Question 1: (40 marks)

A medical research team wants to study the relationship between the **head circumferences of newborns (centimetres) - HC** and the **gestational age (weeks) - GA**. A simple random sample of 57 babies was selected from the records of babies born in a certain hospital, and their head circumferences (centimetres) and gestational age (weeks) were recorded.

In their attempt to find the relationship between head circumferences and gestational age, they plotted the head circumference versus the gestational age and obtained the following graph (Figure 1).

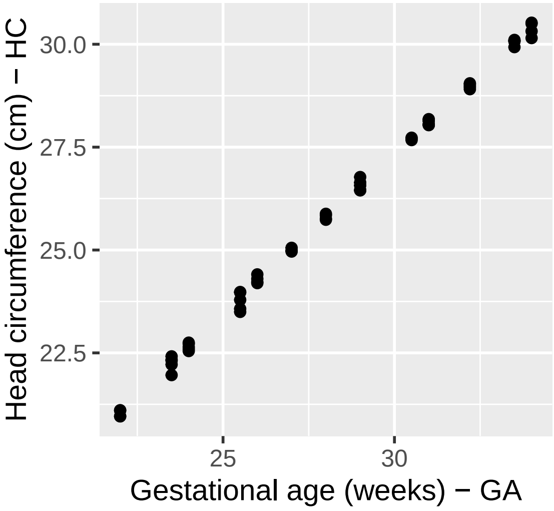


Figure 1: The scatter plot of head circumference versus the gestational age. The Pearson correlation coeffifient is 0.99

i) Comment on the scatter plot (Figure 1) given above.

ii) Write the model you would fit to these data. Define all terms in it and state any assumptions you make regarding the model.

A simple linear regression analysis was performed with these data and the following outputs were obtained using R software.

Output A

```
Call:
lm(formula = HC ~ GA)

Coefficients:
(Intercept)           GA
     3.9707       0.7781
```

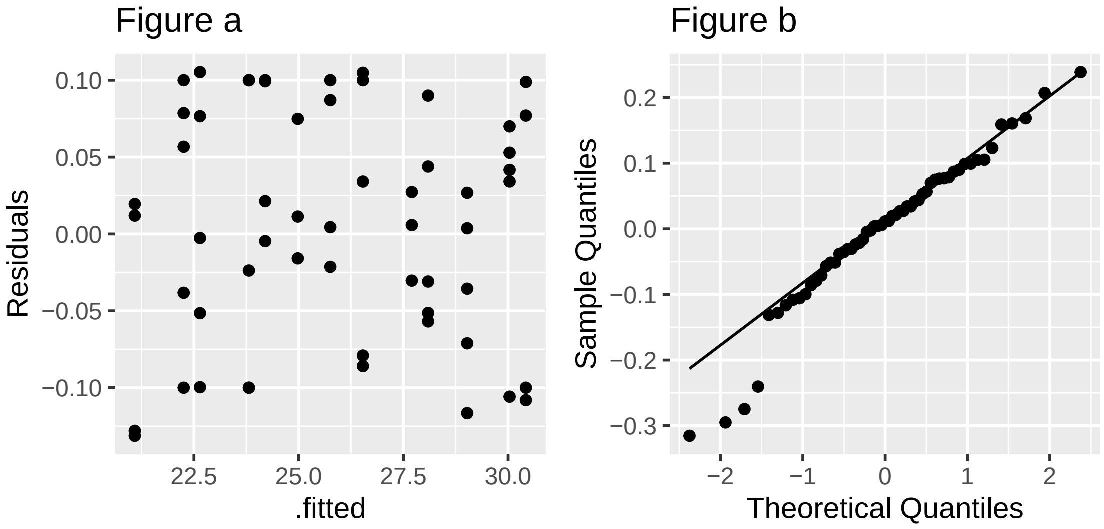
Output B

```
Call:
lm(formula = HC ~ GA)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.31535 -0.05153  0.01130  0.07656  0.23882 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.970711   0.113803   34.89   <2e-16 ***
GA          0.778069   0.004003  194.38   <2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.115 on 55 degrees of freedom
Multiple R-squared:  0.9985,    Adjusted R-squared:  0.9985 
F-statistic: 3.779e+04 on 1 and 55 DF,  p-value: < 2.2e-16
```

Output C

```
Analysis of Variance Table

Response: HC
          Df Sum Sq Mean Sq F value    Pr(>F)    
GA         1 500.02  500.02   37786 < 2.2e-16 ***
Residuals 55   0.73    0.01                      
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Output D



Output E

```
    Shapiro-Wilk normality test

data:  fitmodel$.resid
W = 0.95631, p-value = 0.03828
```

iii) Write the fitted regression model.

iv) Complete the ANOVA table below (Copy the table below in your answer script and complete it.).

Source of variation	DF	Sum of squares (SS)	Mean Square (MS)	F-value	p-value
Regression					
Residual error					
Total					

v) Test the significance of the model fitted. You should clearly write the hypothesis, decision and conclusions.

vi) What proportion of the variation in the response is explained by the model fitted?

vii) Two undergraduates studying Biostatistics were looking at this analysis.

Student A: said that the results strongly suggest that this model is highly significant and can be used for prediction purposes.

Student B: said that the results show the fitted model is not appropriate for this case and this model cannot be used for prediction.

With whom would you agree? Justify your argument using the results given above (Outputs A-E).

viii) The point on the graph (Figure 2, below), coloured in red, represents a baby with gestational age 31 weeks and head circumference 28.03 cm, (GA=31, HC=28.03).

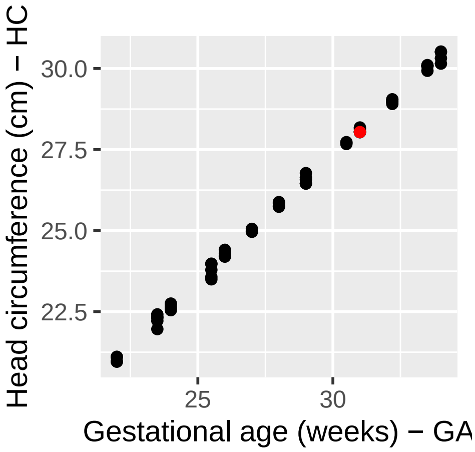


Figure 2: The scatterplot of head circumference versus the gestational age. The point, (GA=31, HC=28.03), is coloured in red on the graph.

a. Find the fitted value for this point.

b. Find the residual for this point.

ix) Later in the study, information regarding the mother's age (AGE) for each baby was obtained. A multiple linear regression model was fitted adding this variable to the earlier model. The R output results are given below. Assume all the assumptions of the multiple linear regression model are satisfied. Interpret the estimated values of the parameters of the model.

```
Call:
lm(formula = HC ~ GA + AGE, data = df1)

Coefficients:
(Intercept)           GA          AGE
    4.36628      0.77764     -0.01095
```

Question 2: (20 marks)

In a soap production factory, two machines: machine A and machine B, are used for the production. Using 32 soaps: 15 from machine A and 17 from machine B, the management wanted to find the relationship between the machine speed and the amount of scrap produced during the production process. To allow the two machines to have different regression lines with different intercepts and slopes the following model was fitted for all 32 observations.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

where,

$$X_1 \text{ is machine speed and}$$

$$X_2 = \begin{cases} 0 & \text{if machine A} \\ 1 & \text{if machine B} \end{cases} \tag{1}$$

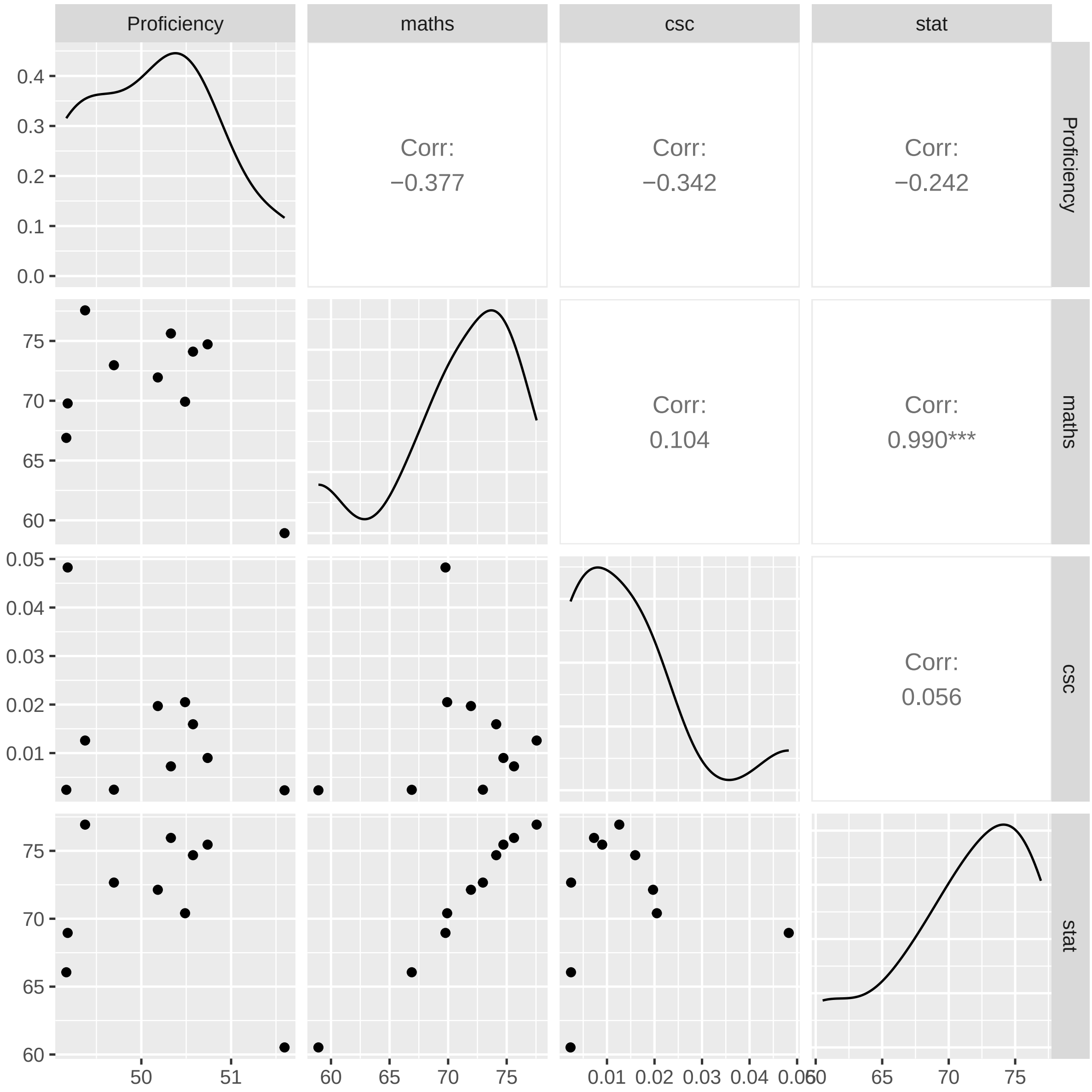
i) Write the regression model equations for each machine.

ii) Draw a sketch of the scatter plot which is expected with the above model along with model equations.

iii) Write the hypotheses that should be tested to find whether the two machines have the same regression model or not, i.e. whether both the intercept and the slope are the same of the two models you wrote in i) in the above.

Question 3: (20 marks)

A group of new graduates who has studied Statistics (stat), Mathematics (maths) and Computer Science (csc) at the Faculty of Applied Sciences, University of Jayewardenepura joined a company. They were given a test for each of the three subjects they have studied for the degree at the final interview from which they were selected for the job. After a probationary period of three months, their proficiency for the job was measured. The tests scores and the measure of proficiency were analysed to find a model to predict proficiency using the test scores. Some results are shown below.



```
model.sjp <- lm(Proficiency ~ maths + csc + stat, data=df)
summary(model.sjp)
```

```
Call:
lm(formula = Proficiency ~ maths + csc + stat, data = df)

Residuals:
       Min         1Q     Median         3Q        Max 
-5.915e-14 -3.388e-15  9.096e-15  1.398e-14  2.057e-14 

Coefficients:
              Estimate Std. Error   t value Pr(>|t|)    
(Intercept)  5.000e+01  1.450e-13  3.449e+14   <2e-16 ***
maths       -1.000e+00  1.397e-14 -7.161e+13   <2e-16 ***
csc          3.142e-13  7.648e-13  4.110e-01    0.695    
stat         1.000e+00  1.457e-14  6.862e+13   <2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
Residual standard error: 2.986e-14 on 6 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:      1
F-statistic: 2.051e+27 on 3 and 6 DF,  p-value: < 2.2e-16
```

```
car::vif(model.sjp)
```

```
    maths       csc      stat
56.313086  1.133357 55.885871
```

A statistician examined these results and claimed that “multicollinearity” has affected this model.

i) What is meant by multicollinearity?

ii) Do you agree with statistician’s claim. Justify your answer.

The following outputs are obtained using R programming software.

```
library(broom)
as.data.frame(augment(model.sjp))
```

```
   Proficiency    maths         csc     stat  .fitted       .resid      .hat
1     49.37355 77.55891 0.012586364 76.93245 49.37355 4.263256e-14 0.3124392
2     50.18364 71.94922 0.019694046 72.13286 50.18364 7.105427e-15 0.1274917
3     49.16437 66.89380 0.002428445 66.05817 49.16437 3.552714e-14 0.6960302
4     51.59528 58.92650 0.002329921 60.52178 51.59528 0.000000e+00 0.8287169
5     50.32951 75.62465 0.007267810 75.95416 50.32951 1.421085e-14 0.2331348
6     49.17953 69.77533 0.048249476 68.95486 49.17953 2.131628e-14 0.8543802
7     50.48743 69.91905 0.020492701 70.40648 50.48743 2.131628e-14 0.1719979
8     50.73832 74.71918 0.008994714 75.45751 50.73832 2.842171e-14 0.2826130
9     50.57578 74.10611 0.015942792 74.68189 50.57578 1.421085e-14 0.2243034
10    49.69461 72.96951 0.002450767 72.66412 49.69461 3.552714e-14 0.2688927
         .sigma     .cooksd .std.resid
1  7.209302e-15 0.648504431 -2.3892369
2  3.255938e-14 0.001948060 -0.2309271
3  3.041727e-14 0.463718803  0.9000332
4  2.142187e-14 4.143791868 -1.8509010
5  3.143744e-14 0.034663857  0.6753431
6  3.199923e-14 0.375759278  0.5061387
7  3.243911e-14 0.005046529  0.3117314
8  3.214039e-14 0.020228372  0.4532013
9  3.233028e-14 0.009883511  0.3697546
10 3.088512e-14 0.059690137  0.8057166
```

iii) Are there any observations that have high leverage values? If so, what are their observation numbers.

Question 4: (20 marks)

i) It was revealed that $\beta_1 = 0$ for a simple linear regression model between the variables X and Y, and therefore, now the model is $Y = \beta_0 + \epsilon$. Draw a **sketch** of the scatter plot for this relationship between X and Y.

The Consumer Affairs Authority (CAA) issued a special gazette notification last September setting a maximum retail price for coconut based on the circumference of coconut due to the high prices in the market. An investigator wants to study how the **circumference** of coconut related to **weight** of the coconut. A simple linear regression model was fitted to the data and the R output is shown below.

```
Call:
lm(formula = weight ~ circumference, data = coconut)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4083 -0.9343 -0.1721  0.7014  4.6157 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -897.2460     4.8095  -186.6   <2e-16 ***
circumference   59.9412     0.1607   373.1   <2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 1.521 on 98 degrees of freedom
Multiple R-squared:  0.9993,    Adjusted R-squared:  0.9993 
F-statistic: 1.392e+05 on 1 and 98 DF,  p-value: < 2.2e-16
```

Figure 4.a

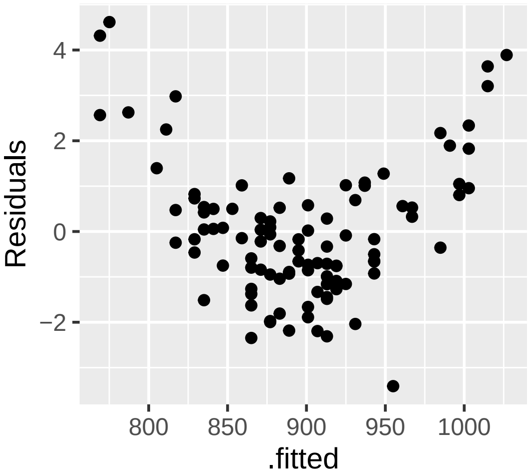
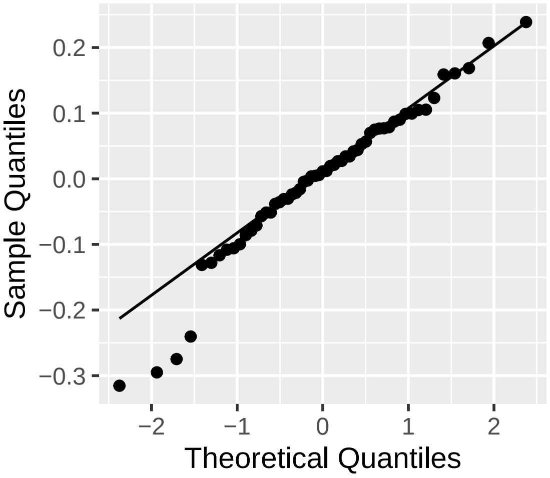


Figure 4.b



```

	Shapiro-Wilk normality test

data:  fitmodel.coco$.resid
W = 0.95463, p-value = 0.001697
```

ii) Are you satisfied with the fitted model? If your answer is “Yes”, write the reasons and give all possible evidence to justify your answer. If your answer is “No”, write the reasons and suggest possible ways to improve the fit of the simple linear regression model.