**# Name: Swapnil Ukey**

**# Id: 22220959**

**# Module: MS5108 Applied Customer Analytics**

***Text

Description automatically generatedIntroduction:***

The code provided generates two linear models that fit the relationship between two variables, x1 and x2, and a response variable, y. The x1 and x2 variables are generated using the R functions ‘rnorm’ and ‘rexp’ respectively, with n = 50, mean = 10, and sd = 10. The linear combination of x1 and x2 is calculated in y and the resulting data is then converted into a data frame. The two linear models, ‘fit\_plus’ and ‘fit\_min’, are fit using the R function ‘lm’.

***Text

Description automatically generatedLinear Model Fit:***

Text

Description automatically generatedChart, scatter chart

Description automatically generatedThe 'lm' function applies a linear regression model to the data, with y represented as a linear combination of the independent variables x1 and x2. The fit\_plus model includes both x1 + x2 as independent variables, whereas the fit\_min model includes difference between x1 and x2. The summary function is used to obtain summary statistics for each model, which includes information such as coefficients, p-values, residuals, and R-squared.

***Summary***

The model's residuals are listed in the first table, with the minimum, first quartile, median, third quartile, and maximum residuals listed. The residuals are the differences between the observed and predicted response values.

The coefficient estimates for the regression model are provided in the second table. The Estimate column contains the estimated regression coefficients, while the Std. Error column contains the estimate's standard error. The t-statistic for each coefficient is given in the t value column, and the p-value for the null hypothesis that the corresponding coefficient is equal to zero is given in the Pr(>|t|) column. The asterisks indicate the p-level value's of significance, with \*\*\* indicating a p-value less than 0.001, \*\* indicating a p-value less than 0.01, \* indicating a p-value less than 0.05, and. indicating a p-value less than 0.1.

The Residual standard error calculates 1.471e-15 on 47 degrees of freedom with the model's residual standard error, which is an estimate of the error variance. Multiple R-squared and Adjusted R-squared return the same value which is 1. The F-statistic and p-value provide the F-statistic and p-value for the model's overall significance.

***Introduction:***

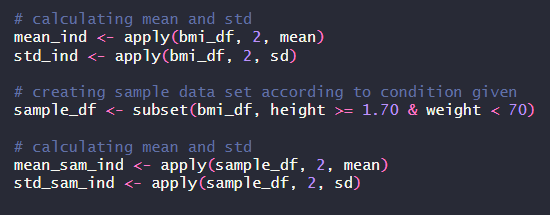
A screenshot of a computer

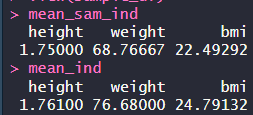
Description automatically generated with medium confidenceThe provided code computes the body mass index (BMI) for a sample of people based on their height and weight. The information is saved in a data frame called bmi\_df. The mean and standard deviation of each variable are computed for the entire sample as well as a subset of individuals with heights greater than or equal to 1.70 and weighing less than 70.

***Calculating BMI:***

The BMI of everyone is calculated using the formula ((weight / height) /2), where weight and height are in kilograms and meters respectively. The resulting values are stored as a new variable in the bmi\_df data frame.

***Mean and Standard Deviation:***

The mean and standard deviation of each variable are calculated using the apply function with mean and sd as arguments. These calculations are carried out for the entire sample as well as for the subset of individuals with a height greater than or equal to 1.70 and a weight less than 70.

***Comparison***

The mean height for the full data set is 1.761, while the mean height for the subset is 1.75. The mean weight for the total data set is 76.68, while the mean weight for the subset is 68.77. The mean BMI for the full data set is 24.79, while the mean BMI for the subset is 22.49.

Text

Description automatically generatedWe can see that the mean values of all three variables are lower in the subset of the data frame than in the complete data frame by comparing the mean values of these three variables between the two data sets.

For the overall data frame, the standard deviation of height is 0.157, while for the subset of the data frame, it is 0.0458. For the total data set, the standard deviation of weight is 11.79, while for the subset, it is 0.7094. For the whole data frame, the standard deviation of BMI is 2.63; for the subgroup, it is 1.39.