STAT 572 **Hw2** **Suggested Solution**

Problem 1:

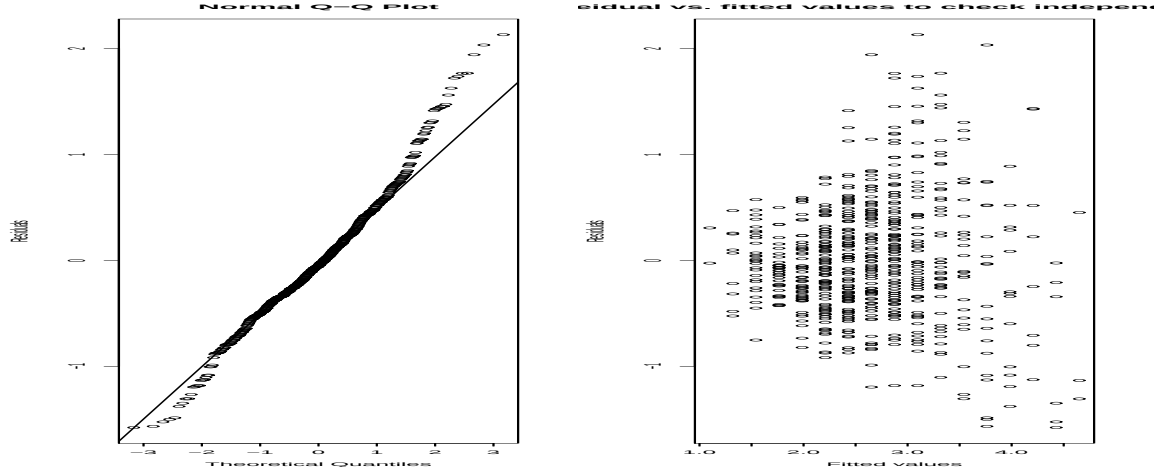
```
>fevdata <- read.table("fevdata.txt", header=T) ## read data fevdata.txt into R
>attach (fevdata)
>fev.lm1<- lm(fev~age) ## fit linear model
>summary(fev.lm1) ## summarize the fit model

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.431648   0.077895   5.541 4.36e-08 ***
age         0.222041   0.007518  29.533  < 2e-16 ***


Residual standard error: 0.5675 on 652 degrees of freedom
Multiple R-Squared: 0.5722,     Adjusted R-squared: 0.5716
F-statistic: 872.2 on 1 and 652 DF,  p-value: < 2.2e-16
```

Hence our simple linear model is : fev = 0.43 + 0.22*(age)

```
R code:
>par(mfrow=c(1,3))
>qqnorm(residuals(fev.lm1), ylab="Residuals") ## Q-Q plot
>qqline(residuals(fev.lm1))
>plot(fev.lm1$resid~fev.lm1$fitted, xlab="Fitted values",ylab="Residuals",
main="Reidual vs. fitted values to check independence")
```



1. Normal Q-Q plot: The residuals can be assessed for normality using a Q-Q plot. Normal residuals should follow the line approximately. From the above Q-Q plot, the distribution of studentized residuals looks heavy-tailed, hence we might just accept the fact that the residuals are **nonnormality**.

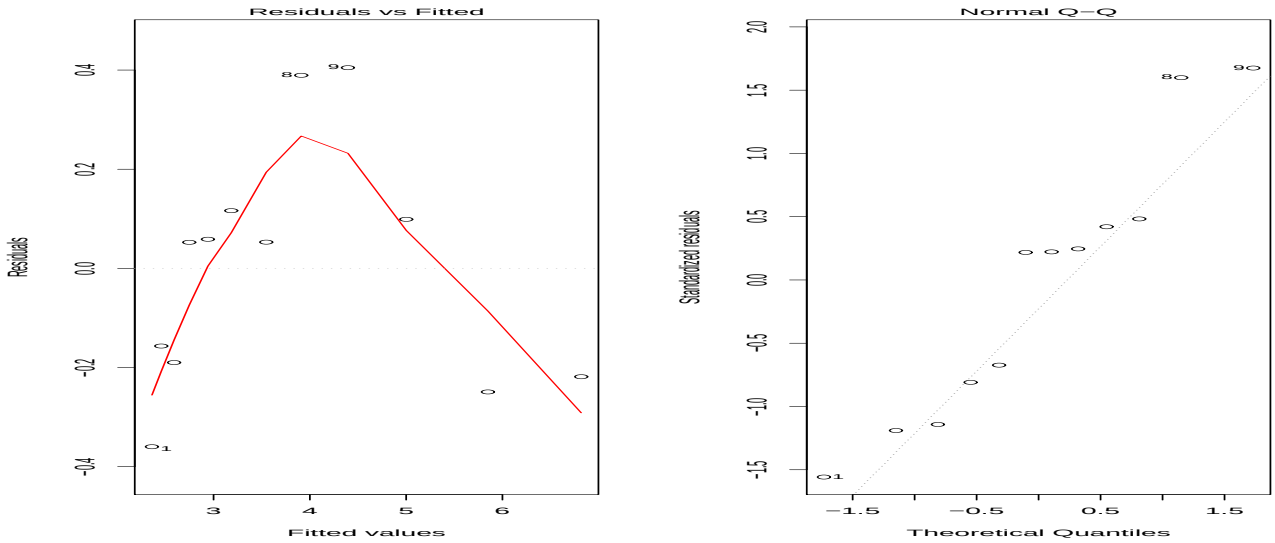
2. Residuals vs. fitted values: check the **Equal Variance**. If observations are equal variance, then there should be no clear pattern in this plot. This plot is fan-shaped pattern, the reason might be that the equal variance assumptions is not fulfilled.

Note: there is no general way to check independence. For independent assumption, people need to go back to check how the data are collected.

From 1 and2, the assumptions of linear regression model are not satisfied, hence the fit model is not acceptable.

Problem 2

Part(a)

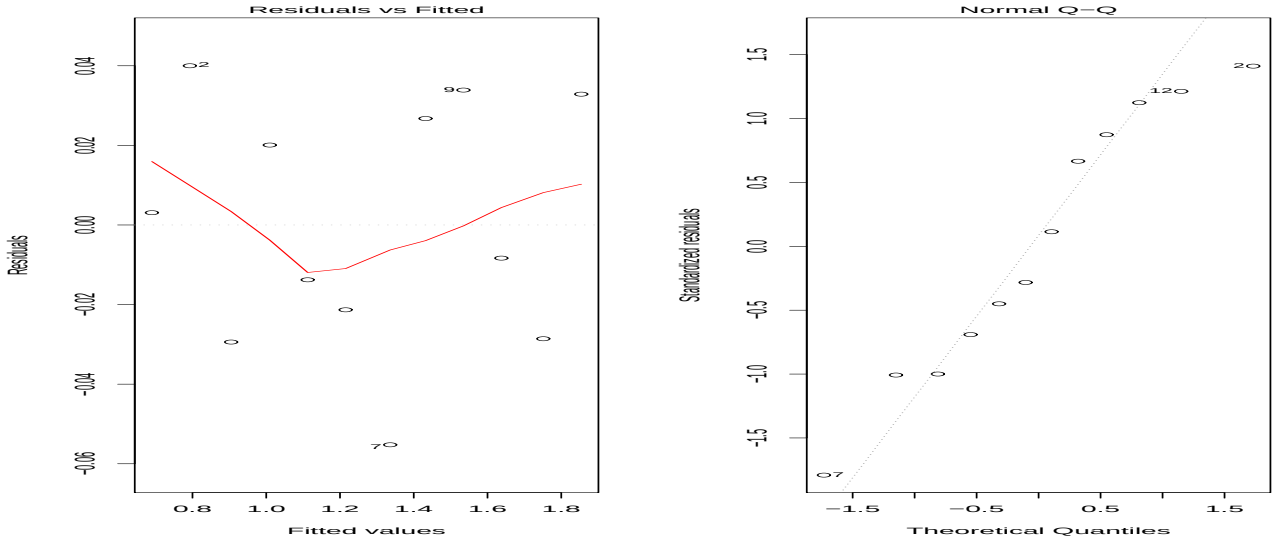


The residual plot shows a convex pattern, which indicates some **nonlinearity**, which should prompt some change in the structural form of the model.

R code:

```
>conc =c(0.32,0.40, 0.51, 0.64, 0.80, 1.0, 1.3, 1.6, 2.0, 2.5, 3.2, 4.0)
>inhibit=c(2.0, 2.3, 2.4, 2.8, 3.0, 3.3, 3.6, 4.3, 4.8, 5.1, 5.6, 6.6)
##fit linear regression model using original data:
>fit.lm <- lm(inhibit~conc)
>par(mfrow=c(1,2))
>plot(fit.lm, which=1)
>plot(fit.lm, which=2)
```

Fit linear regression model after log transformation for both response and predictor:



There is no obvious pattern in the residual plot of log transforms. In addition, normal residuals follows the line approximately.

Hence, the the second model (log-transfrom) is better.

```
R code:
>logfit.lm <- lm(log(inhibit)~log(conc))
>par(mfrow=c(1,2))
>plot(logfit.lm, which=1)
>plot(logfit.lm, which=2)
```

Part(b)

For logfit.lm, the 95% confidence interval for the log of mean inhibition at a concentration of log(2.8) is **(1.66, 1.72)**. This means that we can be 95% confident that the log of mean inhibition at concentration of 2.8 is in the interval (1.66, 1.72).

The 95% confidence interval for the mean inhibition at a concentration of 2.8 is (5.25, 5.59). This means that we can be 95% confident that the mean inhibition at concentration of 2.8 is in the interval (5.25, 5.59).

```
R code and outputs
>exp(predict(logfit.lm, data.frame(conc=c(2.8)), interval="confidence"))
          fit      lwr      upr
[1,] 5.418436 5.250449 5.591797
```

Part (c)

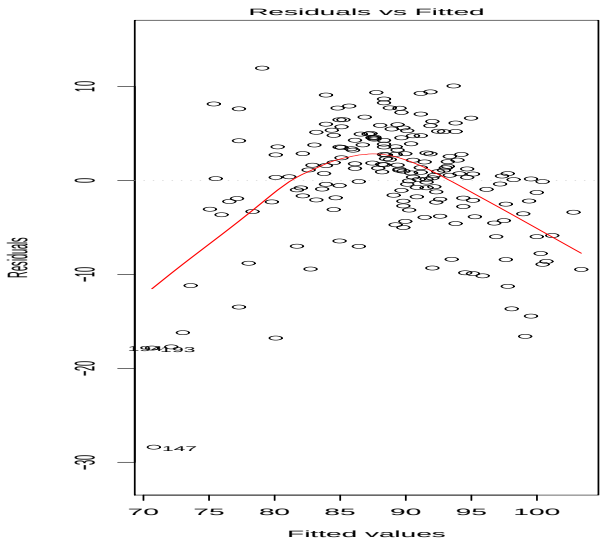
If conc=3.3, the prediction interval is (5.396, 6.331); if conc=1.3, the prediction interval is (3.530, 4.100). The prediction interval of conc=3.3 is wider than the prediction interval of conc=1.3. Please refer to the page 3 of lecture 2 handout. The prediction interval is calculated by $\bar{X} + / - T_\alpha S_n \sqrt{1 + (1/n)}$. The standard error for estimation of conc at 3.3 is greater than that of conc at 1.3 and other terms remain same, so the prediction interval of conc=3.3 is wider than that of conc=1.3.

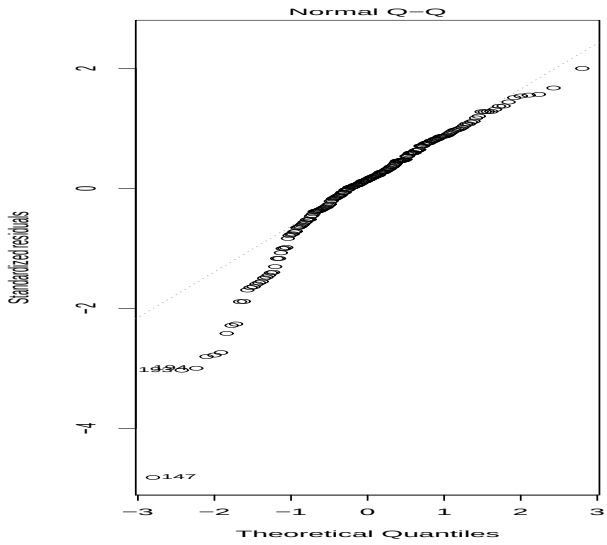
```
R code and outputs
>exp(predict(logfit.lm, data.frame(conc=c(3.3,1.3)), interval="prediction") )
       fit      lwr      upr
1 5.844709 5.396044 6.330678
2 3.804438 3.529928 4.100296
```

Problem 3

Part (a)

After fitting linear model Height=62.5+1.5(DBH), we draw the residual and Q-Q plots. The residual plot is curvature, which indicates some **nonlinearity**. For the Q-Q plot, residuals do **not** follow the line approximately, hence the residuals are **nonnormality**.





R code:

```
>trees<- read.table("tree.txt", header=T)
>attach(trees)
>trees.lm = lm (Height~DBH)
>par(mfrow=c(1,2))
>plot(trees.lm, which=1)
>plot(trees.lm, which=2)
```

Part (b):

For those trees with 15 inches or larger diameter, the linear model is: Height= 83.15+0.46(DBH).

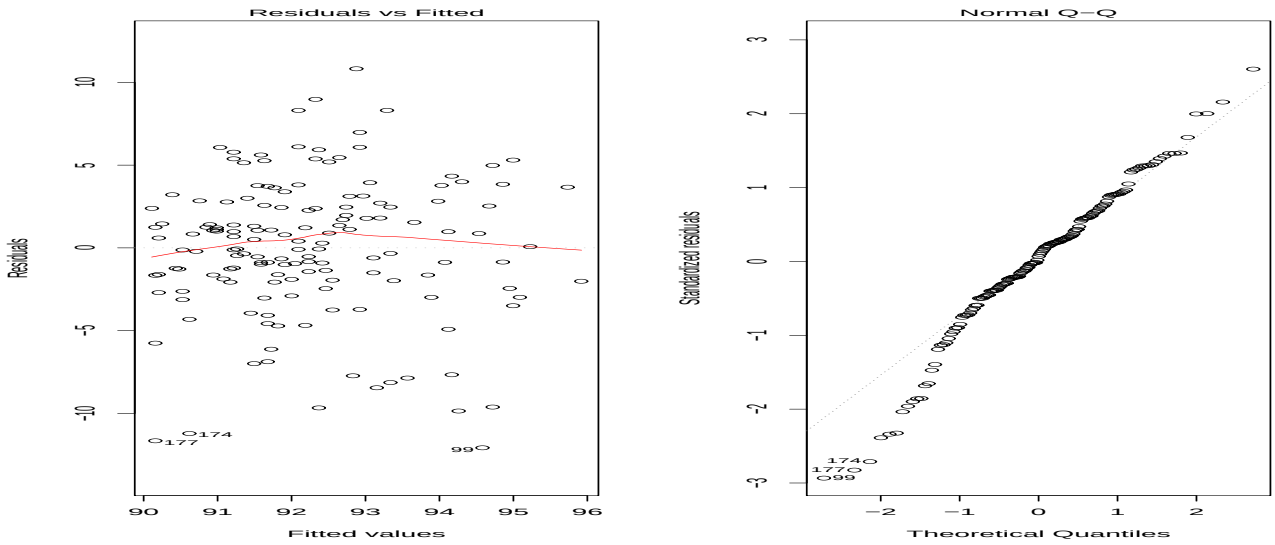
R code:

```
>trees15.lm = lm(Height~DBH, subset=(DBH>15),data=trees)
>summary(trees15.lm)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  83.1549     2.3162   35.90  < 2e-16 ***
DBH           0.4607     0.1160    3.97 0.000111 ***

Residual standard error: 4.173 on 151 degrees of freedom
```

```
Multiple R-Squared: 0.09453,    Adjusted R-squared: 0.08853
F-statistic: 15.76 on 1 and 151 DF,  p-value: 0.0001107
```

Part(c)



The residual plot does not have any pattern except there are three points (99, 174 and 177) with big residuals. The Q-Q plot has a long-tailed error. Hence it is possible that these three points are outliers. Generally, the linear regression assumptions are fulfilled well.

```
R code:
>par(mfrow=c(1,2))
>plot(trees15.lm, which=1)
>plot(trees15.lm, which=2)
```

Part(d)

The confidencd interval for intercept is: (78.58, 87.73)

The confidencd interval for slope is: (0.23,0.69)

```
R code:
83.1549+2.3162*qt(0.975,151)
83.1549-2.3162*qt(0.975,151)
0.4607+0.116*qt(0.975,151)
0.4607-0.116*qt(0.975,151)
```

Part (e):

For DBH=5.5, 95% prediction interval is: (76.8, 94.6)

For DBH=7.5, 95% prediction interval is: (77.9, 95.3)

```
R code:
>predict(trees15.lm, data.frame(DBH=c(5.5,7.5)), interval="prediction")
       fit      lwr      upr
1 85.68897 76.79571 94.58224
2 86.61047 77.87522 95.34571
```

Part (f):

There is one tree (ID=589) with DBH=5.5 and one tree (ID=507) with DBH=7.5. However, the heights **do not** fall into the prediction intervals. One possible reason is the model is fitted by considering the subset of trees for which the diameter at breast height exceeds 15 inches. While DBH =5.5 and 7.5 are less than 15. Using the regression equation to predict values of the dependent variable outside the range of the independent variable is not recommended since we have no evidence that the same linear relationship exists outside the observed range