Assignment 1 - Predicting Software Developer Wages

A. BOGNAR

Introduction

The CPS survey collects employment and demographic information in the US. In my report I try to use it to predict wages of developers. Specifically, I grouped *Computer programmers*, *Software developers* and *Web developers*. These fields are similar in earnings and other characteristics. Thus, predictions hopefully can provide some insight for someone planning to join this field about wage expectations.

Data

For useful predictions, I discarded data points that can not be predicted with high certainty, such as:

- Self-employed workers who have very unique wage calculation.
- Under 20 hour long work weeks usually come with unique situations (i.e. consulting).
- Wages under age 22 are not very stable.
- Education levels are diverse. I included levels above High School except Master's degrees due to lack of data.

Models

Based on patterns of association between wages and other variables I constructed the following predictive models:

Models 1 predicts wages entirely based on education, this variable has the highest association with level of wages.

Model 2 adds age of employees. Older developers earn more into their 50s, after which they tend to earn less.

Model 3 accounts for the actual title of the job (Software developers earn the highest), gender , race and number of children of workers as well as the hours worked weekely and if they were born in the US or not.

Model 4 adds additionaly details to the previous model by accounting for variables that have different patterns together. For example, women on average earn less than man, but women with 3 or more children earn more than fathers of 3+ kids. Granted only 14 such women are included, therefore, more data is needed before generalizing this result.

Table 1: Models to predict software developer wages

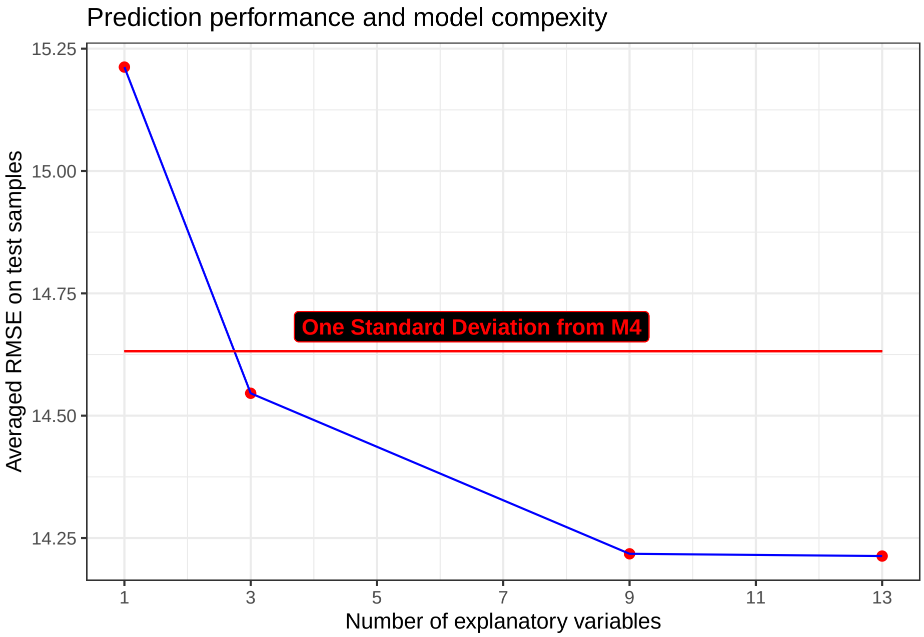
	(1)	(2)	(3)	(4)
Intercept	-83.2***	-127.8***	-219.5***	-222.8***
	(10.2)	(10.3)	(56.0)	(56.1)
Level of Education	2.83***	2.68***	2.44***	2.42***
	(0.239)	(0.225)	(0.232)	(0.232)
Age		2.23***	1.96***	2.08***
		(0.229)	(0.253)	(0.258)
Age squared		-0.022***	-0.019***	-0.019***
		(0.003)	(0.003)	(0.003)
Job Title			0.114*	0.115*
			(0.055)	(0.055)
Gender			-6.30***	-1.09
			(0.802)	(3.12)
Hours worked			-0.237***	-0.235***
			(0.060)	(0.059)
Foreigner (binary)			1.84*	-0.702
			(0.929)	(3.75)
Number of children			1.31***	1.04**
			(0.384)	(0.401)
Race			0.025	1.17
			(0.283)	(1.11)
Age x Gender				-0.142*
				(0.072)
Age x Foreigner (binary)				0.064
				(0.092)
Gender x Number of children				1.11
				(1.02)
Age x Race				-0.030
				(0.028)
BIC	16,258.3	16,083.3	16,029.2	16,052.3
RMSE	15.16	14.44	14.08	14.05
Observations	1,963	1,963	1,963	1,963
No. Variables	1	3	9	13

Which model to choose from the 4? Perhaps the one most closely approximating the patterns in the data, according to BIC measure this is Model 3. Or the one with the lowest average error in it's predictions which is Model 4 according to RMSE.

However, the level of error is sensitive to how we test it, to get a better understanding we can run multiple simulations. The below table shows the results of these. While Model 4 is still the most accurate, the difference is very small and can vary a lot trial to trial. Therefore, it may be better to choose a model built on associations that are likely to not change.

Resample	Model 1	Model 2	Model 3	Model 4
Fold1	15.8	15.12	14.74	14.72
Fold2	15.3	14.61	14.33	14.35
Fold3	14.6	13.88	13.66	13.62
Fold4	15.1	14.55	14.13	14.14
Fold5	15.2	14.25	13.94	13.90
Average	15.2	14.55	14.22	14.21
RMSESD	0.4	0.46	0.41	0.42

The below graph shows that Models 2-4 are really close in performance but Model 2 is lot less complex.

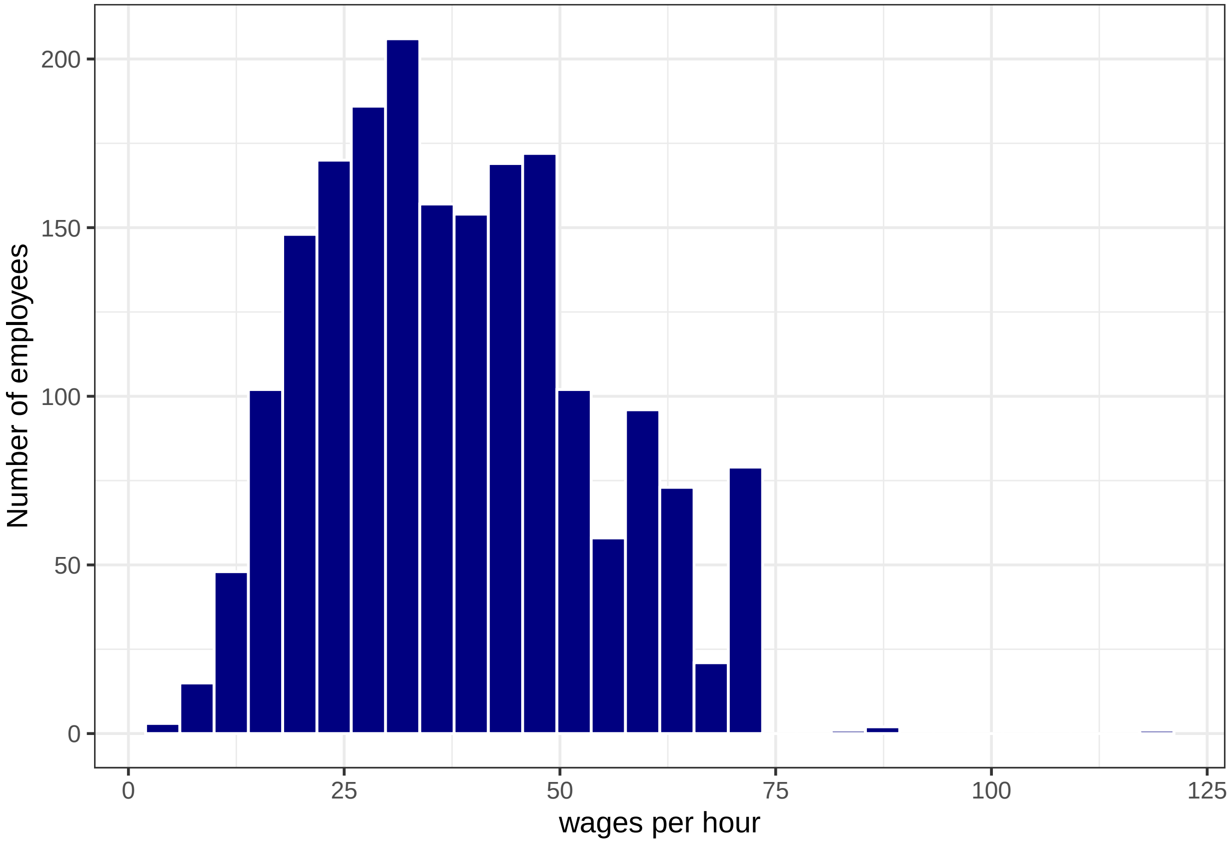


Conclusion

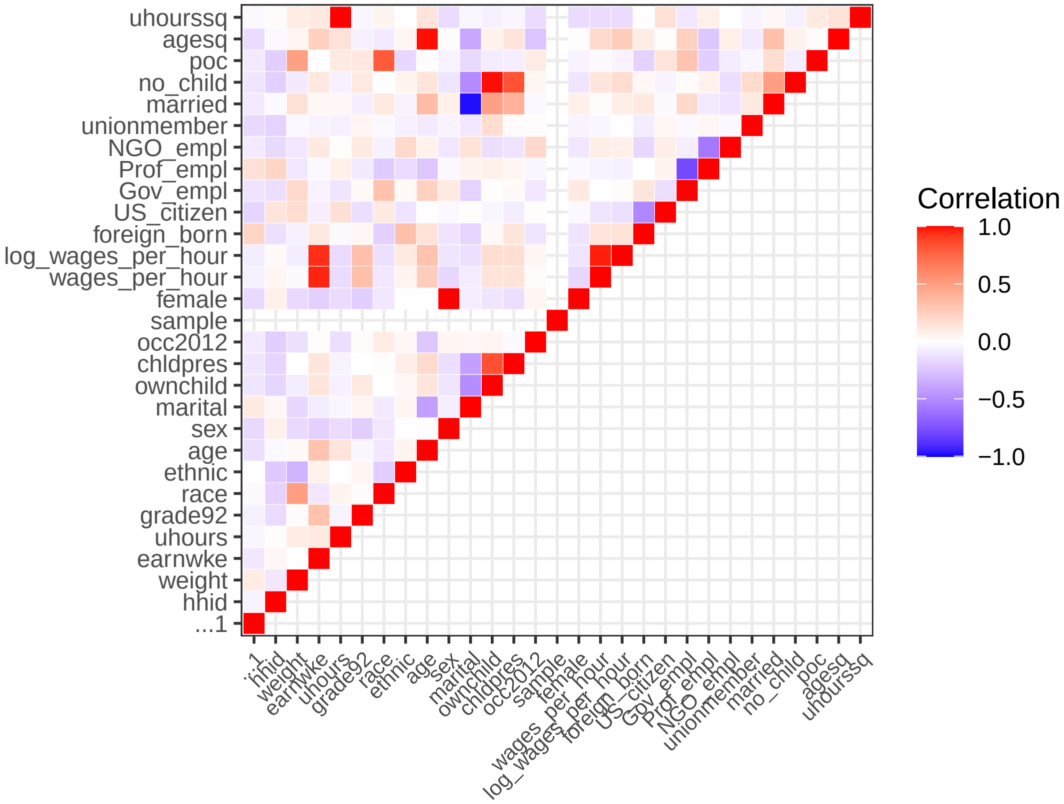
My preferred Model is *Model 3* since it adds considerable accuracy while most associations seem to be generally strong. Some of these associations may become weaker in the future or in specific occupations. Predictions for very specific cases with low sample sizes such as mothers with 3+ kids are not as reliable. In general avg. 14 dollars/h prediction error with 38dollars/h avg. wages is really high, individual wages are difficult to predict with high levels of accuracy.

Appendix

Wages in the sample seem to be closer to normal than lognormal distribution, therefore, I decided not to transform the variable. In addition, there were only a few extremely high wages but this is to be expected in this competitive field, I did not see a reason to exclude them.

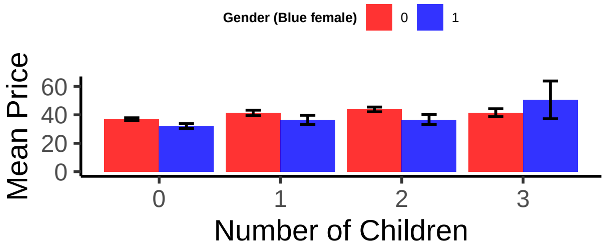


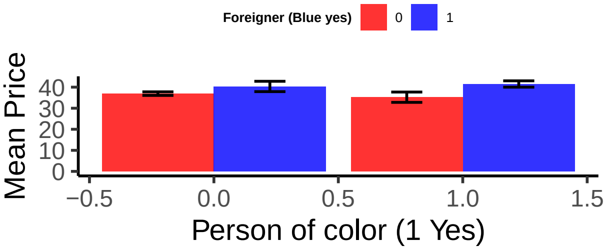
Correlations guided many of my variable choices, I created new variables such US citizenship that turned out to be less predictive than already existing ones.

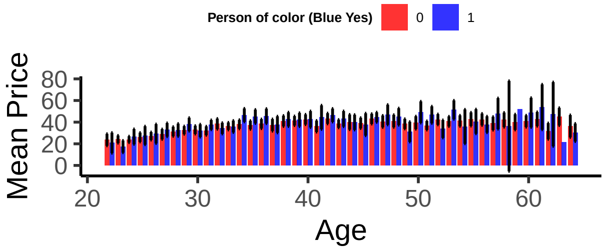


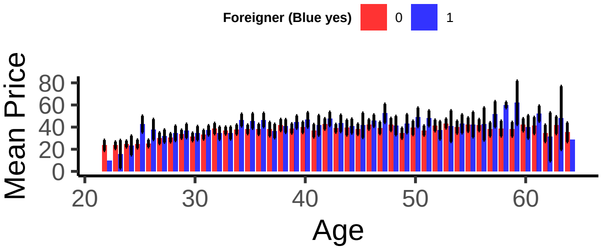
I grouped number of children above 3 for better interpretation, for women even then there were few cases which makes for big variation in this category. However, women with many children tend to earn more than man with many children, further data would be required to validate this pattern.

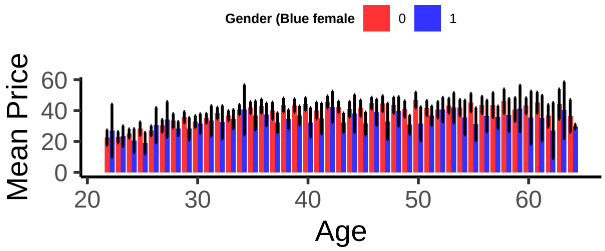
Other interesting patterns below: foreign born developers and people of color earn more than locals and white people for most of their lives but not in really young and really old ages. Women start out earning more than men and fall behind in their late twenties and thirties.

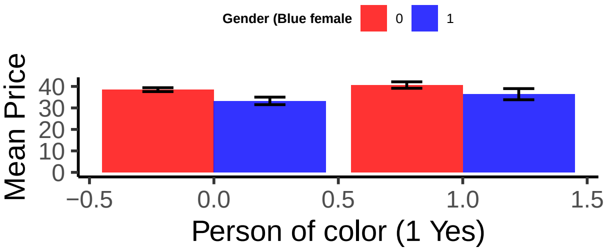




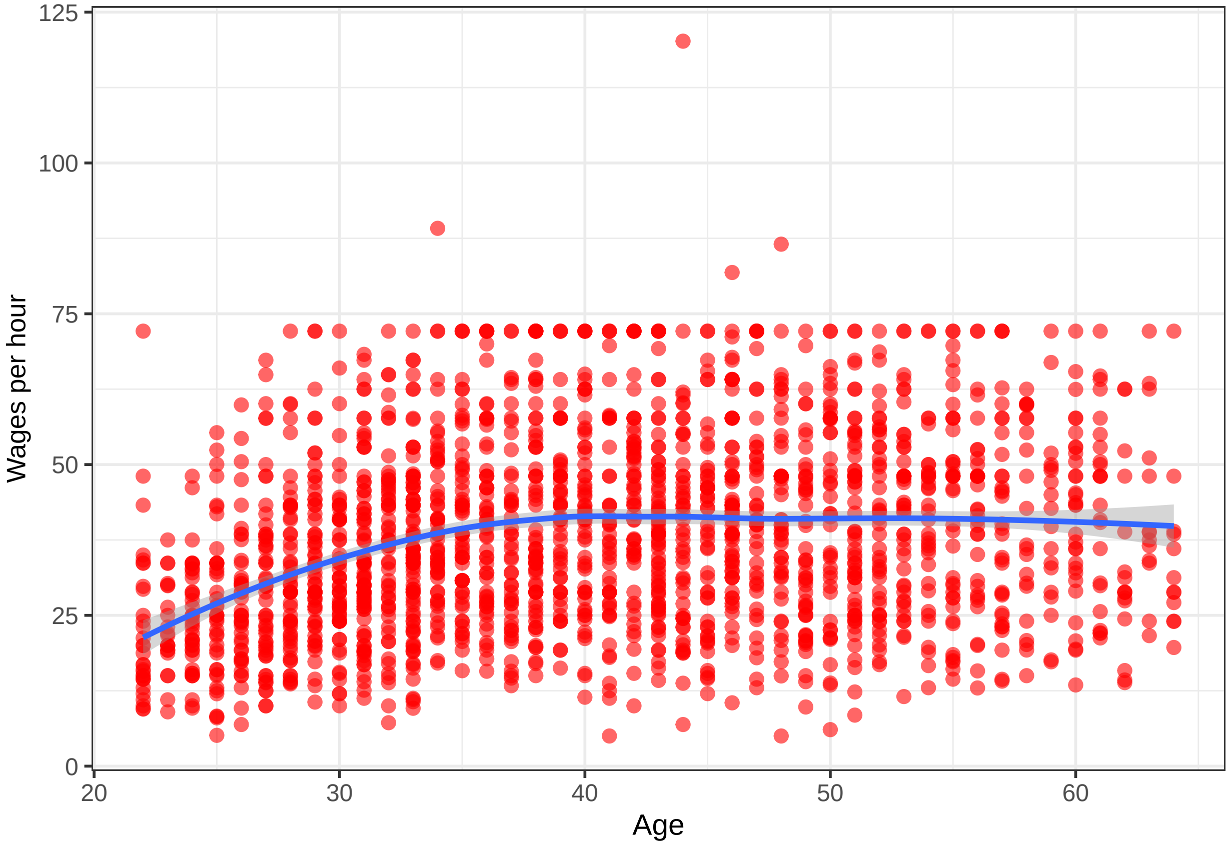








Wages mostly stagnate by age 40, slightly decreasing on average.


Wages per hour
Age