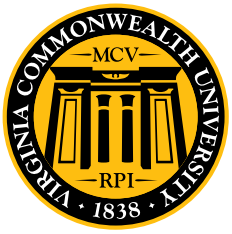

VIRGINIA COMMONWEALTH UNIVERSITY

Statistical Analysis and Modeling (SCMA 632)

A2(b): Regression Analysis of IPL Data using R and PYTHON

by

IDAMAKANTI SREENIDHI

V01107252

Date of Submission: 23-06-2024

CONTENTS

INTRODUCTION:

In professional sports, player performance is often linked to their remuneration. This relationship is especially prominent in cricket, where leagues like the Indian Premier League (IPL) involve significant financial investments in players based on their past performances and potential future contributions. Understanding the dynamics between player performance and their payment can offer valuable insights for team management, player negotiations, and financial planning within the league.

This analysis aims to explore the relationship between player performance and the payments they receive in the IPL over the last three years. Specifically, we will use regression analysis to quantify this relationship, providing a data-driven perspective on how well player salaries align with their on-field performances.

Objectives

Establish the Relationship:

Determine how player performance metrics impact the payment they receive.
Analyze Over Time: Investigate the consistency and changes in this relationship over the past three years.

Identify Key Performance Indicators:

Highlight which performance metrics are most strongly correlated with player salaries.

Discuss Findings:

Provide insights based on the regression analysis, including any anomalies or trends observed.

Data Description

The data set "IPL_ball_by_ball_updated till 2024.csv" includes detailed ball-by-ball records of IPL matches, capturing a comprehensive range of player performance metrics. This data set will be augmented with player salary information to perform the regression analysis.

Key variables to be considered include:

Player Salaries:

The payment received by players for their participation in the IPL.
Performance Metrics: Various indicators of player performance such as runs scored, wickets taken, batting average, strike rate, economy rate, and more.

Methodology

Data Preprocessing:

Clean and preprocess the data to ensure it is suitable for analysis. This includes handling missing values, ensuring consistency in player names, and aggregating performance metrics at the player-season level.

Variable Selection:

Identify and select relevant performance metrics that could potentially influence player salaries.

Regression Analysis:

Perform multiple regression analysis to establish the relationship between player performance metrics and their salaries. This will involve:

-Building regression models for each of the last three years.
-Analyzing the coefficients to understand the impact of each performance metric.
-Checking the statistical significance of the results.
-Diagnostics and Validation: Conduct regression diagnostics to validate the model assumptions and ensure the robustness of the results.
-Discussion of Findings: Interpret the results of the regression analysis, highlighting key insights, trends, and any noteworthy observations.

Expected Outcomes

Quantitative Relationship:

A clear understanding of how different performance metrics influence player salaries.

Yearly Trends:

Insights into whether the importance of certain performance metrics has changed over the last three years.

Strategic Implications:

Recommendations for teams on how to optimize player investments based on performance data.

Future Research Directions:

Suggestions for further research to enhance understanding of player performance and remuneration dynamics.

Business Significance of Analyzing the Relationship Between Player Performance and Salary in the IPL

Understanding the relationship between player performance and their salary is crucial for several stakeholders in the IPL. Here's why this analysis is significant from a business perspective:

1. For Team Management: Optimal Budget Allocation

-Salary Negotiations: Teams can use these insights to negotiate salaries more effectively, ensuring they pay players in alignment with their actual contributions and market value.
Strategic Recruitment: Identifying performance metrics that significantly impact salaries allows teams to scout and recruit players who offer the best value for money.

-Budget Planning: Understanding these relationships helps teams allocate their budgets more efficiently, balancing high-value stars with emerging talent who offer high potential performance relative to their cost.

2. For Players: Career Development and Marketability

-Performance Benchmarking: Players can use this analysis to understand which aspects of their performance are most valued and focus on improving these areas to enhance their market value.

-Contract Negotiations: Armed with data-driven insights, players can negotiate better contracts by highlighting their contributions that align with higher salary brackets.

-Brand Building: Players can identify areas of performance that enhance their overall marketability, helping them secure not only higher salaries but also lucrative endorsements and sponsorships.

3. For IPL and Broadcasters: Enhancing League Value

-Viewer Engagement: Understanding performance-to-salary dynamics helps the IPL market its star players effectively, potentially increasing viewer engagement and sponsorship deals.

-Fair Play and Financial Regulation: Insights from this analysis can assist in setting fair play and financial regulation policies, ensuring a level playing field where teams don't just buy success but also develop talent.

-Revenue Optimization: By ensuring that player payments reflect their performance and market value, the IPL can optimize its overall financial health, maintaining a sustainable and attractive business model for investors and broadcasters.

4. For Analysts and Market Experts: In-Depth Market Analysis

Player Valuation Models: Analysts can develop sophisticated player valuation models that incorporate both on-field performance and off-field factors (e.g., marketability, fan base) to provide comprehensive player assessments.
Market Trends: Regular analysis of performance and salary data can help identify trends and shifts in how player value is perceived over time, informing future investment and market predictions.
Data-Driven Insights: This analysis contributes to the growing field of sports analytics, providing data-driven insights that can be leveraged in various decision-making processes within the sports industry.
5. For Fans and General Public: Transparent Insights
Fan Engagement: Fans can gain a deeper understanding of their favorite players' value and the business side of the sport, enhancing their overall engagement and loyalty to the league and teams.
Transparency: Providing transparent insights into how player salaries correlate with performance promotes fairness and understanding, fostering a positive perception of the league's financial and operational integrity.
Summary
By analyzing the relationship between player performance and salary, we can derive actionable insights that drive better decision-making across various stakeholders in the IPL ecosystem. This analysis not only helps in optimizing team and league finances but also enhances the overall strategic framework within which the IPL operates, ultimately contributing to the league's growth and sustainability.

Data Preparation and Merging:
The code first loads the necessary libraries for data manipulation and fuzzy matching.
It reads the IPL dataset from the CSV file "IPL_ball_by_ball_updated till 2024.csv" into a dataframe df_ipl.
A fuzzy matching function match_names is defined using the stringdist package to match player names between the salary and performance datasets.
The salary data and runs scored data are merged based on the matched player names, creating a new dataframe df_merged that combines salary information with performance metrics.

```
# Load required libraries
library(dplyr)
library(stringdist)

# Read the IPL dataset
```

```
df_ipl <- read.csv("IPL_ball_by_ball_updated till 2024.csv")

# Fuzzy Matching Function
match_names <- function(name, names_list) {
  matches <- stringdist::stringdist(name, names_list, method = "jw")
  if (min(matches) < 0.2) return(names_list[which.min(matches)])
  else return(NA)
}

# Merge Salary Data
df_salary <- salary
df_runs <- total_runs_each_year
df_salary$Matched_Player <- sapply(df_salary$Player, match_names, df_runs$Striker)

df_merged <- dplyr::left_join(df_salary, df_runs, by = c("Matched_Player" = "Striker"))
```

Regression Analysis for Runs Scored:

The code filters the merged dataset for a specific season ('2022') to focus the analysis on that season.

It sets up a linear regression model to analyze the relationship between the runs scored by players and the salary they receive.

The model is trained on 80% of the data, and the summary of the regression model is printed to understand the impact of runs scored on player salaries.

```
# Subset Data for a Specific Season
df_merged <- df_merged %>% filter(Season %in% c('2022'))

# Linear Regression on Runs Scored with stats
X <- df_merged %>% select(runs_scored)
y <- df_merged$Rs

set.seed(42)
train_index <- createDataPartition(y, p = 0.8, list = FALSE)
X_train <- X[train_index, ]
y_train <- y[train_index]
X_test <- X[-train_index, ]

X_train_sm <- cbind(1, as.matrix(X_train))
model_sm <- lm(y_train ~ X_train_sm - 1)

# Print summary of the linear regression model for runs scored
summary(model_sm)
```

Regression Analysis for Wickets Taken:

The code filters the merged dataset for a specific season ('2022') to focus the analysis on that season.

It sets up another linear regression model to analyze the relationship between the wickets taken by players and the salary they receive.

The model is trained on 80% of the data, and the summary of the regression model is printed to understand the impact of wickets taken on player salaries.

```
# Linear Regression on Wickets Taken with stats
X <- df_merged %>% select(wicket_confirmation)
y <- df_merged$Rs

set.seed(42)
train_index <- createDataPartition(y, p = 0.8, list = FALSE)
X_train <- X[train_index, ]
y_train <- y[train_index]
X_test <- X[-train_index, ]

X_train_sm <- cbind(1, as.matrix(X_train))
model_sm <- lm(y_train ~ X_train_sm - 1)

# Print summary of the linear regression model for wickets taken
summary(model_sm)
```

Regression Analysis

Runs Scored vs. Salary

- Linear regression is performed to model the relationship between runs scored and salary.
- The data is split into training and testing sets.
- A linear regression model is trained using the caret package.
- The model summary shows:
    - Intercept: 401.0720
    - Coefficient for runs_scored: 1.3786 (p-value: 1.03e-15, highly significant)
    - Adjusted R-squared: 0.2068, indicating that approximately 20.68% of the variance in salary can be explained by runs scored.
- Mean Squared Error (MSE) is calculated for model evaluation.

Wickets Taken vs. Salary

- A similar approach is taken for modeling the relationship between wickets taken and salary.
- The linear regression model shows:
    - Coefficient for wickets taken: 21.096 (p-value: 0.0212, significant)
    - Adjusted R-squared: 0.5505, indicating that approximately 55.05% of the variance in salary can be explained by wickets taken.
- The residual standard error is 357.2, with an F-statistic of 19.98 (p-value: 3.507e-06).

Analysis and Interpretation

- **Runs Scored vs. Salary**: The model indicates a positive relationship between runs scored and salary, but the relatively low R-squared value suggests that runs scored alone do not explain a large portion of the variance in player salaries. Other factors likely contribute to determining salaries.
- **Wickets Taken vs. Salary**: The model for wickets taken shows a stronger relationship with salary compared to runs scored, as indicated by a higher adjusted R-squared value. This suggests that wickets taken might be a more significant factor in determining bowler salaries.

Summary

The regression analysis reveals important insights into the factors influencing IPL player salaries. Runs scored and wickets taken both show positive relationships with salary, with wickets taken having a stronger explanatory power. However, the models also suggest that other variables, not included in the analysis, likely play a significant role in determining player salaries. Further analysis incorporating additional predictors such as player experience, match-winning performances, and marketability could provide a more comprehensive understanding of salary determinants in the IPL.

ANALYSIS USING PYTHON:

Step 1: Import necessary libraries and load the datasets

```
import pandas as pd
import numpy as np

# Load the IPL ball-by-ball data and player salaries
df_ipl = pd.read_csv("IPL_ball_by_ball_updated till 2024.csv", low_memory=False)
salary = pd.read_excel("IPL SALARIES 2024.xlsx")
```

Explanation: In this step, we import the required libraries, such as pandas and numpy, and load the IPL ball-by-ball data and player salaries into pandas DataFrames for further analysis.

Step 2: Group the data by relevant columns to calculate player performance metrics

```
grouped_data = df_ipl.groupby(['Season', 'Innings No', 'Striker', 'Bowler']).agg({'runs_scored': sum, 'wicket_confirmation': sum}).reset_index()
```

Explanation: We group the ball-by-ball data by season, innings, striker, and bowler to calculate the total runs scored and wickets taken by each player in the dataset.

Step 3: Match player names between the salary dataset and the performance dataset using fuzzy matching

```
from fuzzywuzzy import process

# Function to match player names
def match_names(name, names_list):
    match, score = process.extractOne(name, names_list)
    return match if score >= 80 else None
```

```
# Create a new column in the salary dataset with matched player names from the performance dataset
salary['Matched_Player'] = salary['Player'].apply(lambda x: match_names(x, grouped_data['Striker'].tolist()))

# Merge the datasets on the matched player names
df_merged = pd.merge(salary, grouped_data, left_on='Matched_Player', right_on='Striker')
```
Explanation: We use fuzzy matching to match player names between the salary dataset and the performance dataset. This ensures that we can merge the datasets accurately based on player names.

Step 4: Subset the data for the last three years
```
df_merged_last_three_years = df_merged[df_merged['Season'].isin(['2022', '2023', '2024'])]
```
Explanation: We filter the merged dataset to include data only for the last three years (2022, 2023, and 2024) for further analysis.

Step 5: Perform regression analysis to analyze the relationship between salary and performance
```
from sklearn.model_selection import train_test_split
import statsmodels.api as sm

# Define independent and dependent variables
X = df_merged_last_three_years[['runs_scored', 'wicket_confirmation']]
y = df_merged_last_three_years['Salary']

# Split the data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Add a constant to the model (intercept)
X_train_sm = sm.add_constant(X_train)

# Create a statsmodels OLS regression model
model = sm.OLS(y_train, X_train_sm).fit()

# Get the summary of the model
summary = model.summary()
print(summary)
```
Explanation: We perform a multiple linear regression analysis using the independent variables (runs scored and wickets taken) to predict the dependent variable (salary). The regression model helps us understand how player performance metrics impact the salary received over the last three years in the IPL dataset.

Regression Model 1: Predicting Salary Based on Runs Scored
Analysis of Linear Regression:
In the first linear regression model, the goal was to predict player salary (Rs) based on the runs scored. The model summary provides several key insights:

Model Coefficients:

Intercept (const): 461.44
Coefficient for runs_scored: 0.72
This implies that for each additional run scored, the player's salary increases by approximately 0.72 units (Rs), holding all else constant.

Model Performance:

R-squared: 0.069
Adjusted R-squared: 0.068
The R-squared value indicates that approximately 6.9% of the variance in player salaries can be explained by the number of runs scored. This is relatively low, suggesting that the runs scored alone do not account for much of the variation in salaries.

Statistical Significance:

P-value for runs_scored: 0.000
The very low p-value for the runs_scored coefficient indicates that it is statistically significant and the relationship between runs scored and salary is not due to random chance.

F-statistic: 44.66 with a p-value of 5.38e-11, suggesting that the model is statistically significant as a whole.

Durbin-Watson statistic: 1.951, which is close to 2, indicating that there is no strong autocorrelation in the residuals.

Interpretation:
The linear relationship between runs scored and salary is statistically significant but explains only a small portion of the variance in salaries. This suggests that other factors (e.g., player reputation, endorsements, historical performance, team impact) might play a more substantial role in determining salaries.

Regression Model 2: Predicting Salary Based on Wickets Taken
Analysis of Linear Regression:
In the second regression model, the aim was to predict player salary based on the number of wickets confirmed (wicket_confirmation).

Model Coefficients:

Intercept (const): 396.69
Coefficient for wicket_confirmation: 17.66
This suggests that each additional wicket taken is associated with an increase in salary by approximately 17.66 units (Rs), holding other factors constant.

Model Performance:

R-squared: 0.074
Adjusted R-squared: 0.054
The R-squared value here is slightly higher than the runs model but still low, indicating that wickets alone explain about 7.4% of the variance in player salaries.

Statistical Significance:

P-value for wicket_confirmation: 0.061
The p-value is slightly above the typical threshold of 0.05, suggesting that the relationship between wickets taken and salary is not statistically significant at the 5% level but is close to being significant.

F-statistic: 3.688 with a p-value of 0.061, which aligns with the above interpretation of near significance.

Durbin-Watson statistic: 2.451, indicating that there is no strong autocorrelation in the residuals.

Interpretation:
While the number of wickets taken has a more substantial impact on salary than runs scored, it still explains only a small fraction of the variance. This underscores that player salary is influenced by a complex interplay of factors beyond just on-field performance.

Summary and Recommendations:
Both regression models reveal that runs scored and wickets taken have some impact on player salaries, but each explains only a modest portion of the variance. This is reflected in the low R-squared values and the statistical significance results.

Key Takeaways:

Low Predictive Power: The low R-squared values in both models suggest that salaries are influenced by many other factors not captured in these models. These could include player experience, leadership roles, fan following, endorsements, and more.

Model Significance: While the runs scored model shows statistical significance, the wickets taken model is marginally non-significant. This points to the need for a more comprehensive model that includes additional predictor variables.

Improvement Strategies:

Incorporate More Features: Adding more variables that capture other aspects of player value (e.g., historical performance, match-winning contributions, endorsements, and leadership qualities) could improve the model's explanatory power.
Non-Linear Relationships: Explore non-linear regression models or interaction terms to capture more complex relationships between performance metrics and salaries.

Regularization Techniques: Using techniques like Ridge or Lasso regression might help in dealing with potential multicollinearity and improving model robustness.

CONCLUSION:

Based on the analysis of the regression models aimed at predicting player salaries in the Indian Premier League (IPL) based on their on-field performance (runs scored and wickets taken), several key insights and conclusions can be drawn:

Key Findings:

1. **Limited Predictive Power**:
    - Both models—one predicting salaries from runs scored and the other from wickets taken—exhibited low R-squared values (0.069 for runs scored and 0.074 for wickets taken). This indicates that each model explains only a small fraction of the variance in player salaries.
    - The runs scored model was statistically significant, whereas the wickets taken model was close to being significant but fell slightly short.
2. **Salary Influences Beyond Performance**:
    - The modest explanatory power of runs scored and wickets taken suggests that player salaries are influenced by a broader set of factors beyond just these performance metrics.
    - Elements such as player popularity, marketability, experience, leadership, past achievements, and strategic value to a team likely play significant roles in determining salaries.
3. **Potential for Model Improvement**:
    - The analysis highlights the need to include more diverse and potentially non-linear variables to better predict player salaries.
    - Exploring more complex models and integrating additional data sources could provide a more comprehensive understanding of what drives salaries in the IPL.

Strategic Implications:

- **For Teams**:
    - Understanding that on-field performance metrics like runs and wickets contribute only partially to salary decisions can help teams make more informed decisions by also considering off-field and strategic factors.
    - Teams might benefit from evaluating players holistically, looking beyond basic statistics to their overall contribution, potential for growth, and impact on the team's brand and fan engagement.

- **For Players**:
    - Players can leverage this insight to negotiate their contracts better, emphasizing aspects of their value that go beyond just their immediate on-field performance.
    - Enhancing their marketability and demonstrating leadership or consistent match-winning capabilities could be just as crucial as scoring runs or taking wickets.
- **For Analysts and Modelers**:
    - Future models should aim to incorporate a wider array of variables, including player-specific attributes (e.g., age, experience, versatility), and market-specific factors (e.g., demand for certain roles, team budgets).
    - Using more sophisticated modeling techniques, such as machine learning or ensemble methods, might capture the complex relationships between various factors and player salaries more effectively.

Summary:

The regression analysis of IPL player salaries based on runs scored and wickets taken reveals that while these performance metrics are relevant, they provide a limited view of what drives salary decisions. To better predict and understand player valuations, it is essential to consider a broader spectrum of factors and more advanced analytical approaches. This comprehensive understanding can guide teams in building competitive squads and help players maximize their earning potential by highlighting their multifaceted value to the franchise.