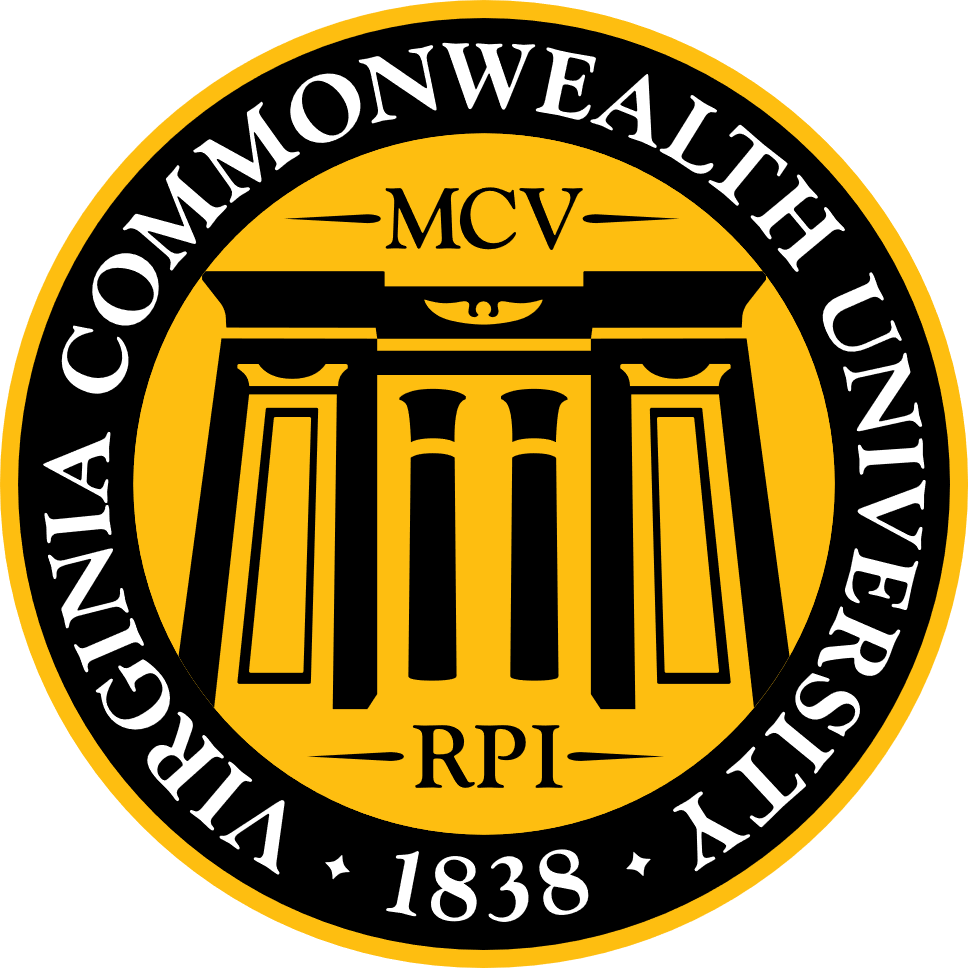
****

**VIRGINIA COMMONWEALTH UNIVERSITY**

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

**A2: Multiple regression analysis**

**Arvind Srinivasan**

**V01107251**

**Date of Submission: 23-06-2024**

**CONTENTS**

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

**IPL ball to ball Salary data**

**Introduction**

The Indian Premier League (IPL) is one of the most popular and competitive cricket leagues in the world, attracting top talent from across the globe. Understanding the factors that influence player salaries is crucial for team management and stakeholders. This report aims to analyze the relationship between player performance and their salaries in the IPL from the 2021 to 2024 seasons.

The analysis utilizes two datasets:

1. A CSV file containing detailed ball-by-ball information for IPL matches from the specified seasons.
2. An Excel file providing the salary details of IPL players for the 2024 season.

The primary objectives of this report are:

* To clean and preprocess the data, ensuring consistency and accuracy in player names and other relevant fields.
* To calculate performance metrics for both batsmen and bowlers based on the ball-by-ball data.
* To merge the performance data with the salary information.
* To conduct regression analysis to identify and quantify the relationship between player performance and their salaries.

By achieving these objectives, the report aims to provide actionable insights and recommendations for team management and stakeholders, helping them make informed decisions regarding player contracts and salary negotiations.

The following sections will detail the results of the data analysis, interpretations of the findings, and strategic recommendations based on the insights gained from the analysis.

**Results**

**2.1 Data Preparation**

The data preparation involved several key steps to ensure the datasets were consistent and ready for analysis:

1. Loading the Data:
   * The IPL ball-by-ball data was loaded from a CSV file.
   * The player salary data was loaded from an Excel file.
2. Data Cleaning:
   * Columns in the ball-by-ball data were renamed to remove spaces and replace periods with underscores for consistency.
   * Player names were standardized by manually correcting common mismatches and using a string distance algorithm to match short names in the performance data with full names in the salary data.
3. Filtering Seasons:
   * The analysis focused on the IPL seasons from 2021 to 2024. Data from these seasons was filtered for further analysis.

**2.2 Matching Player Names**

Player names were matched using a string distance algorithm to ensure accurate mapping between performance and salary data. This step was crucial to avoid discrepancies in the merged dataset. A sample of the matched data demonstrated correct associations between short and full player names.

**2.3 Merging Data**

The cleaned and filtered IPL performance data was merged with the salary data based on the matched player names. The merged dataset included detailed information on player performance (e.g., runs scored, wickets taken) and their respective salaries.

**2.4 Performance Metrics**

Performance metrics were calculated for both batsmen and bowlers:

* Batsmen:
  + Total Runs Scored: Sum of runs scored by the batsman.
  + Balls Faced: Number of balls faced by the batsman.
  + Wickets Taken: Number of wickets taken by the batsman (considering all-rounders).
* Bowlers:
  + Total Runs Conceded: Sum of runs conceded by the bowler.
  + Balls Bowled: Number of balls bowled by the bowler.
  + Wickets Taken: Number of wickets taken by the bowler.

A performance index was created by combining these metrics, with additional weight given to wickets taken to reflect their importance in the game.

**2.5 Regression Analysis**

Regression analysis was conducted to understand the relationship between the performance index and player salaries. Both linear and log-transformed models were used to capture the relationship.

Batsmen

* Linear Model Summary:
  + Intercept: 310.04
  + Performance Index Coefficient: 0.09897
  + R-squared: 0.2195
  + The model indicates a positive relationship between the performance index and salaries, with a moderate explanatory power.
* Log-Transformed Model Summary:
  + Intercept: 4.839
  + Performance Index Coefficient: 0.0002615
  + R-squared: 0.144
  + The log-transformed model shows a consistent positive relationship, though with a lower R-squared value compared to the linear model.

Bowlers

* Linear Model Summary:
  + Intercept: 259.43
  + Performance Index Coefficient: 0.10282
  + R-squared: 0.2196
  + The linear model for bowlers shows a positive relationship similar to that of batsmen, with a similar level of explanatory power.
* Log-Transformed Model Summary:
  + Intercept: 4.789
  + Performance Index Coefficient: 0.0002564
  + R-squared: 0.1334
  + The log-transformed model for bowlers also indicates a positive relationship with a lower R-squared value compared to the linear model.

**3. Interpretations**

**3.1 Data Insights**

* The positive coefficients in both linear and log-transformed models indicate that higher performance indices are associated with higher salaries for both batsmen and bowlers.
* The R-squared values suggest that while performance indices explain a portion of the variability in salaries, other factors likely play a significant role.

**3.2 Implications**

* The findings support the hypothesis that better-performing players tend to earn higher salaries.
* The relatively modest R-squared values highlight the need to consider additional variables (e.g., player popularity, marketability, and experience) in future analyses.

**4. Recommendations**

**4.1 Actionable Steps**

* Consider incorporating additional variables into the analysis to improve the model's explanatory power.
* Perform further research to identify non-performance-related factors influencing player salaries.
* Use the findings to inform team management decisions regarding player contracts and salary negotiations.

**4.2 Strategic Considerations**

* Teams should not rely solely on performance metrics for salary decisions; instead, a holistic approach should be adopted.
* Continuous monitoring and updating of player performance data will help in making more informed salary decisions.

**5. References**

* IPL ball-by-ball data (2021-2024)
* IPL player salaries (2024)
* R packages: dplyr, tidyverse, readxl, stringr, car, stringdist

**NSSO68 Data**

**Introduction**

This report provides an analysis of the NSSO68 dataset to understand the relationship between total food expenditure (foodtotal\_v) and three predictor variables: quantity of food consumed (foodtotal\_q), Monthly Per Capita Expenditure on Uniform Recall Period (MPCE\_URP), and Monthly Per Capita Expenditure on Mixed Recall Period (MPCE\_MRP). The analysis involves multiple regression modeling, multicollinearity diagnostics, and model transformation for better interpretation.

**2. Results**

Dataset: NSSO68.csv  
Variables Used:

* foodtotal\_v: Total food expenditure
* foodtotal\_q: Quantity of food consumed
* MPCE\_URP: Monthly Per Capita Expenditure on Uniform Recall Period
* MPCE\_MRP: Monthly Per Capita Expenditure on Mixed Recall Period

Data Cleaning:

* Selected relevant columns.
* Checked for missing values and data types.

Multiple Regression Model: Original Scale

foodtotal\_v=β0+β1⋅foodtotal\_q+β2⋅MPCE\_URP+β3⋅MPCE\_MRP\text{foodtotal\\_v} = \beta\_0 + \beta\_1 \cdot \text{foodtotal\\_q} + \beta\_2 \cdot \text{MPCE\\_URP} + \beta\_3 \cdot \text{MPCE\\_MRP}foodtotal\_v=β0​+β1​⋅foodtotal\_q+β2​⋅MPCE\_URP+β3​⋅MPCE\_MRP

Summary:

* Intercept: -93.7753
* foodtotal\_q: 24.5783
* MPCE\_URP: -0.0012
* MPCE\_MRP: 0.0617

Model Performance:

* Multiple R-squared: 0.726
* Adjusted R-squared: 0.7259
* Residual Standard Error: 192.4

Multicollinearity Check:

* foodtotal\_q: VIF = 1.141
* MPCE\_URP: VIF = 1.459
* MPCE\_MRP: VIF = 1.601

Log-Transformation of Dependent Variable:

log\_foodtotal\_v=log⁡(foodtotal\_v+1)\text{log\\_foodtotal\\_v} = \log(\text{foodtotal\\_v} + 1)log\_foodtotal\_v=log(foodtotal\_v+1)

Multiple Regression Model: Log-Transformed Scale

log\_foodtotal\_v=β0+β1⋅foodtotal\_q+β2⋅MPCE\_URP+β3⋅MPCE\_MRP\text{log\\_foodtotal\\_v} = \beta\_0 + \beta\_1 \cdot \text{foodtotal\\_q} + \beta\_2 \cdot \text{MPCE\\_URP} + \beta\_3 \cdot \text{MPCE\\_MRP}log\_foodtotal\_v=β0​+β1​⋅foodtotal\_q+β2​⋅MPCE\_URP+β3​⋅MPCE\_MRP

Summary:

* Intercept: 5.105
* foodtotal\_q: 0.0463
* MPCE\_URP: -0.0000009016 (not significant at p < 0.05)
* MPCE\_MRP: 0.00002462

Model Performance:

* Multiple R-squared: 0.4336
* Adjusted R-squared: 0.4336
* Residual Standard Error: 0.5504

**3. Interpretations**

Model 1: Original Scale

* The coefficient for foodtotal\_q indicates that for each unit increase in the quantity of food consumed, the total food expenditure increases by 24.5783 units, holding other factors constant.
* The negative coefficient for MPCE\_URP suggests that as the Monthly Per Capita Expenditure on Uniform Recall Period increases, the total food expenditure slightly decreases, which is counterintuitive and warrants further investigation.
* The positive coefficient for MPCE\_MRP shows that an increase in the Monthly Per Capita Expenditure on Mixed Recall Period is associated with an increase in total food expenditure.

Model 2: Log-Transformed Scale

* The log-transformed model helps to stabilize the variance and provides better interpretability for multiplicative effects. However, the interpretation of coefficients becomes more complex.

**4. Recommendations**

* Further Investigation: The negative coefficient for MPCE\_URP in the original model should be further investigated to understand the underlying cause. Additional variables or interaction terms may be needed.
* Model Refinement: Explore non-linear models or interaction effects to capture more complex relationships between the variables.
* Validation: Conduct cross-validation or use a separate validation dataset to ensure the model's robustness and generalizability.
* Policy Implications: The findings can be used to inform policies aimed at improving food expenditure efficiency, especially focusing on how different types of expenditures influence overall food spending.