**228.371 Assignment 3**

**Assessed lab: June 18, 2014**

**Question 2**

Instructions:

Read through this entire question before starting. You will include your answer for each part within the spaces between <answer starts here> and <answer ends here>. You can add as much space as you need between these lines. Using a fixed width font such as Courier will ensure the answers are easy for the marker to read.

Make sure you copy all relevant R code and output to this template so that the marker can see what you have done. Please do not just copy the whole console window as this might include a lot of unnecessary material. Do not use a screen shot. The marker may need to copy your R code for their use during marking.

2.1 Import the data in the file Reaction.csv which shows the amount of a solid that remains undissolved after an amount of time. The column X is time measured as a fraction of an hour, and Y is the fraction of the solid material that did not dissolve. Plot the data and propose a model (using words to describe the data and your model and how you intend to proceed). You may wish to add detail to your graph to support your thinking.

[2 marks]

<answer starts here>

> plot(Y~X)

> abline(mod)

> summary(mod)

Call:

lm(formula = Y ~ X)

Residuals:

Min 1Q Median 3Q Max

-0.068909 -0.044564 -0.001594 0.038644 0.088762

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.41711 0.01908 21.862 6.95e-14 \*\*\*

X -0.74068 0.10771 -6.877 2.68e-06 \*\*\*

---

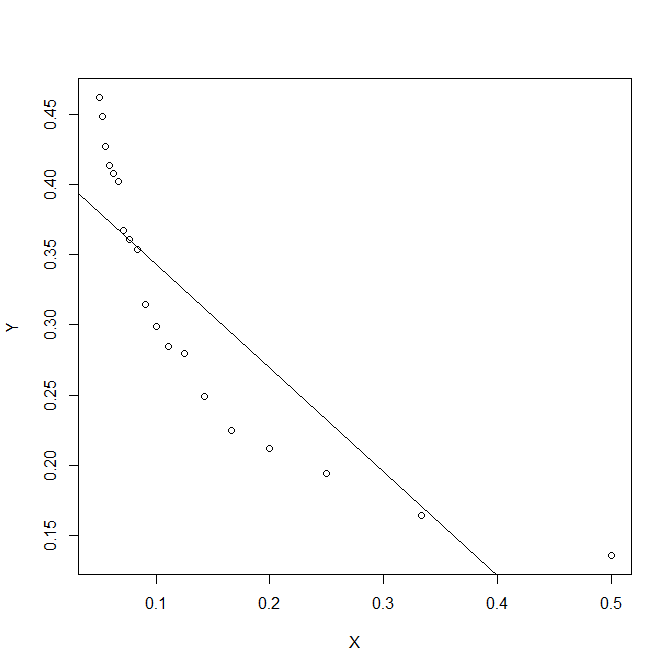
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05288 on 17 degrees of freedom

Multiple R-squared: 0.7356, Adjusted R-squared: 0.72

F-statistic: 47.29 on 1 and 17 DF, p-value: 2.684e-06

From the graph we can see that the linear regression does not fit so perfectly. The relationship between the X and Y is more likely a curve. From the summary, we can see that the residuals are significantly larger at the side and much smaller at the median therefore the relationship is a curve. The R squared value is 0.72 which does not shows a significant linear relationship between X and Y.



<answer ends here>

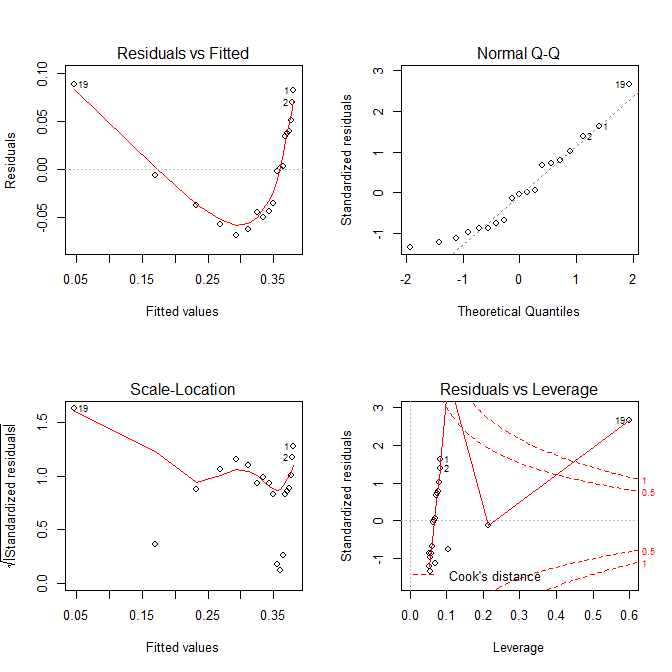
2.2 Fit your model, explaining any actions taken to prepare the data.

[2 marks]

<answer starts here>

The graph and the linear regression line is showed in question 2.1.

The equation of the linear regression line is: Y=-0.74068\*X+0.41711.



These are the plots that can describe the linear model clearly. The residual graph show that the data has larger residual at both sides and less at the middle, therefore a curve relationship would fit more. The normal qq plot shows a diagonal line therefore it is a normal distribution.

<answer ends here>

2.3 Ensure your model is appropriate. Provide suitable graph(s) or summaries and comments to support your preferred model. N.B. this may not end up being the model you originally proposed; if it is not, then include the updated model here.

[4 marks]

<answer starts here>

As I described above, a curve relationship will be more fitted to the data. Therefore I will try several curve lines and compare which of it is the best.

1. Y vs. log(X)

> LX<-log(X)

> mod<-lm(Y~LX)

> plot(Y~LX)

> abline(mod)

> summary(mod)

Call:

lm(formula = Y ~ LX)

Residuals:

Min 1Q Median 3Q Max

-0.02797 -0.01976 0.00003 0.01427 0.04643

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.013288 0.018886 -0.704 0.491

LX -0.147712 0.008144 -18.137 1.47e-12 \*\*\*

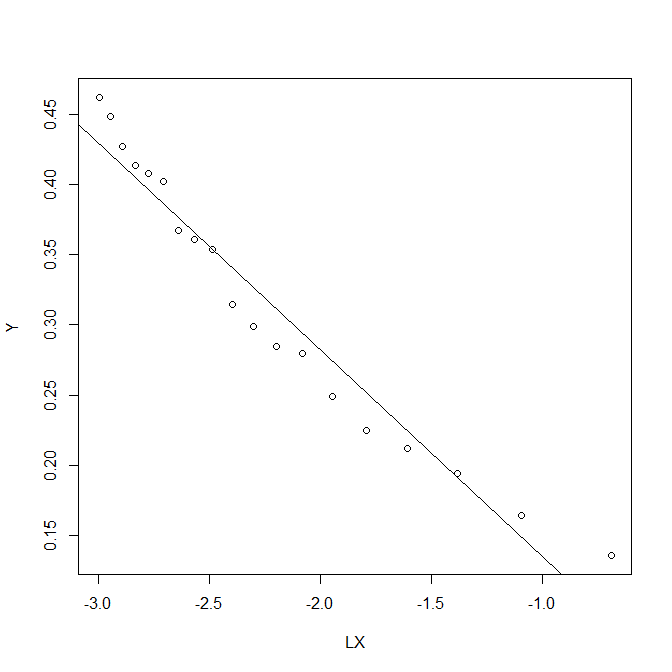
---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02279 on 17 degrees of freedom

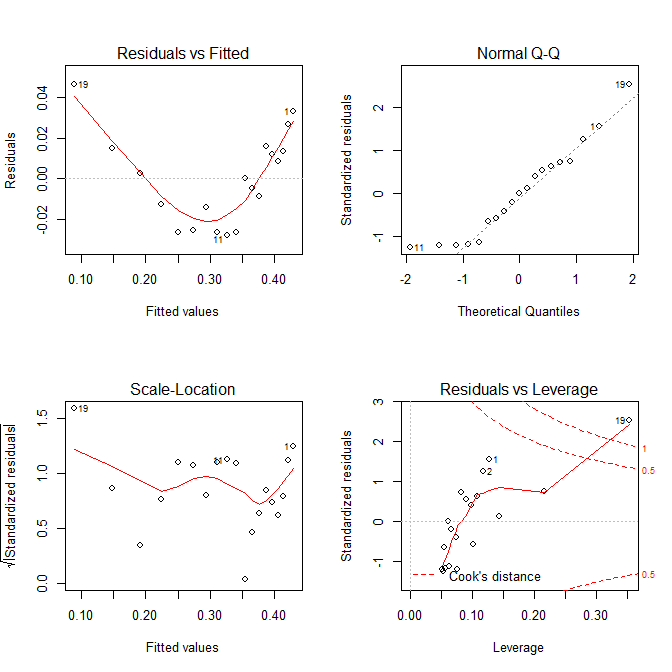
Multiple R-squared: 0.9509, Adjusted R-squared: 0.948

F-statistic: 328.9 on 1 and 17 DF, p-value: 1.47e-12



For Y vs. log(X), it shows a perfect linear relationship. The R squared values are large therefore significant linear relationship.

The expression is: Y=-0.147712\*log(X)-0.013288.



1. Y vs. X1/2

> HX<-X^(1/2)

> mod<-lm(Y~HX)

> plot(Y~HX)

> plot(Y~HX)

> abline(mod)

> summary(mod)

Call:

lm(formula = Y ~ HX)

Residuals:

Min 1Q Median 3Q Max

-0.047828 -0.033352 -0.004295 0.023737 0.074444

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.56149 0.02568 21.86 6.94e-14 \*\*\*

HX -0.70768 0.06945 -10.19 1.18e-08 \*\*\*

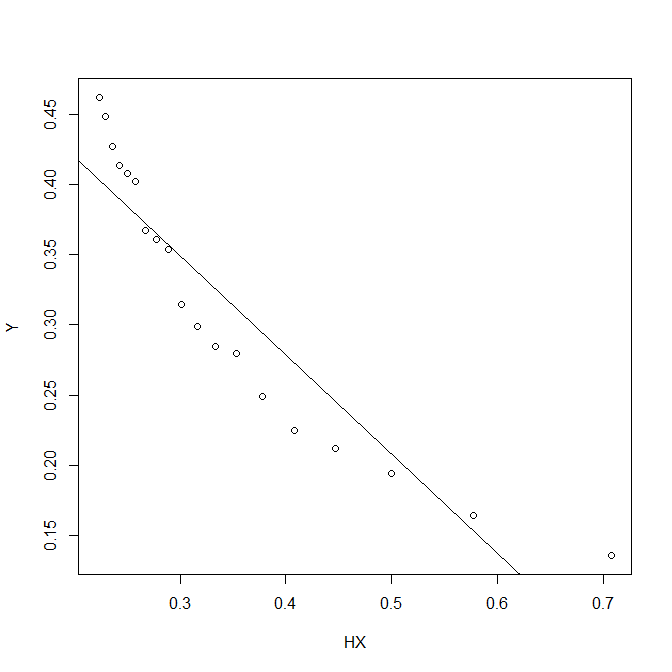
---

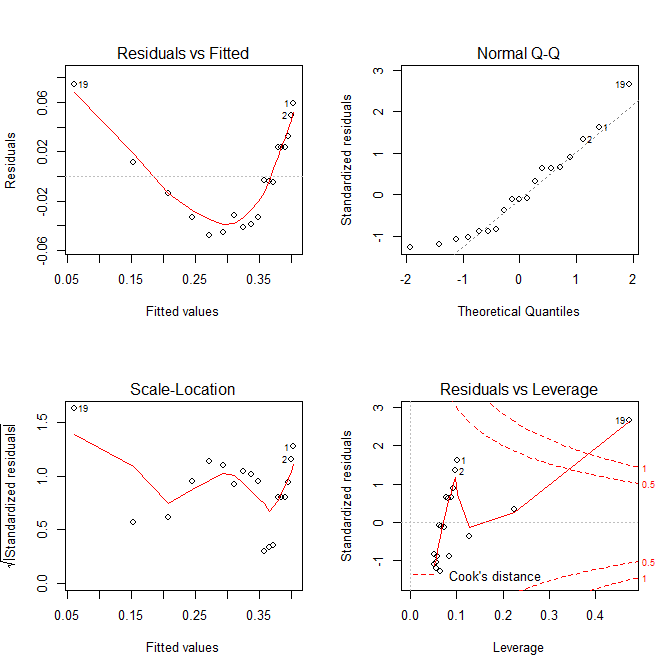
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03857 on 17 degrees of freedom

Multiple R-squared: 0.8593, Adjusted R-squared: 0.851

F-statistic: 103.8 on 1 and 17 DF, p-value: 1.175e-08



This model is better than the linear one, however, not as good as the log one. The R squared value shows that the regression fits well, but the value is not as high as the log model. 

In conclusion, I choose the log model to be my best fit model. The expression of the model is:

Y=-0.147712\*log(X)-0.013288.

<answer ends here>

2.4 Using the best model you have found to this point, make a point prediction about the amount of solid that will remain after 45 minutes.

(3 marks)

<answer starts here>

For X=45min=0.75hr, Y=-0.147712\*log(0.75)-0.013288=0.00516695066

So 0.0052 amount of solid will remain.

<answer ends here>

2.5 Find an interval estimate for the expected amount of solid remaining after 45 minutes. Interpret this interval using a sentence that would appear in a report.

[2 marks]

<answer starts here>

<answer ends here>

2.6 Create a random bivariate sample of size 25, making sure that the y values are a linear function of x with some noise added. You will be using them to demonstrate your knowledge of some key ideas in fitting regression models. Show all the R code used here.

[3 marks]

<answer starts here>

|  |  |
| --- | --- |
| X | Y |
| 1 | 103 |
| 2 | 226 |
| 3 | 354 |
| 4 | 425 |
| 5 | 531 |
| 6 | 642 |
| 7 | 735 |
| 8 | 839 |
| 9 | 921 |
| 10 | 1076 |
| 11 | 1143 |
| 12 | 1224 |
| 13 | 1354 |
| 14 | 1465 |
| 15 | 1543 |
| 16 | 1667 |
| 17 | 1743 |
| 18 | 1865 |
| 19 | 1932 |
| 20 | 2064 |
| 21 | 2176 |
| 22 | 2223 |
| 23 | 2358 |
| 24 | 2464 |
| 25 | 2565 |

<answer ends here>

2.7 Plot your data and add the fitted line.

[2 marks]

<answer starts here>

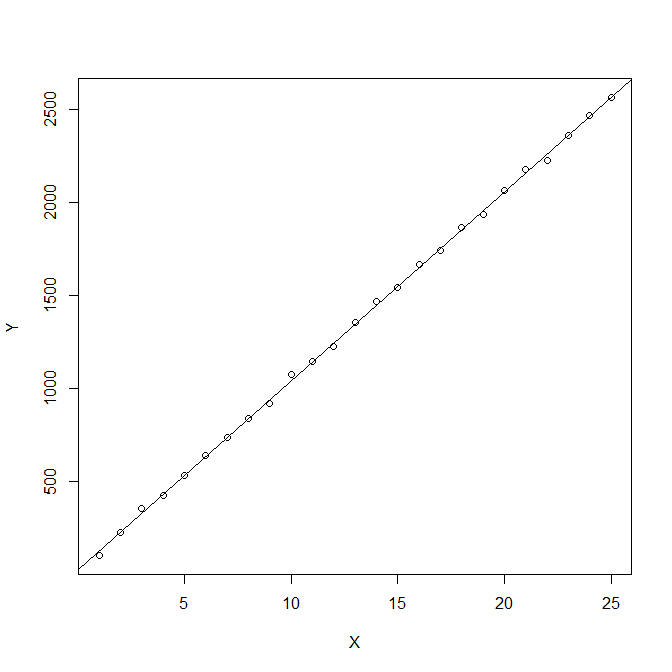
> X<-c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25)

> Y<-c(103,226,354,425,531,642,735,839,921,1076,1143,1224,1354,1465,1543,1667,1743,1865,1932,2064,2176,2223,2358,2464,2565)

> mod<-lm(Y~X)

> plot(Y~X)

> abline(mod)



<answer ends here>

2.8 What is the highest order polynomial that is sensible for this data? Provide evidence to back any assertion you make.

[2 marks]

<answer starts here>

> summary(mod)

Call:

lm(formula = Y ~ X)

Residuals:

Min 1Q Median 3Q Max

-35.757 -7.283 0.422 8.480 34.892

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 26.4000 6.8714 3.842 0.000832 \*\*\*

X 101.4708 0.4622 219.528 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.67 on 23 degrees of freedom

Multiple R-squared: 0.9995, Adjusted R-squared: 0.9995

F-statistic: 4.819e+04 on 1 and 23 DF, p-value: < 2.2e-16

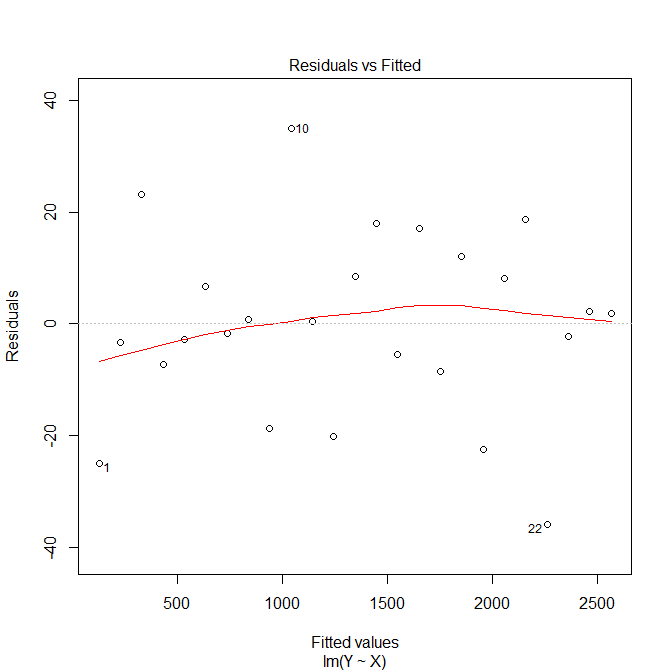
The highest order is X. from the graph we can see a perfect straight line and the linear model fits perfectly. The R squared values are very high. The equation is Y=101.4708\*X+26.4, where the highest order is X.

<answer ends here>

2.9 Assume the point with the maximum absolute residual is an outlier. Show the impact of removing this point from your analysis.

[2 marks]

<answer starts here>



This is the residual plot. The point that has the highest residual is 22.

Remove 22.

> Y<-c(103,226,354,425,531,642,735,839,921,1076,1143,1224,1354,1465,1543,1667,1743,1865,1932,2064,2176,2358,2464,2565)

> X<-c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,23,24,25)

> mod<-lm(Y~X)

> plot(Y~X)

> abline(mod)

> summary(mod)

Call:

lm(formula = Y ~ X)

Residuals:

Min 1Q Median 3Q Max

-25.592 -6.836 -1.768 7.623 34.126

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 24.4084 6.2438 3.909 0.000752 \*\*\*

X 101.7465 0.4308 236.168 < 2e-16 \*\*\*

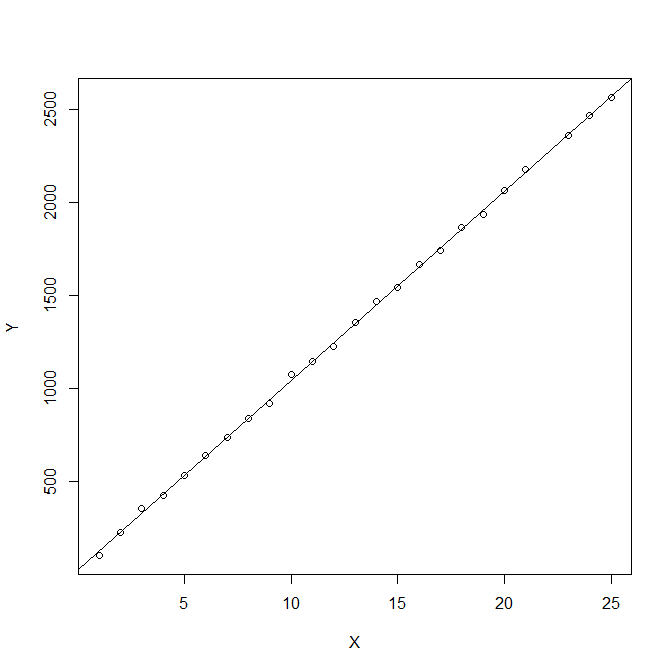
---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.02 on 22 degrees of freedom

Multiple R-squared: 0.9996, Adjusted R-squared: 0.9996

F-statistic: 5.578e+04 on 1 and 22 DF, p-value: < 2.2e-16



There is a slightly increase value in R squared after removing the outlier which means that the regression line fits even more. However the effect is not significant as the original one is already very good.

<answer ends here>

2.10 Alter the 25th point in your original data set so that when you re-fit the chosen model from Part H, the leverage of point 25 becomes a problem. You will need to show what the point was originally and what its new value is, as well as showing that the leverage has become a problem.

[3 marks]

<answer starts here>

The original value is 2565. The new one id 9999.

> Y<-c(103,226,354,425,531,642,735,839,921,1076,1143,1224,1354,1465,1543,1667,1743,1865,1932,2064,2176,2223,2358,2464,9999)

> mod<-lm(Y~X)

> plot(Y~X)

> abline(mod)

> summary(mod)

Call:

lm(formula = Y ~ X)

Residuals:

Min 1Q Median 3Q Max

-1049.9 -628.3 -248.8 189.8 6315.0

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -568.32 589.18 -0.965 0.344785

X 170.09 39.63 4.292 0.000272 \*\*\*

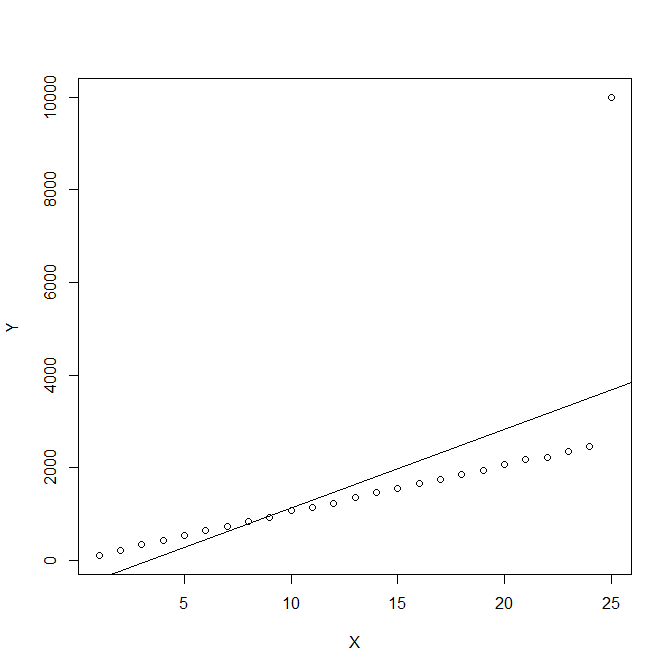
---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1429 on 23 degrees of freedom

Multiple R-squared: 0.4447, Adjusted R-squared: 0.4206

F-statistic: 18.42 on 1 and 23 DF, p-value: 0.0002721



As we can see, the new 25th value drive the regression line towards it. There is a significant reduction of the R squared value.

<answer ends here>

<Question ends>