[redacted]

IT3011 - Take home Assignment III
Regression Analysis.

Q1) Take data in the following table give the heart rate at rest (Y) and body weight (x) in kilograms.

X	Y
90	62
86	45
67	40
89	55
81	64
75	53

X [independent] = body.weight
Y [dependent] = heart.rate

a) Graph these data. Does it appear that there is a linear relationship between body weight and heart rate at rest?

For R studio, first insert above data into the CSV files and write below code to import the csv file.

```
HRBW <- read.csv("C:\\users\\MSI\\Desktop\\HR_BV.csv", header=TRUE)

attach(HRBW)
```

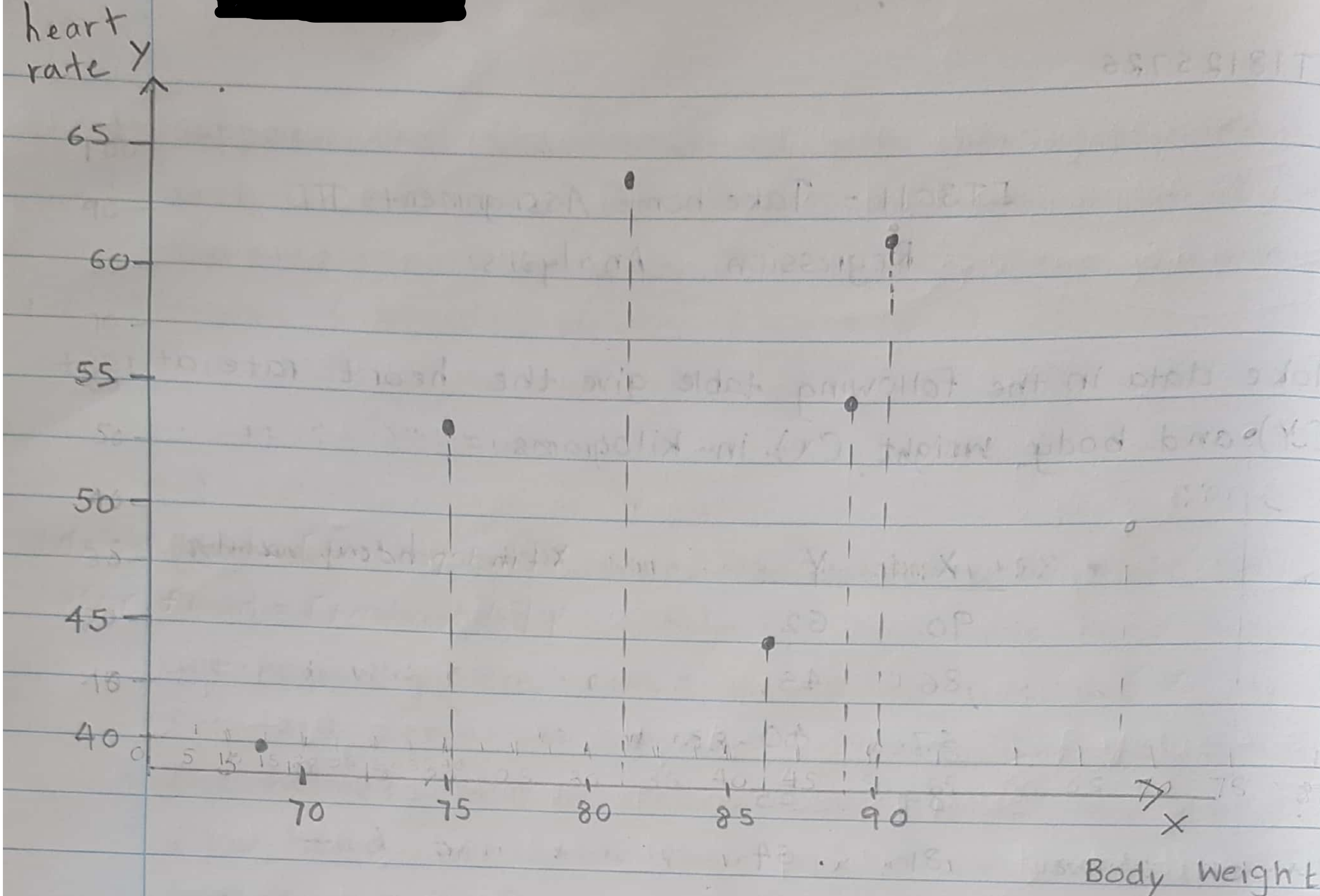
To view the scatter plot used below code

```
plot(body.weight, heart.rate, pch=21, col="blue", bg="red")
```

below shows the scatter plot corrosponding to the data given

T.N.P →

[redacted]



according to the graph x and y have a linear relationship and to check how strong their relationship we can use below R code.

cor (body.weight, heart.rate)

= 0.568705 [Moderate positive linear Relationship]

Further we can perform cor correlation test and get p value to check whether H_0 is rejected or not. for that we can use R code :-

cor.test (body.weight, heart.rate)

* From That we can get p-value as 0.2389. So at 5% significance level we do not reject H_0 ($H_0 \rightarrow \beta_1 = 0$)

∴ we can conclude the x and y have a relationship (Moderate positive linear)

[redacted]

b) Compute $\hat{\beta}_0$ and $\hat{\beta}_1$ and write the regression equation for these data. plot the regression line on the graph Obtained in part (a). Interpret the estimated regression Coefficients.

To compute $\hat{\beta}_0$ and $\hat{\beta}_1$ we have to build the model using R Studio

- HRBWmodel <- lm (heart.rate ~ body.weight)

To view the values corresponding to the $\hat{\beta}_0$ and $\hat{\beta}_1$:-

- Summary (HRBWmodel)

using above R code we can view $\hat{\beta}_0$ and $\hat{\beta}_1$ -

$\hat{\beta}_0 = 4.7990$

$\hat{\beta}_1 = 0.5947$

Regression equation $\Rightarrow$ $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$

$\hat{y} = 4.7990 + 0.5947 X$ //

To plot the regression line on graph used below R code:

abline (4.7990, 0.5947)

Interpret the estimated regression coefficients :-

- A Simple way to grasp regression coefficients is to picture them as linear slopes. The body weight coefficients in the regression equation is 0.5947. This coefficient represents the mean increase of heart rate in units. for every additional one unit in weight. If weight increases by 1 unite. the average heart rate increases by 0.5947 units.

[redacted]

Q1. d) Construct the ANOVA table and test the significance of the regression model at 5 % level of significance [clearly state the hypothesis that you are going to test]

anova (HRBWmodel)

	Df	Sum of sq	Mean sq	F value	Pr (>F)
body.weight	1	141.93	141.931	1.9122	0.2389
Residuals	4	296.90	74.225		

β_1 = body.weight

$H_0 \rightarrow \beta_1 = 0$ vs $H_1 \rightarrow \beta_1 \neq 0$ p value = 0.2389

Reject H_0 if α 0.05 > pvalue

∴ at 5 % significance level p value is greater than α value So we do not reject H_0. $\beta_1 = 0$. therefore our first assumption of the linear relationship between x and y is false

Q1. c) Obtain the residuals and test the model assumption.

Residuals → $e_1 = 3.679$, $e_2 = -10.942$, $e_3 = -4.643$

~~To test residuals~~ $e_4 = -2.726$, $e_5 = 11.032$, $e_6 = 3.600$

To test residuals → plot (HRBWmodel)

1st graph → Residuals vs fitted (dots are allover the graph and no pattern is visible so we can say X & Y are linearly related

2nd graph → Normal Q - Q (can't say exactly ∴ perform Shapilo. test)

3rd graph → Scale.Location (Check the assumption of constant varience) No pattern is visible so we can say There is no constent varience ∴ assumption is not violated)

4th graph → Residuals vs Leverage (Check the outliers) There are outliers in the obtained graph

→

[redacted]

shapiro.test (heart.rate)

⟹ p-value = 0.7116

H_0 = ε is normally distributed $\alpha = 0.05$

Reject H_0 if Pvalue < α at 5% Sig: level

∴ at 5% significance level we do not reject H_0. So we can conclude x and y are normaly distributed. assumption is not violated.

e) What can you say about the fitted model?

1) HRBWmodel ← lm (heart.rate ~ body.weight) from summary

Summary (BPBWmodel), r code

∴ x and y model is 0.76766 away from the fitted model.

2) using plot (body.weight, heart.rate, pch=21, col = "blue", bg="red") and abline (4.7990, 0.5947) we can compare fitted model and our graph,

3) Summary (HRBWmodel)

We can see the values of →

Coefficients :-

Intercept = 4.7990 Body.weight = 0.5947

∴ Model is fitted. So the accuracy level of the fitted model is $[R^2]$ = 0.3234 //

[redacted]

Q2) for a particular brand of automobile tire, an experiment was carried out to examine the relationship between temperature (X, in F) and tread wear of a tire (Y). The following data ware collected

X [independent] = Temperature Y [dependent] = Wear

a) plot the data and interpret the graph.

```
autoM <- read.csv("c:\\users\\MSI\\Desktop\\AutoMobile.csv",
                                              header = TRUE)
attach (autoM)
autoM
plot (Temperature, Wear, pch = 21, col = "blue", bg = "red")

cor (Temperature, Wear)
```

= -0.2907//

This graph interpret Weakly negative linear relationship

b) Compute the pearson Product Moment correlation coefficient and test its significance at 5% level. Interpret the results.

To get correlation coefficient used a r code below ->

```
cor.test (Temperature, Wear)
```

from this it gives the p value as = 0.2575. So at 5% Significence level α value = 0.05, So α < pvalue

We do not reject H_0

∴ We can Conclude that the temperature and wear does not have a linear relationship. So our first assumption of the linear relationship is false.

[redacted]

c) fit a simple linear regression model -

First we have to build the model in R Studio using below code.

```
AutoModel <- lm (Wear ~ Temperature)
Summary (AutoModel)
```

After executing above code we can have a value for $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\hat{\beta}_0 = 2.5678$$
$$\hat{\beta}_1 = -0.00427$$

equation for the simple linear regression model →

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$
$$\hat{y} = 2.5678 - 0.00427x$$ //

d) perform the residual analysis and state your comments on the regression assumptions.

To test residuals → plot (AutoModel)

1st graph → Residuals vs fitted (dots are allover the graph and no pttern is visible so we can say X & Y are linearly related

2nd graph → Normal Q-Q (Check Data is normaly distributed or not) can't exactly come to a conclution so perform Shapiro test

3rd graph → Scale Location (Check the assumption of constant varience) No pattern is visible so we can say There is no constent varience ∴ assumption is not violated.

T → N → P →

[redacted]

graph 04 – Residuals vs Leverage (Check whether there are any out liears)

There are outliears in the graphs which we obtained.

- shapiro. test (wear)

p-value = 0.2734

$H_0 \Rightarrow$ ~~d~~ ε is normally distributed $\alpha = 0.05$

Reject H_0 if p value < α at 5% sig. level

at 5% signicance level we do not reject H_0. So we can conclude Temperature and wear are normally distributed. assumption is not violated.

e) Constract the ANOVA table and test the Significance of the regression model at 5% level of significance. [clearly State the hypothesis that you are going to test]

anova (AutoModel)

Response : wear

	DF	Sum of sq	Mean sq	f value	Pr (>f)
Temperature	1	0.05495	0.05494		0.2575
Residuals	15	0.59495	0.03966		

$H_0 \rightarrow$ Temperature $(\beta_1) = 0$ p value = 0.2575

Reject H_0 if $\alpha = 0.05 >$ p value

Since pvalue (0.2575) > α (0.05) do not reject H_0 at 5% significant level. At 5% significant level we do donot reject H_0 hence we have enough evidence to conclude that $\beta = 0$

[redacted]

f) Perform a Lack of fit test and test Whether the Simple linear regression model is adequate or not

AutoModel ← lm (Wear ~ Temperature) [This is the reduced model]
Summary (AutoModel) ~~[This is the full model]~~
anova (AutoModel)
fac. Temp ← as.factor (Temperature)
Model.aov ← aov(wear ~ fac.Temp) [This is the full model]
Summary (Model.aov)
anova (Model.aov)
anova (AutoModel, Model.aov)

Need execute the above commands to get the anova table of "AutoModel" model and "Model.aov" model. And get the below out put

Model 1: Wear ~ Temperature
Model 2: Wear ~ fact.Temp

Res.DF	RSS	DF	Sum of sq	F	Pr(>F)
15	0.59495				
9	0.09778	6	0.49716	7.6265	0.003965

- Model 1 is the usual linear regression model; SSE(R) = 0.59495
- Model 2 is telling R to consider Temperature as a "factor" instead of a continuous variable.
- SSE(F) = 0.09778 = SSE(PE) [F = full, PE = Pure error]
- The Lack of fit SSE is SSE(LF) = SSE(R) - SSE(F)

$$= 0.59495 - 0.09778$$
$$= 0.49725$$

$$F^* = \frac{0.49725/6}{0.09778/9} = \frac{0.082875}{0.010865}$$
$$= 7.62770$$

Reject H_0 because the p-value = 0.003965. 0.01 = 0.0000 3965

because p value < f value //

[redacted]

3. Consider the following data set.

a) fit a multiple linear regression model.

- attach (Multi)
- Mmodel <- lm (Y ~ X1 + X2 + X3)
- Summary (Mmodel)

using Summary (Mmodel) R code we can find values for as below

$\hat{\beta}_0 = -4.5470$

$\beta_1 = 0.2589$

$\beta_2 = 0.5997$

$\beta_3 = 1.2035$

Regression equation for above data

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

$$= -4.5473 + 0.2589 x_1 + 0.5995 x_2 + 1.2035 x_3 \; //$$

b) According to the model you fitted, which predictor is the most influential in determinging the response variable.

Coefficients :-

	Estimate
(Intercept)	-4.5470
X_3	1.2035
X_2	0.5997
X_1	0.2589

The standardized coefficients show that X_3 has the standardized coefficient with largest absolute value, followed by X_2 and X_1. $\therefore X_3$ is the most influential in determining the response variable

[redacted]

c. compute the adjusted R^2 value and interpret the result.

```
Mmodel <- lm(Y ~ X1 + X2 + X3)
Summary (Mmodel)
```

Adjusted R - Squared : 0.9856 //

Multiple R^2 value = 0.9881 and Adjusted R^2 value = 0.9856
So Adjusted R^2 value decreases because of the predictor improves the model by less than expected.

d) Construct the ANOVA table and test the significance of the parameters.

```
- anova (Mmodel)
```

Analysis Of Variance Table

Source of variation	SS	df	mean Sum of squres	F-value
Regression	9011.7	$p-1=3$	$9011.7/3 = 3003.9$	$F = \frac{3003.9}{7.7214} = 389.0356$
Error / Residual	108.1	$n-p=14$	$108.1/14 = 7.7214$	
Total	9119.8	$n-1=17$		

Regression SS = 8176.0 + 317.8 + 517.9 = 9011.7

p (total no. of variables) = 3 independent + 1 dependent = 4

n = 18

[redacted]

In the above ANOVA table we can find that f value is 389.0356

$p-1 = 3 \longrightarrow$ degree of freedom 1

$n-p = 14 \longrightarrow df_2$

at 5% level of significance,

f - table value = 3.344

$f_{3,14} = 3.344$

Reject H_0 if $f_{cal} > f_{3,14}$

Since $389.0356 > 3.344$ $\therefore$ we can reject H_0 //

[redacted]

04) A student has fitted a multiple linear regression with 6 predictors using 100 observations. The ANOVA table constructed in the analysis is given below. Fill in the blanks of the table.

Source of variation	Degrees of freedom	Sum of squares	Mean Squares (MSR)	f
Regression	$P-1 = 5$	22364	4472.8	63.21130
Residual	$n-P = 94$	6651.4	70.7595	
Total	$n-1 = 99$	29015.4		

$p =$ no. of variables $= 6$

$n =$ no of observations $= 100$

1) $MSR = \dfrac{SSR}{P-1}$

$4472.8 \times 5 = SSR$

$SSR = 22364$ //

2) $Total = SSR + SSE$

$SSE = Total - SSR$

$= 29015.4 - 22364$

$= 6651.4$ //

3) $MSE = \dfrac{SSE}{n-p}$

$= \dfrac{6651.4}{100-6} = 70.759$

4) $f\text{-value} = \dfrac{MSR}{MSE}$

$= \dfrac{4472.8}{70.7595}$

$= 63.2113$ //