Math 4780 - Homework 4

Liam Fruzyna

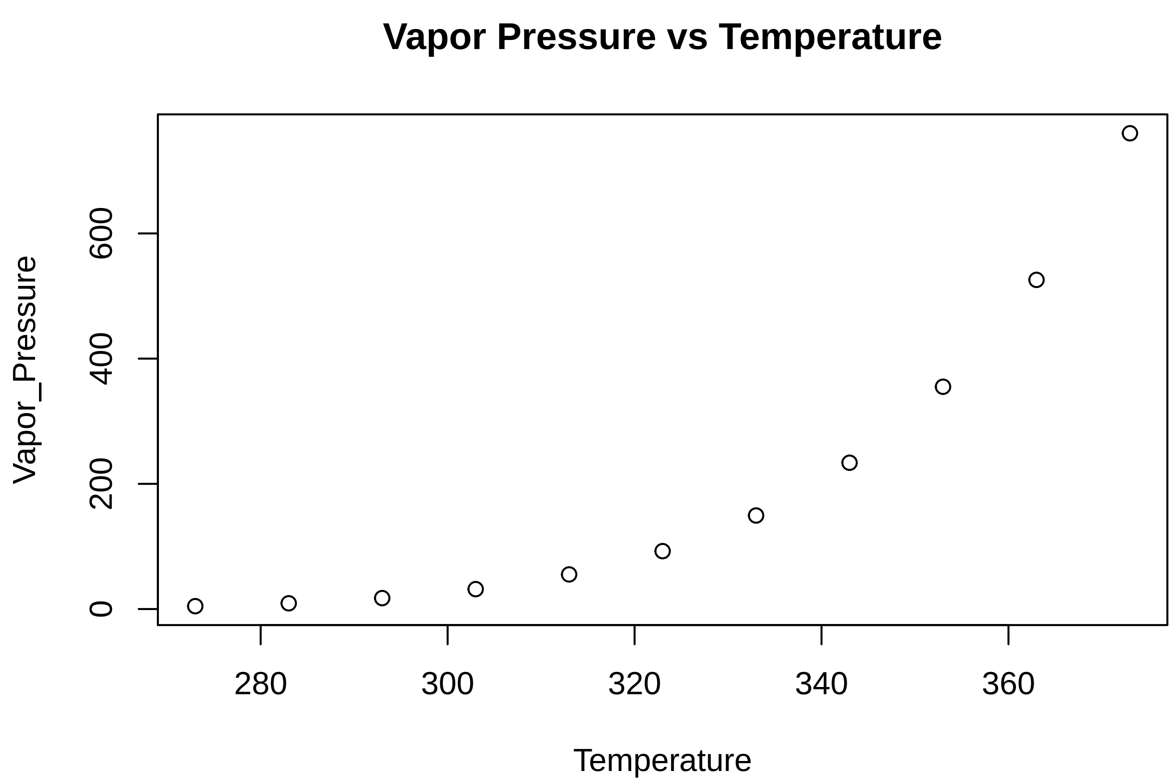
#5.2

Dataset setup

```
temps <- seq(273, 373, 10)
pressures <- c(4.6, 9.2, 17.5, 31.8, 55.3, 92.5, 149.4, 233.7, 355.1, 525.8, 760)
vp <- data.frame(Temperature=temps, Vapor_Pressure=pressures)
```

a. Plot a scatter diagram. Does it seem likely that a straight-line model will be adequate?

```
plot(vp, main='Vapor Pressure vs Temperature')
```



No, it does not appear that a straight line mode will be adequate because the data is non-linear.

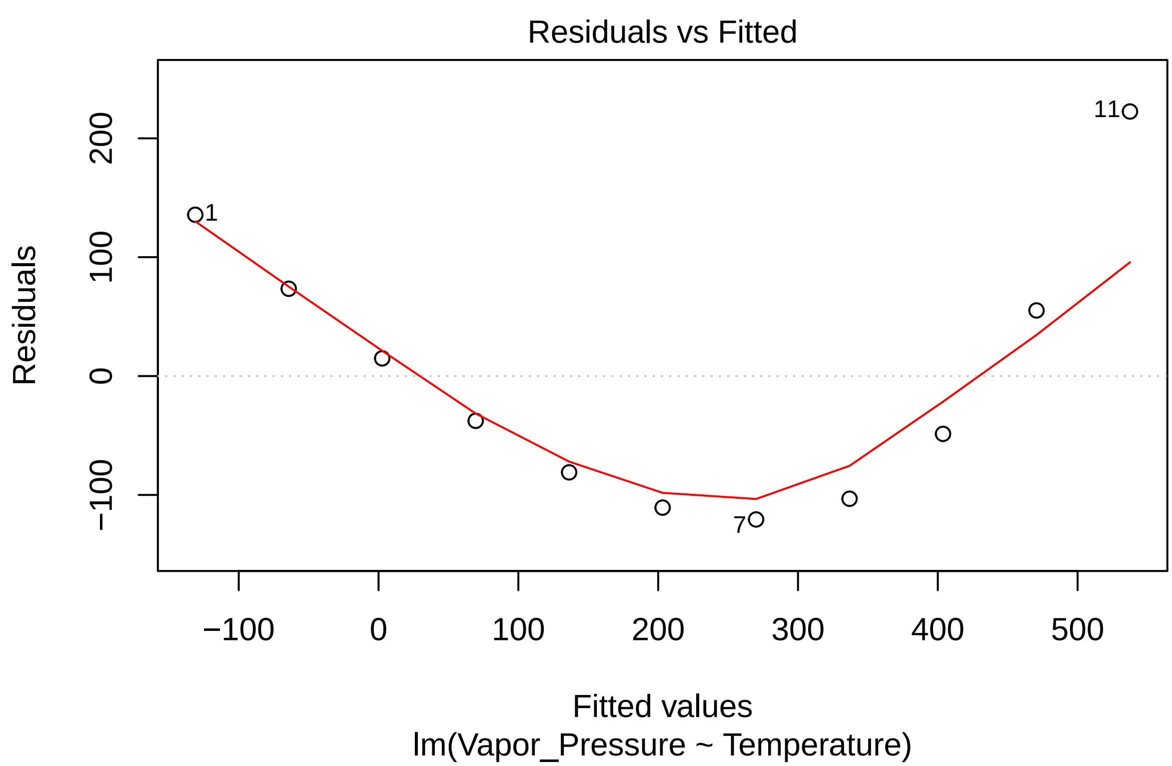
b. Fit the straight-line model. Compute the summary statistics and the residual plots. What are your conclusions regarding the model adequacy?

```
lm_vp <- lm(Vapor_Pressure ~ Temperature, data=vp)
summary(lm_vp)
```

```
##
```

```
## Call:
## lm(formula = Vapor_Pressure ~ Temperature, data = vp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -120.63  -92.10  -37.66   64.33  222.55
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1956.258    363.807  -5.377 0.000446 ***
## Temperature     6.686      1.121   5.964 0.000212 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 117.6 on 9 degrees of freedom
## Multiple R-squared:  0.7981, Adjusted R-squared:  0.7756
## F-statistic: 35.57 on 1 and 9 DF,  p-value: 0.0002117
```

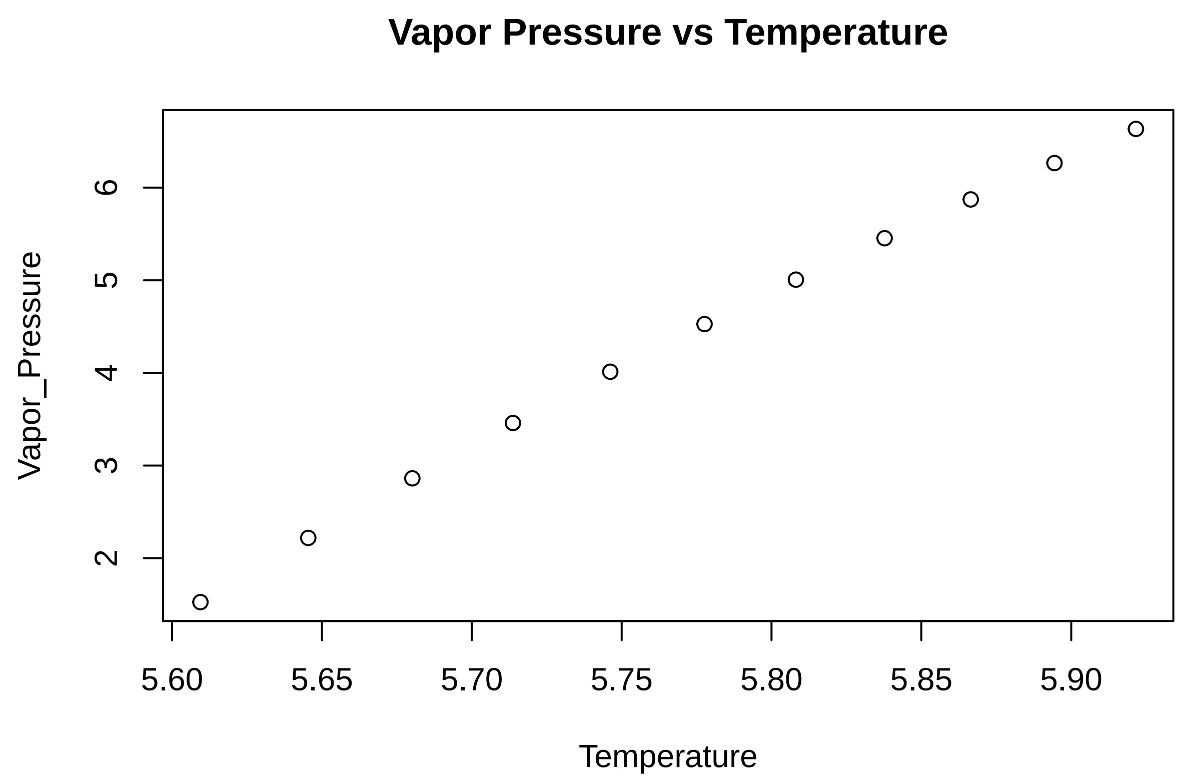
```r
plot(lm_vp, which=1)
```



The model is not accurate.

c. From physical chemistry the Clausius-Clapeyron equation states that $ln(p_v) \propto -\frac{1}{T}$. Repeat part b using the appropriate transformation based on this information.

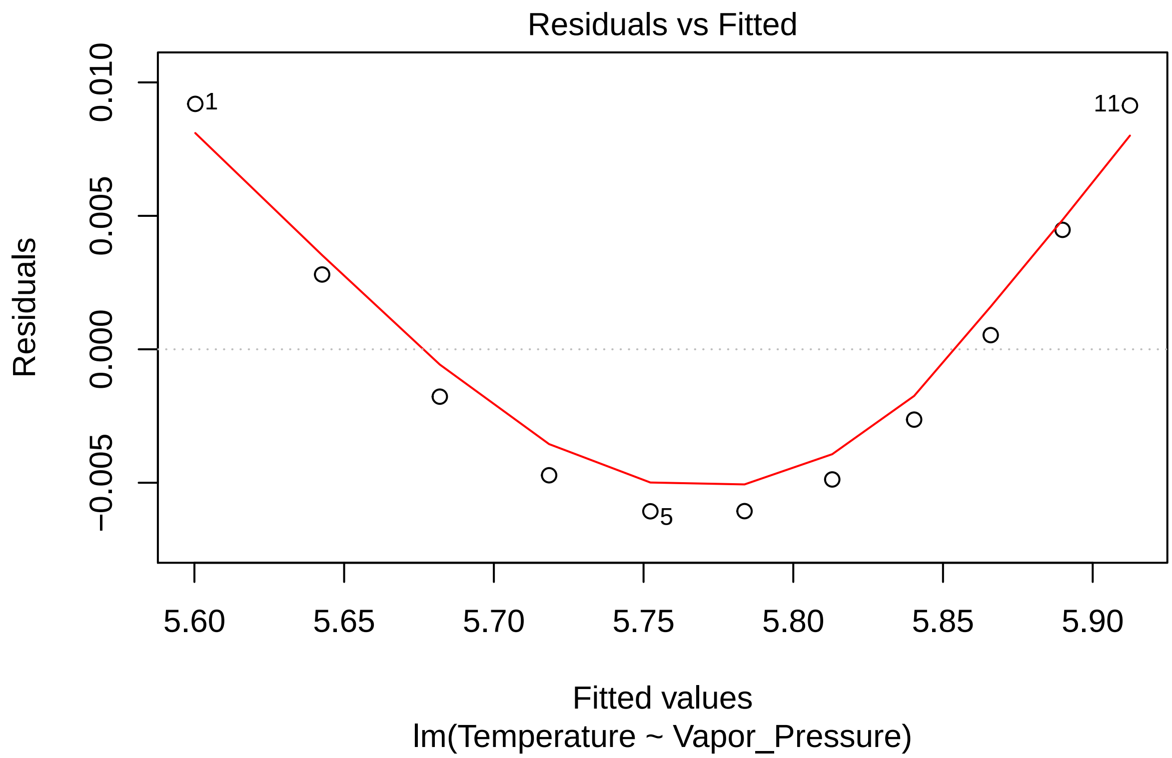
```r
trans_vp <- data.frame(Temperature=log(temps), Vapor_Pressure=log(pressures))
plot(trans_vp, main='Vapor Pressure vs Temperature')
```




```r
lm_trans_vp <- lm(Temperature ~ Vapor_Pressure, data=trans_vp)
summary(lm_trans_vp)
```

```
##
## Call:
## lm(formula = Temperature ~ Vapor_Pressure, data = trans_vp)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.006069 -0.004798 -0.001773  0.003638  0.009194
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.507002   0.005225 1053.93  < 2e-16 ***
## Vapor_Pressure 0.061122   0.001127   54.25 1.24e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.006017 on 9 degrees of freedom
## Multiple R-squared:  0.997,  Adjusted R-squared:  0.9966
## F-statistic:  2943 on 1 and 9 DF,  p-value: 1.236e-12
```

```r
plot(lm_trans_vp, which=1)
```

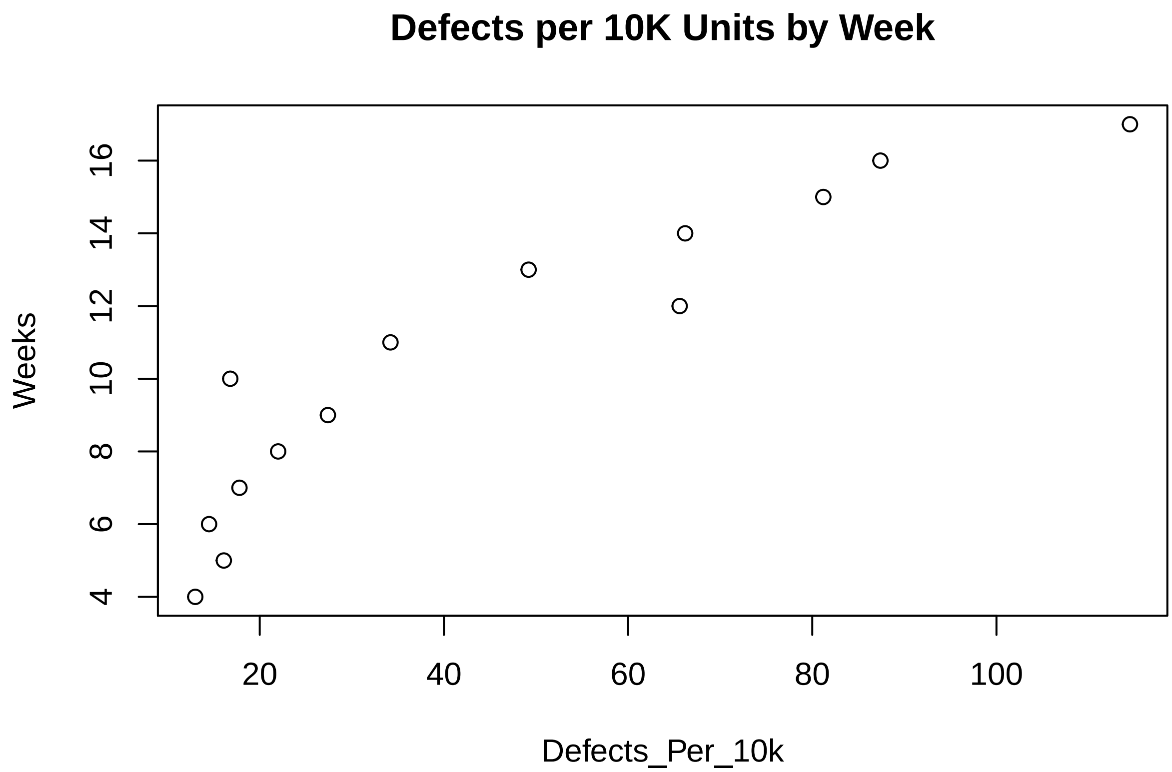


There is some improvement in the model with a log transformation.

#5.5

Dataset setup

```
defects <- c(13, 16.1, 14.5, 17.8, 22, 27.4, 16.8, 34.2, 65.6, 49.2, 66.2, 81.2, 87.4, 114.5)
weeks <- 4:17
glass <- data.frame(Defects_Per_10k=defects, Weeks=weeks)
plot(glass, main='Defects per 10K Units by Week')
```

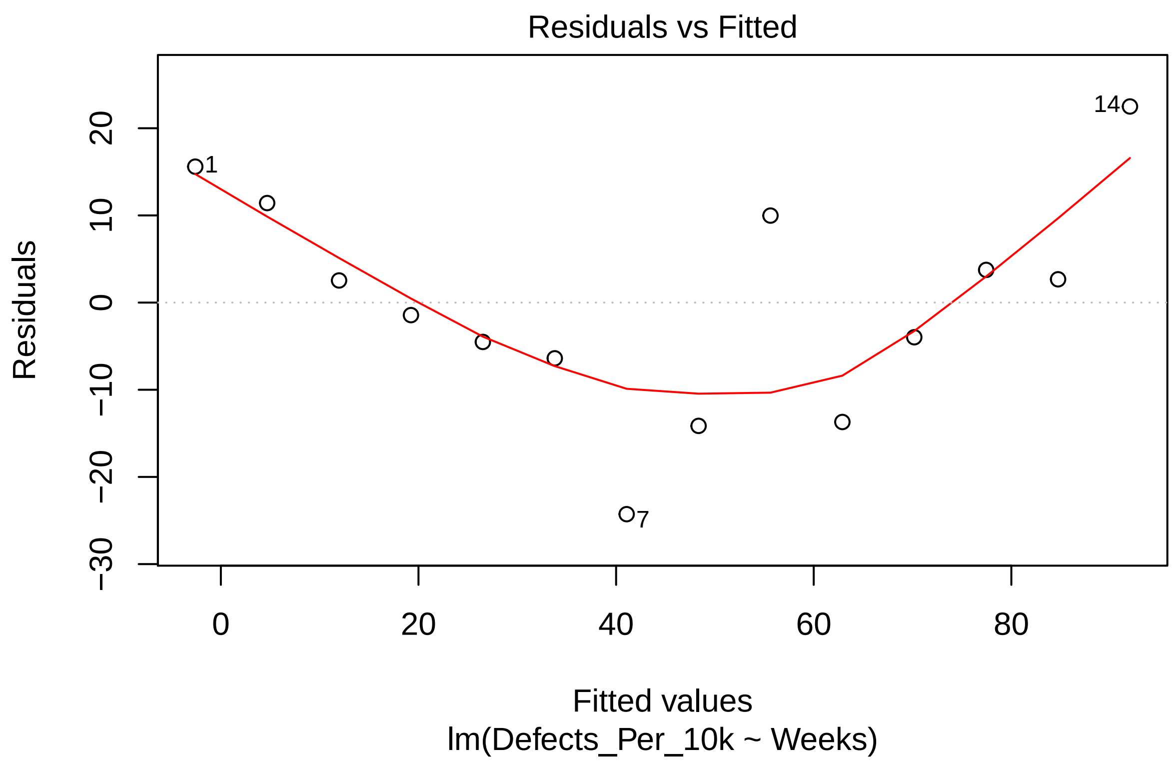


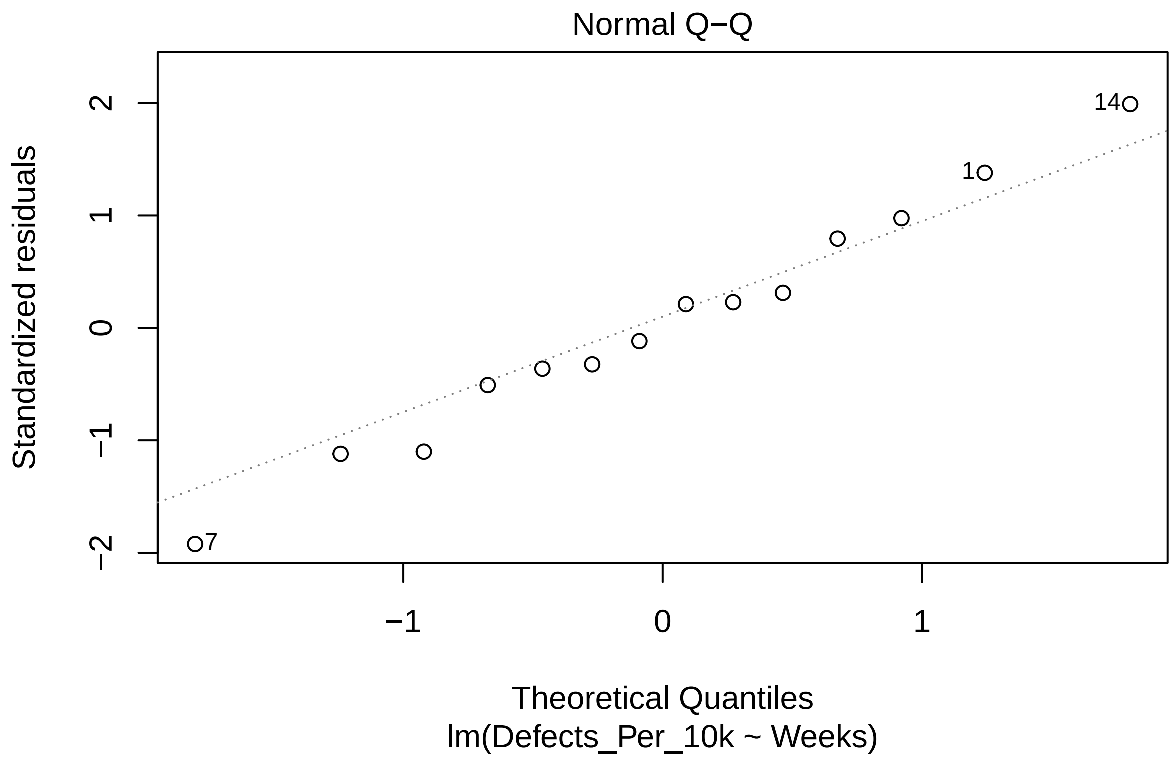
a. Fit a straight-line regression model to the data and perform the standard tests for model adequacy.

```
lm_glass <- lm(Defects_Per_10k ~ Weeks, data=glass)
summary(lm_glass)
```

```
##
## Call:
## lm(formula = Defects_Per_10k ~ Weeks, data = glass)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.2688  -5.9229   0.5497   8.4203  22.4943
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -31.6982     9.7758  -3.243  0.00705 **
## Weeks         7.2767     0.8692   8.372 2.35e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.11 on 12 degrees of freedom
## Multiple R-squared:  0.8538, Adjusted R-squared:  0.8416
## F-statistic: 70.09 on 1 and 12 DF,  p-value: 2.354e-06
```

```r
plot(lm_glass, which=c(1,2))
```

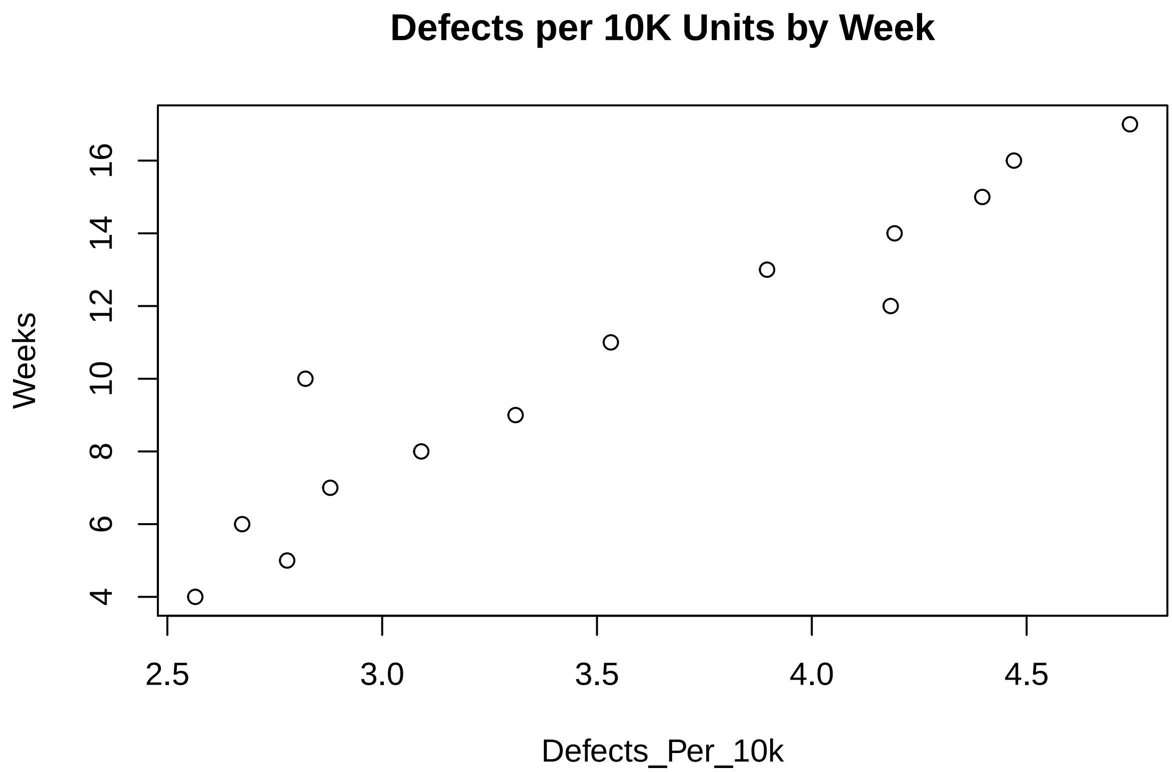




There is a slight non-linear pattern to the residuals.

b. Suggest an appropriate transformation to eliminate the problems encountered in part a. Fit the transformed model and check for adequacy.

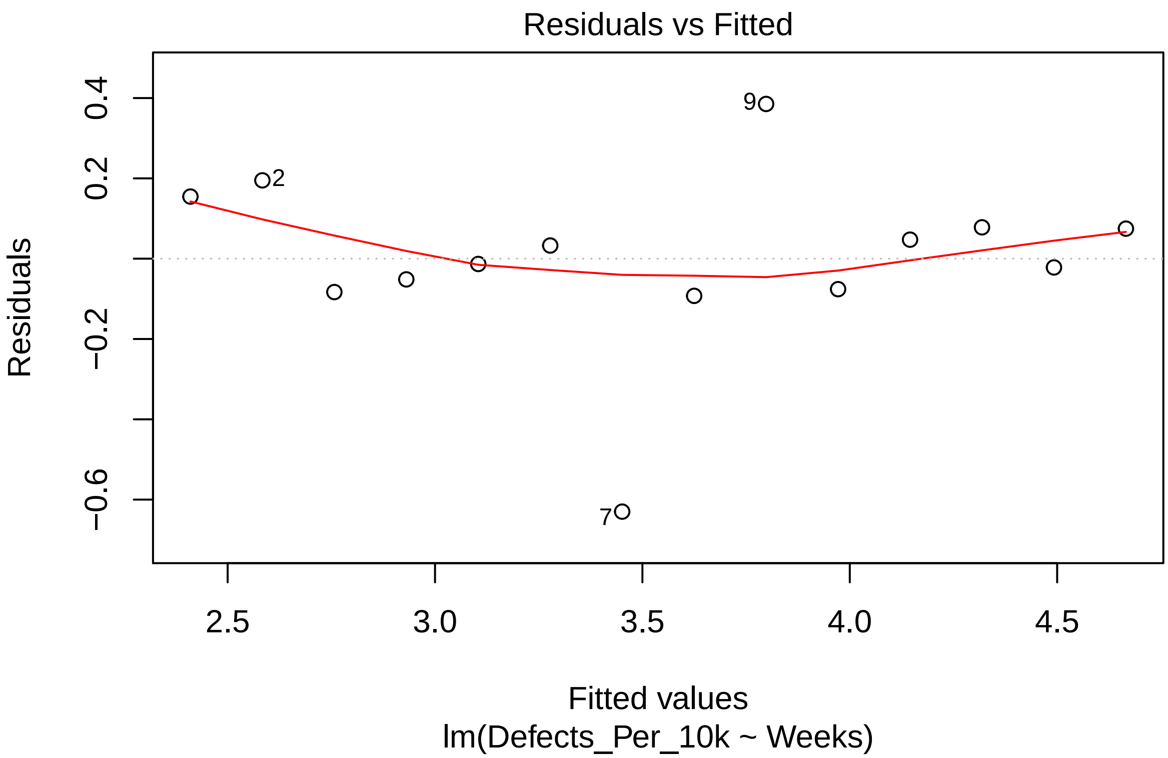
```
trans_glass <- data.frame(Defects_Per_10k=log(defects), Weeks=weeks)
plot(trans_glass, main='Defects per 10K Units by Week')
```

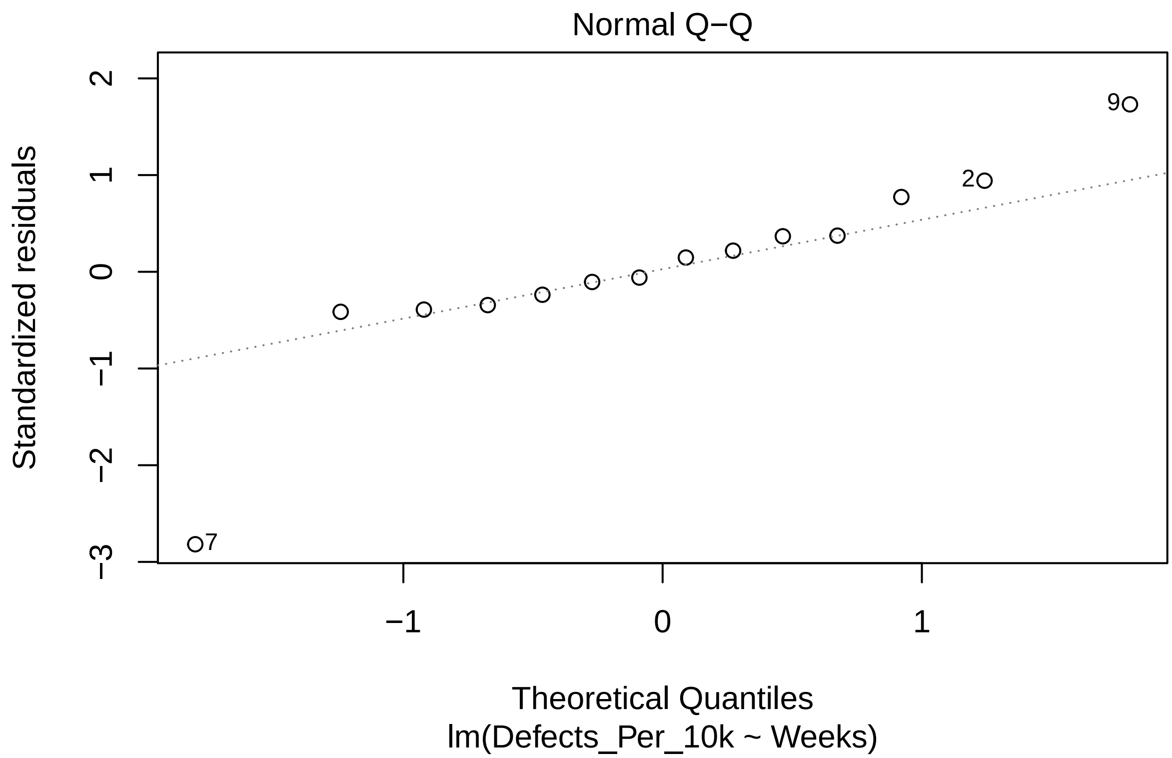



```r
lm_trans_glass <- lm(Defects_Per_10k ~ Weeks, data=trans_glass)
summary(lm_trans_glass)
```

```
##
## Call:
## lm(formula = Defects_Per_10k ~ Weeks, data = trans_glass)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62990 -0.06982  0.00977  0.07727  0.38529
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.71622    0.17311   9.914 3.93e-07 ***
## Weeks        0.17351    0.01539  11.273 9.68e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2322 on 12 degrees of freedom
## Multiple R-squared:  0.9137, Adjusted R-squared:  0.9065
## F-statistic: 127.1 on 1 and 12 DF,  p-value: 9.676e-08
```

```r
plot(lm_trans_glass, which=c(1,2))
```


Residuals vs Fitted
Residuals
0.4
0.2
−0.2
−0.6
2.5
3.0
3.5
4.0
4.5
9
2
7
Fitted values
lm(Defects_Per_10k ~ Weeks)



The model appears to have improved by adding the natural log.

#6.2 Perform a thorough influence analysis of the property valuation data given in B.4. Discuss your results.

```
library(MPV)
prop <- table.b4
lm_prop <- lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8, data=prop)
cooks.distance(lm_prop)
```

```
##            1            2            3            4            5 
## 2.417125e-04 4.128686e-05 1.697418e-02 4.512358e-04 5.457142e-02 
##            6            7            8            9           10 
## 1.101422e-02 3.245172e-01 2.411083e-02 1.729943e-03 1.046164e-01 
##           11           12           13           14           15 
## 7.221803e-02 6.890794e-03 1.340978e-01 3.854853e-01 7.600609e-03 
##           16           17           18           19           20 
## 3.228566e-01 6.667913e-02 4.212734e-02 2.016715e-02 8.609540e-03 
##           21           22           23           24 
## 7.768796e-02 2.641712e-01 8.906766e-02 6.851296e-03
```

No points appear to be influential.

#6.10 Formally show that $D_i = \frac{r_i}{p}\frac{h_{ii}}{1-h_{ii}}$

Using $\hat{\beta} - \hat{\beta}_{(i)} = \frac{(X'X)^{-1}x_ie_i}{1-h_{ii}}$

$D_i = \frac{(\hat{\beta}_{(i)}-\hat{\beta})'X'X(\hat{\beta}_{(i)}-\hat{\beta})}{pMS_{Res}} = \frac{x_i(X'X)^{-1}X'X(X'X)^{-1}x_ie_i^2}{(1-h_{ii})^2pMS_{Res}} = (\frac{e_i}{1-h_{ii}})^2(\frac{h_{ii}}{pMS_{Res}}) = \frac{e_i^2}{MS_{Res}(1-h_{ii})}\frac{1}{p}\frac{h_{ii}}{1-h_{ii}} = \frac{r_i^2}{p}\frac{h_{ii}}{1-h_{ii}}$

#6.15 Table B.14 contains data converning the transient points of an electronic inverter. Fit a regression model to all 25 observations but only use $x_1 - x_4$ as the regressors. Investigate this model for influential observations and comment on your findings.

```
invt <- table.b14
lm_invt <- lm(y ~ x1 + x2 + x3 + x4, data=invt)
cooks.distance(lm_invt)
```

```
##            1            2            3            4            5
## 0.0178354125 1.9820030163 0.0027969280 1.0400671359 0.0040130852
##            6            7            8            9           10
## 0.0161146054 0.0032489766 0.4133792681 0.2553946738 0.0233175192
##           11           12           13           14           15
## 0.0192353611 0.0020805156 0.0038883279 0.1489170953 0.0002687221
##           16           17           18           19           20
## 0.0159570925 0.0266577618 0.0024142074 0.0248865132 0.0245794863
##           21           22           23           24           25
## 0.0003712976 0.0252897585 0.0074672328 0.0259065674 0.0103283125
```

Points 2 and 4 are most influential.