Problem set 3

Exercises to hand in: 1.24, 1.30, 1.44

For this (and all other homework assignments), I am expecting a mixture of text, code, and output.

As you work, I suggested rendering your document frequently to see if you encounter errors. You need to upload the rendered PDF document of your finished homework. This means you should:

- Render your document and look at the preview to make sure it looks good
- Close the preview document
- Go to the Files tab of your RStudio and find the file that ends in .pdf (for this assignment, probably problemset3.pdf)
- Upload the PDF file to Gradescope

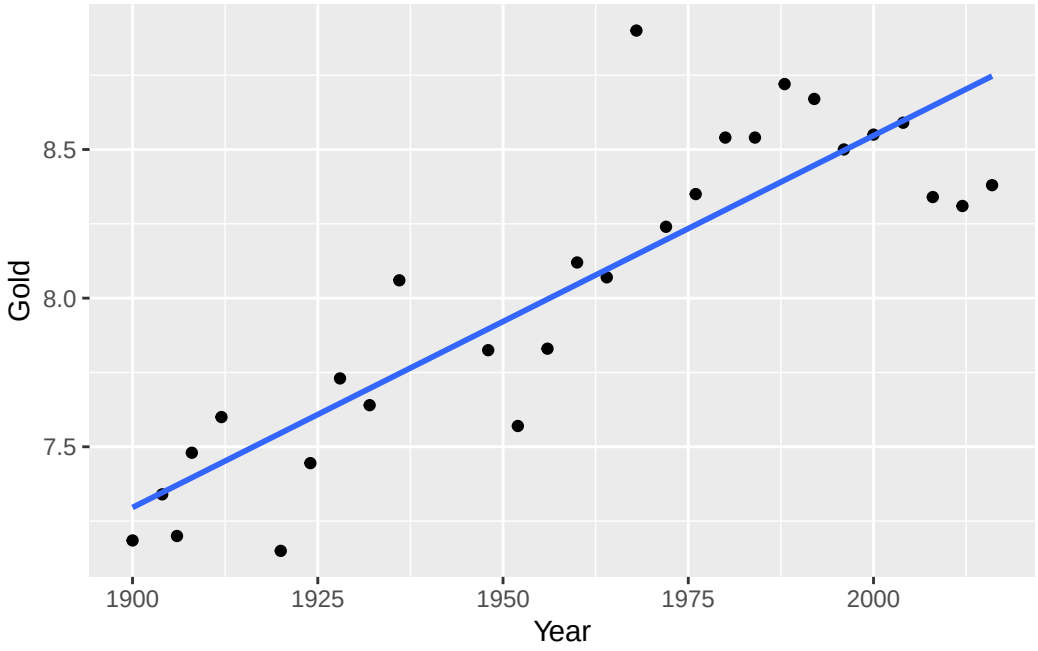
1.24 Olympic long jump residuals

```
data("LongJumpOlympics2016")
```

a. Scatterplot that includes the least squares line. Are there any obvious outliers or influential points in this plot?

```
ggplot(data=LongJumpOlympics2016, aes(x = Year, y = Gold))+ geom_point()+geom_smooth(metho
```

```
`geom_smooth()` using formula 'y ~ x'
```

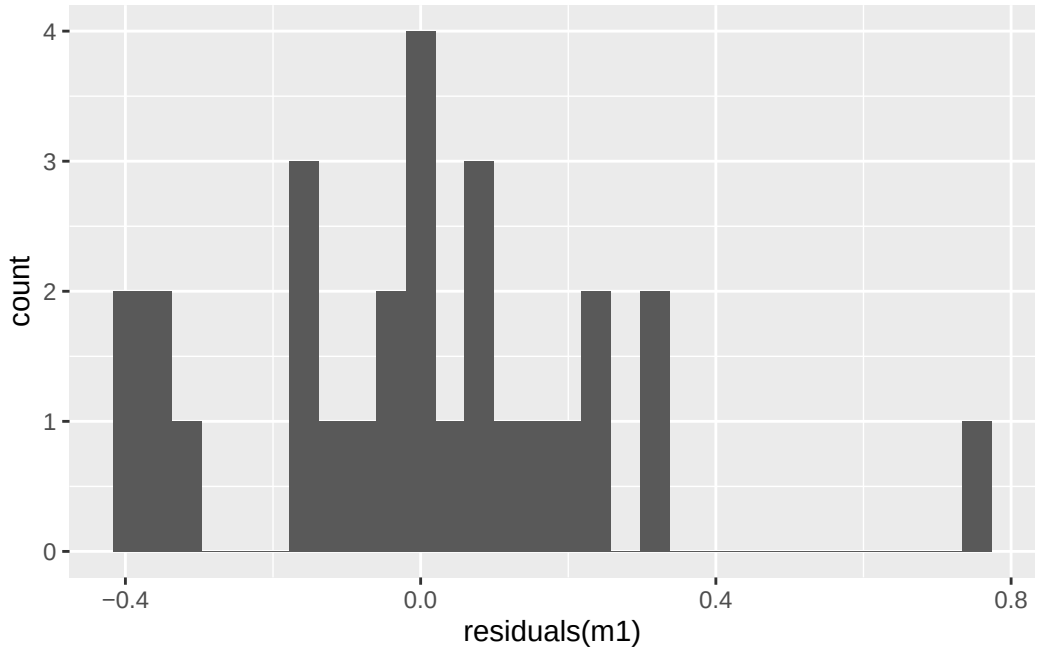


Yes, there are some obvious outlines in this graph. For example, the point that is found at the every top of the graph which is further away from the blue straight line. The graph is also showing a positive linear trend between the 2 variables(Year and Gold). As the years increases the gold medals received also increases.

b. Histogram of the residuals.

```
m1=lm(Gold~Year, data=LongJumpOlympics2016)
ggplot(LongJumpOlympics2016) +
  geom_histogram(aes(x = residuals(m1)))
```

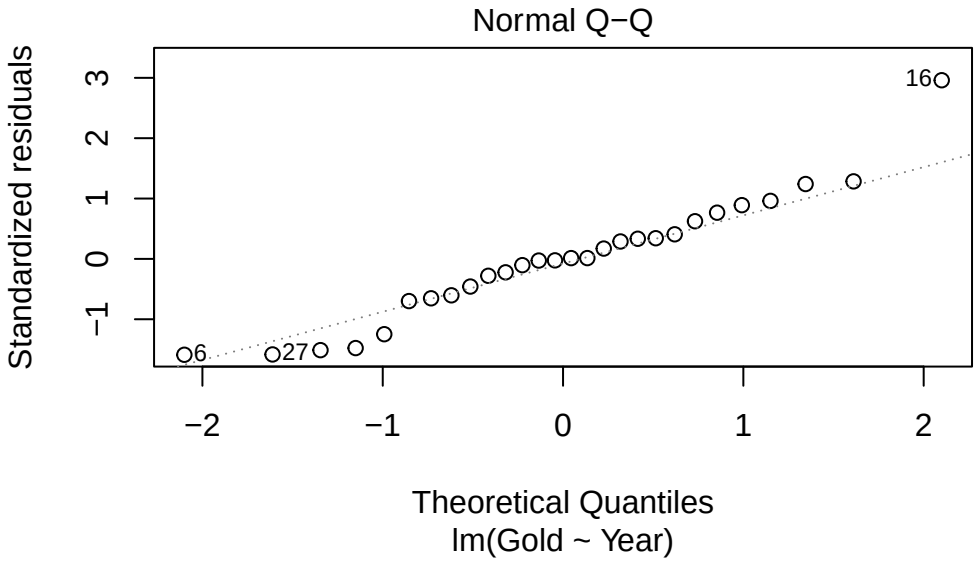
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From the histogram we can see that it does not meet the normality, because the histogram is skewed to the right. Also there is an outlier in the histogram that is found closest to the 0.8 residuals. Another one located at -0.4 residuals.

c. Normal probability plot of the residuals.

```
plot(m1, which = 2)
```



Looking at the Q-Q plot we can say that the graph is tending to meet the linearity. However there is an outlier numbered 16. Moreover there is a positive trend in the Q-Q plot. Saying that variables Year and Gold are directly proportional to each other.

1.30 Caterpillar nitrogen assimilation versus mass

```
data("Caterpillars")
Caterpillars <- Caterpillars %>%
  mutate(Instar = as_factor(Instar)) # for easier plotting
```

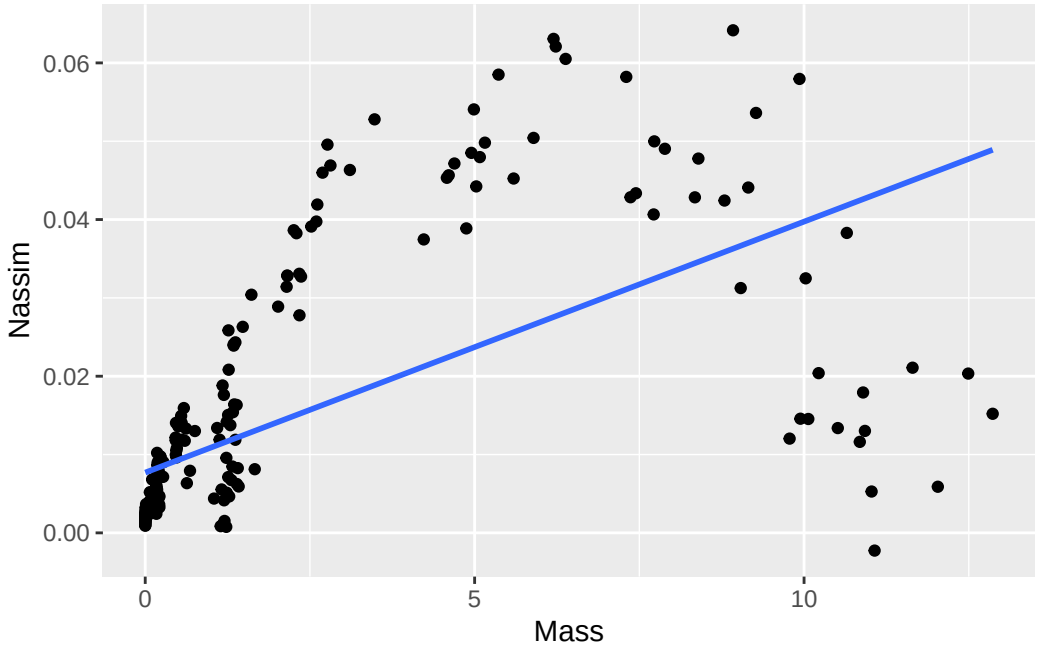
a. Produce a scatterplot for predicting nitrogen assimilation (Nassim) based on Mass. Comment on any patterns.

```
ggplot(data=Caterpillars, aes(x =Mass, y =Nassim))+ geom_point()+geom_smooth(method ="lm",
```

```
`geom_smooth()` using formula 'y ~ x'
```

```
Warning: Removed 13 rows containing non-finite values (stat_smooth).
```

```
Warning: Removed 13 rows containing missing values (geom_point).
```



The graph does not meet the conditions for equality of variance, because the distribution of the points on the graph is uneven on both sides of the blue straight line. The graph also does not meet the normality because the graph is curved and it is more steeper as it goes higher making the graph not meet the condition of linearity.

b. Produce a similar plot using NATURAL LOG transformed variables. Make your own logged variables, don't use LogNassim and LogMass. Again, comment on any patterns.

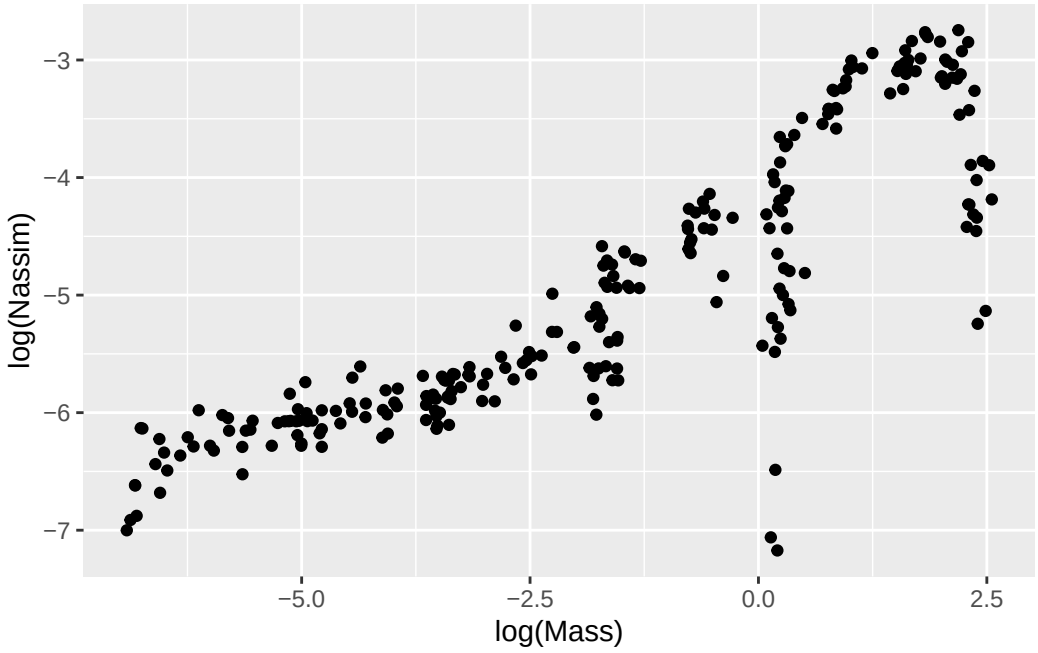
(Notice that I've changed question b above. Please use the natural log, not log base 10.)

```
ggplot(data=Caterpillars)+geom_point(aes(x =log(Mass), y =log(Nassim)))
```

```
Warning in log(Nassim): NaNs produced
```

```
Warning in log(Nassim): NaNs produced
```

```
Warning: Removed 14 rows containing missing values (geom_point).
```



Looking at the graph we can see that it does not meet the condition for linearity., because the graph is curved and is bending downwards. Also we can see that the graph is not meeting the conditions for normality.

c. Would you prefer the plot in part (a) or part (b) to predict the nitrogen assimilation of caterpillars with a linear model? Fit a linear regression model for the plot you chose and write down the prediction equation.

```
m2=lm(log(Nassim)~log(Mass), data=Caterpillars)
```

```
Warning in log(Nassim): NaNs produced
```

```
summary(m2)
```

```
Call:
lm(formula = log(Nassim) ~ log(Mass), data = Caterpillars)

Residuals:
     Min       1Q   Median       3Q      Max
-2.90330 -0.26614  0.04977  0.38511  0.96390

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.34585    0.04239 -102.53   <2e-16 ***
log(Mass)    0.37096    0.01332   27.85   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.576 on 251 degrees of freedom
  (14 observations deleted due to missingness)
Multiple R-squared:  0.7555,    Adjusted R-squared:  0.7545
F-statistic: 775.6 on 1 and 251 DF,  p-value: < 2.2e-16
```

I would choose the second model and the linear equation for this model will be Nassim=0.37096 * mass -4.34585.

d. Add COLOR TO INDICATE the grouping variable Instar on the scatterplot that you chose in (c). Does the linear trend appear consistent for all five stages of a caterpillar's life? (Note: We are not ready to fit more complicated models yet, but we will return to this experiment in Chapter 3.)

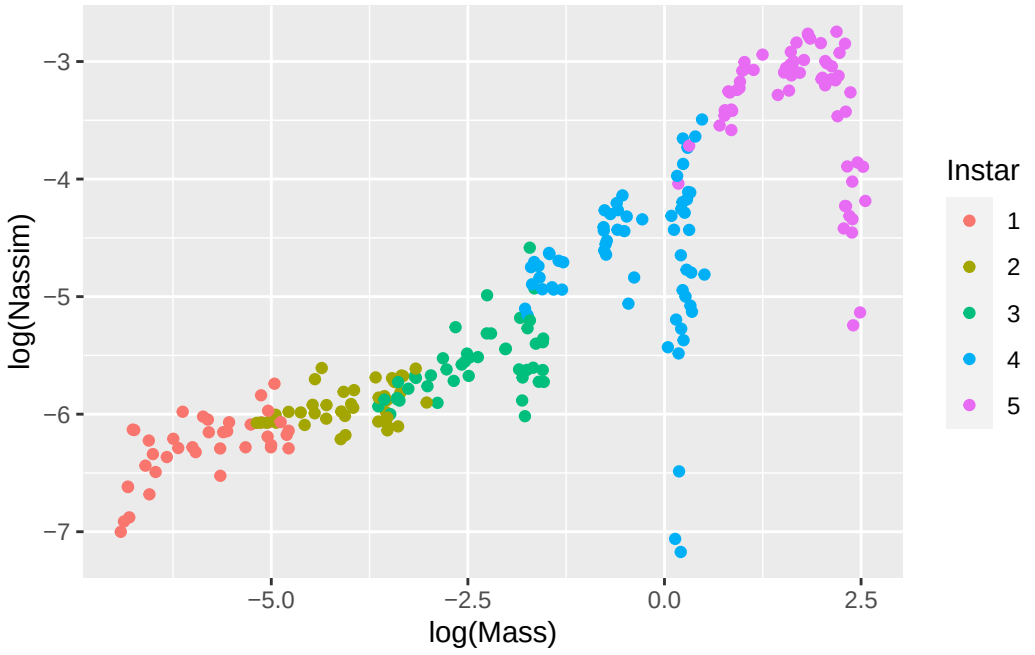
(Notice that I've changed question d above. Use color for Instar, rather than trying to use a different plotting symbol.)

```
ggplot(Caterpillars) +
  geom_point(aes(x = log(Mass), y = log(Nassim), col = Instar ))
```

```
Warning in log(Nassim): NaNs produced
```

```
Warning in log(Nassim): NaNs produced
```

```
Warning: Removed 14 rows containing missing values (geom_point).
```



The linear trend does not appear to be consistent through the stages of the Caterpillar's life.From stages 1 through 4 the data plotting trends to be linear but on the 5th stage the data points looses the linear trend.

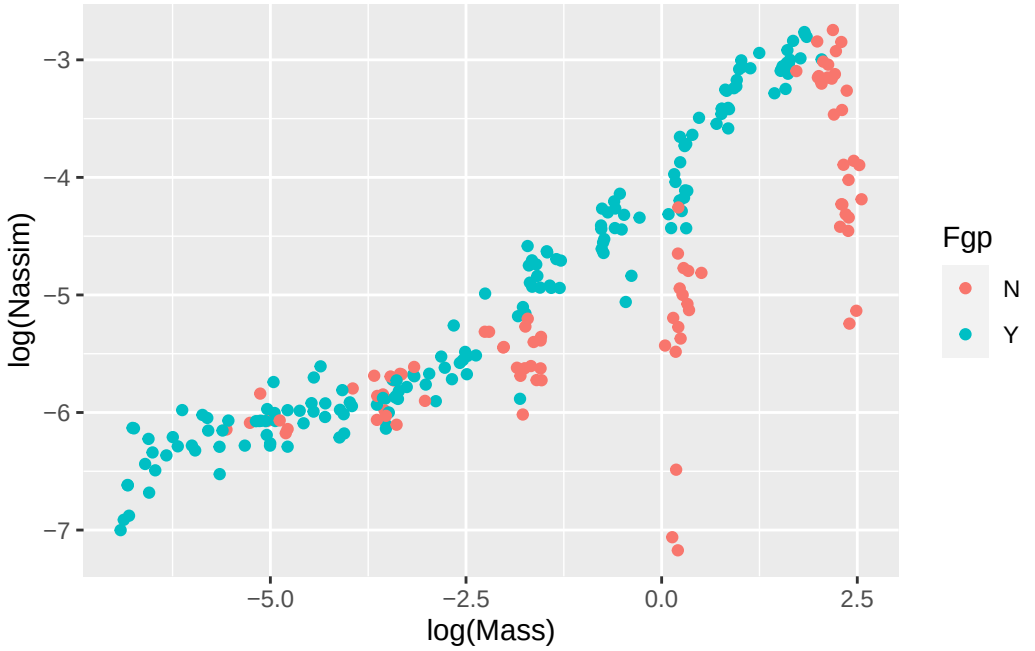
e. Repeat part (d) using COLORS for the groups defined by the free-growth period variable Fgp. Does the linear trend appear to be better when the caterpillars are in a free growth period? (Again, we are not ready to fit more complicated models, but we are looking at the plot for linear trend in the two groups.)

```
ggplot(Caterpillars) +
  geom_point(aes(x = log(Mass), y = log(Nassim), col = Fgp))
```

```
Warning in log(Nassim): NaNs produced
```

```
Warning in log(Nassim): NaNs produced
```

```
Warning: Removed 14 rows containing missing values (geom_point).
```



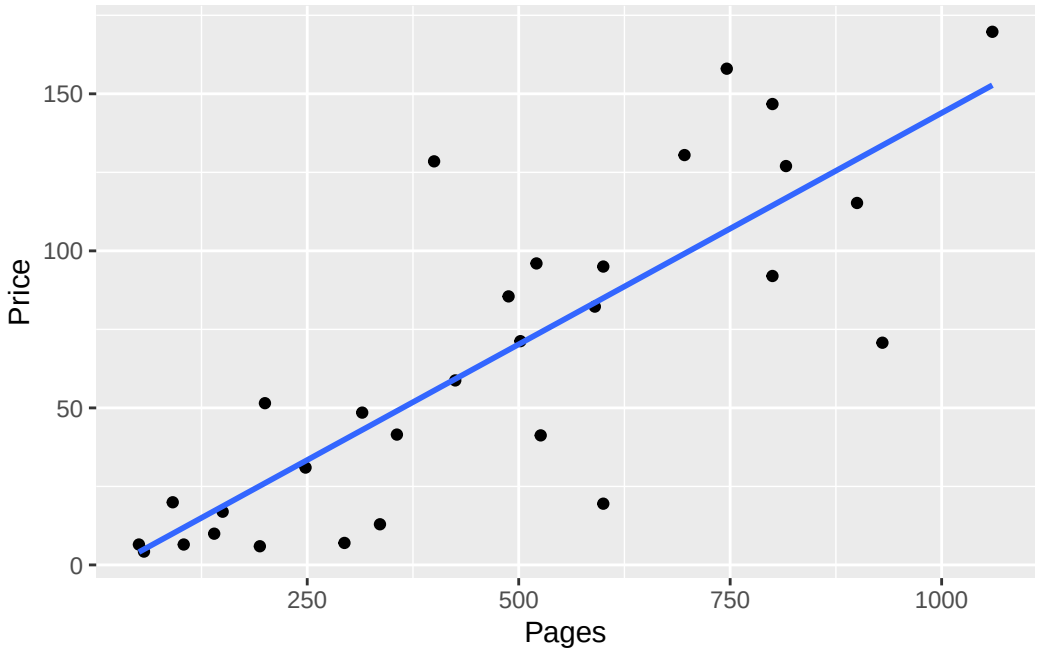
Yes the linear trend seems to get better because if you draw a line of best fit we can see that the points of the graph are closer to the line of best fit when compared to the previous graph.

1.44 Textbook prices

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

```
data("TextPrices")
ggplot(data=TextPrices, aes(x = Pages, y = Price))+ geom_point()+geom_smooth(method ="lm",
```

```
`geom_smooth()` using formula 'y ~ x'
```



Yes the scatter plot can be used to determine or predict the price of the textbook based on the pages it has. From the graph we can see that as he number of pages increases the price of the book increases as well. Meaning they are directly proportional to each other. We can also see that there is positive linear trend in the graph.

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

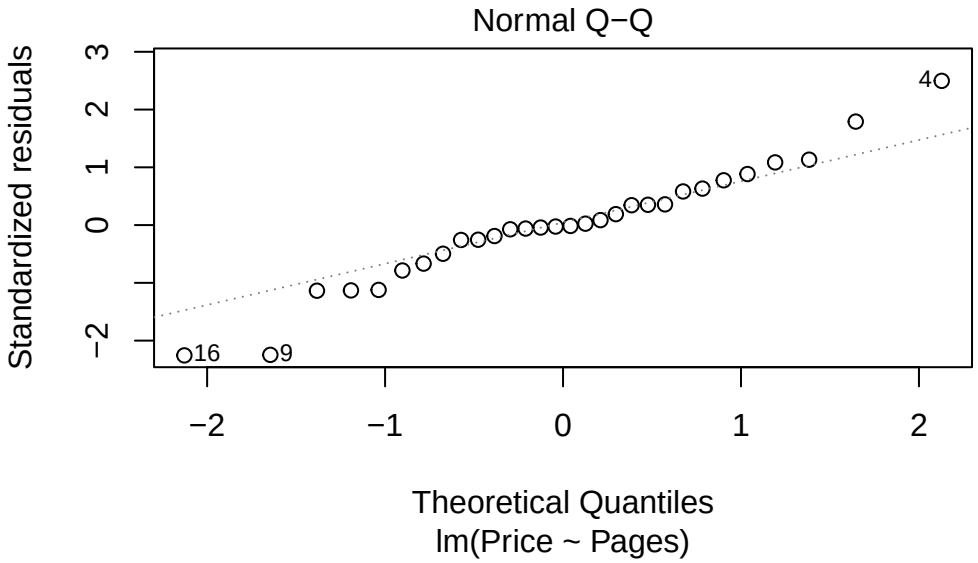
```
m3= lm(Price~Pages, data= TextPrices)
summary(m3)
```

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

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
```

Price= 0.14733 *Pages + -3.42231

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

```
data("TextPrices")
plot(m3,which =2)
```



The conditions for inference are not met with these data because in the graph we can see that there are some points that are below the dotted line and a few of them that are above the the dotted line. The graph is deviating away from the dotted line to 2 different extremes of the graph.