Assignment 1

Zsiros, Gabriella

2023-01-26

In this exercise I am taking one particular occupation group of the CPS dataset, **Marketing and Sales Managers** (occupational code: 50). This selection results in 1033 observations. The salary is captured as weekly earnings in the data set (`earnwke`), completed by usual work hours (`uhourse`). Based on these two, I introduce a calculated variable called `wage`, showing the hourly earnings. Sex, children in household, marital status, ethnic groups, race and State within US is also captured, while`grade92` signifies the highest completed educational grade.

A closed look at the educational distribution shows that the majority of the data is between grades 39 and 44, meaning that the highest education level is between High school and MA University degree, so I am taking these observations in the sample.

A closer look at the ethnic groups shows 94% missing data, other variables mostly do not contain missing values. Extreme values in age are not sowing up anymore, since the education is already filtered.

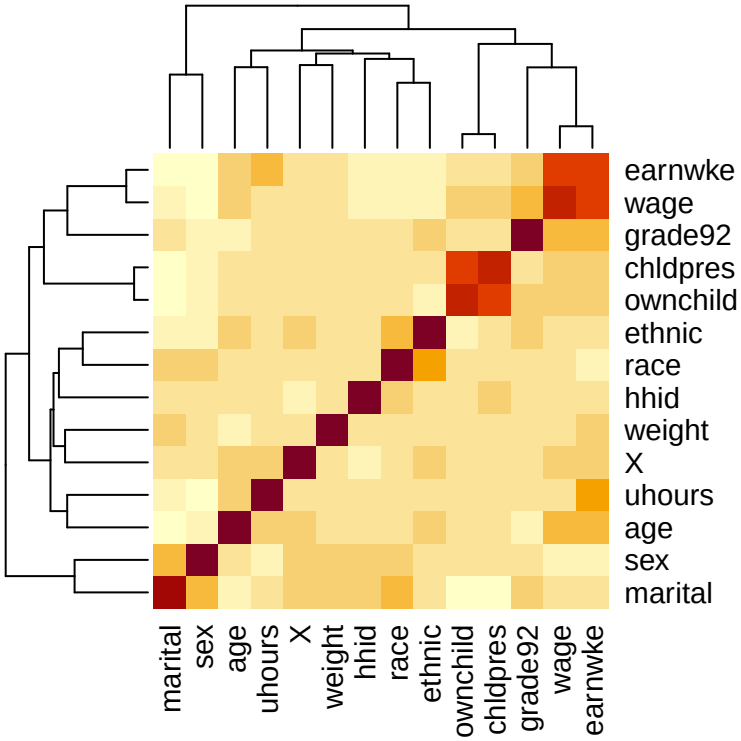
Frequency	Min	P1	D1	Q1	Me	Q3	D9	P99	Max
1014	0.025	8	14.423	20.2783	30.31518	44.1826	57.6922	72.11525	100

A statistical summary of the hourly earnings show some extreme values (1st and 99th percentile) which are dropped from the sample.

Variables

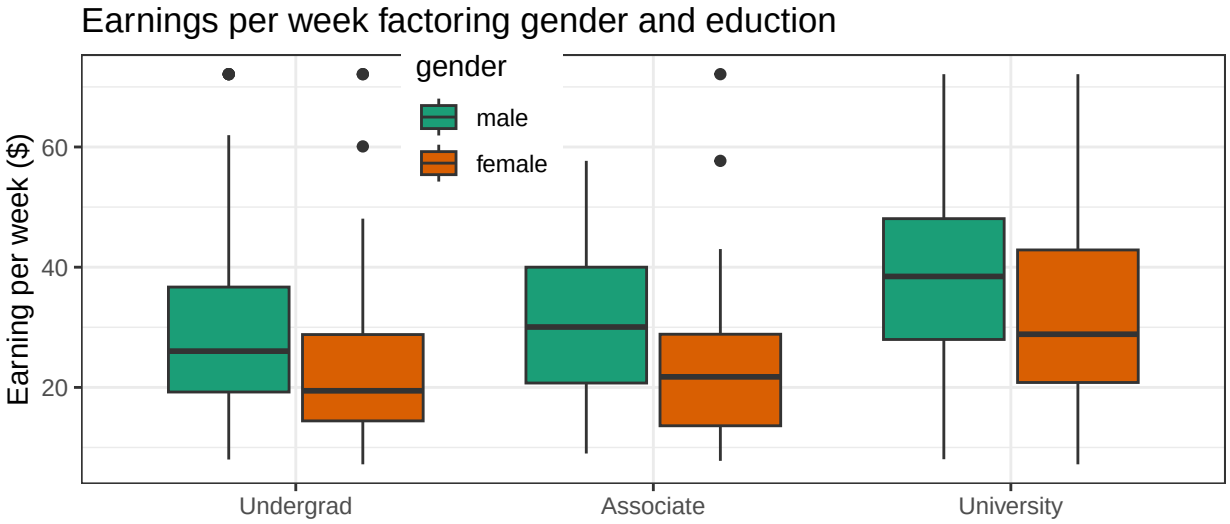
Whiel examining the correlations between the variables, we can see perfect or very high collinearity between variabels that more or less explain each other, like `earnwke` or `wage`, since horuly wage is calculated from the former. (`Chldpres` and `ownchild` both represent children in a household, therefore a very high correlation is also expected.) Education and age seem to have a higher correlation with the hourly wages than other variables.

Education grade	Observations	Proportion
34	1	0.0009681
37	4	0.0038722
39	103	0.0997096
40	131	0.1268151
41	19	0.0183930
42	43	0.0416263
43	553	0.5353340
44	165	0.1597289
45	5	0.0048403
46	9	0.0087125

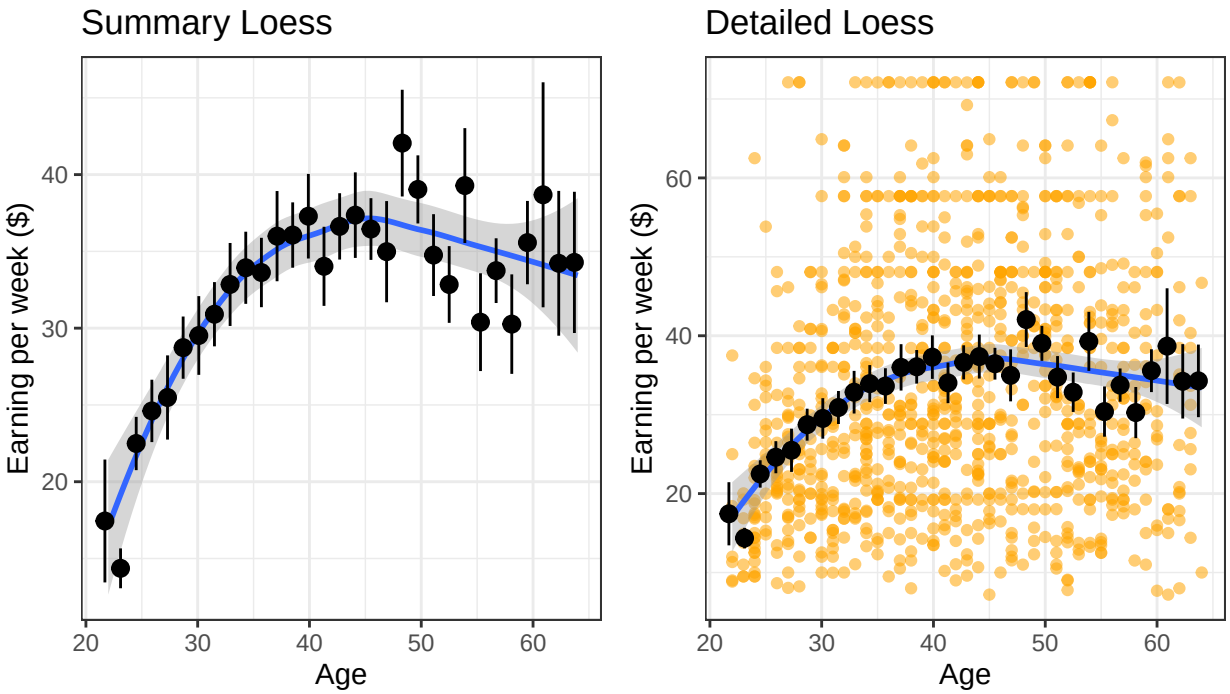


Graphical data overview

Despite the conclusion of the corrleation matrix, when we plot the data factoring the sex, it shows a visible difference between males and females in the same educational category.



Lowess fitting shows a great increase until age 40 when we are using statistical summary, but taking a closer look a plotting all the datappoints,



Linear regression models

The first model had only the regression of hourly wage on age and second introduces gender factor.Even if the correlation has not highlighted this variable, based on the plot, it is worth to add to the model. The third is completed by education as a thee level factor and the fourth adds the usual working hours, which can potentially have a significance, since te wage is already standardized. (i.e Those who put more work hours in on a weekly basis, might earn more hourly, not just as an accumulated weekly salary)

The model shows that adjusted R2 is decreasing with each variable, although the difference between Model 3 and 4 is very low. RMSE remains the same, but BIC shows a slight increase, indicating a potential over fitting.

	Model 1	Model 2	Model 3	Model 4
(Intercept)	21.874	27.243	15.996	15.900
	(2.004)	(2.110)	(2.329)	(3.292)
age	0.277	0.226	0.293	0.293
	(0.046)	(0.046)	(0.044)	(0.044)
genderfemale		−6.713	−5.878	−5.871
		(0.980)	(0.934)	(0.951)
eduedu2			2.026	2.025
			(2.088)	(2.089)
eduedu3			11.128	11.128
			(1.119)	(1.120)
uhours				0.002
				(0.053)
Num.Obs.	1005	1005	1005	1005
R2	0.034	0.078	0.169	0.169
R2 Adj.	0.034	0.076	0.166	0.165
AIC	8385.5	8341.5	8240.4	8242.4
BIC	8400.2	8361.1	8269.9	8276.8
F	35.821	42.201	50.927	40.701
RMSE	15.64	15.29	14.51	14.51

Cross-validation

During cross validation, I split the dataset into 4 folds, each of them containing a different training section.

Model comparison

Comparing the RMSEs of the OLS linear regression model and the cross validated models, there is little difference in the third and second model based on RMSE, altough Model 3 performs best in BIC.