DA3 - Assignment 1 - Mahrukh Khan

Introduction

This assignment uses cps-earnings dataset. It builds predictive models using linear regressions on target variable, earnings per hour, for the Community and Social Services Occupation. It will compare model performance by analyzing the relationship between model complexity and performance by using metrics such as RMSE and BIC in the full sample and k-fold-cross-validated RMSE.

Choice of Predictors

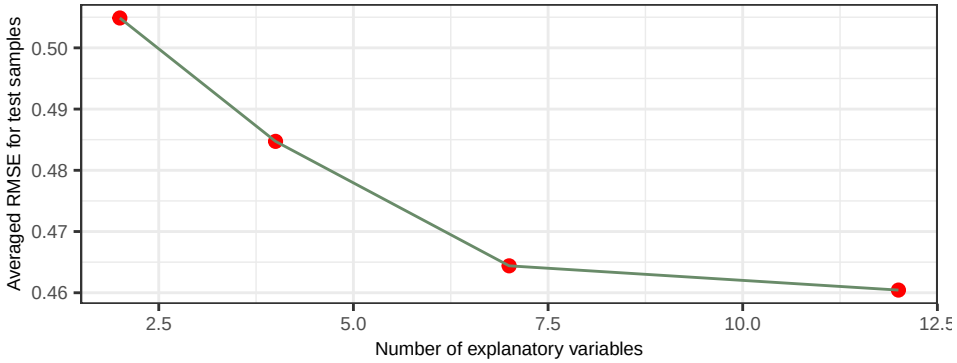
I chose demographic variables such as gender, age, race, education level and job sector. Due to higher female workers, community and social services occupation has been considered a gendered profession. Despite female dominance, the wage gap could potentially be a prevalent problem Also, traditionally, wages in public and private sector have been different. In addition, race discrimination can impact wages earned per hour. Another demographic variable is age which reflects one's experience and position in the work field, being positively associated with wage. Lastly, holding a higher degree can escalate job position in return boosting wage.

Models (Target Variable: Log of earnings per hour)

Model1: *race*;
Model2: *race, age and age squared*;
Model3: *race, age, age squared, gender and education level*;
Model4: *race, age, age squared, age cube, gender, education level, job sector, interaction term of race & education level, and interaction term of race & gender.*

Relationship between model complexity and performance

There is a total of 2,239 observations and, our regression models get more complex with an addition of new variables, the most complex model includes interaction terms. The best predictor model is selected by first comparing the BIC and RMSE values in the full sample followed by k-fold-cross-validation with k set at 5. Both approaches indicate that Model 3 and Model 4 can potentially be chosen as best predictor models. Both models have the lowest BIC, 3049 and 3055, and lowest RMSE, 0.471 and 0.467, in the full sample, respectively. They also have the lowest average cross-validated RMSE being 0.480 and 0.475, with a similar variation amongst the folds. Overall, Model 3 is the less complex model of the two. It has similar RMSE values that show lower prediction error and the lowest BIC which discourages over fitting. In conclusion, Model 3 is a better predictor model. *Please see Figure 6 and Figure 7 in the Appendix.*



A smaller averaged RMSE signals to a better prediction performance of the model. Initially, the drop in average RMSE is greater as more coefficients are added, but it almost flattens between Model 3 and Model 4. After a certain point the average RMSE will start increasing due to the growth in model complexity. Too many variables tend to over fit the original data, hence making it a bad predictor for the general pattern or population.

Appendix

Data Cleaning, Manipulation and EDA

- The frequency distribution for categorical variables was used to combine or keep relevant categories. *Relevant categories kept can be seen in Figure 1.*
- Earnings per hour was set between 1 and 100 dollars. Log transformation was conducted due to the right tailed distribution of hourly wage. *Please see Figure 2.*
- All variables' distributions were made for a better understanding of the data. *Please see Figure 3.*
- Functional forms of continuous variable, age, were explored using the lowess curve to analyze its association with the target variable. *Please see Figure 4.*

Figure 1: Variable Description

Variables Name	Description
Age	Measured in years for each individual.
Female	Takes the value of 1 if individual is female and 0 if male.
HC	Takes the value of 1 if individual graduated from High School or College and 0 otherwise.
MA	Takes the value of 1 if individual graduated with a Master's degree and 0 otherwise.
BA	Takes the value of 1 if individual graduated with a Bachelor's degree and 0 otherwise.
GOV	Takes the value of 1 if individual works in government sector job and 0 otherwise.
PP	Takes the value of 1 if individual works in private for profit sector and 0 otherwise.
PNP	Takes the value of 1 if individual works in private for non-profit sector and 0 otherwise.
White	Takes the value 1 if individual is white and 0 if he/she is black.

Figure 2: Distribution of Earnings Per Hour

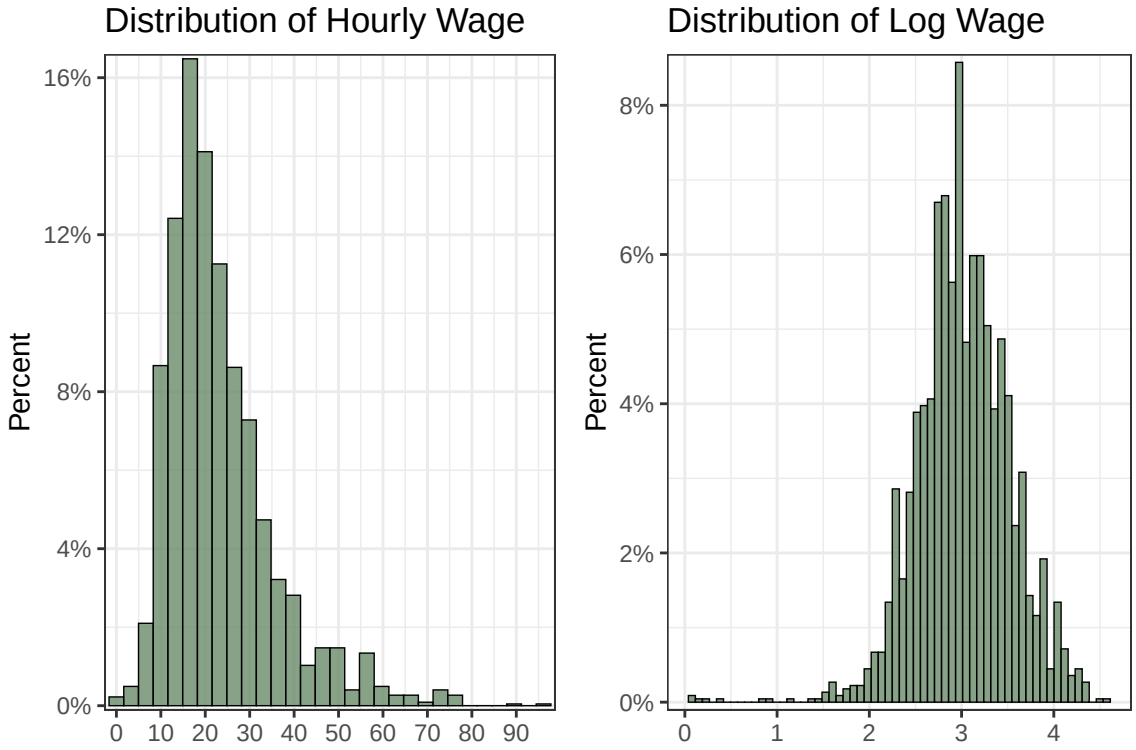


Figure 3: Distribution of Variables

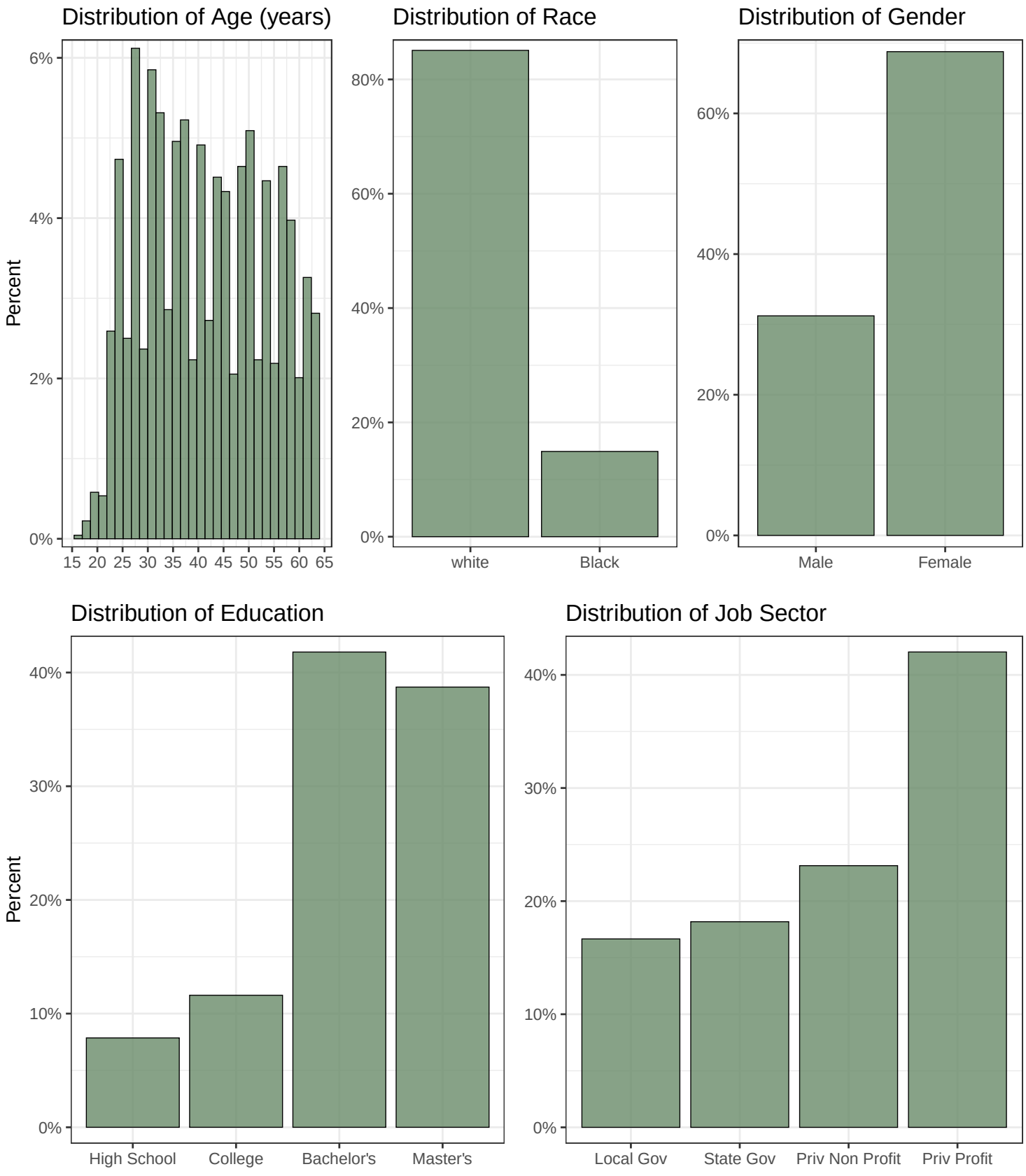


Figure 4: Association of age and its functional forms with hourly wage

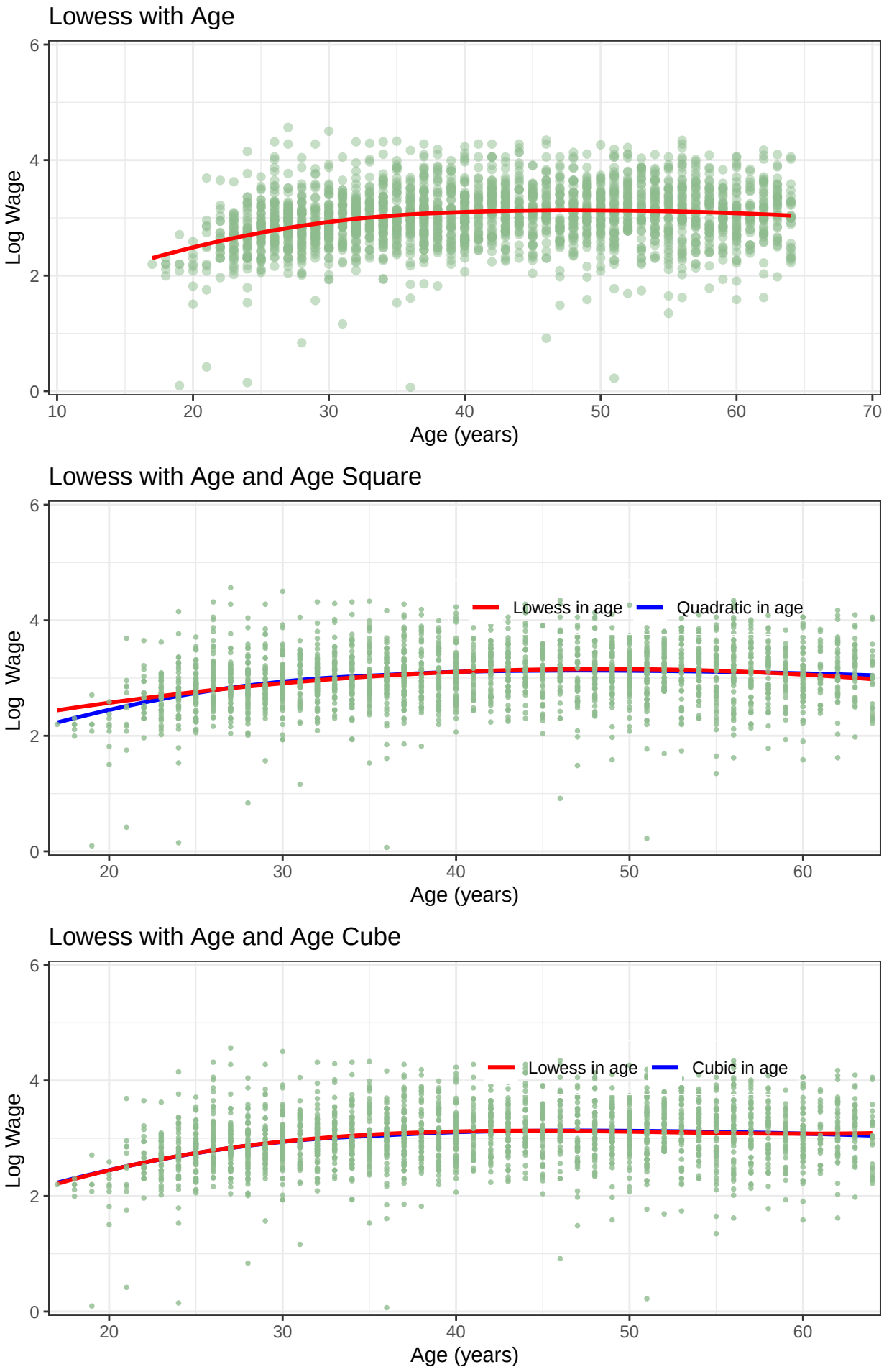


Figure 5: Quantitative Statistics of variables

	Mean	Median	Min	Max	P25	P75	N
wage	23.17	20.00	1.07	96.15	15.00	28.81	2239
age	41.65	41.00	17.00	64.00	31.00	52.00	2239
female	0.69	1.00	0.00	1.00	0.00	1.00	2239
white	0.85	1.00	0.00	1.00	1.00	1.00	2239
HC	0.19	0.00	0.00	1.00	0.00	0.00	2239
MA	0.39	0.00	0.00	1.00	0.00	1.00	2239
BA	0.42	0.00	0.00	1.00	0.00	1.00	2239
GOV	0.18	0.00	0.00	1.00	0.00	0.00	2239
PP	0.23	0.00	0.00	1.00	0.00	0.00	2239
PNP	0.42	0.00	0.00	1.00	0.00	1.00	2239

Figure 6: RMSE AND BIC for full sample

	reg1	reg2	reg3
Dependent Var.:	lnw	lnw	lnw
(Intercept)	2.972*** (0.0268)	1.404*** (0.1421)	1.454*** (0.1366)
white	0.0557. (0.0292)	0.0582* (0.0278)	0.0226 (0.0271)
age		0.0704*** (0.0069)	0.0573*** (0.0066)
agesq		-0.0007*** (8.1e-5)	-0.0006*** (7.74e-5)
female			-0.0188 (0.0223)
BA			0.2588*** (0.0293)
MA			0.3968*** (0.0287)
agecu			
PP			
PNP			
white x MA			
white x BA			
S.E. type	Heteroskeda.-rob.	Heteroskedast.-rob.	Heteroskedasti.-rob.
AIC	3,293.9	3,110.2	2,918.1
BIC	3,305.4	3,133.1	2,958.1
RMSE	0.50447	0.48376	0.46282
R2	0.00154	0.08183	0.15959
Observations	2,239	2,239	2,239
No. Variables	1	3	6

Figure 7: K Fold Cross Validation

Resample	Model1	Model2	Model3	Model4
Fold1	0.5067086	0.4943287	0.4850297	0.4806548
Fold2	0.5088639	0.4812590	0.4578847	0.4544159
Fold3	0.5339318	0.5099337	0.4864160	0.4785671
Fold4	0.4677801	0.4514733	0.4256094	0.4259810
Average	0.5048756	0.4847246	0.4643964	0.4604371