## Comparing Regression Models for Predicting Earnings Per Hour

This report evaluates the performance of four linear regression models predicting hourly earnings (***w***). The models progressively increase in complexity, adding predictors such as gender, education, work hours, and interaction terms. Model performance is compared using Root Mean Squared Error (RMSE), Cross-Validated RMSE, and Bayesian Information Criterion (BIC).

**Model 1**: $w = \beta_0 + \beta_1 Age + \varepsilon$

**Model 2**: $w = \beta_0 + \beta_1 Age + \beta_2 female + \varepsilon$

**Model 3**: $w = \beta_0 + \beta_1 Age + \beta_2 female + \beta_3 grade92 + \beta_4 uhours + \varepsilon$

**Model 4**: $w = \beta_0 + \beta_1 Age + \beta_2 female + \beta_3 grade92 + \beta_4 uhours + \beta_5 age^2 + \beta_6 female * age + \varepsilon$

Model 1 is the baseline model, earnings may increase with age due to experience. It is not enough to explain the earning variation. Model 2 will help to check if there is a gender-based pay gap. Education and work hours are key factors in determining earnings that can be interpreted through Model 3. Due to this, $R^2$ improves significantly. Model 4 accounts for non-linear effects and gender differences in earnings growth. This model helps to capture real world complexity. These variables help us progressively improve the model's predictive power and understand how different factors influence earnings.

## Model Performance Comparison

***Table 1:*** Model Comparison

| | Model | RMSE | Cross_Validated_RMSE | BIC |
|---|---|---|---|---|
| 1 | Model 1 | 25.15046 | 25.15046 | 998.479 |
| 2 | Model 2 | 25.11435 | 25.11435 | 1002.838 |
| 3 | Model 3 | 24.04565 | 24.04565 | 1002.946 |
| 4 | Model 4 | 23.57975 | 23.57975 | 1008.125 |

*(Lower RMSE = Better Accuracy,*
*Lower BIC = Less Complexity)*

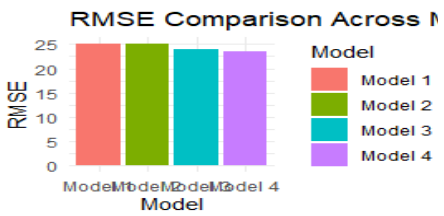


***Figure 1:*** RMSE Comparison

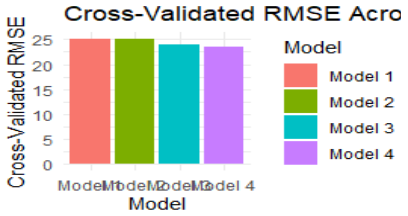


***Figure 2:*** *Cross-Validated RMSE Comparison*

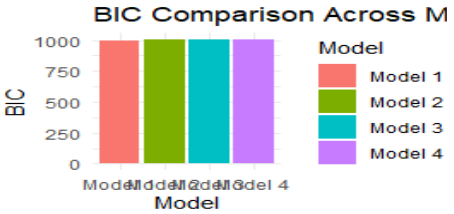


***Figure 3:*** BIC Comparison

From the Table and Figures, it is evident,

***RMSE*** and ***Cross-Validated RMSE*** show that as the model becomes more complex, prediction accuracy improves. Model 4 has the lowest RMSE, meaning it predicts earnings per hour best. However, ***BIC*** increases with complexity, suggesting that while Model 4 is the most accurate, it is also the most complex and may be overfitting. Since Model 4 introduces a squared term for age and an interaction term (female × age), it captures more variation but also risks learning noise rather than general patterns, leading to overfitting. Model 3 offers a good trade-off, reducing RMSE without a large increase in BIC. The graphs support this conclusion: RMSE decreases as we add predictors, but BIC increases, meaning complexity must be balanced. The bias-variance tradeoff is evident here, where Model 1 is too simple (high bias), while Model 4 might be overfitting (high variance). If accuracy is the goal, Model 4 is the best. Model 1 is the best for simplicity (lowest BIC). However, if we look into avoiding overfitting Model 3 is the best balance between Accuracy & Complexity. External validity is a key concern in prediction modeling. While Model 4 has the lowest RMSE, its high complexity may not generalize well to new data. External validity ensures that the model performs well on new, unseen data. A model that overfits (like Model 4) may capture noise in the training set, reducing its predictive power on future data. Model 3 balances complexity and generalizability, making it a more reliable choice for real-world applications.