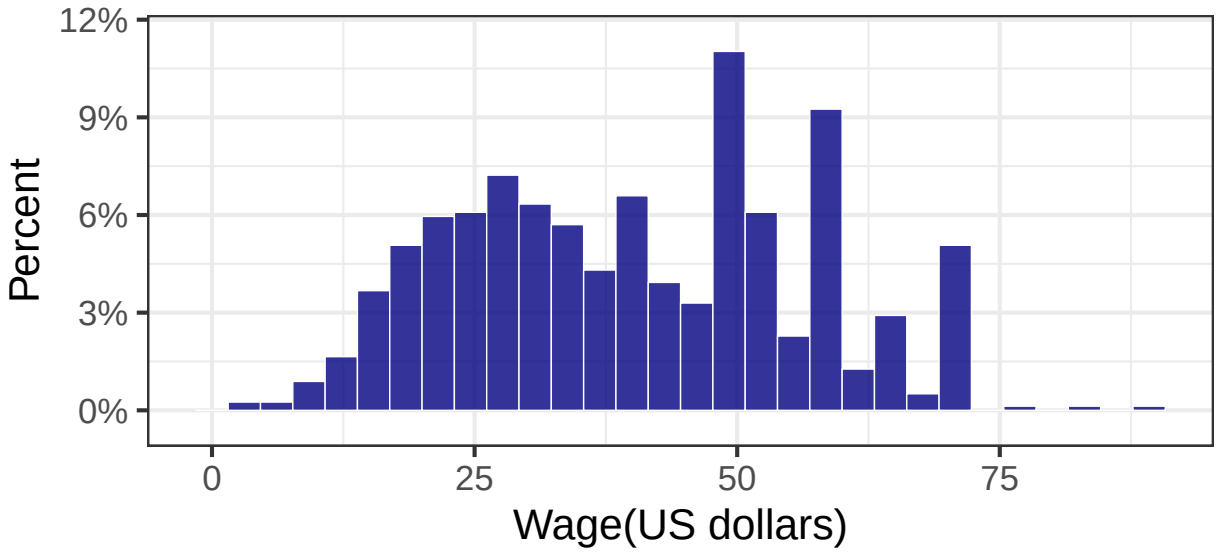
Data cleaning

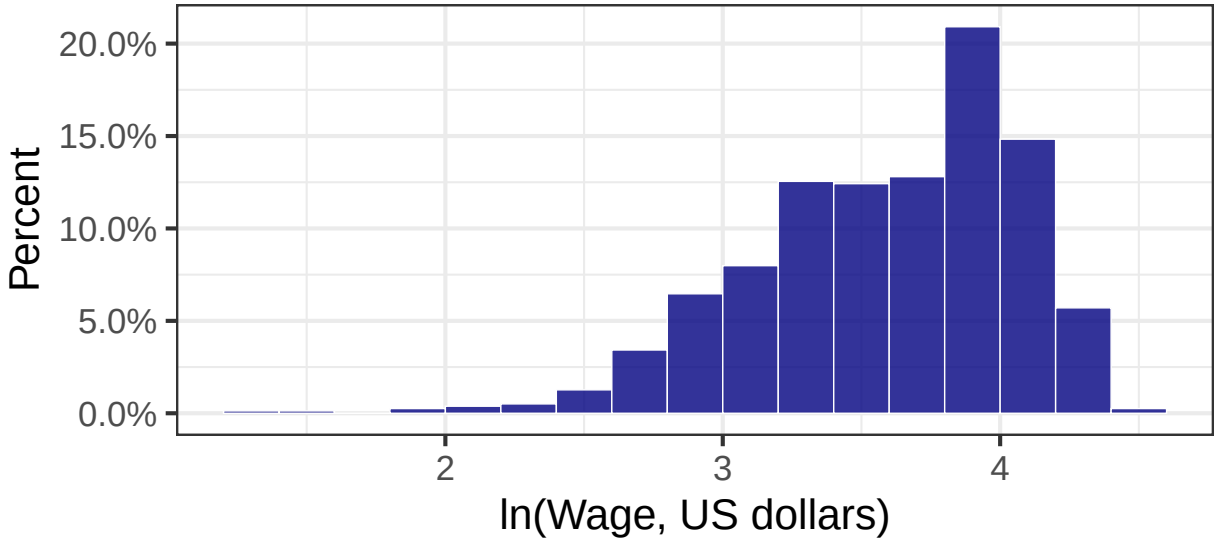
After the data has been loaded from osf.io, first I checked the summary of the variables (e.g. "earnwke", "uhours", "grade92", age, female or not) I intended to use in my analysis. As the first step, I filtered the dataset for lawyers and then I substituted the NA values with "Missing" strings. Second, I created variables for hourly wages, the logarithm of an hourly wage, dummy variables for females, and the square of age and education level (grade92). I also created a dummy variable for every different citizenship. Finally, I applied a filter to set only those values, which have a connection with lawyers.

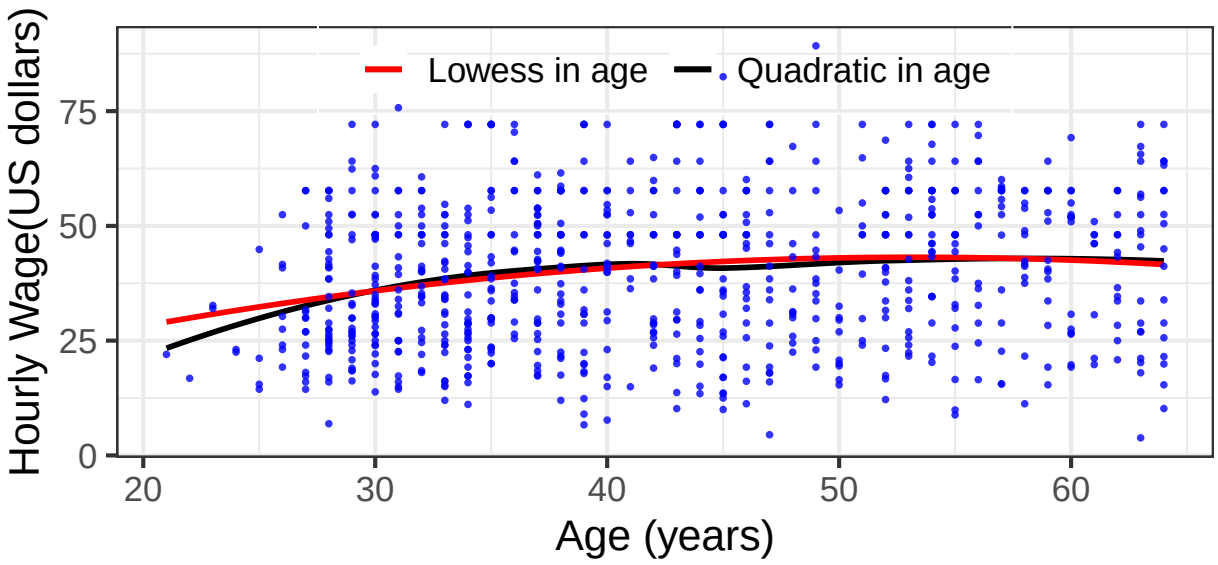
Analysis

Before the analysis, I checked the hourly wage observations (also checked its logarithm) for extreme values with a histogram graph, but I did not find any. After that, I checked the pattern of association between hourly wage and the age of workers. The graph shows what we can expect: As the worker gets older they started to earn more and more. The interesting part of it is that the slope of the curve is much smaller than I expected. At the start of the prediction, I created 4 models: the first model investigates the relationship of hourly wage and the age of the worker. The second model reviews also the connection of the education level. The third also checks variables like race, number of children, and sex. The last model reviews also the influence of different citizenships. The evaluation of these models shows that the values of AIC and BIC of the models are decreasing. At the earlier stage of the analysis, I added too many variables, since the AIC and BIC were increasing by model 4, but I changed it until it became less than model 3. RMSE is between 15.9 and 15.5, but it is also decreasing. R2 is also quite close to 1 in terms of 2, 3, and 4 models. In the case of the first model, it is only 0.35. By cross-validation, the p5 graph shows the prediction performance of the models and it is decreasing dramatically from the first model to the second, but interestingly it strat to raise back up, so after the second model I got worse, higher RMSE. For the prediction, I would use only the second and third models, but in this analysis, I continued to work with all 4 models to see the results as well. By the actual prediction, I wanted to check the possible wage of a lawyer, who is white 35 years old, has one child, native American. With a higher confidence interval (95%) the lawyer would earn around 60-62 dollars and by lower confidence interval this number changed around 49 -52 dollars.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

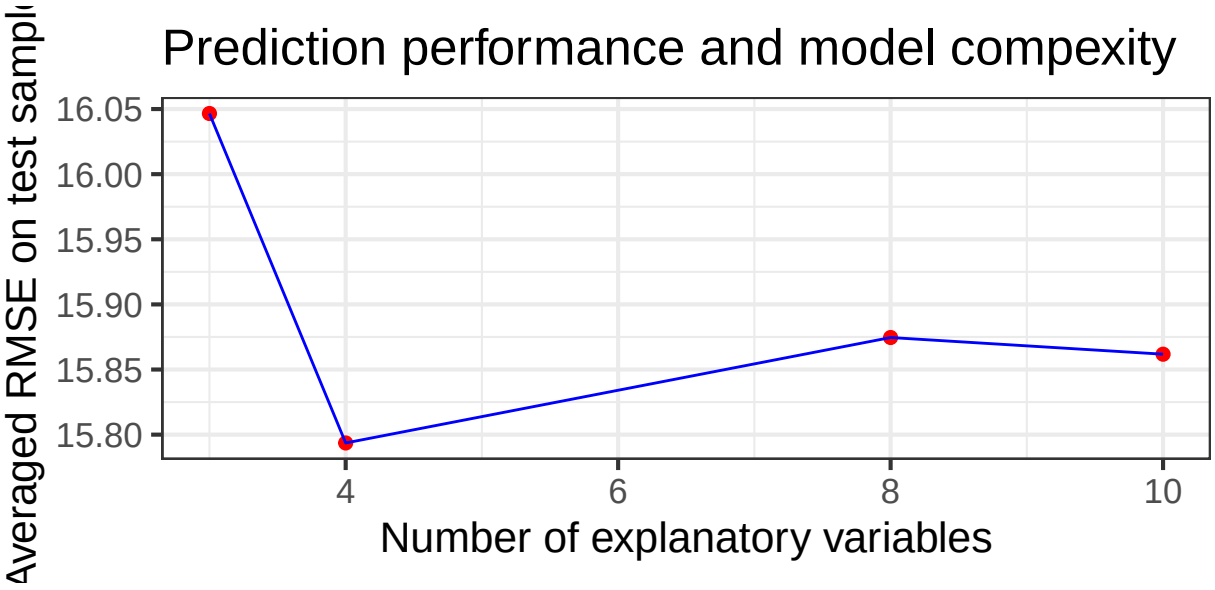

12%
9%
6%
3%
0%
Percent
0
25
50
75
Wage(US dollars)


20.0%
15.0%
10.0%
5.0%
0.0%
Percent
2
3
4
ln(Wage, US dollars)




```
## Linear Regression
##
## 789 samples
##   2 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 632, 632, 630, 631, 631
## Resampling results:
##
##   RMSE      Rsquared    MAE
##   16.01871  0.03377124  13.46495
##
## Tuning parameter 'intercept' was held constant at a value of TRUE

##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -38.15 -12.25  -0.20  12.23  46.30
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.72186    9.66079   0.489  0.62514
## age          1.44466    0.45554   3.171  0.00158 **
## agesq       -0.01357    0.00511  -2.656  0.00807 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16 on 786 degrees of freedom
## Multiple R-squared:  0.03564,    Adjusted R-squared:  0.03319
## F-statistic: 14.53 on 2 and 786 DF,  p-value: 6.385e-07
```




```
##                  Model1    Model2   Model3   Model4
## Predicted     38.66167 29.067900 31.40058 31.60408
## PI_low (80%)  18.11966  8.855733 11.17687 11.40226
## PI_high (80%) 59.20369 49.280067 51.62428 51.80591
```

```
##                   Model1    Model2     Model3   Model4
## Predicted     38.661674 29.067900 31.4005770 31.60408
## PI_low (95%)   7.223364 -1.865623  0.4492874  0.68622
## PI_high (95%) 70.099983 60.001423 62.3518666 62.52195
## PI_low (80%)  18.119657  8.855733 11.1768736 11.40226
## PI_high (80%) 59.203690 49.280067 51.6242804 51.80591
```

reg1	reg2	reg3	reg4
wageh	wageh	wageh	wageh
4.722 (9.387)	-139.5*** (20.39)	-437.3. (262.1)	-487.6. (259.2)
1.445** (0.4527)	1.438** (0.4421)	1.250* (0.5030)	4.244. (2.469)
-0.0136** (0.0052)	-0.0134** (0.0050)	-0.0114* (0.0057)	-0.0686 (0.0577)
	3.206*** (0.4106)	17.51 (12.25)	17.30 (11.93)
		-0.1686 (0.1434)	-0.1661 (0.1398)
		-2.738* (1.195)	6.274 (4.486)
		0.7026 (0.5903)	
		0.2644 (0.5284)	
			0.0004 (0.0004)
			14.38 (13.01)
			-0.2157* (0.1059)
			-0.2907 (0.3329)
Heteroskedas.-rob.	Heteroskedas.-rob.	Heteroskeda.-rob.	Heteroskeda.-rob.
6,617.2	6,588.2	6,587.6	6,586.3
6,631.2	6,606.9	6,624.9	6,633.0
15.969	15.659	15.573	15.521
0.03565	0.07281	0.08289	0.08905
0.03319	0.06926	0.07467	0.07852
789	789	789	789
2	3	7	9

intercept	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
TRUE	16	0.0338	13.5	0.769	0.0236	0.579

RMSE	Rsquared	MAE	Resample
15.7	0.0136	13.1	Fold1
15.8	0.0274	13.4	Fold2
15.3	0.0591	12.9	Fold3
17.3	0.0105	14.4	Fold4
16	0.0583	13.6	Fold5

Resample	Model1	Model2	Model3	Model4
Fold1	15.7	15.2	15.4	15.2
Fold2	15.8	15.3	15.2	15.2
Fold3	15.3	15.1	15.6	15.6
Fold4	17.3	17.4	17.2	17.3
Fold5	16	15.6	15.6	15.7
Average	16	15.8	15.9	15.9