DA3 - Assignment 1

Ersan Kucukoglu

In this Assignment, I analyzed Financial Specialists' wages from the CPS suvery. First, as usual , I started with data cleaning. I kept only the College Non-degree, BA, MA graduated degrees, including employed-at-work people which is common in the dataset. Since the age is mostly distributed between 22 and 64, I filtered the age variable between 22 and 64. As we can see from Figure 1 that shows the Lowess vs. quadratic specification with age, the quadratic line just fit differently after 50 because quadratic tries to force curvature, so I don't think it's relevant, the loess graph is the closest fit on the data so its more reliable.Age has positive effect on wage per hour until some point. I also filtered the Financial Specialists who work as full-time employees (hours >= 40). I created the wage per hour variable by dividing weekly earnings by hours. I only focused on being employed at work because the other earnings are difficult to measure like self-earning and included those who reported 20 hours or more as their usual weekly time worked. As a result, I have Financial Specialists data with 2437 observations which have only College, BA, MA, and as the highest education level.

Four Regression Models

- Model 1 : wage_per_hour ~ age + agesq
- Model 2 : wage_per_hour ~ age + agesq + male + female
- Model 3 : wage_per_hour ~ age + agesq + male + female + College + BA+ MA
- Model 4 : wage_per_hour ~ age + agesq + male + female + College + BA+ MA + female* College + female* BA + female*MA +Private + Government

Using the data, four prediction models were built which can be seen above. I started by adding age and gender and then gradually added more variables. Table 1 shows the regression coefficients of the four regression models. To find the best model, i first evaluated the BIC values along with R-squared and the RMSE in Table 1. Second, by using k-fold cross validation I set k=4, it means splitting the data into four in a random fashion to define the four test sets. According to the both approaches, Model 3 and Model 4 have the best prediction properties. They have the lowest BIC (19,833.6 and 19,854.7), and also they have the lowest average cross- validated RMSE values(14.054 and 14.097), in the Table 2. In addition to the Table 1, we can see the number of the variables and averaged RMSE on the test samples from the Prediction Performance and model complexity graph. Model performance is better as number of predictor variables is larger from the beginning. However, after a certain point (6), model performance is worse, as number of predictor variables is getting larger. According to performance measures, the actual difference between two models is very small. Choosing a simple model can be valuable as it may help us avoid overfitting the live data. Since the Model 3 and Model 4 have RMSE values that are very close, it make sense to choose Model 3.

0.1 Appendix

Table 1: Regression Models for predicting earning per hour

	reg1	reg2	reg3	reg4
Dependent Var.:	wage_per_hour	wage_per_hour	wage_per_hour	wage_per_hour
(Intercept)	-9.699* (3.850)	-15.84*** (3.742)	-10.59** (3.785)	-11.26** (3.922)
age	1.754*** (0.2024)	1.880*** (0.1948)	1.801*** (0.1927)	1.805*** (0.1933
agesq	-0.0168*** (0.0025)	-0.0181*** (0.0024)	-0.0169*** (0.0023)	-0.0170*** (0.00
male		7.091*** (0.5875)	6.044*** (0.5892)	6.824*** (1.329)
College			-10.22*** (0.9760)	-12.46*** (1.693
BA			-3.491*** (0.7519)	-3.791*** (1.022
Private				0.2075 (0.8379)
female x College				3.266 (2.122)
female x BA				0.6946 (1.510)
S.E. type	Heteroskedast.-rob.	Heteroskedast.-rob.	Heteroskedast.-rob.	Heteroskedast.-r
AIC	20,031.9	19,890.2	19,798.8	19,802.6
BIC	20,049.3	19,913.4	19,833.6	19,854.7
RMSE	14.728	14.301	14.023	14.017
R2	0.08064	0.13328	0.16656	0.16732
Observations	2,437	2,437	2,437	2,437
No. Variables	2	3	5	8

Table 2: 4-fold cross-validation and RMSE

Resample	Model1	Model2	Model3	Model4
Fold1	14.48274	13.93398	13.57886	13.56791
Fold2	14.48202	14.33291	14.02649	14.12810
Fold3	15.01655	14.46690	14.28230	14.27980
Fold4	15.01504	14.61914	14.39369	14.40076
Average	14.75150	14.34049	14.07383	14.09775

Figure 1. Earning per hour vs. age

Lowess vs. quadratic specification with age

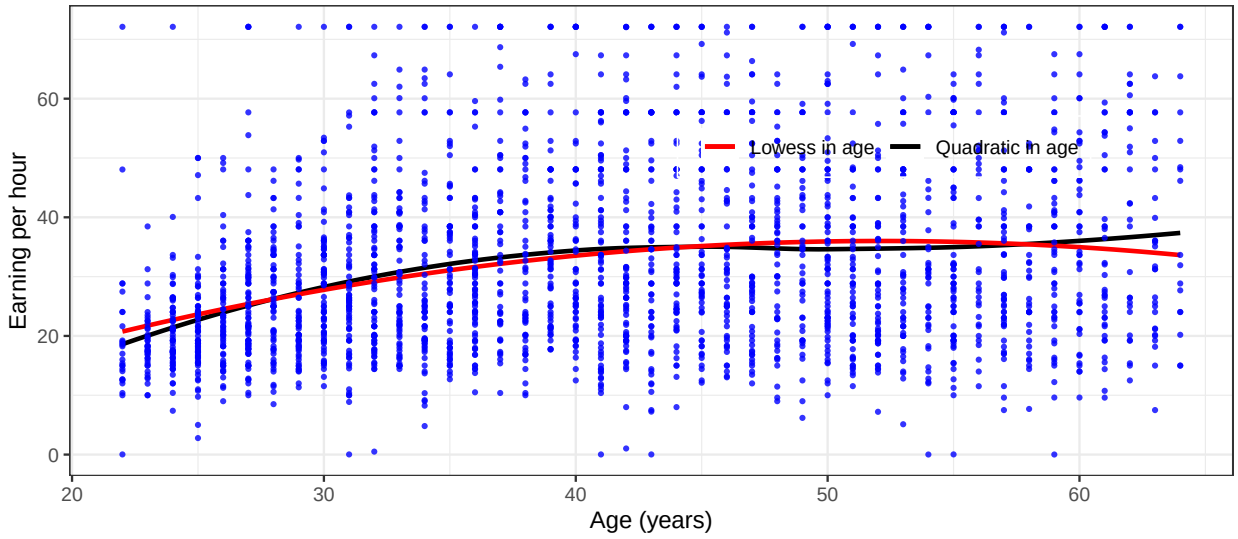

Figure 2. Prediction performance and model compexity