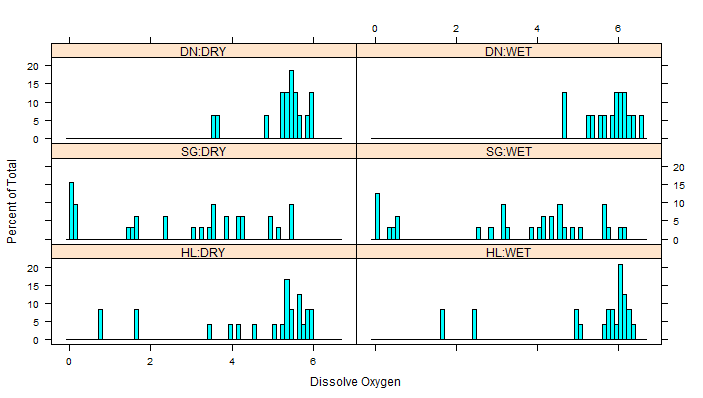
DISSOLVED OXYGEN DATA SET.

* **Fit the data set with a 2 factor model:**

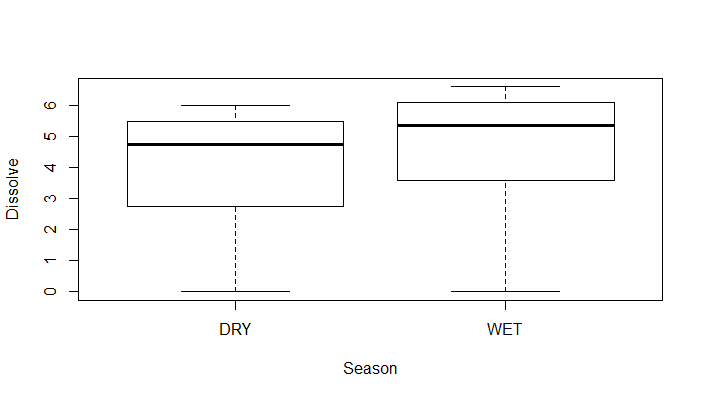
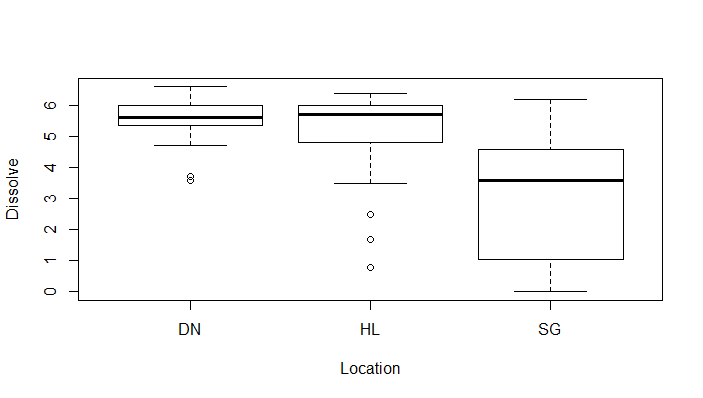
Factor 1: Season, with 2 levels, DRY, and WET

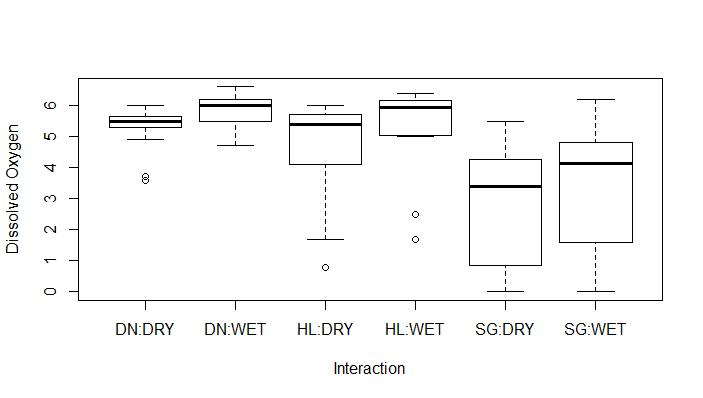
Factor 2: Location, with 3 levels, DN, SG, and HL

Response variable: Dissolved Oxygen (DO)

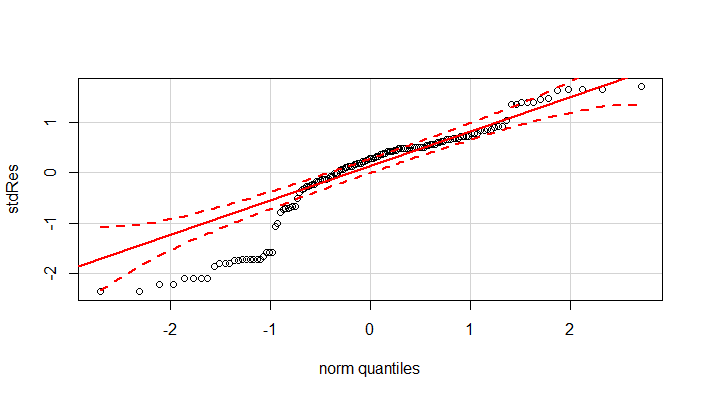
**Histogram of Data, separated by the factors:** 

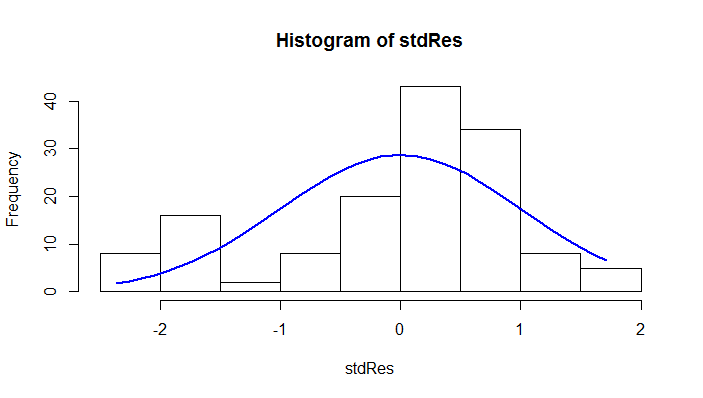
**Boxplots of the data:**





Residual Analysis: To check the assumption of normality, I produce a qq plot. If this assumption is justified, then my points should be in the bandwidth. As can be seen, the plot is highly skewed. The standardized residuals are significantly lower than they should be for the assumption to be justified. Also, according to the analysis of variance table provided below, it suggests that the interaction term is not significant, though the Pvalue is a crude estimate since our data is not following a normal distribution.





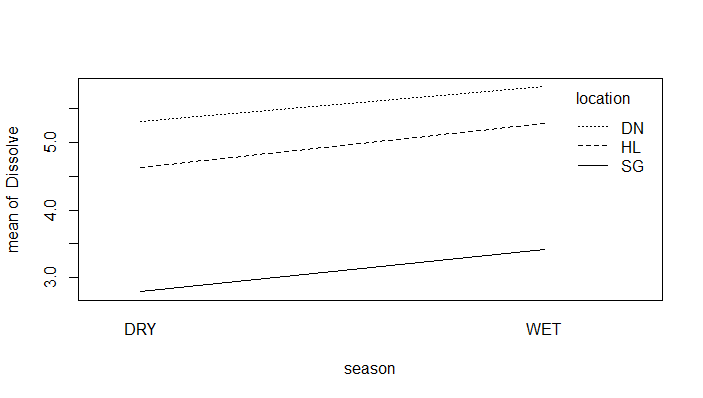
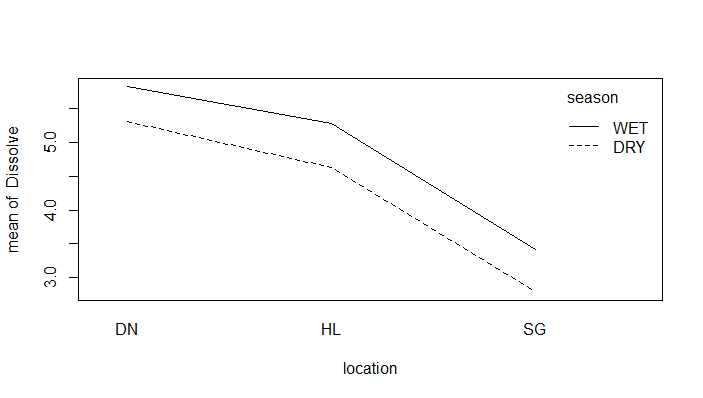
Df Sum Sq Mean Sq F value Pr(>F)

location 2 163.2 81.59 29.841 1.7e-11 \*\*\*

season 1 13.6 13.57 4.962 0.0275 \*

location:season 2 0.1 0.05 0.018 0.9825

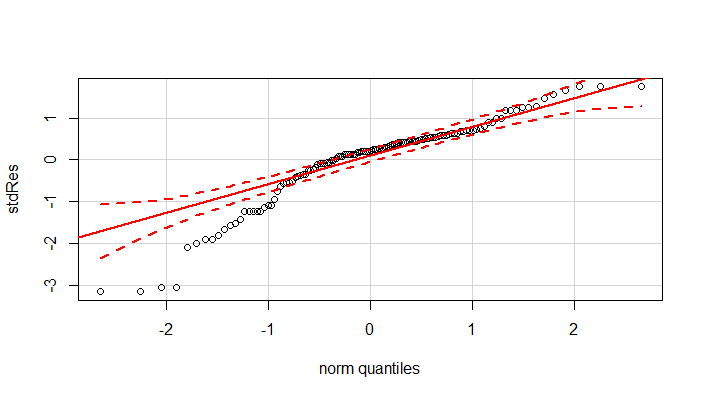
Residuals 138 377.3 2.73

These above interaction plots also show no interation as we move from location to location and from season to

Season. The fact that these lines are parallel confirm that location:season interaction terms are not significant.

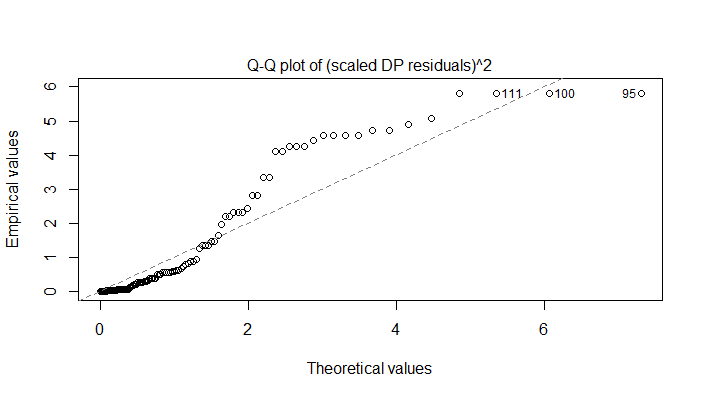
* Similar assumptions as above, **but the “outliers” are taken out of the model** (20 data points: SG(7), SG(9), SG(12), SG(13), SG(23), SG(25), SG(28), SG(29), SG(39), SG(41), SG(44), SG(45), SG(55), SG(57), SG(60), SG(61), HL(10), HL(22),HL(34), HL(46)). These points are taken out with the expectation that rerunning the analysis above, we get a better fit with the qq plot. Below is the plot:



While this does not give a significantly better plot, do note that the points between -2 and -1 (norm quantiles axis) get closer to the 45 degree line, though not close enough for this to be a good stopping point.

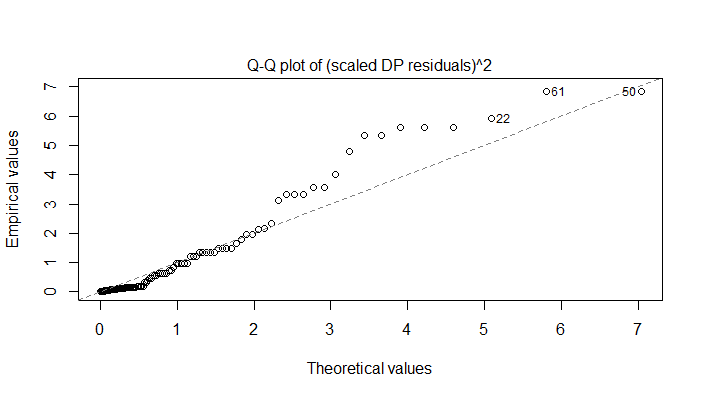
* Since the normality assumption made was not valid, we can implement a generalization of the distribution, called the standard **skew normal distribution, which involves introducing a skew parameter .**

**This is implemented on the full data set of 144 observations.**



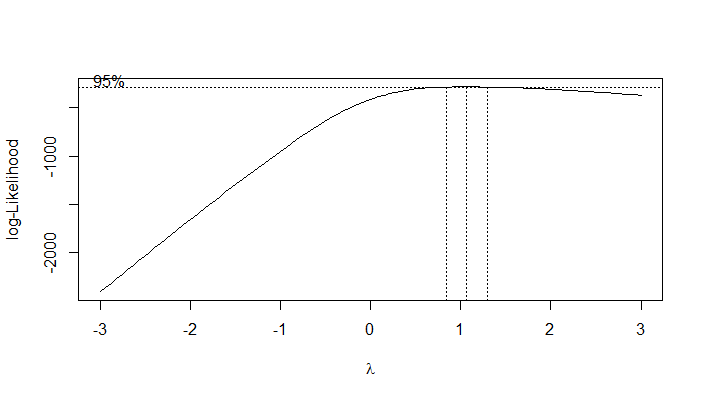
The skewness on the left, has been corrected, but this does not do a good job of estimating the rest of the data. This could be due to the fact that we might have some outliers in the data set, so this is done again, but with those 20 observations taken out, so we have 124 observations.

* **Skew normal distribution with 124 observations (No outliers)**



Similar to when this was fitted with a normal distribution, we see that taking out these observations, does not provide a significantly better plot.

* Next, try to transform the entire data set using the **Box-Cox transformation**. The 8 data points that are 0, are replaced by 0.01 so this transformation can be applied.



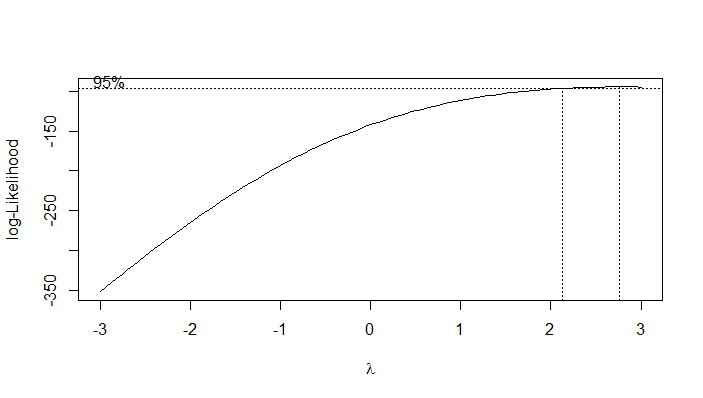
According to the analysis of this plot, the optimum value of lambda = 1.060606, so let lambda = 1, thus no

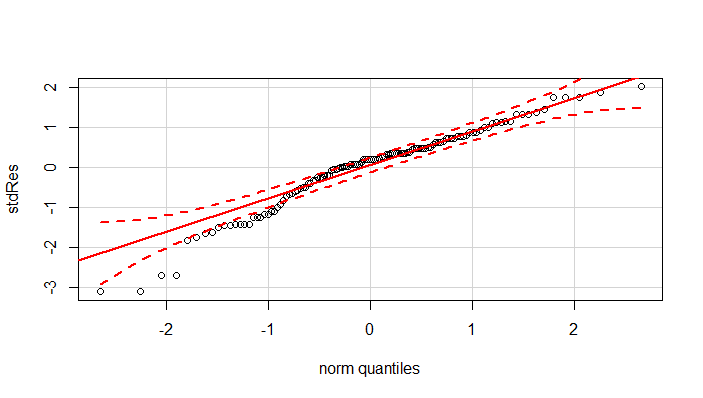
transformation is performed

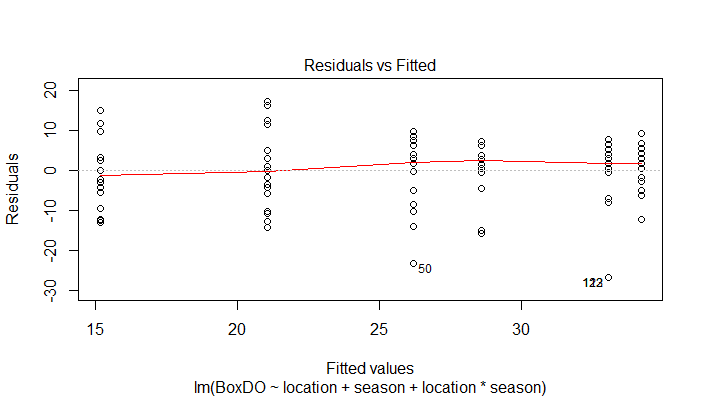
* The next approach, is to try to transform this data using the **Box-Cox transformation on the 124**

**observations**. According to the analysis done, the best transformation is to square the observations.

These squared observations are fitted with a normal distribution. The only reservation with applying this transformation, is the interpretation of the square of dissolved oxygen and its units. Below is the plot obtained:







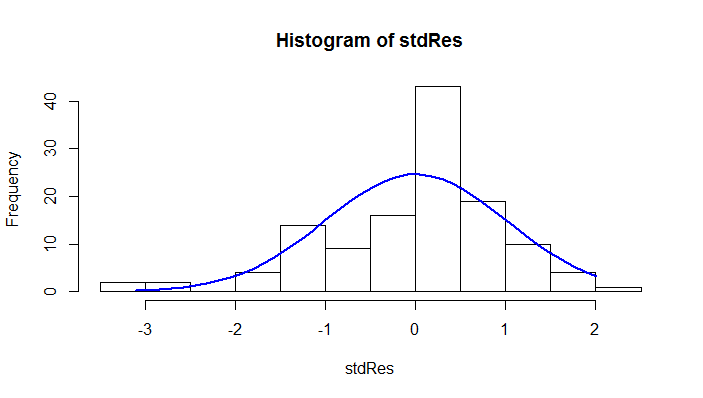
This looks good, especially since the plot of the residuals versus the fitted values give a pretty equal variance which is further supported by Levene’s test on this transformed data set.

Levene's Test for Homogeneity of Variance (center = median)

Df F value Pr(>F)

group 5 0.912 0.4758

118



The square of the observed values almost follows a normal distribution as seen by the residual plot and the histogram of the standardized residuals. Due to the reservation about the transformation, we keep analyzing the data so see if we can get a reasonable distribution without transforming.

* **Fit the data set with a 3 factor model:**

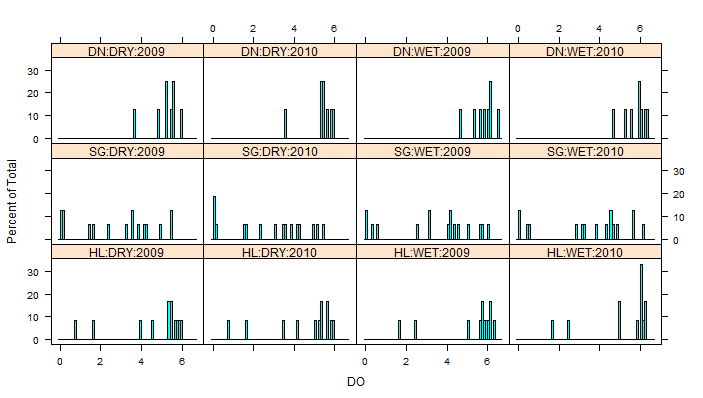
Factor 1: Season, with 2 levels, DRY, and WET

Factor 2: Location, with 3 levels, DN, SG, and HL

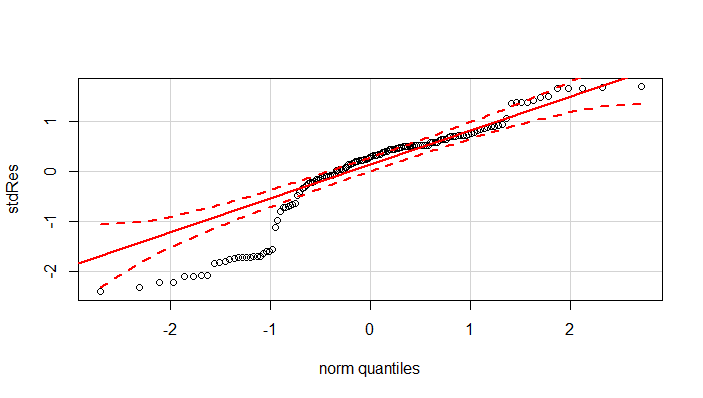
Factor 3: Year, with 2 levels, 2009, and 2010

Response variable: Dissolved Oxygen (DO)

**Histogram of Data, separated by the factors:**



Analyzing the residuals, we get the following plot:



Introducing the year as a factor, does not have a significant effect (if any) on the model.

Df Sum Sq Mean Sq F value Pr(>F)

loc 2 163.2 81.59 28.562 4.93e-11 \*\*\*

seas 1 13.6 13.57 4.749 0.0311 \*

YR 1 0.0 0.00 0.001 0.9764

loc:seas 2 0.1 0.05 0.017 0.9832

loc:YR 2 0.1 0.04 0.013 0.9869

seas:YR 1 0.0 0.02 0.006 0.9372

loc:seas:YR 2 0.1 0.07 0.024 0.9759

Residuals 132 377.1 2.86

* **2 Normal distributions**

We can try to fit this as 2 normal distributions, with majority of the data points being greater than 3. This way, we do not lose any information from excluding “outliers”.

The general idea behind this is to have

We perform the analysis on these sets separately, they are assumed to follow normal distributions, get the residuals, and then combine these residuals to get the residuals for the entire data set.

**Majority of the data points in each category > 3 (112 observations):**

Residuals:

Min 1Q Median 3Q Max

-1.8611 -0.3200 0.0850 0.3563 1.6850

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.30625 0.17384 30.523 < 2e-16 \*\*\*

locHL 0.05486 0.23892 0.230 0.819

locSG -0.97125 0.23324 -4.164 6.38e-05 \*\*\*

seasWET 0.43750 0.24585 1.780 0.078 .

locHL:seasWET 0.16048 0.33058 0.485 0.628

locSG:seasWET -0.25750 0.32984 -0.781 0.437

Residual standard error: 0.6954 on 106 degrees of freedom

Multiple R-squared: 0.4527, Adjusted R-squared: 0.4269

F-statistic: 17.54 on 5 and 106 DF, p-value: 1.266e-12

ANOVA TABLE

Df Sum Sq Mean Sq F value Pr(>F)

loc 2 37.00 18.500 38.259 3.1e-13 \*\*\*

seas 1 4.51 4.512 9.331 0.00285 \*\*

loc:seas 2 0.88 0.442 0.913 0.40436

Residuals 106 51.26 0.484

**The rest of the data points in each category < 3.**

location Diss

1 SG 1.5

2 SG 0.0

3 SG 0.2

4 SG 1.7

5 SG 2.4

6 SG 0.0

7 SG 0.2

8 SG 1.6

9 SG 0.0

10 SG 0.2

11 SG 1.7

12 SG 2.4

13 SG 0.0

14 SG 0.1

15 SG 0.0

16 SG 0.6

17 SG 2.6

18 SG 0.0

19 SG 0.4

20 SG 0.0

21 SG 0.6

22 SG 2.9

23 SG 0.0

24 SG 0.5

25 HL 0.8

26 HL 1.7

27 HL 0.8

28 HL 1.7

29 HL 1.7

30 HL 2.5

31 HL 1.7

32 HL 2.5

Residuals:

Min 1Q Median 3Q Max

-0.8750 -0.8167 -0.2667 0.7937 2.0833

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.6750 0.3255 5.146 1.54e-05 \*\*\*

locationSG -0.8583 0.3759 -2.284 0.0296 \*

Residual standard error: 0.9207 on 30 degrees of freedom

Multiple R-squared: 0.1481, Adjusted R-squared: 0.1197

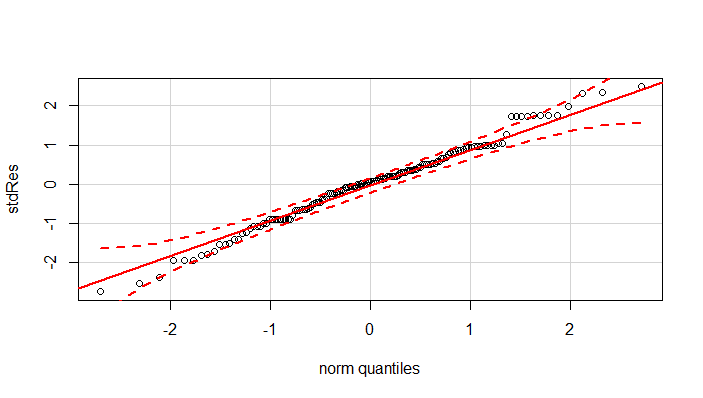
F-statistic: 5.215 on 1 and 30 DF, p-value: 0.02964

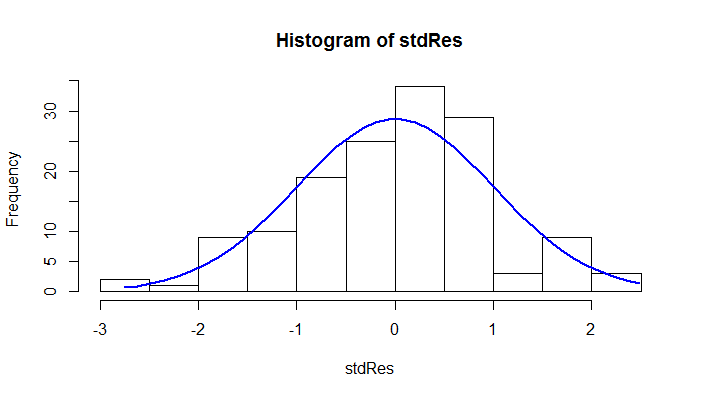
ANOVATABLE

Df Sum Sq Mean Sq F value Pr(>F)

location 1 4.42 4.420 5.215 0.0296 \*

Residuals 30 25.43 0.848

Combining these to get a residual plot for the entire data set, we get: 



This is the best result gotten so far, since most of the points are in that bandwidth, and we were able to use

all 144 observations without transforming the data.

Note : Only the main effects affects the amount of dissolved oxygen according to this analysis, the interactions and the year are not significant factors.