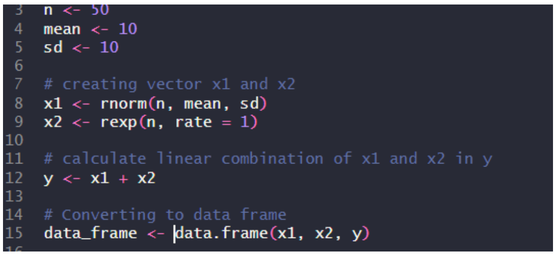
Name: Swapnil Ukey
Id: 22220959
Module: MS5108 Applied Customer Analytics

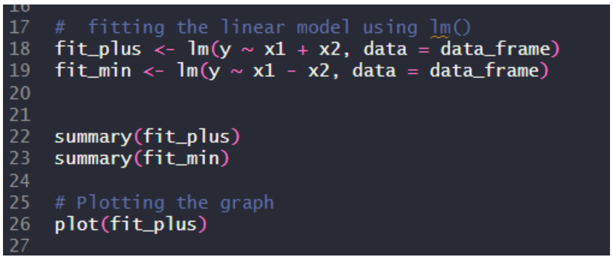
Introduction:



```
n <- 50
mean <- 10
sd <- 10

# creating vector x1 and x2
x1 <- rnorm(n, mean, sd)
x2 <- rexp(n, rate = 1)

# calculate linear combination of x1 and x2 in y
y <- x1 + x2

# Converting to data frame
data_frame <- data.frame(x1, x2, y)
```

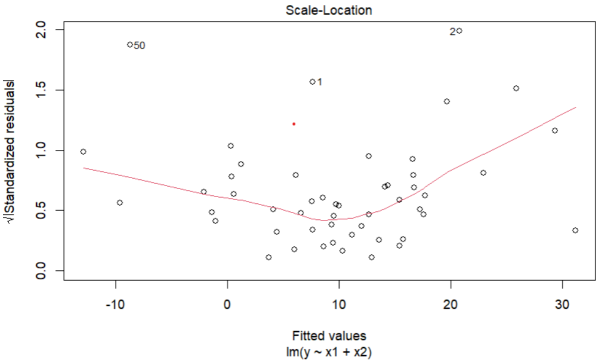
The code provided generates two linear models that fit the relationship between two variables, x1 and x2, and a response variable, y. The x1 and x2 variables are generated using the R functions ‘rnorm’ and ‘rexp’ respectively, with n = 50, mean = 10, and sd = 10. The linear combination of x1 and x2 is 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’.

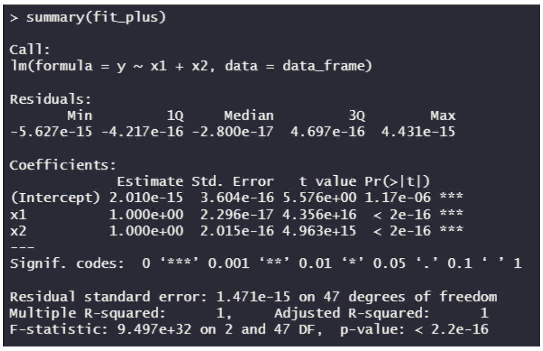
Linear Model Fit:



```
#  fitting the linear model using lm()
fit_plus <- lm(y ~ x1 + x2, data = data_frame)
fit_min <- lm(y ~ x1 - x2, data = data_frame)


summary(fit_plus)
summary(fit_min)

# Plotting the graph
plot(fit_plus)
```

The '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(fit_plus)

Call:
lm(formula = y ~ x1 + x2, data = data_frame)

Residuals:
       Min         1Q     Median         3Q        Max 
-5.627e-15 -4.217e-16 -2.800e-17  4.697e-16  4.431e-15 

Coefficients:
             Estimate Std. Error   t value Pr(>|t|)    
(Intercept) 2.010e-15  3.604e-16 5.576e+00 1.17e-06 ***
x1          1.000e+00  2.296e-17 4.356e+16  < 2e-16 ***
x2          1.000e+00  2.015e-16 4.963e+15  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.471e-15 on 47 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1 
F-statistic: 9.497e+32 on 2 and 47 DF,  p-value: < 2.2e-16
```

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.

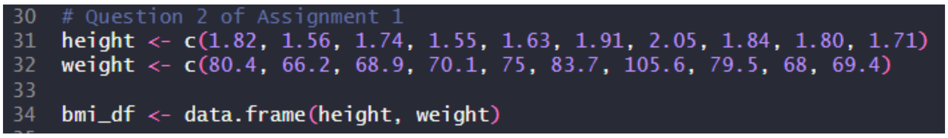
The coefficient estimates for the regression model are provided in the second table. The Estimate column contains the estimated regression coefficients, while the Std. Error column contains the estimate's standard error. The t-statistic for each coefficient is given in the t value column, and the p-value 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:

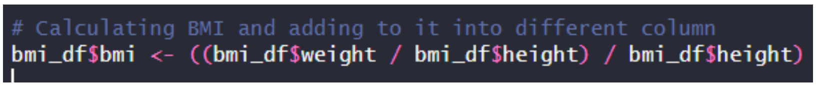
The provided code computes the body mass index (BMI) for a sample of people based on their height and weight. The information is saved in a data frame called bmi_df. The mean and standard deviation of each variable are computed for the entire sample as well as a subset of individuals with heights greater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



```
# Question 2 of Assignment 1
height <- c(1.82, 1.56, 1.74, 1.55, 1.63, 1.91, 2.05, 1.84, 1.80, 1.71)
weight <- c(80.4, 66.2, 68.9, 70.1, 75, 83.7, 105.6, 79.5, 68, 69.4)

bmi_df <- data.frame(height, weight)
```

Calculating BMI:

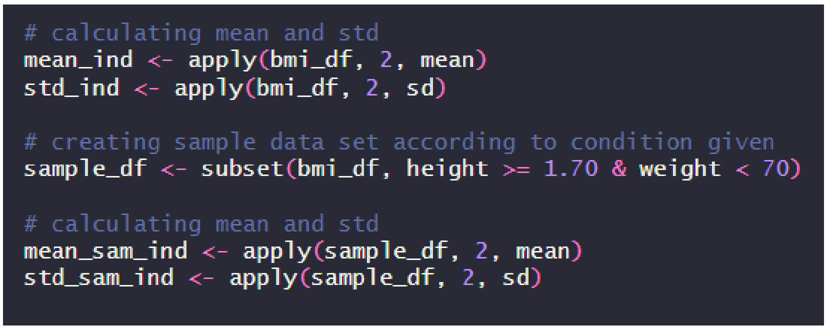
The BMI of everyone is calculated using the formula ((weight / height) /2), where weight and height are in kilograms and meters respectively. The resulting values are stored as a new variable in the bmi_df data frame.



```
# Calculating BMI and adding to it into different column
bmi_df$bmi <- ((bmi_df$weight / bmi_df$height) / bmi_df$height)
```

Mean and Standard Deviation:

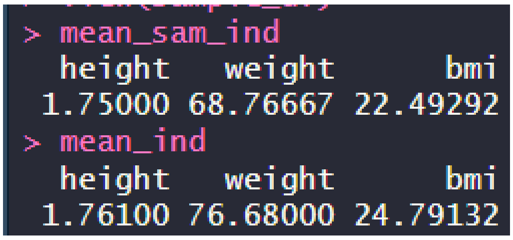
The mean and standard deviation of each variable are calculated using the apply function with mean and sd as arguments. These calculations are carried out for the entire sample as well as for the subset of individuals with a height greater than or equal to 1.70 and a weight less than 70.



```
# calculating mean and std
mean_ind <- apply(bmi_df, 2, mean)
std_ind <- apply(bmi_df, 2, sd)

# creating sample data set according to condition given
sample_df <- subset(bmi_df, height >= 1.70 & weight < 70)

# calculating mean and std
mean_sam_ind <- apply(sample_df, 2, mean)
std_sam_ind <- apply(sample_df, 2, sd)
```

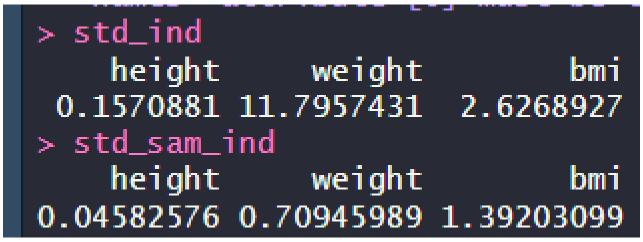
Comparison

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.



```
> mean_sam_ind
  height   weight      bmi 
 1.75000 68.76667 22.49292 
> mean_ind
  height   weight      bmi 
 1.76100 76.68000 24.79132 
```

We 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.



```
> std_ind
    height     weight        bmi 
 0.1570881 11.7957431  2.6268927 
> std_sam_ind
    height     weight        bmi 
0.04582576 0.70945989 1.39203099 
```

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.