Problem 1.1

1.1 Equation of a line. Consider the fitted regression equation Å¶= $100 + 15X$. Which of the following is *false*?

a. The sample slope is 15.

b. The predicted value of Y when $X = 0$ is 100.

c. The predicted value of Y when $X = 2$ is 110.

d. Larger values of X are associated with larger values of Y .

C is false, because the predicted value of Y when X=2 is 130 not 110. Since,

Y = 100 + 15(2) = 130

Problem 1.2

1.2 Residual plots to check conditions. For which of the following conditions for inference in regression does a residual plot *not* aid in assessing whether the condition is satisfied?

a. Linearity

b. Constant variance

c. Independence

d. Zero mean

C, Independence

Problem 1.3

1.3 Sparrows slope. Priscilla Erickson from Kenyon College collected data on a stratified random sample of 116 Savannah sparrows at Kent Island. The weight (in grams) and wing length (in mm) were obtained for birds from nests that were reduced, controlled, or enlarged. The data[5] are in the file **Sparrows**. Based on the following computer output (which you will also use for the odd exercises through Exercise 1.11), what is the slope of the least squares regression line for predicting sparrow weight from wing length? Sparrow

The regression equation is Weight = 1.37 + 0.467 WingLength

Predictor	Coef	SE Coef	T	P
Constant	1.3655	0.9573	1.43	0.156
WingLength	0.4674	0.03472	13.46	0.000

S = 1.39959 R-Sq = 61.4% R-Sq(adj) = 61.1%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	355.05	355.05	181.25	0.000
Residual Error	114	223.31	1.96		
Total	115	578.36			

$\beta_1 = 0.4674$

Problem 1.16

1.16 Glow-worms. The paper "I'm Sexy and I Glow It: Female Ornamentation in a Nocturnal Capital Breeder" is about glow-worms. Female glow-worms attract males by glowing, and flying males search for females. The authors write, "We found brightness to correlate with female fecundity." The file **GlowWorms** has data on 26 female glow-worms captured in Finland. The variable *Lantern* is the size, in mm, of the part of the female abdomen that glows. The variable *Eggs* is number of eggs laid by the glow-worm. **Glow**

a. Fit the regression of *Eggs* on *Lantern*. What is the fitted regression model?

b. Interpret the coefficient of *Lantern* in the context of this setting.

c. Suppose a glow-worm has a lantern size of 14 mm. What is the predicted number of eggs she will lay?

```
> summary(model)

Call:
lm(formula = Eggs ~ Lantern)

Residuals:
   Min     1Q Median     3Q    Max 
-69.50 -23.59  -3.20  22.95  63.33 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -8.977     21.869  -0.410 0.685087    
Lantern        7.325      1.757   4.169 0.000343 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 32.71 on 24 degrees of freedom
Multiple R-squared:  0.4201,	Adjusted R-squared:  0.3959 
F-statistic: 17.38 on 1 and 24 DF,  p-value: 0.0003431

> 
```

<u>Part a</u>

According to the summary above, the equation for fitted regression model is:

Eggs hat = (-8.977) + (7.325 x *Lantern*)

<u>Part b</u>

When the glow of bee (*Lantern*) increases by 1 mm, there is an increase in eggs by 7.325 on average

<u>Part c</u>

If the glow worm has a lantern size of 14mm then:

Eggs hat = (-8.977) + (7.325 x *Lantern*)

Eggs hat = (-8.977) + (7.325 x *14mm*)

Eggs hat = 93.573

That is, the glow worm will lay 93-94 eggs.

Problem 1.18

1.18 Male body measurements. The file **Faces** has data on grip strength (*MaxGripStrength*) and shoulder-to-hip ratio (*SHR*) for each of 38 college men. Grip strength, measured in kilograms, is the maximum of three readings for each hand from the man squeezing a handheld dynamometer. *SHR* is the ratio of shoulder circumference to hip circumference. Faces

a. Fit the regression of *MaxGripStrength* on *SHR*. What is the fitted regression model?

b. Interpret the coefficient of *SHR* in the context of this setting.

c. Predict the *MaxGripStrength* of a man with *SHR* equal to 1.5.

C:\Users\Admin\Documents\R\M349R\HW1.R - R Editor

```
setwd("C:\\Users\\Admin\\Downloads")
data <- read.csv("ex01-18Faces.csv",header=TRUE)
attach(data)
fix(data)
hist(MaxGripStrength)
hist(SHR)
plot(MaxGripStrength~SHR)
model=lm(MaxGripStrength~SHR)
summary(model)
```

```
Call:
lm(formula = MaxGripStrength ~ SHR)

Residuals:
    Min      1Q  Median      3Q     Max 
-13.148  -6.068  -1.977   6.668  22.242 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)    9.298     15.574   0.597   0.5542  
SHR           28.959     11.721   2.471   0.0184 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.664 on 36 degrees of freedom
Multiple R-squared:  0.145,	Adjusted R-squared:  0.1212 
F-statistic: 6.104 on 1 and 36 DF,  p-value: 0.01836
```

<u>Part a</u>

According to the summary above, the equation for fitted regression model is:

MaxGripStrength hat = (9.298) + (28.959 x *SHR*)

<u>Part b</u>

If the shoulder to hip ratio increases by 1, there is an increase in maximum grip strength by 28.959 on average

<u>Part c</u>

If a man has a SHR of 1.5 then:

MaxGripStrength hat = (9.298) + (28.959 x *SHR*)

MaxGripStrength hat = (9.298) + (28.959 x *1.5*)

MaxGripStrength hat = 52.7365

That is, the man has maximum grip strength of 52.7365 kgs.

Problem 1.20

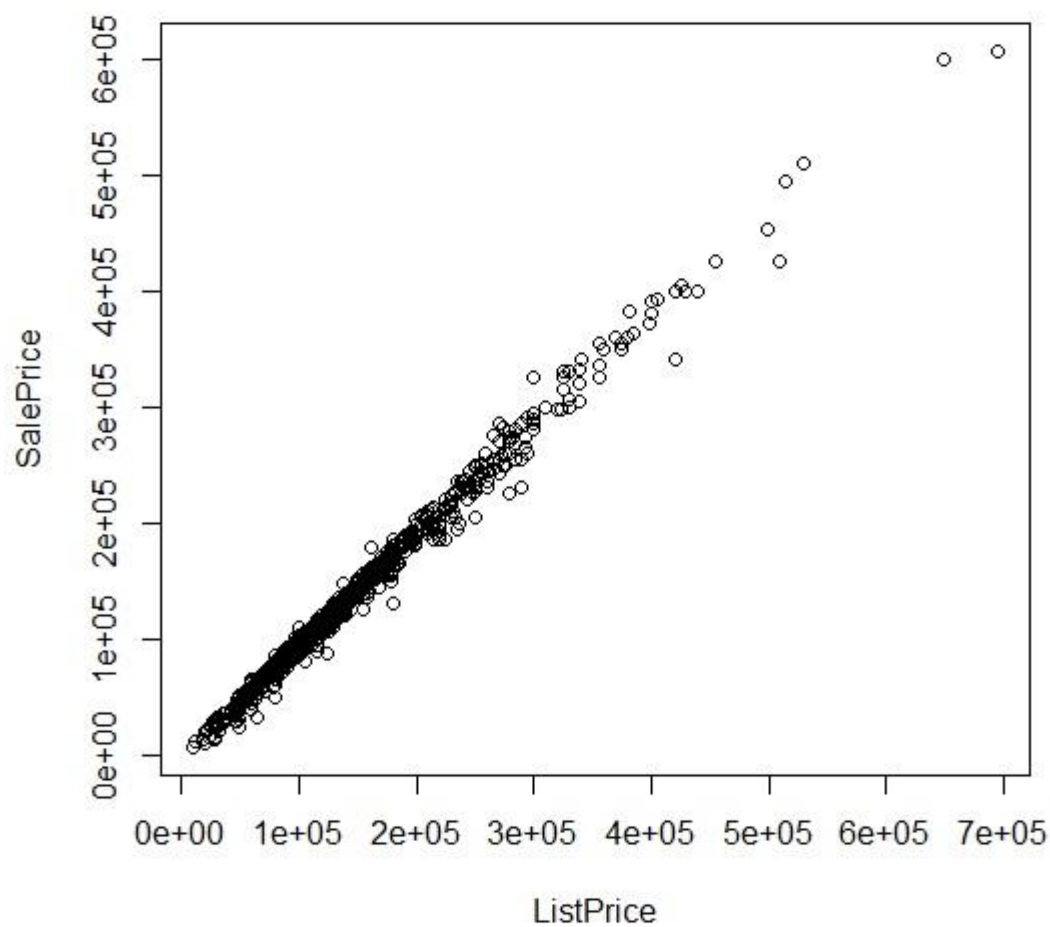
1.20 Houses in Grinnell, CHOOSE/FIT. The file **GrinnellHouses** contains data from 929 house sales in Grinnell, Iowa, between 2005 and into 2015. In this question we investigate the relationship of the list price of the home (what the seller asks for the home) to the final sale price. One would expect there to be a strong relationship. In most markets, including Grinnell during this period, list price almost always exceeds sale price. Grinnell

a. Make a scatterplot with ListPrice on the horizontal axis and SalePrice on the vertical axis. Comment on the pattern.

b. Find the least squares regression line for predicting sale price of a home based on list price of that home.

c. Interpret the value (not just the sign) of the slope of the fitted model in the context of this setting.

Part a



The graph in general seems linear, but we can clearly see that list price of a house is slightly higher than the sales price almost consistently.

<u>Part b</u>

From the model (see below) we deduce that:

ListPrice hat = (1647) + (1.049 x *SalePrice*)

```
Call:
lm(formula = ListPrice ~ SalePrice)

Residuals:
   Min     1Q Median     3Q    Max
-44674  -4513  -1104   3175  61638

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.647e+03  5.496e+02   2.996  0.00281 **
SalePrice   1.049e+00  3.562e-03 294.578  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8458 on 927 degrees of freedom
Multiple R-squared:  0.9894,    Adjusted R-squared:  0.9894
F-statistic: 8.678e+04 on 1 and 927 DF,  p-value: < 2.2e-16
```

Part c

ListPrice hat = (1647) + (1.049 x *SalePrice*)

For every dollar in selling price of the home, the list price is higher by $1.049

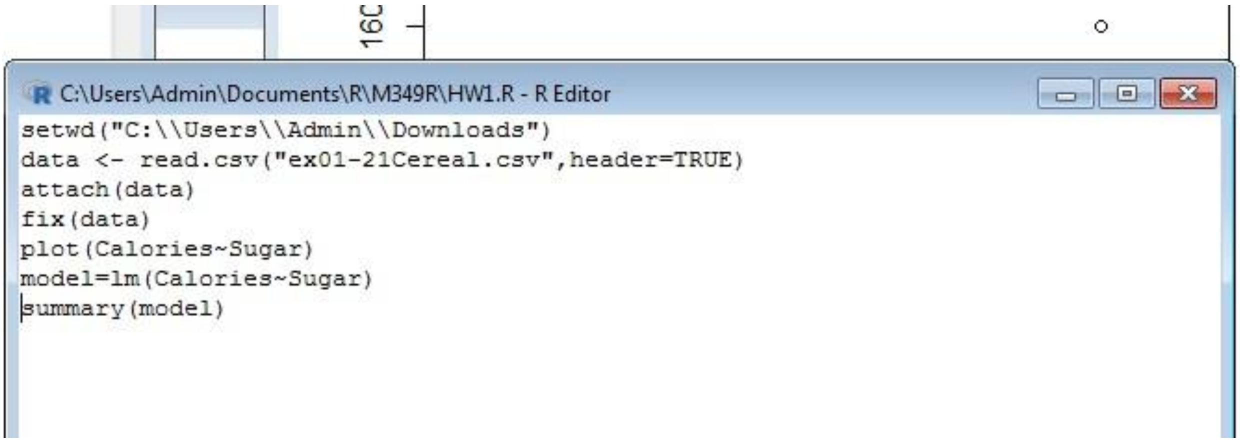
Problem 1.21

1.21 Breakfast cereal, ASSESS. Refer to the data on breakfast cereals that is described in Exercise 1.19. The number of calories and number of grams of sugar per serving were measured for 36 breakfast cereals. The data are in the file **Cereal**. We are interested in trying to predict the calories using the sugar content. **Cereal**

a. How many calories would the fitted model predict for a cereal that has 10 grams of sugar?

b. Cheerios has 110 calories but just 1 gram of sugar. Find the residual for this data point.

c. Does the linear regression model appear to be a good summary of the relationship between calories and sugar content of breakfast cereals?




```
Call:
lm(formula = Calories ~ Sugar)

Residuals:
    Min      1Q  Median      3Q     Max
-37.428  -9.832   0.245   8.909  40.322

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  87.4277     5.1627  16.935   <2e-16 ***
Sugar         2.4808     0.7074   3.507   0.0013 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.27 on 34 degrees of freedom
Multiple R-squared:  0.2656,    Adjusted R-squared:  0.244
F-statistic:  12.3 on 1 and 34 DF,  p-value: 0.001296
```

Part a

According to the summary above, the equation for fitted regression model is:

Calories hat = (87.4277) + (2.4808 x *Sugar*)

Then,

Calories hat = (87.4277) + (2.4808 x *10*)

Calories hat = 112.2357 calories

Part b

Calories hat = (87.4277) + (2.4808 x *Sugar*)

Calories hat = (87.4277) + (2.4808 x *1*)

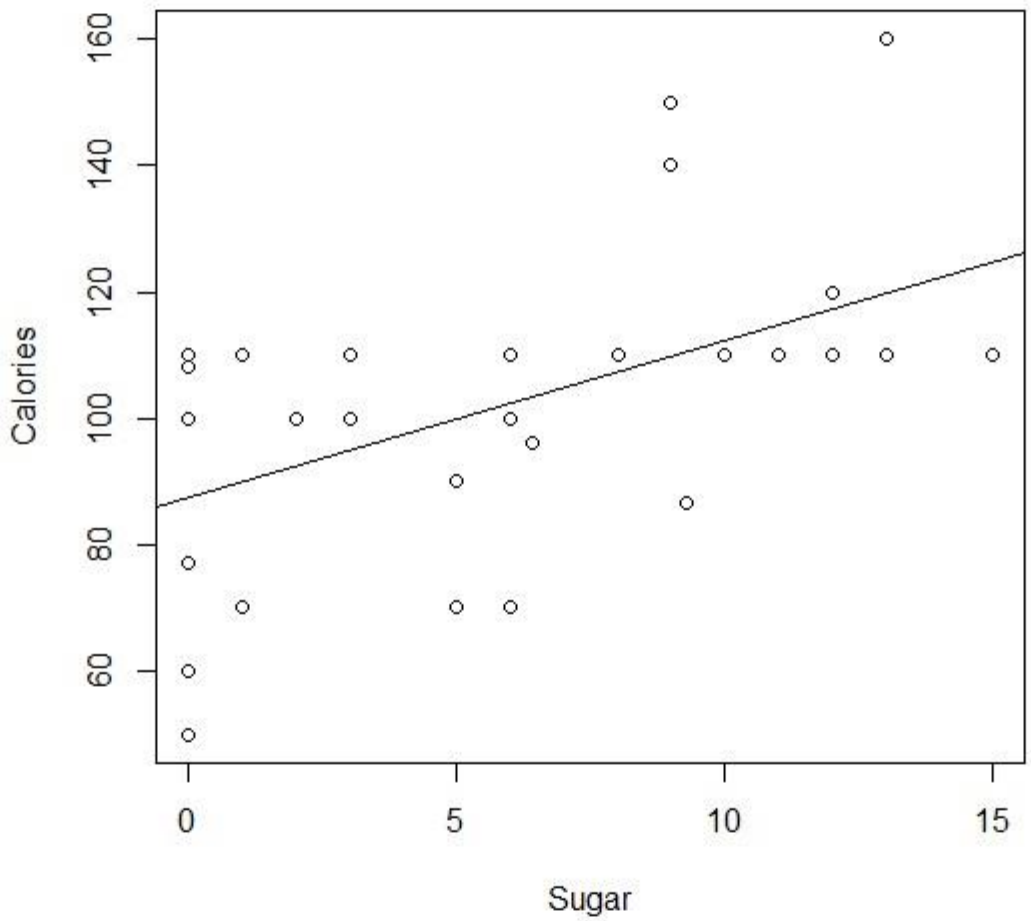
Calories hat = 89.9085 calories

Then,

Residual = 110- 89.9085

Residual = 20.0915 calories Part

c



As the amount of sugar increases, calories increase. However, there are some outliers which makes our points scatter away from the line.

Problem 1.27

1.27 Capacitor voltage. A capacitor was charged with a 9-volt battery and then a voltmeter recorded the voltage as the capacitor was discharged. Measurements were taken every 0.02 second. The data are in the file **Volts**. Volts

a. Make a scatterplot with *Voltage* on the vertical axis versus *Time* on the horizontal axis. Comment on the pattern.

b. Create a residuals versus fits plot for predicting *Voltage* from *Time*. What does this plot tell you about the idea of fitting a linear model to predict *Voltage* from *Time*? Explain.

c. Transform *Voltage* using a log transformation and then plot log(*Voltage*) versus *Time*. Comment on the pattern.

d. Regress log(*Voltage*) on *Time* and write down the prediction equation.

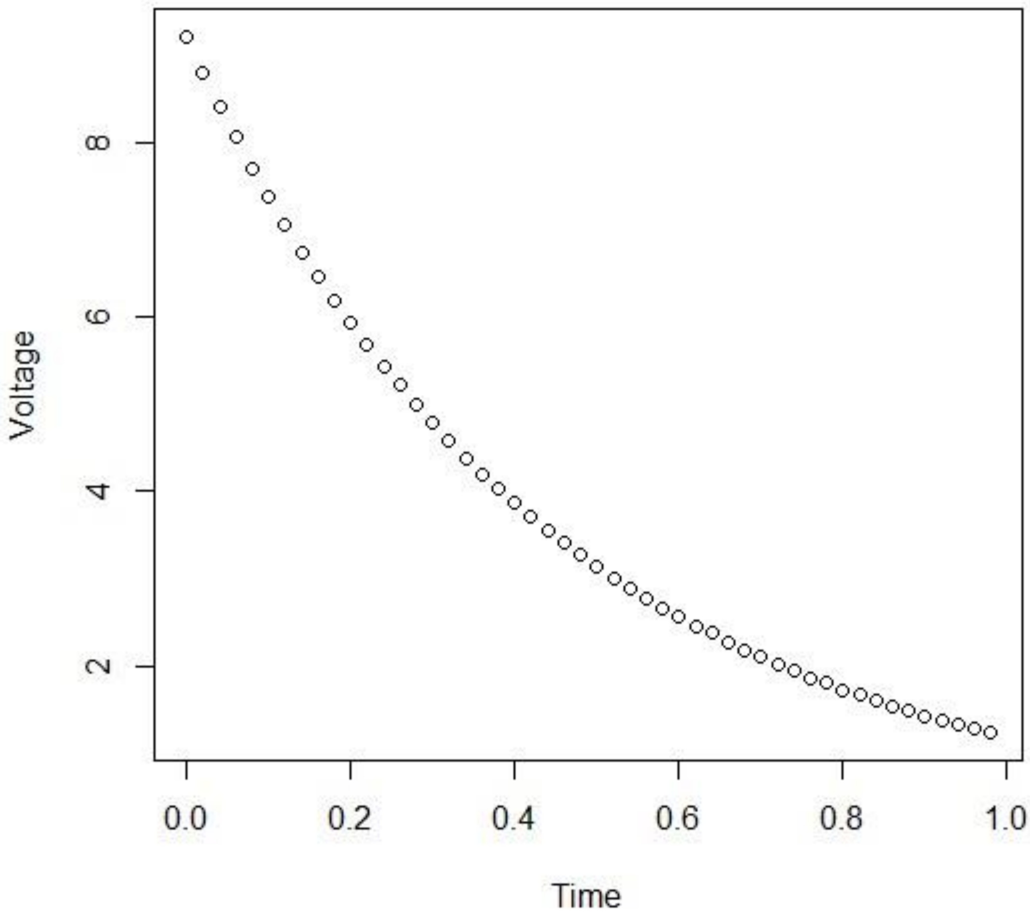
e. Make a plot of residuals versus fitted values from the regression from part (c). Comment on the pattern.

C:\Users\Admin\Documents\R\M349R\HW1.R - R Editor

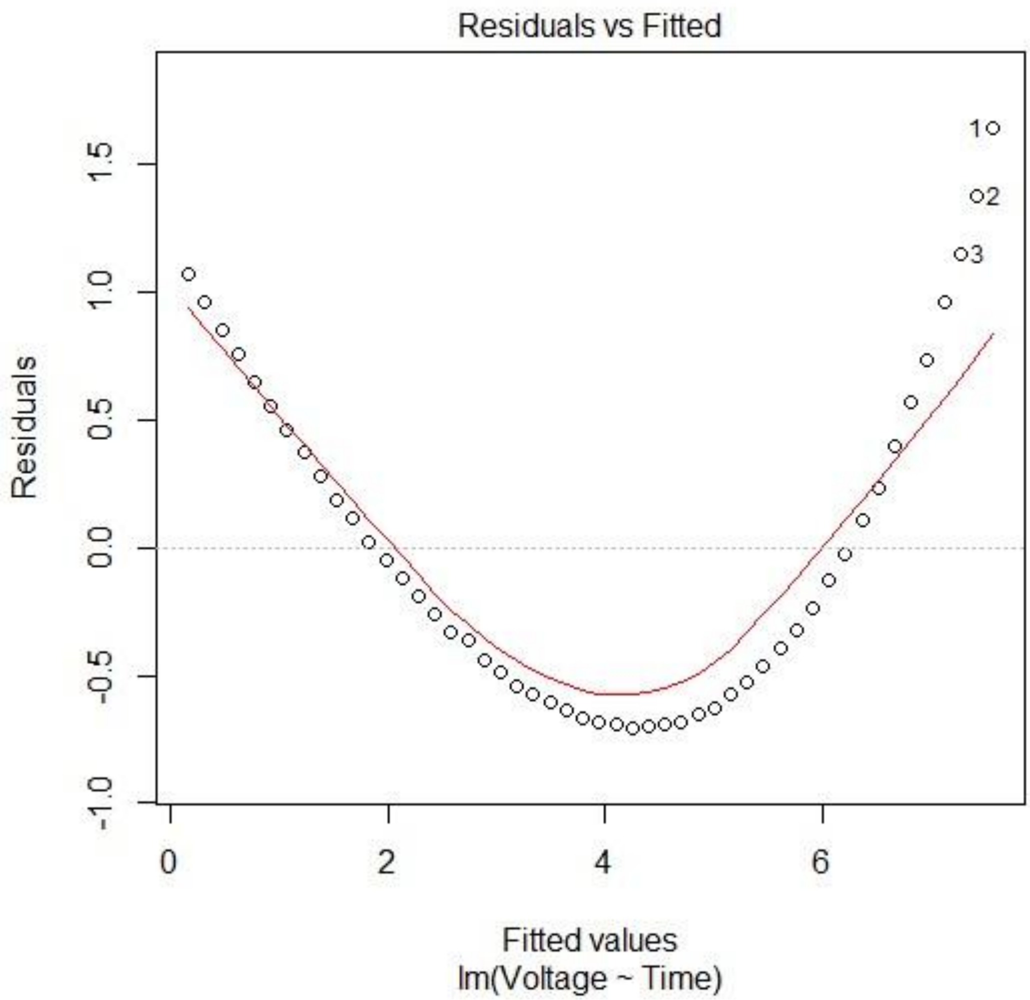
```
setwd("C:\\Users\\Admin\\Downloads")
data <- read.csv("ex01-27Volts.csv",header=TRUE)
attach(data)
fix(data)
plot(Voltage~Time)
model=lm(Voltage~Time)
plot(model, which = 1)
summary(model)
LogTran <- log(data$Voltage)
plot(LogTran~Time)
model2=lm(LogTran~Time)
summary(model2)
plot(model2, which =1)
```

Part A

We can see that, as time increases voltage decreases exponentially (not linearly).


Voltage
8
6
4
2
0.0
0.2
0.4
0.6
0.8
1.0
Time

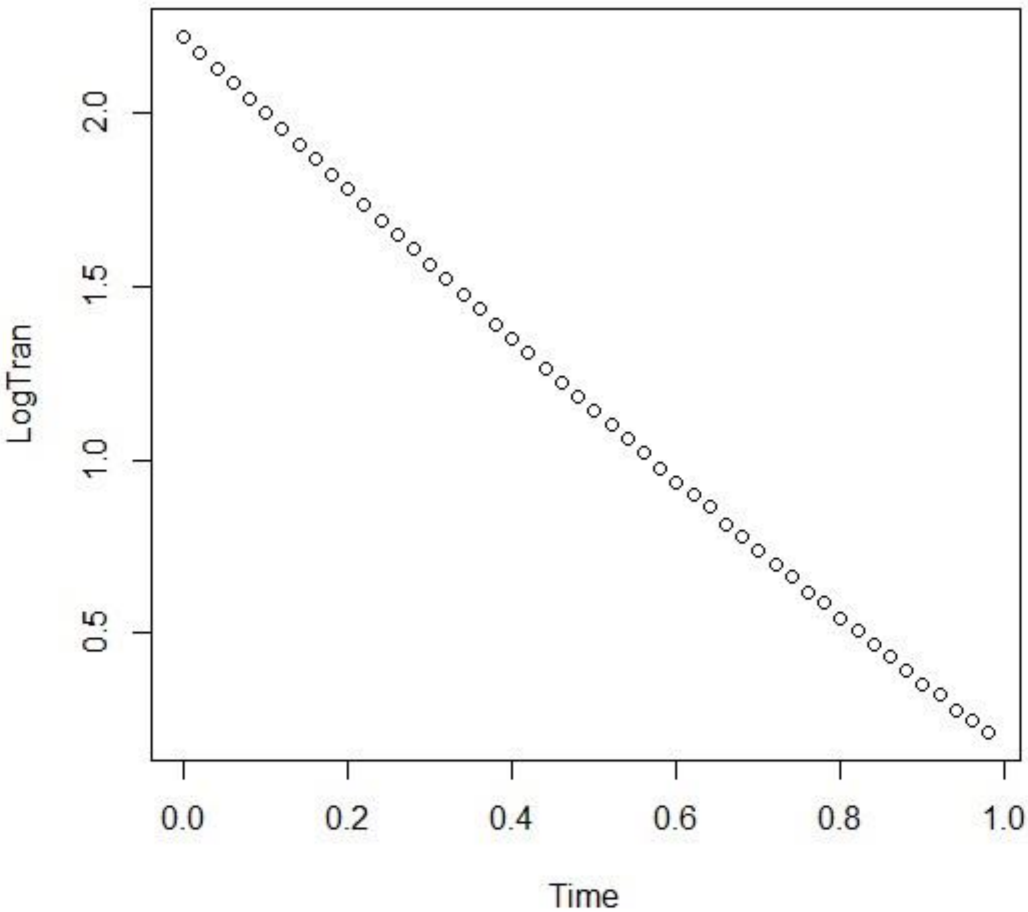
Part B



The Residuals vs fitted plot shows us curve that is nonlinear, which means this is not a good model.

Part C

The natural log of Voltage plotted against time gives us an linearly declining graph instead of an exponential decline.



2.0
1.5
1.0
0.5
LogTran
0.0
0.2
0.4
0.6
0.8
1.0
Time

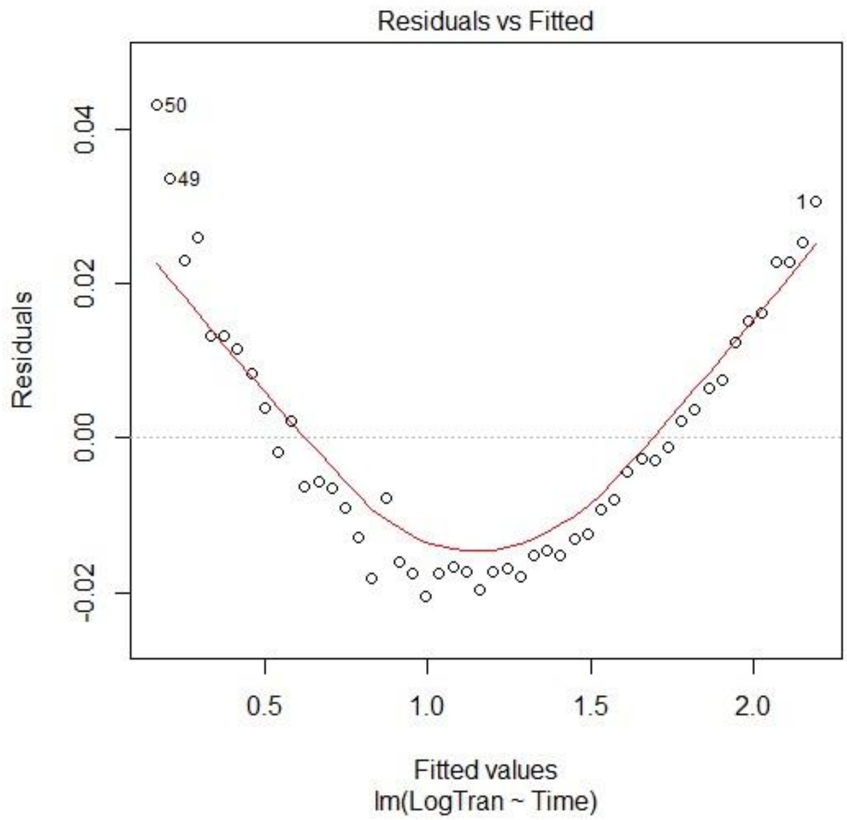
Part D

```
Call:
lm(formula = LogTran ~ Time)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.020448 -0.015084 -0.003621  0.012190  0.043212 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.189945   0.004637   472.3   <2e-16 ***
Time        -2.059065   0.008154  -252.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01664 on 48 degrees of freedom
Multiple R-squared:  0.9992,    Adjusted R-squared:  0.9992 
F-statistic: 6.377e+04 on 1 and 48 DF,  p-value: < 2.2e-16
```

According to the summary above, the equation for fitted regression model is:

LogVoltage hat = (2.189945) – (2.059065 x *Time*) Part

E



The residuals vs fitted graph is a curved instead of a linear. This means we have some outliers we can take out to make our model better.

Problem 1.28

1.28 Arctic sea ice. Climatologists have been measuring the amount of sea ice in both the Arctic and Antarctic regions for a number of years. The datafile **SeaIce** gives information about the amount of sea ice in the arctic region as measured in September (the time when the amount of ice is at its least) since 1979. The basic research question is to see if we can use time to model the amount of sea ice.

In fact, there are two ways to measure the amount of sea ice: *Area* and *Extent*. *Area* measures the actual amount of space taken up by ice. *Extent* measures the area inside the outer boundaries created by the ice. If there are areas inside the outer boundaries that are not ice (think about a slice of Swiss cheese), then the *Extent* will be a larger number than the *Area*. In fact, this is almost always true. Both *Area* and *Extent* are measured in 1,000,000 square km.

We will focus on the *Extent* of the sea ice in this exercise and see how it has changed over time since 1979. Instead of using the actual year as our explanatory variable, to keep the size of the coefficients manageable, we will use a variable called t, which measures time since 1978 (the value for 1979 is 1). SeaIce

a. Produce a scatterplot that could be used to predict *Extent* from t. Comment on the pattern.

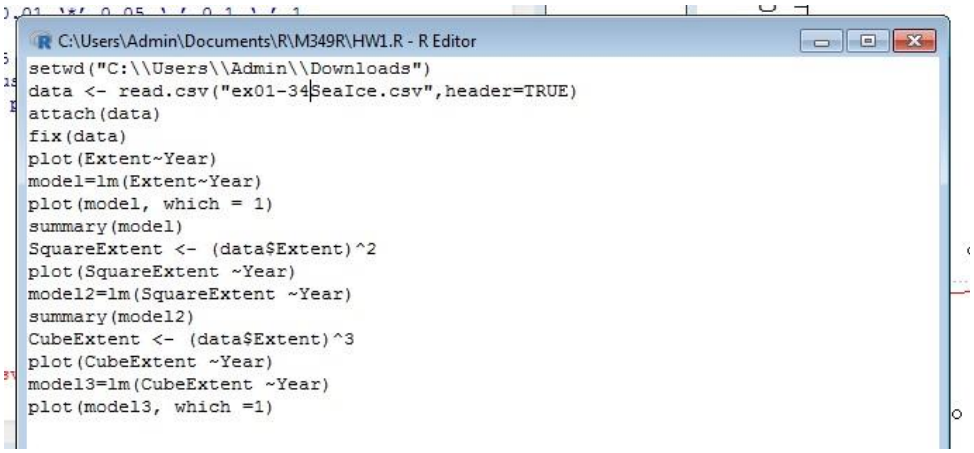
b. Create a residuals versus fits plot for predicting *Extent* from t. What does this plot tell you about the idea of fitting a linear model to predict *Extent* from t? Explain.

c. Transform *Extent* by squaring it. Plot $Extent^2$ versus t. Comment on the pattern.

d. Create a residuals versus fits plot for this new response variable using t. Discuss whether there is improvement in this plot over the one you created in part (b).

e. Redo parts (c) and (d) using the cube of *Extent*.

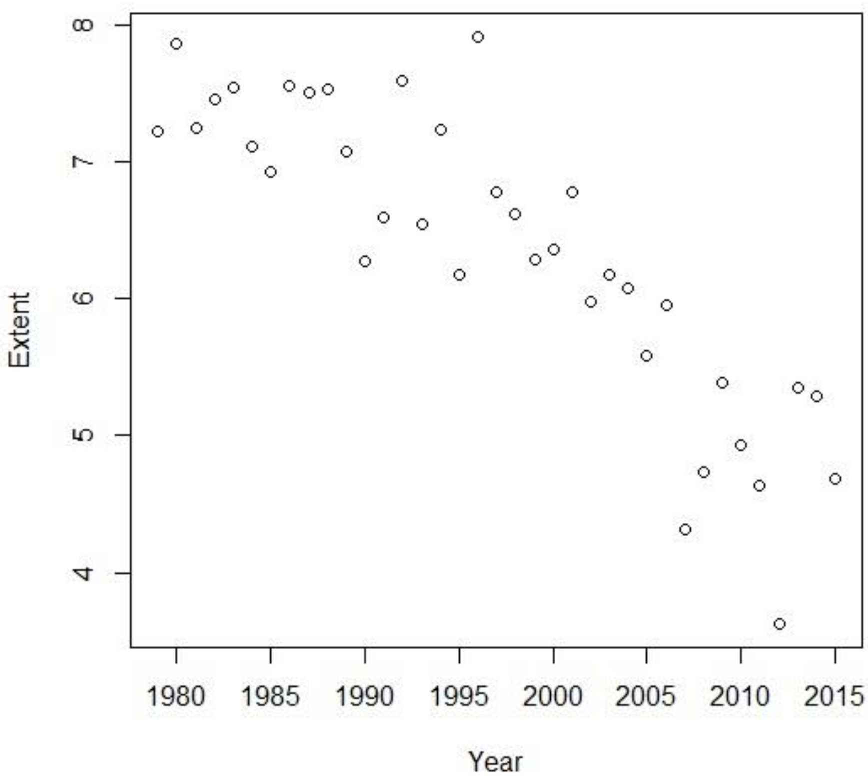
f. Would you be comfortable using a linear model for any of these three response variables? Explain.



C:\Users\Admin\Documents\R\M349R\HW1.R - R Editor

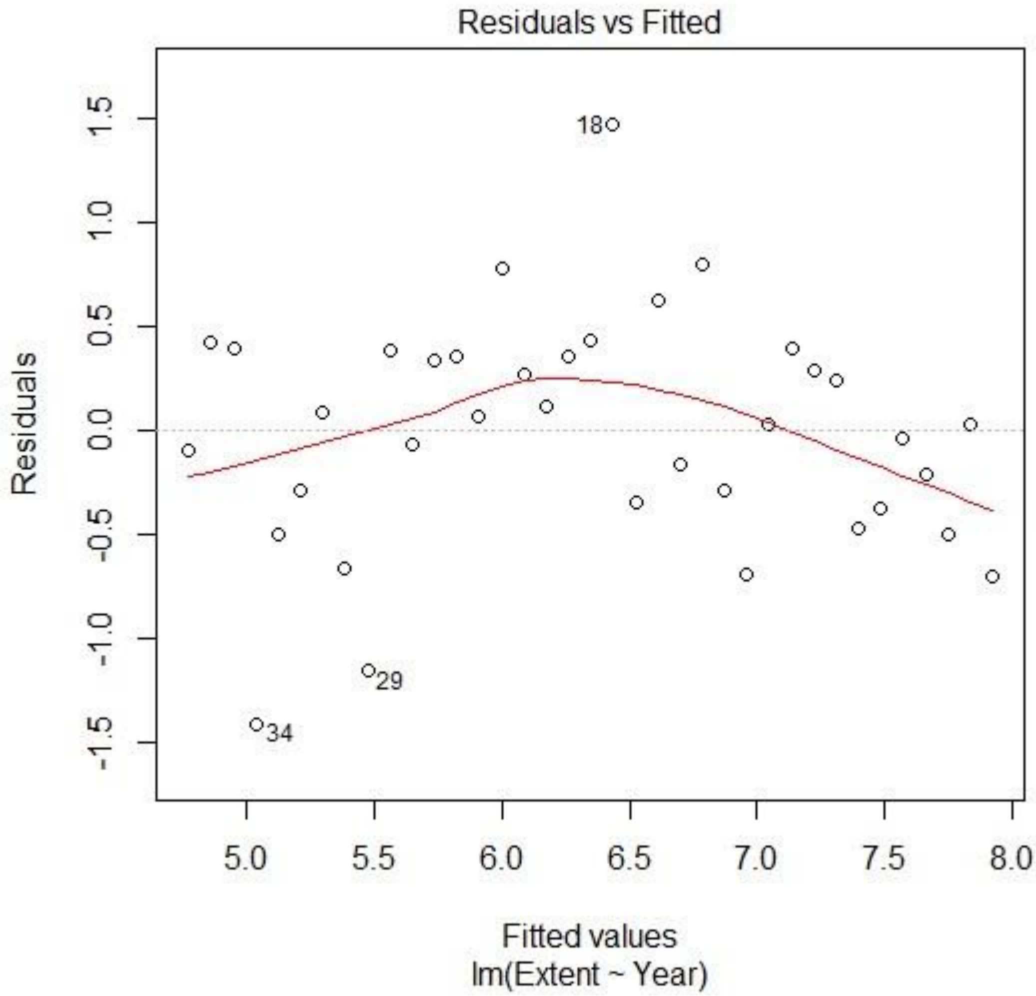
```
setwd("C:\\Users\\Admin\\Downloads")
data <- read.csv("ex01-34SeaIce.csv",header=TRUE)
attach(data)
fix(data)
plot(Extent~Year)
model=lm(Extent~Year)
plot(model, which = 1)
summary(model)
SquareExtent <- (data$Extent)^2
plot(SquareExtent ~Year)
model2=lm(SquareExtent ~Year)
summary(model2)
CubeExtent <- (data$Extent)^3
plot(CubeExtent ~Year)
model3=lm(CubeExtent ~Year)
plot(model3, which =1)
```

Part A



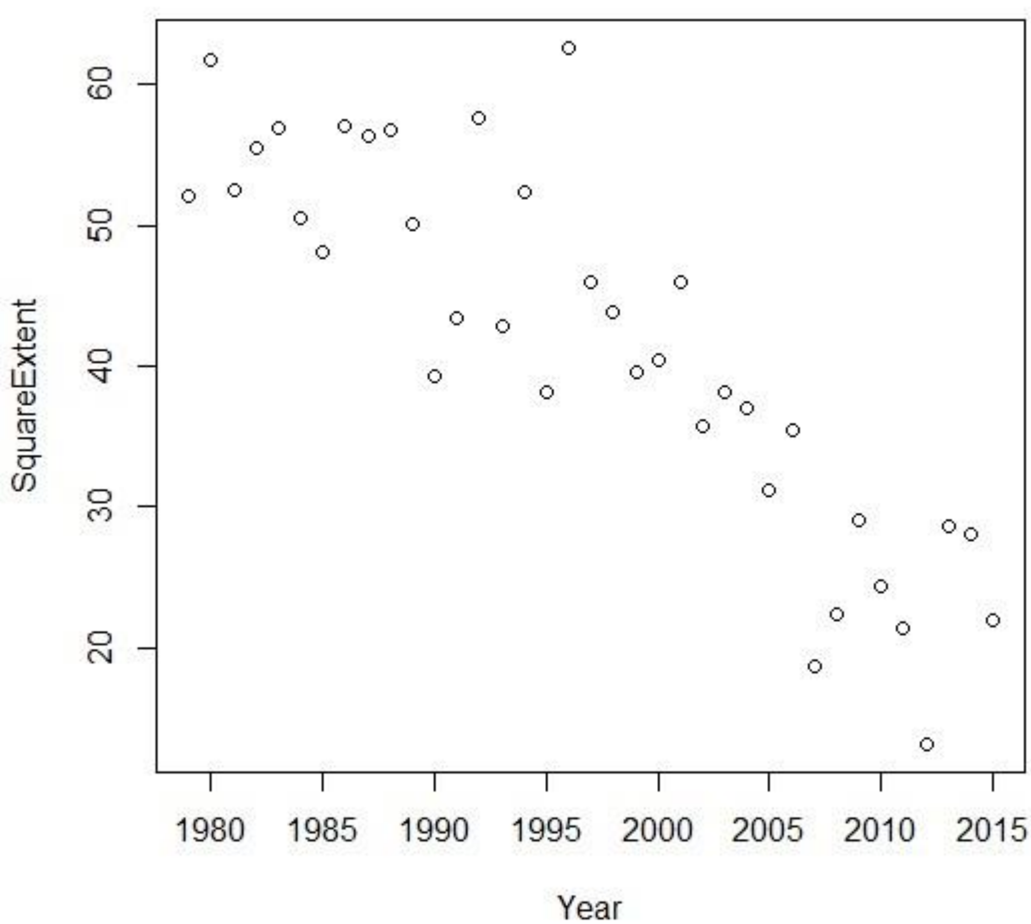
The scatterplot shows that the extent of the ice has been declining per decade. However, in the short term (inside a decade), Extent seems to go up and down.

Part B



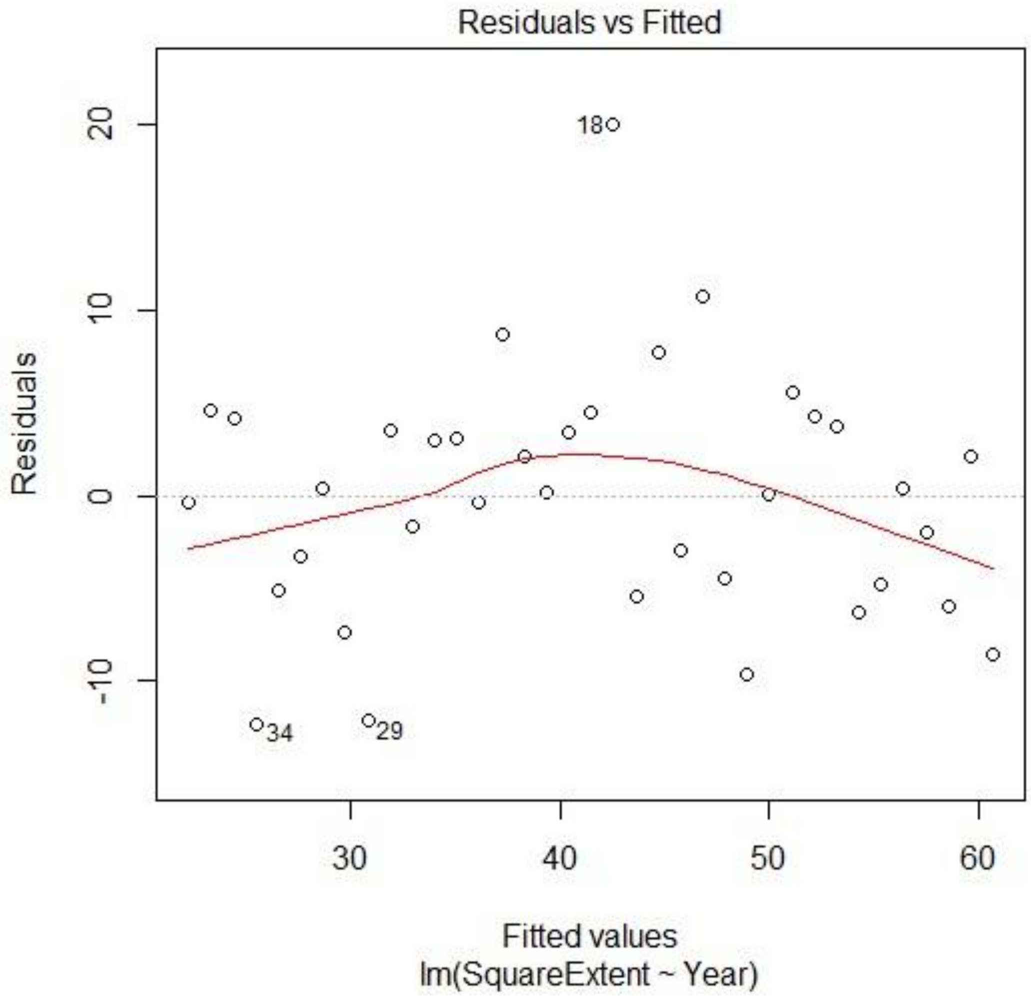
The residual vs fitted plot shows us that this is an okay model, as it seems to be linear in nature. This could become a better model if removed some outliers.

Part C



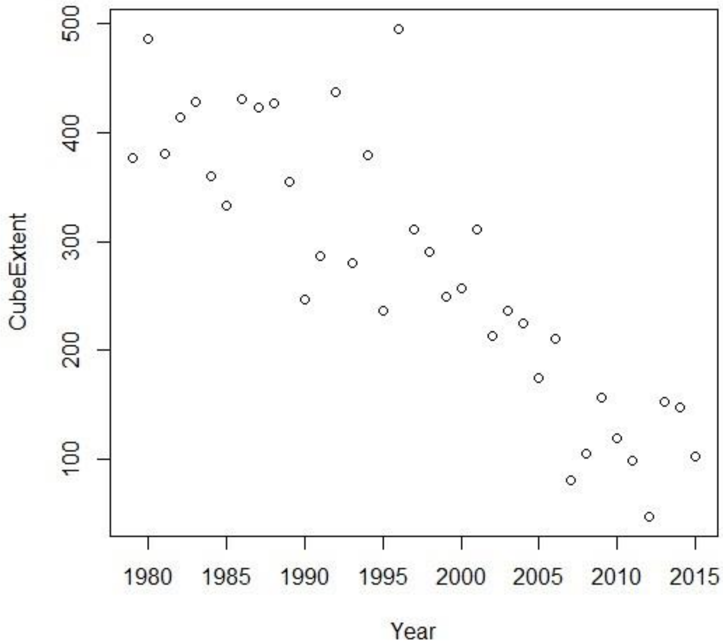
The decline seems much more linear than without squaring it.

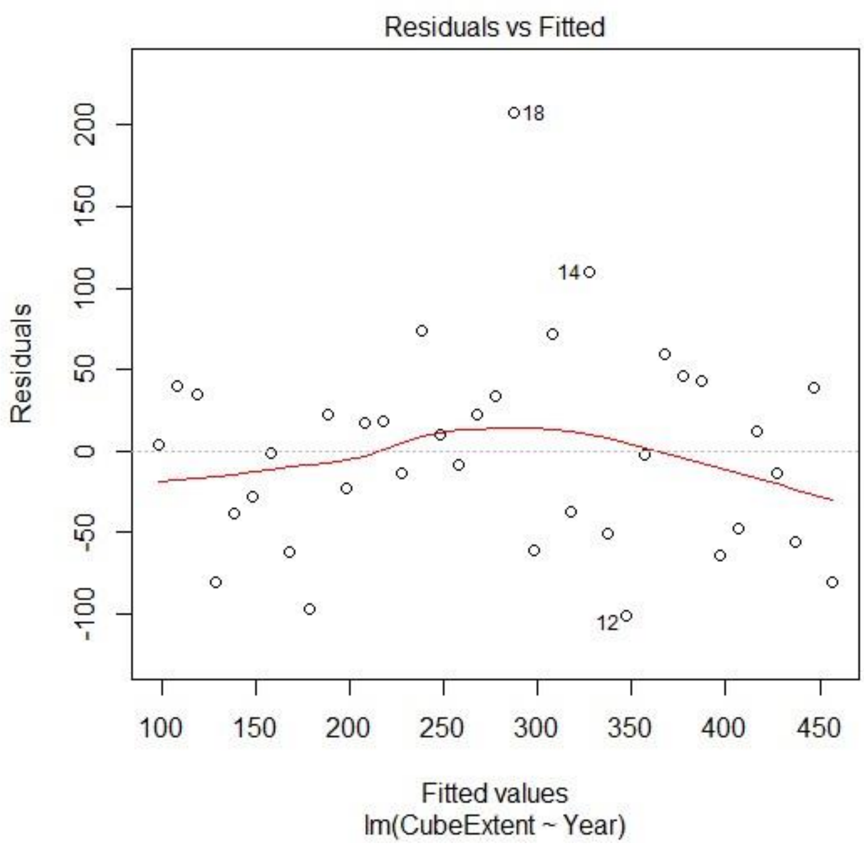
Part D



There is an improvement by squaring the extent, as the red line seems straighter.

Part E


CubeExtent
Year
1980
1985
1990
1995
2000
2005
2010
2015
100
200
300
400
500


Residuals vs Fitted
Residuals
Fitted values
lm(CubeExtent ~ Year)
18
14
12

By cubing Extent instead of squaring, there does not seem to be much improvement in scatterplot. However we can see in residual vs fitted plot that model3 is a little bit staighter. However, from residual vs fitted plot we can also see that our model3 is skewed, and could become better if we take out points 14, 18 and 12.

<u>Part E</u>

I would be comfortable using a linear model for both squaring extent (model2), and cubing Extent (model3). However, I would prefer model2 over model3.

Problem 1.34

1.34 Enrollment in mathematics courses. Total enrollments in mathematics courses at a small liberal arts college[9] were obtained for each semester from Fall 2001 to Spring 2012. The academic year at this school consists of two semesters, with enrollment counts for *Fall* and *Spring* each year as shown in Table 1.4. The variable

AYear indicates the year at the beginning of the academic year. The data are also provided in the file **MathEnrollment.** MthEnr

TABLE 1.4 Math enrollments

AYear	Fall	Spring
2001	259	246
2002	301	206
2003	343	288
2004	307	215
2005	286	230
2006	273	247
2007	248	308
2008	292	271
2009	250	285
2010	278	286
2011	303	254

a. Plot the mathematics enrollment for each semester against time. Is the trend over time roughly the same for both semesters? Explain.

b. A faculty member in the Mathematics Department believes that the fall enrollment provides a very good predictor of the spring enrollment. Do you agree?

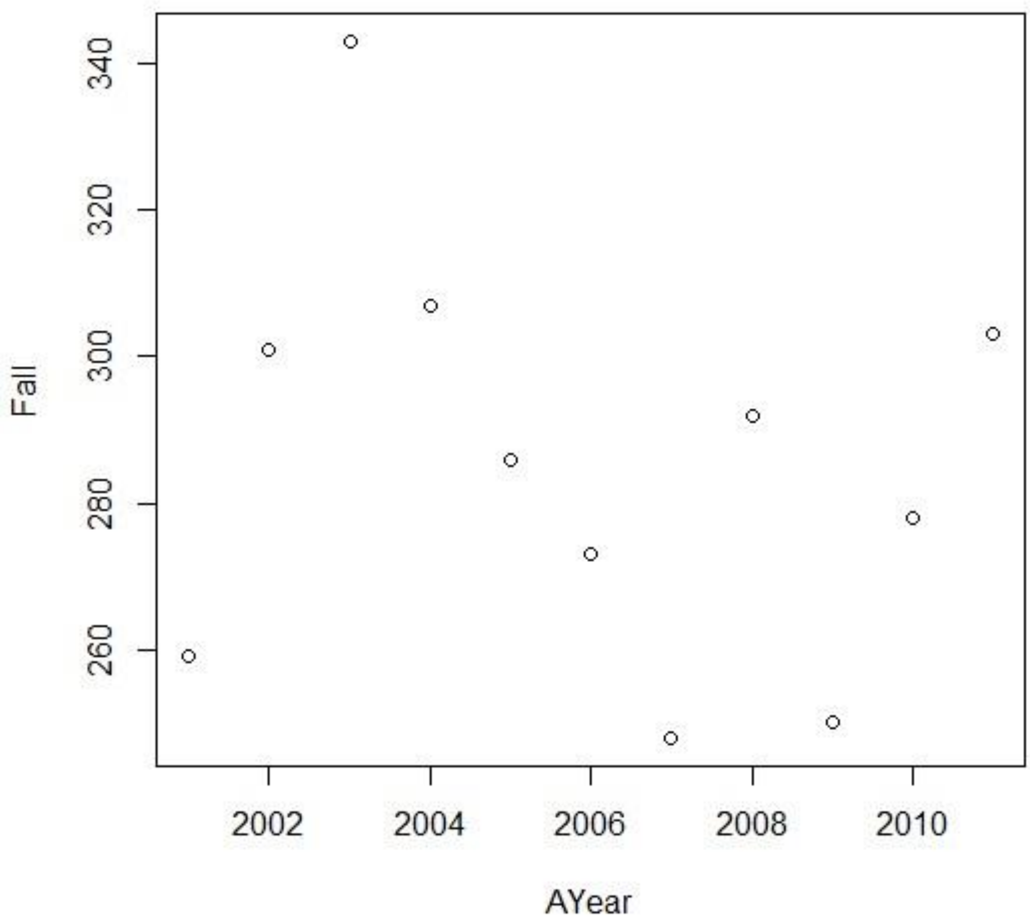
c. After examining a scatterplot with the least squares regression line for predicting spring enrollment from fall enrollment, two faculty members begin a debate about the possible influence of a particular point. Identify the point that the faculty members are concerned about.

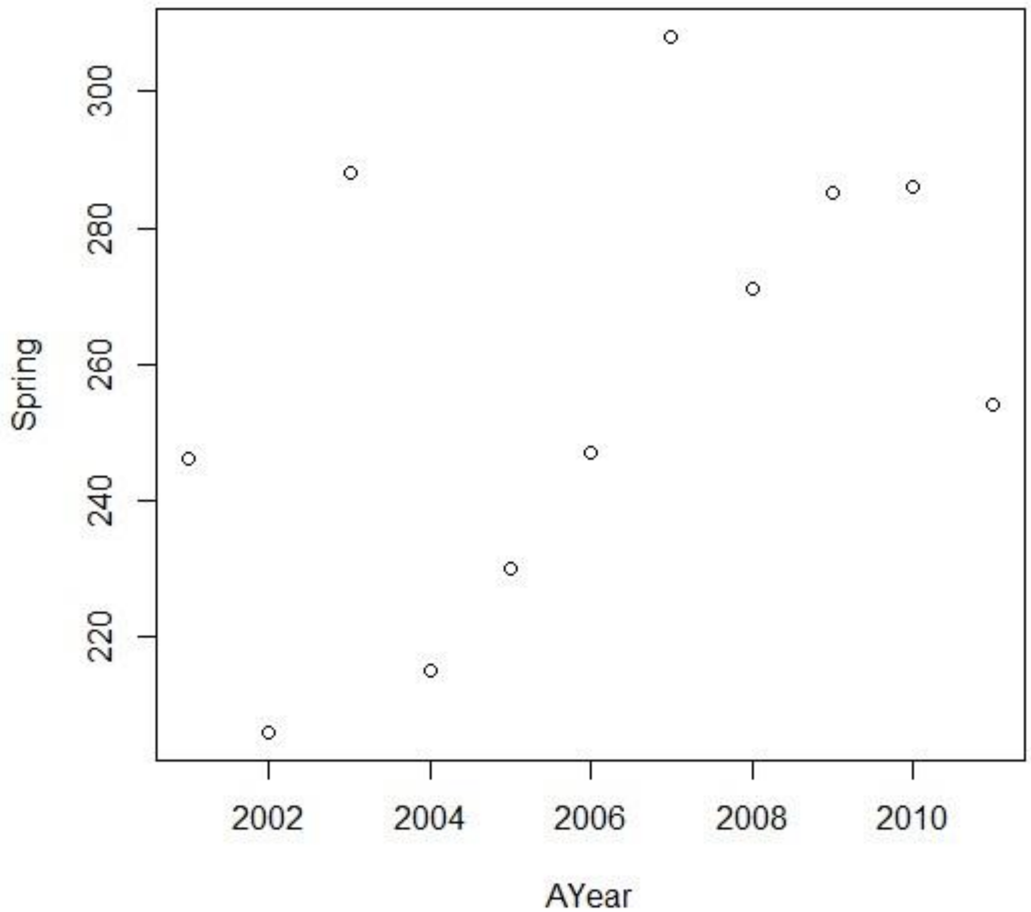
d. Fit the least squares line for predicting math enrollment in the spring from math enrollment in the fall, with and without the point you identified in part (c). Would you tag this point as influential? Explain.

C:\Users\Admin\Documents\R\M349R\HW1.R - R Editor

```
setwd("C:\\Users\\Admin\\Downloads")
data <- read.csv("ex01-34MthEnr.csv",header=TRUE)
attach(data)
fix(data)
plot(Fall~AYear)
plot(Spring~AYear)
model=lm(Spring~Fall, data=data)
summary(model)
plot(Fall~Spring)
abline(model)
plot(model, which =1)
```

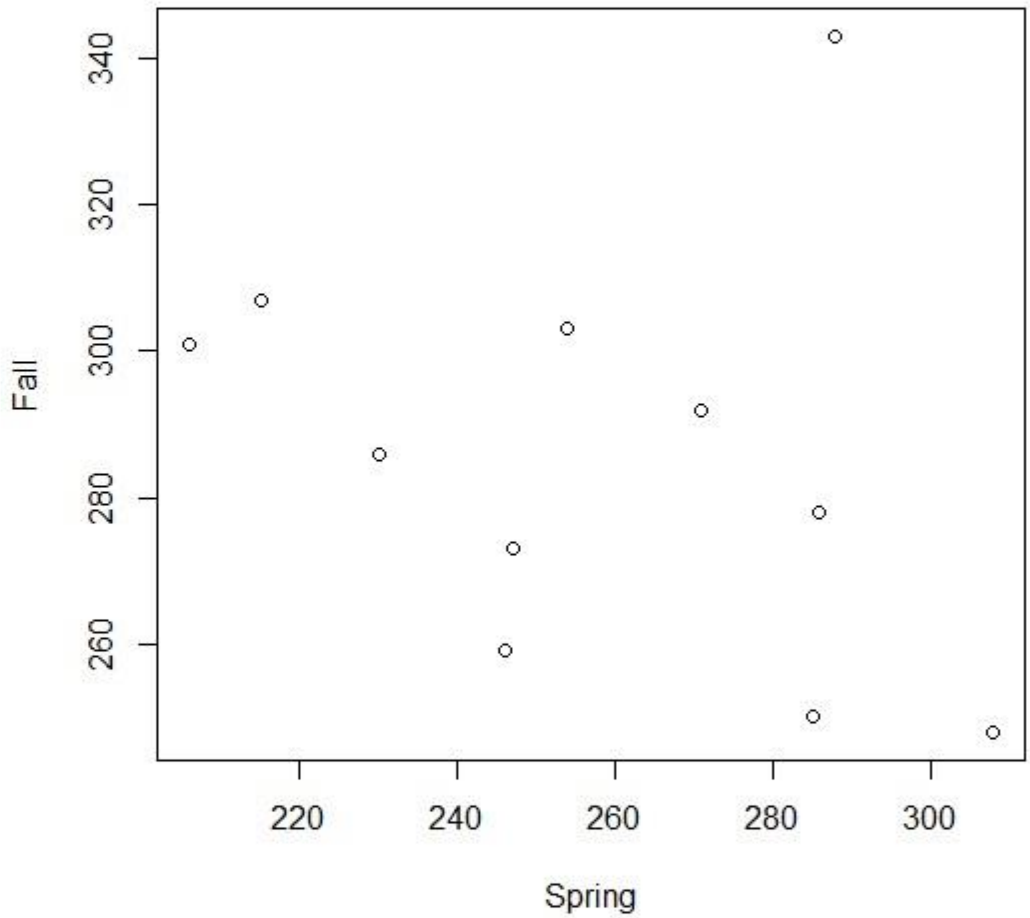
Part A





Both scatterplots seem that enrollments go up and down like a curve. It seems there are more students in fall, but the pattern is similar.

Part B



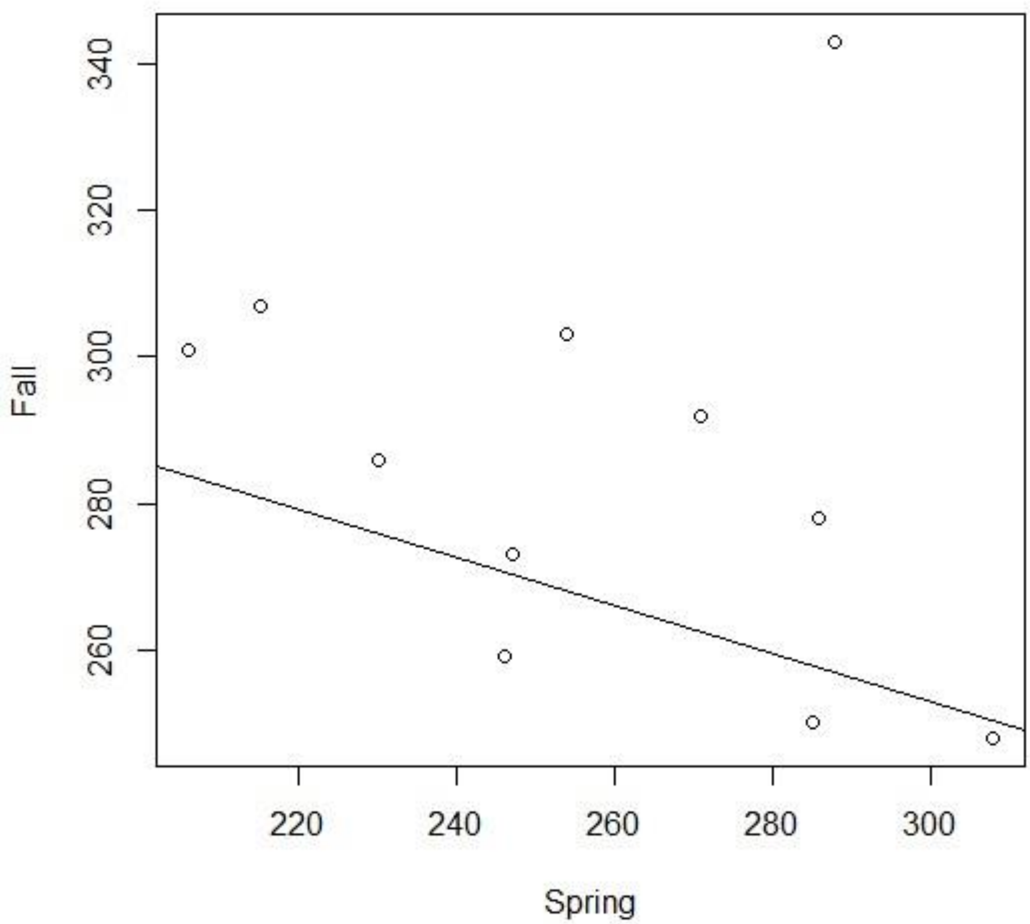

```
Call:
lm(formula = Spring ~ Fall)

Residuals:
    Min      1Q  Median      3Q     Max 
-46.740 -24.050   1.913  20.674  48.978 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) 351.0585   106.4710   3.297  0.00927 **
Fall         -0.3266     0.3713  -0.880  0.40195   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.09 on 9 degrees of freedom
Multiple R-squared:  0.07916,   Adjusted R-squared:  -0.02315 
F-statistic: 0.7737 on 1 and 9 DF,  p-value: 0.4019
```

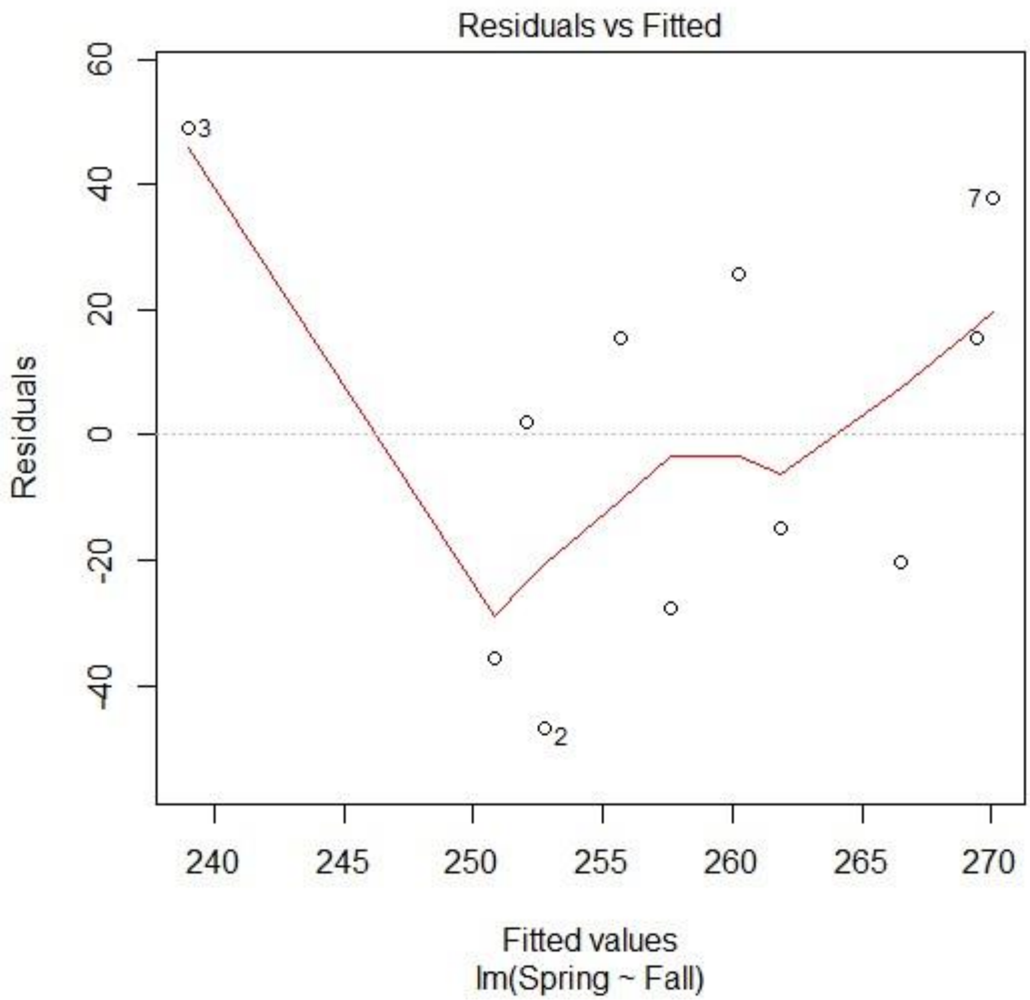
From the scatter plot of Fall vs Spring, we can see that there is no pattern. So, we can't confirm if fall enrollment is a good indicator of spring. However, from a linear model of spring against fall, we can see that spring enrollment goes down by 0.3266 x *Fall* enrollment.

Part C

I think they are talking about the year 2007, where spring enrollment was considerably higher than the fall.

Part D





Looking at only the scatterplots I would have eliminated 2007 data only. After, looking at residual vs fitted I would eliminate 2003, and 2002 also. So, yes, I would tag 2007 as influential.

Problem 1.44

1.44 Textbook prices. Two undergraduate students at Cal Poly took a random sample[13] of 30 textbooks from the campus bookstore in the fall of 2006. They recorded the price and number of pages in each book in order to investigate the question of whether the number of pages can be used to predict price. Their data are stored in the file **TextPrices** and appear in Table 1.5. **TxtPrc**

TABLE 1.5 Pages and price for textbooks

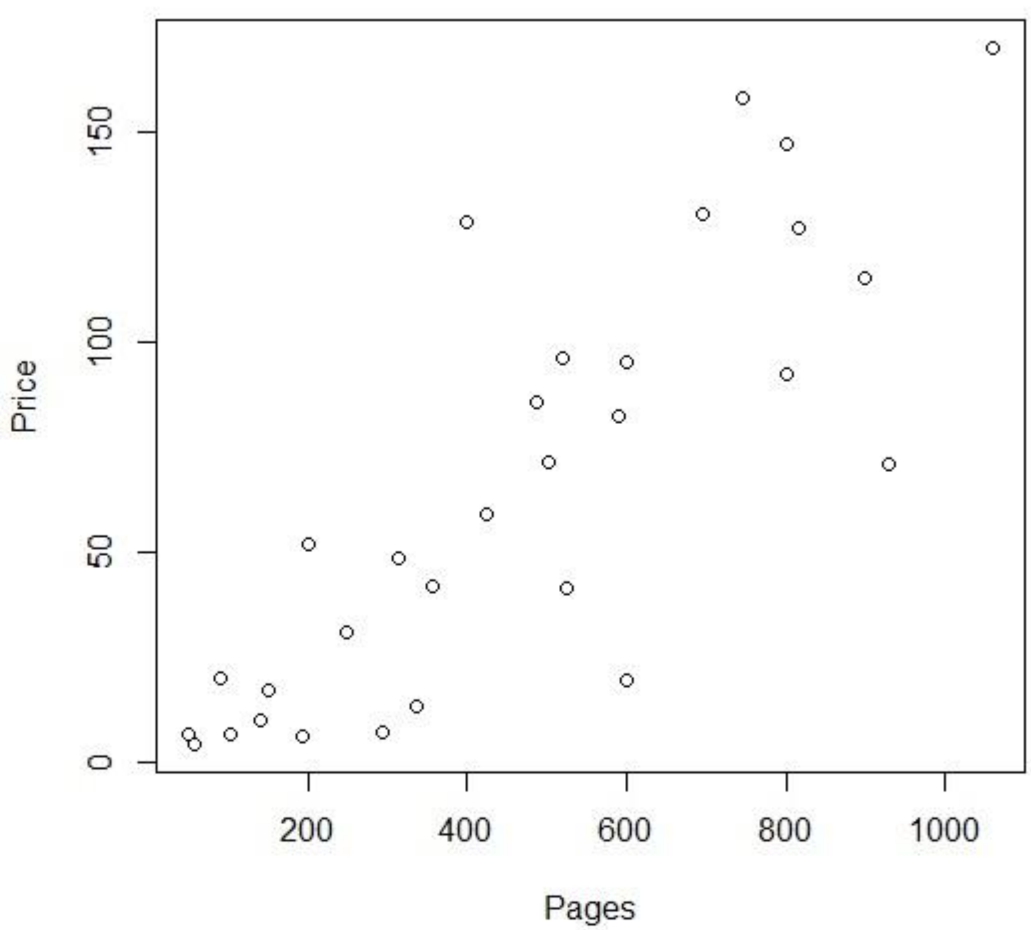
Pages	Price	Pages	Price	Pages	Price
600	95.00	150	16.95	696	130.50
91	19.95	140	9.95	294	7.00
200	51.50	194	5.95	526	41.25
400	128.50	425	58.75	1060	169.75
521	96.00	51	6.50	502	71.25
315	48.50	930	70.75	590	82.25
800	146.75	57	4.25	336	12.95
800	92.00	900	115.25	816	127.00
600	19.50	746	158.00	356	41.50
488	85.50	104	6.50	248	31.00

a. Produce the relevant scatterplot to investigate the students' question. Comment on what the scatterplot reveals about the question.

b. Determine the equation of the regression line for predicting price from number of pages.

c. Produce and examine relevant residual plots, and comment on what they reveal about whether the conditions for inference are met with these data.

Part A



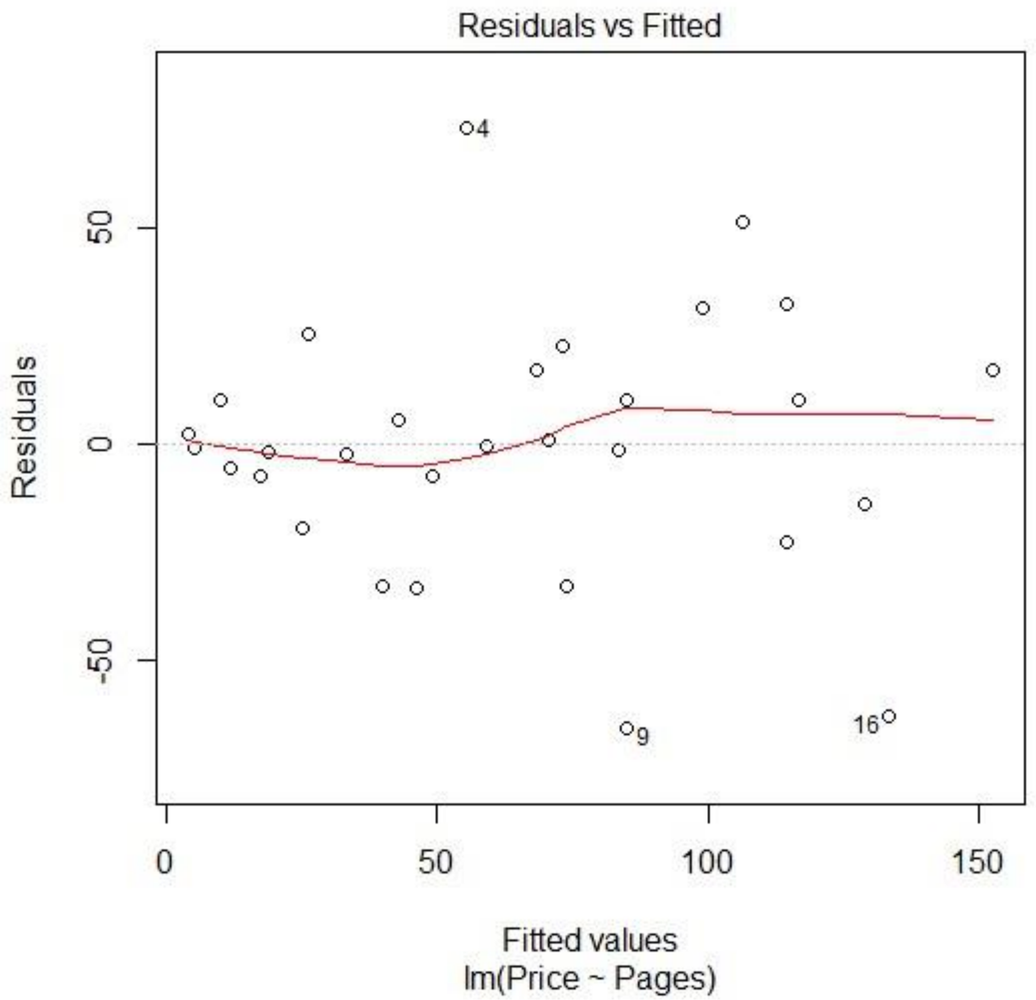
From scatterplot, we can see that in general there is a steady increase of price, as the number of pages increases.

```
Call:
lm(formula = Price ~ Pages, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-65.475 -12.324  -0.584  15.304  72.991 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.42231   10.46374  -0.327    0.746    
Pages        0.14733    0.01925   7.653 2.45e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.76 on 28 degrees of freedom
Multiple R-squared:  0.6766,	Adjusted R-squared:  0.665 
F-statistic: 58.57 on 1 and 28 DF,  p-value: 2.452e-08
```

According to the summary above, the equation for fitted regression model is:

Price hat = -3.42231 + 0.14733 x (*Pages*)

Part C



The red line is acceptable level of straight, confirming that this is a good linear model. So, we can conclude that as pages increases price is likely to increase.