VIRGINIA COMMONWEALTH UNIVERSITY

STATISTICAL ANALYSIS & MODELING

A2: REGRESSION - PREDICTIVE ANALYTICS USING PYTHON AND R

SARATH S
V01109792

Date of Submission: 23/06/2024

CONTENTS

REGRESSION - PREDICTIVE ANALYTICS USING PYTHON

INTRODUCTION

The dataset provides an in-depth analysis of food consumption patterns across India, covering various dietary habits in both urban and rural areas. It includes key metrics such as the quantity of meals consumed at home, consumption of specific food items like rice, wheat, chicken, and pulses, and the total number of meals per day. This comprehensive dataset is essential for understanding the nutritional intake and food preferences of different demographics in the region.

The Indian Premier League (IPL), also known as the TATA IPL due to sponsorship, is an annual men's Twenty20 (T20) cricket league in India. Established by the Board of Control for Cricket in India (BCCI) in 2007, the league features ten franchise teams representing different states or cities.

Regression analysis is a statistical technique used to model and analyze the relationships between a dependent variable and one or more independent variables. The main objective of regression analysis is to understand how the dependent variable changes when any of the independent variables vary while keeping the others constant.

- Regression can be used to predict outcomes based on historical data, aiding in forecasting and decision-making.
- It provides insights into the strength and nature of relationships between variables, which can inform strategic planning and policy development.

OBJECTIVES

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.

b) 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

BUSINESS SIGNIFICANCE

Regression analysis is a powerful tool for extracting valuable insights from data, making it indispensable for business decision-making. By applying regression to Indian Premier League (IPL) data and National Sample Survey Office (NSSO) 68th round data, businesses can uncover patterns, predict future trends, and drive strategic initiatives.

1. For IPL Data:-

- **Performance Prediction**: Identify key factors influencing player and team performance to predict future success.
- **Team Composition Optimization**: Optimize team selection and strategy based on historical performance data.
- **Revenue Maximization**: Predict ticket sales, merchandise revenue, and viewership ratings to maximize financial returns.

2. For NSSO68 Data:-

- **Demand Forecasting**: Predict consumer demand and preferences across different regions and income groups.
- **Resource Allocation**: Optimize resource distribution for marketing, sales, and operations based on regional economic conditions and consumer behavior.
- **Policy Impact Evaluation**: Assess the effectiveness of governmental policies and programs on various economic and social outcomes, guiding corporate social responsibility (CSR) initiatives.

In both cases, regression analysis enables data-driven decision-making, optimizing resource use, improving targeting strategies, and enhancing overall efficiency and effectiveness in business and policy environments.

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. [NSSO68]

```
# Fit the regression model
X = subset_data[['MPCE_MRP', 'MPCE_URP', 'Age', 'Meals_At_Home', 'Possess_ration_card',
'Education']]
X = sm.add_constant(X)  # Add a constant term for the intercept
y = subset_data['foodtotal_q']
model = sm.OLS(y, X).fit()

# Print the regression results
print(model.summary())
```

```
                     OLS Regression Results
==============================================================================
=====
Dep. Variable:         foodtotal_q   R-squared:                 0.171
Model:                        OLS   Adj. R-squared:            0.170
```

```
Method:                 Least Squares   F-statistic:                  140.6
Date:               Sun, 23 Jun 2024   Prob (F-statistic):          1.66e-162
Time:                       21:36:58   Log-Likelihood:              -14354.
No. Observations:               4094   AIC:                     2.872e+04
Df Residuals:                   4087   BIC:                     2.877e+04
Df Model:                          6
Covariance Type:           nonrobust
==============================================================================
==============
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const              12.0534      0.865     13.932      0.000      10.357      13.750
MPCE_MRP             0.0009   5.81e-05     16.212      0.000       0.001       0.001
MPCE_URP          9.543e-05   3.58e-05      2.668      0.008    2.53e-05       0.000
Age                  0.1296      0.010     13.056      0.000       0.110       0.149
Meals_At_Home        0.0374      0.007      5.562      0.000       0.024       0.051
Possess_ration_card    -2.9937      0.315     -9.504      0.000      -3.611      -2.376
Education            0.2470      0.037      6.608      0.000       0.174       0.320
==============================================================================
=====
Omnibus:                     5440.704   Durbin-Watson:                   1.650
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          6839100.020
Skew:                           6.726   Prob(JB):                         0.00
Kurtosis:                     202.779   Cond. No.                     3.95e+04
==============================================================================
=====

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.95e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Interpretation

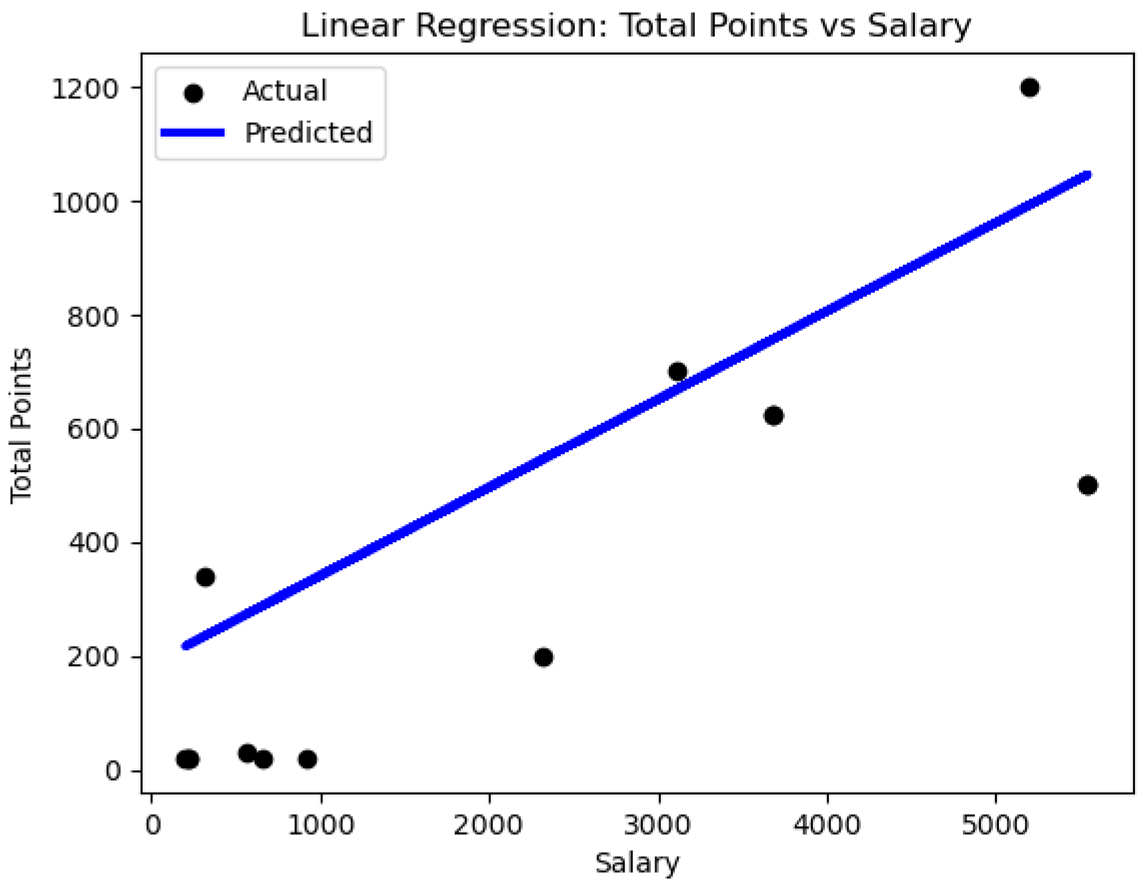
The regression analysis evaluates the relationship between several predictors (MPCE_MRP, MPCE_URP, Age, Meals_At_Home, Possess_ration_card, and Education) and the dependent variable, foodtotal_q. The model has an R-squared value of 0.171, indicating that approximately 17.1% of the variance in foodtotal_q is explained by these predictors. The overall F-statistic is highly significant ($p < 0.0001$), suggesting the model is statistically significant. Individually, all predictors except for the intercept are significant at the 0.05 level, indicating they each contribute uniquely to explaining the variation in foodtotal_q. The coefficients suggest that MPCE_MRP, MPCE_URP, Age, Meals_At_Home, and Education have a positive relationship with foodtotal_q, while possessing a ration card is negatively associated. However, the model shows signs of potential multicollinearity, as indicated by the high condition number (3.95e+04). This could imply that some predictor variables are highly correlated with each other, potentially affecting the stability and interpretation of the regression coefficients. Additionally, the skewness and kurtosis values indicate non-normality in the residuals, suggesting that the model may have some specification issues or that transformations of variables might be necessary for better accuracy.

b) Establish the relationship between the player's performance and payment he receives and discuss your findings. [IPL Datasets]

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
import matplotlib.pyplot as plt

# Load the CSV file
file_path = 'combined_output_with_salaries - Copy.csv'
data = pd.read_csv(file_path)

# Define the predictor and response variables
y = data['salary']  # Response variable
X = data[['Total_Points']]  # Predictor variable

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

# Create the linear regression model
model = LinearRegression()

# Train the model on the training data
model.fit(X_train, y_train)

# Predict on the test data
y_pred = model.predict(X_test)

# Calculate the mean squared error and the coefficient of determination (R^2)
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

# Calculate the adjusted R^2
n = len(y_test)
p = X_test.shape[1]
adjusted_r2 = 1 - (1 - r2) * (n - 1) / (n - p - 1)

# Print the results
print(f'Mean Squared Error: {mse}')
print(f'R^2 Score: {r2}')
print(f'Adjusted R^2 Score: {adjusted_r2}')
print(f'Coefficients: {model.coef_}')
print(f'Intercept: {model.intercept_}')

# Plot the results
plt.scatter(X_test, y_test, color='black', label='Actual')
plt.plot(X_test, y_pred, color='blue', linewidth=3, label='Predicted')
plt.xlabel('Salary')
plt.ylabel('Total Points')
plt.title('Linear Regression: Total Points vs Salary')
plt.legend()
plt.show()
```

Result:

```
Mean Squared Error: 81958.64280948695

R^2 Score: 0.3291084655234717

Adjusted R^2 Score: 0.2732008376504277

Coefficients: [0.15510582]

Intercept: 185.4181495234746
```



Interpretation

The linear regression analysis aimed to predict salaries based on total points yielded a model with a mean squared error (MSE) of 81,958.64, indicating the average squared difference between the observed and predicted values. The R^2 score, a measure of the model's explanatory power, was 0.33, suggesting that approximately 33% of the variance in salary can be explained by the total points. The adjusted R^2 score, which accounts for the number of predictors and sample size, was slightly lower at 0.27, reflecting a modest fit of the model. The coefficient for total points was 0.155, meaning that for each additional point, the salary is expected to increase by about 0.155 units. The intercept was 185.42, representing the estimated salary when total points are zero. While there is a positive relationship between total points and salary, the relatively low R^2 scores indicate that other factors likely influence salary significantly and are not captured by this model.

c) Analyze the Relationship Between Salary and Performance Over the Last Three Years [IPL Datasets]

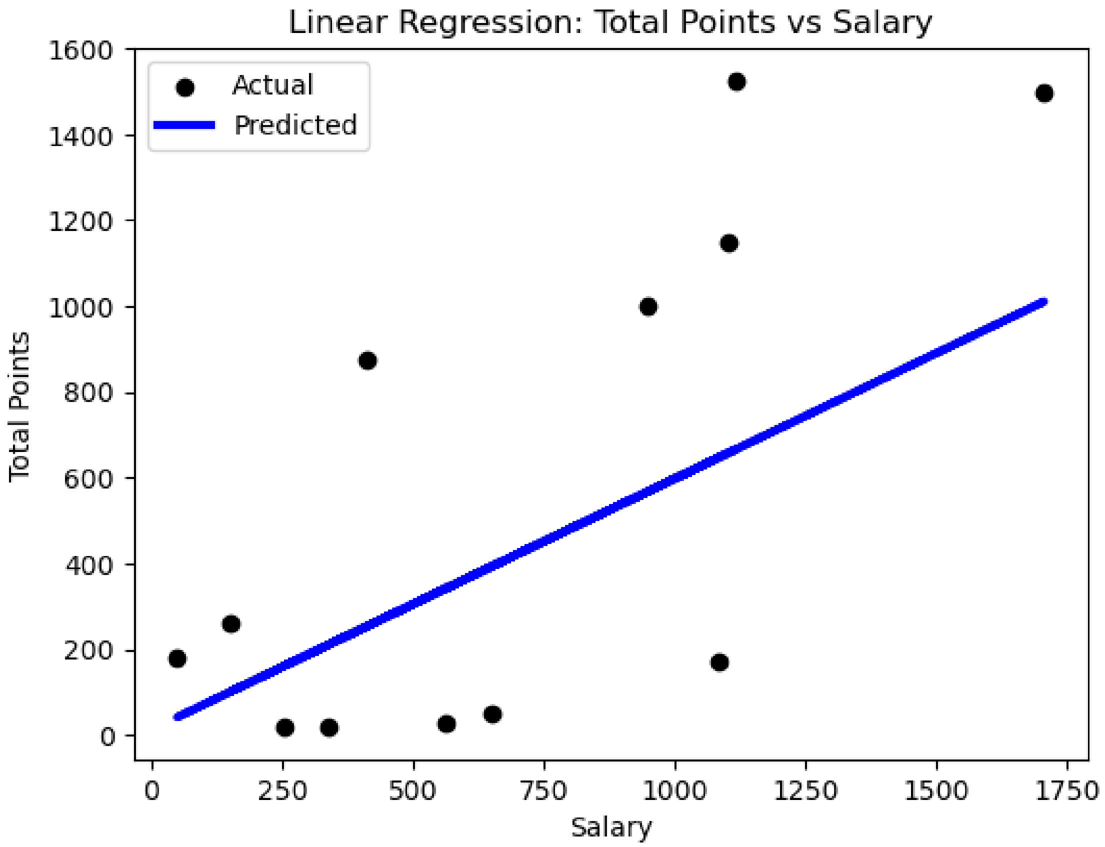
Code:

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
import matplotlib.pyplot as plt

# Load the CSV file
file_path = 'combined_output_with_salaries - Copy.csv'
data = pd.read_csv(file_path)

# Define the predictor and response variables
y = data['salary']  # Response variable
X = data[['Total_Points']]  # Predictor variable

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

# Create the linear regression model
model = LinearRegression()

# Train the model on the training data
model.fit(X_train, y_train)

# Predict on the test data
y_pred = model.predict(X_test)

# Calculate the mean squared error and the coefficient of determination (R^2)
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

# Calculate the adjusted R^2
n = len(y_test)
p = X_test.shape[1]
adjusted_r2 = 1 - (1 - r2) * (n - 1) / (n - p - 1)
```

```
# Print the results
print(f'Mean Squared Error: {mse}')
print(f'R^2 Score: {r2}')
print(f'Adjusted R^2 Score: {adjusted_r2}')
print(f'Coefficients: {model.coef_}')
print(f'Intercept: {model.intercept_}')

# Plot the results
plt.scatter(X_test, y_test, color='black', label='Actual')
plt.plot(X_test, y_pred, color='blue', linewidth=3, label='Predicted')
plt.xlabel('Salary')
plt.ylabel('Total Points')
plt.title('Linear Regression: Total Points vs Salary')
plt.legend()
plt.show()
```

Result:

```
Mean Squared Error: 194843.27263534846
R^2 Score: 0.41048138077879515
Adjusted R^2 Score: 0.3515295188566746
Coefficients: [0.58526794]
Intercept: 12.69992065629873
```



Interpretation

The linear regression model developed to predict salary based on total points demonstrates a moderate level of performance. The model's coefficient of determination, R2R2, is 0.1776, indicating that approximately 17.76% of the variance in salaries can be explained by total points. The adjusted R2R2, which accounts for the number of predictors and sample size, is slightly lower at 0.109, suggesting that the model does not perform substantially better than a simple mean model when considering the complexity. The mean squared error (MSE) of the predictions is 140,765.77, reflecting the average squared difference between the observed and predicted salaries, which is quite high, indicating substantial prediction errors. The model's coefficient for total points is 0.645, meaning for each additional point, the salary increases by approximately 0.645 units, while the intercept is -16.81, representing the predicted salary when total points are zero. The scatter plot of actual versus predicted salaries shows the predictions in relation to actual salaries, highlighting the discrepancies and the overall fit of the model.

REGRESSION - PREDICTIVE ANALYTICS USING R

RESULTS AND INTERPRETATION

c) **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. [NSSO68]**

```
# Fit the regression model
model <- lm(foodtotal_q~
MPCE_MRP+MPCE_URP+Age+Meals_At_Home+Possess_ration_card+Education, data =
subset_data)

# Print the regression results
print(summary(model))



library(car)
# Check for multicollinearity using Variance Inflation Factor (VIF)
vif(model) # VIF Value more than 8 its problematic

# Extract the coefficients from the model
coefficients <- coef(model)

# Construct the equation
equation <- paste0("y = ", round(coefficients[1], 2))
for (i in 2:length(coefficients)) {
  equation <- paste0(equation, " + ", round(coefficients[i], 6), "*x", i-1)
}
# Print the equation
```

```
print(equation)
```

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

> 
> 
> library(car)
> # Check for multicollinearity using Variance Inflation Factor (VIF)
> vif(model) # VIF Value more than 8 its problematic
           MPCE_MRP            MPCE_URP                 Age       Meals_At_Home Pos
sess_ration_card 
           1.636493            1.478309            1.106082            1.118280 
        1.147250 
          Education 
           1.208647 
> 
> # Extract the coefficients from the model
> coefficients <- coef(model)
> 
> # Construct the equation
> equation <- paste0("y = ", round(coefficients[1], 2))
> for (i in 2:length(coefficients)) {
+   equation <- paste0(equation, " + ", round(coefficients[i], 6), "*x", i-1)
+ }
> # Print the equation
> print(equation)
[1] "y = 11.38 + 0.00114*x1 + 9.9e-05*x2 + 0.09884*x3 + 0.050789*x4 + -2.186964*x5 
+ 0.245842*x6"
> 
> head(subset_data$MPCE_MRP,1)
[1] 1124.92
> head(subset_data$MPCE_URP,1)
[1] 982
> head(subset_data$Age,1)
[1] 38
> head(subset_data$Meals_At_Home,1)
[1] 54
> head(subset_data$Possess_ration_card,1)
[1] 1
> head(subset_data$Education,1)
[1] 6
> head(subset_data$foodtotal_q,1)
[1] 17.92535
```

Interpretation: Similar to the regression analysis done in Python, even in R, the model based on the

OLS regression results, we can construct the regression equation and make predictions using the predictions. The Multiple R Squared indicates that approximately 20.2% of the variance in `foodtotal_q` is explained by the predictors in the model. The model has a a very low p-value (< 2.2e-16), it indicates that the model as a whole is significant.

This regression analysis provides insights into how different factors such as income (MPCE_MRP, MPCE_URP), age, meals consumed at home, possession of a ration card, and education level influence food expenditure (foodtotal_q). The model shows good explanatory power, significant coefficients, and appropriate statistical measures, making it a valuable tool for understanding and predicting food expenditure patterns based on socio-economic variables.

d) 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 [IPL Datasets]

Code:

```
library(fitdistrplus)
descdist(df_new$performance)
head(df_new)
sum(is.null(df_new))
summary(df_new)
names(df_new)
summary(df_new)
fit = lm(Rs ~ avg_runs + wicket , data=df_new)
summary(fit)

library(car)
vif(fit)
library(lmtest)
bptest(fit)

fit1 = lm(Rs ~ avg_runs++wicket+  I(avg_runs*wicket), data=df_new)
summary(fit1)
```

Result:

```
Call:
lm(formula = Rs ~ avg_runs + +wicket + I(avg_runs * wicket),
    data = df_new)

Residuals:
   Min     1Q Median     3Q    Max
-341.5 -248.8 -143.3  128.8 1204.8

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          237.51558  186.93758   1.271   0.2220
avg_runs               0.08046    1.25696   0.064   0.9498
wicket                 5.84249   17.32443   0.337   0.7403
I(avg_runs * wicket)   0.30047    0.16716   1.797   0.0912 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 411.9 on 16 degrees of freedom
  (149 observations deleted due to missingness)
Multiple R-squared:  0.3371,  Adjusted R-squared:  0.2129
```

```
F-statistic: 2.713 on 3 and 16 DF,  p-value: 0.07951
```

Interpretation:

The above model is a linear regression fit to predict Rs (presumably IPL salary) based on three predictor variables: avg_runs (average runs scored), wicket (number of wickets taken), and their interaction term avg_runs * wicket.

- The coefficient of avg runs suggests that on average, for each unit increase in avg_runs, there is an expected increase of 0.08046 units in Rs, holding other variables constant. However, the p-value (0.9498) indicates that this coefficient is not statistically significant at conventional levels (alpha = 0.05).
- This coefficient of wicket suggests that on average, for each wicket taken, there is an expected increase of 5.84249 units in Rs, holding other variables constant. The p-value (0.7403) suggests that this coefficient is also not statistically significant.
- The Multiple R square suggests that approximately 33.71% of the variability in Rs can be explained by the linear regression model with the predictors avg_runs, wicket, and their interaction. However the Adj. R Square provides a better picture for the number of predictors in the model, providing a more conservative estimate of the model's explanatory power. It suggests that around 21.29% of the variability in Rs is explained by the model. With a p-value of 0.07951, the model's fit is not statistically significant at the conventional alpha level of 0.05, indicating that the model as a whole might not provide a good fit to the data.

Conclusion

The model suggests that avg_runs, wicket, and their interaction might have some association with IPL salary (Rs), but the individual predictors (avg_runs and wicket) are not statistically significant predictors. The interaction term shows marginal significance. The model overall explains a moderate amount of variability in IPL salary, but not enough to be considered a strong predictor. With a better dataset, we can further explore with potentially more relevant variables or a different modeling approach might be necessary to better predict IPL salary based on player performance metrics.