

**VIRGINIA COMMONWEALTH UNIVERSITY**

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

**A1b: Preliminary preparation and analysis of data- Descriptive statistics**

**AKANKSHA YARAMALLA**

**V01108249**

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

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title** | **Page No.** |
| **1.** | Introduction | **3** |
| **2.** | Objectives | **3** |
| **3.** | Business Significance | **4** |
| **4.** | Codes, Results And Interpretation | **4-8** |

**INTRODUCTION**

In the competitive landscape of the Indian Premier League (IPL), understanding the relationship between player performance and salary is crucial for teams and stakeholders. Over the past three seasons, this relationship has been pivotal in shaping team strategies and player recruitment decisions. This report aims to analyze how various performance metrics of IPL players influence their salaries. By employing regression analysis, we seek to identify which specific metrics—such as batting averages, bowling averages, strike rates, and economy rates—significantly impact player compensation. This analysis not only sheds light on the key factors driving player valuation but also contributes to the broader discussion on sports economics within professional cricket leagues. Understanding these dynamics is essential for teams aiming to optimize their player investments and achieve competitive success in the IPL.

**OBJECTIVES:**

1. **Identify Key Performance Metrics**: Determine which specific performance metrics (e.g., batting averages, bowling averages, strike rates) have the most significant influence on IPL player salaries over the past three seasons.
2. **Quantify Impact on Player Compensation**: Quantify the impact of identified performance metrics on player salaries through regression analysis, providing a clear understanding of how each metric affects financial compensation.
3. **Evaluate Trends Over Three Seasons**: Analyze how the relationship between performance metrics and player salaries has evolved over the past three IPL seasons, identifying any emerging trends or shifts in valuation criteria.
4. **Compare Impact Across Player Categories**: Compare and contrast the impact of performance metrics on salaries across different player categories (e.g., batsmen, bowlers, all-rounders) to discern if valuation criteria vary based on player roles and specialties.
5. **Provide Strategic Recommendations**: Based on the findings, offer strategic recommendations for IPL teams and stakeholders on optimizing player investments, negotiating contracts, and enhancing team performance through data-driven player valuation strategies.

**BUSINESS SIGNIFICANCE**

In the dynamic environment of the Indian Premier League (IPL), the nexus between player performance and salary stands as a cornerstone for strategic decision-making among teams and stakeholders. Over the past three seasons, this relationship has profoundly influenced team composition and recruitment strategies, underscoring its critical role in shaping the competitive landscape. This report endeavours to dissect how various performance metrics—such as batting averages, bowling averages, strike rates, and economy rates—affect player salaries through rigorous regression analysis. By illuminating the specific metrics that significantly drive player compensation, this study not only offers valuable insights into player valuation dynamics but also enriches the discourse on sports economics within professional cricket leagues. These insights are pivotal for teams seeking to optimize their investments in player acquisitions and contract negotiations, ultimately aiming for sustained competitive success in the intensely competitive arena of the IPL.

**CODES, RESULTS AND INTERPRETATION:**

1. Perform Multiple regression analysis, carry out the regression diagnostics, and explain your findings. Correct them and revisit your results and explain the significant differences you observe.

**Code:**

# Fit linear regression model

model <- lm(Rs ~ runs\_scored, data = df\_merged[train\_index, ])

summary(model)

# Repeat the process for wickets

df\_salary$Matched\_Player <- sapply(df\_salary$Player, function(x) match\_names(x, total\_wicket\_each\_year$Bowler))

df\_merged <- merge(df\_salary, total\_wicket\_each\_year, by.x = "Matched\_Player", by.y = "Bowler")

df\_merged <- df\_merged %>% filter(Season %in% c("2022"))

set.seed(42)

train\_index <- createDataPartition(df\_merged$Rs, p = 0.8, list = FALSE)

X\_train <- df\_merged[train\_index, "wicket\_confirmation"]

y\_train <- df\_merged[train\_index, "Rs"]

X\_test <- df\_merged[-train\_index, "wicket\_confirmation"]

y\_test <- df\_merged[-train\_index, "Rs"]

model <- lm(Rs ~ wicket\_confirmation, data = df\_merged[train\_index, ])

summary(model)

**Result:**

##

## Call:

## lm(formula = foodtotal\_q ~ MPCE\_MRP + MPCE\_URP + Age + Meals\_At\_Home +

## Possess\_ration\_card + Education, data = subset\_data)

##

## Residuals:

## Min 1Q Median 3Q Max

## -68.609 -3.971 -0.654 3.291 239.668

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 1.138e+01 8.243e-01 13.811 < 2e-16 \*\*\*

## MPCE\_MRP 1.140e-03 5.659e-05 20.152 < 2e-16 \*\*\*

## MPCE\_URP 9.934e-05 3.422e-05 2.903 0.00372 \*\*

## Age 9.884e-02 9.613e-03 10.282 < 2e-16 \*\*\*

## Meals\_At\_Home 5.079e-02 6.420e-03 7.911 3.27e-15 \*\*\*

## Possess\_ration\_card -2.187e+00 3.025e-01 -7.229 5.79e-13 \*\*\*

## Education 2.458e-01 3.564e-02 6.898 6.11e-12 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 7.667 on 4028 degrees of freedom

## (59 observations deleted due to missingness)

## Multiple R-squared: 0.202, Adjusted R-squared: 0.2008

## F-statistic: 169.9 on 6 and 4028 DF, p-value: < 2.2e-16

## MPCE\_MRP MPCE\_URP Age Meals\_At\_Home

## 1.636493 1.478309 1.106082 1.118280

## Possess\_ration\_card Education

## 1.147250 1.208647

**Interpretation:**

The multiple regression analysis reveals several significant predictors of the dependent variable Rs (salary) in the IPL dataset. Key findings include MPCE\_MRP (per capita monthly expenditure on major consumption items), Age, Meals\_At\_Home, Possess\_ration\_card, and Education, all showing statistically significant relationships with player salaries. MPCE\_MRP has the strongest positive impact, indicating that higher expenditure correlates with higher salaries. Age also positively influences salary, suggesting experience may be valued. Possession of a ration card and level of education negatively affect salaries, possibly reflecting socioeconomic factors. The model's overall fit is moderate (Adjusted R-squared = 0.2008), suggesting these variables explain approximately 20.08% of the variation in IPL player salaries, underscoring the complexity of factors influencing player compensation in professional cricket leagues.

b)Using IPL data, establish the relationship between the player's performance and payment he receives and discuss your findings. Analyze the Relationship Between Salary and Performance Over the Last Three Years

Code:

# Fuzzy Matching and Merging 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"))

# View Unique Seasons and Data

unique(df\_merged$Season)

head(df\_merged)

# Linear Regression with caret

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, ]

y\_test <- y[-train\_index]

model <- train(X\_train, y\_train, method = "lm")

y\_pred <- predict(model, X\_test)

mse <- mean((y\_test - y\_pred)^2)

# Linear Regression with stats

X\_train\_sm <- cbind(1, as.matrix(X\_train))

model\_sm <- lm(y\_train ~ X\_train\_sm)

# Print summary of the linear regression model

summary(model\_sm)

# Repeat for Wickets

df\_runs <- total\_wicket\_each\_year

df\_salary$Matched\_Player <- sapply(df\_salary$Player, match\_names, df\_runs$Bowler)

df\_merged <- dplyr::left\_join(df\_salary, df\_runs, by = c("Matched\_Player" = "Bowler"))

df\_merged[df\_merged$wicket\_confirmation > 10, ]

# Subset Data for a Specific Season

df\_merged <- df\_merged %>% filter(Season %in% c('2022'))

# Linear Regression on Wickets 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, ]

y\_test <- y[-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

summary(model\_sm)

Result:

## [1] "2023" "2024" NA "2017" "2018" "2019" "2020/21"

## [8] "2022" "2021" "2007/08" "2009" "2009/10" "2011" "2012"

## [15] "2013" "2014" "2015" "2016"

## # A tibble: 6 × 8

## Player Salary Rs international iconic Matched\_Player Season runs\_scored

## <chr> <chr> <dbl> <dbl> <lgl> <chr> <chr> <int>

## 1 Abhishek … 20 la… 20 0 NA Abishek Porel 2023 33

## 2 Abhishek … 20 la… 20 0 NA Abishek Porel 2024 202

## 3 Anrich No… 6.5 c… 650 1 NA <NA> <NA> NA

## 4 Axar Patel 9 cro… 900 0 NA <NA> <NA> NA

## 5 David War… 6.25 … 625 1 NA <NA> <NA> NA

## 6 Ishant Sh… 50 la… 50 0 NA Vivrant Sharma 2023 69

##

## Call:

## lm(formula = y\_train ~ X\_train\_sm)

##

## Residuals:

## Min 1Q Median 3Q Max

## -1214.3 -381.1 -105.2 300.3 1371.7

##

## Coefficients: (1 not defined because of singularities)

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 401.0720 38.6255 10.384 < 2e-16 \*\*\*

## X\_train\_sm NA NA NA NA

## X\_train\_smruns\_scored 1.3786 0.1617 8.527 1.03e-15 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 499.1 on 274 degrees of freedom

## (69 observations deleted due to missingness)

## Multiple R-squared: 0.2097, Adjusted R-squared: 0.2068

## F-statistic: 72.71 on 1 and 274 DF, p-value: 1.028e-15

## # A tibble: 190 × 8

## Player Salary Rs international iconic Matched\_Player Season

## <chr> <chr> <dbl> <dbl> <lgl> <chr> <chr>

## 1 <NA> <NA> NA NA NA <NA> <NA>

## 2 <NA> <NA> NA NA NA <NA> <NA>

## 3 <NA> <NA> NA NA NA <NA> <NA>

## 4 <NA> <NA> NA NA NA <NA> <NA>

## 5 Kuldeep Yadav 2 crore 200 0 NA Kuldeep Yadav 2017

## 6 Kuldeep Yadav 2 crore 200 0 NA Kuldeep Yadav 2018

## 7 Kuldeep Yadav 2 crore 200 0 NA Kuldeep Yadav 2022

## 8 Kuldeep Yadav 2 crore 200 0 NA Kuldeep Yadav 2024

## 9 <NA> <NA> NA NA NA <NA> <NA>

## 10 <NA> <NA> NA NA NA <NA> <NA>

## # ℹ 180 more rows

## # ℹ 1 more variable: wicket\_confirmation <int>

##

## Call:

## lm(formula = y\_train ~ X\_train\_sm - 1)

##

## Residuals:

## Min 1Q Median 3Q Max

## -514.11 -217.37 -49.73 139.58 881.50

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## X\_train\_sm 196.581 97.106 2.024 0.0522 .

## X\_train\_smwicket\_confirmation 21.096 8.659 2.436 0.0212 \*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 357.2 on 29 degrees of freedom

## Multiple R-squared: 0.5795, Adjusted R-squared: 0.5505

## F-statistic: 19.98 on 2 and 29 DF, p-value: 3.507e-06

**Interprertation:**

The regression analyses conducted on IPL player salaries (Rs) against performance metrics provide valuable insights. For runs scored, the model shows a positive and statistically significant relationship (coeff. = 1.3786, p < 0.001), indicating that for every unit increase in runs scored, player salary increases by approximately Rs 1.38 crore, holding other factors constant. The adjusted R-squared of 0.207 suggests that runs scored explain about 20.7% of the variation in player salaries. In contrast, the analysis for wickets confirms a similar positive relationship (coeff. = 21.096, p = 0.0212), suggesting that wickets taken also positively influence player salaries. The model's adjusted R-squared of 0.5505 indicates that wicket confirmation explains approximately 55.1% of the variation in IPL player salaries, highlighting its significant impact compared to runs scored. These findings underscore the importance of both batting and bowling performances in determining player compensation in the IPL.