Homework Chapter 3

Jinhong Du 15338039

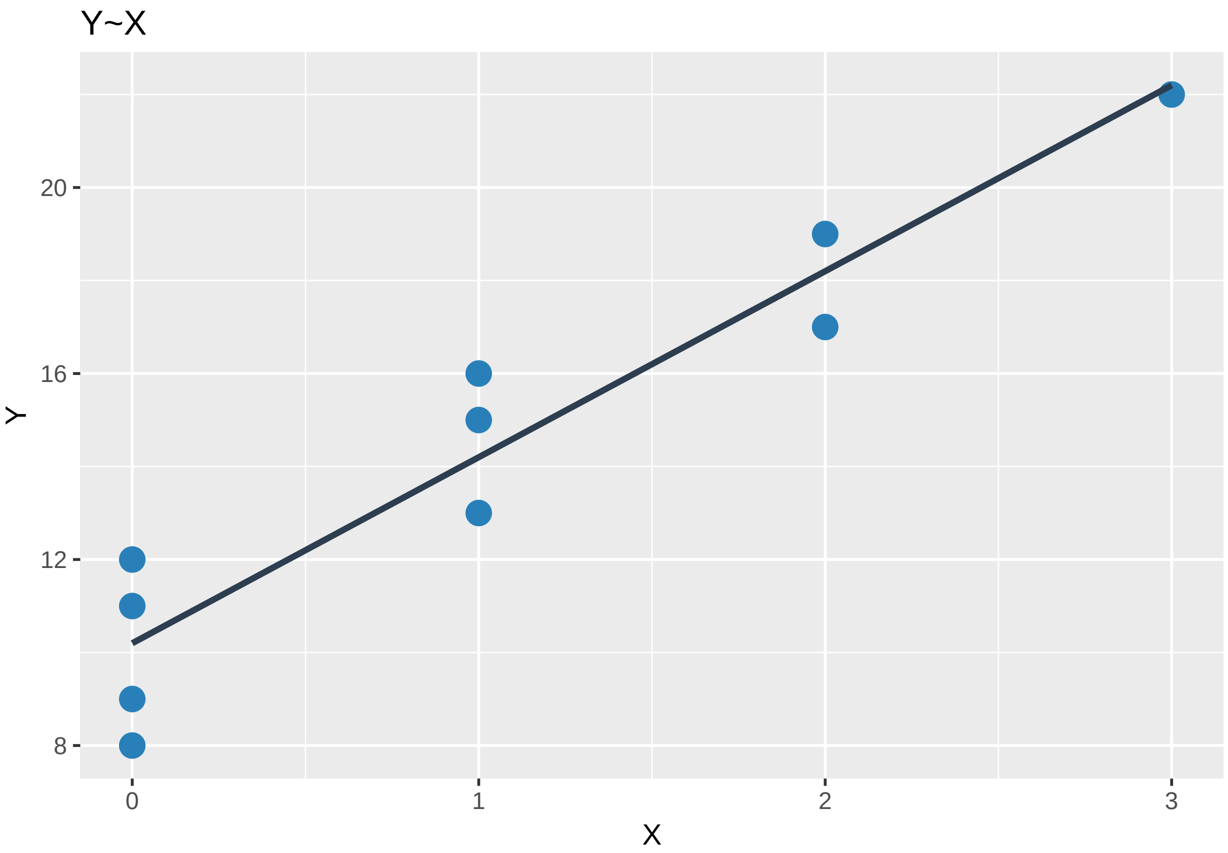
3.5 Refer to Airfreight breakage Problem 1.21.

(e) Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality to ascertain whether the normality assumption is reasonable here. Use Table B.6 and $\alpha = .01$. What do you conclude?

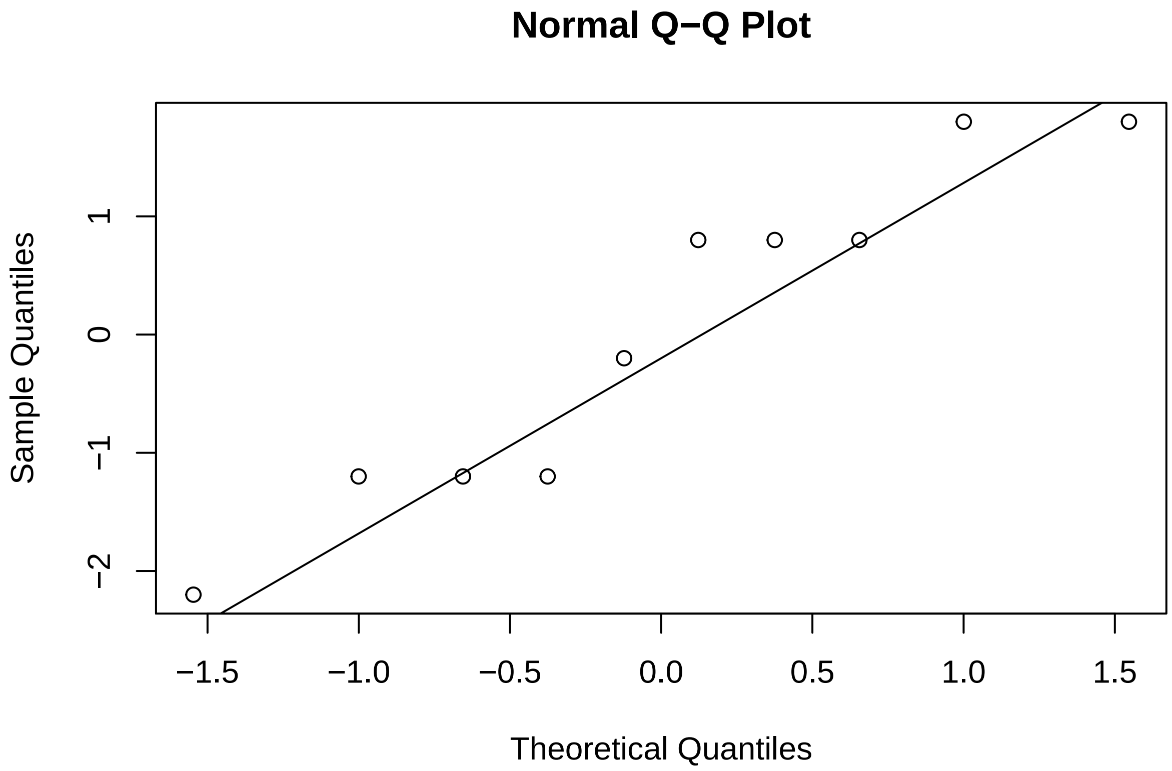
```
library(ggplot2)
library(gridExtra)
data1 <- read.table("CH01PR21.txt",head=FALSE,col.names = c('Y','X'))
fit <- lm('Y~X',data1)
summary(fit)
```

```
## 
## Call:
## lm(formula = "Y~X", data = data1)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
##   -2.2   -1.2    0.3    0.8    1.8 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  10.2000     0.6633  15.377 3.18e-07 ***
## X             4.0000     0.4690   8.528 2.75e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.483 on 8 degrees of freedom
## Multiple R-squared:  0.9009, Adjusted R-squared:  0.8885 
## F-statistic: 72.73 on 1 and 8 DF,  p-value: 2.749e-05
```

```
lm.scatter <- ggplot(data1, aes(x=X, y=Y)) +
  geom_point(color='#2980B9', size = 4) + xlim(c(0, 3)) +
  geom_smooth(method = lm, se=FALSE, fullrange=TRUE, color='#2C3E50', size=1.1) +
  labs(title='Y~X')
grid.arrange(lm.scatter)
```



```r
q <- qqnorm(fit$residuals)
qqline(fit$residuals)
```



```
library(olsrr,warn.conflicts=FALSE)
resid=sort(fit$residuals)
n = length(data1$X)
k <- c(1:n)
z=qnorm((k-0.375)/(n+0.25))
MSE = sum(fit$residuals^2)/(fit$df.residual)
expect_residual = z* sqrt(MSE)
print(sprintf('The expected values of residuals is :'))
```

```
## [1] "The expected values of residuals is :"
```

```
expect_residual
```

```
##  [1] -2.2940308 -1.4839673 -0.9721502 -0.5568998 -0.1818168  0.1818168
##  [7]  0.5568998  0.9721502  1.4839673  2.2940308
```

```
ols_corr_test(fit)
```

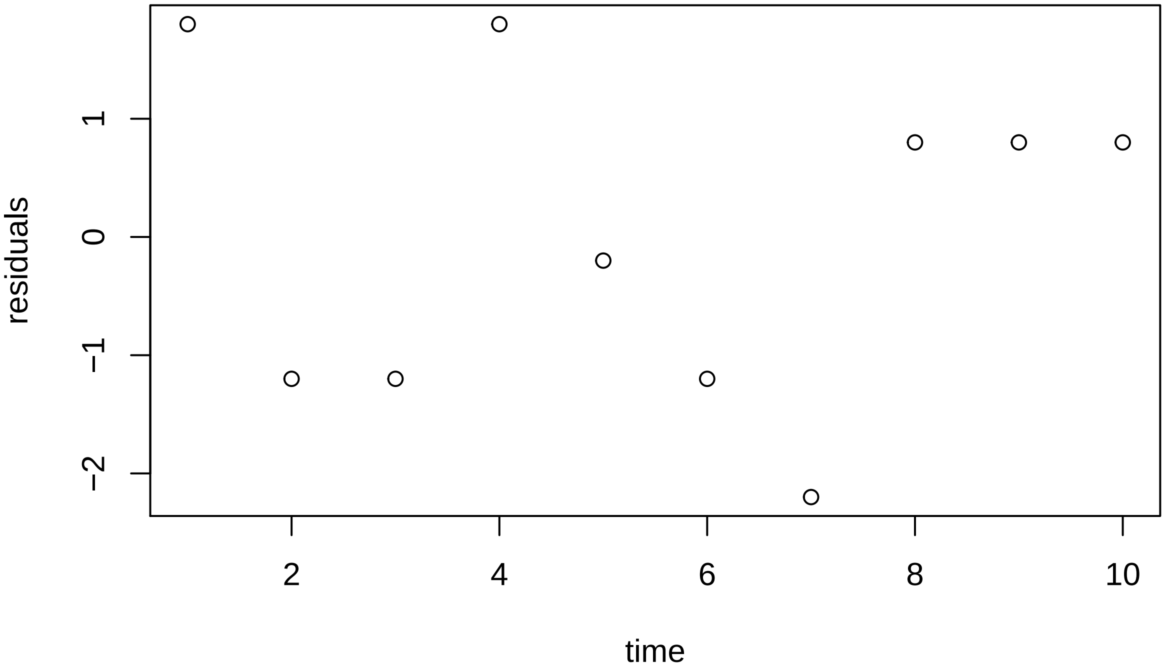
```
## [1] 0.9609751
```

$$H_0 : \text{Normal} \qquad H_a : \text{not normal}$$

$r = 0.9609751 \geqslant 0.879$, conclude H_0.

(f) Prepare a time plot of the residuals. What information is provided by your plot?

```
plot(k,fit$residuals,xlab = 'time',ylab = 'residuals')
```



The residual versus time plot did not show any evidence that the error terms were correlated over time.

(g) Assume that (3.10) is applicable and conduct the Breusch-Pagan test to determine whether or not the error variance varies with the level of X. Use $\alpha = .10$. State the alternatives, decision rule, and conclusion. Does your conclusion support your preliminary findings in part (d)?

```r
library(lmtest,warn.conflicts = FALSE )
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
library(latex2exp,warn.conflicts = FALSE )
data2 = data.frame(x=data1$X,y=fit$residuals^2)
fit2 = lm('y~x',data2)
SSR_star = sum((fitted(fit2)-mean(data2$y))^2)
SSE = sum(fit$residuals^2)
XBP = (SSR_star/2) / (SSE/n)^2
print(sprintf('SSR* :%f',SSR_star))
```

```
## [1] "SSR* :6.400000"
```

```r
print(sprintf('SSE  :%f',SSE))
```

```
## [1] "SSE  :17.600000"
```

```r
print(sprintf('XBP^2:%f',XBP))
```

```
## [1] "XBP^2:1.033058"
```

```r
bptest(data1$Y ~ data1$X,studentize=FALSE)
```

```
##
##  Breusch-Pagan test
##
## data:  data1$Y ~ data1$X
## BP = 1.0331, df = 1, p-value = 0.3094
```

$\because$ when H_0 holds,

$$X_{BP}^2 \dot{\sim} \chi^2(1)$$

$\therefore$

$$H_0 : \gamma_1 = 0 \qquad H_a : \gamma_1 \neq 0$$

where

$$\log_e \sigma_i^2 = \gamma_0 + \gamma_1 X_i$$

If $X_{BP}^2 < \chi^2(0.9; 1) = 2.71$, then conclude H_0; otherwise conclude H_a.

Here $X_{BP}^2 = 1.0331 < 2.71$, conclude H_0.

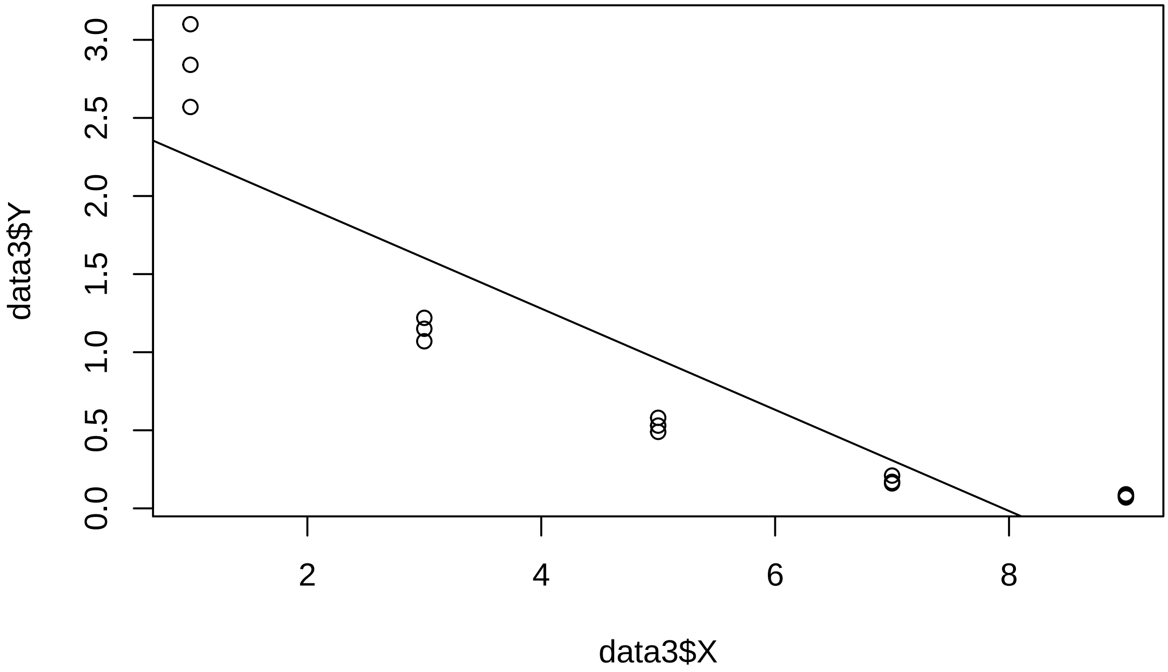
3.15 Solution concentration. A chermist studied the concentration of a solution (Y) over time (X). Fifteen identical solutions were prepared. The 15 solutions were randomly divided into five sets of three, and the five sets were measured, respectively, after 1, 3, 5, 7, and 9 hours. The results follow.

(a) Fit a linear regression function.

```r
data3 <- read.table("CH03PR15.txt",head=FALSE,col.names = c('Y','X'))
fit3 <- lm('Y~X',data3)
summary(fit3)
```

```
##
## Call:
## lm(formula = "Y~X", data = data3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5333 -0.4043 -0.1373  0.4157  0.8487
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.5753     0.2487  10.354 1.20e-07 ***
## X            -0.3240     0.0433  -7.483 4.61e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4743 on 13 degrees of freedom
## Multiple R-squared:  0.8116, Adjusted R-squared:  0.7971
## F-statistic: 55.99 on 1 and 13 DF,  p-value: 4.611e-06
```

```
plot(data3$X,data3$Y)
abline(fit3)
```



(b) Perform the F test to determine whether or not there is lack of fit of a linear regression function; use $\alpha = .025$. State the alternatives, decision rule, and conclusion.

```
c = length(unique(data3$X))
n = length(data3$X)
SSER = sum(fit3$residuals^2)
SSEF = 0
for (i in unique(data3$X)) {
  SSEF <- SSEF + sum((data3[data3$X==i,]$Y-mean(data3[data3$X==i,]$Y))^2)
}
Fvalue = (SSER-SSEF)/(c-2)/(SSEF/(n-c))
```

```r
print(sprintf('SSE of Reduced Model :%f',SSER))
```

```
## [1] "SSE of Reduced Model :2.924653"
```

```r
print(sprintf('SSE of Full Model    :%f',SSEF))
```

```
## [1] "SSE of Full Model    :0.157400"
```

```r
print(sprintf('F-value              :%f',Fvalue))
```

```
## [1] "F-value              :58.603417"
```

```r
print(sprintf('0.975 Quantile F(%d,%d) value:%f',c-2,n-c,qf(0.975,c-2,n-c)))
```

```
## [1] "0.975 Quantile F(3,10) value:4.825621"
```

$$H_0 : \text{the regression function is linear} \qquad H_a : \text{the regression function is not linear}$$

$$F^* = \frac{\frac{SSE(R) - SSE(F)}{3}}{\frac{SSE(F)}{10}} \sim F(3, 10)$$

If $F^* \leqslant F(0.975; 3, 10)$, then conclude H_0; otherwise conclude H_a.

Here, $F^* = 58.603417 > 4.825621$, conclude H_a.

(c)

Does the test in part (b) indicate what regression function is appropriate when it leads to the conclusion that lack of fit of a linear regression function exists? Explain.

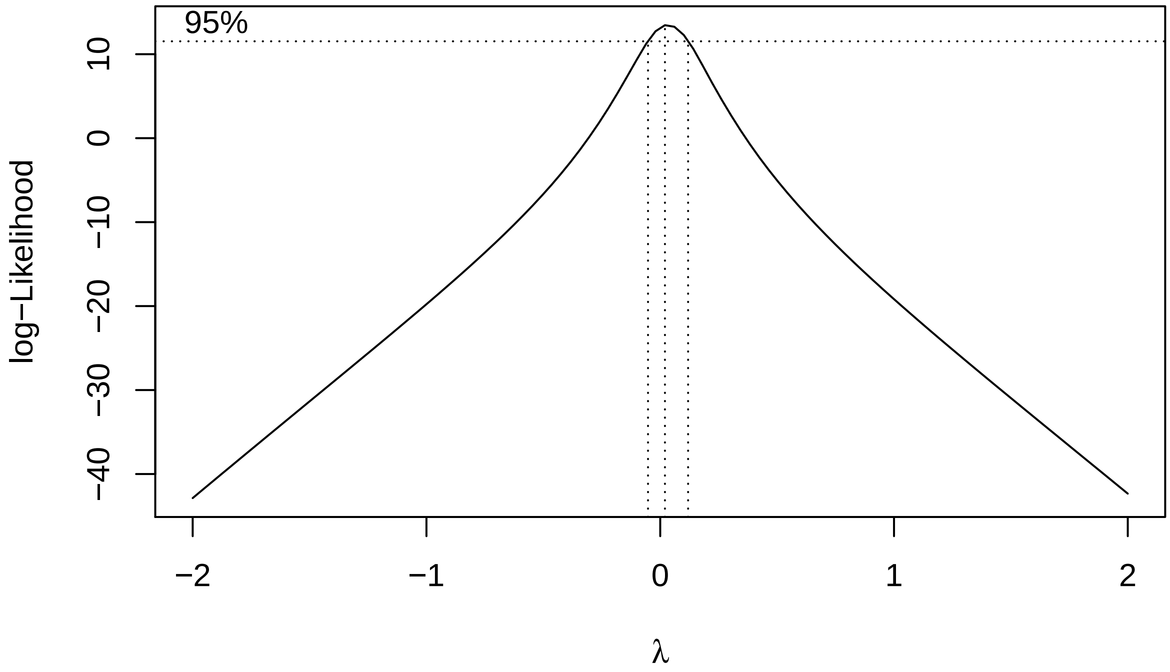
No, since the alternative H_a includes all regrssion functions other than a linear one.

3.16

Refer to **Solution concentration** Problem 3.15.

(b) Use the Box—Cox procedure and standardization (3.36) to find an appropriate power transformation. Evaluate SSE for $\lambda = -.2, -.1, 0, .1, .2$. What transformation of Y is suggested?

```r
library(MASS,warn.conflicts = FALSE)
a = boxcox(fit3)
```



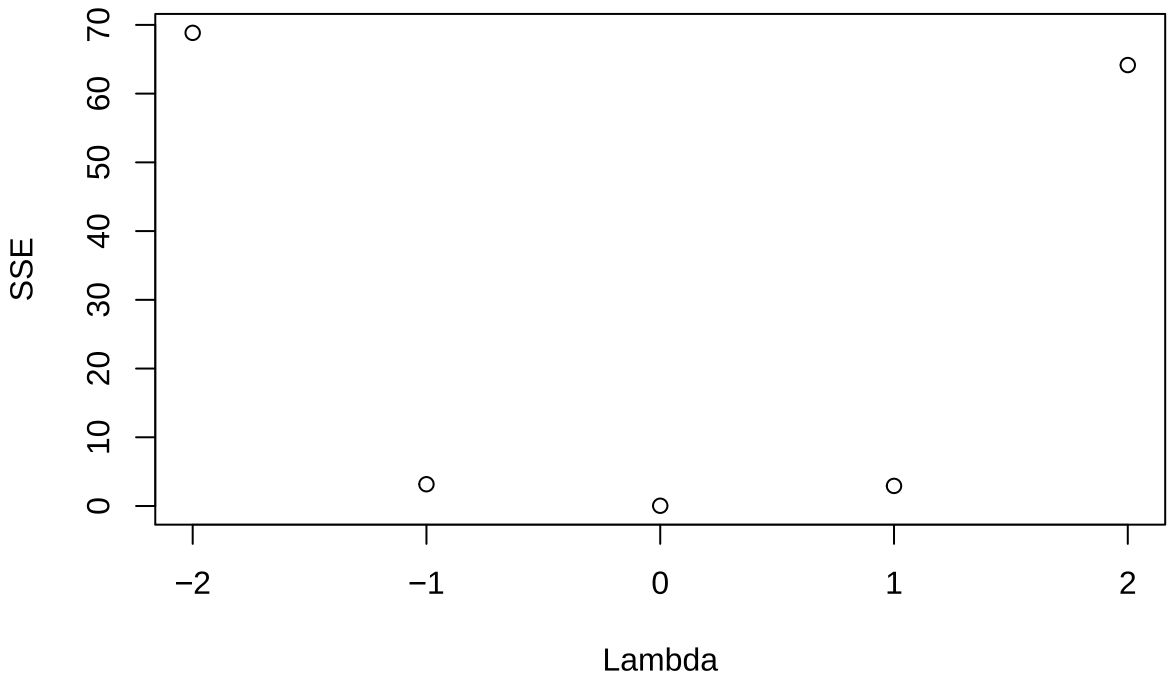
```
print(sprintf('The best power returned by Box-Cox is %f',a$x[which.max(a$y)]))
```

```
## [1] "The best power returned by Box-Cox is 0.020202"
```

```
SSE = c(0,0,0,0,0);
lam = c(-2,-1,0,1,2);
K2 = (prod(data3$Y))^(1/length(data3$Y));
print(K2)
```

```
## [1] 0.4762974
```

```
for (i in c(1:5)){
  data4 = data3
  if(lam[i]==0){
    data4$Y = K2*log(data4$Y)
  }
  else{
    K1 = 1/lam[i]/K2^(lam[i]-1)
    data4$Y = K1*(data4$Y^lam[i] - 1)
  }
  SSE[i] = sum(lm('Y~X',data4)$residuals^2);
}
plot(c(-2:2),SSE,xlab='Lambda')
```

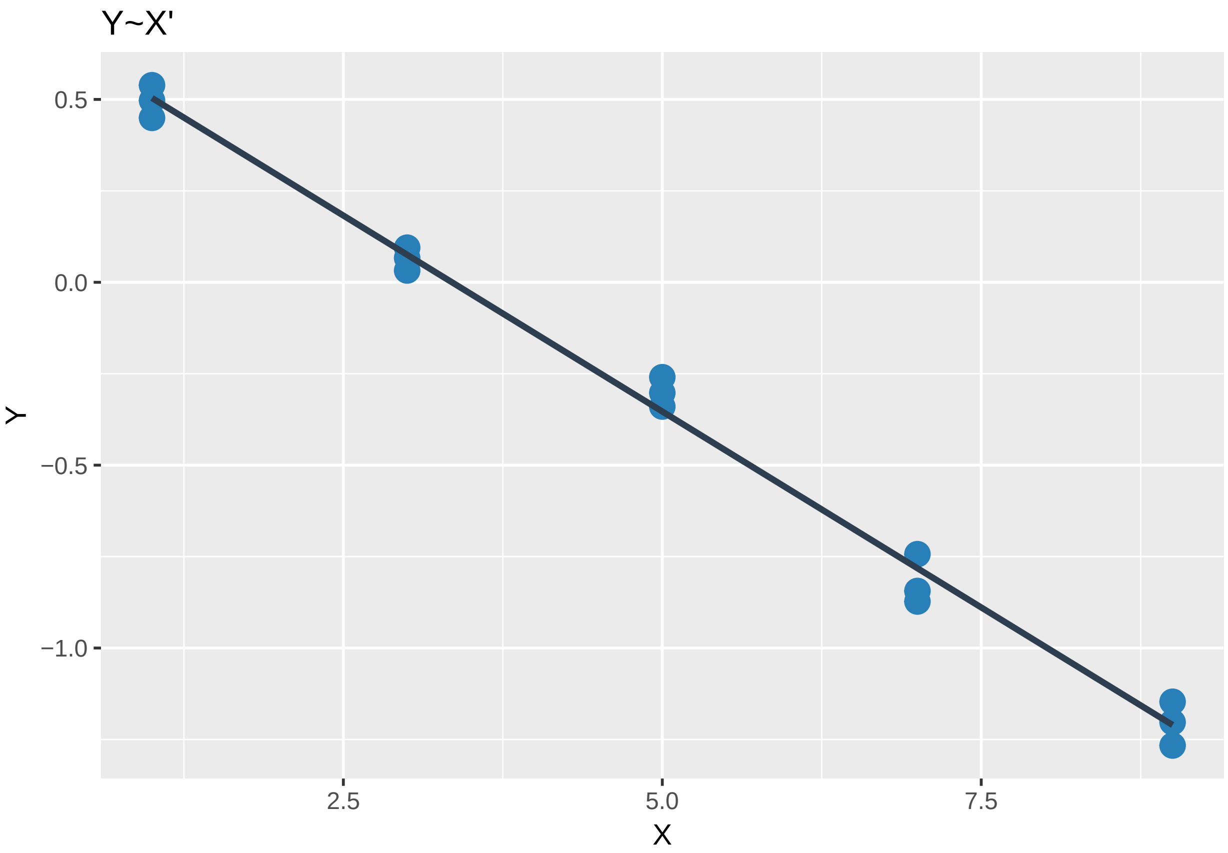


(c) Use the transformation $Y' = \log_{10} Y$ and obtain the estimated linear regression function for the transformed data.

```
data4 = data3
data4$Y = K2*log(data4$Y)
fit4 =lm('Y~X',data4)
summary(fit4)
```

```
##
## Call:
## lm(formula = "Y~X", data = data4)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.090981 -0.048717  0.007472  0.036753  0.093824
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.718217   0.028713   25.01 2.22e-12 ***
## X           -0.214299   0.004998  -42.88 2.19e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05475 on 13 degrees of freedom
## Multiple R-squared:  0.993,  Adjusted R-squared:  0.9924
## F-statistic:  1838 on 1 and 13 DF,  p-value: 2.188e-15
```

```
lm.scatter <- ggplot(data4, aes(x=X, y=Y)) +
  geom_point(color='#2980B9', size = 4)  +
  geom_smooth(method = lm, se=FALSE, fullrange=TRUE, color='#2C3E50', size=1.1) +
  labs(title='Y~X\'')
grid.arrange(lm.scatter)
```



The regression function is

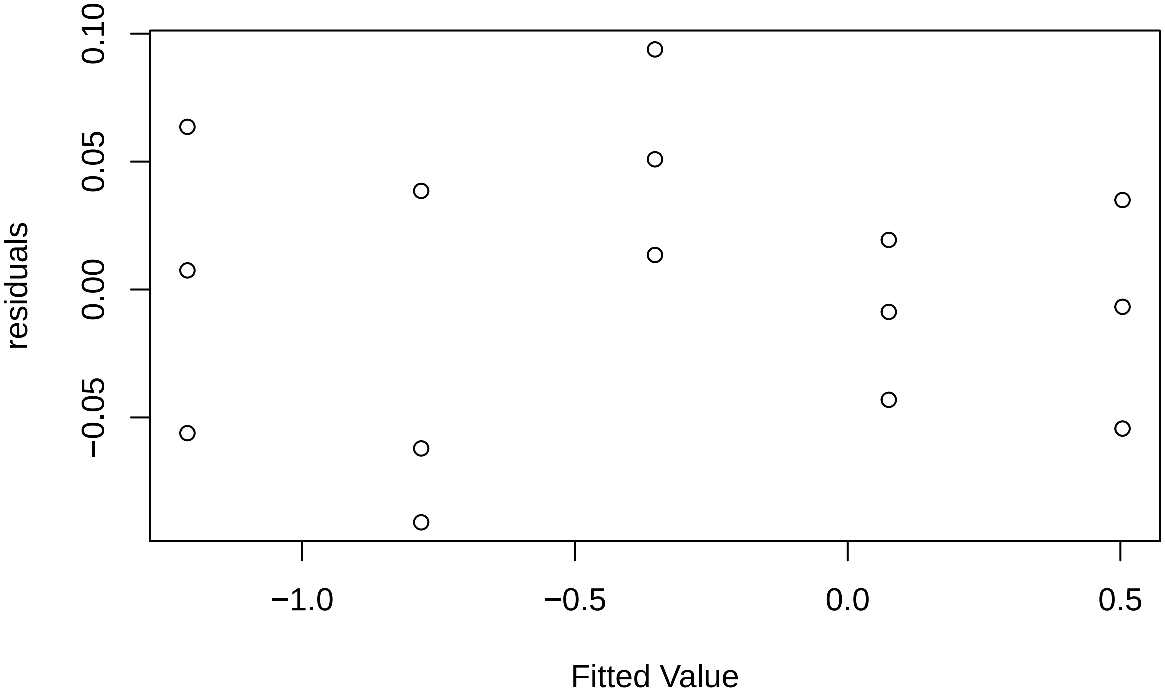
$$Y' = -0.214299 + 0.718217$$

i.e.

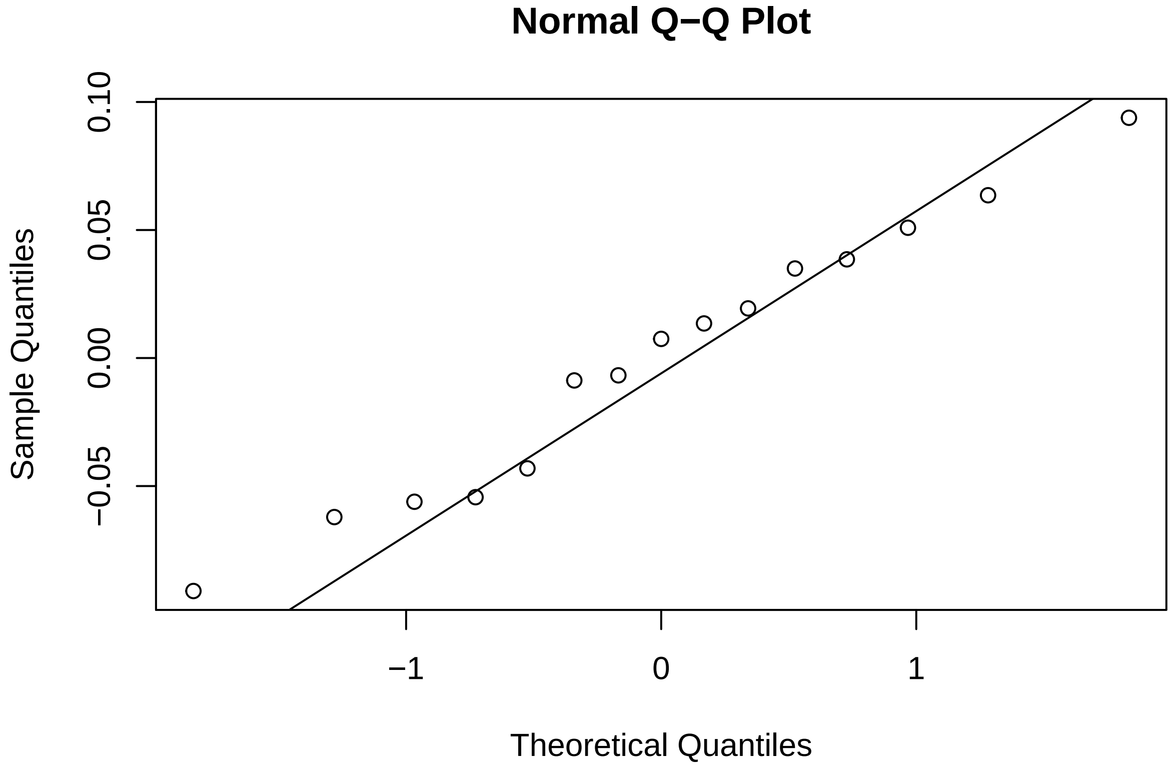
$$Y = e^{\frac{-0.44993X + 1.50792}{0.4762974}}$$

(e) Obtain the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?

```
plot(fit4$fitted.values,fit4$residuals,xlab = 'Fitted Value',ylab = 'residuals')
```



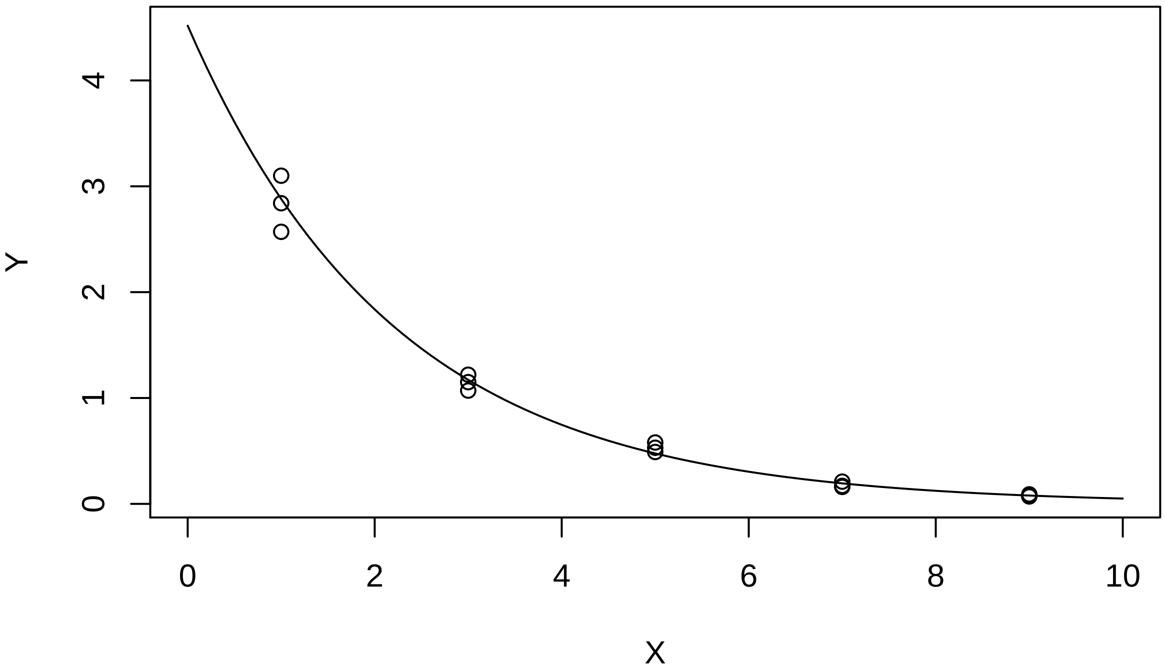
```
q <- qqnorm(fit4$residuals)
qqline(fit4$residuals)
```



In the first plot, residuals lie randomly near x-axis. In the second plot, the points lie near a straight line. Therefore, the plots show that the transformed model is linear and normal.

(f) Express the estimated regression function in the original units.

```r
d = data.frame('X'=seq(0,10,0.1))
tyh = predict(fit4,d)
plot(d$X,exp(tyh/K2),xlab = 'X',ylab = 'Y',type = 'l')
points(data3$X,data3$Y)
```



$$Y = e^{\frac{-0.44993X+1.50792}{0.4762974}}$$