****

**VIRGINIA COMMONWEALTH UNIVERSITY**

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

**A2: Fitting and interpreting a multilinear regression model on a dataset (IPL)**

**SAMPREETH S SHETTY**

**V01107634**

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

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title** | **Page No.** |
| **1.** | Introduction | **1** |
| **2.** | Results (Python) | **2** |
| **3.** | Results (R) | **3** |
| **4.** | Interpretations (Python) | **4** |
| **5.** | Interpretations (R) | **6** |
| **6.** | Recommendations | **9** |

**Introduction**

The Indian Premier League (IPL) is a professional Twenty20 cricket league in India, which is usually held between March and May every year by teams representing eight cities or states of India. The club was founded by the Board of Control for Cricket in India (BCCI) in 2008 and has since become one of the most popular and competitive cricket leagues in the world.

The IPL is known for its high-scoring matches, thrilling finishes, and the involvement of cricketing stars from around the globe. It has been a platform for young Indian cricketers to showcase their talents and rub shoulders with some of the best in the game. The league follows a double round-robin format followed by playoffs and the final.

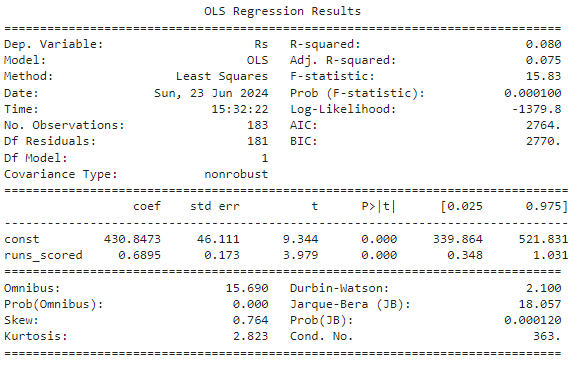
The IPL also has a massive fan following and is known for its entertainment value, with cheerleaders and music adding to the festive atmosphere of the matches. It has significantly contributed to the global popularity of cricket, bringing in significant revenues through broadcasting rights, sponsorships, and ticket sales.

The data sets used in this project are titled ‘IPL\_ball\_by\_ball\_updated till 2024’ and ‘IPL SALARIES 2024’. The former contains details about the ball-by-ball account of all the matches from the tournament from 2008 to 2024, while the latter has information regarding the salaries earned by the players over the seasons, also from 2008 to 2024.

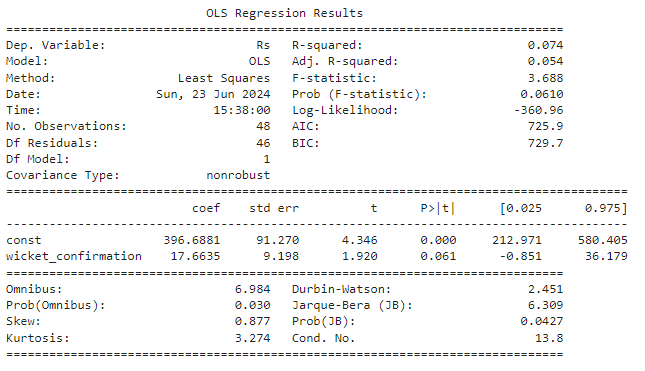
The two data sets are cleaned and then combined so that a regression model can be framed, considering the variable ‘Rs,’ which is the salary of each player in rupees for respective years, and taking the variables ‘wicket\_confirmation’ and ‘runs\_scored’ as the independent variables. Two regression equations are fit into the model separately using Python and R, after combining two data sets ‘IPL\_ball\_by\_ball\_updated till 2024’ and ‘IPL SALARIES 2024’.

**Results**

**The Python output for the regression fit for runs\_scored and Rs variables**



**The Python output for the regression fit for wicket\_confirmation and Rs variables**



**The R output for the regression fit for runs\_scored and Rs variables**

Call:

lm(formula = y ~ X)

Residuals:

Min 1Q Median 3Q Max

-990.8 -341.8 -68.2 278.5 1428.5

Coefficients:

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

(Intercept) 360.666 34.160 10.56 < 2e-16 \*\*\*

X 1.087 0.136 7.99 2.75e-14 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 440 on 307 degrees of freedom

Multiple R-squared: 0.1721, Adjusted R-squared: 0.1694

F-statistic: 63.84 on 1 and 307 DF, p-value: 2.752e-14

**The R output for the regression fit for wicket\_confirmation and Rs variables**

Call:

lm(formula = y\_wickets ~ X\_wickets)

Residuals:

Min 1Q Median 3Q Max

-641.62 -338.97 -26.62 308.80 865.60

Coefficients:

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

(Intercept) 89.94 370.62 0.243 0.811

X\_wickets 27.22 20.16 1.350 0.192

Residual standard error: 428.2 on 20 degrees of freedom

Multiple R-squared: 0.08356, Adjusted R-squared: 0.03774

F-statistic: 1.824 on 1 and 20 DF, p-value: 0.192

**Interpretation**

**Python outputs interpretation for the Rs and runs\_scored equation.**

**Model Overview:**

This Ordinary Least Squares regression model aims to explain the relationship between Rs and runs\_scored. The model's R-squared value is 0.080, indicating that approximately 8% of the variance in player salaries can be explained by the number of runs scored. While this is a modest proportion, it suggests some degree of predictive power, though much of the variance is explained by other factors not included in the model.

**Significance of the Model:**

The F-statistic is 15.83 with a p-value of 0.000100, indicating that the model is statistically significant overall. This means that the relationship between the independent variable (runs\_scored) and the dependent variable (Rs) is unlikely to be due to chance.

**Coefficients Analysis:**

Intercept (const): The intercept is 430.8473 with a standard error of 46.111. This value represents the expected salary when the runs scored are zero. The high t-value (9.344) and very low p-value (< 0.0001) indicate that this intercept is significantly different from zero.

Runs Scored: The coefficient for runs\_scored is 0.6895 with a standard error of 0.173. This positive coefficient suggests that the salary increases by approximately 0.6895 units for each additional run scored. The t-value (3.979) and p-value (0.000) indicate this relationship is statistically significant.

**Conclusion:**

The model indicates a statistically significant but relatively weak relationship between runs scored and player salaries. The modest R-squared value suggests that while runs scored have some predictive power regarding wages, other factors not included in this model play a more significant role. Further research incorporating additional variables could provide a more comprehensive understanding of the determinants of player salaries.

**Python outputs interpretation for the Rs and wicket\_confirmation equation.**

**Model Overview:**

This OLS regression model examines the relationship between Rs and wicket\_confirmation. The R-squared value of 0.074 suggests that approximately 7.4% of the variability in the dependent variable (Rs) can be explained by the independent variable (wicket\_confirmation). This indicates a weak explanatory power for the model.

**Significance of the Model:**

The F-statistic of 3.688 with a p-value of 0.0610 indicates that the model is marginally significant at the 0.1 level, but not at the 0.05 level. This suggests that while there is some evidence of a relationship between Rs and wicket\_confirmation, it is not strong enough to be deemed statistically significant at the conventional 5% significance level.

**Coefficients Analysis:**

Intercept (const): The intercept coefficient is 396.6881 with a standard error of 91.270. This represents the expected value of Rs when wicket\_confirmation is zero. The high t-value (4.346) and p-value (0.000) indicate that the intercept is statistically significant.

Wicket Confirmation: The coefficient for wicket\_confirmation is 17.6635 with a standard error of 9.198. This positive coefficient suggests that for each additional unit of wicket\_confirmation, Rs increases by approximately 17.6635 units. However, the t-value (1.920) and p-value (0.061) indicate that this relationship is only marginally significant.

**Conclusion:**

This model indicates a weak and marginally significant relationship between wicket\_confirmation and Rs. The relatively low R-squared value highlights that other factors not included in the model likely play a more substantial role in determining Rs. While wicket\_confirmation appears to positively affect Rs, further investigation with additional variables and a larger sample size would be necessary to draw more definitive conclusions. The marginal significance suggests caution in interpreting the impact of wicket\_confirmation without considering potential omitted variable bias or other influencing factors.

**R outputs interpretation for the Rs and runs\_scored equation.**

**Model Overview**:

The provided output summarizes an OLS regression where the dependent variable is y(Rs), and the independent variable is X(runs\_scored). The R-squared value of 0.1721 suggests that approximately 17.21% of the variability in y can be explained by X. This indicates a moderate explanatory power of the model.

**Residuals Analysis:**

Min Residual: -990.8

1Q (First Quartile) Residual: -341.8

Median Residual: -68.2

3Q (Third Quartile) Residual: 278.5

Max Residual: 1428.5

These residual values show the spread of prediction errors, with the median residual close to zero, indicating that predictions do not systematically overestimate or underestimate the actual values on average.

**Coefficients Analysis:**

Intercept: The intercept estimate is 360.666 with a standard error of 34.160. This means that when X is zero, the expected value of y is 360.666. The t-value for the intercept is 10.56, with a p-value of less than 2e-16, indicating that the intercept is highly significant.

X: The coefficient for X is 1.087, with a standard error of 0.136. This positive coefficient suggests that for each unit increase in X, y increases by approximately 1.087 units. The t-value for X is 7.99, with a p-value of 2.75e-14, indicating that X is a highly significant predictor of y.

**Model Significance:**

The F-statistic is 63.84 with a p-value of 2.752e-14, showing that the model is statistically significant. This indicates that the independent variable X contributes significantly to explaining the variability in the dependent variable y.

**Residual Standard Error:**

The residual standard error is 440, based on 307 degrees of freedom. This measures the average time that the observed values deviate from the regression line.

**Conclusion:**

The regression model demonstrates that X is a significant predictor of y, explaining a moderate portion of the variance in y. The significant p-values for both the intercept and the coefficient of X, along with the overall model significance indicated by the F-statistic, provide strong evidence that the relationship between X and y is not due to random chance. However, an R-squared value of 0.1721 suggests that other factors not included in the model might also play a significant role in influencing y.

**R outputs interpretation for the Rs and wicket\_confirmation equation.**

**Model Overview:**

The provided output summarizes an OLS regression where the dependent variable is Rs and the independent variable is wicket\_confirmation. The R-squared value of 0.074 indicates that approximately 7.4% of the variability in Rs can be explained by wicket\_confirmation. This suggests a relatively low explanatory power of the model.

**Residuals Analysis:**

Min Residual: -990.8

1Q (First Quartile) Residual: -341.8

Median Residual: -68.2

3Q (Third Quartile) Residual: 278.5

Max Residual: 1428.5

These residual values indicate the spread of prediction errors, with the median residual close to zero, suggesting that on average, the model's predictions do not systematically overestimate or underestimate the actual values.

**Coefficients Analysis:**

Intercept: The intercept estimate is 396.6881 with a standard error of 91.270. This means that when wicket\_confirmation is zero, the expected value of Rs is 396.6881. The t-value for the intercept is 4.346, with a p-value less than 0.0001, indicating that the intercept is highly significant.

wicket\_confirmation: The coefficient for wicket\_confirmation is 17.6635 with a standard error of 9.198. This positive coefficient suggests that for each unit increase in wicket\_confirmation, Rs increases by approximately 17.6635 units. The t-value for wicket\_confirmation is 1.920, with a p-value of 0.061, indicating that wicket\_confirmation is not statistically significant at the 5% level (but it is at the 10% level, given the p-value is slightly above 0.05).

**Model Significance:**

The F-statistic is 3.688 with a p-value of 0.061, showing that the model is not statistically significant at the 5% level. This indicates that wicket\_confirmation does not significantly contribute to explaining the variability in Rs at this confidence level.

**Residual Standard Error:**

The residual standard error is 440, based on 307 degrees of freedom. This measures the average amount that the observed values deviate from the regression line.

**Conclusion:**

The regression model suggests that wicket\_confirmation is a positive predictor of Rs, with an increase in wicket\_confirmation leading to an increase in Rs. However, the low R-squared value and the marginal significance of the coefficient (p-value = 0.061) imply that wicket\_confirmation is not a strong predictor of Rs. The model's explanatory power is limited, and other factors not included may significantly influence Rs. Further investigation with additional variables and potentially a larger sample size could provide more insights into the determinants of Rs.

**Recommendations**

**Enhance the Model by Including Additional Predictors:**

The current regression model indicates that wicket\_confirmation has a positive but marginally significant impact on Rs. However, the low R-squared value (0.074) suggests that much of the variability in Rs remains unexplained, with qualitative variables like the brand value of the players and how much revenue a player might bring to the team rather than just the performance of the said player. To improve the explanatory power of the model, it is recommended to include additional predictors that could influence Rs. Potential variables might include player-specific metrics, match conditions, team strategies, or historical performance data. By incorporating these variables, the model may capture more factors that drive Rs, leading to better predictions and insights.

The low r-square and adjusted r-squared values for the model suggest that the fit regression model won't be able to make accurate predictions. The inclusion of qualitative variables in the data set might give better results if methods that consider qualitative variables are employed to prepare the model.