Data Analysis3 Assignment-1 Jo Kudo

▶ Code

Introduction

In this report, four predictive models are built with linear regression for earnings per hour, especially for Personal Care and Service Occupations. The four models are compared in three different ways; (1) RMSE in the full sample, (2) crossvalidated RMSE and (3) BIC in the full sample. The result is that RMSE in the full sample monotonically decreased, while crossvalidated RMSE and BIC in the full sample showed a U curve by comlexitiy.

Making Models

4 models are made with following variables.

- model1: This is the simple model with the assumption that age is the most effective element for earnings.

$$Earnings = \beta_0 + \beta_1 Age + \beta_2 (Age)^2$$

- model2: As well as age, grade can be the next powerful variables.

$$Earnings = \beta_0 + \beta_1 Age + \beta_2 (Age)^2 + \beta_3 Grade + \beta_4 (Grade)^2$$

- model3: Gender is also important, so it should be added to model3

$$Earnings = \beta_0 + \beta_1 Age + \beta_2 (Age)^2 + \beta_3 Grade + \beta_4 (Grade)^2 + \beta_5 Female$$

- model4: This is the most complex model of 4 models, including any variables such as the length of working, the number of children, and others times age.

$$Earnings = \beta_0 + \beta_1 Age + \beta_2 (Age)^2 + \beta_3 Grade + \beta_4 (Grade)^2 + \beta_5 Female$$
$$+\beta_6 WorkingHours + \beta_7 Children + \beta_8 (Grade)(Age) + \beta_9 (Female)(Age)$$
$$+\beta_10 (WorkingHours)(Age) + \beta_11 (Children)(Age)$$

(1)Comparing 4 models in RMSE in the full sample

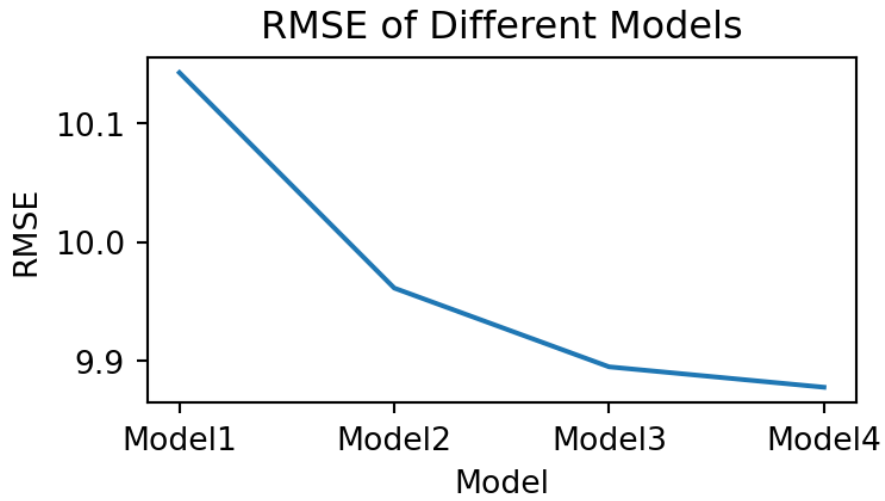
According to the figure below, complexity makes less rmse.

▶ Code

```
      Model1    Model2    Model3    Model4
0  10.142604  9.961411  9.895169  9.878004
```

RMSE in the full sample in 4 models

▶ Code



(2)Comparing 4 models in liner regression with k-fold cross validation

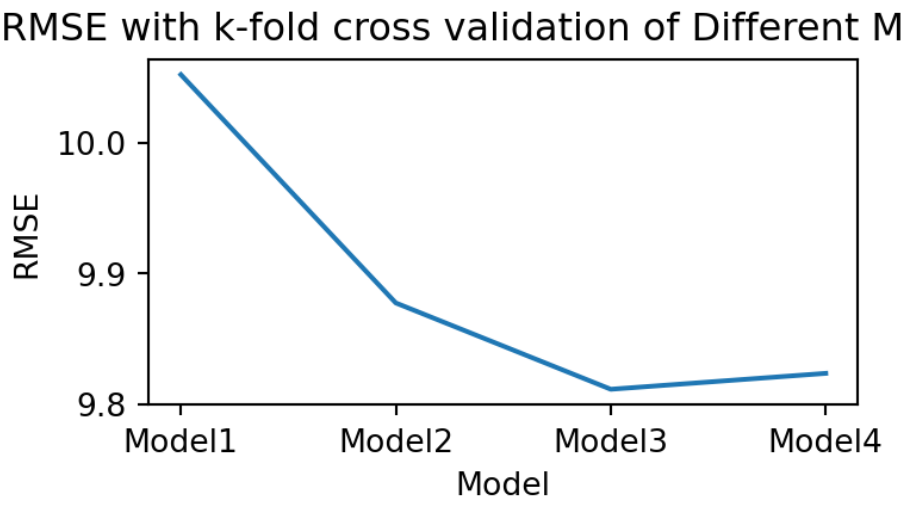
This is the direct approach to find the best model. In k-fold cross validation, sample is randomly divided into test set and train set. This time, k is defined as 4. Model 3 shows the lowest figure, which means model3 can be the best for prediction without overfitting.

▶ Code

	Model1	Model2	Model3	Model4
Fold1	10.010594	9.741763	9.689723	9.701351
Fold2	12.388202	12.252362	12.207471	12.203284
Fold3	8.699050	8.637001	8.497963	8.516327
Fold4	9.109672	8.878014	8.850003	8.872851
Average	10.051879	9.877285	9.811290	9.823453

RMSE with k-fold cross validation in 4 models

▶ Code



(3)Comparing 4 models in Bayesian Information Criterion(BIC)

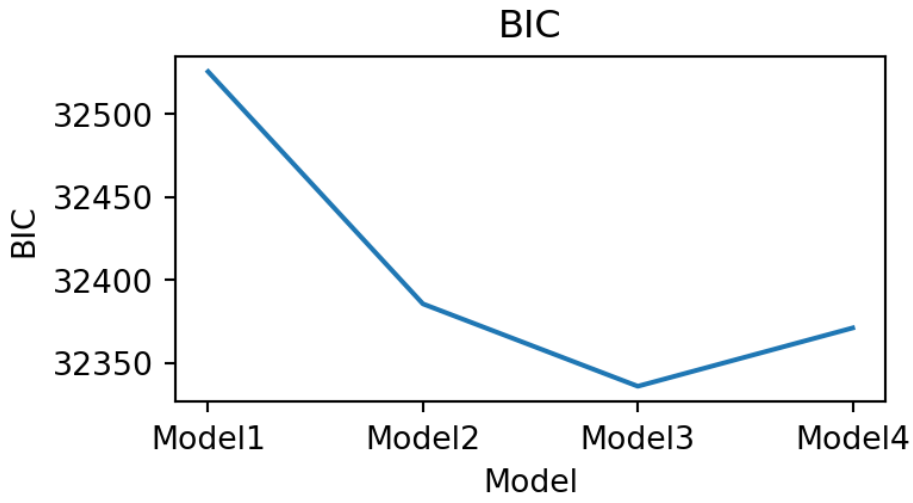
BIC is one of good approaches to find indirectly the best model by test fit and penalty. The result is the same as that in RMSE with cross validation; BIC of model 3 is the lowest, which means that model 3 can be the best model minimizing error in live data. Another point to is that in BIC and k-fold cross validation, the result from 4 models made a U curve by complexity.

▶ Code

```
      Model1    Model2    Model3   Model4
0  32525.58  32385.51  32335.84  32371.0
```

BIC of 4 models

▶ Code



Conclusion

In this analysis, four linear regression models for estimating hourly earnings in Personal Care and Service Occupations were compared using RMSE on the full sample, cross-validated RMSE, and BIC. Results showed that while full-sample RMSE decreased with complexity, cross-validated RMSE and BIC displayed a U-shaped pattern, indicating a trade-off between model complexity and predictive accuracy.

Full-sample RMSE decreased with more complex models. However, cross-validation and BIC analyses identified Model 3 as optimal, balancing accuracy and overfitting by its lowest BIC and cross-validated RMSE.