VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A1b: Preliminary preparation and analysis of data- Descriptive statistics

RAKSHITH HARISH KUMAR
V01107367

Date of Submission: 23-06-2024

CONTENTS

INTRODUCTION

The Indian Premier League (IPL) is a highly competitive league, and clubs and other stakeholders need to grasp the link between player performance and compensation. This connection has been crucial in determining club tactics and player acquisition choices throughout the last three seasons. This study looks at how different IPL players' performance measures affect their pay. Regression analysis is the method we use to determine whether particular metrics—like strike rates, economy rates, batting averages, and bowling averages—have a substantial influence on player remuneration. The main drivers of player valuation are clarified by this research, which also adds to the larger conversation on sports economics in professional cricket leagues. A team's ability to maximise player investments and succeed competitively depends on its understanding of these dynamics.

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:

a) **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:

```
# 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
```

Interpretation:

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.