Assignment 1

Alima Dzhanybaeva

In this assignment, four predictive models for earnings per hour will be constructed using the CPS-earnings dataset. For this work, all 'Computer and Mathematical' occupations were chosen for the analysis.

In order to obtain the variable for earnings per hour the original *earnwke* (weekly earnings) variable was divided by *uhours* (number of hours). Further, the dataset was filtered for an hourly wage less than 73, as the 99 percentile for this variable is 72.12 while the maximum value is 120 (Appendix: Table 3).

For the explanatory variables, *education* was proven to be a strong driver of people's wages in many studies, as better-educated individuals are considered and indeed are more productive. This correlation was also found in the dataset (Graph 1). Therefore, the dummy variables for different levels of education level were created and chosen for the simplest *Model 1*.

Another variable that affects people's wages is age, as younger workers have less experience. However, as it is shown in Graph 2 after one point the earnings per hour start to decline with age. Consequently, *age* and newly generated *age squared* were also added to *Model 2*.

For *Model 3* three dummy variables were generated: *female*, *married*, and *child*. *Female* is believed to be a good explanatory variable for earnings, as the gender pay gap is still present nowadays and women are often under-represented in decision-making roles. The proof of it is shown in Graph 3. As for *married* and *child*, people in wedlock and with children tend to work longer hours and put in increased effort at work as they're well aware that their current wage affects not only them. These correlations are indeed present for the chosen occupations (Graph 4-5).

***Model 4* includes all previously mentioned variables with the addition of four interaction terms: *education*female*, *education*married*, *female*child*, and *female*married*.**

RMSE in the full sample, cross-validated RMSE and BIC in the full sample are presented in the tables below.

Table 1: Linear regression evaluation

Model	N predictors	R-squared	Training RMSE	BIC
(1)	4	0.1044428	14.74245	38944.69
(2)	6	0.1979340	13.95173	38439.88
(3)	9	0.2255230	13.70968	38299.64
(4)	19	0.2275340	13.69187	38371.95

[H]

Table 2: Cross-validation

Resample	Model1	Model2	Model3	Model4
Fold1	14.68559	13.73177	13.39637	13.39637
Fold2	14.92069	14.06386	13.87710	13.87710
Fold3	14.66523	13.86202	13.67343	13.67343
Fold4	14.76237	14.24060	14.01734	14.01734
Average	14.75881	13.97591	13.74304	13.76449

As we can see from the Table 1, due to a better fit RMSE decreases as more variables are added to the model. Nevertheless, my choice falls on the *Model 3*, as the difference in RMSE in the full sample between the two most complex models is negligible, and both RMSE in cross-validation and BIC in the full sample are the lowest for *Model 3* indicating that *Model 4* is overfitted.

Appendix

Table 1: Descriptive statistics

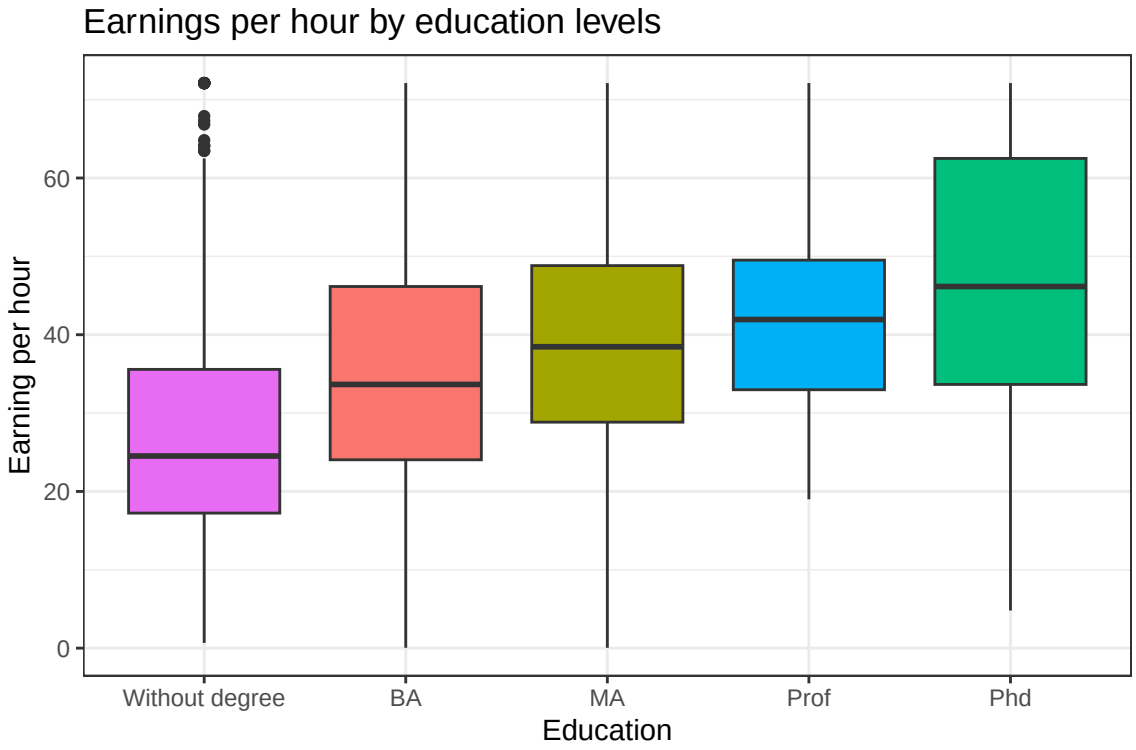
	Median	SD	Min	Max	P99	N
Weekly earnings	1346.00	678.49	1.78	2884.61	2884.61	4740
Weekly hours worked	40.00	6.56	1.00	80.00	60.00	4740
Earning per hour	32.05	15.74	0.04	120.19	72.12	4740
Female	0.00	0.45	0.00	1.00	1.00	4740
BA Degree	0.00	0.50	0.00	1.00	1.00	4740
MA Degree	0.00	0.41	0.00	1.00	1.00	4740
Professional Degree	0.00	0.08	0.00	1.00	0.00	4740
PhD	0.00	0.13	0.00	1.00	1.00	4740
Age	40.00	11.14	16.00	64.00	63.00	4740
Has child	0.00	0.49	0.00	1.00	1.00	4740

Table 4: Results of linear regressions

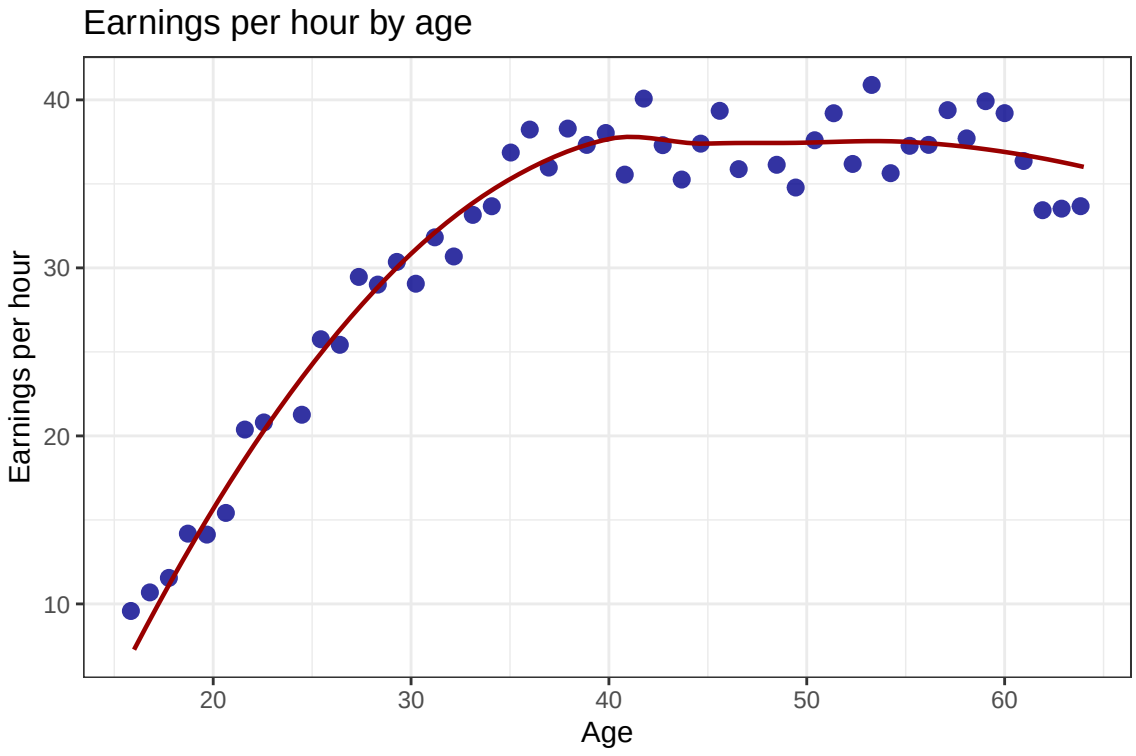
	Model 1	Model 2	Model 3	Model 4
(Intercept)	27.206**	−19.387**	−14.438**	−14.686**
	(0.389)	(2.709)	(2.843)	(2.853)
BA	8.538**	8.514**	8.447**	8.653**
	(0.500)	(0.476)	(0.468)	(0.794)
MA	12.415**	11.365**	11.355**	12.109**
	(0.611)	(0.582)	(0.574)	(1.099)
Prof	16.434**	14.472**	14.041**	12.522*
	(2.721)	(2.578)	(2.535)	(5.397)
PhD	19.935**	18.130**	17.935**	12.400**
	(1.725)	(1.638)	(1.610)	(3.191)
age		2.026**	1.784**	1.785**
		(0.137)	(0.148)	(0.148)
age_sq		−0.020**	−0.017**	−0.017**
		(0.002)	(0.002)	(0.002)
female			−5.311**	−5.741**
			(0.451)	(0.975)
married			1.281**	1.809*
			(0.496)	(0.853)
child			1.091*	1.514**
			(0.500)	(0.587)
BA × female				1.439
				(1.055)
MA × female				0.934
				(1.266)
Prof × female				−3.641
				(6.108)
PhD × female				2.581
				(3.694)
BA × married				−0.970
				(0.952)
MA × married				−1.514
				(1.232)
Prof × married				3.102
				(5.838)
PhD × married				6.848
				(3.531)
female × child				−1.481
				(1.024)
female × married				0.147
				(1.006)
Num.Obs.	4732	4732	4732	4732
BIC	38 944.7	38 439.9	38 299.6	38 372.0
RMSE	14.74	13.95	13.71	13.69

* $p < 0.05$, ** $p < 0.01$

Graph 1: Earnings per hour by education levels

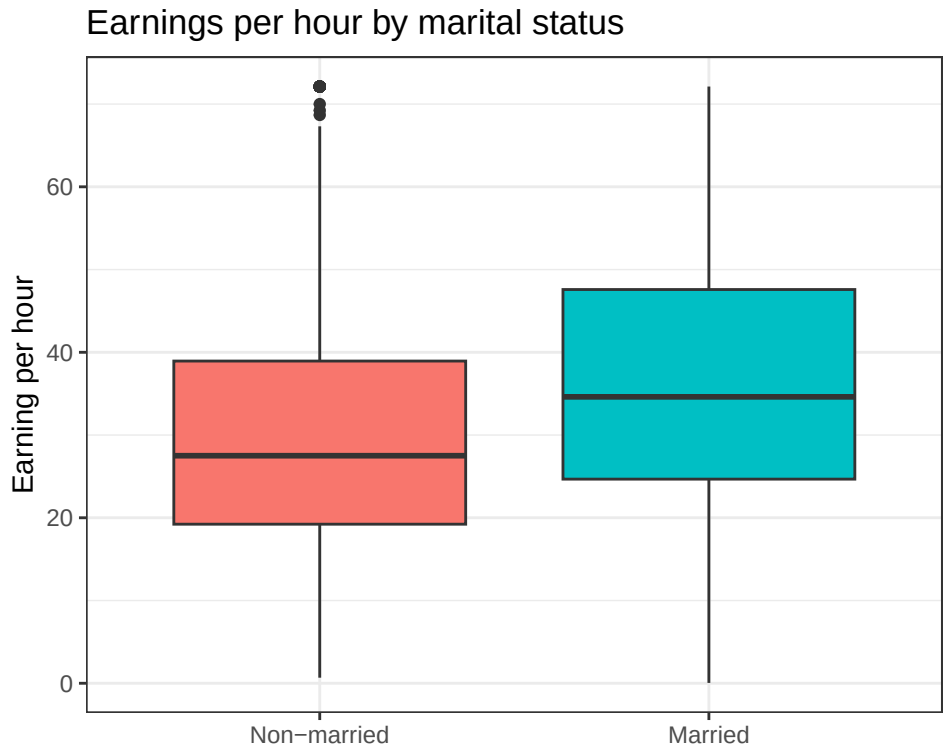


Graph 2: Earnings per hour by age

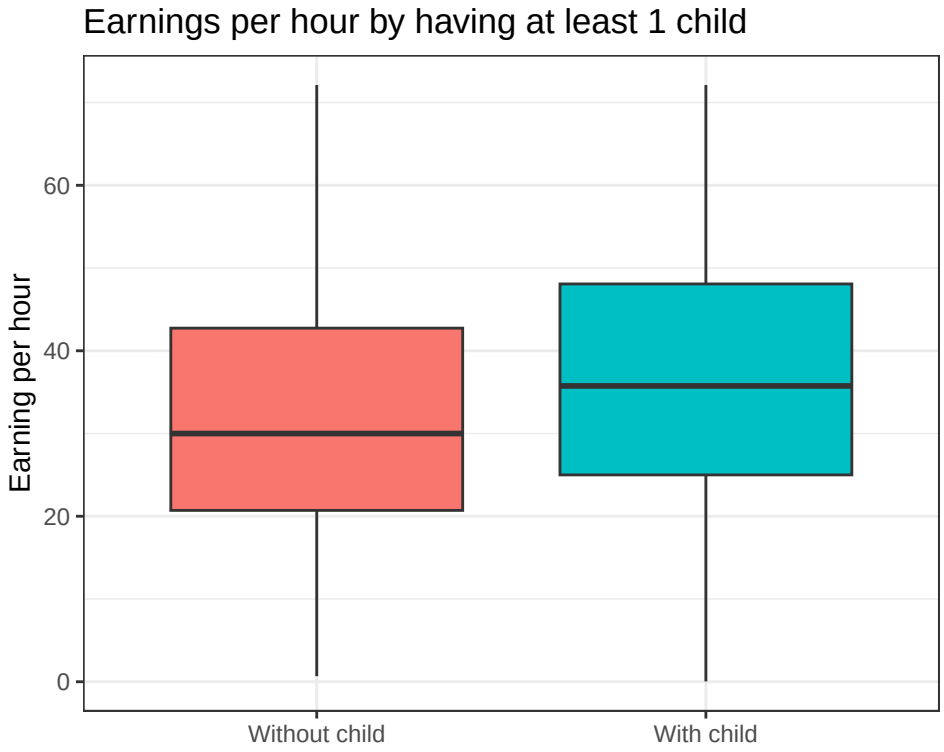


Graph 3: Earnings per hour by gender

Graph 4: Earnings per hour by marital status



Graph 5: Earnings per hour by having at least 1 child



Graph 6: Boxplots by two variables