HW1

Markdown for HW1 of Applied Regression and Time Series course.

1.1 Equation of a line

C

1.2 Residual plot to check conditions

C

1.3 Sparrows slope

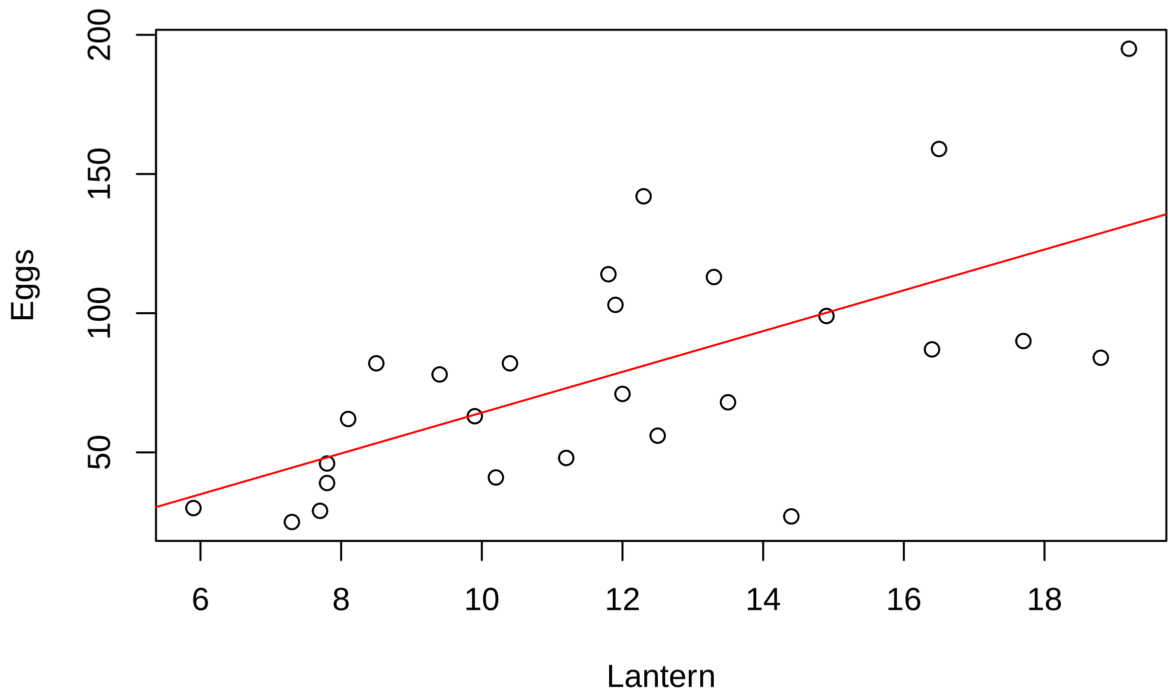
$\hat{B}_1 = .4674$

1.16 Glow-worms

```
data = read.csv("C:/Users/adhri/OneDrive/Documents/R/App_Reg_and_Time_Series/chpt1/datasets/GlowWorms.cs
attach(data)
plot(Eggs~Lantern)

#linear model
model = lm(Eggs~Lantern)
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
```

```
#create line
abline(model, col = 'red')
```



```
#predict num of eggs
predict(model, data.frame(Lantern = 14))
```

```
##       1
## 93.5751
```

a) $\hat{Eggs} =$ -8.977 + 7.325*Lantern

b) For every 1 mm increase in female lantern size, the number of eggs laid increases by approximately on the average 7.325.

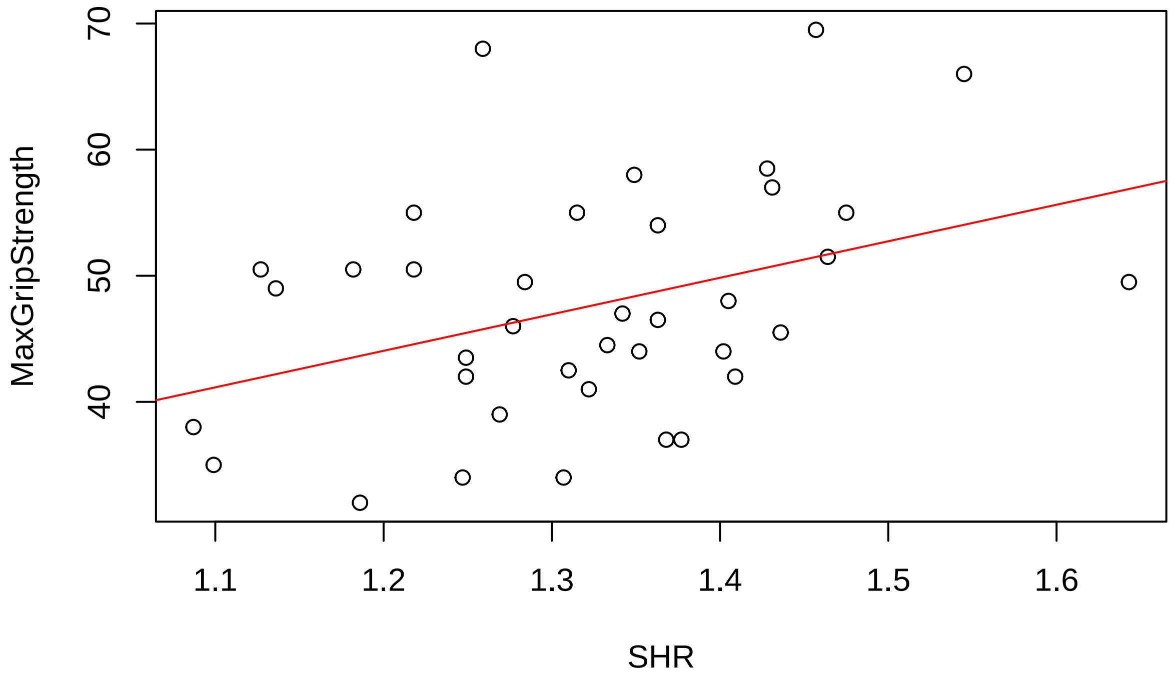
c) If the glow-worm has a lantern size of 14 mm, the predicted number of eggs she will lay is 93.5751, or rounded down to 93 eggs.

1.18 Male body Measurements

```
data1 = read.csv("C:/Users/adhri/OneDrive/Documents/R/App_Reg_and_Time_Series/chpt1/datasets/Faces.csv"
attach(data1)
plot(MaxGripStrength ~ SHR)
model1 = lm(MaxGripStrength ~ SHR)
summary(model1)
```

```
##
## 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
```

```
abline(model1, col = 'red')
```



```
predict(model1, data.frame(SHR = 1.5))
```

```
##        1
## 52.73736
```

a) $Max\hat{Grip}Strength$ = 9.298 + 28.959*SHR

b) For every additional 1 increase in SHR, the MaxGripStrength increases by approximately on the average by 28.959 kilograms.

c) If the SHR is 1.5, the MaxGripStrength of the man is predicated to be 52.73736 or 52.7 when rounded like the data.

1.20 Houses in Grinnell

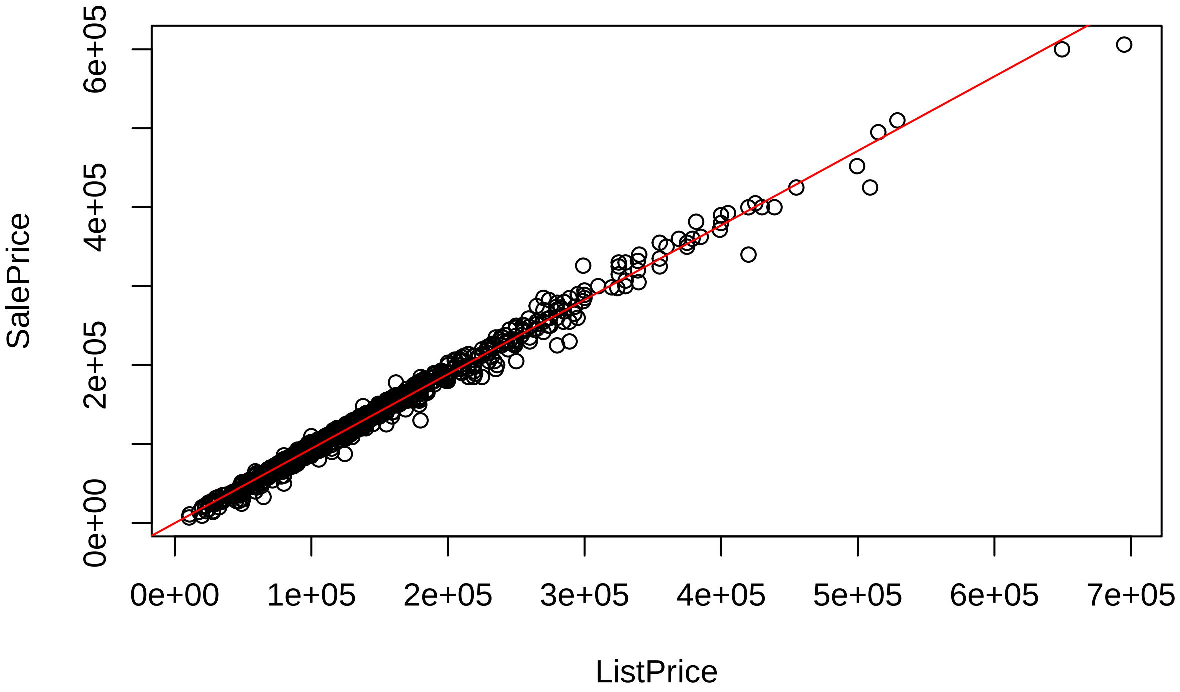
```
data2 = read.csv("C:/Users/adhri/OneDrive/Documents/R/App_Reg_and_Time_Series/chpt1/datasets/GrinnellHou
attach(data2)
names(data2)  # names of data columns
```

```
##  [1] "Date"         "Address"      "Bedrooms"     "Baths"        "SquareFeet"
##  [6] "LotSize"      "YearBuilt"    "YearSold"     "MonthSold"    "DaySold"
## [11] "CostPerSqFt" "OrigPrice"    "ListPrice"    "SalePrice"    "SPLPPct"
```

```
#a
plot(SalePrice ~ ListPrice)

model2 = lm(SalePrice ~ ListPrice)
summary(model2)
```

```
##
## Call:
## lm(formula = SalePrice ~ ListPrice)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -55942  -3275    846   4141  44168
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.448e+02  5.236e+02  -0.277    0.782
## ListPrice    9.431e-01  3.201e-03 294.578   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 8019 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
```

```
abline(model2, col = 'red')
```



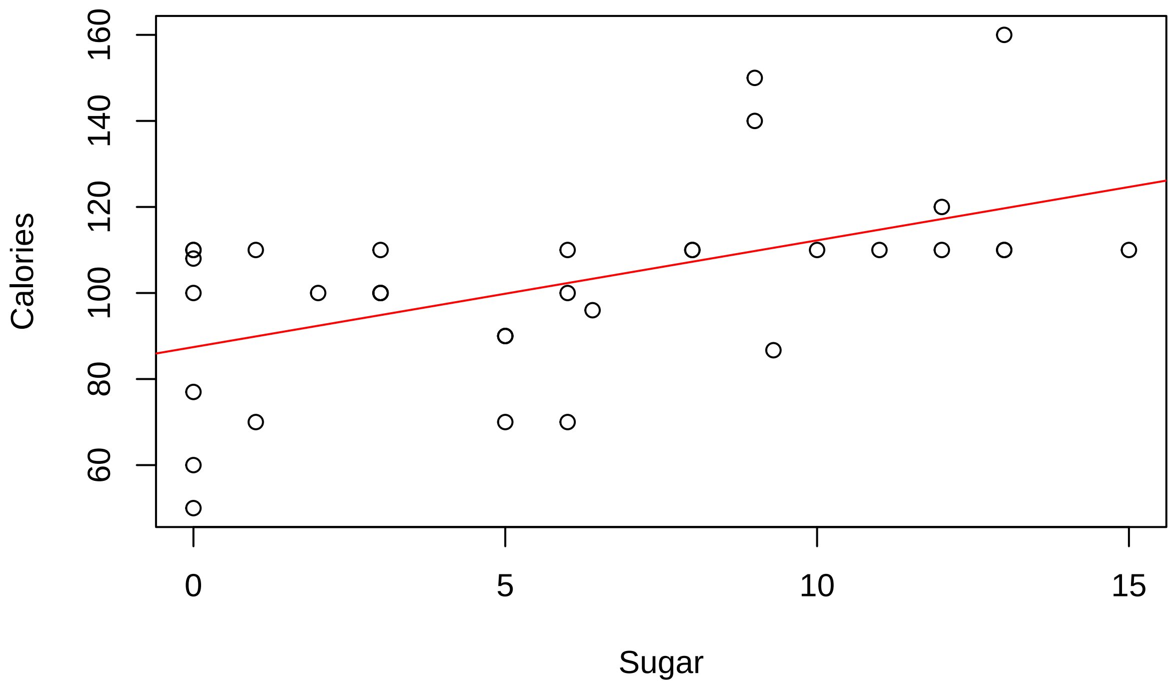
a) There appears to be a positive linear relationship between the ListPrice and SalePrice. THere is also a high density of data points at the ower listed and sold priced houses. This could be due to the higher frequency of less expensive houses.

b) $\hat{SalePrice}$ = -144.8 + .9431*ListPrice

c) For every additional $1 on the ListPrice, the SalePrice increases by approximatedly on the average by .9431 dollars.

1.21 Breakfast Cereal

```
data3 = read.csv("C:/Users/adhri/OneDrive/Documents/R/App_Reg_and_Time_Series/chpt1/datasets/Cereal.csv
attach(data3)
plot(Calories ~ Sugar)
model3 = lm(Calories ~ Sugar)
abline(model3, col = 'red')
```



```
predict(model3, data.frame(Sugar = 10))
```

```
##        1
## 112.2358
```

```
summary(model3)
```

```
##
## 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
```

```
#predict calories for just 1 gram of sugar to find residual for Cheerios data point
predict(model3, data.frame(Sugar = 1))
```

```
##       1
## 89.9085
```

a) $\widehat{Calories}$ = 87.4277 + 2.4808*Sugar; A cereal with 10 grams of sugar is predicted to have 112.2358 calories or when rounded like the data, 112.0 calories.

b) The residual for the Cheerios data point is 20.0915 (110 - 89.91).

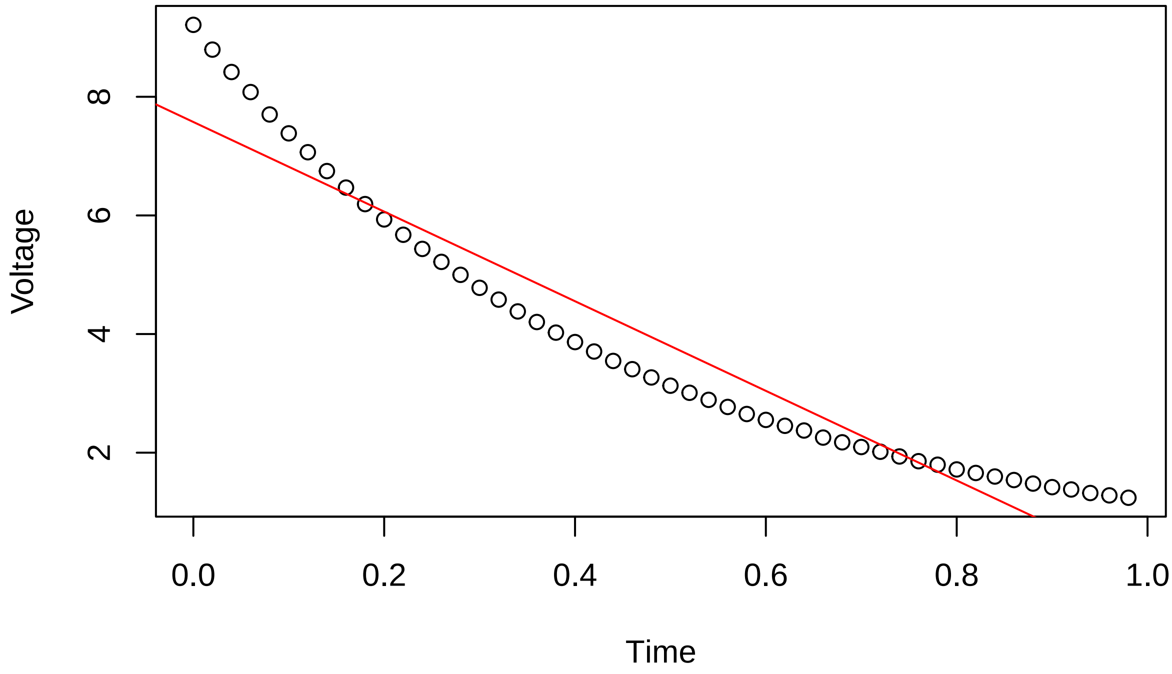
c) The data is very widely spready out and has a relatively weak positive relationship/slope, so it might not be the best summary of the relationship between calories and sugar content. The residual standard error is also 19.27 which is quite high.

1.27 Capacitor voltage

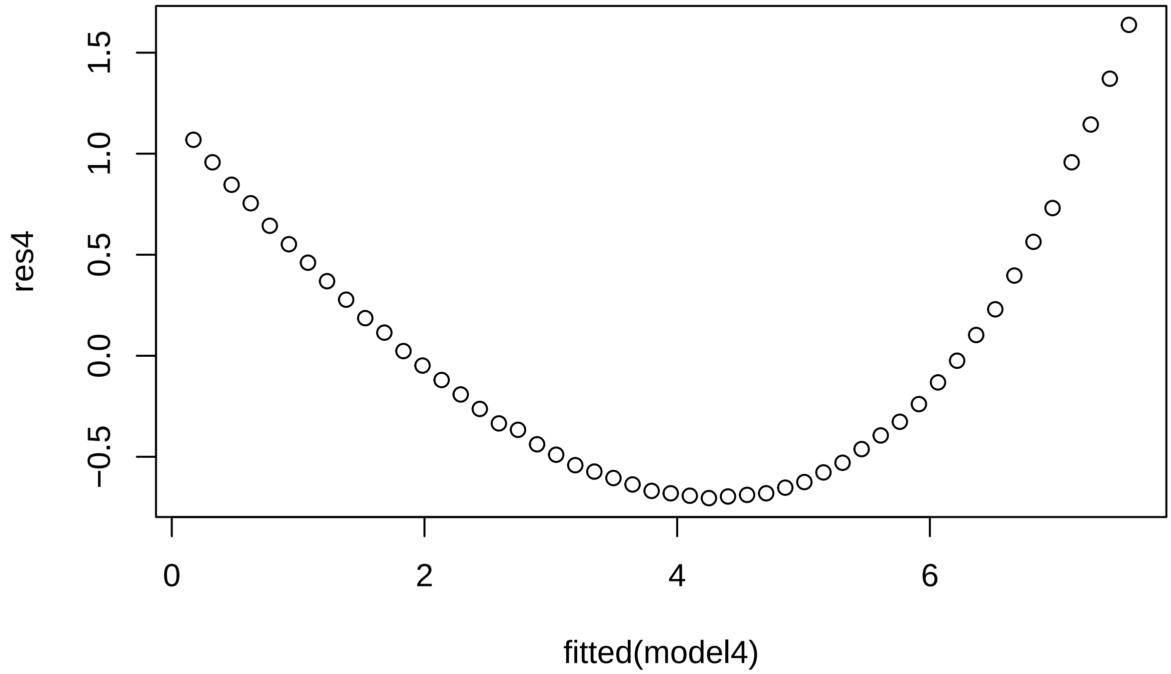
```
data4 = read.csv("C:/Users/adhri/OneDrive/Documents/R/App_Reg_and_Time_Series/chpt1/datasets/Volts.csv"
attach(data4)

plot(Voltage ~ Time)
model4 = lm(Voltage ~ Time)
summary(model4)
```

```
##
## Call:
## lm(formula = Voltage ~ Time)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7046 -0.5655 -0.1618  0.4446  1.6375
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.5753     0.1793   42.24   <2e-16 ***
## Time         -7.5549     0.3154  -23.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6436 on 48 degrees of freedom
## Multiple R-squared:  0.9228, Adjusted R-squared:  0.9212
## F-statistic: 573.9 on 1 and 48 DF,  p-value: < 2.2e-16
```

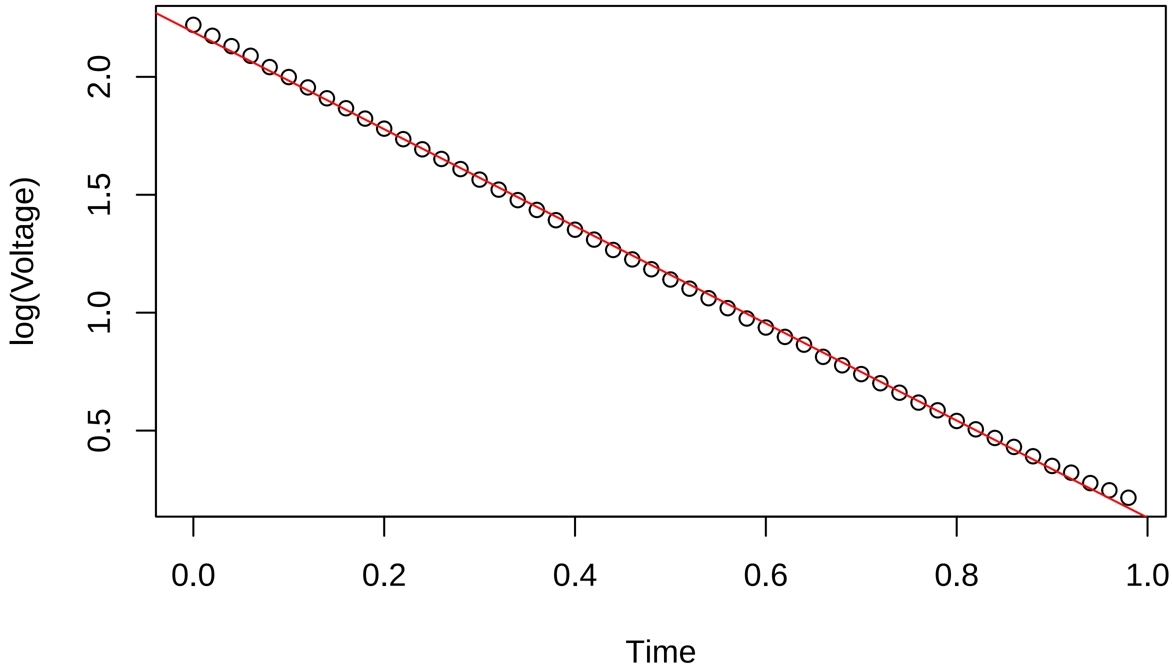
```
abline(model4, col = 'red')
```



```
# get list of residuals
res4 = resid(model4)

# create residual vs fitted plot
plot(fitted(model4), res4)
```



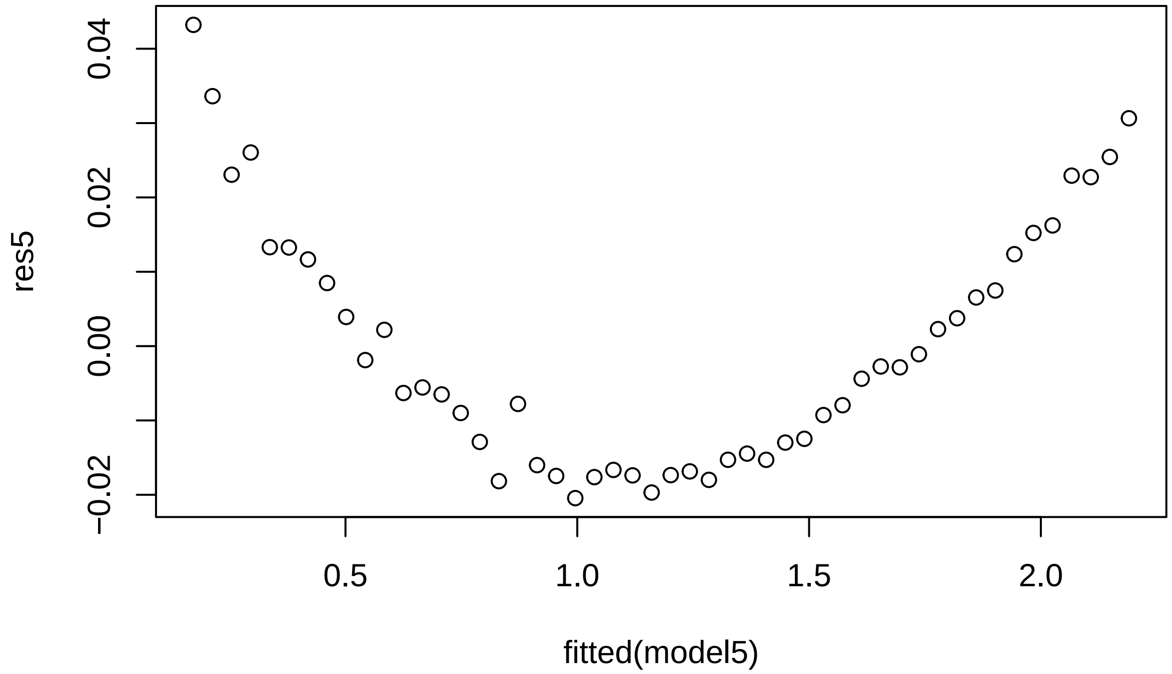
```
# transformed log Voltage and plot versus time
plot(log(Voltage) ~ Time)
model5 = lm(log(Voltage) ~ Time)
abline(model5, col = 'red')
```



```
summary(model5)
```

```
##
## Call:
## lm(formula = log(Voltage) ~ 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
```

```
# create residual plot
res5 = resid(model5)
plot(fitted(model5), res5)
```



a) The scatterplot has a negatively exponential pattern in decreasing voltage. The data has a curved, decreasing trend.

b) The residuals versus fits plot is extremely curved which indicates a lack of linearity in the data. This means a linear model will not do well in predicting Voltage from Time.

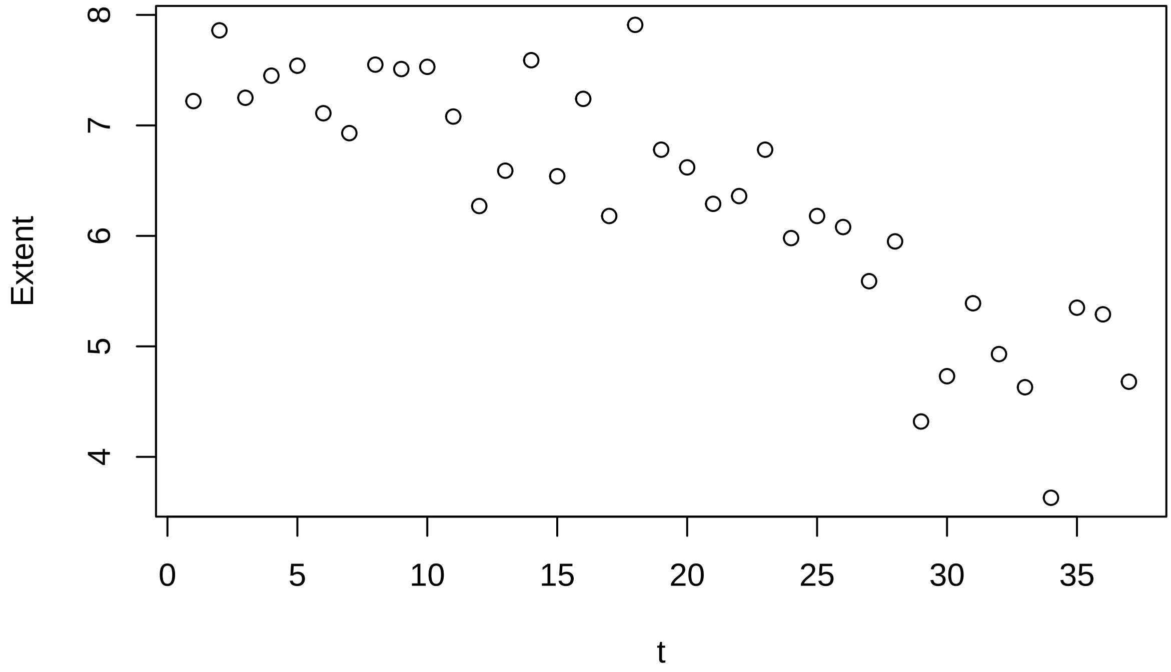
c) The pattern in the data has become more linear.

d) $log(\hat{Voltage})$ = 2.189945 + (-2.059065)*Time

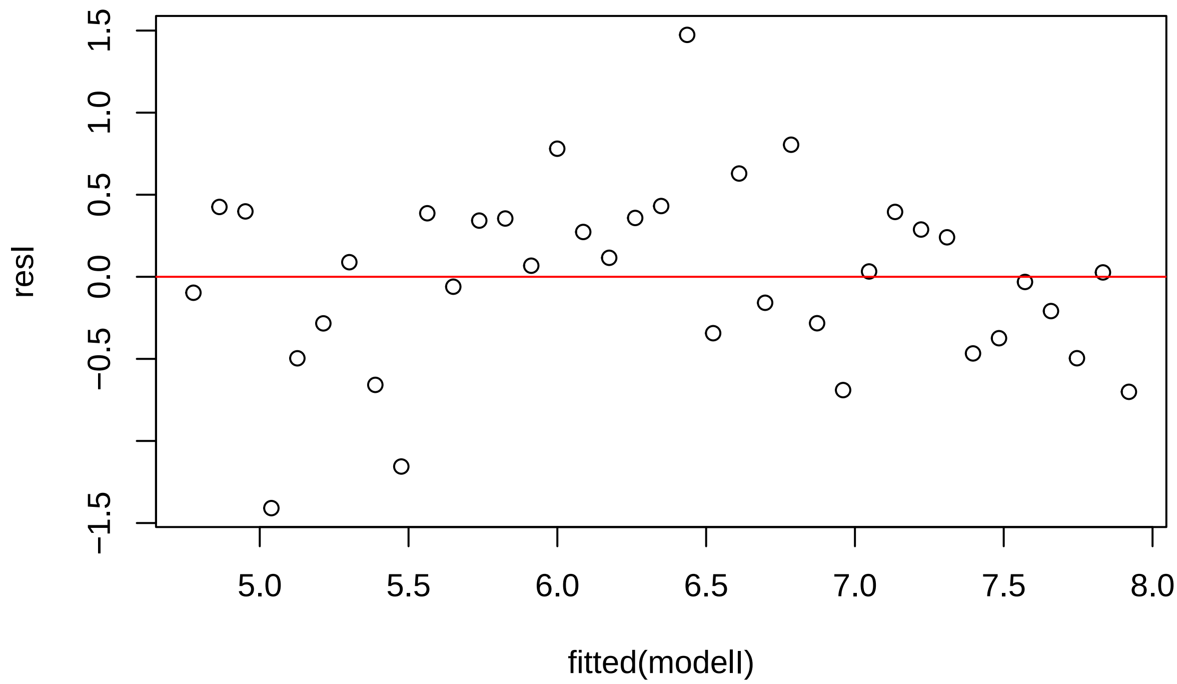
e) The plot of the residuals versus fitted values is still very curved which still indicates lack of linearity in the data despite transformation via logarithm.

1.28 Arctic sea ice

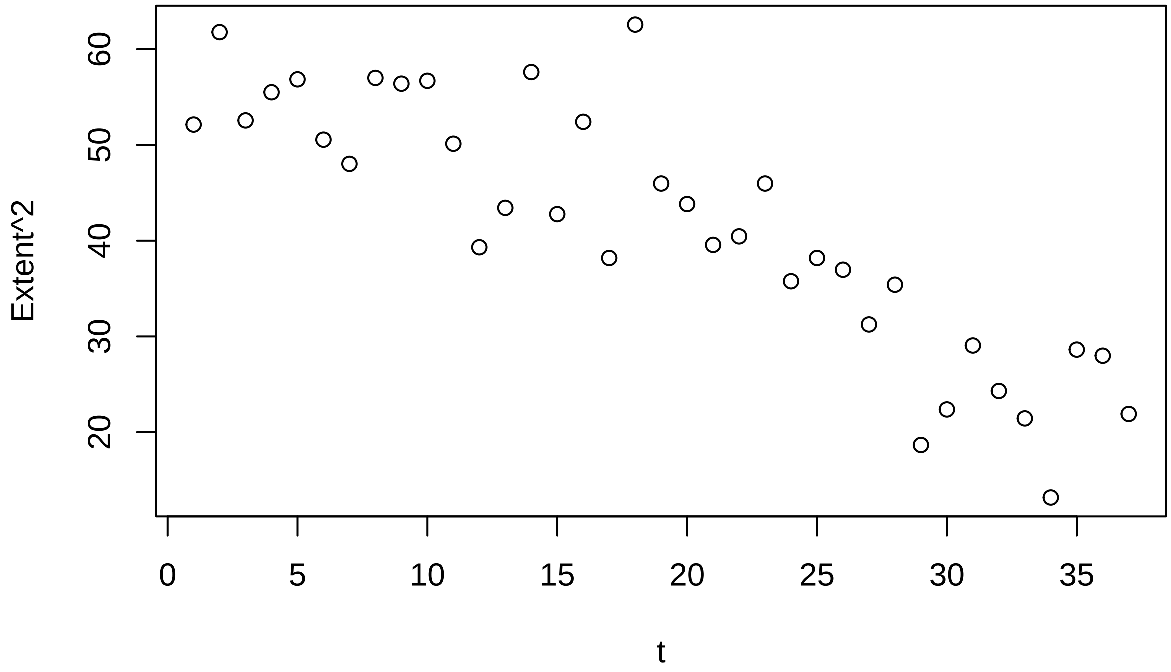
```
SeaIce = read.csv("C:/Users/adhri/OneDrive/Documents/R/App_Reg_and_Time_Series/chpt1/datasets/SeaIce.cs
attach(SeaIce)
plot(Extent ~ t)
```



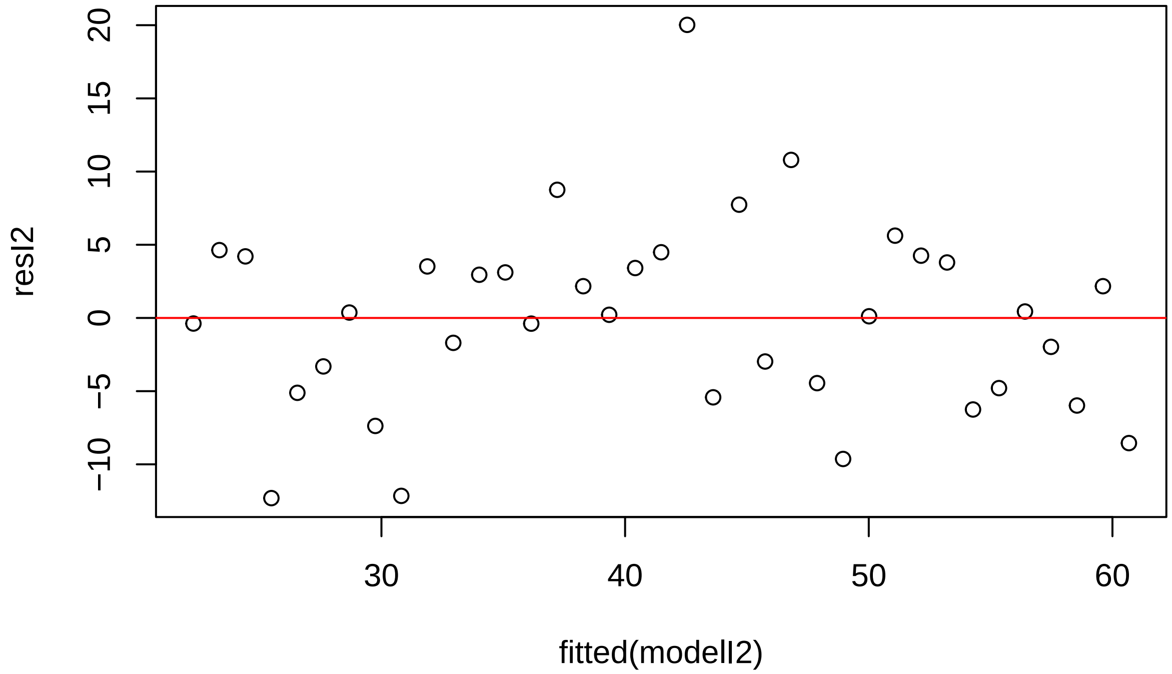
```
modelI = lm(Extent ~ t)

# residual plot
resI = resid(modelI)
plot(fitted(modelI), resI)
abline(h = 0, col = 'red')
```



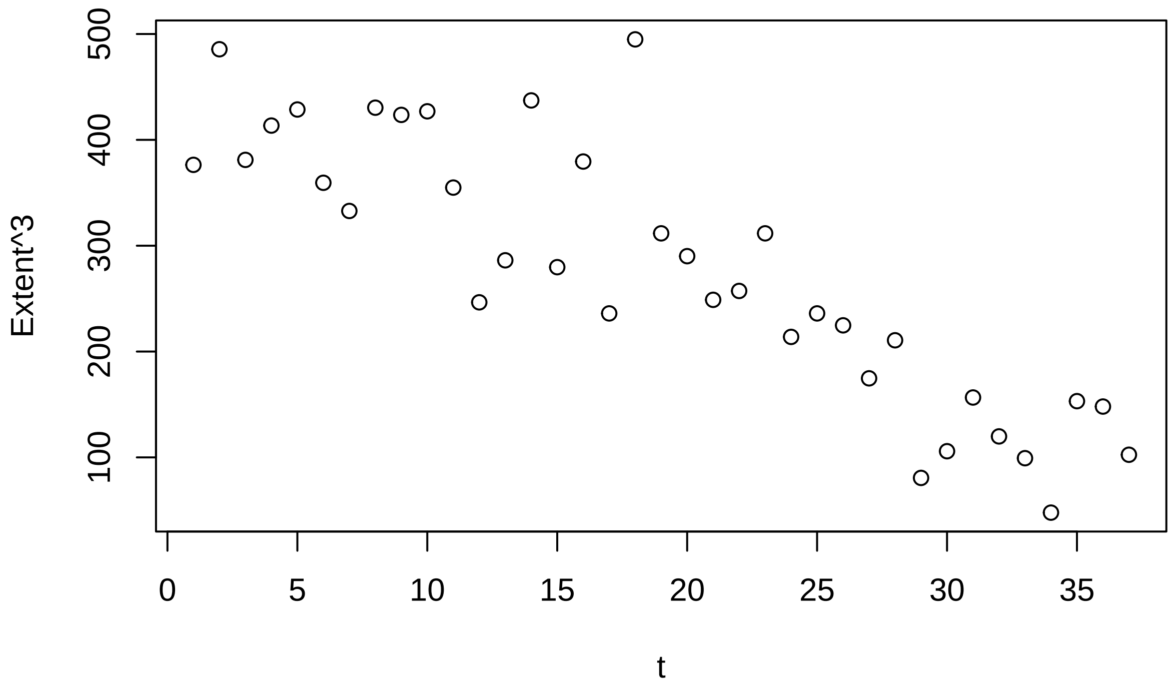
```
# transformed squared plot
plot(Extent^2 ~ t)
```



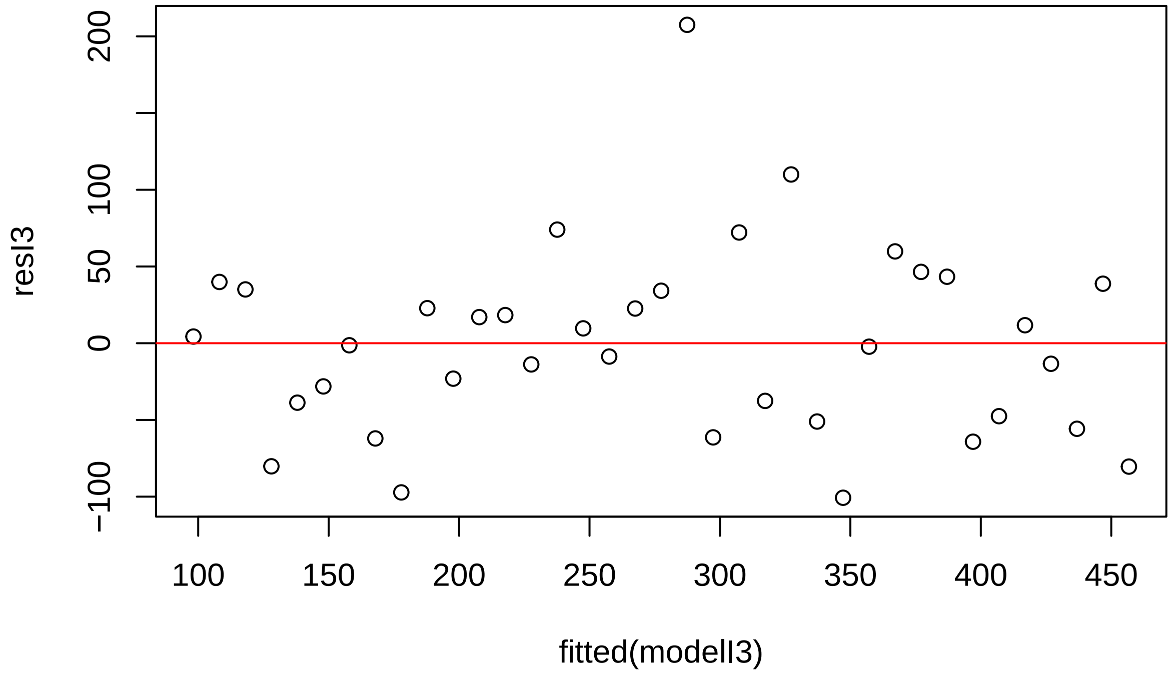
```
modelI2 = lm(Extent^2 ~ t)

# residual plot transformed
resI2 = resid(modelI2)
plot(fitted(modelI2), resI2)
abline(h = 0, col = 'red')
```



```
# residual plot transformed cubed
plot(Extent^3 ~ t)
```



```
modelI3 = lm(Extent^3 ~ t)
resI3 = resid(modelI3)
plot(fitted(modelI3), resI3)
abline(h = 0, col = 'red')
```



a) The scatterplot of data points has a slight curve.

b) The residual plot violates the linearity and constant variance condition, thus fitting a linear model will not be a good idea.

c) The pattern is straighter, but it is still curved.

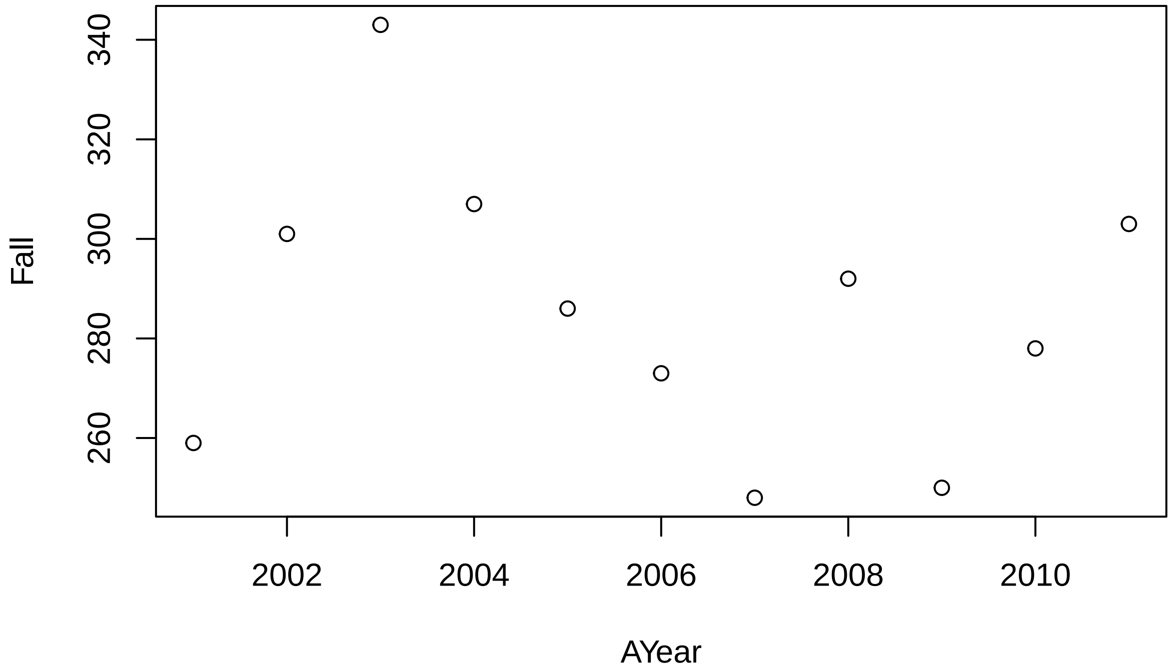
d) There is improvement in the residual plot, but the linearity condition is still not passed.

e) The pattern of the transformed data is less curved and more linear. There is improvement in the residual plot and it is most likely to pass the linearity condition out of all of the residual plots we've created.

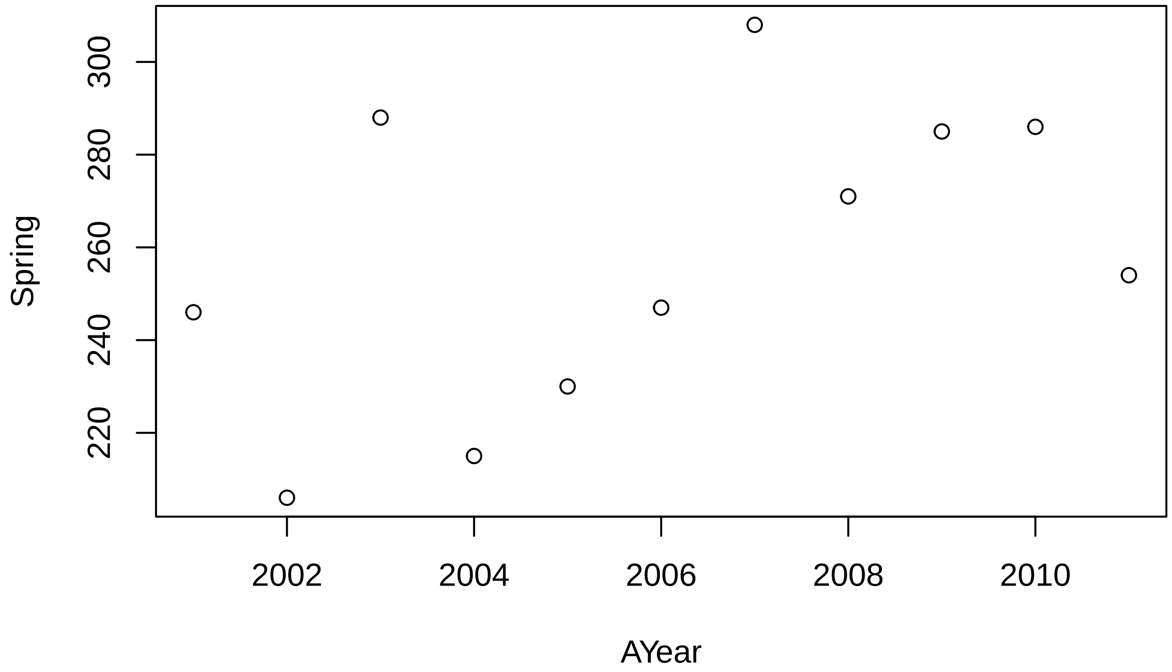
f) Out of all of the residual plots we've built, I would be the most comfortable using a linear model on the last transformation. It had the straightest pattern off all the transformations thus I thiknk it would perform best.

1.34 Enrollment in mathematics courses

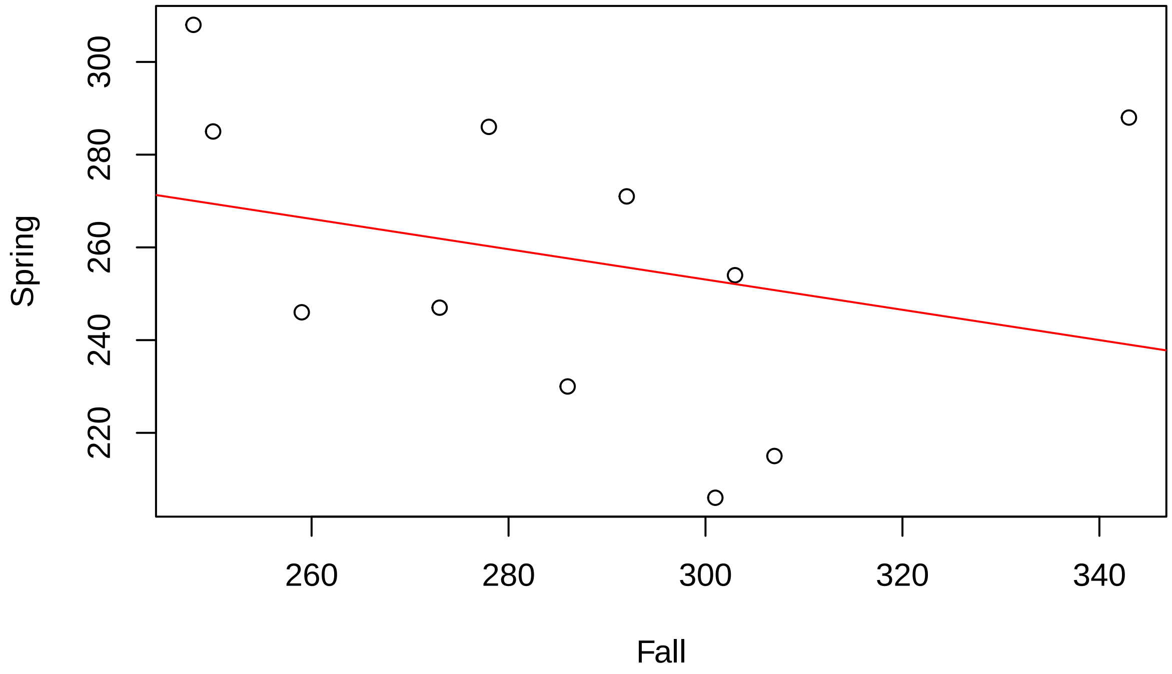
```
MathEnrollment = read.csv("C:/Users/adhri/OneDrive/Documents/R/App_Reg_and_Time_Series/chpt1/datasets/Ma
attach(MathEnrollment)
plot(Fall ~ AYear)
```



```
plot(Spring ~ AYear)
```



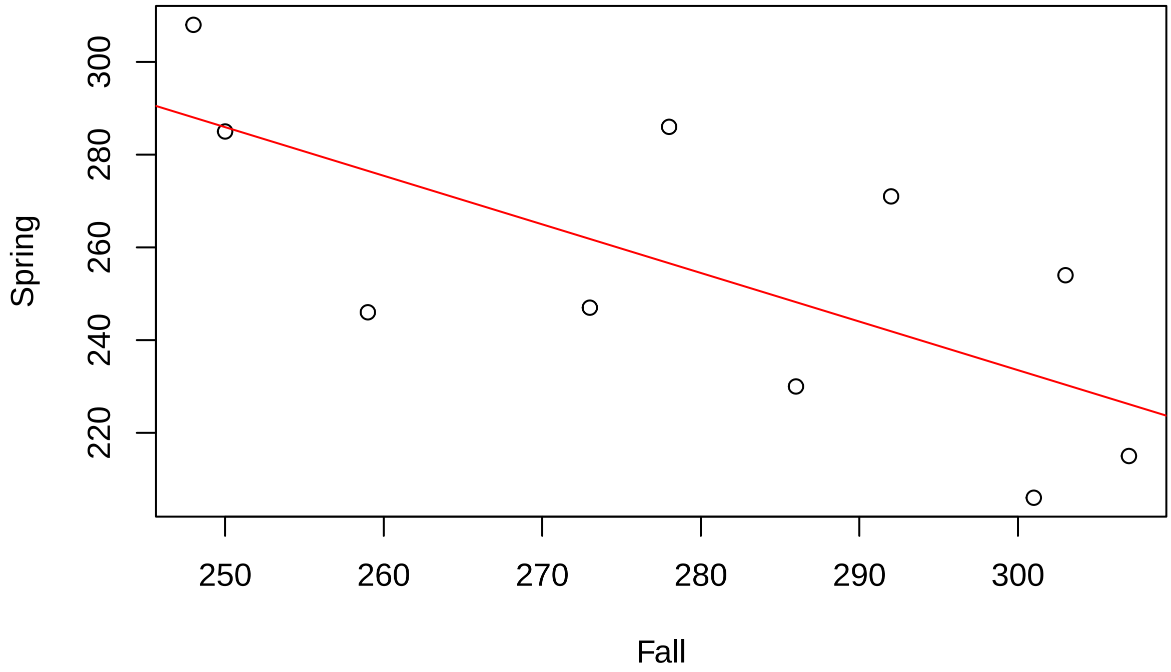
```
plot(Spring ~ Fall)
modelM = lm(Spring ~ Fall)
abline(modelM, col = "red")
```



```
# remove row 2 (AYear 2003) to get rid of the outlier and replot
MathEnrollmentR = MathEnrollment[-3,]
attach(MathEnrollmentR)
```

```
## The following objects are masked from MathEnrollment:
##
##     AYear, Fall, Spring
```

```
plot(Spring ~ Fall)
modelMR = lm(Spring ~ Fall)
abline(modelMR, col = "red")
```



a) For the fall there appears to be an upwards hump and then a downwards hump, while in the spring there appears to be a downwards hump and then a positive hump. Thus, the trend over time does not appear to be the same for both semesters.

b) I think there is some extent to which the fall enrollment proveds a decent predictor of spring enrollment. If we plot fall enrollments against spring, we can see there is a slight negative relationship between the two: as number of enrollments in fall increase, the number of enrollments in spring decrease. Theoretically, if more people attend in fall, then less people are likely to attend in spring.

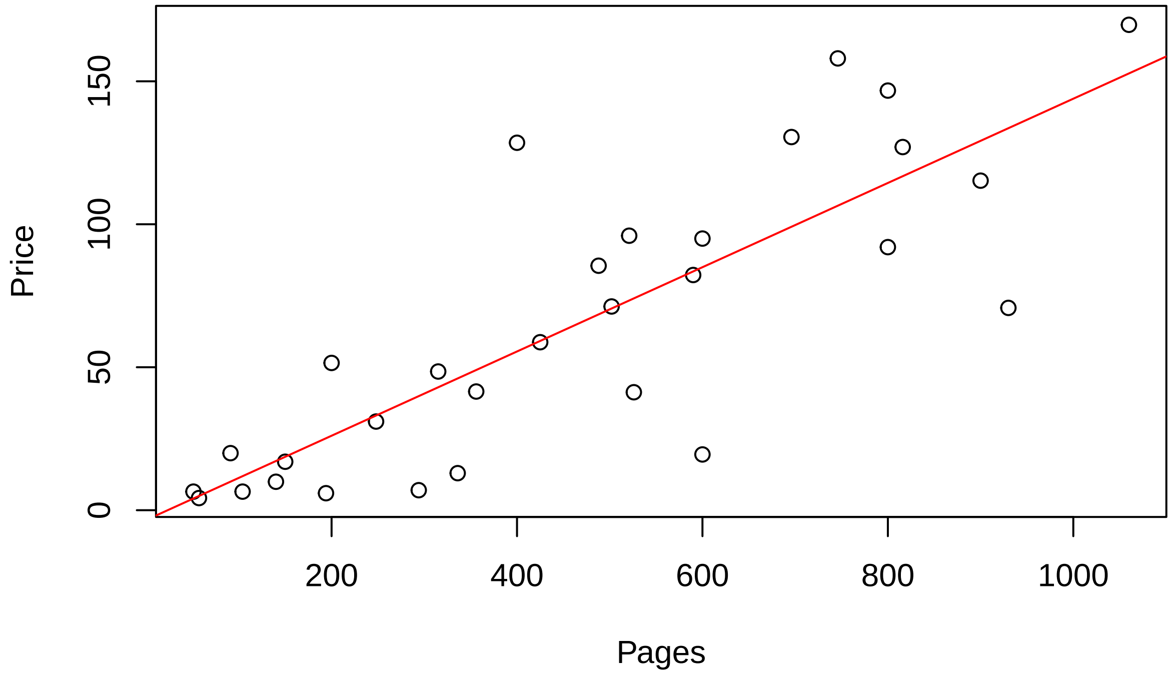
c) The point the faculty members are most likely concerned about is the outlier point at Fall-343, Spring-288 from AYear 2003. It has a very high amount of both fall and spring enrollments, whereas the trend would indicate that there should be less spring enrollments because there are so many fall enrollments.

d) I would tag this point as influential because the least squares regression line became more fitted. The removal of the outlier allowed the least squares regression line to be less influenced and skewed and permitted a better fitted model. The residual standard error also decreased when we removed the outlier.

1.44 Textbook prices

```
TextPrices = read.csv("C:/Users/adhri/OneDrive/Documents/R/App_Reg_and_Time_Series/chpt1/datasets/TextP
attach(TextPrices)
plot(Price ~ Pages)
```

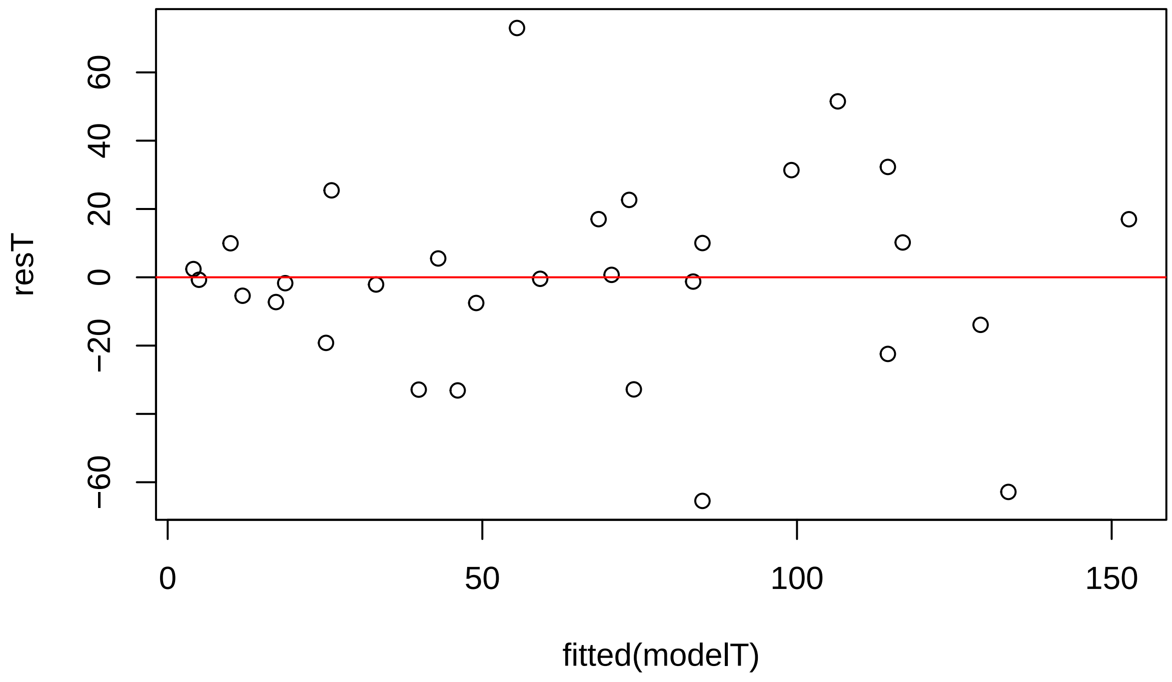
```
modelT = lm(Price ~ Pages)
abline(modelT, col = "red")
```



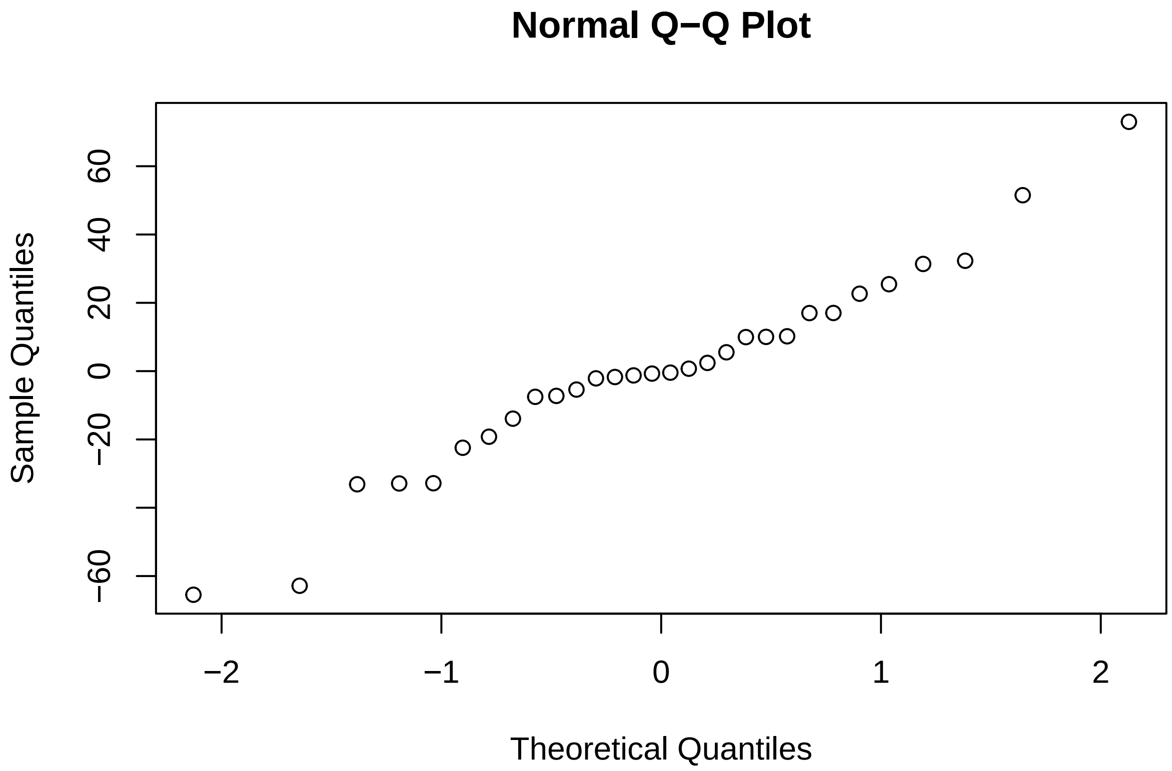
```
summary(modelT)
```

```
##
## Call:
## lm(formula = Price ~ Pages)
##
## 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
```

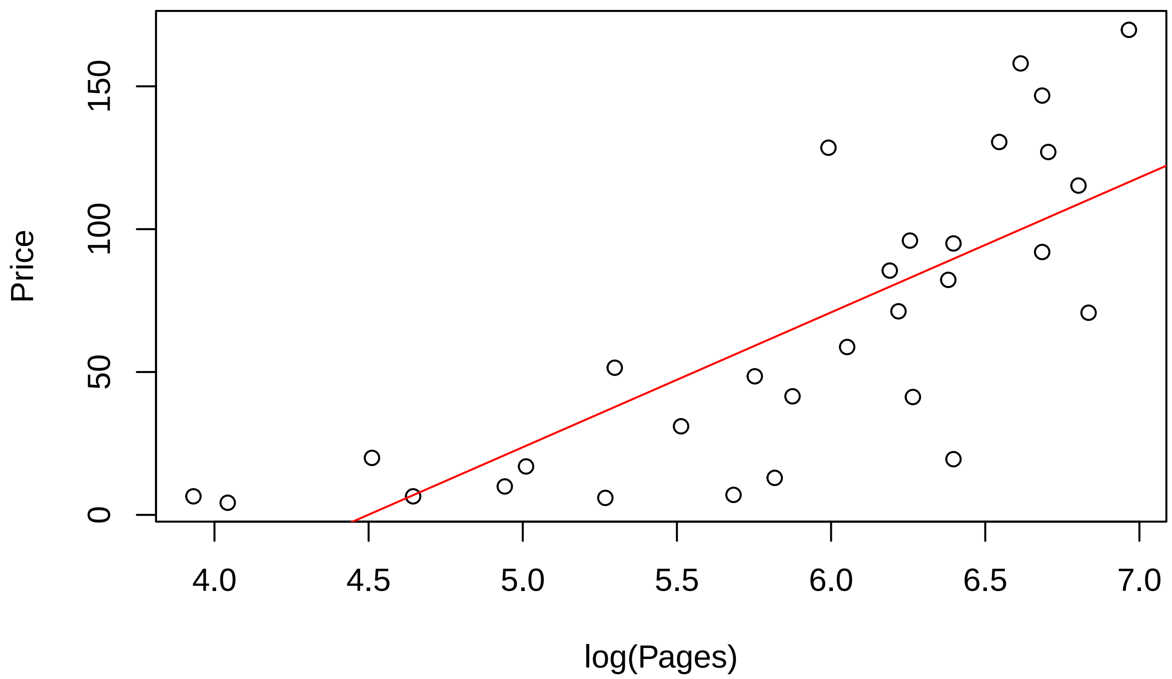
```
# residual plot; kind of fans out. Maybe violates equal variance
resT = resid(modelT)
plot(fitted(modelT), resT)
abline(h = 0, col = 'red')
```



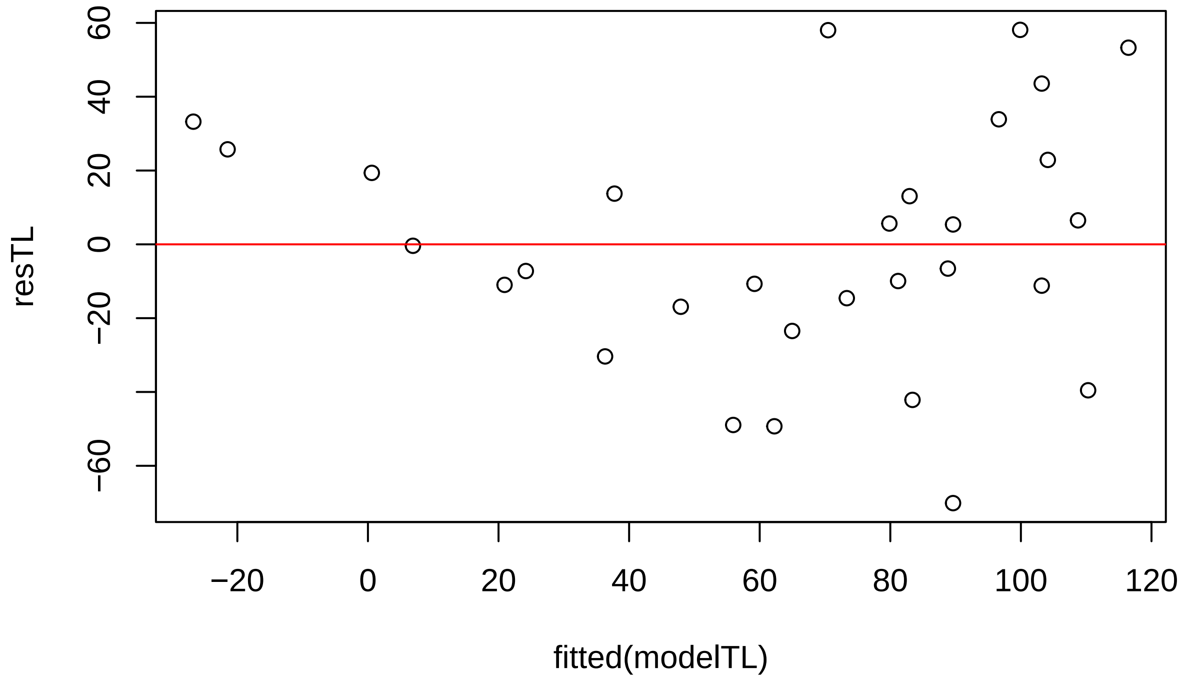
```
# qqnorm
qqnorm(resT)
```



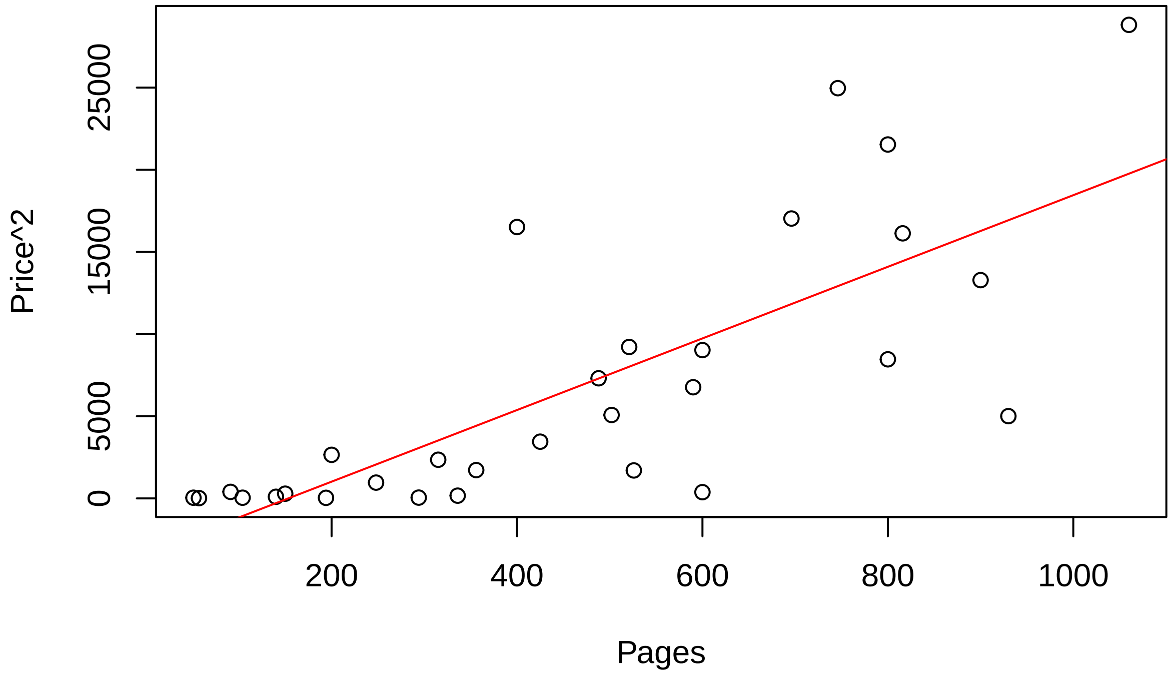
```
# residual plot log transformed; residual plot fans out and fails/violates equal variance
plot(Price ~ log(Pages))
modelTL = lm(Price ~ log(Pages))
abline(modelTL, col = 'red')
```



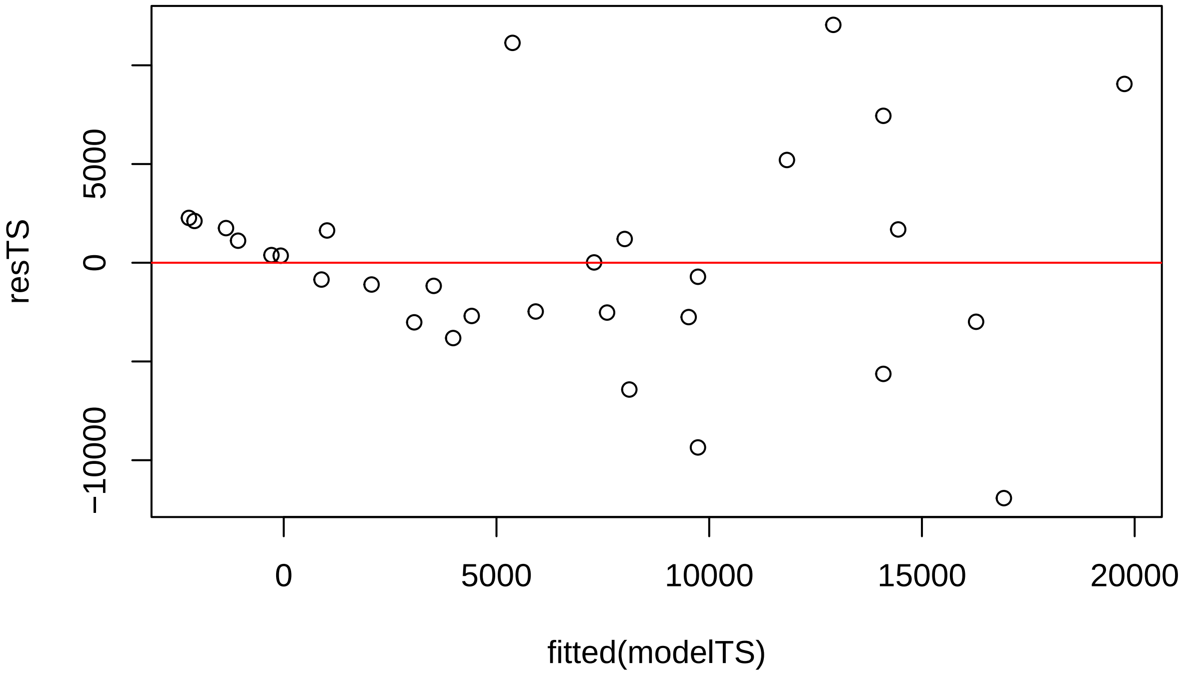
```
resTL = resid(modelTL)
plot(fitted(modelTL), resTL)
abline(h = 0, col = 'red')
```



```
# residual plot squared transformed; residual plot fans out and fails/violates equal variance
plot(Price^2 ~ Pages)
modelTS = lm(Price^2 ~ Pages)
abline(modelTS, col = 'red')
```



```
resTS = resid(modelTS)
plot(fitted(modelTS), resTS)
abline(h = 0, col = 'red')
```



a) The scatterplot has a positive, increasing relationship. As the number of pages in the textbook increases, the price of the textbook increases as well.

b) $\hat{Price}$ = -3.42231 + .14733*Pages

c) I created a normal residual plot, a logarithmically transformed residual plot, and a squared residual plot. There is no clumping in data points, so they are independent and random. The Q-Q plot appears to be straight, so the normality condition is met. The normal plot had some degree of fanning out, while the logarithm and squared plots had large and obvious degrees of fanning out. The transformations violated equal variance and so did the original plot. Thus the conditions for inference are not fully met with these data.