****

**VIRGINIA COMMONWEALTH UNIVERSITY**

**Statistical analysis and modelling (SCMA 632)**

**A2: Regression Analysis – P2**

**Manoranjan Mohankumar**

**V01107254**

**Date of Submission: 25-06-2024**

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title** | **Page No.** |
| **1.** | Introduction | **1** |
| **2.** | Results | **2** |
| **3.** | Interpretations | **4** |
| **4.** | Recommendations | **6** |
| **5.** | Codes | **NA** |
| **6.** | References | **6** |

**Introduction**

The Indian Premier League (IPL) is one of the most lucrative and popular cricket leagues globally, attracting top players from around the world. Players' salaries in the IPL can vary significantly, influenced by various factors including their performance, marketability, and experience. Understanding the relationship between players' on-field performance and their earnings is crucial for teams to make informed decisions about player retention, recruitment, and salary negotiations.

This report aims to analyze the performance and salary data of IPL players from the 2021 to 2024 seasons. The objective is to determine how players' performance metrics, such as runs scored and wickets taken, impact their salaries. By using fuzzy matching techniques, we reconcile player names across different datasets to ensure accuracy in the analysis.

The analysis is conducted using both Python and R programming languages to ensure robustness and reproducibility of the results. Python is utilized for its powerful data manipulation capabilities and libraries like fuzzywuzzy for name matching, while R is leveraged for its strong statistical analysis packages. By employing both languages, we can cross-verify the results and provide a comprehensive understanding of the relationship between performance and salary.

**Methodology**

* Data Preparation
* Fuzzy Matching and Name Reconciliation
* Performance Metrics Calculation
* Regression Analysis
* Extended Regression Analysis

**Results**

Python Code Results

1. Data Preparation and Fuzzy Matching:
   1. The IPL ball-by-ball data was filtered for the seasons 2021 to 2024.
   2. Player names from the IPL data were matched with the names in the salary data using fuzzy matching, with common mismatches manually corrected.
   3. The matched names were used to merge the IPL performance data with the salary data.
2. Performance Metrics Calculation:
   1. Performance metrics were calculated for each player, including total runs scored, balls faced, and wickets taken.
   2. A Performance Index was derived for each player, combining runs scored and wickets taken with a weighting factor.
3. Regression Analysis:
   1. A regression analysis was performed to understand the relationship between the Performance Index and player salaries.
   2. For batsmen, the regression model showed a significant positive relationship between the Performance Index and salaries, with an R-squared value of 0.258.
   3. For bowlers, the regression model also indicated a significant positive relationship, although the R-squared value was lower at 0.100.
4. Extended Regression Analysis:
   1. Variance Inflation Factor (VIF) analysis indicated no significant multicollinearity issues.
   2. Log-transformed regression models were also fitted to account for potential skewness in the salary data.
   3. The log-transformed models confirmed the positive relationship between performance and salaries, although the fit was slightly weaker.

R Code Results

1. Data Preparation and Fuzzy Matching:
   1. Similar to the Python analysis, the IPL ball-by-ball data was filtered for the seasons 2021 to 2024.
   2. Fuzzy matching was used to align player names between the IPL data and the salary data, with manual corrections for common mismatches.
2. Performance Metrics Calculation:
   1. Performance metrics were calculated for each player, including total runs scored, balls faced, and wickets taken.
   2. A Performance Index was derived, similar to the Python analysis.
3. Regression Analysis:
   1. A regression analysis was conducted to explore the relationship between the Performance Index and player salaries.
   2. For batsmen, the regression model showed a significant positive relationship, with an R-squared value of 0.2195.
   3. For bowlers, the regression model indicated a significant positive relationship, with an R-squared value of 0.2196.
4. Extended Regression Analysis:
   1. The log-transformed models were also fitted, and the results indicated a consistent positive relationship between performance and salaries.
   2. The VIF analysis showed no significant multicollinearity issues.

**Interpretations**

The analysis reveals several important insights into the relationship between IPL players' performance and their salaries, derived from both the Python and R analyses. Here, we highlight the key findings and their implications.

Python Analysis Interpretations

1. Significant Positive Relationship for Batsmen:
   1. The regression analysis for batsmen showed a significant positive relationship between the Performance Index and their salaries, with an R-squared value of 0.258. This indicates that approximately 25.8% of the variation in batsmen's salaries can be explained by their performance metrics (runs scored and wickets taken).
   2. The high coefficient value in the regression model suggests that batsmen who perform better on the field are rewarded with higher salaries. This is consistent with the expectation that runs scored, being a primary measure of a batsman’s success, heavily influence salary decisions.
2. Positive but Weaker Relationship for Bowlers:
   1. For bowlers, the regression analysis revealed a positive but weaker relationship, with an R-squared value of 0.100. This suggests that only 10% of the variation in bowlers' salaries is explained by their performance metrics.
   2. The lower R-squared value indicates that while performance (runs conceded and wickets taken) influences salaries, other factors such as experience, reputation, and marketability might play a more significant role in determining bowlers' earnings.
3. Consistency in Log-Transformed Models:
   1. The log-transformed regression models confirmed the positive relationship between performance and salaries for both batsmen and bowlers, though the fit was slightly weaker.
   2. The log transformation, used to address potential skewness in salary data, did not drastically change the nature of the relationship, reaffirming the robustness of the findings.

R Analysis Interpretations

1. Reinforcement of Positive Relationship for Batsmen:
   1. Similar to the Python results, the R analysis showed a significant positive relationship between the Performance Index and batsmen's salaries, with an R-squared value of 0.2195. This reinforces the finding that better on-field performance correlates with higher earnings for batsmen.
   2. The consistency of this result across both programming environments strengthens the confidence in the robustness of the finding.
2. Similar Insights for Bowlers:
   1. The R analysis for bowlers also indicated a positive relationship, with an R-squared value of 0.2196, slightly higher than the Python result. This suggests that performance metrics are indeed a significant factor in determining bowlers' salaries, although not as strongly as for batsmen.
   2. The slightly higher R-squared value in R may be due to differences in how the data was processed or slight variations in the modeling approach, but it still points to the same overall conclusion.
3. Extended Regression and VIF Analysis:
   1. The extended regression analysis and VIF (Variance Inflation Factor) checks in R confirmed the absence of significant multicollinearity issues, indicating that the independent variables used in the models are not highly correlated.
   2. The log-transformed models in R also reaffirmed the positive relationship, consistent with the findings from Python.

Common Interpretations

1. Performance as a Key Salary Determinant:
   1. Across both analyses, performance metrics (runs scored for batsmen and wickets taken for bowlers) emerge as key determinants of player salaries in the IPL. This is expected as these metrics directly reflect a player’s contribution to the game.
2. Additional Factors for Bowlers:
   1. The weaker relationship for bowlers suggests that additional factors, possibly including leadership qualities, experience, team role, and marketability, significantly influence their salaries. Teams might value bowlers' strategic importance, versatility, and off-field contributions more than just on-field statistics.
3. Need for Comprehensive Metrics:
   1. The findings suggest that while basic performance metrics are important, a more comprehensive set of metrics might be needed to fully understand salary determinants. This could include advanced performance analytics, player fitness, and contributions in high-pressure situations.

**Recommendations**

1. Player Evaluation and Salary Negotiations:
   1. Teams should consider the Performance Index as a key metric during player evaluations and salary negotiations. Players with higher Performance Indices should be rewarded with higher salaries.
   2. For bowlers, additional metrics may need to be considered to fully capture their value, as the current model indicates a slightly weaker relationship between performance and salary.
2. Data-Driven Decision Making:
   1. Teams should continue to employ data-driven approaches in their decision-making processes, leveraging performance metrics to make informed decisions about player retention and recruitment.
   2. Regular updates to the data and continuous monitoring of player performance metrics will help in maintaining a fair and competitive salary structure.

**References**

* IPL\_Ball\_by\_Ball dataset
* IPL\_Salaries\_2024 dataset
* ChatGPT