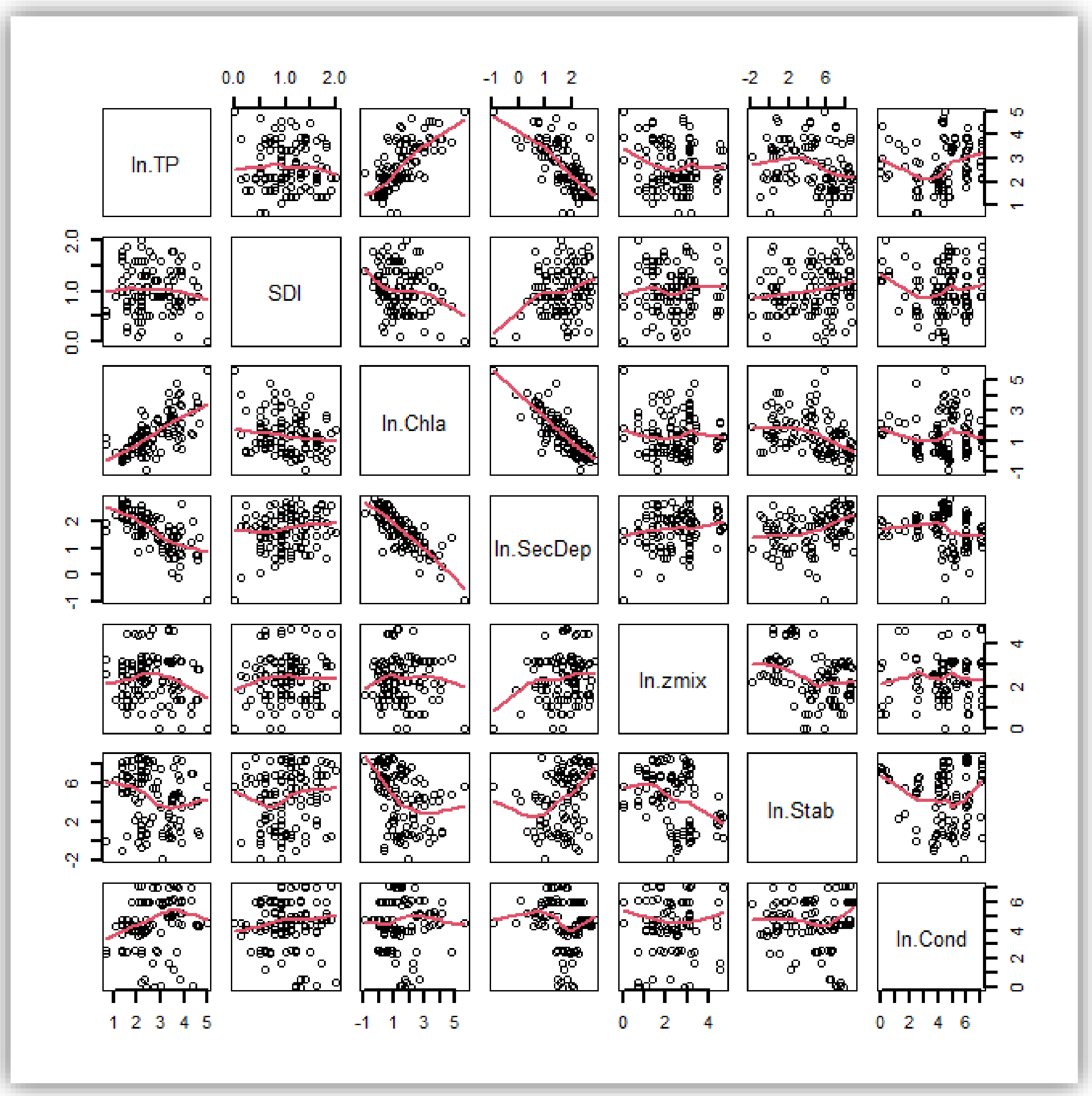
Q1.

Below is the scatterplot-matrix for the variables

```
lakewater.df = read.csv('lakewater.csv', header = TRUE)

x11()

plot(lakewater.df, panel = panel.smooth, lwd = 2)
```



Predictor variables that seem to be related to ln.TP are

- ln.Chla appears to have a strong positive relationship with ln.TP
- ln.SecDep seems to have a negative relationship with ln.TP.

Predictor variables that seem to be related to each other are:

- ln.Chla and ln.SecDep seem to have a strong negative relationship
- ln.Stab seem to have relationship with ln.Chla and ln.SecDep

Most of the relationships mentioned in answer seem to be linear. With ln.Chla and ln.SecDep having very strong linear relationship.

Q2.

Correlation Matrix

```
> round(cor(lakewater.df), 3)

           ln.TP    SDI ln.Chla ln.SecDep ln.zmix ln.Stab ln.Cond
ln.TP      1.000 -0.063   0.723    -0.715  -0.112  -0.252   0.203
SDI       -0.063  1.000  -0.217     0.257   0.073   0.172   0.001
ln.Chla    0.723 -0.217   1.000    -0.824  -0.049  -0.402   0.019
ln.SecDep -0.715  0.257  -0.824     1.000   0.219   0.349  -0.192
ln.zmix   -0.112  0.073  -0.049     0.219   1.000  -0.344   0.030
ln.Stab   -0.252  0.172  -0.402     0.349  -0.344   1.000  -0.069
ln.Cond    0.203  0.001   0.019    -0.192   0.030  -0.069   1.000
```

- So we can see the green highlighted correlations are high suggesting strong relationship with ln.TP
- And the correlation between ln.Chla and ln.SecDep(highlighted in blue) is also very strong
- ln.Stab also have some significant correlation values with ln,Chla and ln.SecDep. Suggesting some relationship in them.
- Yes, these confirm the relations between variables we saw in the scatterplot-matrix

Q3.

So after doing stepwise regression using AIC decision criteria to find Best Model we got following result:

```
> lakewater.step.aic.lm = step(lakewater.full.lm, direction = "both", trac
e = TRUE)
Start:  AIC=-102.7
ln.TP ~ SDI + ln.Chla + ln.SecDep + ln.zmix + ln.Stab + ln.Cond

            Df Sum of Sq    RSS      AIC
- ln.zmix    1    0.0166 42.434 -104.655
- ln.Stab    1    0.1400 42.557 -104.318
```

```
<none>                   42.417 -102.700
- SDI        1    1.4513 43.868 -100.798
- ln.Cond    1    1.7101 44.127 -100.115
- ln.SecDep  1    2.6992 45.116  -97.544
- ln.Chla    1    7.3537 49.771  -86.155

Step:  AIC=-104.65
ln.TP ~ SDI + ln.Chla + ln.SecDep + ln.Stab + ln.Cond

            Df Sum of Sq    RSS      AIC
- ln.Stab    1    0.2268 42.661 -106.036
<none>                   42.434 -104.655
- SDI        1    1.4380 43.872 -102.789
+ ln.zmix    1    0.0166 42.417 -102.700
- ln.Cond    1    1.6948 44.129 -102.112
- ln.SecDep  1    3.3035 45.737  -97.958
- ln.Chla    1    7.4811 49.915  -87.819

Step:  AIC=-106.04
ln.TP ~ SDI + ln.Chla + ln.SecDep + ln.Cond

            Df Sum of Sq    RSS      AIC
<none>                   42.661 -106.036
+ ln.Stab    1    0.2268 42.434 -104.655
+ ln.zmix    1    0.1035 42.557 -104.318
- SDI        1    1.5616 44.222 -103.866
- ln.Cond    1    1.6231 44.284 -103.705
- ln.SecDep  1    3.3026 45.963  -99.387
- ln.Chla    1    7.2742 49.935  -89.773
```

The Best Model chosen here by AIC decision criterion includes following variables:

- SDI
- ln.Cond
- ln.SecDep
- ln.Chla

This model has the lowest AIC (-106.036), indicating it provides the best balance between goodness of fit and model complexity according to the AIC criterion.

Q4.

After doing Step-wise regression using the BIC decision criterion to find the “best” model to fit the data we got following result:

```
> lakewater.step.bic.lm = step(lakewater.full.lm, direction = "both", trac
e = TRUE, k = log(nrow(lakewater.df)))
Start:  AIC=-83.43
ln.TP ~ SDI + ln.Chla + ln.SecDep + ln.zmix + ln.Stab + ln.Cond

            Df Sum of Sq    RSS     AIC
- ln.zmix    1    0.0166 42.434 -88.133
- ln.Stab    1    0.1400 42.557 -87.797
- SDI        1    1.4513 43.868 -84.276
- ln.Cond    1    1.7101 44.127 -83.594
<none>                   42.417 -83.425
- ln.SecDep  1    2.6992 45.116 -81.022
- ln.Chla    1    7.3537 49.771 -69.633
```

```
Step:  AIC=-88.13
ln.TP ~ SDI + ln.Chla + ln.SecDep + ln.Stab + ln.Cond

            Df Sum of Sq    RSS     AIC
- ln.Stab    1    0.2268 42.661 -92.268
- SDI        1    1.4380 43.872 -89.021
- ln.Cond    1    1.6948 44.129 -88.344
<none>                   42.434 -88.133
- ln.SecDep  1    3.3035 45.737 -84.191
+ ln.zmix    1    0.0166 42.417 -83.425
- ln.Chla    1    7.4811 49.915 -74.051

Step:  AIC=-92.27
ln.TP ~ SDI + ln.Chla + ln.SecDep + ln.Cond

            Df Sum of Sq    RSS     AIC
- SDI        1    1.5616 44.222 -92.852
- ln.Cond    1    1.6231 44.284 -92.690
<none>                   42.661 -92.268
- ln.SecDep  1    3.3026 45.963 -88.372
+ ln.Stab    1    0.2268 42.434 -88.133
+ ln.zmix    1    0.1035 42.557 -87.797
- ln.Chla    1    7.2742 49.935 -78.759

Step:  AIC=-92.85
ln.TP ~ ln.Chla + ln.SecDep + ln.Cond

            Df Sum of Sq    RSS     AIC
- ln.Cond    1    1.7989 46.021 -92.980
<none>                   44.222 -92.852
+ SDI        1    1.5616 42.661 -92.268
- ln.SecDep  1    2.7148 46.937 -90.694
+ ln.Stab    1    0.3504 43.872 -89.021
+ ln.zmix    1    0.0934 44.129 -88.343
- ln.Chla    1    7.3011 51.523 -79.880

Step:  AIC=-92.98
ln.TP ~ ln.Chla + ln.SecDep

            Df Sum of Sq    RSS     AIC
<none>                   46.021 -92.980
+ ln.Cond    1    1.7989 44.222 -92.852
+ SDI        1    1.7374 44.284 -92.690
+ ln.Stab    1    0.2622 45.759 -88.889
+ ln.zmix    1    0.0125 46.009 -88.258
- ln.SecDep  1    4.7164 50.738 -86.416
- ln.Chla    1    5.9762 51.997 -83.571
```

The Best Model chosen here by BIC decision criterion includes variables:

- ln.Chla
- ln.SecDep

This model has the lowest BIC (-92.98). The BIC tends to favor simpler models with fewer predictors compared to the AIC, which explains why the final model includes fewer variables than the model selected using the AIC.

Q5

To find the best subsets of regression variables of different sizes, function myallpossregs.func() was used. Below is the output:

Keeping n=3 we will find the best model using different variables of size 1,2,3,4 and 5

```
> View(myallpossregs.func)
> models=myallpossregs.func(lakewater.full.lm,3)
> models
   Vars   Rsq Rsq(adj) PRESS   AIC   BIC    Cp      s SDI ln.Chla ln.SecDep ln.zmix ln.Stab ln.Cond
1     1 0.523    0.519  52.3 239.3 247.5  18.4 0.6671   0       1         0       0       0       0
2     1 0.511    0.507  53.6 242.1 250.4  21.6 0.6754   0       0         1       0       0       0
3     1 0.064    0.055 103.0 317.6 325.8 144.1 0.9350   0       0         0       0       1       0
4     2 0.568    0.560  48.2 230.0 241.0   8.3 0.6382   0       1         1       0       0       0
5     2 0.559    0.551  50.0 232.2 243.3  10.6 0.6445   0       1         0       0       0       1
6     2 0.533    0.524  52.1 239.0 250.0  17.8 0.6635   1       1         0       0       0       0
7     3 0.585    0.573  47.8 227.3 241.1   5.6 0.6284   0       1         1       0       0       1
8     3 0.584    0.573  47.1 227.5 241.3   5.8 0.6288   1       1         1       0       0       0
9     3 0.570    0.559  48.9 231.3 245.1   9.6 0.6392   0       1         1       0       1       0
10    4 0.599    0.585  46.8 225.2 241.7   3.6 0.6199   1       1         1       0       0       1
11    4 0.588    0.573  48.3 228.4 244.9   6.7 0.6287   0       1         1       0       1       1
12    4 0.585    0.570  47.7 229.1 245.6   7.4 0.6305   1       1         1       0       1       0
13    5 0.601    0.583  47.3 226.5 245.8   5.0 0.6211   1       1         1       0       1       1

14    5 0.600    0.582  47.6 226.9 246.2   5.4 0.6220   1       1         1       1       0       1
15    5 0.588    0.569  49.3 230.4 249.7   8.7 0.6315   0       1         1       1       1       1
16    6 0.601    0.580  48.3 228.5 250.5   7.0 0.6238   1       1         1       1       1       1
```

1. Predictors = 1

   Best Model: ln.Cha (Model 1)

   Criteria: lowest PRESS, AIC, BIC and Cp

2. Predictors = 2

   Best Model: ln.Chla, ln.SecDep (Model 4)

   Criteria: Highest Adj R^2, lowest PRESS,AIC, BIC and Cp

3. Predictors = 3

   Best Model: ln.Chla, ln.SecDep, ln.Cond (Model 7)

   Criteria: Highest Adj R^2, lowest AIC, BIC and Cp

4. Predictors = 4

   Best Model: SDI, ln.Chla, ln.SecDep, ln.Cond (Model 10)

   Criteria: Highest Adj R^2, lowest PRESS,AIC, BIC and Cp

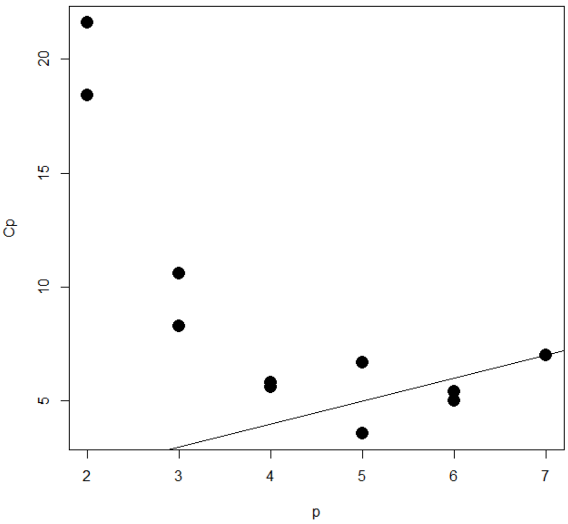
5. Predictors = 5

   Best Model: SDI, ln.Chla, ln.SecDep, ln.Stab, ln.Cond (Model 13)

   Criteria: Highest Adj R^2, lowest PRESS,AIC, BIC and Cp

Q6.

```
# Plot Mallow's Cp with p
        x11()
        plot(models$Vars + 1,
         models$Cp,
           pch = 16, ylab = "Cp", xlab = "p", cex = 2)
        abline(0, 1)
```



Q7.

R output:

```
     Vars   Rsq Rsq(adj) PRESS   AIC   BIC    Cp      s SDI ln.Chla ln.SecDep ln.zmix ln.Stab ln.Cond
4       2 0.568    0.560  48.2 230.0 241.0   8.3 0.6382   0       1         1       0       0       0
10      4 0.599    0.585  46.8 225.2 241.7   3.6 0.6199   1       1         1       0       0       1
```

```
> which.max(models$"Rsq(adj)")
[1] 10
> which.min(models$PRESS)
[1] 10
> which.min(models$AIC)
[1] 10
> which.min(models$BIC)
[1] 4
> which.min(models$s)
[1] 10
```

a. Adjusted R-squared (Adj R^2)

- *Criterion: Choose the model with the highest Adjusted R-squared.*
- *Best Model: SDI, ln.Chla, ln.SecDep, ln.Cond (Model 10)*
- *Adjusted R-squared: 0.585*

b. Predicted Residual Sum of Squares (PRESS)

- *Criterion: Choose the model with the lowest PRESS.*
- *Best Model: SDI, ln.Chla, ln.SecDep, ln.Cond (Model 10)*

- *PRESS: 46.8*

c. Akaike Information Criterion (AIC)

- *Criterion: Choose the model with the lowest AIC*
- *Best Model: SDI, ln.Chla, ln.SecDep, ln.Cond(Model 10)*
- *AIC: 225.2*

d. Bayesian Information Criterion (BIC)

- *Criterion: Choose the model with the lowest BIC.*
- *Best Model: ln.Chla, ln.SecDep (Model 4)*
- *BIC: 241.0*

e. Mallows' Cp (Cp)

- *Criterion: Choose the model with the Cp value closest to the number of predictors plus the intercept.*
- *Best Model: SDI, ln.Chla, ln.SecDep, ln.Cond(Model 10)*
- *Cp: 3.6 (since there are 4 predictors, Cp ≈ 4 is ideal)*

f. Standard Error of the Regression (s)

- *Criterion: Choose the model with the lowest standard error (s).*
- *Best Model: SDI, ln.Chla, ln.SecDep, ln.Cond (Model 10)*
- *Standard Error (s): 0.6199*

Q8.

In question 7 the identified models are:

1. Best Model: SDI, ln.Chla, ln.SecDep, ln.Cond
2. Second Best Model: ln.Chla, ln.SecDep

In question 3 the identified model Through ACI criteria is:

1. SDI, ln.Chla, ln.SecDep, ln.Cond (Same as the above best model)

In question 4 the identified model through BIC criteria is:

1. ln.Chla, ln.SecDep (Same as the above Second Best Model)

Q9.

So based on all the analysis so far (Qs 3, 4, 7 & 8), the model I would choose as the 'best' model would be the 4 variable model SDI, ln.Chla, ln.SecDep, ln.Cond.

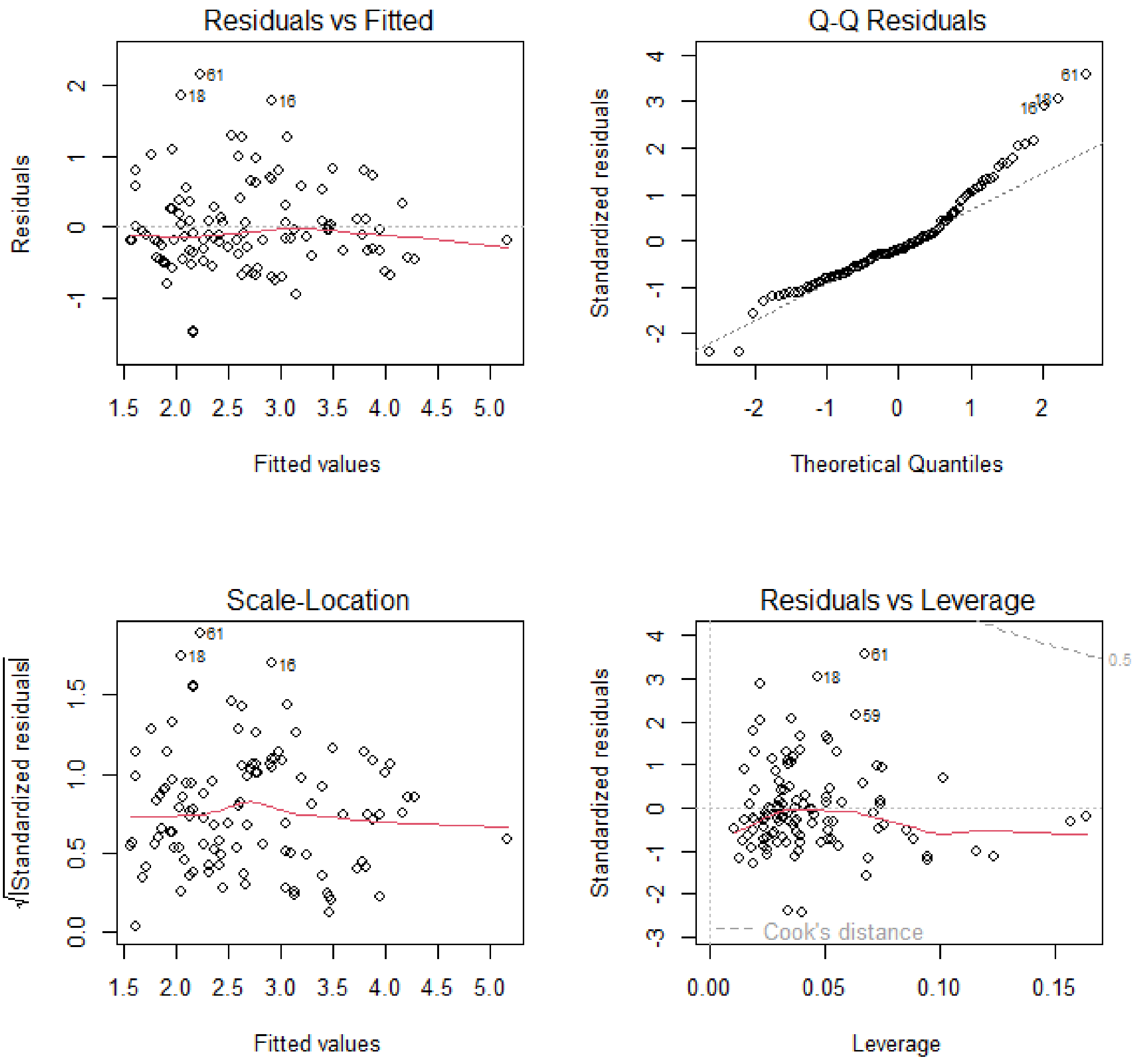
Reasons for chossing this model:

- It has the **highest adjusted R-sq**,
- **lowest AIC**,
- **lowest residual standard error**,
- has a **low C_p** and the

- **best PRESS** value
- The two variable model chosen by BIC only fits one criteria i.e. decrease in complexity of model.

Q10

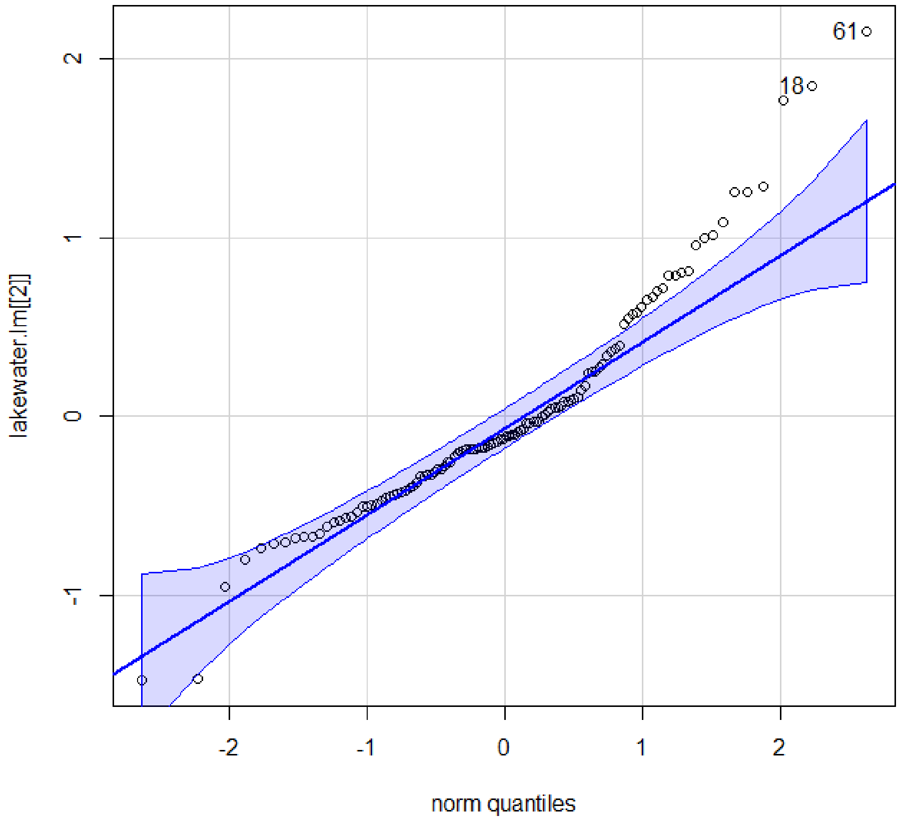
Error Assumptions can be made by looking at the residuals. For checking the assumptions we will create a 4 in 1 residual plot



From the plots we can conclude following error assumptions:

- The relationship between the independent variables and the dependent variable is linear. This condition is satisfied because Residuals vs. Fitted plot is showing no pattern as there is random scatter of points.

- All the errors are independent. This condition is satisfied because there is no pattern left in the errors that we could model with the variables we have identified. In both Residual vs Fitted plot and Scale-Location plot we don't any pattern. This means that the errors are independent.
- The residuals have constant variance across all levels of the independent variables. This condition is also satisfied because when we see the Scale-Location plot (or Residual vs Fitted Plot) both show a somewhat horizontal line with equally spread points.
- The errors come from a common Normal distribution with mean 0 and standard deviation s This condition can be further checked using QQ plot and then Shapiro Test. In Shapiro Test we will test following hypothesis: H0: data is consistent with a Normal distribution, HA: It is not.



```
Shapiro-Wilk normality test

data:  residuals(lakewater.lm)
W = 0.93631, p-value = 3.299e-05
```

- QQPlot of residuals shows that approximately 85% are inside confidence interval In Shapiro test the p-value is very less. Hence we have a strong evidence against the assumption the residuals were sampled from a Normal distribution.

Conclusion: While linear regression is robust to mild deviations from normality, significant non-normality can affect the validity of statistical tests, confidence intervals, and p-values. This is particularly relevant if the sample size is small, as the Central Limit Theorem (which helps in larger samples) might not sufficiently mitigate the impact of non-normality.

Q11.

```
> summary(lakewater.lm)

Call:
lm(formula = ln.TP ~ SDI + ln.Chla + ln.SecDep + ln.Cond, data = lakewater
.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.4726 -0.3962 -0.1206  0.2561  2.1503 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.30152    0.46192   4.983 2.33e-06 ***
SDI          0.27192    0.13490   2.016   0.0462 *  
ln.Chla      0.37119    0.08532   4.350 3.03e-05 ***
ln.SecDep   -0.47262    0.16123  -2.931   0.0041 ** 
ln.Cond      0.07450    0.03625   2.055   0.0422 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6199 on 111 degrees of freedom
Multiple R-squared:  0.5992,	Adjusted R-squared:  0.5847 
F-statistic: 41.48 on 4 and 111 DF,  p-value: < 2.2e-16
```

We will look at T-test of individual predictor

1. SDI: t-value=2.016

   p-value=0.0462(Significant)

   Conclusion: SDI is a statistically significant predictor of ln.TP as p-value is smaller than 0.05 level, suggesting that SDI contributes to the explanation of ln.TP variability.

2. ln.Chla: t-value=4.350

   p-value = 3.03e-05 (highly significant)

   Conclusion: Here the p-value is very small suggesting that there is strong evidence to reject null hypothesis that ln.Chla does not have any significant relationship with ln.Tp. So ln.Chla has a statistically significant relationship with ln.Tp

3. ln.SecDep : t-value= -2.931

   p-value = 0.0041

   Conclusion: low p-value hence significant relationship with ln.Tp
4. ln.Cond : t-value=2.055

   p-value=0.0422

   Conclusion: low p-value hence significant relationship with ln.Tp

F-test for Overall Model Significance

The F-test assesses whether the overall regression model is statistically significant:

F-statistic: 41.48

p-value: < 2.2e-16
Conclusion:
The very small p-value indicates that the overall model is highly significant. This means that at least one of the predictors is significantly related to ln.TP, and the model as a whole provides a good fit to the data.

Summary

- The individual t-tests show that all predictors (SDI, ln.Chla, ln.SecDep, ln.Cond) are significant.
- The F-test indicates that the overall model is highly significant.
- Based on these tests, we can conclude that the model is robust and each predictor contributes significantly to explaining the variation in ln.TP.

Q12.

```
> sort(cooks.distance(lakewater.lm))
          27           85           10           96           89           93            9           50 
1.173151e-08 1.315311e-06 1.016551e-05 1.628403e-05 2.037304e-05 2.167717e-05 2.171404e-05 2.714033e-05 
          43            3           45           26           23           60           30           92 
2.733176e-05 3.627443e-05 3.792734e-05 9.678568e-05 1.048441e-04 1.438395e-04 1.707085e-04 1.787695e-04 
          73           94           91          104           40           33           20           44 
1.797076e-04 1.891755e-04 2.013977e-04 2.063982e-04 2.320967e-04 2.341649e-04 2.702696e-04 2.902307e-04 
          67           68          103           19           57           35           36           55 
2.922400e-04 2.922400e-04 2.951483e-04 3.354469e-04 3.471955e-04 3.860948e-04 3.860948e-04 4.025666e-04 
          32          105           71           39          114           66           78           99 
4.982846e-04 5.954811e-04 6.328933e-04 6.452963e-04 6.830011e-04 6.885283e-04 7.336530e-04 8.075712e-04 
           8           31          109          100            7           21           24          116 
9.948135e-04 1.031992e-03 1.054664e-03 1.073301e-03 1.093003e-03 1.099254e-03 1.224900e-03 1.342987e-03 
           2           54           42           95            6          102           98           63 
1.398044e-03 1.492753e-03 1.628462e-03 1.741934e-03 1.752139e-03 1.958704e-03 2.121629e-03 2.306197e-03 
          28           22          101           86           70          106          111           37 
2.321348e-03 2.359769e-03 2.365728e-03 2.448506e-03 2.633970e-03 2.877382e-03 2.970356e-03 3.093022e-03 
         107            5           79            1          110           90           47           69 
3.168488e-03 3.540656e-03 3.575835e-03 3.939932e-03 3.963498e-03 4.005709e-03 4.209520e-03 4.356251e-03 
          14           13           29           49           65           87          113           34 
4.439819e-03 4.546161e-03 4.733144e-03 5.649010e-03 5.708380e-03 5.917742e-03 6.082907e-03 6.156165e-03 
          17           51           74           41           82           88           84           58 
6.635356e-03 6.757887e-03 7.031482e-03 7.104763e-03 7.144743e-03 8.264168e-03 8.403769e-03 8.434737e-03 
          72           77           64           12           15           11           38           83 
9.451399e-03 9.715912e-03 1.017846e-02 1.031570e-02 1.034283e-02 1.144994e-02 1.194990e-02 1.469626e-02 
          52           25           48           75           53          108           56           80 
1.471034e-02 1.475396e-02 1.867318e-02 1.996846e-02 2.049830e-02 2.251200e-02 2.659331e-02 2.719551e-02 
          46           62            4           76           97           81           16          115 
2.737599e-02 2.916387e-02 3.064316e-02 3.079626e-02 3.591940e-02 3.695609e-02 3.716611e-02 4.143534e-02 
         112           59           18           61 
4.834133e-02 6.219555e-02 9.102073e-02 1.855259e-01 
```

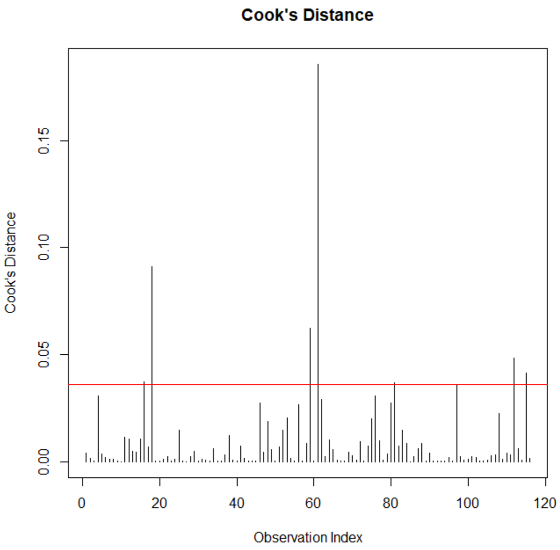
Cook's Distance

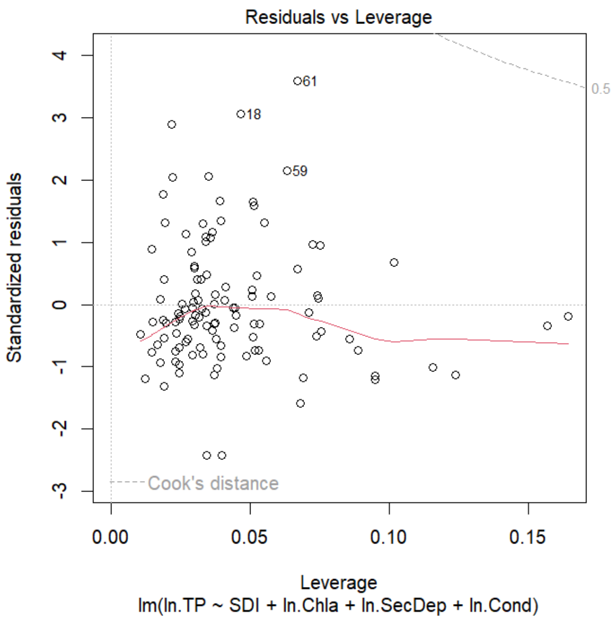
- Cook's distance for an observation is a measure of influence that describes how much the least squares estimates of the regression coefficients change when you delete the ith observation.
- Cook's distance takes into account the correlation structure of the least squares estimates.
- Here point 61 have highest cooks distance followed by point 18.

- These points however lie below the threshold line of 0.5 hence they are moderately influential and may not significantly affect the regression model.

Difference between having high leverage and having high influence.

- High leverage points have extreme predictor values, potentially affecting the regression line's slope, but may not significantly influence parameter estimates**.**
- High influence points, typically with both high leverage and large residuals, exert substantial impact on the model's fit and coefficients.
- While high leverage points can occur without high influence, influential points often have both high leverage and substantial residual deviations, significantly altering the regression model.


Cook's Distance
Cook's Distance
Observation Index


Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
lm(ln.TP ~ SDI + ln.Chla + ln.SecDep + ln.Cond)