CASE STUDY:

FITNESS ANALYSIS

Suppose that you would like to model a person's aerobic fitness as measured by the ability to consume oxygen. The data set analyzedin this example is called”Fitness”, and it contains measurements made on three groups of men involved in a physical fitness course at North Carolina State University.

The goal of the study is to predict fitness as measured by oxygen consumption. Thus, the response variable for the analysis is the variableoxygen.

Data Available to you:

Age in years, Weight in kilograms

Oxygen: Oxygen intake rate in millilitersper kilogram of body weight per minute

Runtime: Time taken to run 1.5 miles, in minutes

Runpulse: heart rate while running

Maxpulse: maximum heart rate recorded while running

Problem : Model person’s aerobic fitness in terms of Oxygen Intake.

Solution: Age ,Weight,Runtime,Runpulse,Maxpulse are five Independent variable and Oxygen is response variable.

Analytical Technique Used: Linear Regression Analysis as response variable is continueous in nature and relation between dependent varaiables and Independent variables is casual.

Data is already well prepared so skip the data exploration and Data preparation Step.

Predictive Model: Multiple Linear Regression analysis.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | SUMMARY OUTPUT | |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.920870139 |  |  |  |  |  |  |  |
| R Square | 0.848001814 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.817602176 |  |  |  |  |  |  |  |
| Standard Error | 2.275156705 |  |  |  |  |  |  |  |
| Observations | 31 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 5 | 721.9731 | 144.3946 | 27.89513 | 1.81E-09 |  |  |  |
| Residual | 25 | 129.4085 | 5.176338 |  |  |  |  |  |
| Total | 30 | 851.3815 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 102.2042752 | 11.97929 | 8.531748 | 7.13E-09 | 77.53247 | 126.8761 | 77.53247 | 126.8761 |
| Age | -0.219621383 | 0.095502 | -2.29964 | 0.030097 | -0.41631 | -0.02293 | -0.41631 | -0.02293 |
| Weight | -0.072302339 | 0.05331 | -1.35626 | 0.18714 | -0.1821 | 0.037492 | -0.1821 | 0.037492 |
| RunTime | -2.682522973 | 0.340985 | -7.86697 | 3.19E-08 | -3.3848 | -1.98025 | -3.3848 | -1.98025 |
| RunPulse | -0.373400848 | 0.117141 | -3.18762 | 0.00383 | -0.61466 | -0.13214 | -0.61466 | -0.13214 |
| MaxPulse | 0.304907835 | 0.133936 | 2.276512 | 0.03164 | 0.029061 | 0.580755 | 0.029061 | 0.580755 |

|  |  |  |
| --- | --- | --- |
| RESIDUAL OUTPUT | |  |
|  |  |  |
| *Observation* | *Predicted Oxygen* | *Residuals* |
| 1 | 44.59963286 | 0.009367 |
| 2 | 48.30746951 | -2.99447 |
| 3 | 56.10466181 | -1.80766 |
| 4 | 56.59616683 | 2.974833 |
| 5 | 51.10750594 | -1.23351 |
| 6 | 43.02974154 | 1.781258 |
| 7 | 45.03459965 | 0.6464 |
| 8 | 49.12834911 | -0.03735 |
| 9 | 40.25870879 | -0.81671 |
| 10 | 58.02381001 | 2.03119 |
| 11 | 48.57991059 | 1.961089 |
| 12 | 37.43723917 | -0.04924 |
| 13 | 45.63239675 | -0.8784 |
| 14 | 47.23854408 | 0.034456 |
| 15 | 46.47427894 | 5.380721 |
| 16 | 50.74318718 | -1.58719 |
| 17 | 46.30835137 | -5.47235 |
| 18 | 49.27885857 | -2.60686 |
| 19 | 47.05547255 | -0.28147 |
| 20 | 47.59134703 | 2.796653 |
| 21 | 39.19279998 | 0.2142 |
| 22 | 46.70083969 | -0.62084 |
| 23 | 48.81011404 | -3.36911 |
| 24 | 54.91522495 | -0.29022 |
| 25 | 44.63809955 | 0.4799 |
| 26 | 38.88156643 | 0.321434 |
| 27 | 45.45827894 | 0.331721 |
| 28 | 50.77416994 | -0.22917 |
| 29 | 48.57911218 | 0.093888 |
| 30 | 46.57127414 | 1.348726 |
| 31 | 45.59828787 | 1.868712 |

**COEFFICIENT**

We can see that Age ,weight ,Runtime, Runpulse all four variables have negative coefficient it shows if these variables will increase by Unit then oxygen intake will decrease by respective amount .

Maxpulse has positive coefficient it means unit increase in maxpulse lead 0.304907835 increase In oxygen intake.

**P-VALUE**

If we set the significance level=5% then only four variables are significant P-value and weight variable has insignificant P-Value.

Means to say age ,runtime, runpulse and maxpulse have impact on oxygen intake by one. Here Null Hypothesis is rejected.

**LINEAR EQUATION**

**Predicted Oxygen = 102.2042752 + -0.219621383\*Age + -0.072302339\*Weight +-2.682522973\*Runtime + 0.304907835\*maxpulse**

**Goodness Of fit Test**

**To measure how model is correct and efficient to predicted the future or forecast.**

1. **R^2 =**0.848001814

|  |
| --- |
|  |

1. **Adjusted R^2=** 0.817602176

**As R square have high value it shows model is quite good and efficient to predict.**

**Adjusted R^2 is less than R^2 because prior consider only significant variable and in this case weight is not significant.**

**C) Graph between Residual and dependent variable**

**It shows Residuals follow homoscedasticity as it does not follow any pattern**

**D) Fit chart-In line graph Actual Oxygen and predicted Oxygen overlapped each other It shows our model is good enough to predict accurate value.**

**E) Mean absolute percentage error(MAPE)**

**MAPE=100\*((ACTUALVALUE-FORECASTED VALUE)/ACTUAL VALUE))/n**

|  |  |
| --- | --- |
| ABS Residual | MAPE |
| 0.009367 | 0.020998 |
| 2.99447 | 6.608412 |
| 1.807662 | 3.329211 |
| 2.974833 | 4.993761 |
| 1.233506 | 2.473244 |
| 1.781258 | 3.975047 |
| 0.6464 | 1.415031 |
| 0.037349 | 0.076081 |
| 0.816709 | 2.070658 |
| 2.03119 | 3.382216 |
| 1.961089 | 3.880195 |
| 0.049239 | 0.131698 |
| 0.878397 | 1.962722 |
| 0.034456 | 0.072887 |
| 5.380721 | 10.37647 |
| 1.587187 | 3.228878 |
| 5.472351 | 13.4008 |
| 2.606859 | 5.585487 |
| 0.281473 | 0.601771 |
| 2.796653 | 5.550236 |
| 0.2142 | 0.543558 |
| 0.62084 | 1.347308 |
| 3.369114 | 7.41426 |
| 0.290225 | 0.531304 |
| 0.4799 | 1.063656 |
| 0.321434 | 0.819921 |
| 0.331721 | 0.72444 |
| 0.22917 | 0.453398 |
| 0.093888 | 0.192895 |
| 1.348726 | 2.814536 |
| 1.868712 | 3.936866 |
| MAPE | 2.999289 |

**Here MAPE is 3% which proves model is Good .**

MULTICOLLINEARITY

Correlation between different dependent variables.

|  |  |
| --- | --- |
| corelation(age,weight) | -0.23354 |
| corelation(age,Runtime) | 0.188745 |
| corelation(age,maxpluse) | -0.43292 |
| corelation(age,Runpluse) | -0.33787 |

**Sample is not following multicollinearity so it does not creates standard error that influence dependent variable value therefore we can say that out model is not biased.**