Analysis and Recommendations for Lawyer Earnings Data

Student: Thi Mai Pham
GitHub link: https://github.com/thimaipham/CEU_DA3_Assignment1

1. Introduction

Start from the original data (Cross section. N=149 316 individuals), I chose ***"Lawyers, Judges, magistrates, and other judicial workers"*** occupation (occupation code 2100). The chosen data includes 1027 observations.

This report details an analytical journey through a dataset of lawyer earnings. It delved into data preprocessing, feature engineering, and the development of linear regression models. The objective is to accurately predict lawyers' hourly earnings, balancing model performance and complexity.

2. Data Feature Engineering

I started by refining the dataset:

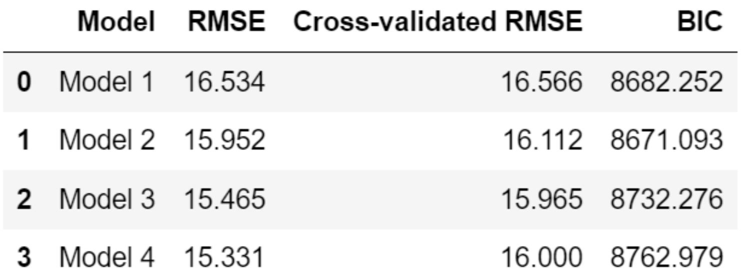
- Removal of superfluous columns such as 'ethnic', 'unioncov', 'intmonth', etc.
- Addressed missing values to enhance data integrity.
- Created 'earnings_per_hour' as our target variable.
- Applied One-Hot Encoding to categorical variables ('race', 'sex', 'marital', etc.).
- Filtered the dataset to include instances where 'earnings_per_hour' was below 100.

3. Models and Explanation for Each Model

Four linear regression models were constructed:

- **Model 1:** Solely based on 'age'.
- **Model 2:** Incorporated 'age' and specific 'grade92' categories.
- **Model 3:** Expanded on Model 2 by adding 'sex_2', 'prcitshp', 'class', and 'race' categories.
- **Model 4:** Further included 'ownchild' and 'marital' status categories.

  The results show in the below table:



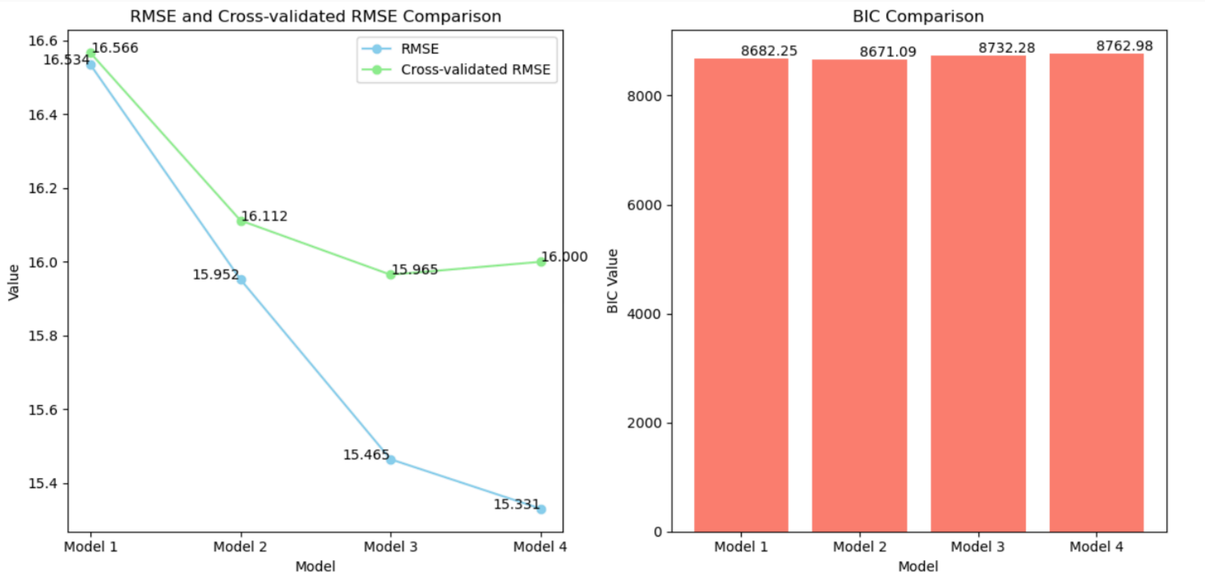
	Model	RMSE	Cross-validated RMSE	BIC
0	Model 1	16.534	16.566	8682.252
1	Model 2	15.952	16.112	8671.093
2	Model 3	15.465	15.965	8732.276
3	Model 4	15.331	16.000	8762.979

4. Discussion

4.1. Model Performance Comparison:

- **RMSE (Full sample):** Model 4 demonstrated the lowest RMSE (15.331), indicating the best fit on the training data.
- **Cross-Validated RMSE:** Contrary to initial expectations, Model 3 (15.965) exhibited a slightly better cross-validated RMSE than Model 2 (16.112), suggesting a marginally superior generalization capability.
- **BIC (Full sample):** Despite this, Model 2 (867.093) achieved the lowest BIC, indicating an efficient balance between model complexity and fit.

4.2 Graphical Observations:



- The graphs for RMSE and BIC show a trend where Model 4, being the most complex, has the lowest RMSE on the training data but does not generalize as well as Models 2 or 3.
- It's evident that Model 3 has the lowest RMSE, but its cross-validated RMSE is slightly higher than its RMSE, which might indicate a bit of overfitting.

4.3. Relationship Between Model Complexity and Performance:

- Higher complexity models, like Model 4, did not necessarily translate to better generalization, as seen in the cross-validation results.
- Model 3, despite its slightly better cross-validated RMSE compared to Model 2, has a higher BIC, which could indicate overfitting due to its complexity.
- Model 2, with its lower complexity compared to Model 3, provides a more balanced approach. It achieves a slightly higher cross-validated RMSE but has the lowest BIC, suggesting it is the most efficient model in balancing fit and complexity.

5. Conclusion/Recommendation

Given these insights, while Model 3 shows marginally better performance in cross-validation, Model 2 is still preferable due to its optimal balance of complexity, performance on unseen data, and overall model efficiency as indicated by its lower BIC. This makes Model 2 a more practical choice for predicting lawyers' earnings per hour. Model 2 is recommended for predicting lawyers' earnings per hour.