Data Analysis 3 - Assignment 1

Xibei Chen

15 January 2022

Introduction

The aim of this assignment is to use cps-earnings dataset to build multiple predictive models using linear regression for hourly wage for **Computer and Mathematical Occupations**, compare model performance of these models in terms of (a) RMSE in the full sample, (b) BIC in the full sample and (c) cross-validated RMSE, and discuss the relationship between model complexity and performance.

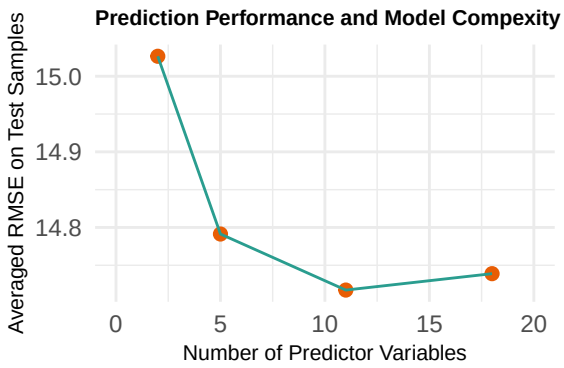
Data Cleaning, EDA and Transformation (see details in Appendix)

- Target variable: Hourly wage
- Predictor variables(cover several demographic indicators and one anthropometric indicator): Sex(male as base category), Marital status(married as base category), Race(white as base category), Highest degree(bachelor as base category), Age(squared and cubic), Weight(squared and cubic).

Build Models (using linear regression with increasing model complexity)

- Model 1: age and age squared;
- Model 2: age, age squared, sex, and highest degree;
- Model 3: age, age squared, sex, highest degree, race, and marital status;
- Model 4: age, age squared, age cubic, sex, highest degree, race, marital status, weight, weight squared, weight cubic, interaction term of sex and age, interaction term of sex and highest degree.

For the full sample, Model 4 is the best according to RMSE, while Model 2 is the best according to BIC (see model summary details in Appendix).



In terms of 5-fold Cross-validated RMSE, Model 3 is the best (see details in Appendix). And from the above visual we can see the relationship between model complexity and performance. The smaller the averaged RMSE, the better the model performance. Therefore, model performance tends to get better as number of predictor variables gets larger from the beginning. However, after a certain point, model performance starts to get worse, as number of predictor variables keeps getting larger, this is because too complex models tends to overfit the test sample hence do not give the best prediction for the population or the general pattern represented by the original data.

Appendix

Date Cleaning

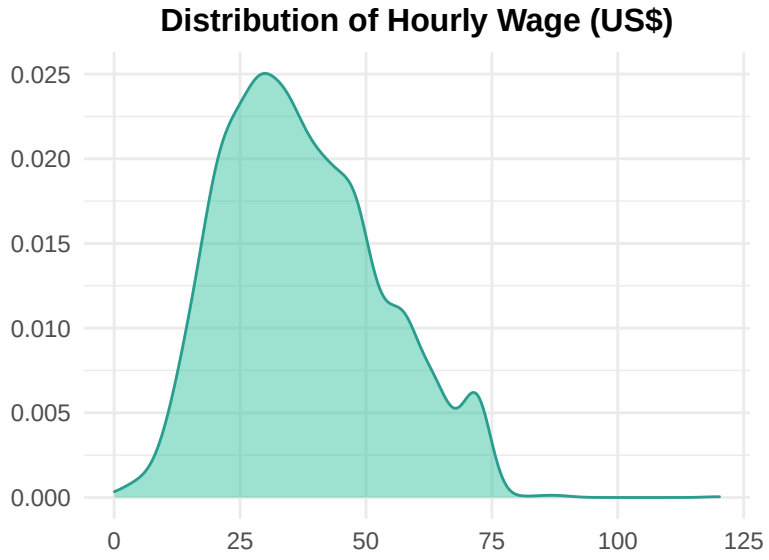
In the data cleaning process, I selected only the observations that meet the following criteria.

- Occupation type: Computer and Mathematical Occupations;
- Weekly working hours greater than 20;
- Weekly earning greater than 0;
- Age greater than 24, less than 64;
- Education categories: Bachelor's, Master's or Doctorate degree.

Explain choice of predictors

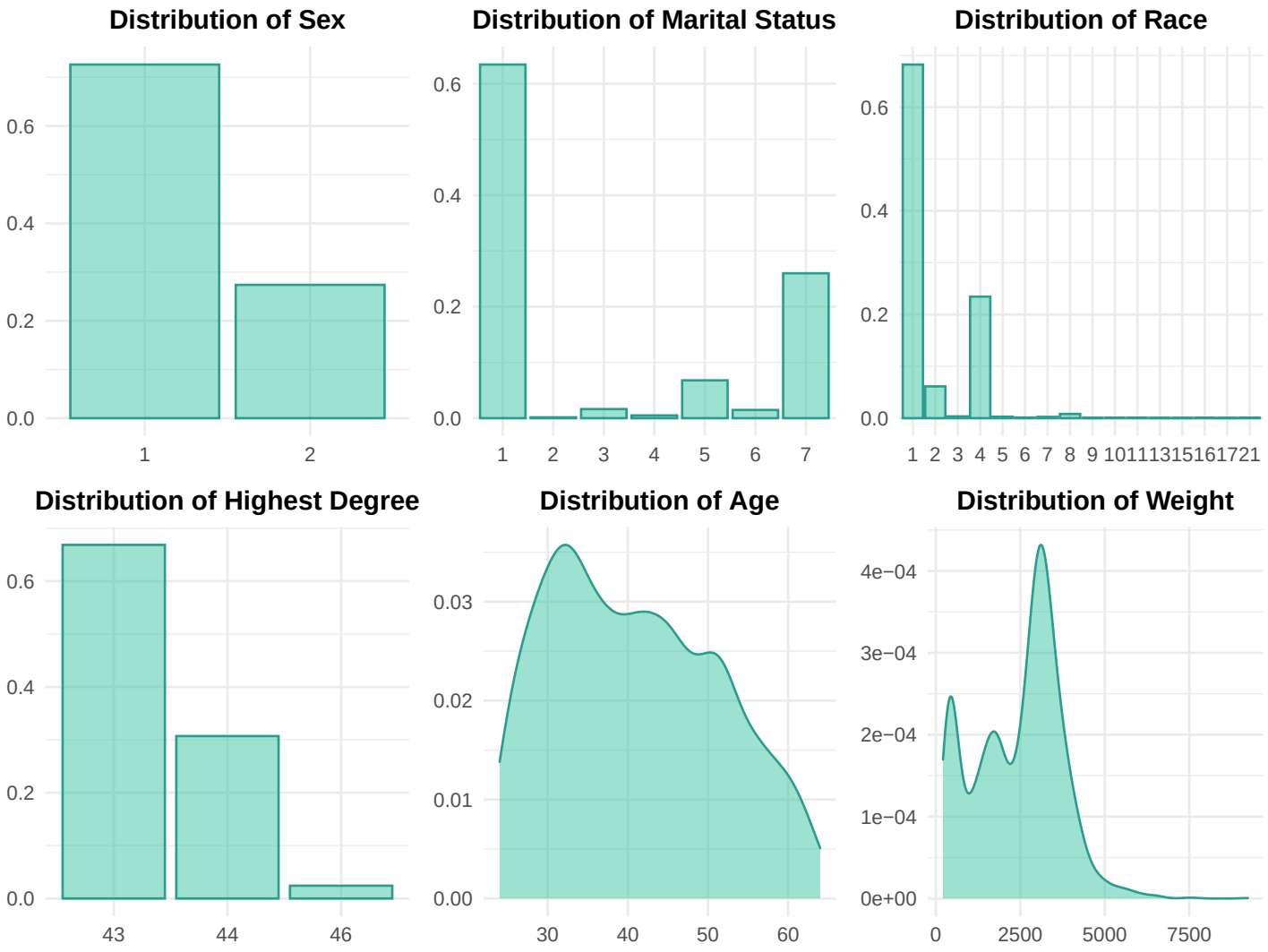
1. Computer and Mathematical Occupations are traditionally male-dominated, which is also confirmed in the distribution graph of sex, so I assume sex might have something to do with wage;
2. Marital status can to some extent explain how much a person is dedicated to his or her job, which is just my assumption;
3. Race can cover for example immigration background, the pattern of racial discrimination if there is any, or if people of certain race is more likely to do well in Computer and Mathematical Occupations;
4. Highest degree might be very closely related one's wage, as it can explain how much academical achievement someone has;
5. Age is supposed to be positively related to wage, although the association turns to be weaker when one gets older;
6. I also included weight, which might seems odd at the first glance, but I was hoping that it can catch some information about one's health, for example diet habit or how much exercise one does.

Distributions of target variable - hourly wage



From the graph we can tell that hourly wage is close to normally distributed.

Distributions of predictor variables

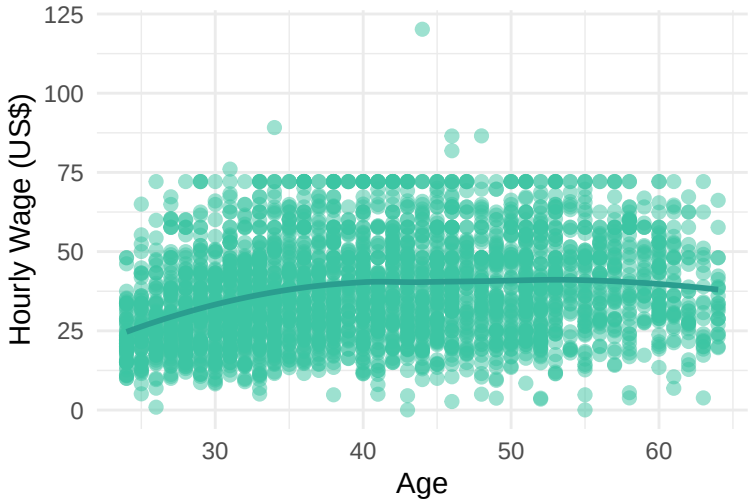


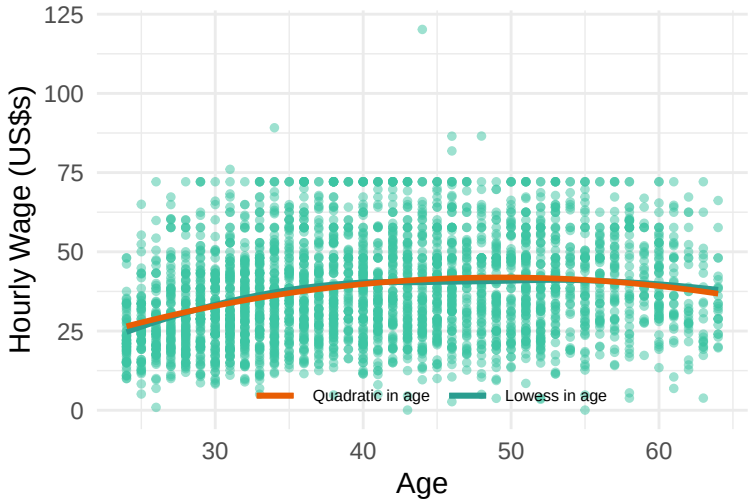
- According to the distributions, combine some categories into one: (1) Marital status: combine 1 Married civilian spouse present, 2 Married AF spouse present, 3 Married spouse absent or separated as married(m); keep 4 as widowed; combine 5 Divorced and 6 Separated as separated/divorced(sd) and keep 7 as never married(nm); (2) Race: keep 1 white, 2 black, and 4 asian, combine the rest as other.
- Decide the base category that has the most observations: (1) Sex(male as base category); (2) Marital status(married as base category); (3) Race(white as base category); (4) Highest degree(bachelor as base category).

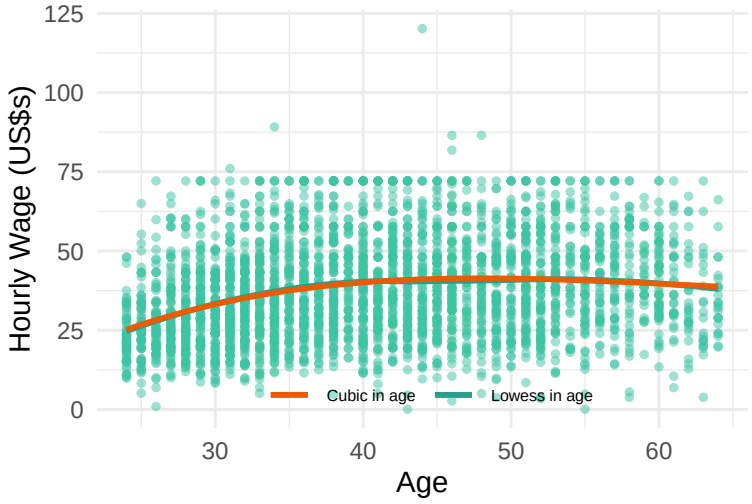
Descriptive Statistics of Quantitative Variables

	Mean	Median	Min	Max	P25	P75	N
Hourly Wage	37.61	36.00	0.04	120.19	25.48	48.08	3181
Age	40.99	40.00	24.00	64.00	32.00	49.00	3181
Weight	2456.17	2761.36	212.48	9233.90	1472.46	3316.78	3181

Association between hourly wage and lowess age/quadratic age

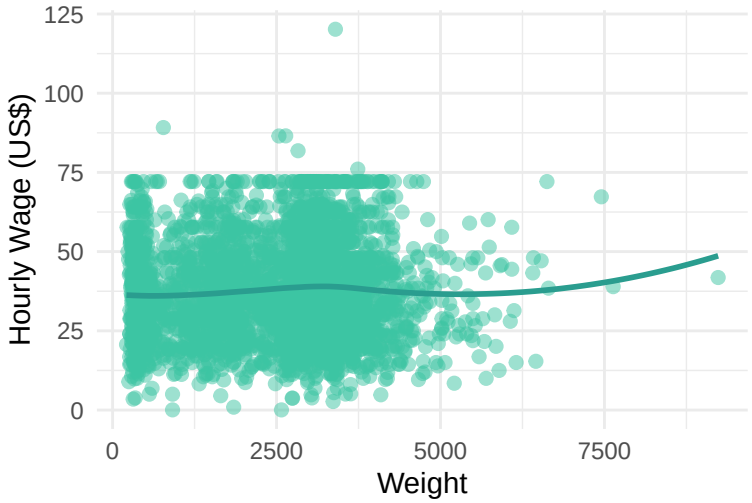


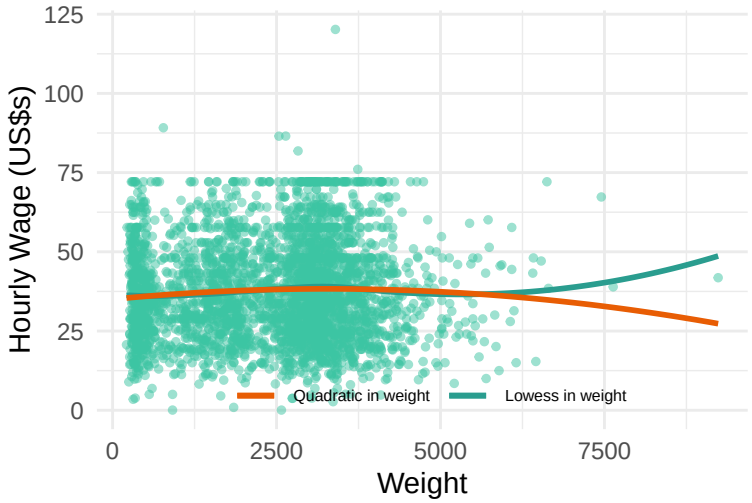


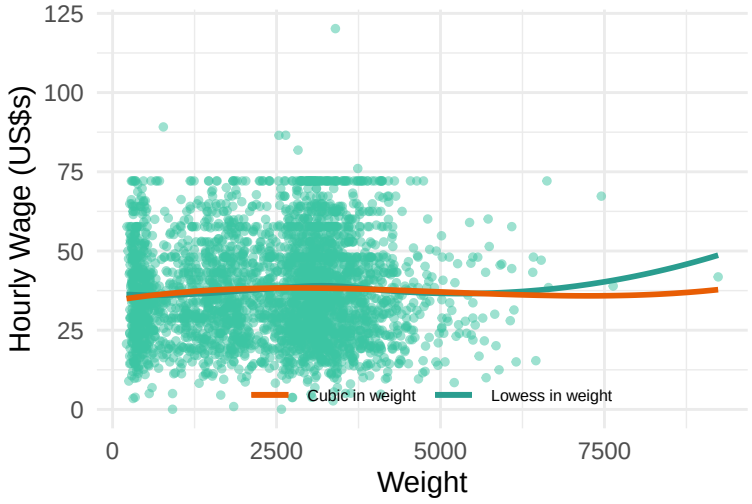


From the graphs we can see that both quadratic and cubic age work quite well.

Association between hourly wage and lowess weight/quadratic weight/cubic weight







From the graphs we can tell that cubic weight works better than quadratic weight especially for extreme values.

Compare models according to RMSE and BIC for the full sample

Table 1: Running linear regressions using all observations

	(1)	(2)	(3)	(4)
Intercept	-16.48***	-14.10***	-9.606*	-26.36
	(3.929)	(3.878)	(4.248)	(15.43)
age	2.366***	2.253***	2.047***	3.163**
	(0.1967)	(0.1945)	(0.2056)	(1.149)
age squared	-0.0240***	-0.0226***	-0.0203***	-0.0466
	(0.0023)	(0.0023)	(0.0024)	(0.0276)
female		-4.750***	-4.408***	-1.341
		(0.5889)	(0.5920)	(2.161)
master		2.888***	2.301***	2.373***
		(0.5757)	(0.5838)	(0.6958)
doctorate		9.318***	8.410***	8.228***
		(1.932)	(1.904)	(2.259)
black			-3.339**	-3.623**
			(1.217)	(1.236)
asian			2.817***	2.548***
			(0.6502)	(0.6656)
other race			-1.718	-1.981
			(1.866)	(1.860)
widowed			-3.147	-2.667
			(3.680)	(3.608)
separated/divorced			0.0694	0.1618
			(1.007)	(1.015)
never married			-1.700*	-1.528*
			(0.6677)	(0.6750)
age cubic				0.0002
				(0.0002)
weight				0.0008
				(0.0012)
weight squared				-1.13e-7
.				(4.26e-7)
weight cubic				9.99e-12
.				(4.14e-11)
age x female				-0.0697
.				(0.0515)
female x master				-0.2791
.				(1.265)
female x doctorate				0.6953
.				(4.213)
AIC	26,267.3	26,163.1	26,130.8	26,137.9
BIC	26,285.5	26,199.4	26,203.6	26,253.1
RMSE	15.013	14.755	14.653	14.637
R2	0.07624	0.10769	0.12001	0.12193
Observations	3,181	3,181	3,181	3,181
No. Variables	2	5	11	18

For the full sample, Model 4 is the best according to RMSE, while Model 2 is the best according to BIC.

Compare models in terms of 5-fold Cross-validated RMSE

Table 2: 5-fold Cross Validated RMSE

Resample	Model1	Model2	Model3	Model4
Fold1	15.19471	15.06850	14.97784	14.95459
Fold2	15.65343	15.37384	15.29071	15.28561
Fold3	14.92100	14.64900	14.61177	14.61831
Fold4	14.80463	14.52023	14.39939	14.46437
Fold5	14.53424	14.32034	14.28255	14.35169
Average	15.02642	14.79133	14.71719	14.73888

In terms of 5-fold Cross-validated RMSE, Model 3 is the best.