DA3 - Assignment 1

Haaris Afzal Cheema

Introduction

The aim of this assignment is to use the cps-earnings dataset to compare and contrast different predictive models (using linear regressions) based on multiple criterion such as RMSE and BIC in the full sample and the cross-validated RMSE. Also, the final section will discuss the model complexity and performance. This analysis will focus on **Financial Specialists** and the target variable will be **earnings per hour**, whereas for predictors, demographic variables such as age, gender, education level, race and marital status will be used.

Choice of Predictor Variables

Firstly, highest degree attained (education level) was included as one can expect that with every successive degree, you acquire more knowledge and become worth more and paid better. Next demographic variables such as age and sex were kept to include any patterns within their respective categories due to any potential discrimination. Age was also included as it is expected to be strongly and positively correlated with one's earnings. Lastly, current marital status was also included to explore whether familial responsibilities have an association with how much one earns.

Data Cleaning, Manipulation and Exploratory Data Analysis

As far as the categorical variables were concerned, their respective distributions were analyzed and consequently some categories with relatively low frequency were either dropped or clubbed together. The minimum earning per hour was set at 0, and a single value of 500, was dropped as it was drastically different from all of the observations in the data and it is possible that the input value could have been an error. Next, a summary output for the relevant variables was created (Fig 1). Moreover, variable distributions were chalked out (Fig 2 & 3) and for the age variable, the functional form was explored via a lowess curve against the target variable (Fig 4). It was decided to incorporate the quadratic and cubic terms for age.

Comparing model performance and relationship with model complexity

The models were built with increasing levels of complexity. The variables contained in each model can be viewed in the appendix (pg 2). Based on the full sample comparison (Fig 5), we can see that model 4 (the most complex one), has the lowest RMSE, which indicates that we expect it to have the lowest prediction error among all four models. Even when we evaluate the models based on the BIC metric, we can see that the metric has the lowest value for model 4, despite the fact that it is the most complex model. This indicates that even after the penalty imposed due to the higher model complexity of model 4, it still has a lower BIC value due to its superior model fit. Lastly, the cross-validated RMSE (Fig 6) also points towards model 4 as the best choice, as it has the lowest average RMSE. Model 4 has more or less the same RMSE variation amongst the different folds compared to model 3 and also has a slightly lower average RMSE, so it will remain our preferred model choice.

Based on the visual in the appendix (Fig 7), we can see that as the number of explanatory variables (model complexity) increases, the average RMSE of the test samples decreases, resulting in an increasingly better fit for the prediction. Initially there is a significant decrease in the RMSE for an increase in the explanatory variables, however, this becomes less steep for greater number of explanatory variables. This is because more complex models tend to over fit the original data and may represent detailed patterns in the original data and not in the live data which we are interested in. We can expect that for even more complex models, the average RMSE will start to increase.

Appendix

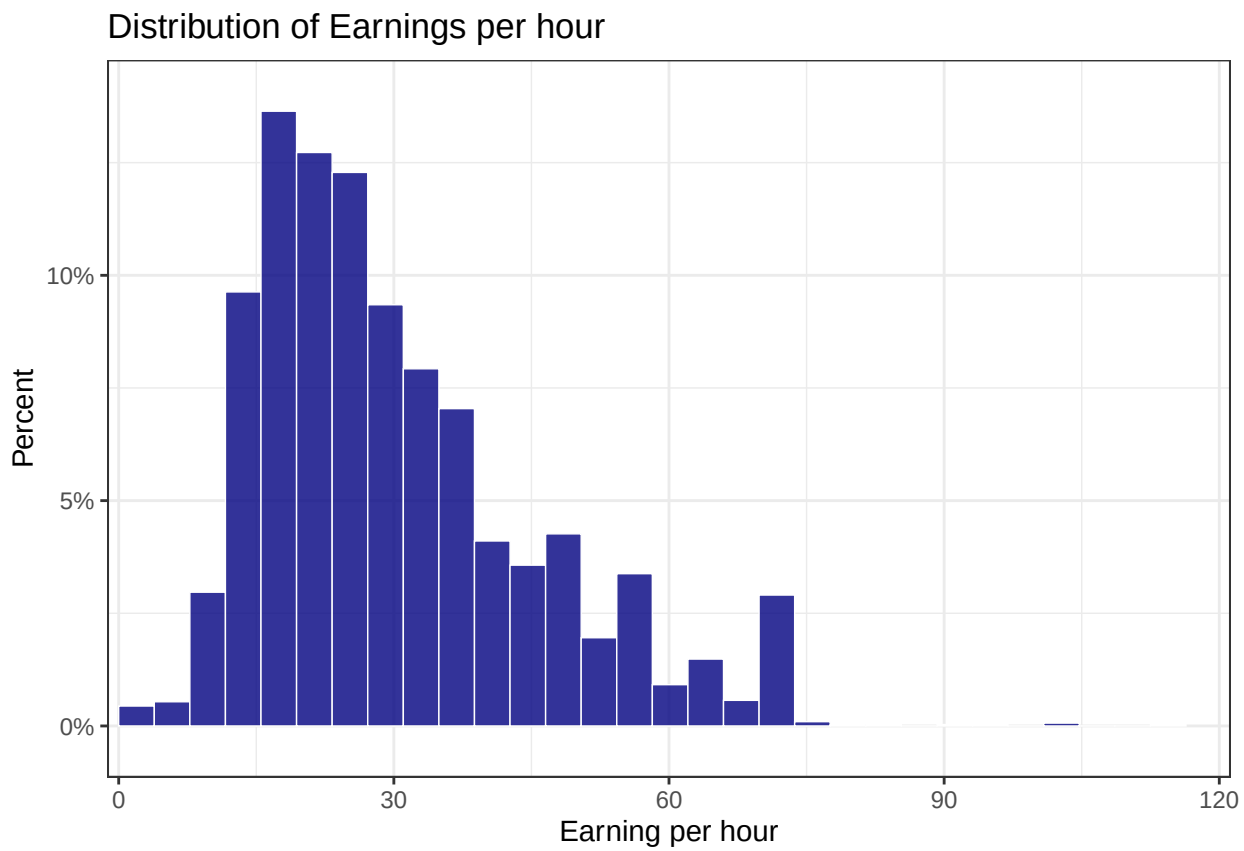
Fig 1: Data Summary Table

	Mean	Median	Min	Max	P25	P75	N
Earnings per hour	30.36	26.42	0.01	112.50	19.23	38.46	3167
Lower than Bachelors Degree	0.25	0.00	0.00	1.00	0.00	1.00	3167
Bachelors Degree	0.57	1.00	0.00	1.00	0.00	1.00	3167
Masters Degree	0.18	0.00	0.00	1.00	0.00	0.00	3167
Whites	0.84	1.00	0.00	1.00	1.00	1.00	3167
Blacks	0.08	0.00	0.00	1.00	0.00	0.00	3167
Asians	0.08	0.00	0.00	1.00	0.00	0.00	3167
Age	41.96	42.00	18.00	64.00	32.00	51.00	3167
Female	0.60	1.00	0.00	1.00	0.00	1.00	3167
Current Marital Status	0.63	1.00	0.00	1.00	0.00	1.00	3167

Model Details

- Model 1: Age and Age squared
- Model 2: Age, Age squared, Sex and Education Level (Highest Degree)
- Model 3: Age, Age squared, Sex, Education Level, Race and Current Marital Status
- Model 4: Age, Age squared, Age cubic, Sex, Education Level, Race, Current Marital Status and interaction terms between Sex and Education Level and Sex and Age

Fig 2: Distribution of Target Variable



The distribution for the target variable resembles a normal curve with the exception of a few extreme values
These extreme values are not frequent enough nor are they 'extreme' enough to warrant a log-transformation in this case
Generally wage distributions are considered to be right-tailed but for this subsetted data, the distribution does not resemble a typical wage distribution as there aren't too extreme values as mentioned above

Fig 3: Distributions of Predictor Variables

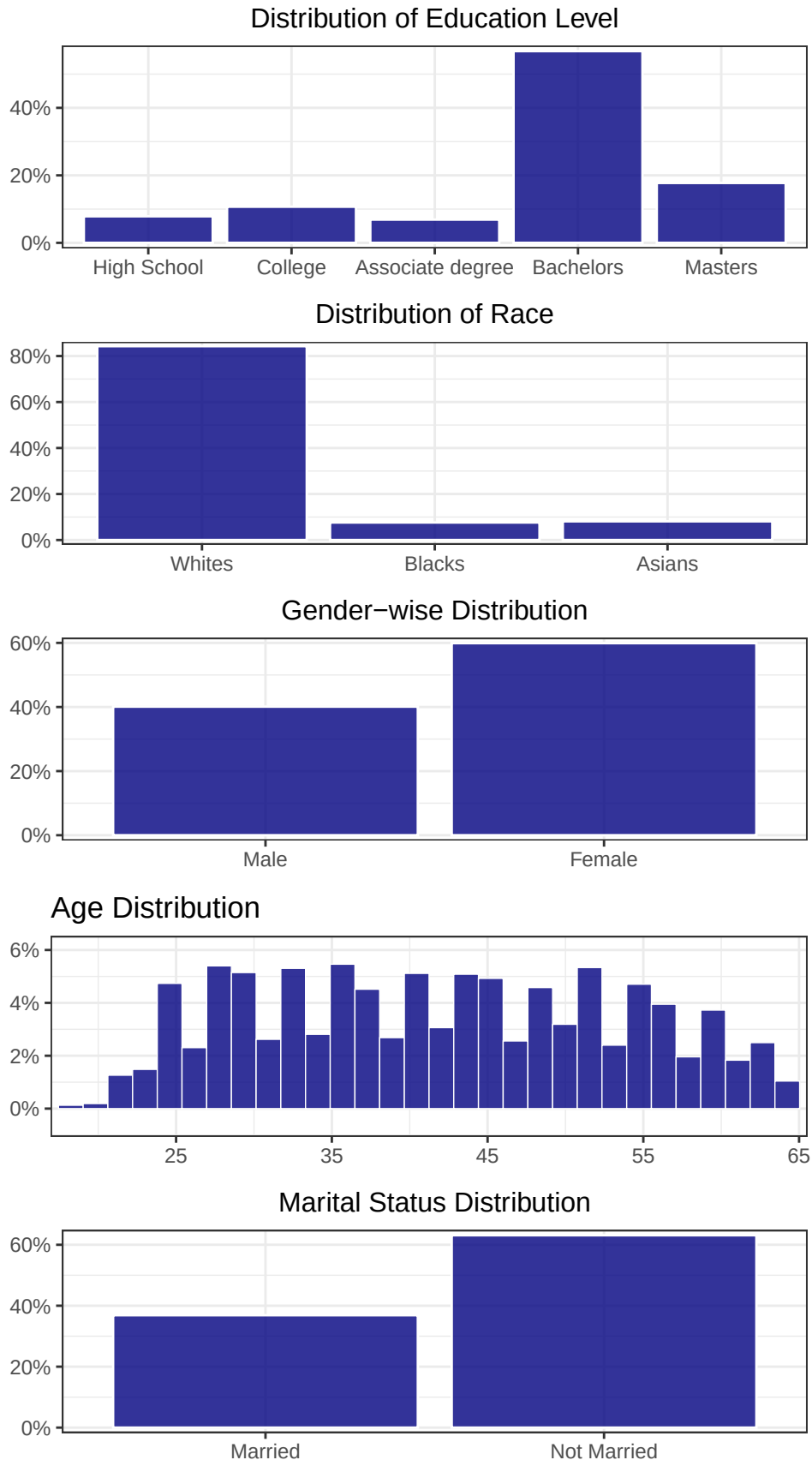
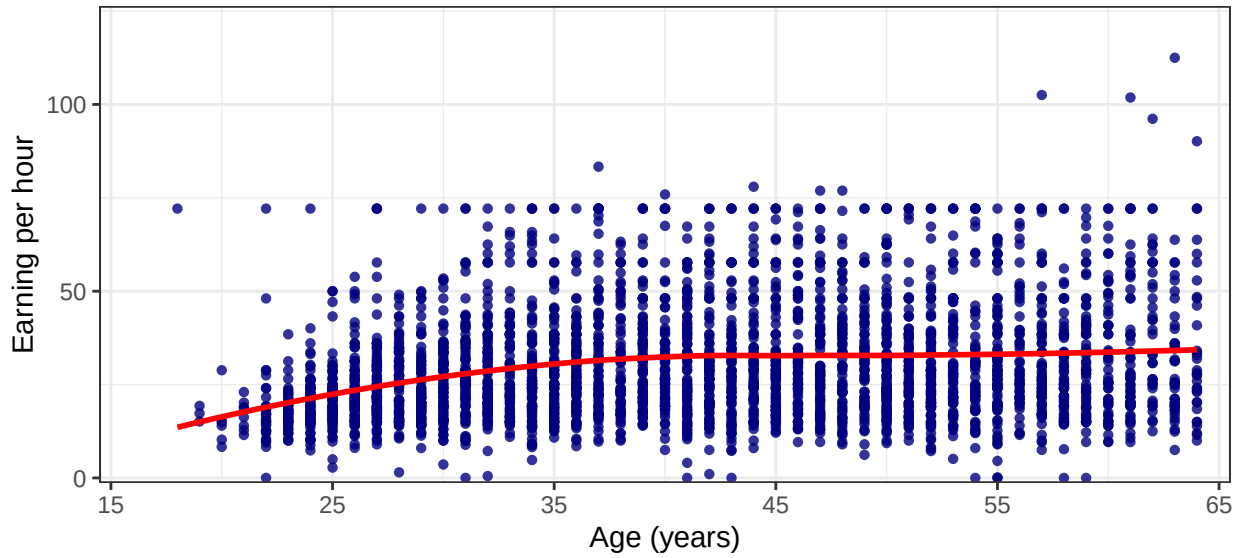
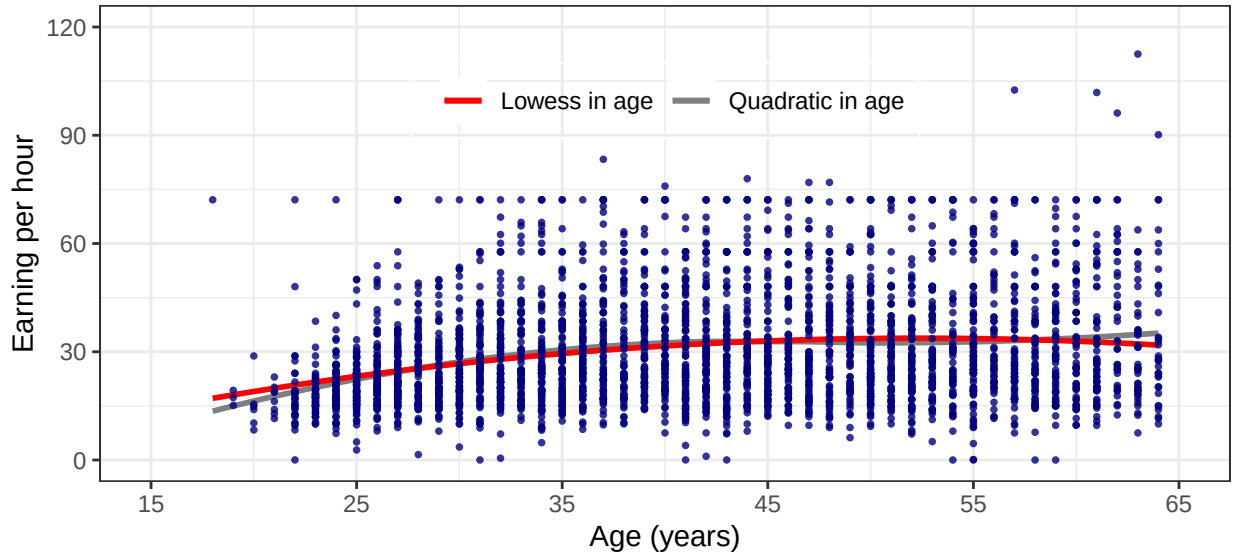


Fig 4: Association between Target Variable and Age (Lowess Explorations)





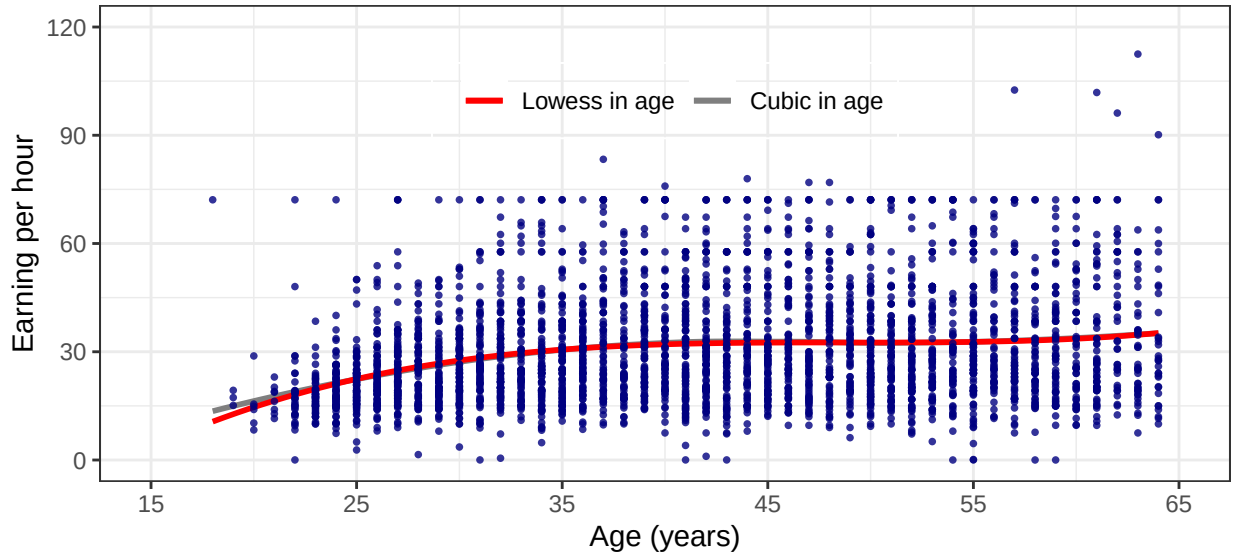


Fig 5: Comparing Predictive Models - full sample RMSE and BIC

	(1)	(2)	(3)	(4)
(Intercept)	-4.923	-2.776	-2.027	-34.35**
	(3.479)	(3.254)	(3.359)	(13.31)
age	1.478***	1.552***	1.500***	3.730***
	(0.1819)	(0.1695)	(0.1783)	(1.022)
agesq	-0.0141***	-0.0142***	-0.0137***	-0.0650**
	(0.0022)	(0.0021)	(0.0021)	(0.0250)
female		-6.206***	-6.110***	4.377*
		(0.5412)	(0.5390)	(1.852)
lower_BA		-7.737***	-7.607***	-8.788***
		(0.5429)	(0.5439)	(1.090)
MA		3.672***	3.613***	3.844***
		(0.7427)	(0.7387)	(1.020)
black			-3.253***	-3.452***
			(0.8632)	(0.8622)
asian			1.489	1.469
			(0.9894)	(0.9837)
current_marital			0.7859	0.2258
			(0.5458)	(0.5662)
agecu				0.0004*
				(0.0002)
female x lower_BA				1.987
				(1.256)
female x MA				-0.7961
				(1.463)
age x female				-0.2589***
				(0.0455)
AIC	26,131.0	25,674.8	25,661.2	25,626.3
BIC	26,149.2	25,711.1	25,715.7	25,705.1
RMSE	14.964	13.911	13.868	13.774
Observations	3,167	3,167	3,167	3,167
No. Variables	2	5	8	12

lower_BA reflects those individuals with a degree lower than a bachelors
MA reflects those individuals with a masters degree
current_marital indicates the current marital status of the individual
agesq and agecu are the quadratic and cubic terms respectively for the age variable

Fig 6: Comparing Predictive Models - cross-validated RMSE

Resample	Model1	Model2	Model3	Model4
Fold1	15.30926	14.24894	14.24860	14.13316
Fold2	14.53078	13.42208	13.32935	13.25253
Fold3	15.30382	14.01557	13.94419	13.83863
Fold4	15.16456	14.40886	14.41841	14.38756
Fold5	14.53333	13.57392	13.58934	13.52891
Average	14.97268	13.93906	13.91183	13.83414

Fig 7: Model Complexity and Performance