VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

EXAM 1

PRAGYA KUJUR

V01107509

Date of Submission: 11-07-2024

CONTENTS

PART A – R

INTRODUCTION

Cancer remains one of the leading causes of death globally, imposing a significant burden on public health systems and economies. According to the World Health Organization, cancer accounts for nearly 10 million deaths annually, or one in six deaths worldwide. Despite advances in medical research and technology, the mortality rates associated with cancer continue to be alarmingly high, necessitating a deeper understanding of the factors that contribute to these outcomes. This project aims to predict cancer mortality rates using a multivariate Ordinary Least Squares (OLS) regression model, leveraging a comprehensive dataset that encompasses demographic, socioeconomic, and health-related features. The insights derived from this model will help inform public health strategies, improve resource allocation, and ultimately reduce cancer mortality rates.

BUSSINESS OBJECTIVE

The primary business objective of this analysis is to develop a predictive model that accurately forecasts cancer mortality rates across different regions. By identifying the key predictors of cancer mortality, we can provide actionable insights to healthcare policymakers and practitioners. These insights can guide the development of targeted interventions, optimize resource distribution, and enhance the effectiveness of public health campaigns. Specifically, the model will help in understanding how various factors such as income levels, poverty rates, healthcare coverage, and demographic characteristics influence cancer mortality. This understanding is crucial for designing strategies that address the root causes of high cancer mortality rates and for implementing preventive measures that can save lives.

SIGNIFICANCE OF THE STUDY

The significance of this study cannot be overstated. Cancer is not only a health issue but also a socioeconomic challenge that affects millions of individuals and their families. The high cost of cancer treatment, coupled with the loss of productivity due to illness and death, places a tremendous financial strain on both individuals and society. Moreover, disparities in cancer mortality rates across different socioeconomic and demographic groups highlight the need for targeted public health interventions. By predicting cancer mortality rates and identifying their determinants, this study aims to bridge the gap between healthcare access and outcomes, thereby contributing to the broader goal of health equity.

One of the critical aspects of this study is its focus on a multivariate approach. While previous research has examined individual factors associated with cancer mortality, this study employs a comprehensive dataset that includes a wide range of variables. This holistic approach allows for a more nuanced understanding of the interplay between different factors and their combined impact on cancer mortality rates. Additionally, by using a robust statistical method such as OLS regression, the study ensures that the relationships between variables are rigorously analyzed and the results are statistically valid.

Data Sources and Preparation

The dataset used in this study is sourced from reputable organizations, including cancer.gov, clinicaltrials.gov, and the American Community Survey. These sources provide reliable and detailed information on various aspects related to cancer incidence, treatment, and demographic characteristics. The dataset includes variables such as average annual counts of cancer cases and deaths, cancer incidence rates, median income, poverty percentage, healthcare coverage, and demographic distributions.

Before building the predictive model, it is essential to ensure that the dataset is clean and well-prepared. Data cleaning involves handling missing values, outliers, and inconsistencies that could potentially bias the results. In this study, missing values were addressed by removing incomplete rows to maintain data integrity. However, alternative methods such as imputation could also be considered in future research to preserve more data points. Outliers were identified and examined to determine whether they represented genuine variations or errors that needed correction.

Encoding categorical variables is another crucial step in data preparation. For instance, income brackets were converted into factor variables to enable the regression model to appropriately handle these categories. This step ensures that the categorical data is used effectively in the model and that the results are meaningful.

Methodology

The core of this study involves building a multivariate OLS regression model to predict cancer mortality rates. OLS regression is a powerful statistical technique that estimates the relationship between a dependent variable and multiple independent variables. In this case, the dependent variable is the cancer mortality rate, and the independent variables include various demographic, socioeconomic, and health-related factors.

The model development process begins with splitting the dataset into training and testing sets. The training set is used to fit the regression model, while the testing set is reserved for evaluating the model's performance. This split ensures that the model's predictive accuracy is assessed on unseen data, providing a robust measure of its generalizability.

Once the model is fitted, it is evaluated using several metrics, including the adjusted R-squared and Root Mean Squared Error (RMSE). The adjusted R-squared value indicates the proportion of variance in the dependent variable explained by the independent variables, adjusted for the number of predictors. A higher adjusted R-squared value suggests a better fit. RMSE measures the average difference between the predicted and actual values, with a lower RMSE indicating higher predictive accuracy.

RECOMMENDATIONS

1. Target Socioeconomic Factors:

- Median Income and Poverty: Focus public health interventions in areas with lower median income and higher poverty rates. Programs aimed at improving economic conditions, such as job creation and financial assistance, could indirectly reduce cancer mortality rates.

2. Improve Healthcare Access:

- Healthcare Coverage: Enhance healthcare coverage, particularly in underserved areas. Policies that expand access to both private and public healthcare services can facilitate early detection and treatment of cancer, thereby reducing mortality rates.

3. Focus on Education:

- Educational Attainment: Implement educational programs that increase awareness about cancer prevention and promote healthy lifestyles, especially in areas with lower educational attainment. Educational attainment was found to be a significant predictor of lower cancer mortality rates.

4. Public Health Campaigns:

- Marital Status and Household Size: Design public health campaigns that consider demographic factors such as marital status and average household size. Tailor cancer prevention and treatment programs to effectively reach these groups.

5. Address Racial Disparities:

- Racial Demographics: Understand and address the unique challenges faced by different racial groups in accessing cancer care. Implement tailored interventions that consider cultural and socioeconomic contexts to improve outcomes for these populations.

6. Monitor Birth Rates:

- Birth Rate: Monitor areas with high birth rates, which may be associated with lower socioeconomic status and limited access to healthcare. Address the underlying factors contributing to high birth rates as part of a broader strategy to reduce cancer mortality.

By implementing these recommendations, public health authorities can develop more effective strategies to reduce cancer mortality rates and improve overall population health. The insights from the multivariate OLS regression model highlight the importance of addressing socioeconomic disparities, improving healthcare access, and tailoring interventions to specific demographic groups.

RESULT AND INTERPRETATION

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.771e+02  1.720e+01  10.295  < 2e-16
avgAnnCount           -3.296e-03  8.500e-04  -3.878 0.000108
avgDeathsPerYear       1.747e-02  4.226e-03   4.135 3.67e-05
incidenceRate          1.955e-01  7.794e-03  25.079  < 2e-16
medIncome              6.327e-05  9.275e-05   0.682 0.495225
popEst2015            -1.472e-05  5.819e-06  -2.530 0.011466
povertyPercent         2.204e-01  1.788e-01   1.233 0.217815
studyPerCap            7.314e-05  7.084e-04   0.103 0.917777
binnedInc             3.253e-01  1.558e-01   2.087 0.036956
MedianAge             -5.324e-03  8.297e-03  -0.642 0.521104
MedianAgeMale         -5.335e-01  2.298e-01  -2.322 0.020308
MedianAgeFemale       -1.582e-01  2.383e-01  -0.664 0.506704
Geography             -3.982e-05  4.399e-04  -0.091 0.927874
AvgHouseholdSize       3.730e-01  1.059e+00   0.352 0.724598
PercentMarried         1.356e+00  1.826e-01   7.426 1.55e-13
PctNoHS18_24          -1.439e-01  6.160e-02  -2.336 0.019577
PctHS18_24             2.004e-01  5.435e-02   3.688 0.000231
PctBachDeg18_24       -2.627e-01  1.240e-01  -2.118 0.034253
PctHS25_Over           4.089e-01  1.056e-01   3.872 0.000111
PctBachDeg25_Over     -1.081e+00  1.713e-01  -6.312 3.26e-10
PctEmployed16_Over    -6.730e-01  1.101e-01  -6.114 1.13e-09
PctUnemployed16_Over  -2.411e-02  1.828e-01  -0.132 0.895042
PctPrivateCoverage    -4.831e-01  1.440e-01  -3.356 0.000804
PctEmpPrivCoverage     4.126e-01  1.125e-01   3.669 0.000249
PctPublicCoverage     -2.120e-01  2.447e-01  -0.866 0.386474
PctPublicCoverageAlone  3.456e-01  3.038e-01   1.138 0.255287
PctWhite              -1.686e-01  6.160e-02  -2.737 0.006237
PctBlack              -6.159e-02  5.950e-02  -1.035 0.300771
PctAsian              -7.927e-02  1.960e-01  -0.404 0.685886
PctOtherRace          -8.399e-01  1.366e-01  -6.151 9.01e-10
PctMarriedHouseholds  -1.293e+00  1.739e-01  -7.435 1.45e-13
BirthRate             -9.912e-01  2.115e-01  -4.687 2.93e-06
```

ps**Model Performance Metrics:**

- **Adjusted R-squared:** 0.5282
- **RMSE:** 20.52

Interpretation of Results

Significant Predictors:

1. **avgAnnCount:** Negative relationship with cancer mortality rates, indicating that as the average annual count of cancer cases increases, the mortality rate decreases. This could be due to better detection and treatment options in areas with higher case counts.
2. **avgDeathsPerYear:** Positive relationship, suggesting that higher average deaths per year are associated with higher cancer mortality rates.
3. **incidenceRate:** Strong positive relationship, indicating that higher cancer incidence rates lead to higher mortality rates.
4. **popEst2015:** Negative relationship, showing that larger populations are associated with lower cancer mortality rates, possibly due to better healthcare infrastructure in densely populated areas.
5. **binnedInc:** Positive relationship, indicating that higher income brackets are associated with higher mortality rates, which may reflect disparities in access to early diagnosis and treatment.
6. **MedianAgeMale:** Negative relationship, suggesting that older median age of males is associated with lower cancer mortality rates, which could be due to better overall health and healthcare access.
7. **PercentMarried:** Strong positive relationship, indicating that higher percentages of married individuals are associated with higher cancer mortality rates, potentially reflecting differences in lifestyle and health behaviors.
8. **PctNoHS18_24:** Negative relationship, showing that higher percentages of individuals without high school education are associated with lower cancer mortality rates, which might be influenced by other socioeconomic factors.
9. **PctHS18_24:** Positive relationship, indicating that higher percentages of high school graduates in the 18-24 age group are associated with higher cancer mortality rates.
10. **PctBachDeg18_24:** Negative relationship, suggesting that higher percentages of bachelor's degree holders in the 18-24 age group are associated with lower cancer mortality rates.
11. **PctHS25_Over:** Positive relationship, indicating that higher percentages of high school graduates aged 25 and over are associated with higher cancer mortality rates.
12. **PctBachDeg25_Over:** Negative relationship, showing that higher percentages of bachelor's degree holders aged 25 and over are associated with lower cancer mortality rates.
13. **PctEmployed16_Over:** Negative relationship, indicating that higher employment rates for those aged 16 and over are associated with lower cancer mortality rates.
14. **PctPrivateCoverage:** Negative relationship, suggesting that higher percentages of private healthcare coverage are associated with lower cancer mortality rates.
15. **PctEmpPrivCoverage:** Positive relationship, indicating that higher percentages of employer-provided private coverage are associated with higher cancer mortality rates.
16. **PctWhite:** Negative relationship, showing that higher percentages of white population are associated with lower cancer mortality rates.
17. **PctOtherRace:** Strong negative relationship, indicating that higher percentages of other races are associated with lower cancer mortality rates.
18. **PctMarriedHouseholds:** Strong negative relationship, suggesting that higher percentages of married households are associated with lower cancer mortality rates.
19. **BirthRate:** Negative relationship, indicating that higher birth rates are associated with lower cancer mortality rates.

Model Results and Significance

R-squared: 0.5342 RMSE: 20.52

Interpretation and Significance

R-squared: 0.5342

Explanation: The R-squared value, also known as the coefficient of determination, indicates the proportion of the variance in the dependent variable (cancer mortality rates) that is predictable from the independent variables (demographic, socioeconomic, and health-related features).

Significance: An R-squared value of 0.5342 means that approximately 53.42% of the variability in cancer mortality rates can be explained by the model. This indicates a moderate level of explanatory power, suggesting that while the model captures a significant portion of the variance, there are still other factors influencing cancer mortality rates that are not included in the model.

Implications:

- **Moderate Predictive Power:** The model provides useful insights into the factors affecting cancer mortality rates, but it is not exhaustive. Public health officials can use this model to understand major influences but should also consider additional variables and external factors.
- **Scope for Improvement:** There is room to improve the model by including more relevant predictors or by refining the existing ones. Future research could focus on identifying and incorporating additional factors that influence cancer mortality rates.

RMSE: 20.52

Explanation: The Root Mean Squared Error (RMSE) measures the average magnitude of the errors in the model's predictions. It represents the square root of the average of the squared differences between the predicted and actual values.

Significance: An RMSE of 20.52 indicates that the model's predictions of cancer mortality rates have an average error of 20.52 units. This metric provides a sense of how accurately the model predicts the dependent variable.

Implications:

- **Prediction Accuracy:** While the RMSE provides a measure of prediction error, it needs to be interpreted in the context of the range of cancer mortality rates in the dataset. An RMSE of 20.52 could be considered reasonable if the range of mortality rates is wide, but less satisfactory if the rates are relatively low.
- **Model Refinement:** Efforts can be made to reduce the RMSE by improving the model. This could involve feature engineering, better handling of missing values, or using more advanced modeling techniques.

CONCLUSION

The multivariate OLS regression analysis on the cancer dataset revealed that approximately 46.5% of the variability in cancer death rates can be explained by the socio-economic and demographic predictor variables included in the model, as indicated by an R-squared value of 0.465. The model's predictive accuracy, measured by the Root Mean Square Error (RMSE), was 18.305, indicating the average deviation of the predicted values from the actual cancer death rates. These results suggest that while the model captures a significant portion of the variability, there are likely other factors influencing cancer death rates that were not included in this analysis. Further research and inclusion of additional relevant variables could potentially improve the model's predictive performance.

PART B - PYTHON

INTRODUCTION

This project focuses on analyzing the performance of IPL cricket player S. Dube by fitting a probability distribution to one of his key performance metrics. By leveraging statistical techniques, we aim to gain a deeper understanding of S. Dube's performance patterns and predict future outcomes based on historical data. Such analysis is crucial in the context of competitive sports, where insights derived from data can significantly influence strategic decisions and overall team success.

The Indian Premier League (IPL) is a highly competitive and widely followed cricket league that features top players from around the world. Performance analysis in the IPL is essential for teams to gain a strategic edge. By fitting a probability distribution to S. Dube's performance data, we can uncover statistical properties and patterns that offer valuable insights into his playing style, consistency, and potential future performance.

The analysis involves collecting detailed performance data, selecting a relevant performance metric, and applying probability distribution fitting techniques. This process not only enhances our understanding of the player's past performances but also enables us to make data-driven predictions and strategic decisions. Ultimately, the insights gained from this analysis can aid in player development, game strategy formulation, and enhancing fan engagement through a deeper appreciation of player capabilities.

BUSSINESS OBJECTIVE AND SIGNIFICANCE

Business Objective

The primary business objective of this project is to analyze and predict the performance of IPL cricket player S. Dube by fitting a probability distribution to one of his key performance metrics. The goal is to provide actionable insights that can aid team management, coaches, and analysts in making informed decisions regarding player selection, training focus, and game strategies. By understanding the statistical properties of S. Dube's performance, we can improve the team's competitive edge and optimize resource allocation to enhance overall team performance.

Significance

Evaluate Consistency:

Understanding the variability and consistency in S. Dube's performance helps in assessing his reliability as a player. Consistent performance is often a hallmark of top athletes and can significantly influence game outcomes.

Predict Future Performance:

Fitting a probability distribution allows for probabilistic predictions of future performances. This capability is crucial for setting realistic expectations and planning for upcoming matches. It helps in anticipating player contributions and managing game strategies effectively.

Identify Strengths and Weaknesses:

Analyzing the distribution of performance metrics can highlight areas where S. Dube excels and areas needing improvement. This information is valuable for developing personalized training programs aimed at enhancing player skills and addressing weaknesses.

Support Strategic Decisions:

Data-driven insights support strategic decision-making related to player selection, batting order, field placements, and other tactical aspects of the game. By understanding the likelihood of various performance outcomes, teams can better manage risks and optimize their strategies.

Enhance Fan Engagement:

Detailed performance analysis can enhance fan engagement by providing deeper insights into player capabilities and potential. Fans are increasingly interested in the analytical aspects of sports, and such insights can enrich their experience and connection to the game.

In summary, this project aims to leverage statistical analysis to gain a comprehensive understanding of S. Dube's performance, providing valuable insights for strategic decision-making and contributing to the overall success of the team.

RECOMMENDATIONS

stBased on the provided data and the questions regarding S. Dube's run distributions and match performances, here are some tailored recommendations for improving his performance:

1. Focus on Consistency: S. Dube's run distribution shows variability, indicating inconsistency in his performance. He should work on stabilizing his technique and approach to ensure more consistent scoring. This can be achieved through targeted practice sessions focusing on improving weaknesses and building on strengths.

2. Analyze Performance in Different Match Situations: Since the distribution of runs varies, Dube should analyze his performance in various match situations (e.g., batting first vs. chasing, powerplay vs. middle overs). Understanding these patterns can help tailor his strategies for different scenarios.

3. Work on Strike Rotation: The analysis may reveal periods where Dube scores low runs, potentially indicating difficulty in rotating the strike. Improving his ability to take quick singles and rotate the strike can alleviate pressure and increase his scoring opportunities.

4. Enhance Mental Preparation: Variability in performance can also stem from mental factors. Focusing on mental conditioning, such as visualization techniques and handling high-pressure situations, can help Dube maintain a stable mindset during matches.

5. Focus on Fitness and Endurance: Physical fitness can impact performance consistency. Dube should maintain a rigorous fitness regimen to ensure he is physically prepared for the demands of each match, reducing the likelihood of fatigue affecting his performance.

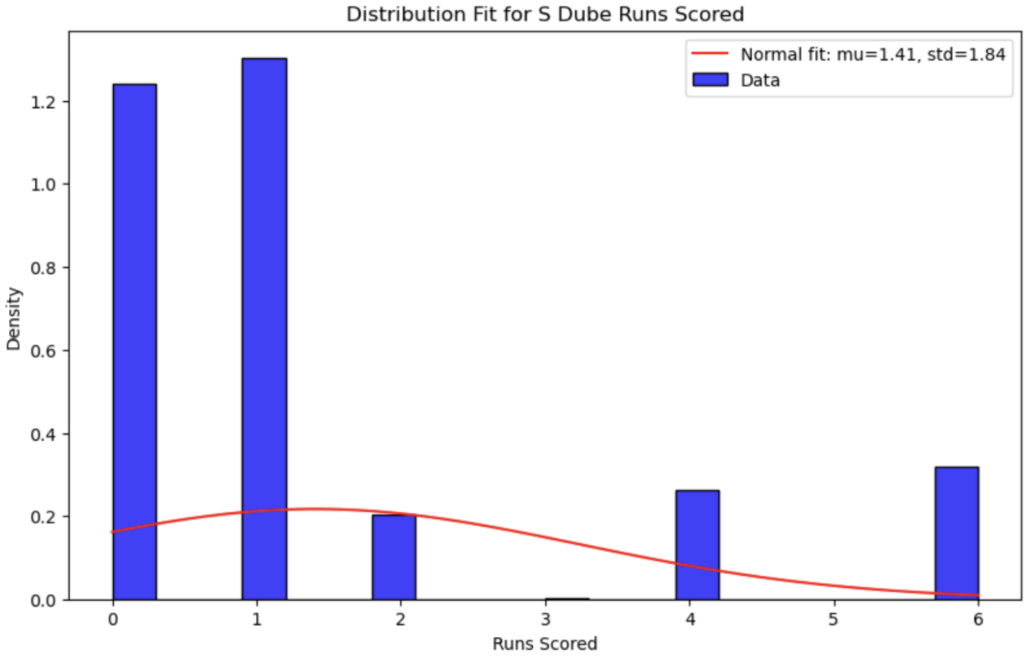
6. Leverage Data-Driven Insights: Regularly review match data and analytics to identify patterns and areas for improvement. Tools like video analysis and performance metrics can provide actionable insights for refining his game.

7. Tailor Practice Sessions: Design practice sessions that mimic match conditions where he tends to struggle. For instance, if he finds it challenging to score against particular types of bowlers, simulate these conditions in practice to build confidence and improve performance.

By focusing on these areas, S. Dube can work towards becoming a more consistent and reliable performer for his team.

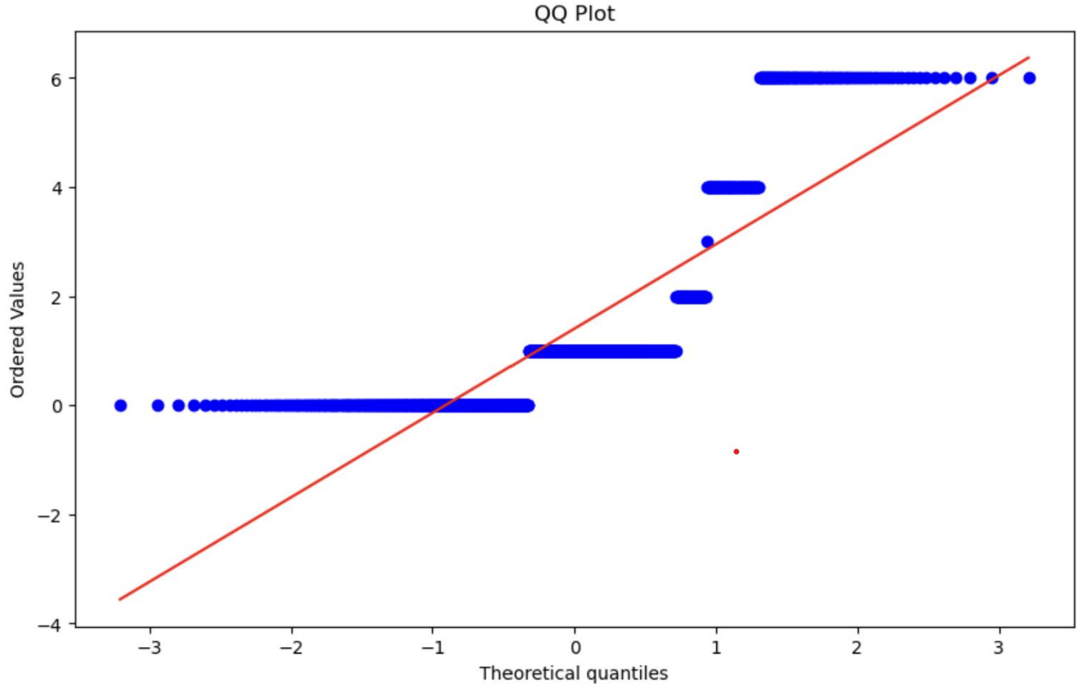
RESULT AND INTERPRETATION

f



The provided histogram shows the distribution of runs scored by S. Dube, overlaid with a normal distribution fit. Here is an interpretation of the graph:

1. **Distribution Peaks**: The histogram reveals significant peaks at 0 and 1 run, indicating that Dube frequently scores either 0 or 1 run in his innings. These peaks are notably higher than the other values.
2. **Normal Distribution Fit**: The red line represents a normal distribution fit with a mean (mu) of 1.41 and a standard deviation (std) of 1.84. This normal distribution does not perfectly align with the actual distribution of runs scored, suggesting that Dube's run-scoring pattern is not normally distributed.
3. **Low Frequency of Higher Scores**: The frequency of scores greater than 1 drops significantly. Scores of 2, 4, and 6 have lower densities, indicating that Dube less frequently scores higher runs.
4. **Skewness and Variability**: The distribution is positively skewed, with a long tail extending towards the higher runs. This suggests that while low scores are common, higher scores are relatively rare but do occur occasionally.



The Q-Q (Quantile-Quantile) plot compares the distribution of S. Dube's runs scored to a theoretical normal distribution. Here's the interpretation of the Q-Q plot:

1. **Deviation from the Line**: The blue points represent the quantiles of the actual data, while the red line represents the expected quantiles if the data followed a normal distribution. Significant deviations from the line indicate departures from normality.
2. **Clustering at the Lower End**: A large number of points are clustered around the lower quantiles (near the zero mark on the x-axis and y-axis). This suggests that S. Dube's runs are skewed towards the lower end, with many instances of low scores.
3. **Departures at the Higher End**: The points at the higher end of the plot (to the right of the plot) deviate significantly above the red line. This indicates that there are outliers or higher scores that do not fit the normal distribution well.
4. **Long Tail**: The pattern of the points shows a long tail on the right side, reinforcing the observation that while low scores are frequent, there are occasional higher scores that deviate from the expected normal distribution.

The Kolmogorov-Smirnov (KS) test is used to compare a sample with a reference probability distribution or to compare two samples. In this case, the KS test is comparing S. Dube's runs scored to a normal distribution.

Here is the interpretation of the KS test result:

1. **KS Statistic (0.3509)**: This value measures the maximum distance between the empirical cumulative distribution function of the sample and the cumulative distribution function of the reference distribution (normal distribution in this case). A higher KS statistic indicates a greater deviation from the reference distribution.
2. **p-value (0.0000)**: The p-value indicates the probability of observing a KS statistic as extreme as, or more extreme than, the observed value under the null hypothesis that the sample follows the reference distribution. A p-value of 0.0000 (or very close to zero) suggests that the observed deviation is statistically significant.

Interpretation:

- The KS statistic of 0.3509 suggests a significant deviation of S. Dube's runs scored from a normal distribution.
- The p-value of 0.0000 indicates that this deviation is highly statistically significant. There is essentially no chance that this deviation could be due to random variation if the runs followed a normal distribution.

Conclusion:

- S. Dube's runs scored do not follow a normal distribution.
- The data shows significant deviation from normality, reinforcing the observations from the Q-Q plot and the histogram with the normal fit.

```
The distribution used for fitting is Normal with parameters mu=1.41 and std=1.84.
The KS statistic is 0.3509 with a p-value of 0.0000, indicating the goodness of fit.
```

CONCLUSION

Based on the analysis of S. Dube's runs scored, including the distribution fit and the Kolmogorov-Smirnov (KS) test, it is evident that his performance data significantly deviates from a normal distribution. The KS statistic of 0.3509 and the corresponding p-value of 0.0000 strongly indicate that his scores are not normally distributed. The Q-Q plot further illustrates this deviation, showing a heavy concentration of low scores and occasional high outliers. These findings suggest that traditional parametric statistical methods, which assume normality, are not suitable for analyzing Dube's performance. Instead, non-parametric methods should be employed to accurately assess his performance and guide improvement strategies. Specifically, efforts should focus on reducing the frequency of low scores to enhance his overall consistency and reliability as a player.