Introduction

This analysis examines the influences of demographic and socioeconomic factors on variations in hourly earnings. With hourly earnings positioned as the dependent variable, this analysis includes six predictors: age, sex, race, educational attainment, native status, and government sector employment.

The approach involves constructing a series of statistical models that incrementally integrate these predictors to understand their combined effect on earnings. The first model employs simple linear regression, focusing solely on age as a determinant, suggesting that earnings could correlate with experience. Additional models progressively include sex, race (specifically White individuals), education levels (Bachelor to Doctorate), native status, and government employment, to discern their respective impacts on wage levels.

Model accuracy was evaluated using Root Mean Squared Error (RMSE) to see how well they predict earnings within the dataset and through cross-validation, which tests the models on different parts of the data to make sure they work well in general; ensuring that our models' predictive power holds when applied to new subsets of data. The Bayesian Information Criterion (BIC) was also applied to balance model simplicity and fit, making sure we're not making our models too complicated without good reason.

Outputs from this analysis shed light on the factors that shape hourly earnings and establish a basis for economic analysis. The study dissects the relationships between various predictors and earnings, aiming to highlight the economic dynamics involved.

Comparing Model Performance

(a) RMSE in the Full Sample

Root Mean Squared Error (RMSE) is a measure of the difference between the predicted values from a model and the actual values. Lower RMSE values indicate a better fit.

From the RMSE outputs:

- Model 1: 19.4837
- Model 2: 19.3482
- Model 3: 18.9603
- Model 4: 18.9585

Observation: As we move from Model 1 to Model 4, the RMSE decreases slightly, suggesting that the models are improving in predicting the actual wage values.

(b) Cross-validated RMSE

Cross-validated RMSE provides a more robust estimate of model performance on unseen data.

From the cross-validation table:

- Model 1: Average 19.3712
- Model 2: Average 19.2316
- Model 3: Average 18.8369
- Model 4: Average 18.8346

Observation: The cross-validated RMSE also shows a decreasing trend as the model complexity increases, similar to the full-sample RMSE. This indicates that additional variables may be providing useful information for predicting wages.

(c) BIC in the Full Sample

The Bayesian Information Criterion (BIC) is a criterion for model selection among a finite set of models; the model with the lowest BIC is preferred. It is based not only on the goodness of fit but also includes a penalty term for the number of parameters in the model, to avoid overfitting.

From the regression output:

- Model 1: 87622.06
- Model 2: 87491.95
- Model 3: 87106.02
- Model 4: 87122.59

Observation: Model 3 has the lowest BIC, suggesting that it is the preferred model when considering both fit and complexity. Model 4 has a slightly higher BIC, which may be due to adding more variables without sufficient improvement in the fit to justify the increased complexity.

Relationship Between Model Complexity and Performance

The relationship between model complexity and performance is evident in the results. Adding more predictors (increasing complexity) has led to a decrease in RMSE, both in the full sample and cross-validated results, up to Model 3. This suggests that additional variables are capturing more information about the variability in wages, thus improving the model's predictions.

However, there is a point where adding more complexity does not necessarily lead to better model performance. Model 4 has a very similar RMSE to Model 3, but a slightly higher BIC, indicating that the additional complexity may not be justifiable as it doesn't provide a better fit.

In summary, it seems that Model 3 strikes a good balance between complexity and predictive performance. It has the lowest BIC and a lower RMSE compared to simpler models. Model 4

does not seem to offer a significant improvement in RMSE to warrant its additional complexity. This illustrates the principle of parsimony in model selection: choose the simplest model that adequately explains the data.