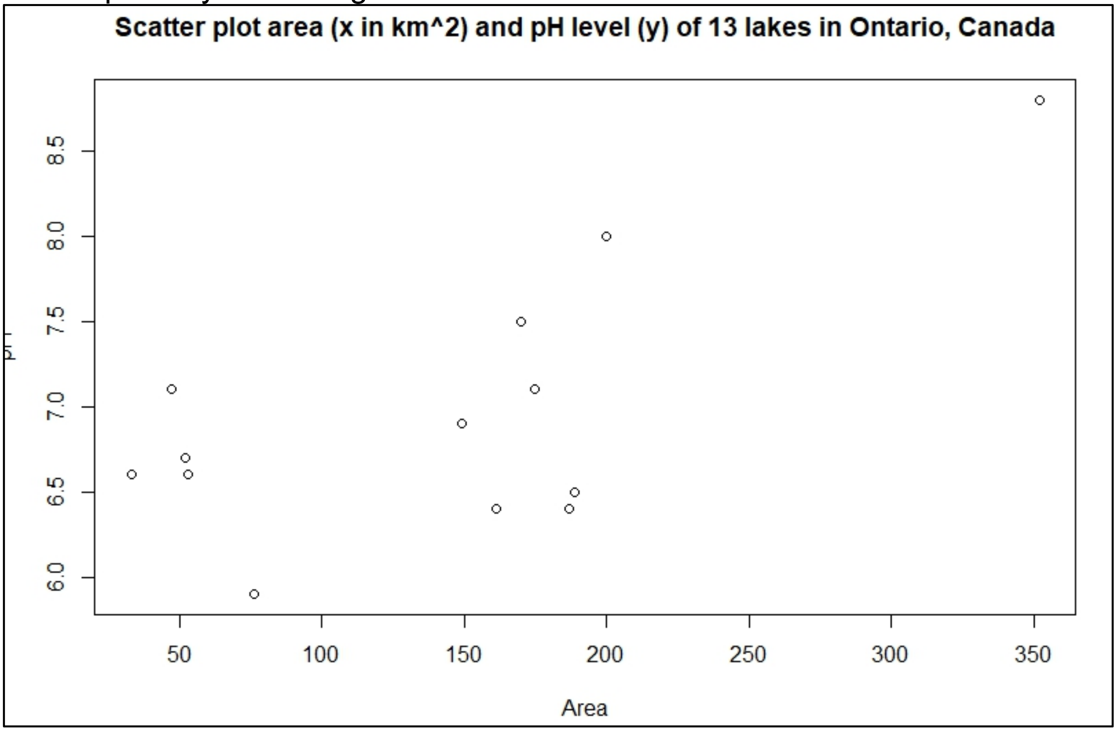
QUESTION 1:

1 A:

Scatterplot of y vs x using R



For displaying the scatterplot in R created a dataframe called Q1 and code used to plot the scatterplot using R is mentioned below:

```
# Q1 contains all data mentioned in the question 1
Q1 = data.frame(Area = c( 33, 161, 189, 149, 47, 170, 352, 187, 76, 52, 175, 53,
200) , pH = c( 6.6, 6.4, 6.5, 6.9, 7.1, 7.5, 8.8, 6.4, 5.9, 6.7, 7.1, 6.6, 8.0))
Q1

x = Q1[,1]
y = Q1[,2]

n = length(x)
plot(x,y,xlab="Area", ylab="pH", main="Scatter plot area (x in km^2) and pH level (y)
of 13 lakes in Ontario, Canada")
```

Comment:

- From the scatterplot between the Area and pH value of 13 lakes its evident that the pH value is not that much dependant on the Area of lakes.The pH value is from 5.9 to 8.8 where value can round to 6.5.
- In general if the pH value is less than 7 is considered as acidic and pH greater than 7 is considered as alkaline.The lakes which are having less area are having pH values less than 7 which is acidic and higher area lakes water is acidic as well as basic.
- The highest area value lakes (200 and 352) which are having pH value 8 and 8.8, are alkaline and they are having exceptionally higher value.

1 B:

Calculation by hand:

For applying the principal of least squares to fit the simple linear regression model to the data the pH is the dependent variable (y axis) and area is the independent variable (x axis). The calculations required are:

No	Area(x)	pH(y)	Area*pH(x*y)	Area^2 (x^2)	pH^2 (y^2)
1	33	6.6	217.8	1089	43.56
2	161	6.4	1030.4	25921	40.96
3	189	6.5	1228.5	35721	42.25
4	149	6.9	1028.1	22201	47.61
5	47	7.1	333.7	2209	50.41
6	170	7.5	1275	28900	56.25
7	352	8.8	3097.6	123904	77.44
8	187	6.4	1196.8	34969	40.96
9	76	5.9	448.4	5776	34.81
10	52	6.7	348.4	2704	44.89
11	175	7.1	1242.5	30625	50.41
12	53	6.6	349.8	2809	43.56
13	200	8	1600	40000	64
Total	1844	90.5	13397	356828	637.11

From the above manual calculation,

Here number of observations are 13, so n = 13
Sum(x) = 1844
Sum(y) = 90.5
Sum(x^2) = 356828
Sum(y^2) = 637.11
Sum(x*y) = 13397

S_xx = Sum(x^2) - ((Sum(x)^2) / n)
= 356828 - ((1844 ^2) / 13) = 95263.69

S_yy = Sum(y^2) - ((Sum(y)^2) / n)
= 637.11 - ((9.5^2)/ 13) = 7.09

S_xy = Sum(x*y) - ((Sum(x) * Sum(y)) / n)
= 13397- ((1844 * 90.5) / 13) = 559.92

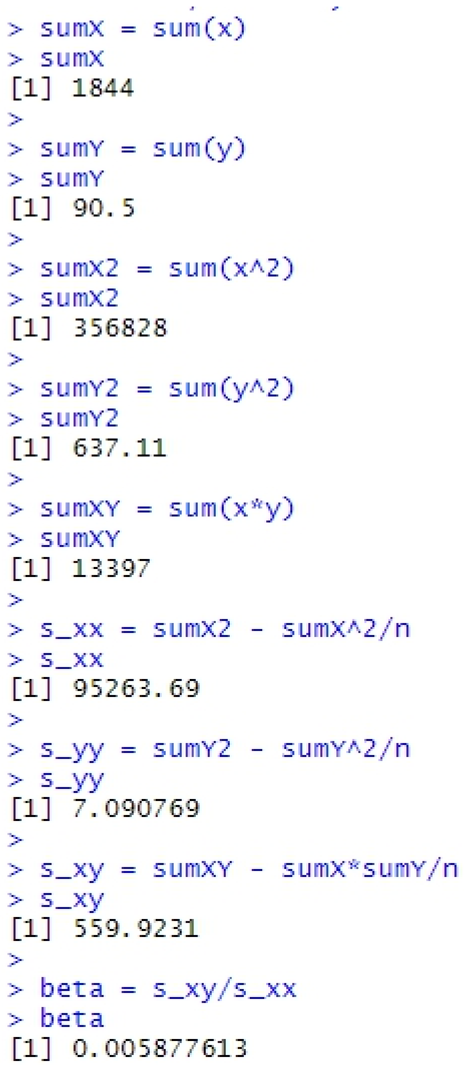
Slope or Beta = S_xy / S_xx = 559.92 / 95263.69 = 0.0059

Alpha = Mean(y) -(Beta * Mean(x))

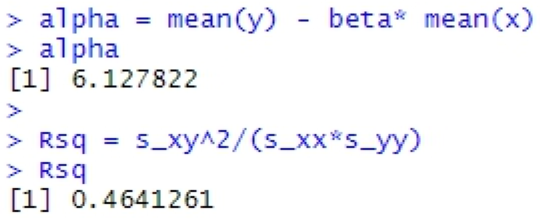
Mean(y) = 6.9620
Mean(x) = 141.85
Intercept / Alpha = 6.9620 - (0.0059 * (1844/13)) = 6.125

We can write the regression line as y = 6.125 + 0.0059x , where Y which is the dependant or response variable is pH value . X which is the independant (regressor/predictor) variable is Area.

Following codes are just to verify the calculations done by hand in the R, inorder to fit the regression

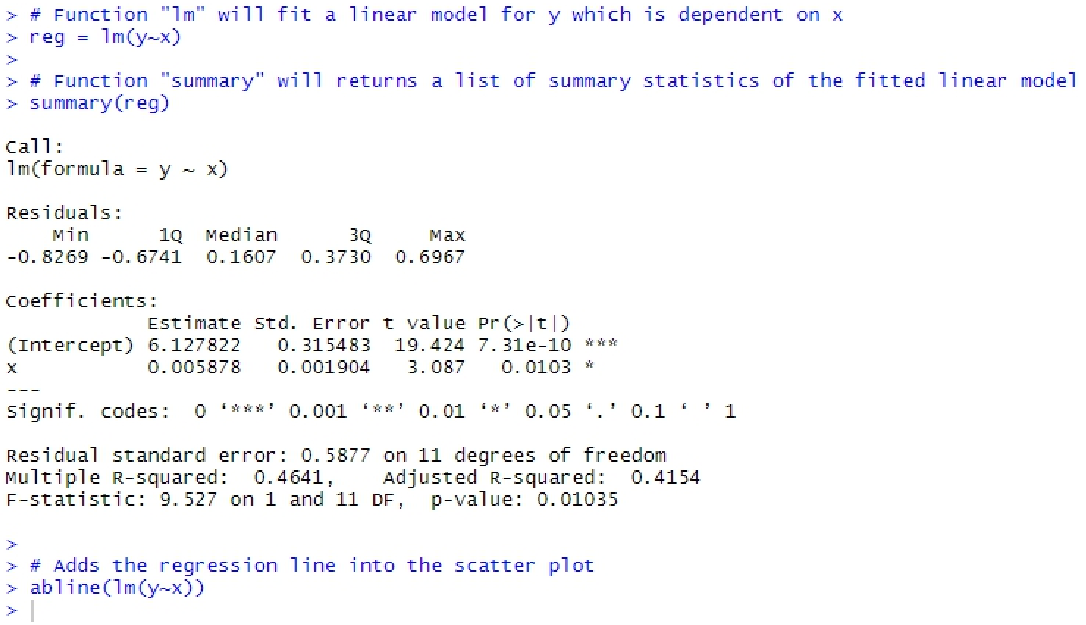


```
> sumX = sum(x)
> sumX
[1] 1844
>
> sumY = sum(y)
> sumY
[1] 90.5
>
> sumX2 = sum(x^2)
> sumX2
[1] 356828
>
> sumY2 = sum(y^2)
> sumY2
[1] 637.11
>
> sumXY = sum(x*y)
> sumXY
[1] 13397
>
> s_xx = sumX2 - sumX^2/n
> s_xx
[1] 95263.69
>
> s_yy = sumY2 - sumY^2/n
> s_yy
[1] 7.090769
>
> s_xy = sumXY - sumX*sumY/n
> s_xy
[1] 559.9231
>
> beta = s_xy/s_xx
> beta
[1] 0.005877613
```



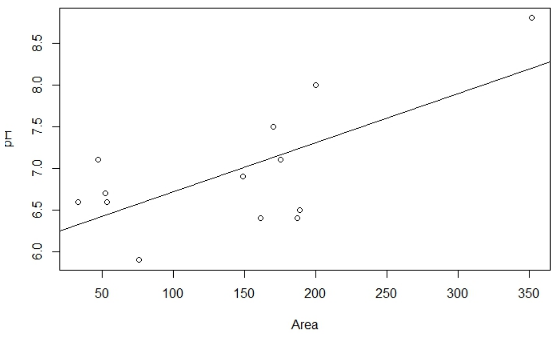
```
> alpha = mean(y) - beta* mean(x)
> alpha
[1] 6.127822
>
> Rsq = s_xy^2/(s_xx*s_yy)
> Rsq
[1] 0.4641261
```

That is 46.4% of the variation in pH value is explained by linear variations in the Area.

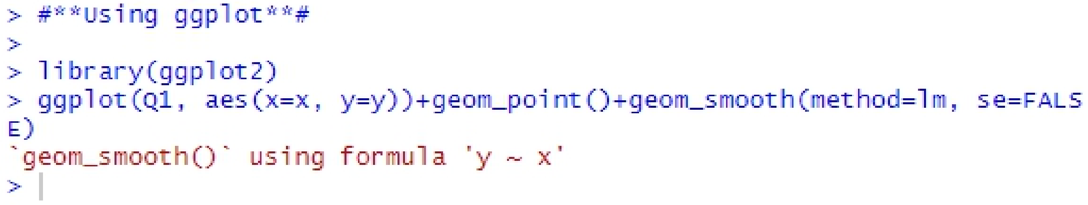


```
> # Function "lm" will fit a linear model for y which is dependent on x
> reg = lm(y~x)
> 
> # Function "summary" will returns a list of summary statistics of the fitted linear model
> summary(reg)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.8269 -0.6741  0.1607  0.3730  0.6967 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.127822   0.315483  19.424 7.31e-10 ***
x           0.005878   0.001904   3.087   0.0103 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5877 on 11 degrees of freedom
Multiple R-squared:  0.4641,	Adjusted R-squared:  0.4154 
F-statistic: 9.527 on 1 and 11 DF,  p-value: 0.01035

> 
> # Adds the regression line into the scatter plot
> abline(lm(y~x))
> 
```

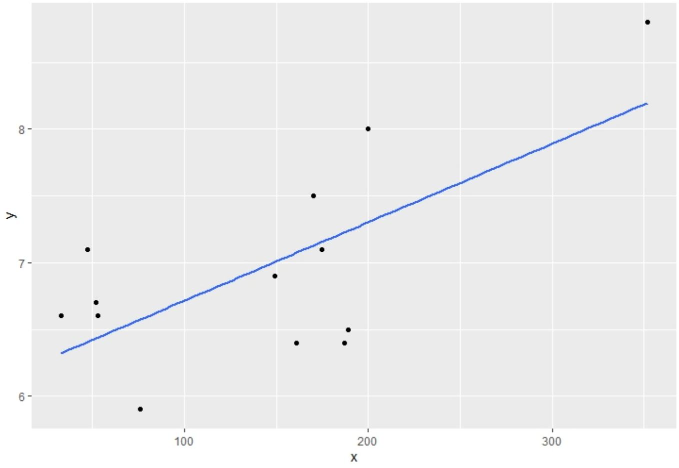
Scatter plot area (x in km^2) and pH level (y) of 13 lakes in Ontario, Canada



Using the ggplot



```
> #**Using ggplot**#
> 
> library(ggplot2)
> ggplot(Q1, aes(x=x, y=y))+geom_point()+geom_smooth(method=lm, se=FALS
E)
`geom_smooth()` using formula 'y ~ x'
> 
```



1. C

Hand calculations to perform the ANOVA Test to deduce whether there is a linear relationship between area and pH level:

To perform the ANOVA Test the following table has to be calculated:

	DF	Sum of Squares	MS	F
Regression	1	SSR	MSR=SSR/1	F = MSR / MSE
Residual	n - 2	SSE	MSE=SSE/(n−2)	
Total	n - 1	SST		

Considering the above table we need to find the SSR, SSE, SST.
SSR = Square(Difference of estimated y and average of y)
SSE = Square(Difference of y and estimated y value)
SST = Square(Difference of y and average of y)

For calculating those values the below table has to be calculated:

No	Area (x axis)	ph (y axis)	xy	x^2	y^2	Estimated y	(y-Estimated y)^2	(Estimated y - y average)^2	(y- y average)^2
1	33	6.6	217.8	1089	43.56	6.32	0.08	0.41	0.13
2	161	6.4	1030	25921	40.96	7.07	0.46	0.01	0.32
3	189	6.5	1229	35721	42.25	7.24	0.55	0.08	0.21
4	149	6.9	1028	22201	47.61	7.00	0.01	0.00	0.00
5	47	7.1	333.7	2209	50.41	6.40	0.48	0.31	0.02
6	170	7.5	1275	28900	56.25	7.13	0.14	0.03	0.29
7	352	8.8	3098	123904	77.44	8.20	0.36	1.53	3.38
8	187	6.4	1197	34969	40.96	7.23	0.68	0.07	0.32
9	76	5.9	448.4	5776	34.81	6.57	0.46	0.15	1.13
10	52	6.7	348.4	2704	44.89	6.43	0.07	0.28	0.07
11	175	7.1	1243	30625	50.41	7.16	0	0.04	0.02
12	53	6.6	349.8	2809	43.56	6.44	0.03	0.27	0.13
13	200	8	1600	40000	64	7.30	0.48	0.12	1.08
Sum	1844	90.5	13397	356828	637.1	90.51	3.8	3.29	7.09

As per the above table the values the first table can be tabulate as:
(The values are not rounded off here as it have to be cross verify with output from the R)

	DF	SS	MS	F
Regression	1	3.291011	3.291011	9.527217
Residual	11	3.799758	0.345433	
Total	12	7.090769		

To test whether there exists linear relationship between the variables Area and pH value we have to test the hypothesis as mentioned below,

Null hypothesis H0 : Beta is equal to 0
Alternative hypothesis H1 : Beta is not equal to 0

F-Test for overall goodness of fit:

The F test statistic is defined as mean square due to regression/MSE.It is used to test the null hypothes is

H0= A linear relationship does not exist between y and x
Against the alternative
H1= A linear relationship does exist between y and x.

As per method of rejection points in the F test;

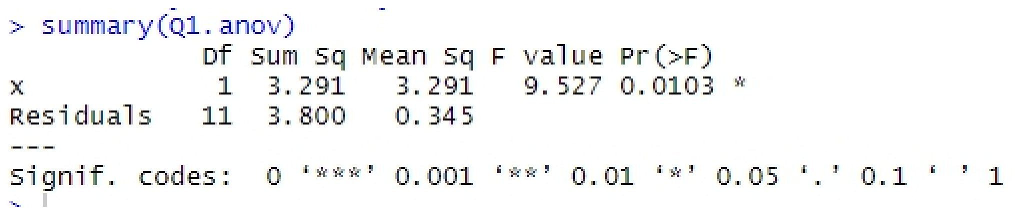
If Beta = 0 then F = 1 which means there is no linear relationship.

If F > 1 then b ≠ 0, which means there is a linear relationship

The F test statistic is defined as Mean Square due to Regression / MSE and used to test the null hypothesis.If the Beta is zero then F would be 1.A large value of F indicates that regression is significant.Here the value of F is 9.527217 which is greater than 1 indicates Beta value is not equal to 0.Hence we can say that they are having linear relationship.

ANOVA plot in R for analysis:

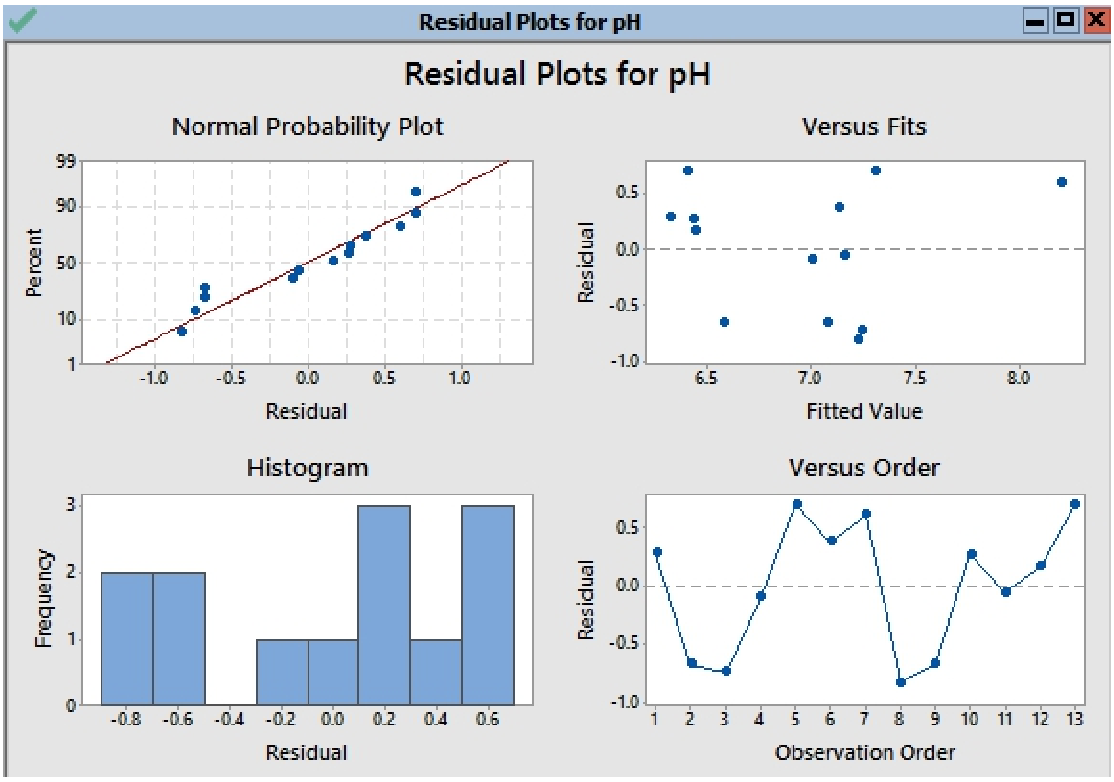
Q1.anov <- aov(y ~ x, data = Q1)
# Summary of the analysis
summary(Q1.anov)



```
> summary(Q1.anov)
            Df Sum Sq Mean Sq F value Pr(>F)
x            1  3.291   3.291   9.527 0.0103 *
Residuals   11  3.800   0.345
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1. **D**

Residul plots are plotted using Minitab



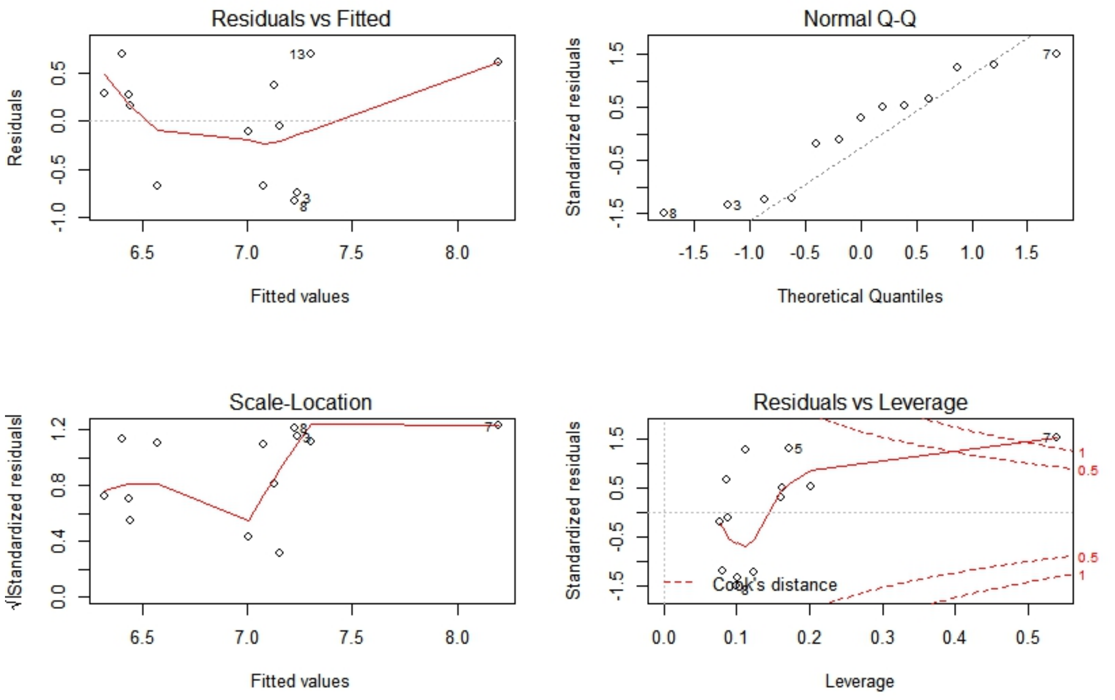
From here we can infer that:
Linearity and Additivity indicates the relationship between the dependant and independant variables.Here the expected value of dependent variable exhibits a straight-line function of each independent variable whereas the other values are fixed.This can be infer from Normal Probablity plot.The points are not that much symmetrically distributed around a diagonal line in the plot or around horizontal line with a roughly constant variance.

Statistical Independance of the errors : Here there is no correlationbetween consecutive errors and its evident from the Versus fits.

Normality : assumption can checked in the histograph.The errors are normally distributed.If the distribution is normal, the points on the plot are falling close to the diagonal reference line.As the bow-shaped pattern of deviations from the diagonal indicates that the residuals have excessive skewness from this plot its evident that normality is assumption is not violated.

Homoscedasticity : means the constant variance of the errors and it can observed in Versus Order graph.The residuals are growing larger as a function of the predicted value.

And Plot in R:



The four principal assumptions based on above graph are

First plot have a non linear pattern whereas the expected one is the straight line.The residuals are not spread uniformly in the horizontal line.

From the Normal Q-Q plot its evident that the residuals are normally distributed.

From the Scale location plot the residuals are not spread across uniformly.They are exhibiting non equal variance which can be termed as heteroscedastic.

From the Residuals vs Leverage Outliers are present that can be influential against the regression line.

Inference:

There are four model assumptions :

1. Variables are uncorrelated
2. The residuals follow a normal probablity distribution
3. The residuals are having uniform variance
4. The mean is zero

Here the model assumptions have been violated apart from the point The residuals follow a normal probablity distribution.

1.E

The regression line is denoted as y = 6.125 + 0.0059x.

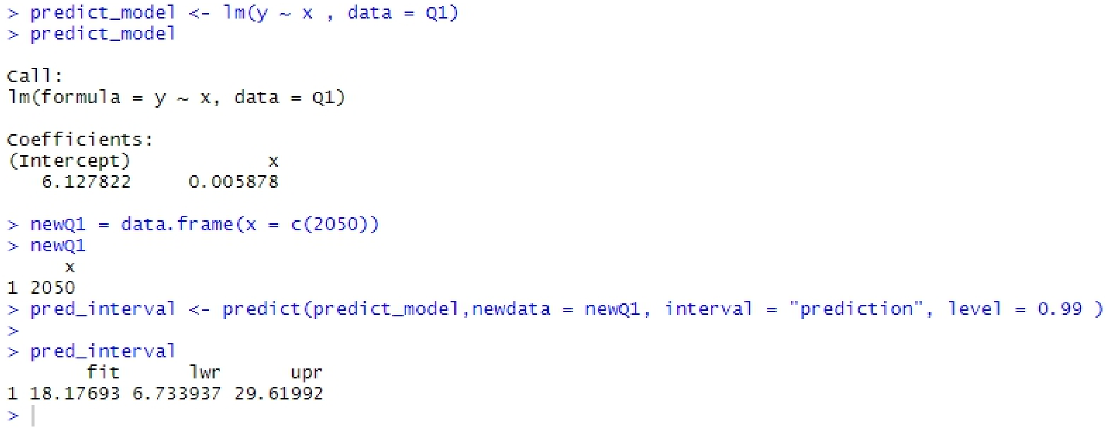
If the area = 2050
Ph Level / y = 6.125 + (0.0059 * 2050) = 18.22

The 99% confidence interval of this prediction is :

$$\ddot{y} \pm t\left(\frac{\alpha}{2}, n-2\right) * \sqrt{MSE} * \sqrt{\frac{(x-\overline{x})}{\Sigma(x-\overline{xi})} + 1/n}$$

Substituting the values the confidence interval is (6.88, 29.47)

Verifying with the R code:



```
> predict_model <- lm(y ~ x , data = Q1)
> predict_model

Call:
lm(formula = y ~ x, data = Q1)

Coefficients:
(Intercept)            x
   6.127822     0.005878

> newQ1 = data.frame(x = c(2050))
> newQ1
     x
1 2050
> pred_interval <- predict(predict_model,newdata = newQ1, interval = "prediction", level = 0.99 )
>
> pred_interval
       fit      lwr      upr
1 18.17693 6.733937 29.61992
> |
```

The (lwr,upr) CI is (6.733, 29.62) for the 99% interval

QUESTION 2:

2.A

Hand calculation to fit the linear regression model and least square estimates for the constant and slope includes below:

Variables are named as Steam in the X axis and mean atmospheric temperature is denoted as Pressure in Y axis

No	Steam	Pressure	xy	X square	Y square
1	35.3	10.98	387.59	1,246.09	120.56
2	29.7	11.13	330.56	882.09	123.88
3	30.8	12.51	385.31	948.64	156.5
4	58.8	8.4	493.92	3,457.44	70.56
5	61.4	9.27	569.18	3,769.96	85.93
6	71.3	8.73	622.45	5,083.69	76.21
7	74.4	6.36	473.18	5,535.36	40.45
8	76.7	8.5	651.95	5,882.89	72.25
9	70.7	7.82	552.87	4,998.49	61.15
10	57.5	9.14	525.55	3,306.25	83.54
11	46.4	8.24	382.34	2,152.96	67.9
12	28.9	12.19	352.29	835.21	148.6
13	28.1	11.88	333.83	789.61	141.13
14	39.1	9.57	374.19	1,528.81	91.58
15	46.8	10.94	511.99	2,190.24	119.68
16	48.5	9.58	464.63	2,352.25	91.78
17	59.3	10.09	598.34	3,516.49	101.81
18	70	8.11	567.7	4,900.00	65.77
19	70	6.83	478.1	4,900.00	46.65
20	74.5	8.88	661.56	5,550.25	78.85
21	72.1	7.68	553.73	5,198.41	58.98
22	58.1	8.47	492.11	3,375.61	71.74
23	44.6	8.86	395.16	1,989.16	78.5
24	33.4	10.36	346.02	1,115.56	107.33
25	28.6	11.08	316.89	817.96	122.77
Total	1315	235.6	11821.43	76323.42	2284.11

From the above table :

There are 25 observations in the above table , so n = 25

Sum(x) = 1315

Sum(y) = 235.6
Mean(x) = 52.6
Mean(y) = 9.424
Sum(xy) = 11821.43
Sum(x^2) = 76323.42
Sum(y^2) = 2284.11

S_xx = Sum(x^2) - ((Sum(x)^2) / n)
= 76323.42 - ((1315 ^2) / 25) = 7154.42

S_yy = Sum(y^2) - ((Sum(y)^2) / n)
= 2284.11 - ((235.6^2)/ 25) = 63.8158

S_xy = Sum(x*y) - ((Sum(x) * Sum(y)) / n)
= 11821.43 - ((1315 * 235.6) / 25) = -571.128

Beta = S_xy / S_xx = -571.128 / 7154.42 = -0.07982869, It can be round off to -0.079

Alpha = Mean(y) -(Beta * Mean(x))
= 9.424 - (-0.079 * 52.6) = 13.5794 , It can be round off to 13.62

The equation for the regression line can be represented as

Y = alpha + beta * x
= 13.62 - 0.079x

The least square estimate for the constant is 13.62 and the slope is -0.079.

2. B

The residuals means the difference between the y value and estimated y value and they are used to determine the adequacy of model.

No	Steam(x)	Pressure(y)	xy	X square	Y square	Estimated y	Residuals
1	35.3	10.98	387.59	1,246.09	120.56	10.81	0.17
2	29.7	11.13	330.56	882.09	123.88	11.25	-0.12
3	30.8	12.51	385.31	948.64	156.5	11.16	1.35
4	58.8	8.4	493.92	3,457.44	70.56	8.93	-0.53
5	61.4	9.27	569.18	3,769.96	85.93	8.72	0.55
6	71.3	8.73	622.45	5,083.69	76.21	7.93	0.8
7	74.4	6.36	473.18	5,535.36	40.45	7.68	-1.32
8	76.7	8.5	651.95	5,882.89	72.25	7.5	1
9	70.7	7.82	552.87	4,998.4	61.15	7.98	-0.16

				9			
10	57.5	9.14	525.55	3,306.25	83.54	9.03	0.11
11	46.4	8.24	382.34	2,152.96	67.9	9.92	-1.68
12	28.9	12.19	352.29	835.21	148.6	11.32	0.87
13	28.1	11.88	333.83	789.61	141.13	11.38	0.5
14	39.1	9.57	374.19	1,528.81	91.58	10.5	-0.93
15	46.8	10.94	511.99	2,190.24	119.68	9.89	1.05
16	48.5	9.58	464.63	2,352.25	91.78	9.75	-0.17
17	59.3	10.09	598.34	3,516.49	101.81	8.89	1.2
18	70	8.11	567.7	4,900.00	65.77	8.03	0.08
19	70	6.83	478.1	4,900.00	46.65	8.03	-1.2
20	74.5	8.88	661.56	5,550.25	78.85	7.68	1.2
21	72.1	7.68	553.73	5,198.41	58.98	7.87	-0.19
22	58.1	8.47	492.11	3,375.61	71.74	8.98	-0.51
23	44.6	8.86	395.16	1,989.16	78.5	10.06	-1.2
24	33.4	10.36	346.02	1,115.56	107.33	10.96	-0.6
25	28.6	11.08	316.89	817.96	122.77	11.34	-0.26
Total	1315	235.6	11821.43	76323.42	2284.11	235.6	

2.C

Hand calculations to perform the ANOVA Test:

To perform the ANOVA Test the following table has to be calculated:

	DF	Sum of Squares	MS	F
Regression	1	SSR	MSR=SSR/1	F = MSR / MSE
Residual	n - 2	SSE	MSE=SSE/(n−2)	
Total	n - 1	SST		

Considering the above table we need to find the SSR, SSE, SST. For calculating tho se values the below table has to be calculated:

SSR = Square(Difference of estimated y and average of y)

SSE = Square(Difference of y and estimated y value)
SST = Square(y and average of y)

No	Steam(x)	Pressure(y)	x*y	x^2	y^2	Estimated y	(y-Estimated y)^2	(Estimated y- average. of y)^2	(y- average. of y)^2	Residual
1	35.3	10.98	387.59	1,246.09	120.56	10.81	0.03	1.91	2.42	0.17
2	29.7	11.13	330.56	882.09	123.88	11.25	0.02	3.35	2.91	-0.12
3	30.8	12.51	385.31	948.64	156.5	11.17	1.81	3.03	9.52	1.34
4	58.8	8.4	493.92	3,457.44	70.56	8.93	0.28	0.24	1.05	-0.53
5	61.4	9.27	569.18	3,769.96	85.93	8.72	0.3	0.49	0.02	0.55
6	71.3	8.73	622.45	5,083.69	76.21	7.93	0.63	2.22	0.48	0.8
7	74.4	6.36	473.18	5,535.36	40.45	7.69	1.76	3.02	9.39	-1.33
8	76.7	8.5	651.95	5,882.89	72.25	7.5	1	3.69	0.85	1
9	70.7	7.82	552.87	4,998.49	61.15	7.98	0.03	2.08	2.57	-0.16
10	57.5	9.14	525.55	3,306.25	83.54	9.03	0.01	0.15	0.08	0.11
11	46.4	8.24	382.34	2,152.96	67.9	9.92	2.82	0.25	1.4	-1.68
12	28.9	12.19	352.29	835.21	148.6	11.32	0.76	3.58	7.65	0.87
13	28.1	11.88	333.83	789.61	141.13	11.38	0.25	3.83	6.03	0.5
14	39.1	9.57	374.19	1,528.81	91.58	10.5	0.87	1.16	0.02	-0.93
15	46.8	10.94	511.99	2,190.24	119.68	9.89	1.11	0.22	2.3	1.05
16	48.5	9.58	464.63	2,352.25	91.78	9.75	0.03	0.11	0.02	-0.17
17	59.3	10.09	598.34	3,516.49	101.81	8.89	1.44	0.28	0.44	1.2
18	70	8.11	567.7	4,900.00	65.77	8.04	0.01	1.92	1.73	0.07
19	70	6.83	478.1	4,900.00	46.65	8.04	1.46	1.92	6.73	-1.21
20	74.5	8.88	661.56	5,550.25	78.85	7.68	1.45	3.05	0.3	1.2
21	72.1	7.68	553.73	5,198.41	58.98	7.87	0.04	2.42	3.04	-0.19
22	58.1	8.47	492.11	3,375.61	71.74	8.99	0.27	0.19	0.91	-0.52
23	44.6	8.86	395.16	1,989.16	78.5	10.06	1.45	0.41	0.32	-1.2
24	33.4	10.36	346.02	1,115.56	107.33	10.96	0.36	2.35	0.88	-0.6
25	28.6	11.08	316.89	817.96	122.77	11.34	0.07	3.67	2.74	-0.26
Tot	1315	235.6	11821.43	76323.42	2284.11	235.638	18.2234617	45.55969	63.8158	-0.038

Substituting the values in the ANOV Table we will get as:

	DF	SS	MS	F
Regression	1	45.559	45.559	57.543
Residual	23	18.223	0.792	
Total	24	63.816		

- Analysis of Variance ANOVA is statistical analysis and ANOVA table is used for test the hypotheses about the regression linearity.
- This table compares the amound of variation between the groups. The mean square is defined by sum of degrees / Degrees of freedom.
- The SSR and SSE depend on the model used to fit data and by obtaining these values we can examine whether the model provides a good fit to data.
- If the null hypothesis of equal means is true, two mean squares estimates the same quantitycalled as error variance and data would be of approximately equal magnitude. The ratio should be close to 1.

2.D.

The Correlation coefficient r can be found by the below equation,

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

This is also a measure of goodness of fit where the value is in between -1 and 1.
Resolving the above equation using the values from table of 2 A we will get as : -0.845
From the value of r we can determine the strength of linear relationship.The magnitude of r is 0.845 and if value of r is in between 0.8≤|r|≤1 the strength of linear relationship is strong.

Coefficient of determination is = square(-0.845) = 0.714 ,

$$r^2 = \frac{S_{XY}^2}{S_{XX}S_{YY}}$$

This is in between 0 and 1 value and its a measure of goodness of fit.If this value is 1 they exhibits a perfect correlation that explains 100% of the linear variations in the data.
That is 71.4% of the variation in Pressure is explained by linear variations in the Steam.

2.E.

Standard error of a statistic is the standard deviation of its sampling distribution or an estimate of that standard deviation
Std for the error = sqrt((s_yy - beta*s_xy)/(n-2)) = 0.8901245

Std of slope is represented the average distance that observed values deviate from the regression line.

S_alpha = s*sqrt(sumX2/(n*s_xx)) = 0.5814635

Std of constant can be find out by

s_beta = s/sqrt(s_xx) = 0.01052358

2.F

Here we need to test the hypothesis that pressure is linearly related to the Steam.

Hypothesis

Null hypothesis H0: Beta = 0
Alternative hypothesis Ha: Beta ≠ 0.

F-Test for overall goodness of fit:

The F test statistic is defined as mean square due to regression/MSE.It is used to test the null hypothes is

H0= A linear relationship does not exist between y and x
Against the alternative
H1= A linear relationship does exist between y and x.

As per method of rejection points in the F test;

If Beta = 0 then F = 1 which means there is no linear relationship.

If F > 1 then b ≠ 0, which means there is a linear relationship

The F value is 57.543 which > 1 hence we can reject null hypothesis and we can say the slope and constant are significant.There exists a linear relationship between the variables.

2.G.

Confidence interval for the slope:

The 100(1 - alpha)% confidence interval for the slope estimate Beta ^ is given by

$$\hat{\beta} \pm t_{n-2}^{\alpha/2} s_{\hat{\beta}}$$

$s_{\hat{\beta}}$ is the standard deviation of estimate which is called standard of the estimate.For calculation we need SSE which is the sum of squared estimates and can find out from the last column

SSE = Sum = 18.22
S square = SSE / n - 2 = 18.22 / 23 = 0.79
S = sqrt(S square) = 0.889

Standard of the estimate is S/ sqrt(Sxx) = 0.889 / sqrt(7154.2)
= 0.0105
Beta value is -0.079

Substituting in the main equation the Confidence intervals will be

(-0.079 - (2.807 * 0.0105)) , (-0.079 + (2.807 * 0.0105))

The CI is (-0.112 , -0.05)

Confidence interval for the constant:
The 100(1 - alpha)% confidence interval for the slope estimate Alpha ^ is given by:

$$\hat{\alpha} \pm t_{n-2}^{\alpha/2} s_{\hat{\alpha}}$$

Where the standard error of the intercept estimate is given by,

= s * sqrt(sum of x^2 / (n * Sxx))
= 0.889 * sqrt(76323.42 / (25 * 7154.42)) = 0.5807

Here we considered for the 99% interval and t value is 2.807.
Confidence interval for the slope estimate is:

(13.62 - (2.807 * 0.5807)), (13.62 - (2.807 * 0.5807))

= (11.99 , 15.25)

QUESTION 3:

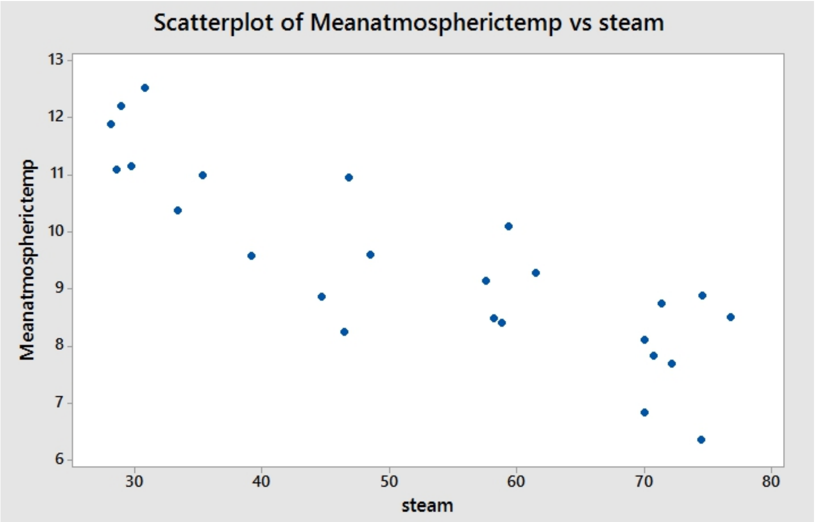
3 A.

Steps in Minitab:

- ❖ Get the data loaded into the Minitab(First column denotes the steam in pounds per months and Second column denotes the mean atmospheric temperature.

- ❖ Select Graph
- ❖ Select Scatterplot and click on Simple, Then OK
- ❖ From the displayed table all the column variables get displayed.Click on C3 pressure and click on Select.It will get select under the Y variables
- ❖ From the displayed table all the column variables get displayed.Click on C2 steam and click on Select.It will get select under the X variables
- ❖ Click OK

This is the Scatterplot of the data using Minitab:



Comment : From the scatterplot its evident that there exists a negative correlation between the mean atmospheric temperature and steam where the steam in pounds per months and Meanatmospherictemp denotes the mean atmosphere temperature measured in Fahrenheit.Negative correlation indicates the relationship between these two variables are inverse.When one variable increases other will get decreases.The relationship exists between the steam and Meanatmospherictemp is opposite all the time. As the correlation is negative the regression slope would be negative and obviously the correlation coefficient value would be less than zero

3. B

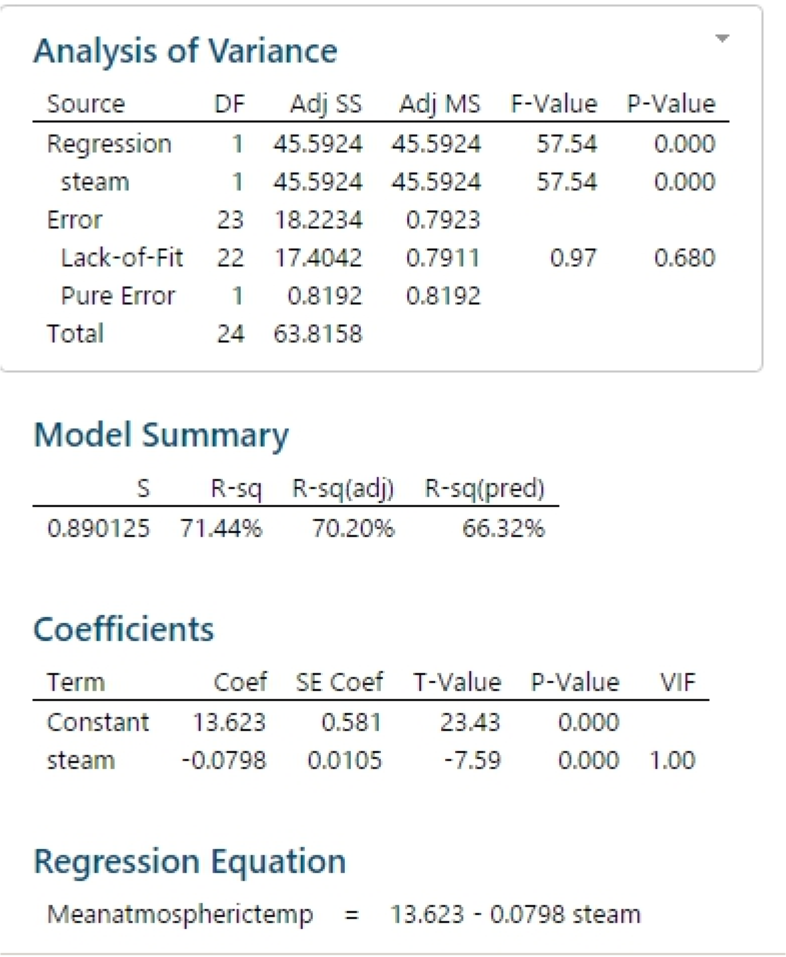
1)Referring to A Question:

The least square estimates for the constant and slope can obtained by Minitab as mentioned below:

- Get the data in the worksheet and click on Stat
- Select Regression and Regression from the side arrow list
- Select Fit Regression Model

- Select the Steam as the Responses as its the dependant variable
- Select Meanatmospherictemp as the Continous predictor
- Click on Storage
- From the Regression Storage tick the check boxes Fits and Residuals
- Click OK and OK in the dialog box

We will get the Least square estimates as mentioned below



Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	45.5924	45.5924	57.54	0.000
steam	1	45.5924	45.5924	57.54	0.000
Error	23	18.2234	0.7923		
Lack-of-Fit	22	17.4042	0.7911	0.97	0.680
Pure Error	1	0.8192	0.8192		
Total	24	63.8158			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.890125	71.44%	70.20%	66.32%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	13.623	0.581	23.43	0.000	
steam	-0.0798	0.0105	-7.59	0.000	1.00

Regression Equation

Meanatmospherictemp = 13.623 - 0.0798 steam

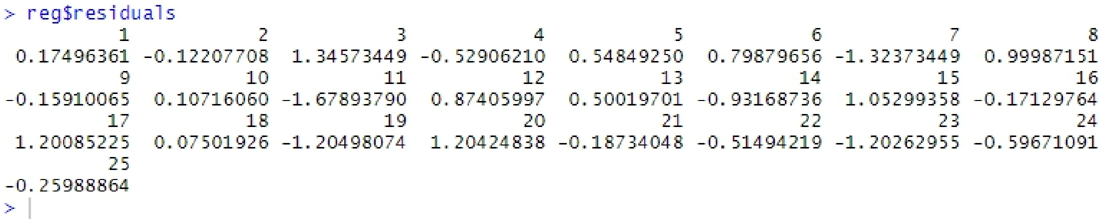
2) Referring to B

Residuals for each of the 25 Observations can be find out in R by:

```
Q3 = data.frame(
  x = c(35.3, 29.7, 30.8, 58.8, 61.4, 71.3, 74.4, 76.7, 70.7, 57.5, 46.4, 28.9, 28.1,
        39.1, 46.8, 48.5, 59.3, 70, 70, 74.5, 72.1, 58.1, 44.6, 33.4, 28.6),
  y = c(10.98, 11.13, 12.51, 8.40, 9.27, 8.73, 6.36, 8.50, 7.82, 9.14, 8.24, 12.19,
11.88,
        9.57, 10.94, 9.58, 10.09, 8.11, 6.83, 8.88, 7.68, 8.47, 8.86, 10.36, 11.08))

x = Q3[,1]
y = Q3[,2]
n = length(x)

# Function "lm" will fit a linear model for y which is dependent on x
```

reg = lm(y~x)

# Returns the residuals.
reg$residuals

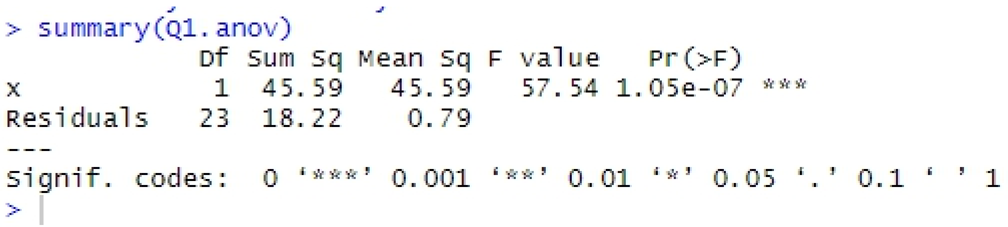


```
> reg$residuals
          1           2           3           4           5           6           7           8 
 0.17496361 -0.12207708  1.34573449 -0.52906210  0.54849250  0.79879656 -1.32373449  0.99987151 
          9          10          11          12          13          14          15          16 
-0.15910065  0.10716060 -1.67893790  0.87405997  0.50019701 -0.93168736  1.05299358 -0.17129764 
         17          18          19          20          21          22          23          24 
 1.20085225  0.07501926 -1.20498074  1.20424838 -0.18734048 -0.51494219 -1.20262955 -0.59671091 
         25 
-0.25988864 
> |
```

3)Referring to C

Making the ANOVA Table in R:

Q1.anov <- aov(y ~ x, data = Q1)
# Summary of the analysis
summary(Q1.anov)



```
> summary(Q1.anov)
            Df Sum Sq Mean Sq F value   Pr(>F)    
x            1  45.59   45.59   57.54 1.05e-07 ***
Residuals   23  18.22    0.79                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> |
```

Analysis of Variance ANOVA is statistical analysis and ANOVA table is used for test the hypotheses about the regression linearity. This table compares the amound of variation between the groups. The SSR and SSE depend on the model used to fit data and by obtaining these values we can examine whether the model provides a good fit to data.

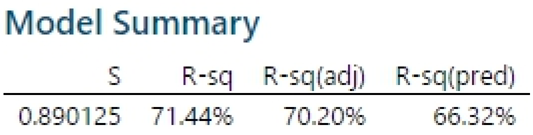
If the null hypothesis of equal means is true, two mean squares estimates the same quantitycalled as error variance and data would be of approximately equal magnitude. The ratio should be close to 1.

4)Referring to D

Minitab:

The least square estimates for the constant and slope can obtained by Minitab as mentioned below:

- Get the data in the worksheet and click on Stat
- Select Regression and Regression from the side arrow list
- Select Fit Regression Model
- Select the Steam as the Responses as its the dependant variable
- Select Meanatmospherictemp as the Continous predictor
- Click on Storage
- From the Regression Storage tick the check boxes Fits and Residuals
- Click OK and OK in the dialog box
- In here we will get the model summary too:

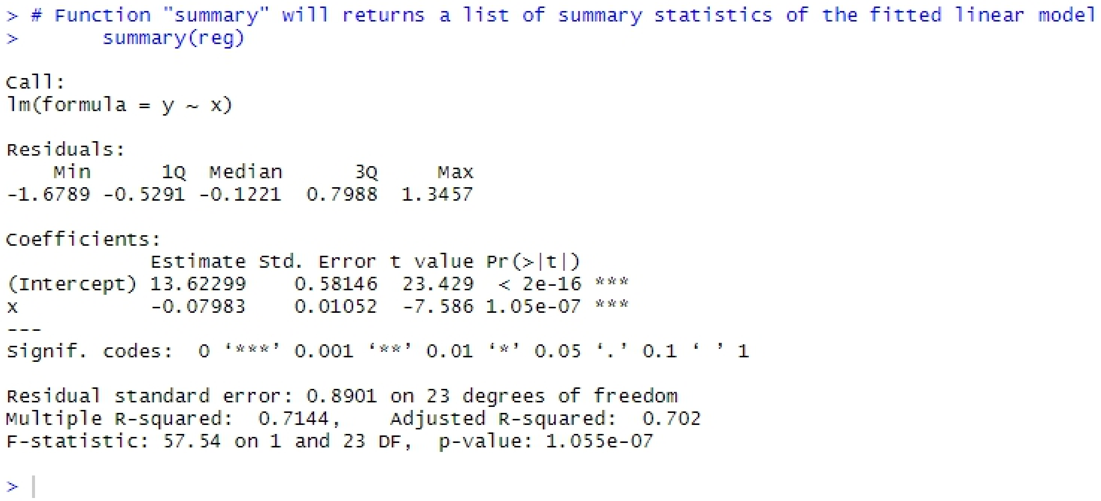


Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.890125	71.44%	70.20%	66.32%

OR

In R:



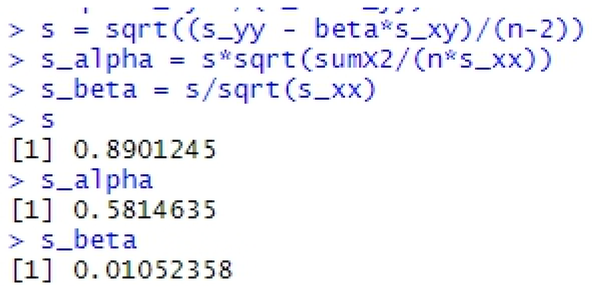
```
> # Function "summary" will returns a list of summary statistics of the fitted linear model
>       summary(reg)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.6789 -0.5291 -0.1221  0.7988  1.3457 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13.62299    0.58146  23.429  < 2e-16 ***
x           -0.07983    0.01052  -7.586 1.05e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8901 on 23 degrees of freedom
Multiple R-squared:  0.7144,	Adjusted R-squared:  0.702 
F-statistic: 57.54 on 1 and 23 DF,  p-value: 1.055e-07

> |
```

From the value of r we can determine the strength of linear relationship.The magnitude of r is 0.89 and if value of r is in between 0.8≤|r|≤1 the strength of linear relationship is strong.

Coefficient of determination is = 0.714.

That is 71.4% of the variation in Mean atmospheric temperature is explained by linear variations in the Steam.

5)Referring to E:



```
> s = sqrt((s_yy - beta*s_xy)/(n-2))
> s_alpha = s*sqrt(sumX2/(n*s_xx))
> s_beta = s/sqrt(s_xx)
> s
[1] 0.8901245
> s_alpha
[1] 0.5814635
> s_beta
[1] 0.01052358
```

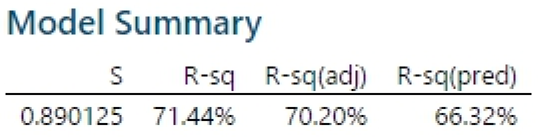
S is the std for the error which is 0.89
S_alpha is the std for the intercept which is 0.58
S_beta is the std for the slope is 0.010

6)Referring to F:

If there is a linear relationship between the independent variable and the dependent variable slope will not be equal to zero. The null hypothesis states that the slope is equal to zero, and the alternative hypothesis states that the slope is not equal to zero.The P value is an important parameter where its the probablity of observing the sample statistic as extreme as test statistic.

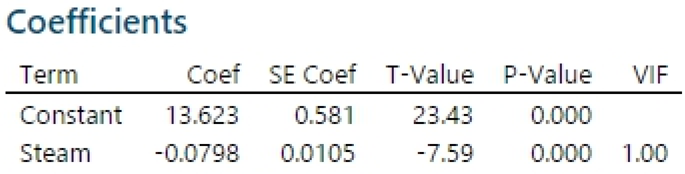
The least square estimates for the constant and slope can obtained by Minitab as mentioned below:

- Get the data in the worksheet and click on Stat
- Select Regression and Regression from the side arrow list
- Select Fit Regression Model
- Select the Steam as the Responses as its the dependant variable
- Select Meanatmosphericttemp as the Continous predictor
- Click on Storage
- From the Regression Storage tick the check boxes Fits and Residuals
- Click OK and OK in the dialog box
- In here we will get the Coeffiecients too:



Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.890125	71.44%	70.20%	66.32%



Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	13.623	0.581	23.43	0.000	
Steam	-0.0798	0.0105	-7.59	0.000	1.00

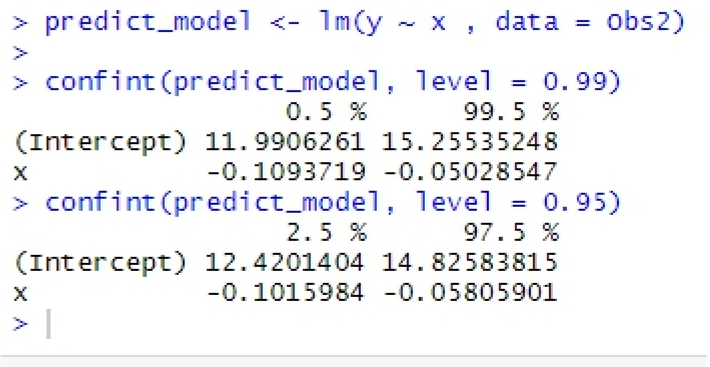
Hypothesis:
Null hypothesis H0: Beta = 0
Alternative hypothesis Ha: Beta ≠ 0.

If the relation between slope and constant are significant the slope will not be equal to zero.

Interpret results:
The P-value is zero for 0 for the constant and slope which is less than significance level 0.05 we cannot accept the null hypothesis.
The Slope is not equal to zero an there is a significant relationship between Slope and constant.

7) Referring to G:

The CI of contant and slope can be found by Confint in the R by:



```
> predict_model <- lm(y ~ x , data = obs2)
>
> confint(predict_model, level = 0.99)
                 0.5 %      99.5 %
(Intercept) 11.9906261 15.25535248
x           -0.1093719 -0.05028547
> confint(predict_model, level = 0.95)
                 2.5 %      97.5 %
(Intercept) 12.4201404 14.82583815
x           -0.1015984 -0.05805901
> |
```

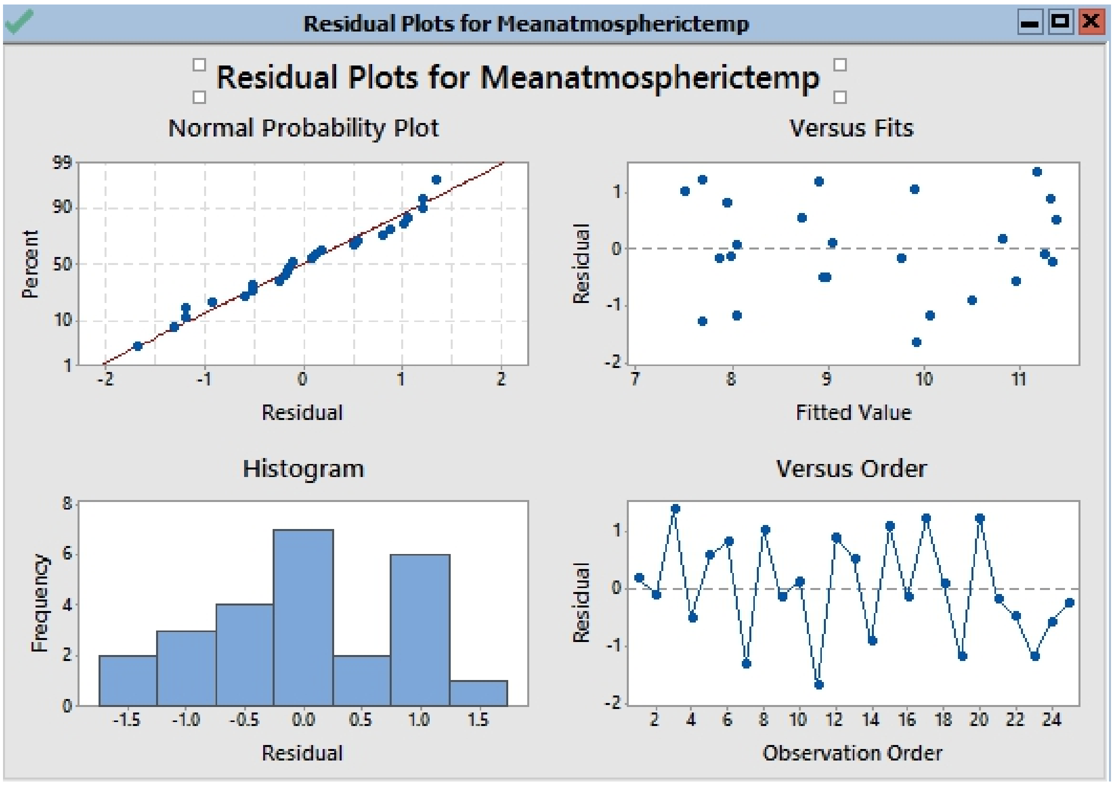
Here we can observe the values for the 99% and 95% interval.

The CI of the slope is (-0.109 , -0.050)
The CI of the intercept is (11.99 , 15.26)

Question 3 - Part 3

Residual plots can be generated by:

- Click on Stat
- Click on Regression and select Regression from the side options
- Select Fit Regression Model
- Give the Response variable as Meanatmosphericternp and Select Steam as the Continous predictors
- Click on Graph
- Under the Residuals for plots Select Regular
- Select the Four in one
- Click OK and OK in the final dialogue box



This can be seen From Normal Probability plot the residual plots are almost linear.The expected value of Atmospheric temperature is a straight line of each value of Steam.Plot is having correct functional form.

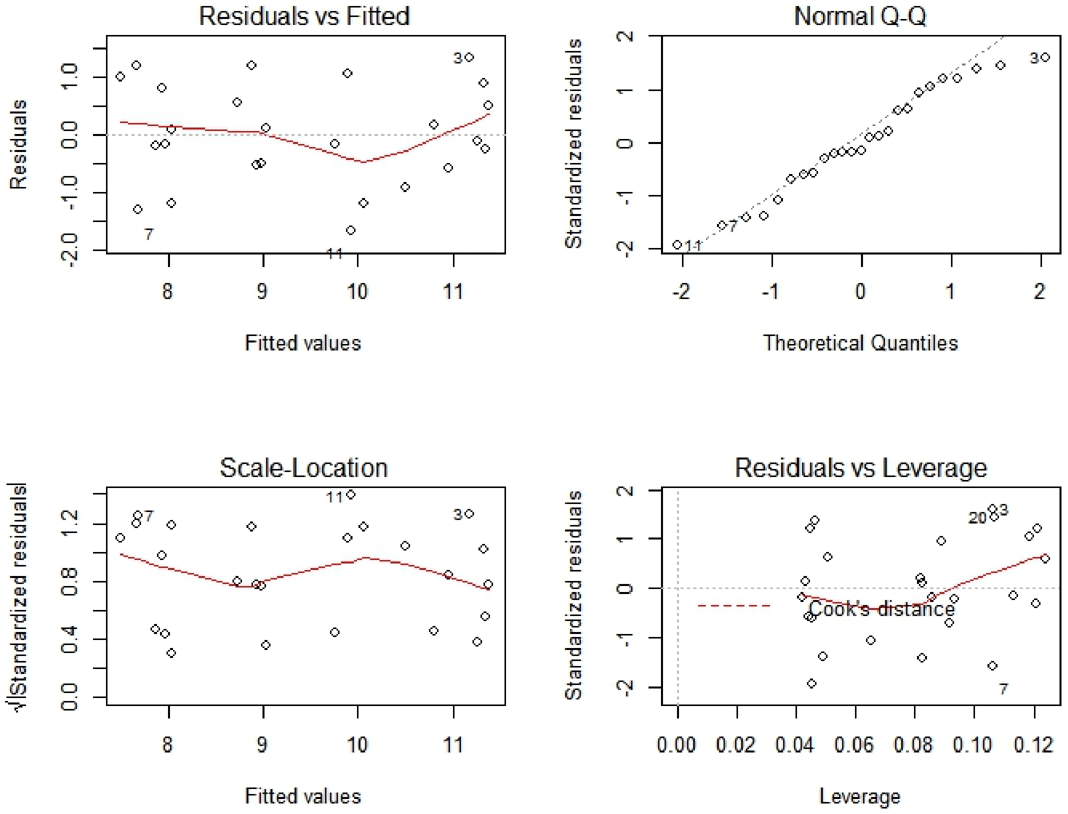
From the Versus fits graph its evident that there is no correlation between the consecutive errors and exhibits statistical independence, constant variance

Histogram plot exhibits normality of the error distribution.Residual terms are normally distributed.

From the Versus Orders graph the constant variance of the errors are exhibited.Homoscedasticity is not violated

And in R:

```
model <-lm(y ~ x, data = Obs2)

par(mfrow = c(2,2))

plot(model)
```



There are four model assumptions :

Variables are uncorrelated

The residuals follow a normal probablity distribution

The residuals are having uniform variance

The mean is zero

Inference: Here there is a slight nonlinearity as per the Residuals vs Fitted graph,The residuals follow a normal probablity distribution in the Normal Q-Q plot,The spread points are not that much got spreaded equaly along with the ranges of predictors.From the residuals vs Leverage graph the values are not that much influential to determine the regression line.