VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modeling (SCMA 632)

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

POLAMREDDY MADHUMITHA

V01107517

Date of Submission: 23-06-2024

CONTENTS

Introduction

Teams from eight different Indian cities or states compete in the professional Twenty20 cricket tournament known as the Indian Premier tournament (IPL), which is typically played between March and May each year. The Board of Control for Cricket in India (BCCI) established the team in 2008, and it has since grown to be of the most well-liked and fiercely competitive cricket leagues worldwide.

The Indian Premier League (IPL) is renowned for its spectacular endings, high-scoring matches, and participation by international cricket stars. Young Indian cricket players have used it as a platform to show off their skills and mingle with some of the greatest players in the world. The league has a double round robin system, with the final and playoffs coming after.

With cheerleaders and music adding to the lively atmosphere of the game, the Indian Premier League (IPL) is renowned for its entertainment value and has a sizable fan base as well. It has made a substantial financial contribution to cricket's worldwide appeal through sponsorships, ticket sales, and television rights.

The project makes use of two data sets: "IPL SALARIES 2024" and "IPL_ball_by_ball_updated till 2024." The former offers information on the ball-by-ball breakdown of every tournament match from 2008 to 2024, and the later has data on player pay over the course of the seasons, also from 2008 to 2024.

After the two data sets are cleaned and joined, a regression model can be constructed with the variables "wicket_confirmation" and "runs_scored" as the independent variables and the variable "Rs," which represents each player's wage in rupees for the corresponding year. Using Python and R, two regression equations are fitted independently into the model after two data sets, "IPL SALARIES 2024" and "IPL ball by ball updated till 2024," are combined.

Results

The Python output for the regression fit for runs_scored and Rs variables

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     Rs   R-squared:                       0.118
Model:                            OLS   Adj. R-squared:                  0.113
Method:                 Least Squares   F-statistic:                     24.26
Date:                Sun, 23 Jun 2024   Prob (F-statistic):           1.89e-06
Time:                        21:48:37   Log-Likelihood:                -1384.2
No. Observations:                 183   AIC:                             2772.
Df Residuals:                     181   BIC:                             2779.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        440.4557     46.237      9.526      0.000     349.223     531.689
runs_scored    0.9448      0.192      4.925      0.000       0.566       1.323
==============================================================================
Omnibus:                       13.531   Durbin-Watson:                   1.967
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               15.137
Skew:                           0.695   Prob(JB):                     0.000517
Kurtosis:                       2.772   Cond. No.                         321.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

The Python output for the regression fit for wicket_confirmation and Rs variables

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     Rs   R-squared:                       0.090
Model:                            OLS   Adj. R-squared:                  0.071
Method:                 Least Squares   F-statistic:                     4.569
Date:                Sun, 23 Jun 2024   Prob (F-statistic):             0.0379
Time:                        21:49:58   Log-Likelihood:                -358.29
No. Observations:                  48   AIC:                             720.6
Df Residuals:                      46   BIC:                             724.3
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        357.4774     89.737      3.984      0.000     176.846     538.109
wicket_confirmation    16.4952      7.717      2.137      0.038       0.961      32.030
==============================================================================
Omnibus:                        4.371   Durbin-Watson:                   1.723
Prob(Omnibus):                  0.112   Jarque-Bera (JB):                4.228
Skew:                           0.701   Prob(JB):                        0.121
Kurtosis:                       2.616   Cond. No.                         16.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

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:

The purpose of this Ordinary Least Squares regression model is to elucidate the correlation between runs_scored and Rs. The amount of runs scored accounts for about 11.8% of the variance in player wages, according to the model's R-squared value of 0.118. Even while this is a small percentage, it does indicate some predictive value, even though the model does not account for a large fraction of the variance.

Significance of the Model:

The F-statistic is 24.26 with a p-value of `1.89e-06`, strongly suggests that the regression mo del as a whole is statistically significant. This means that the variables 'wicket_confirmation' and 'runs_scored' are significant predictors of players' salaries in the IPL.

Coefficients Analysis:

Intercept (const): The intercept is 440.4557 with a standard error of 46.237. This value represe nts the expected salary when the runs scored are zero. The high t-value (9.526) and very low p-value (< 0.0001) indicate that this intercept is significantly different from zero.

Runs Scored: The coefficient for runs_scored is 0.9448 with a standard error of 0.192. This po sitive coefficient suggests that the salary increases by approximately 0.9448 units for each additional run scored. The t-value (4.925) and p-value (0.000) indicate this relationship is statistically significant.

Conclusion:

The model shows that there is a weak but statistically significant correlation between player wages and runs scored. The weak R-squared value indicates that other characteristics not included in this model are more important in predicting earnings, even though runs scored has some predictive power in this regard. A more thorough understanding of the factors influencing player earnings may be possible with more study that takes into account other variables.

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 0.090 suggests that approximately 9% 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 4.569 with a p-value of 0.0379 indicates that the model is marginally significant at the 0.1 level, but not at the 0.05 level. This implies that although a correlation between Rs and wicket confirmation may exist, it is not significant enough to be considered statistically significant at the standard significance threshold of 5%.

Coefficients Analysis:

Intercept (const): The intercept coefficient is 357.4774 with a standard error of 89.737. This represents the expected value of Rs when wicket confirmation is zero. The high t-value (3.984) and p-value (0.000) indicate that the intercept is statistically significant.

Wicket Confirmation: The coefficient for wicket confirmation is 16.4952 with a standard error of 7.717. This positive coefficient suggests that for each additional unit of wicket confirmation, Rs increases by approximately 16.4952 units. However, the t-value (2.137) and p-value (0.038) 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:

```
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
```

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:

According to the regression model, X accounts for a moderate amount of the variance in y, making it a significant predictor of y. The correlation between X and y is not the result of chance, as demonstrated by the significant p-values for the intercept and the coefficient of X as well as the overall model significance shown by the F-statistic. An R-squared value of 0.1721, however, indicates that y might possibly be significantly influenced by other factors that are not included in the model.

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 standard error of the intercept estimate is 91.270, and it is 396.6881. This indicates that the expected value of Rs is 396.6881 when wicket_confirmation is zero. With a p-value of less than 0.0001 and an intercept t-value of 4.346, the intercept is highly significant.

wicket_confirmation: For wicket_confirmation, the coefficient is 17.6635 with a standard error of 9.198. According to this positive coefficient, Rs rises by about 17.6635 units for every unit increase in wicket_confirmation. Indicating that wicket_confirmation is not statistically significant at the 5% level (but is at the 10% level, provided the p-value is slightly over 0.05), the t-value for wicket_confirmation is 1.920 and the p-value is 0.061.

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. Therefore,This measures the average amount that the observed values deviate from the regression line.

Conclusion:

According to the regression model, there is a positive correlation between wicket_confirmation and Rs, meaning that a rise in wicket_confirmation corresponds to an increase in Rs. The marginal significance of the relationship (p-value = 0.061) and the low R-squared value, however, suggest that wicket_confirmation is not a very good predictor of Rs. The explanatory power of the model is limited, and Rs may be strongly influenced by other factors that are not included. More research with more variables and possibly a bigger sample size could shed more light on the factors that influence Rs.

Recommendations

Enhance the Model by Including Additional Predictors:

According to the present regression model, Rs is positively but only minimally impacted by wicket_confirmation. Nonetheless, the low R-squared value (0.074) indicates that qualitative factors other than player performance, such as player brand value and potential revenue contribution, may account for a large portion of the variability in Rs. It is advised to incorporate other predictors that potentially affect Rs in order to increase the model's explanatory power. Possible variables could be measurements unique to each player, the circumstances of the contest, team tactics, or past performance information. By adding these variables, the model might be able to account for more factors that influence Rs, which would improve forecasts and insights.

It appears from the model's low r-square and modified r-squared values that the fit regression model won't be able to produce reliable predictions. If techniques that take into account qualitative factors are used to create the model, adding qualitative variables to the data set may result in improved outcomes.