Omar
Prediction with Machine Learning for Economists
Assignment 1
Report

For this Assignment, I have choosen Financial Analysts (Occupation Code: 0840) as my occupation. As the very first phase of cleaning the data, I calculate the hourly earnings and then remove missing and infinite values to ensure data quality. Additional variables such as log-transformed earnings and age squared are created for better regression modelling. Lastly, categorical variables like sex, race, and employment type are converted to the appropriate data type for analysis.

Model 1 includes education (grade92) and age (age), as these are fundamental determinants of wages. Model 2 introduces a squared age term (age_squared) to account for potential non-linear effects and adds sex (sex) to explore gender differences. Model 3 incorporates employment category (employment_category), recognizing that different job sectors may impact earnings. Model 4 is the most comprehensive, further including union membership (unionmme).

Regression Result: In the case of financial analyst, Higher education (`grade92`) consistently has a positive and significant impact on earnings across all models, confirming that higher education leads to higher wages. Age also has a significant positive effect, meaning earnings increase with age, though the squared age term suggests diminishing returns over time. Government employment (local and state) is associated with lower earnings in Models 3 and 4, indicating that financial analysts in government roles tend to earn less than those in other sectors. However, sex, employment category, and union membership are not statistically significant, suggesting no strong evidence that these factors influence earnings in this sample. Also, the models explain only a small portion of earnings variation (R^2 ~5.6%), highlighting the importance of unobserved factors such as experience, firm size, or industry-specific dynamics.

Model	R-squared	Adjusted R-squared	RMSE (Full Sample)	BIC (Full Sample)	Residual Std. Error
Model 1	0.027	0.019	0.92	692	0.925
Model 2	0.049	0.034	0.909	697	0.918
Model 3	0.054	0.023	0.907	718	0.923
Model 4	0.056	0.021	0.906	723	0.924

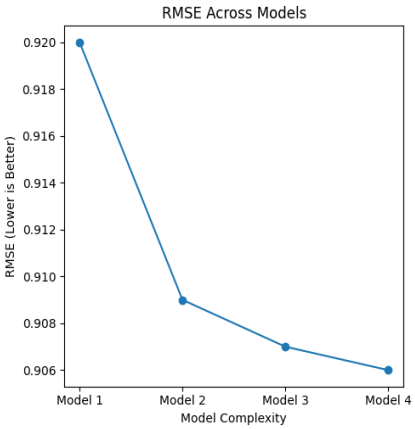
The RMSE values remain almost the same across models, meaning that adding more variables does not significantly improve prediction accuracy. Meanwhile, BIC increases with model complexity, suggesting that more complex models may be overfitting without providing much additional explanatory power. Since the best model balances accuracy and simplicity, Model 1 or Model 2 might be preferable because they have lower BIC and similar RMSE compared to more complex models.

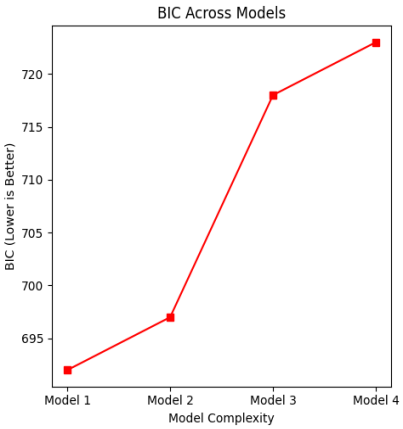
Cross-Validation RMSE Table

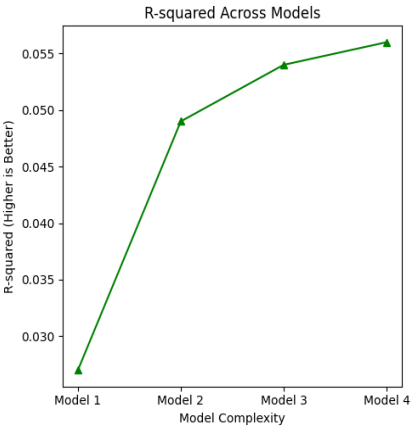
Fold	Model 1	Model 2	Model 3	Model 4
Fold 1	1.11	1.168	1.169	1.169
Fold 2	0.669	0.666	0.705	0.701
Fold 3	1.098	1.087	1.097	1.096
Fold 4	0.416	0.39	0.396	0.398
Fold 5	1.132	1.11	1.114	1.12
Average	0.885	0.884	0.896	0.897

The average cross-validation RMSE values are very close across all models, meaning adding more variables does not significantly improve predictive accuracy. Model 1 and Model 2 have the lowest average RMSE, suggesting they generalize slightly better than the more complex models. Model 3 and Model 4 have slightly higher RMSE, indicating that the additional variables do not necessarily lead to better predictions. Since, simplicity is preferable when performance is similar, Model 1 or Model 2 might be the best choice.

Also, to validate the arguments, as advised in the question, I have also drafted some plots that support my arguments. They are as such:


RMSE Across Models
RMSE (Lower is Better)
0.920
0.918
0.916
0.914
0.912
0.910
0.908
0.906
Model 1
Model 2
Model 3
Model 4
Model Complexity


BIC Across Models
BIC (Lower is Better)
720
715
710
705
700
695
Model 1
Model 2
Model 3
Model 4
Model Complexity


R-squared Across Models
R-squared (Higher is Better)
0.055
0.050
0.045
0.040
0.035
0.030
Model 1
Model 2
Model 3
Model 4
Model Complexity