Project 1 - Oskar Andersson.

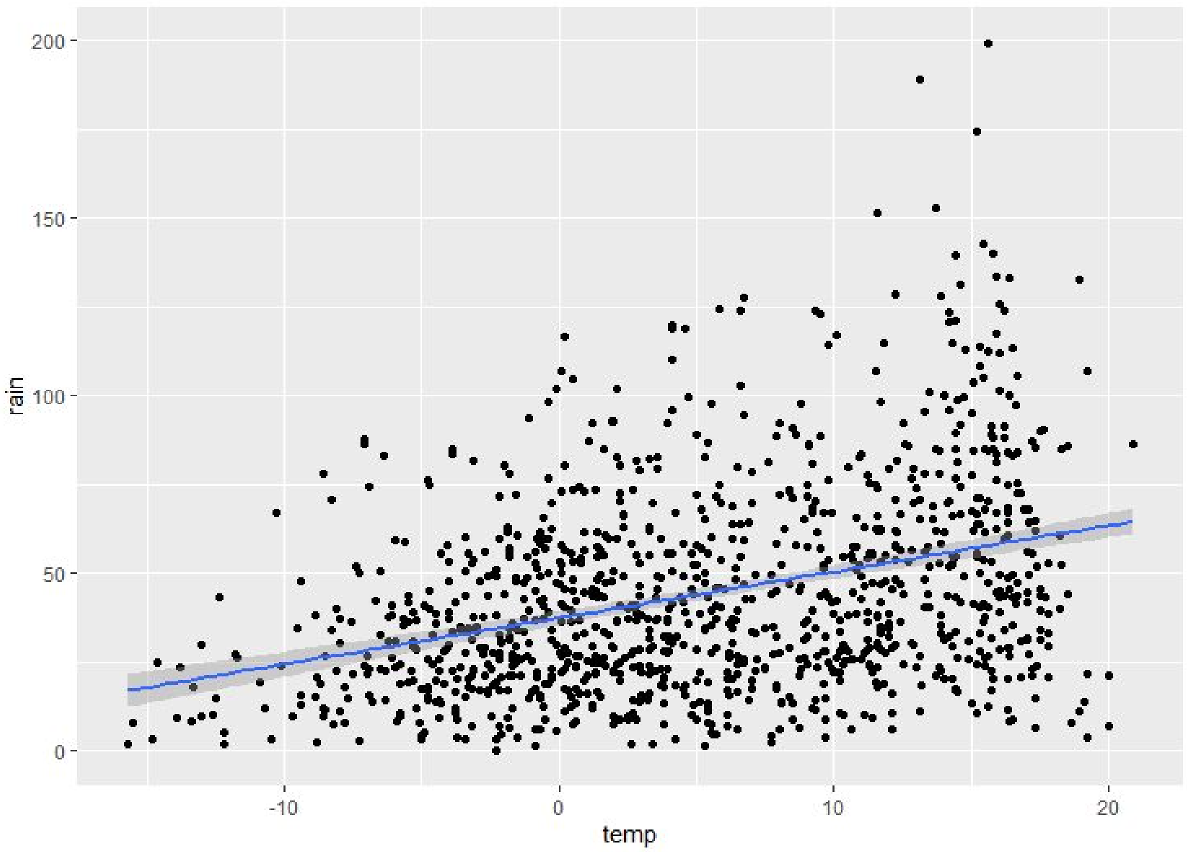
Introduction

This project is about using linear regression to explain precipitation using various explanatory variables such as air pressure, temperature and location. In order to come up with a good model of the data various techniques are used such as variable transformation, removal of outliers and backward/forwards selection.

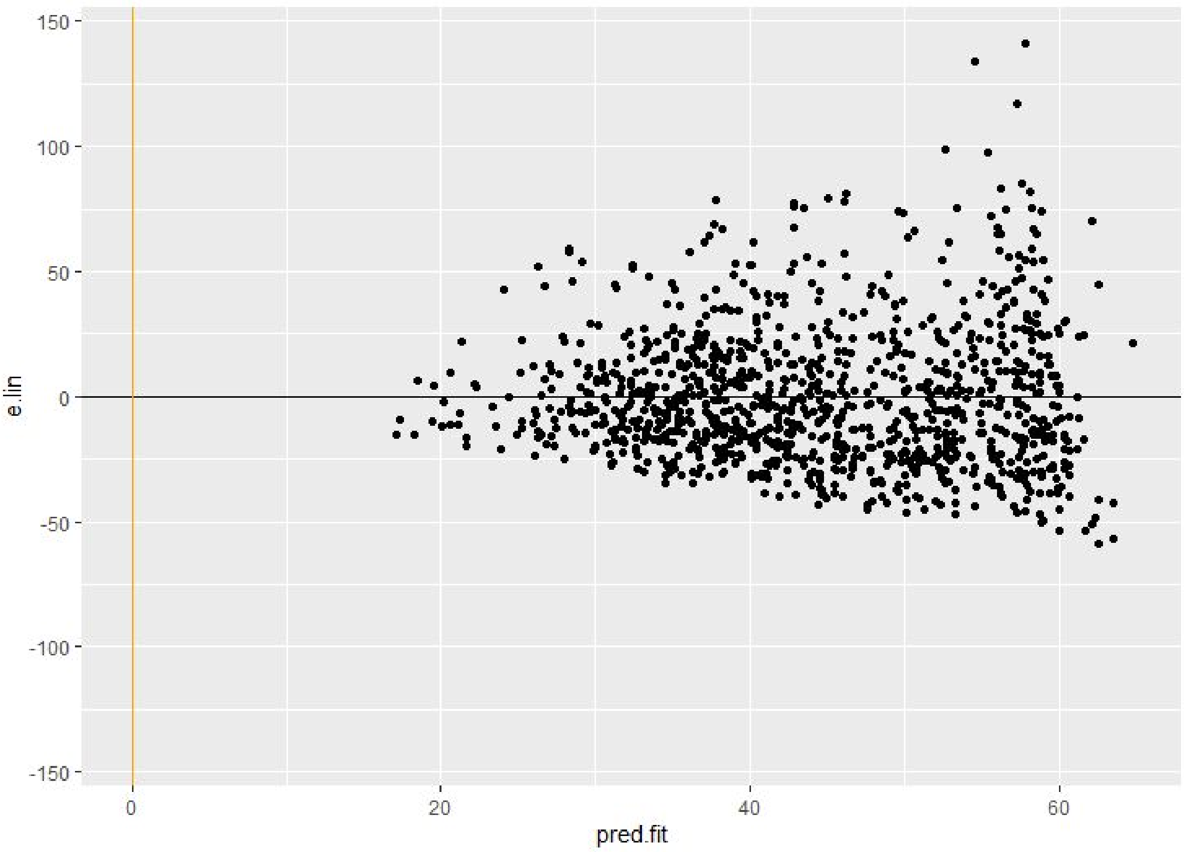
1(a)

As a first attempt at explaining the data I use linear regression without any transformations. One problem that the plot below highlights is that there does not seem to be constant variance, rather variance seems to be increasing with higher temperatures.

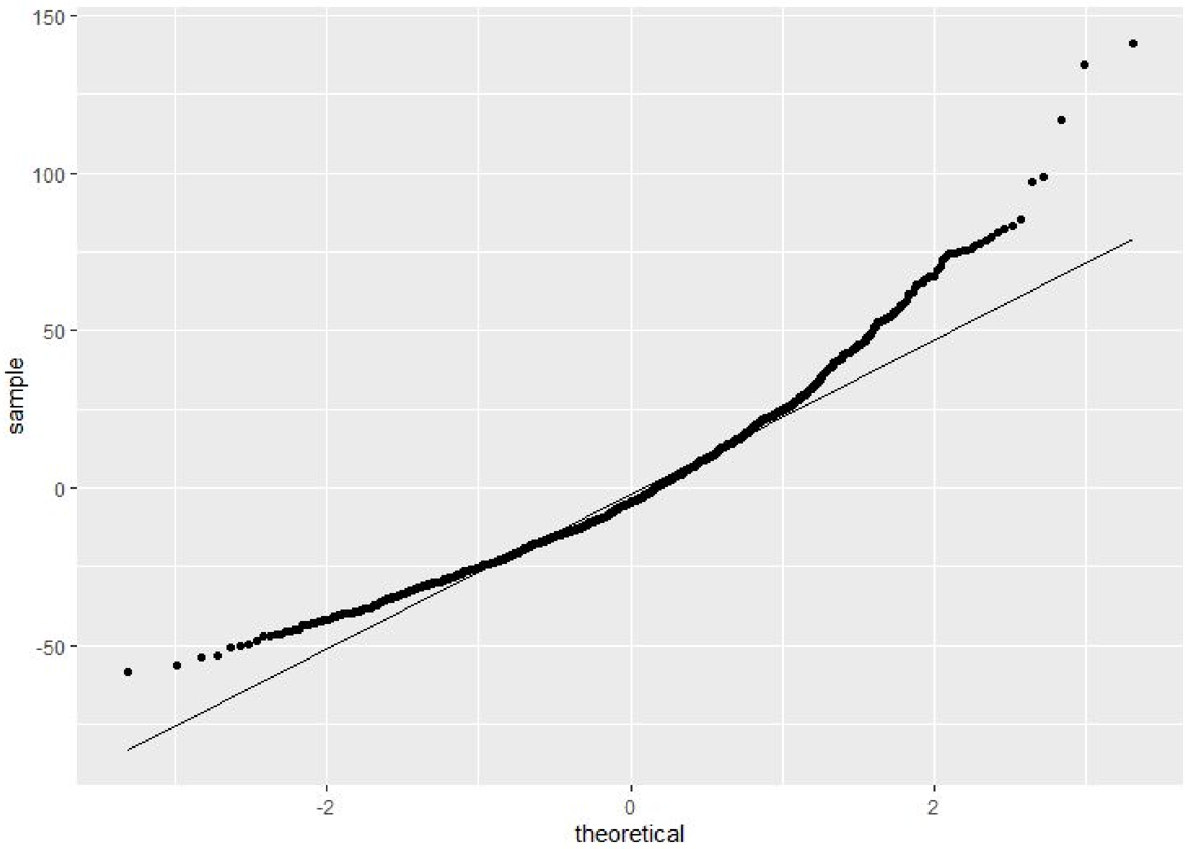
Data plotted with regression line and confidence interval:



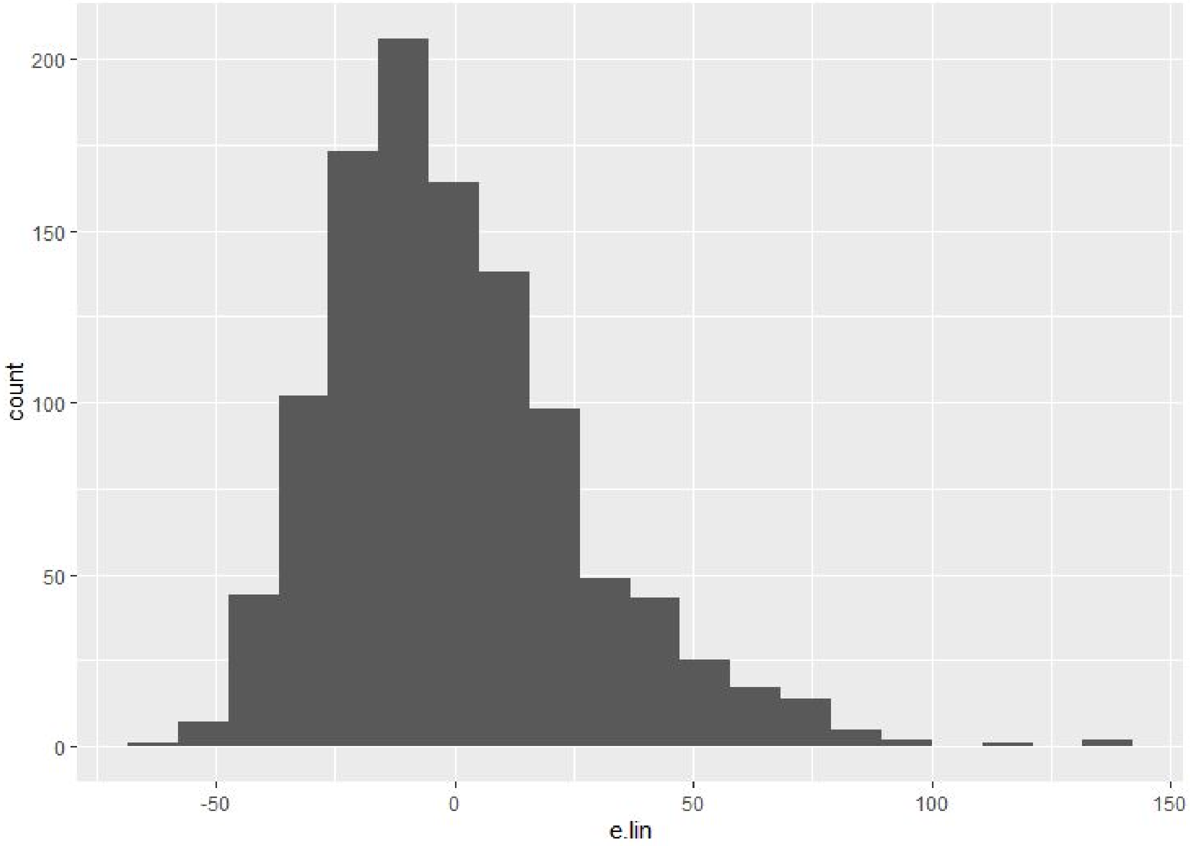
Residuals plotted against yhat.
The heteroscadascity of the data is a bit clearer in this plot.



The QQ-plot below shows that the residuals are not normally distributed but rather right skewed. This can also be observed in the histogram of the residuals.

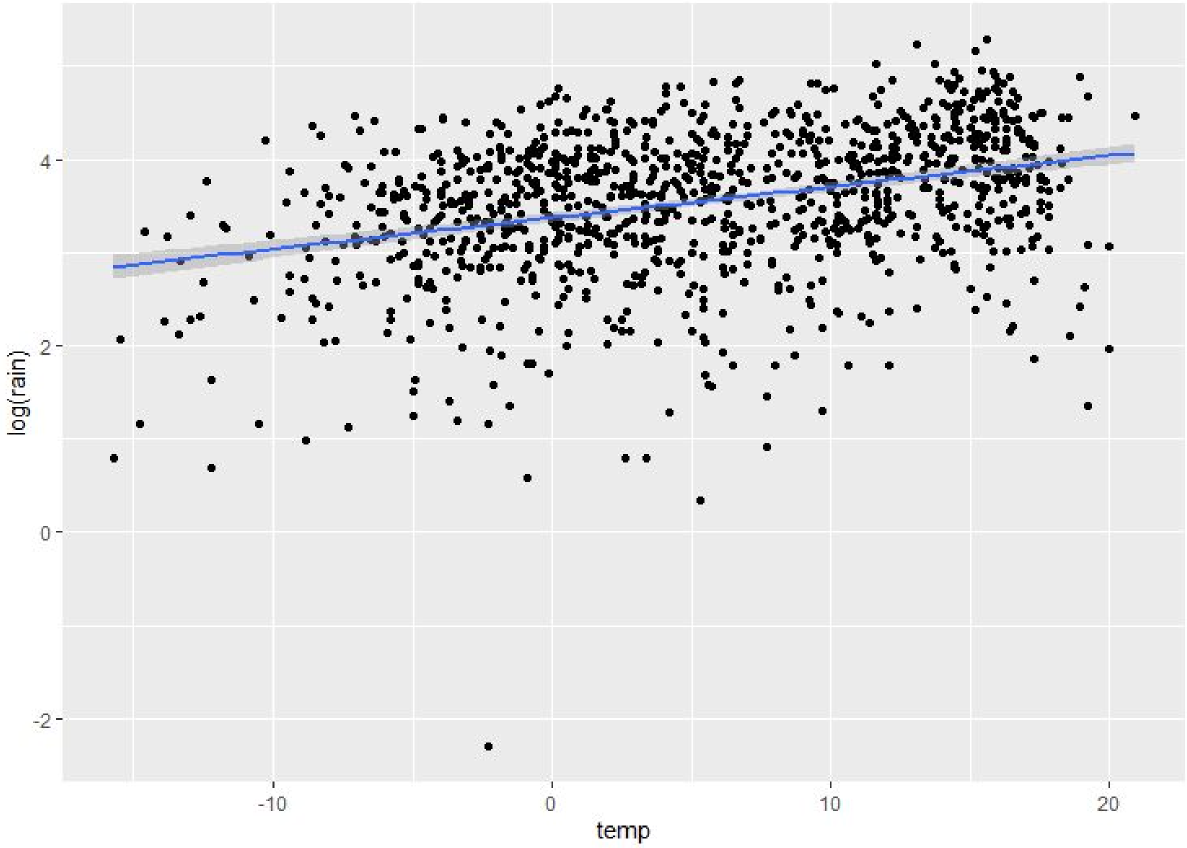


Histogram of residuals:



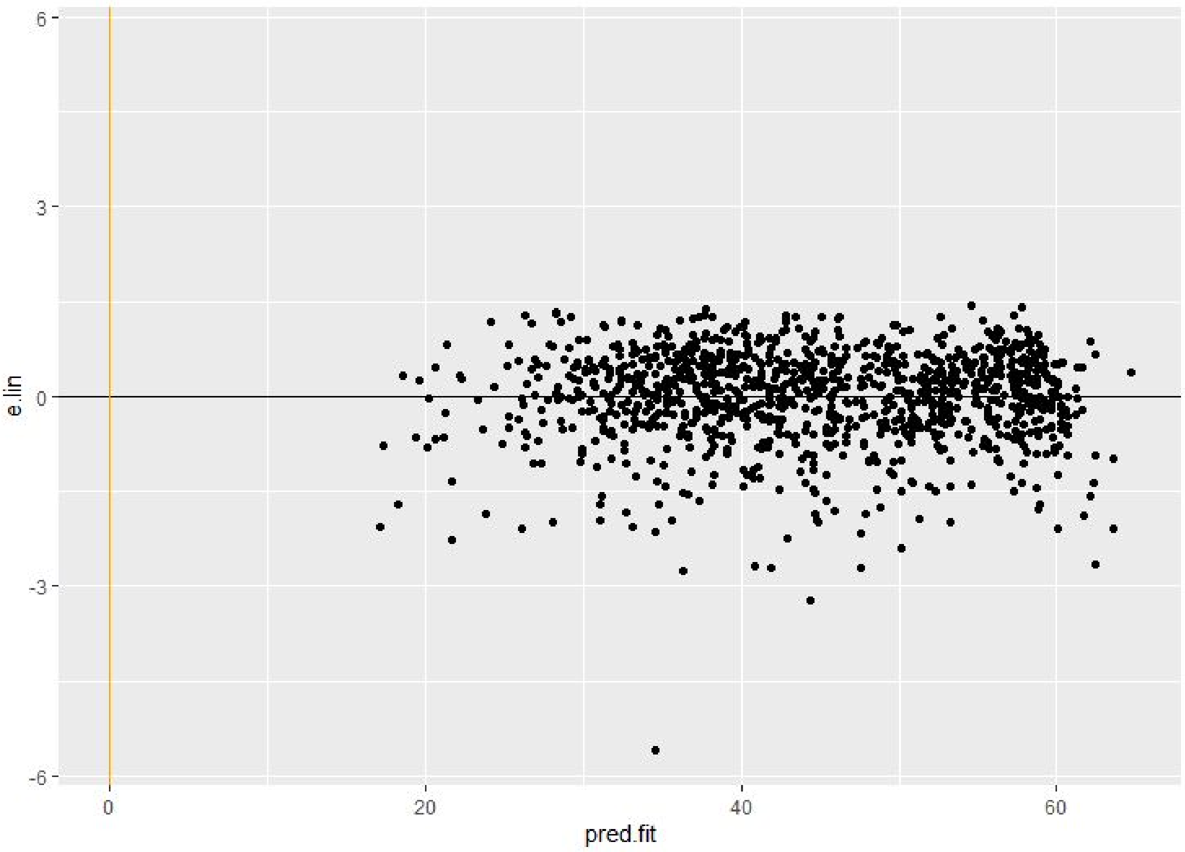
1.(b).

Using log transform on explanatory variable and plot against temperature (with regression line):

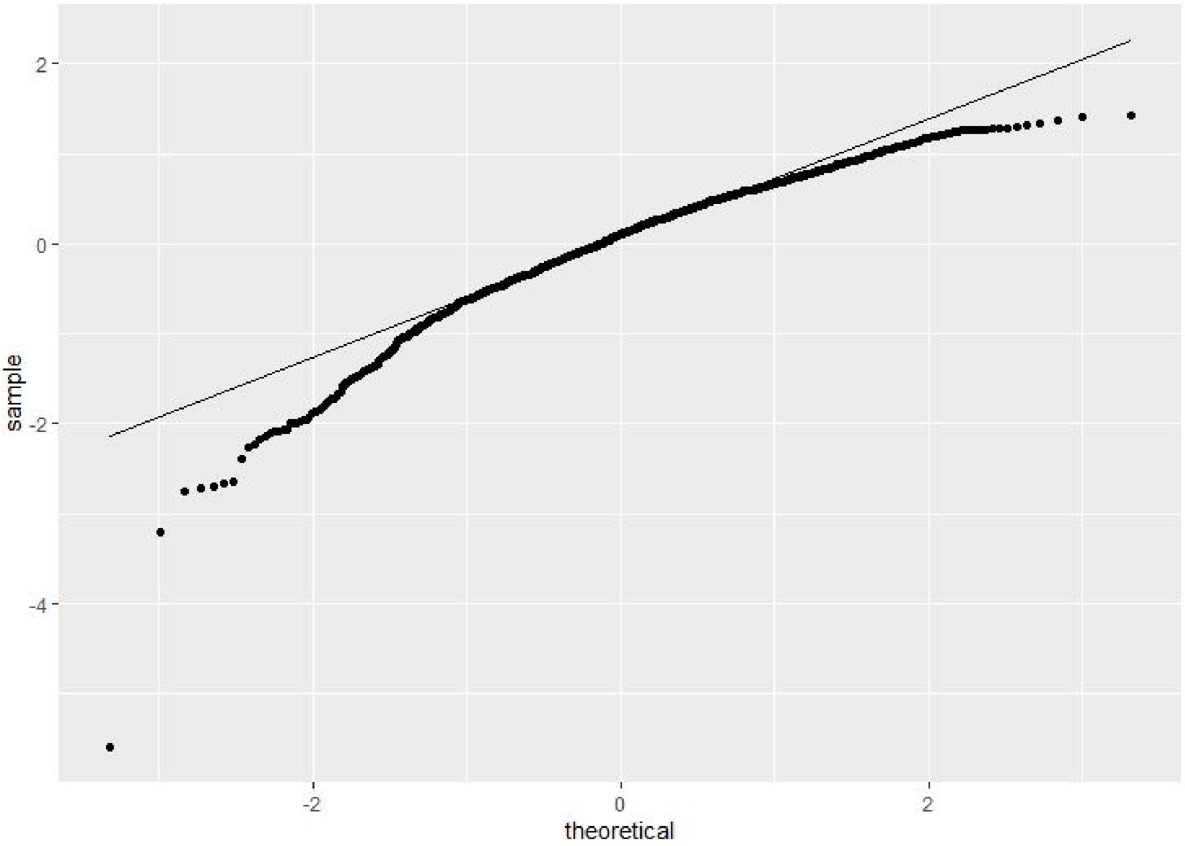


Heteroscedascity is, if not gone, greatly reduced.

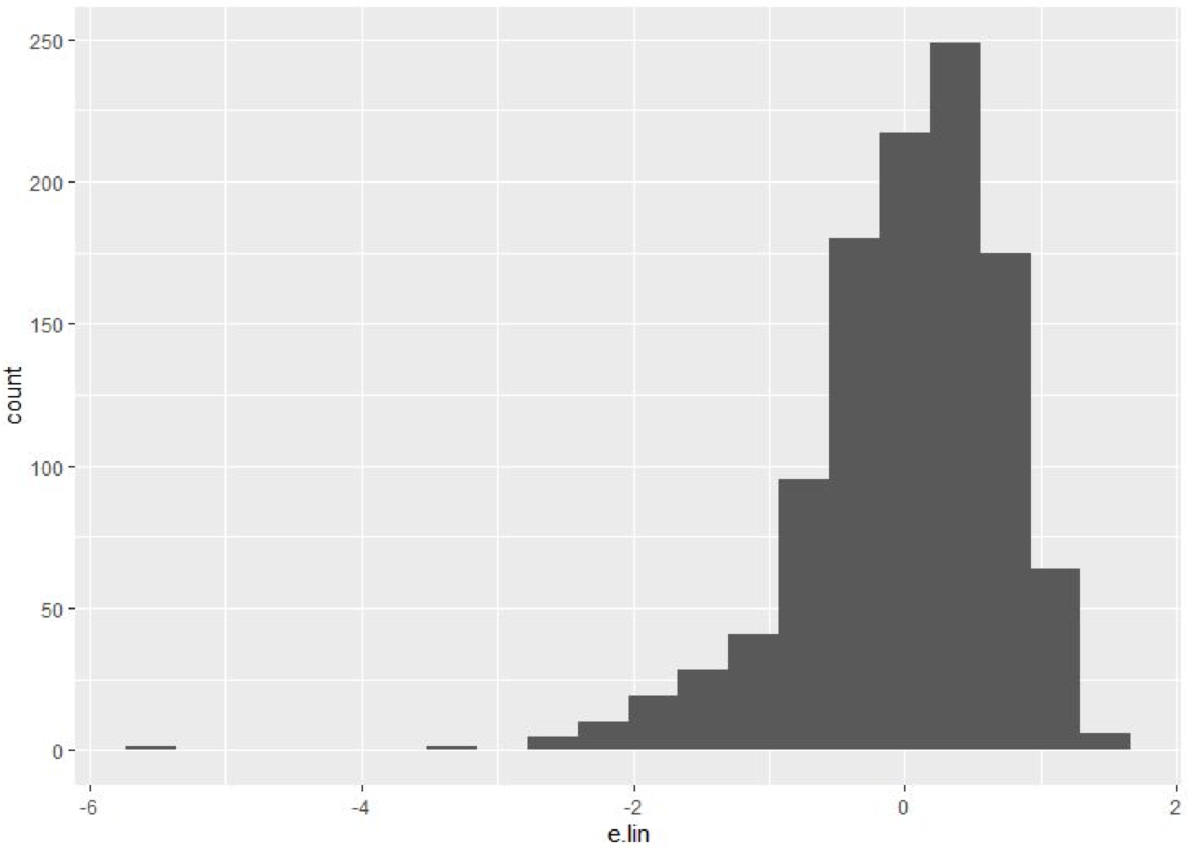
With the transform residuals look more like they are more normally distributed than before.



However, if we look at the qq-plot of the transformed residuals it is clear that they are not normal. The curve shape of the residuals indicate that instead of a heavy right tail we now have a heavy left tail.



The skewedness of the residuals is perhaps better illustrated by the following histogram:



1.C

The model is slightly better with the transformation. The QQ-plots shows that the residuals adhere more to a normal distribution with the transformation. Furthermore, the reduced heteroscedascity of the transformed residuals shows us that using a log-transform is a good idea.

1.D

Our slightly improved model is

$$log(y) = \beta_0 + \beta_1 * x$$

Where y is precipitation and x is temperature.

Or equivalently using the original scale:

$y = C * e^{\beta_1 * x}$

Beta estimates:

β_0: 3.371602
β_1: 0.033374

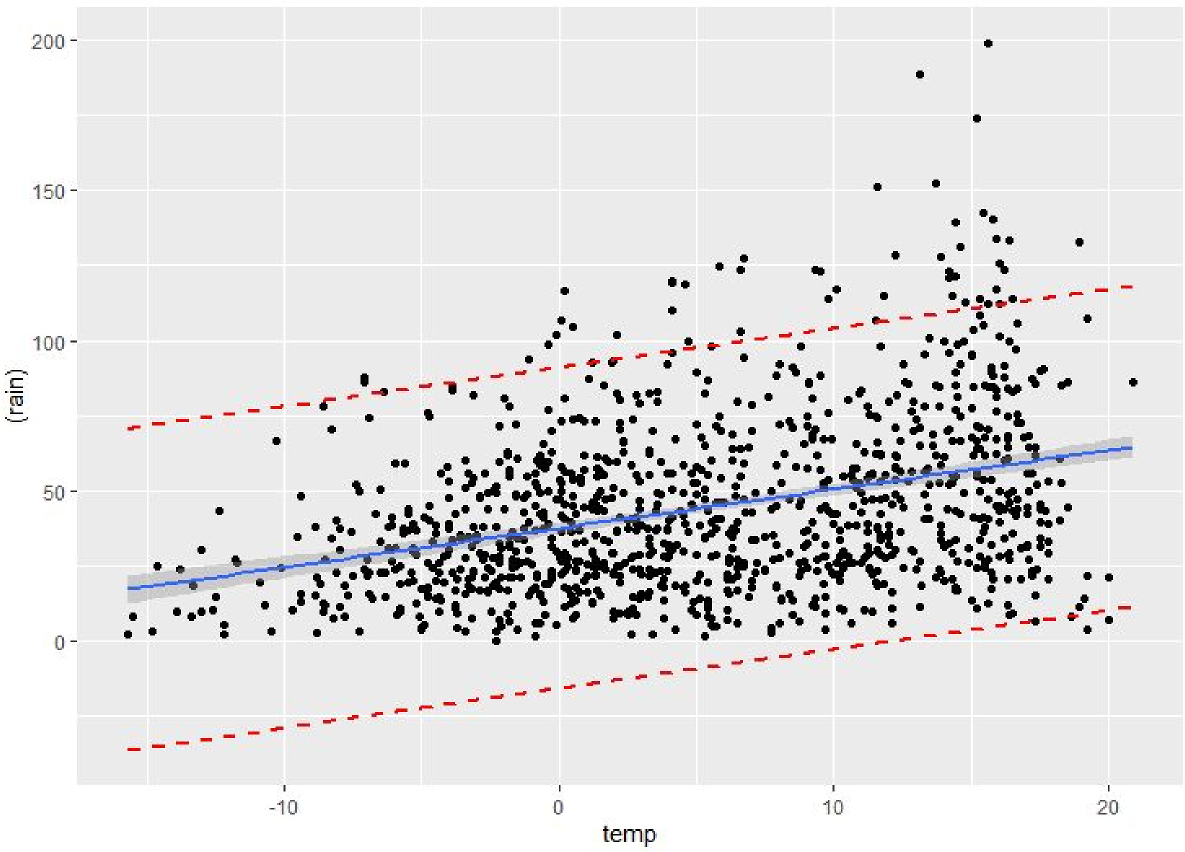
Confidence Intervals:

β_0: (3.31871385 , 3.42448920)
β_1: (0.02774783, 0.03900001)

According to this model, precipitation increases with approximately 0.0334 mm for every centigrade increase in temperature.

1.E

Below is a plot of the model with 95% prediction intervals included. Given that we have a lot of observations it is reasonable that quite a few of them fall outside of the prediction interval. However, some of the higher temperature observations fall way outside of the prediction interval and almost all of the observations outside of the prediction interval are on the “north side” of the prediction interval.



1.F

The model is not perfect but has some use. For example we can come up with a prediction of the amount of precipitation for a month where the average temperature is 5 degrees centigrade:

3.371602 + 5*0.02774783 = 3.51034115

3.371602 + 5*0.03900001 = 3.56660205

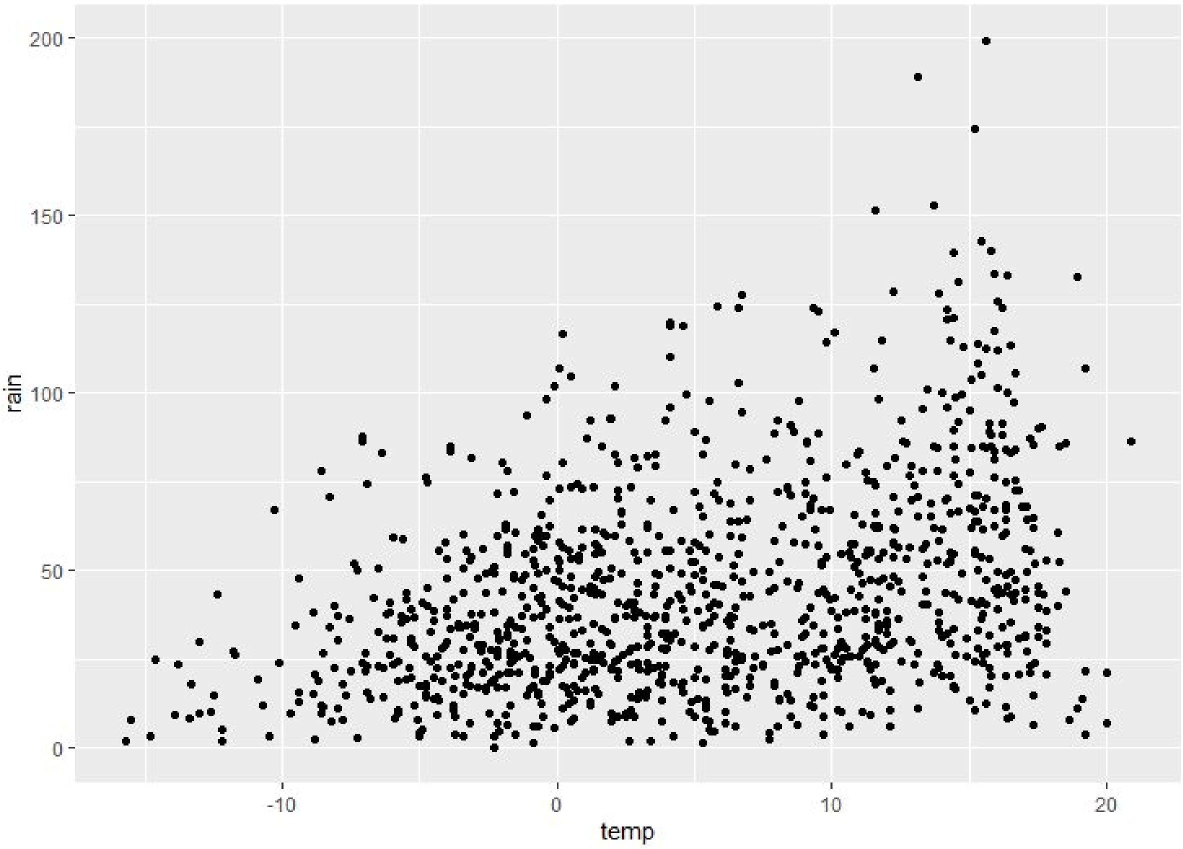
2.1.

2.A

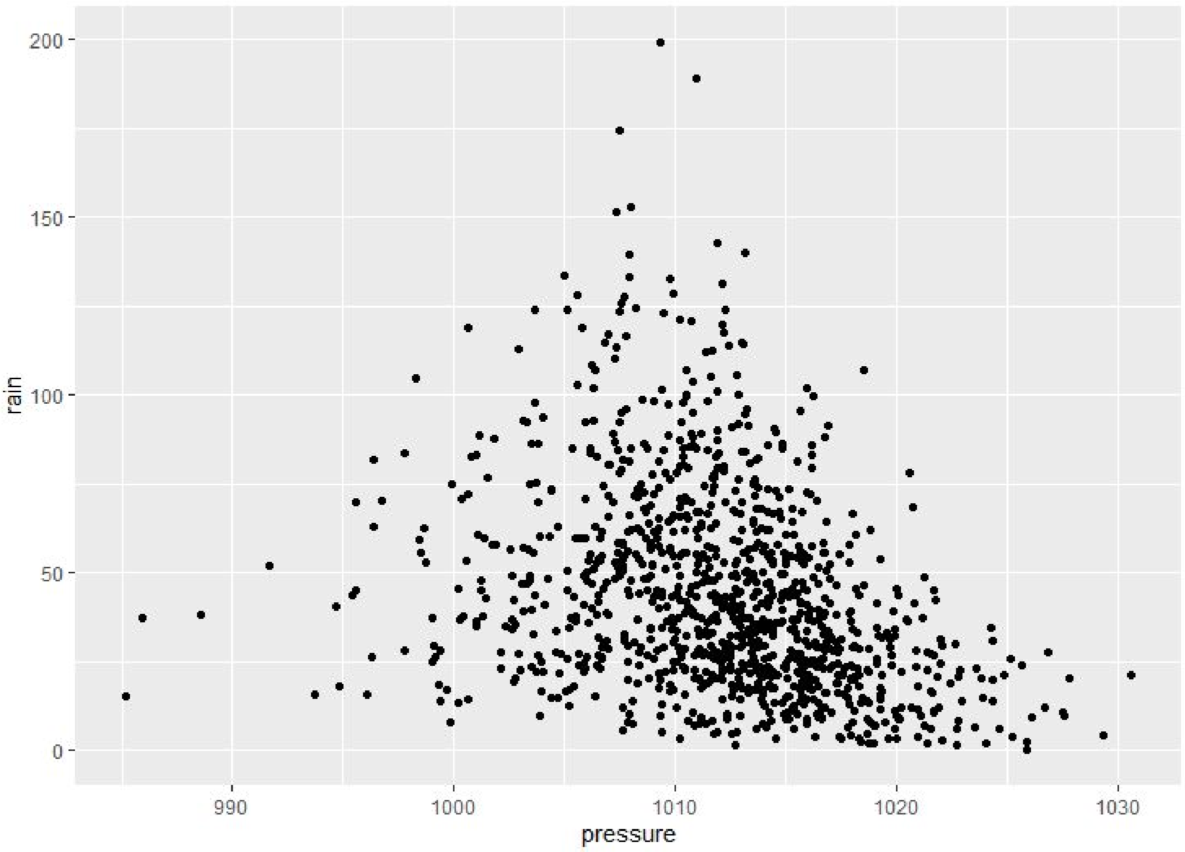
A very important fact to establish is whether or not the results of our model are statistically significant. Using a t-test on the β_1 coefficient shows that the likelihood of it being equal to zero is less than 0.1%. i.e. it is very likely that temperature has an impact on precipitation.

2.2

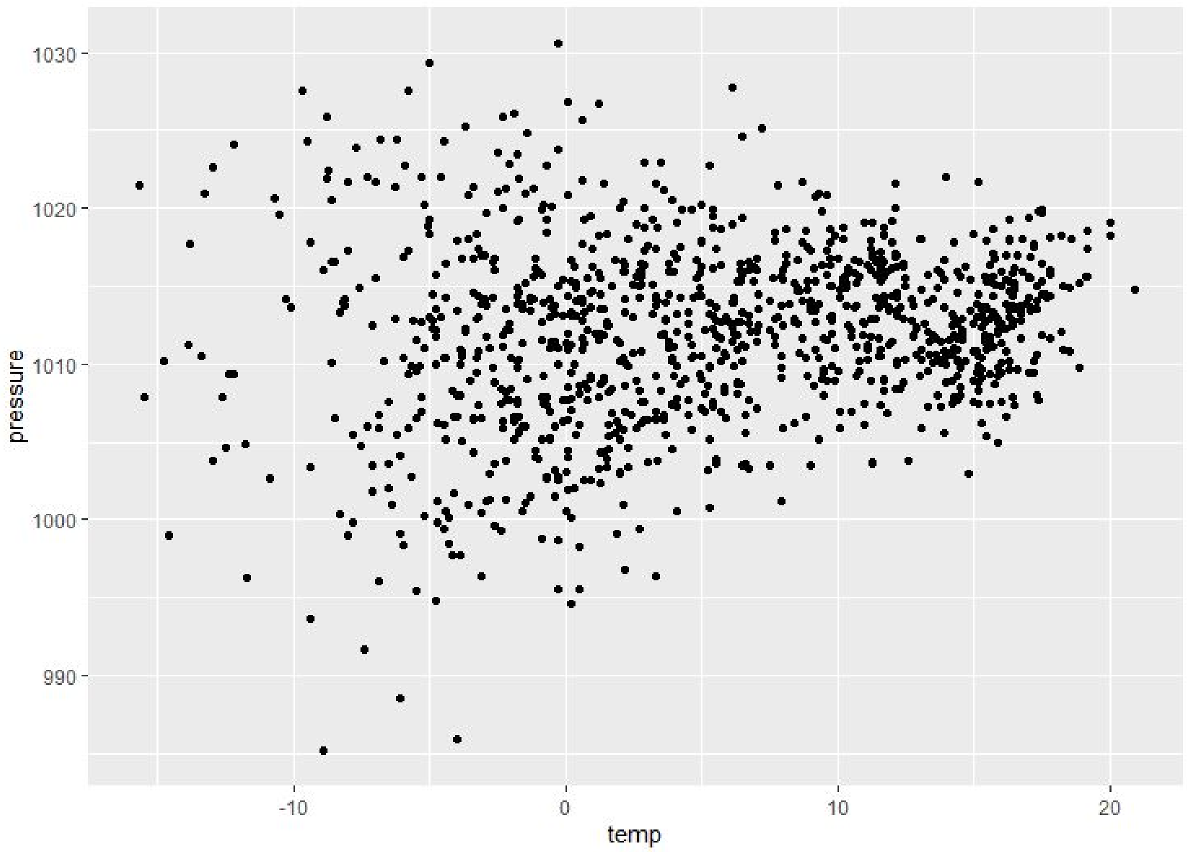
2.B



It looks like there could be a negative linear relationship between pressure and precipitation but it is hard to tell.



There might be a relationship between pressure and temperature. Not certain but not completely random either.



2.C

It is reasonable to think that precipitation depends not only on pressure but on temperature also. When fitting a linear model with both temperature and pressure as explanatory variables it turns out that inclusion of temperature is significant on a 0.001 level. The coefficient estimates and their confidence intervals are as follow.

Beta estimates:

β_0: 61.902608
β_1: 0.040443
β_2: -0.057873

Confidence Intervals:

β_0: (54.92106690 , 68.88414831)

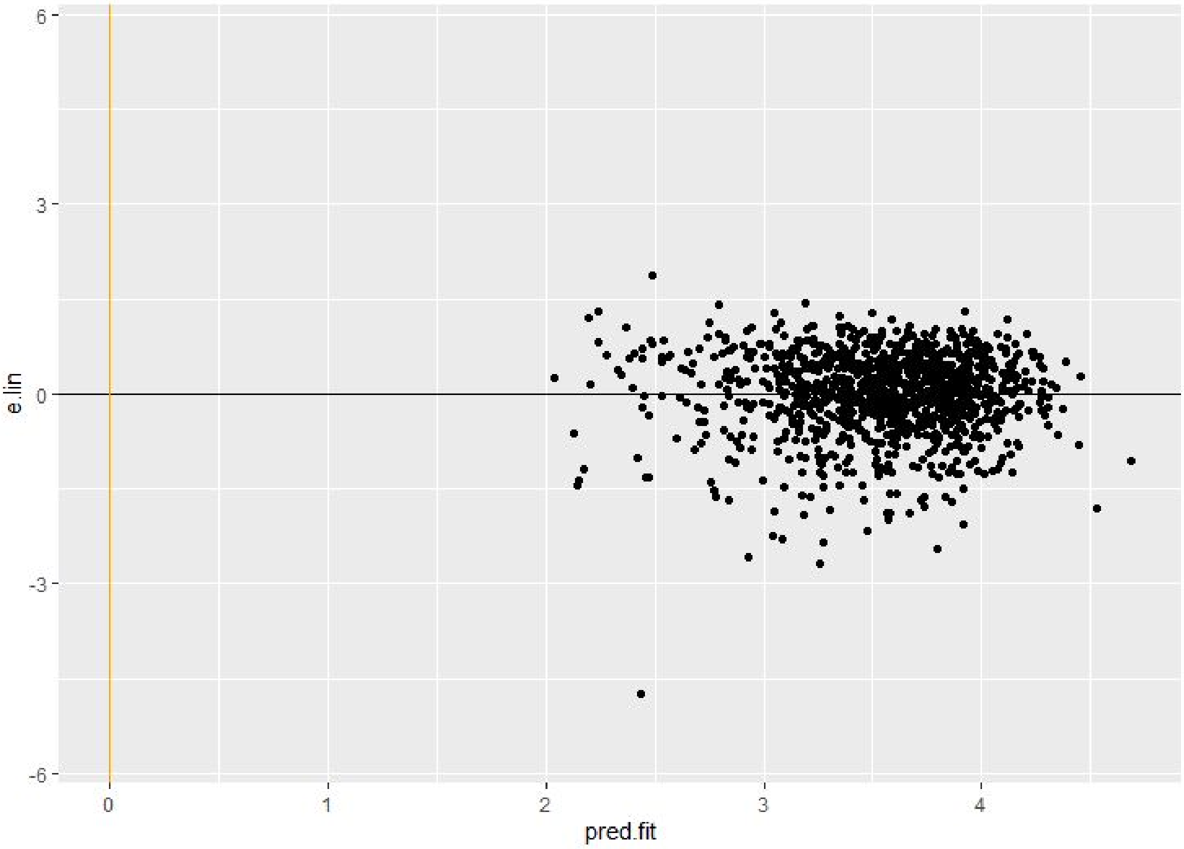
β_1: (0.03533587 , 0.04555006)

β_2: (-0.06477596 , -0.05097015)

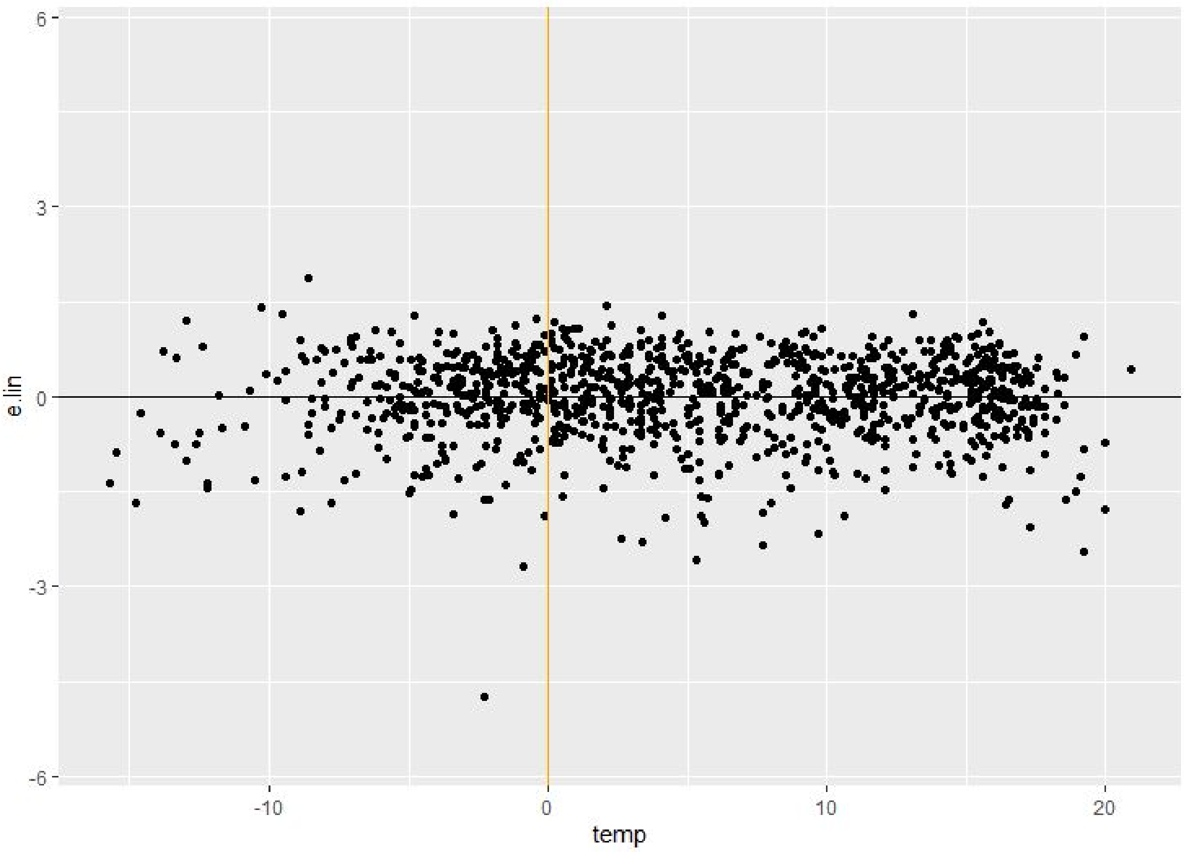
2.D

In order to determine if the data obeys the assumption of normalcy I investigate the residuals. These should be normally distributed.

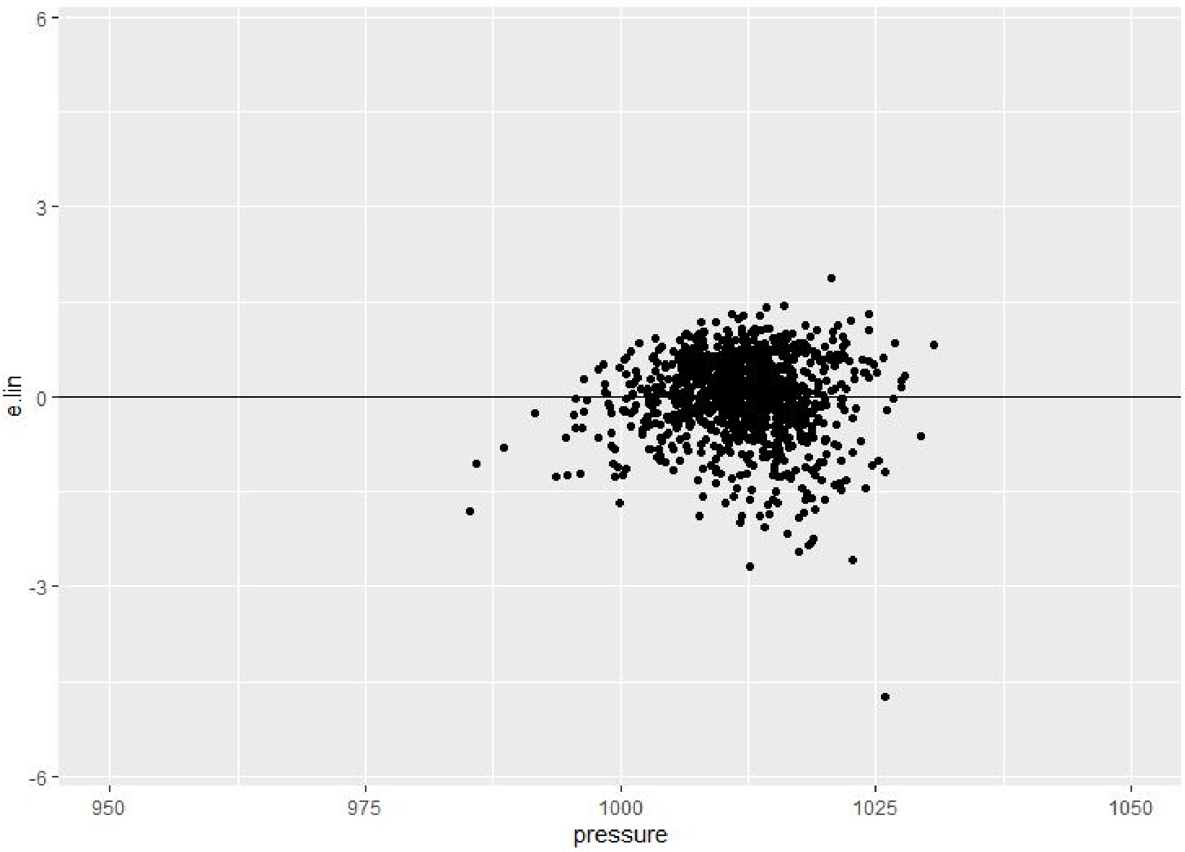
Against fitted values:



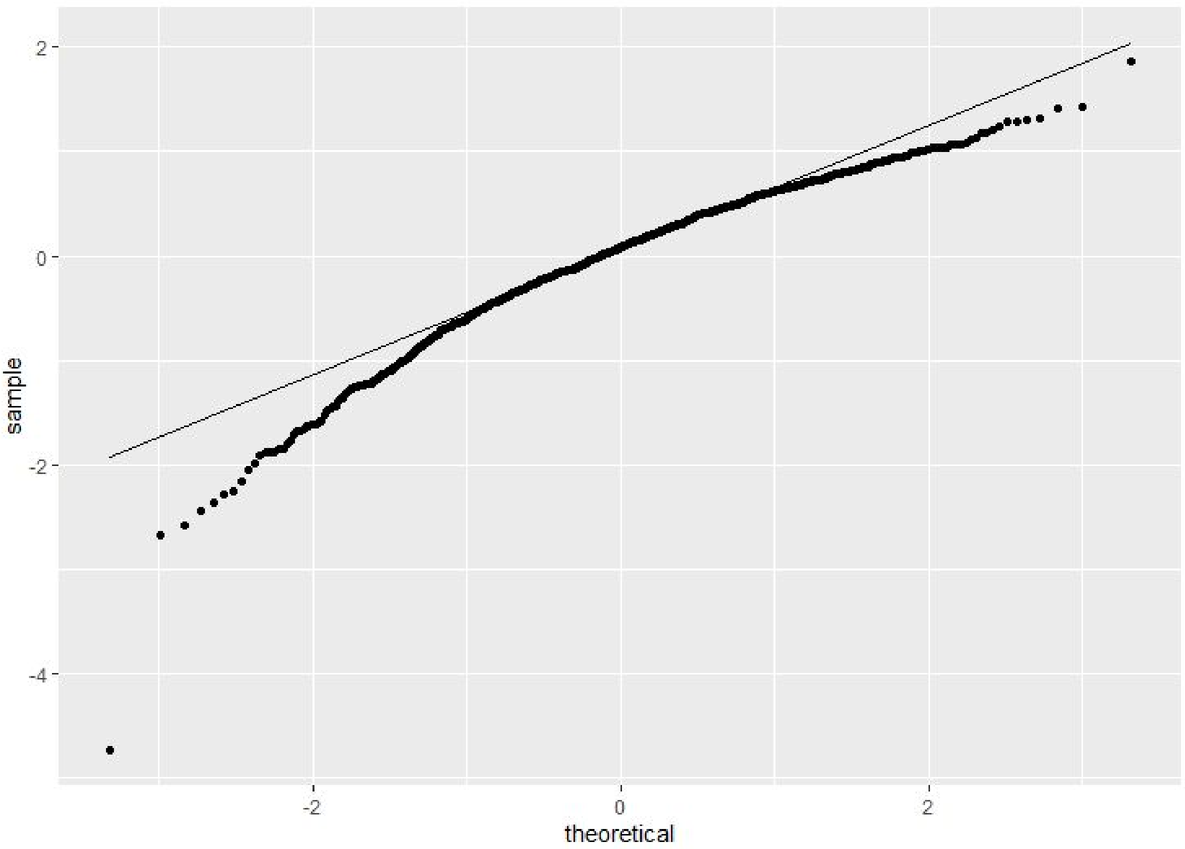
Against Temperature:



Against air pressure:



QQ-plot:



The residuals are closer to being normally distributed than before but the QQ-plot quite clearly shows that they are not perfectly normal.

2.E

Log of the precipitation changes by 0.040443 with a one degree temperature change. This is quite different from the previous model where it changed by 0.033374.

2.F

A 20 hPa increase would lead to a -1.15746 mm change in precipitation.

2.G

—-

2.3.

2.H

Beta estimates:

β_0: 61.0413685
β_1: 3.2717976
β_2: -0.0570093
β_3: -0.0031910

Confidence Intervals:

β_0:   54.192433567 67.890303441

β_1:   2.327347658 4.216247530

β_2:   0.063781200 -0.050237355

β_3:   -0.004123633 -0.002258353

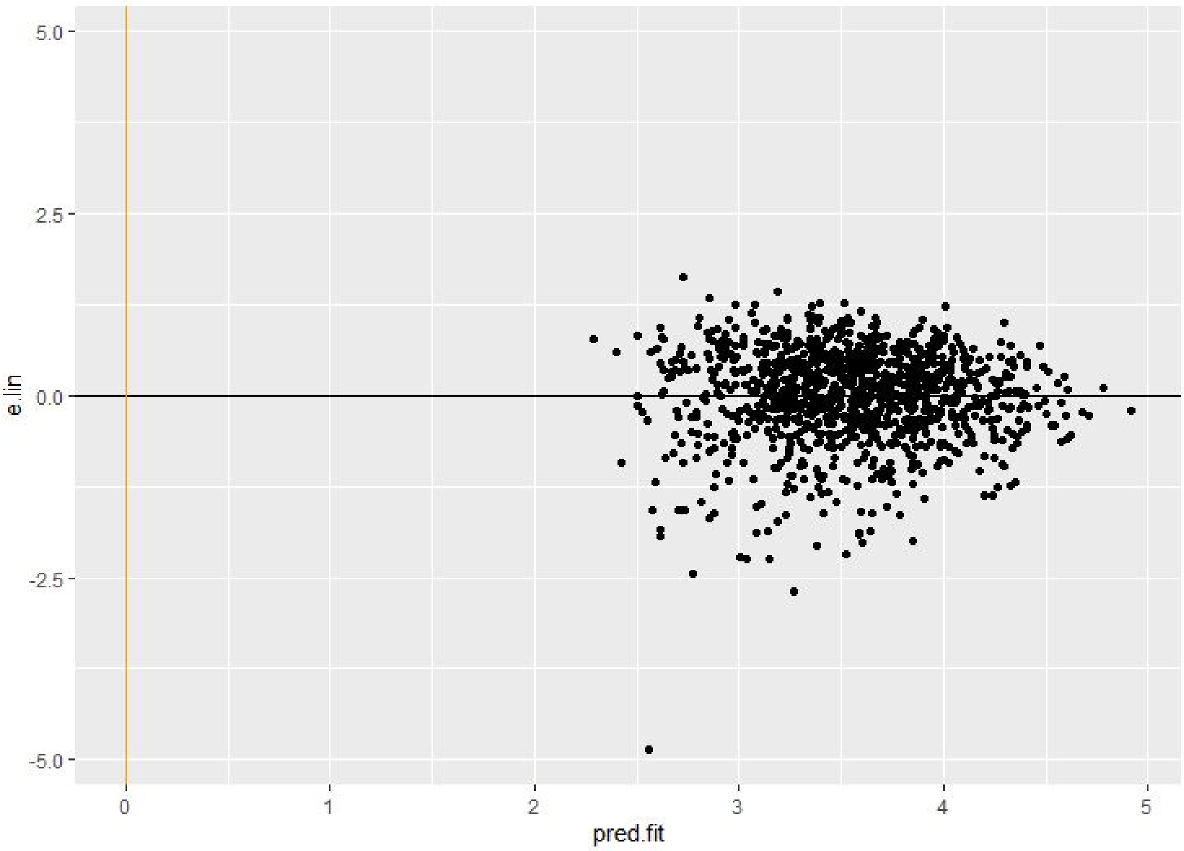
The interaction term is statistically significant.

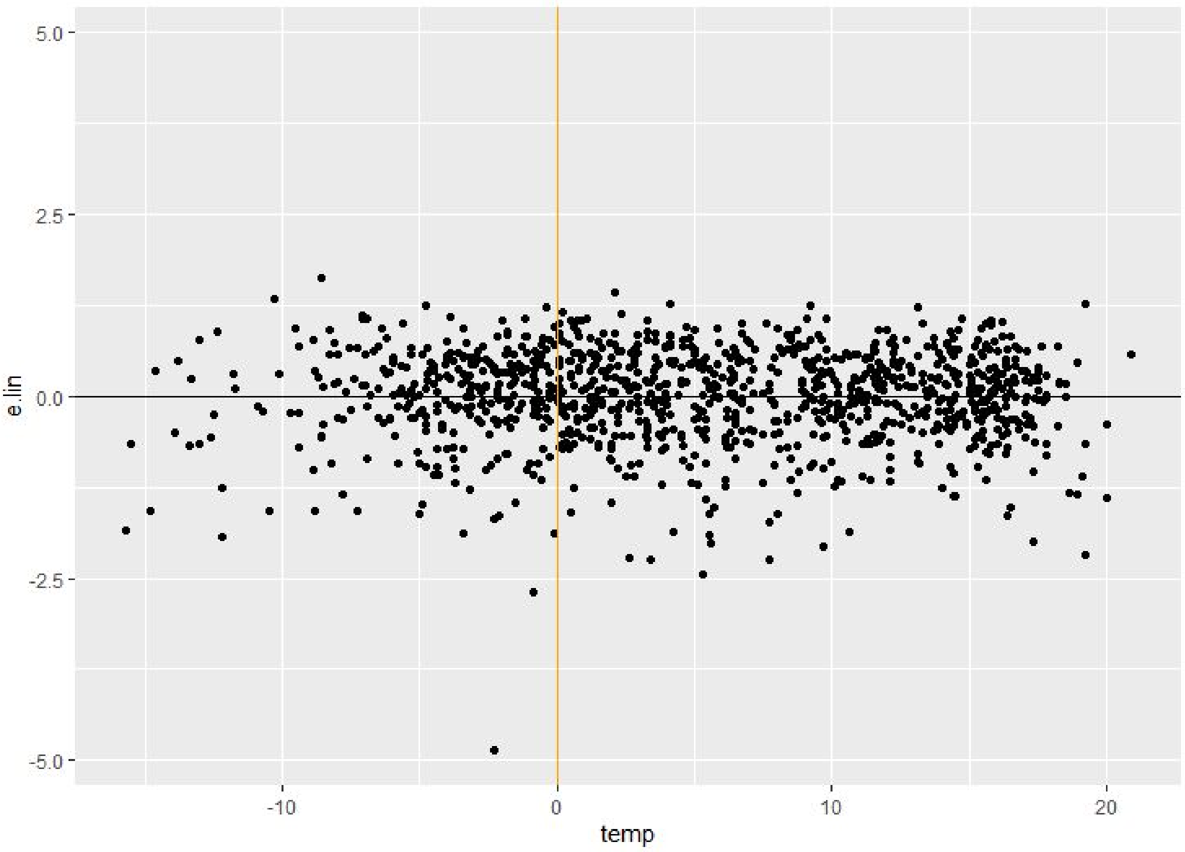
2.I

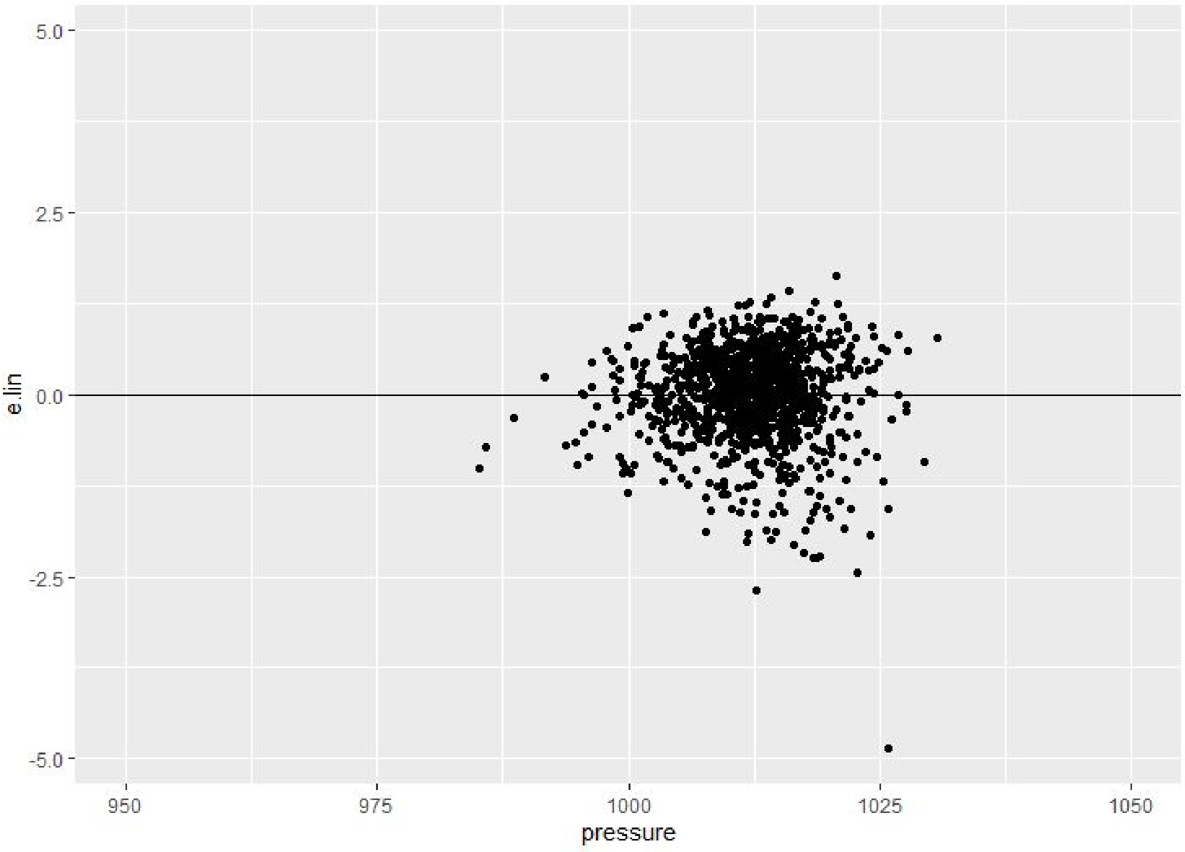
```
Call:
lm(formula = log(rain) ~ temp * I(pressure - 1012), data = weather)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.8632 -0.3287  0.0856  0.4500  1.6308 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)            3.3479801  0.0238350 140.465  < 2e-16 ***
```

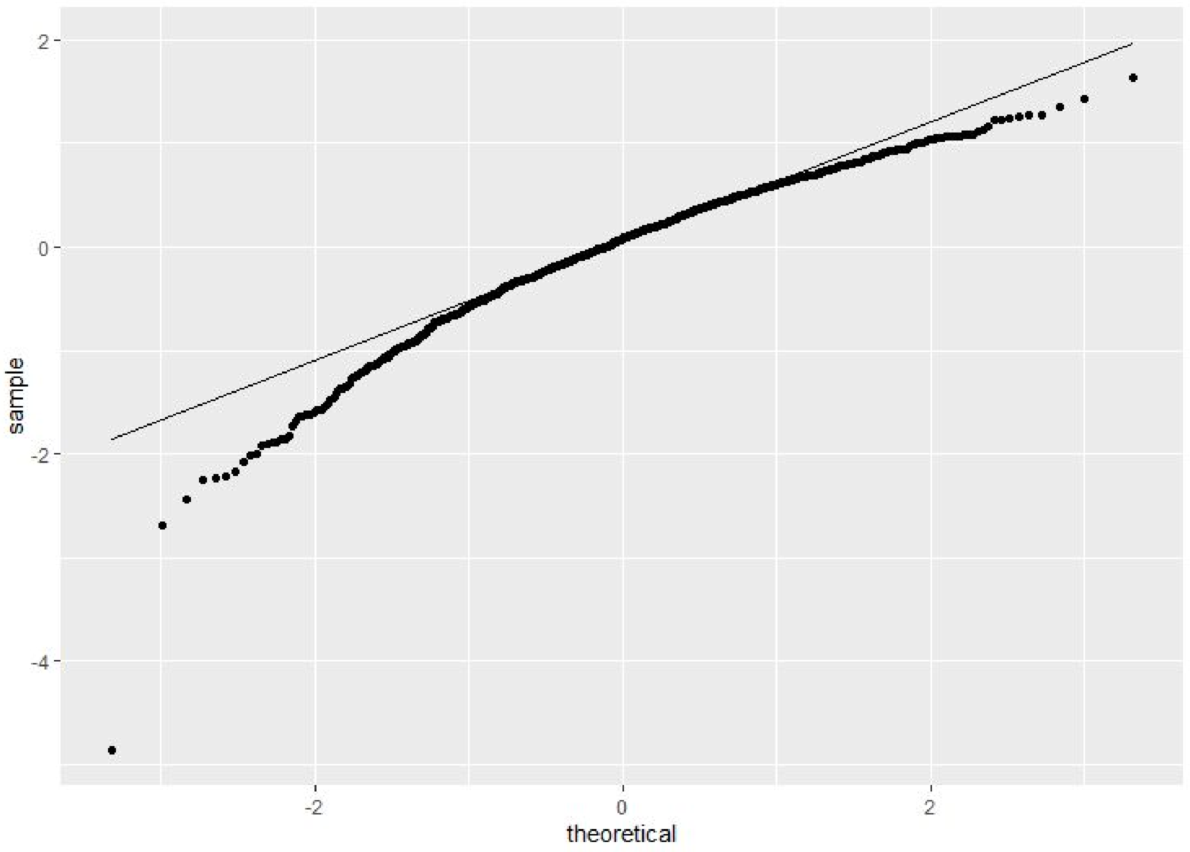
```
temp                    0.0425128  0.0025702  16.541  < 2e-16 ***
I(pressure - 1012)      -0.0570093  0.0034513 -16.518  < 2e-16 ***
temp:I(pressure - 1012) -0.0031910  0.0004753  -6.713 3.06e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.647 on 1087 degrees of freedom
Multiple R-squared:  0.3161,  Adjusted R-squared:  0.3143
F-statistic: 167.5 on 3 and 1087 DF,  p-value: < 2.2e-16
```

2.J




5.0
2.5
0.0
-2.5
-5.0
e.lin
-10
0
10
20
temp


5.0
2.5
0.0
-2.5
-5.0
e.lin
950
975
1000
1025
1050
pressure



Residuals seem better than before….

2.K

2.4

2.M

Lund: 207
Uppsala: 738
Abisko: 146

2.N

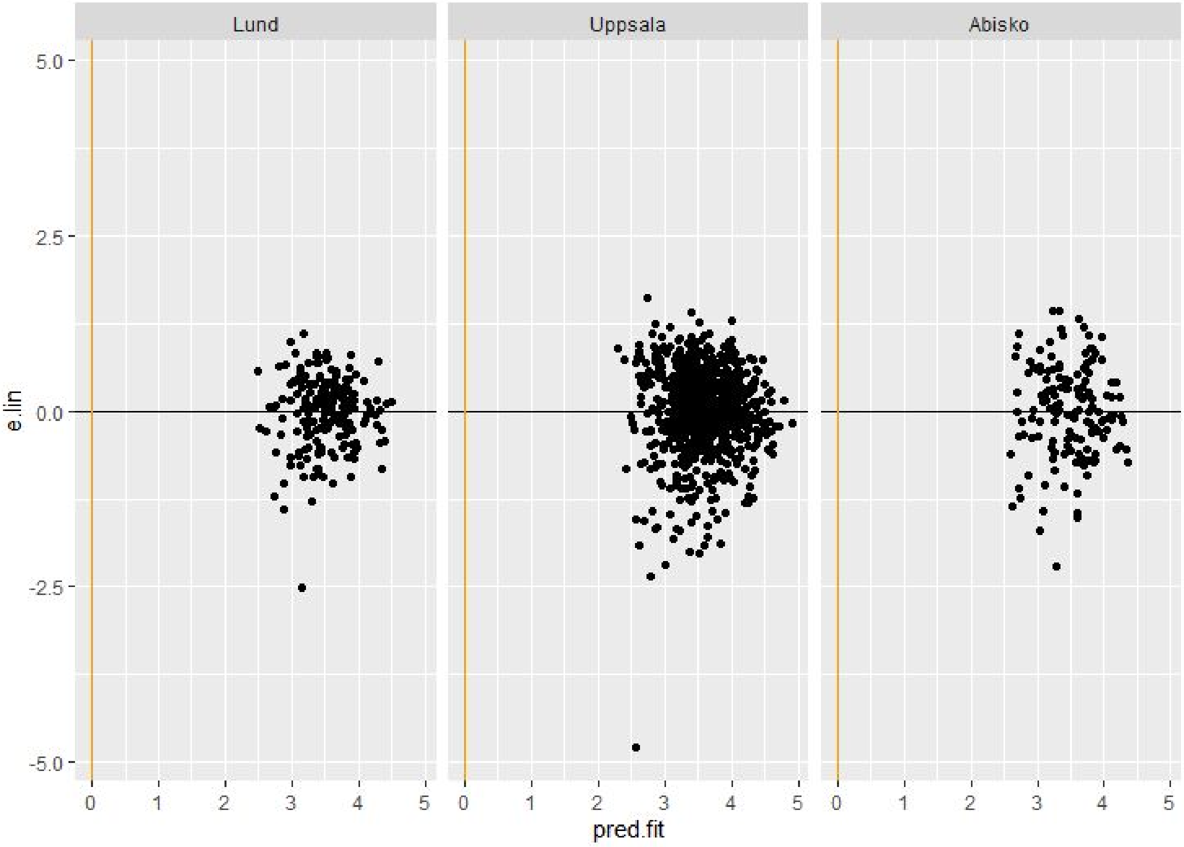
Beta estimates:

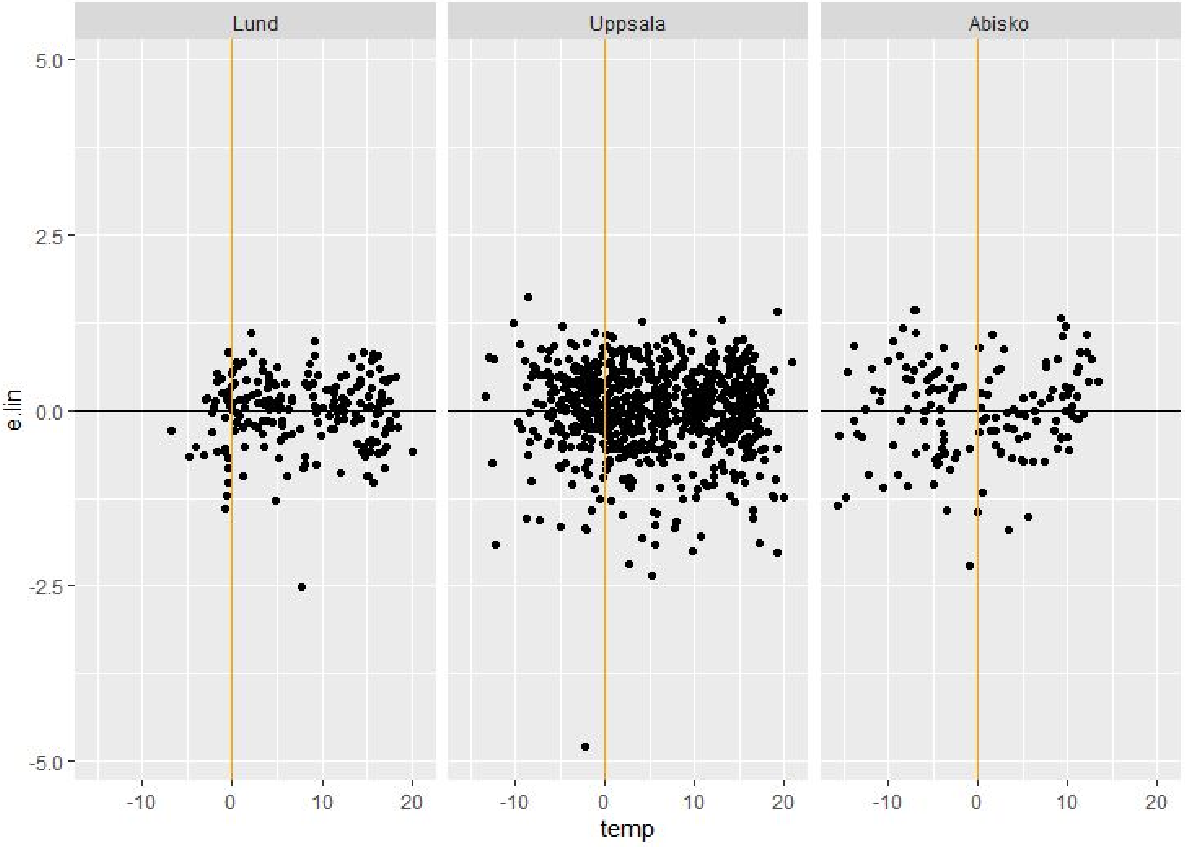
β_0 :(Intercept)	3.3913013
β_1 : temp	0.0346847
β_2 : I(pressure - 1012)	-0.0658602
β_3 :locationLund	0.245181167
β_4 locationAbisko	-0.5112770
β_5 :temp:I(pressure - 1012)	-0.0029452

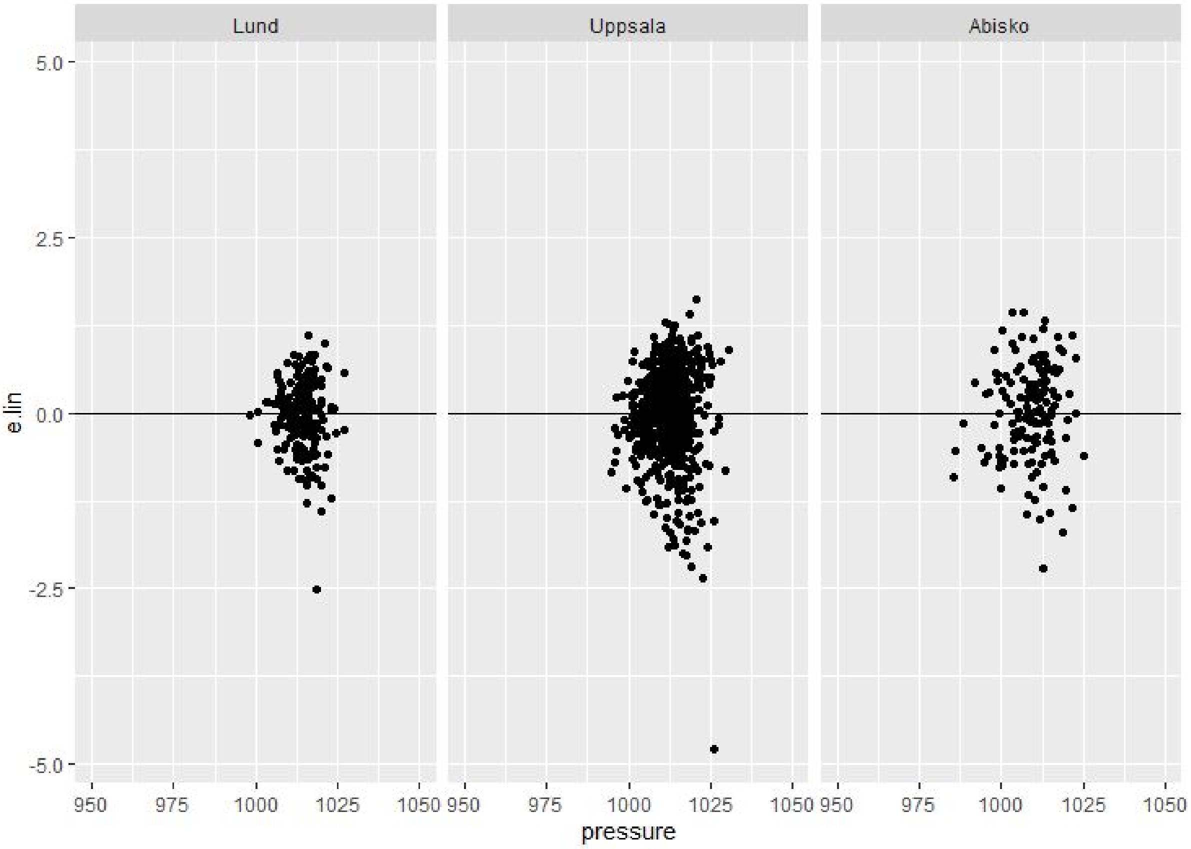
Confidence Intervals:

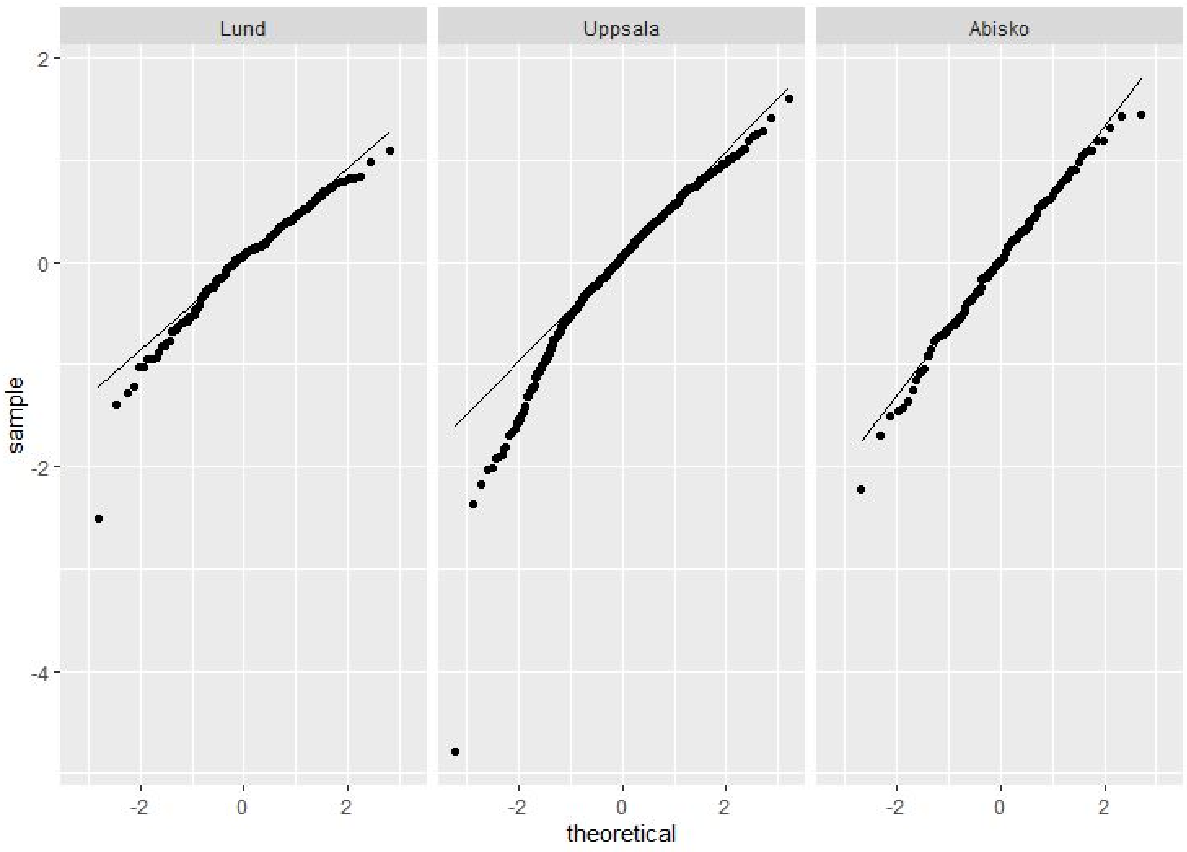
β_0 : (Intercept)	3.339462535	3.443140134
β_1 : temp	0.029761372	0.039608100
β_2 : I(pressure - 1012)	-0.072383230	-0.059337074
β_3 : locationLund	0.245181167	0.436189923
β_4 : locationAbisko	-0.625978812	-0.396575285
β_5 : temp:I(pressure - 1012)	-0.003835445	-0.002055026

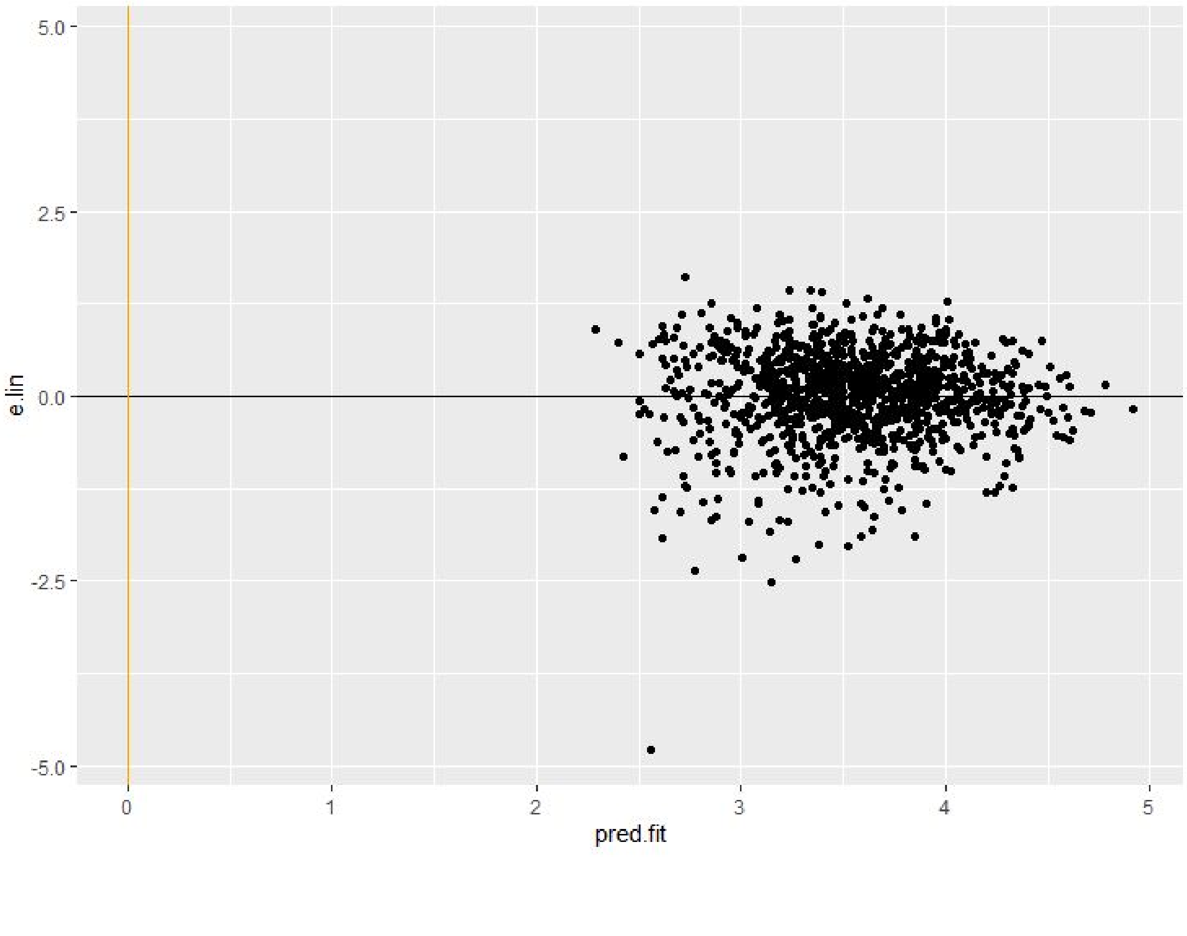
Anova analysis shows that the addition of location is significant.
Pvalue less than 2.2e-16

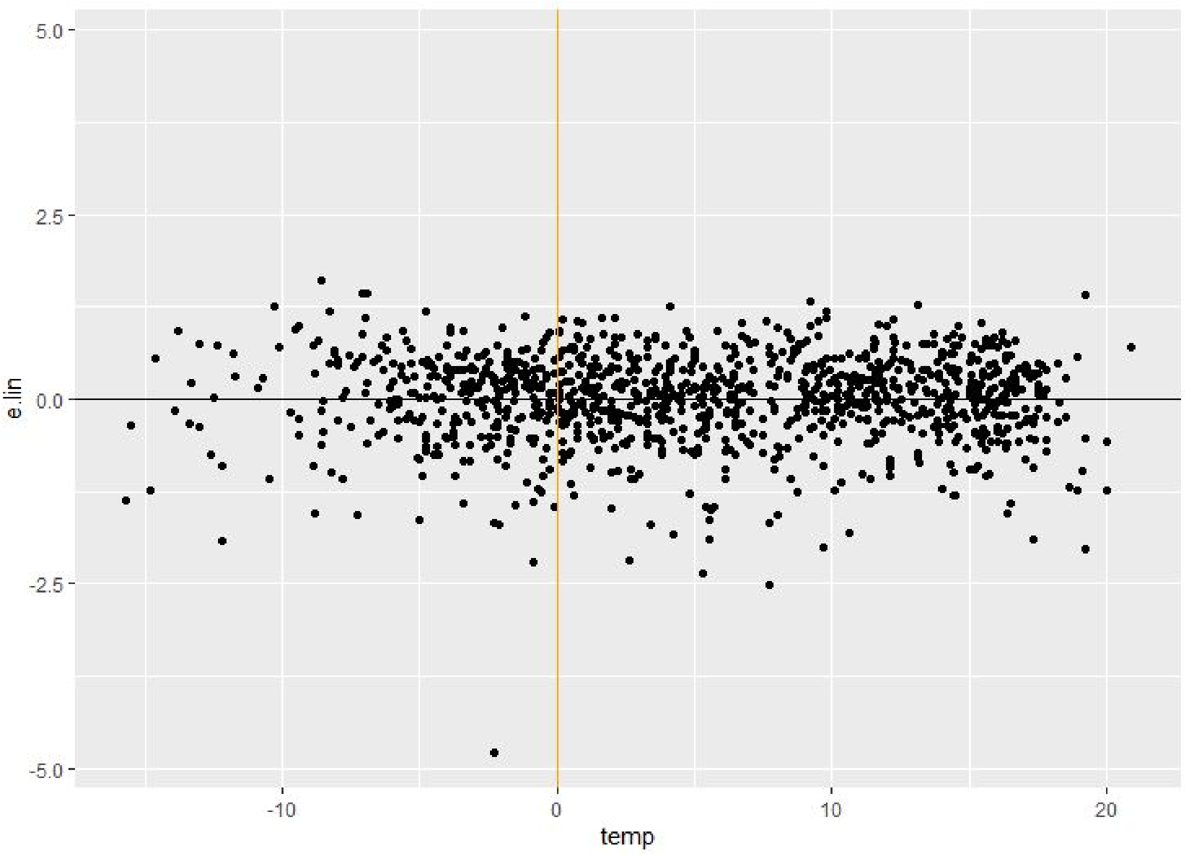

Lund
Uppsala
Abisko
5.0
2.5
0.0
-2.5
-5.0
e.lin
0
1
2
3
4
5
pred.fit

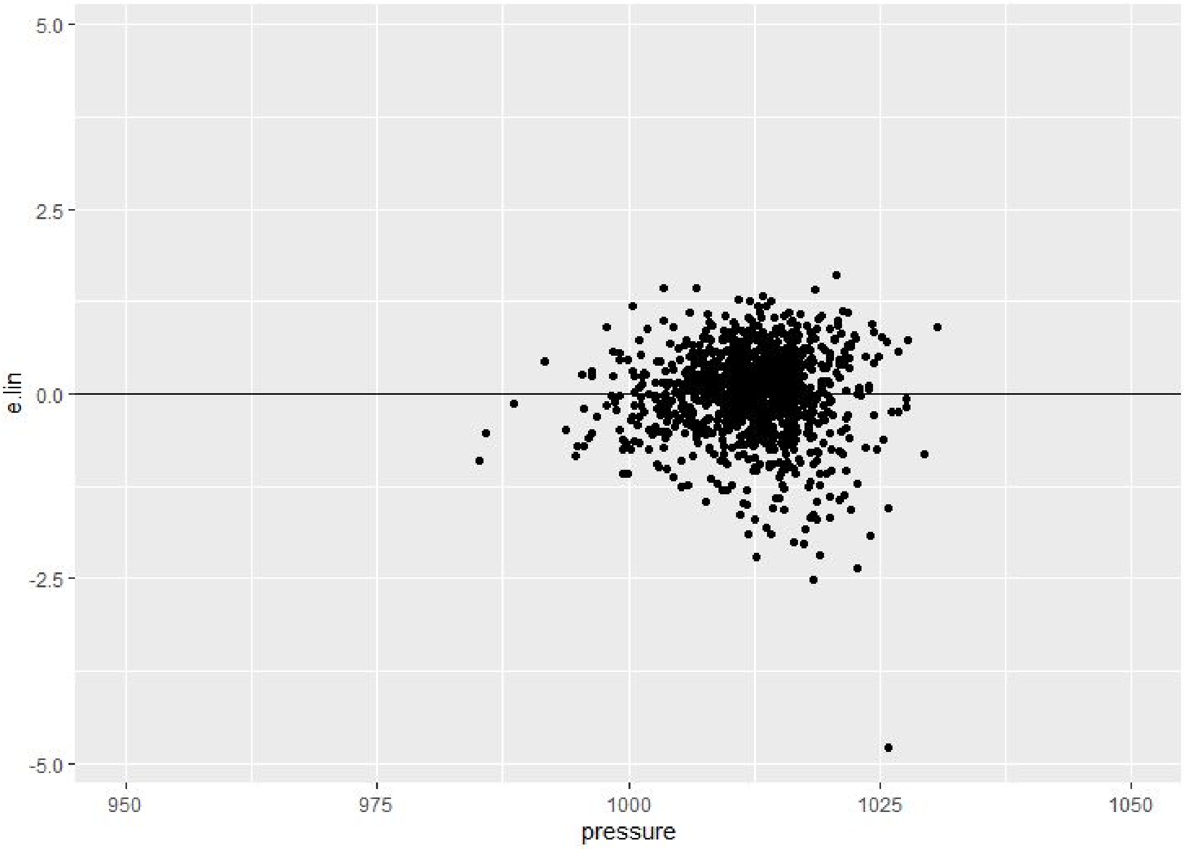

Lund
Uppsala
Abisko
5.0
2.5
0.0
-2.5
-5.0
e.lin
-10
0
10
20
temp

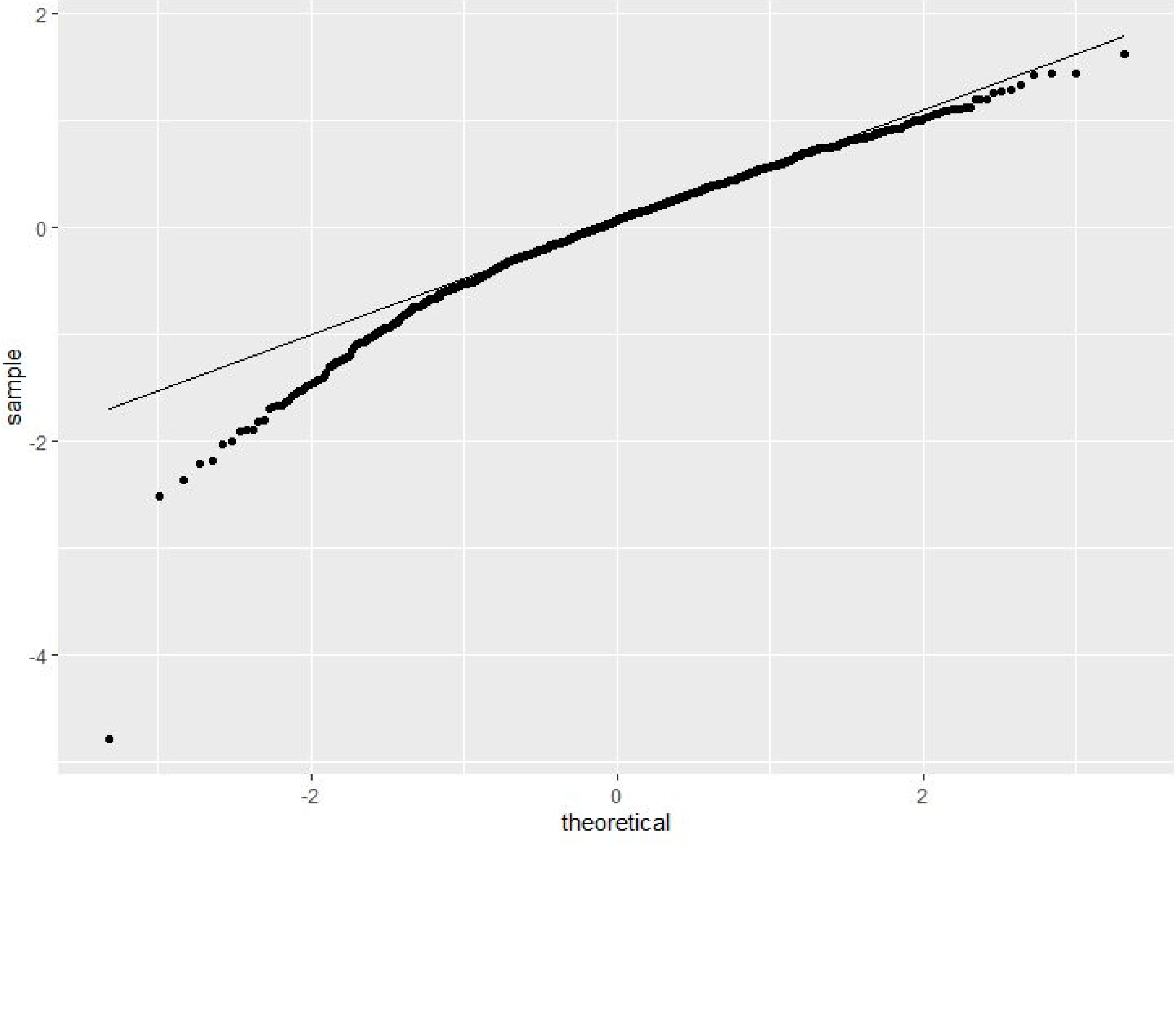

Lund
Uppsala
Abisko
5.0
2.5
0.0
-2.5
-5.0
e.lin
950
975
1000
1025
1050
pressure


Lund
Uppsala
Abisko
sample
theoretical
2
0
-2
-4
-2
0
2


e.lin
pred.fit
5.0
2.5
0.0
-2.5
-5.0
0
1
2
3
4
5


5.0
2.5
0.0
-2.5
-5.0
e.lin
-10
0
10
20
temp


5.0
2.5
0.0
-2.5
-5.0
e.lin
950
975
1000
1025
1050
pressure


2
0
-2
-4
sample
-2
0
2
theoretical

2.O

2.P

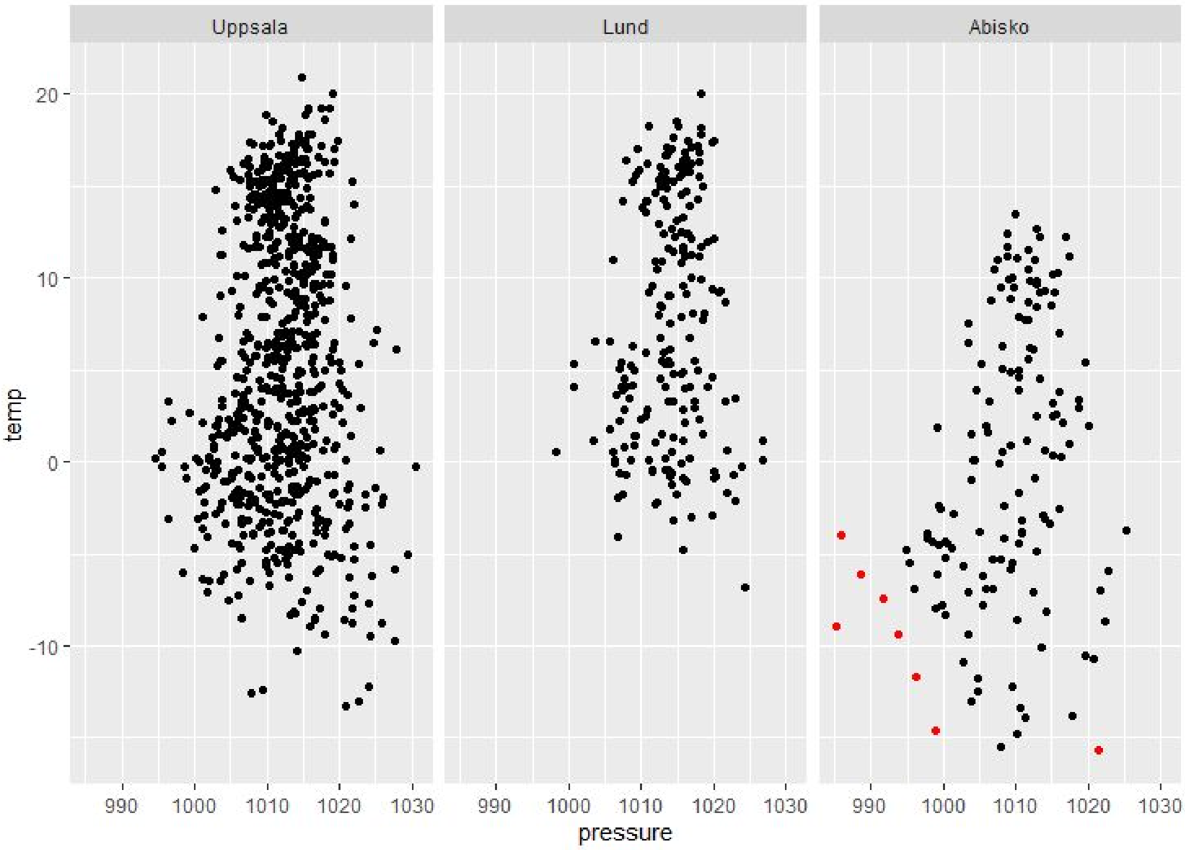
Except for some of the observations from uppsala the residuals look very normally distributed. There is also an outlier in Uppsala.

3

3.A

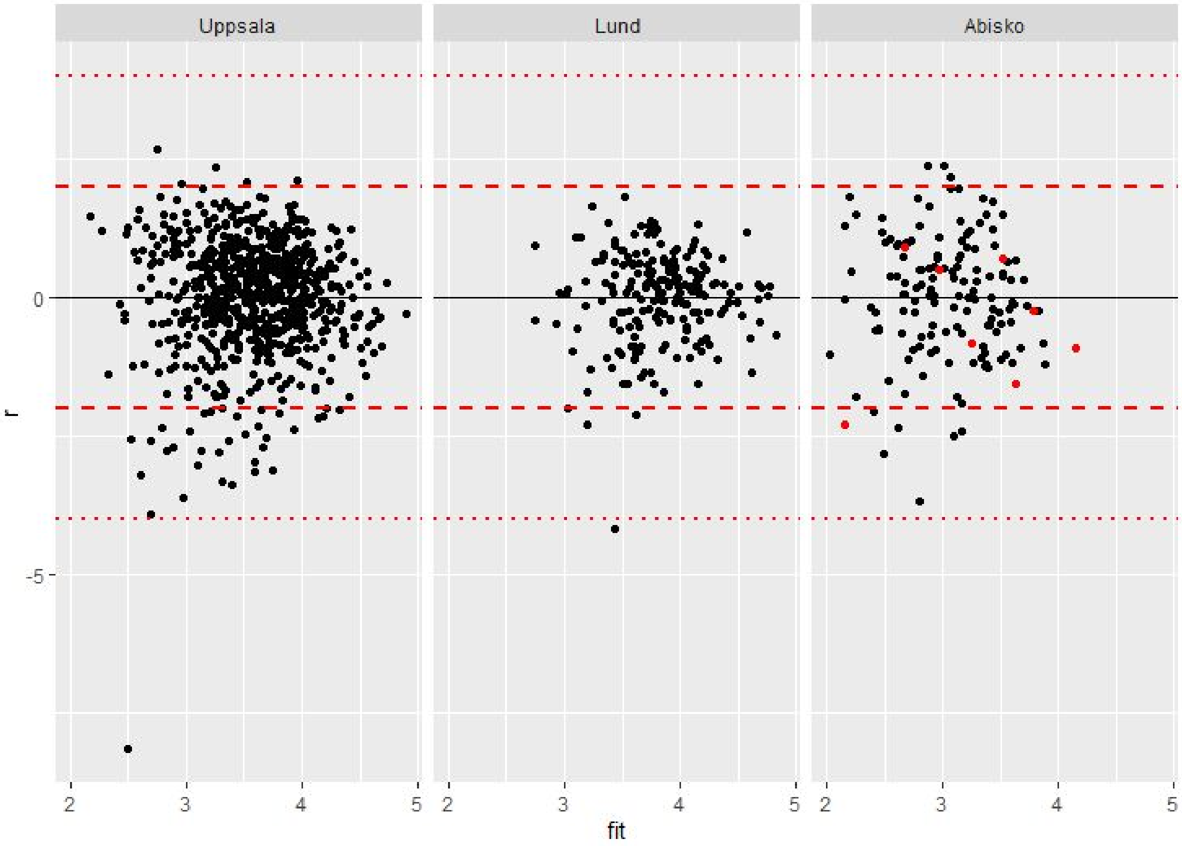
The reason why the smallest leverages in Uppsala are smaller than those of Lund and Abisko is that there are more observations in Uppsala. Naturally each observation in Uppsala is going to have less of an impact. If all else is equal the leverages are going to be smaller.

3.B



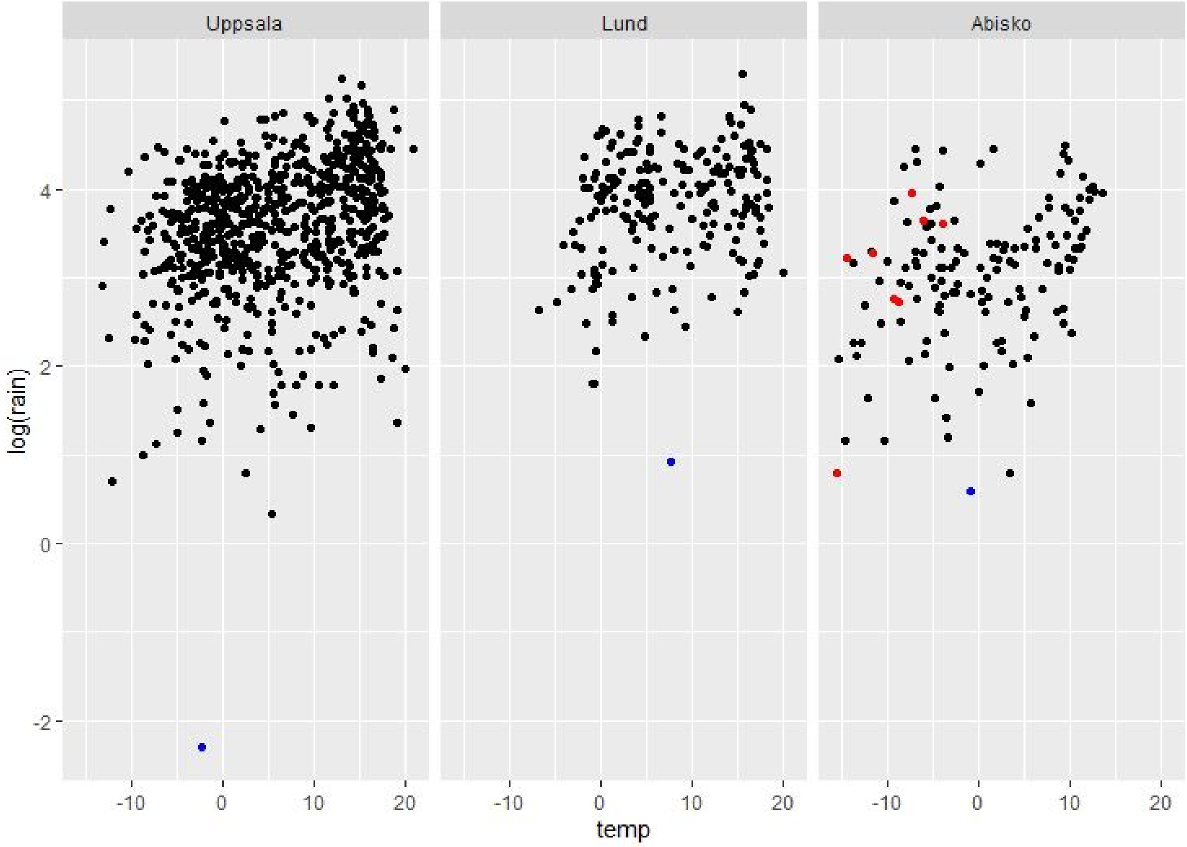
It seems like combo of low pressure and temp leads to high leverage.

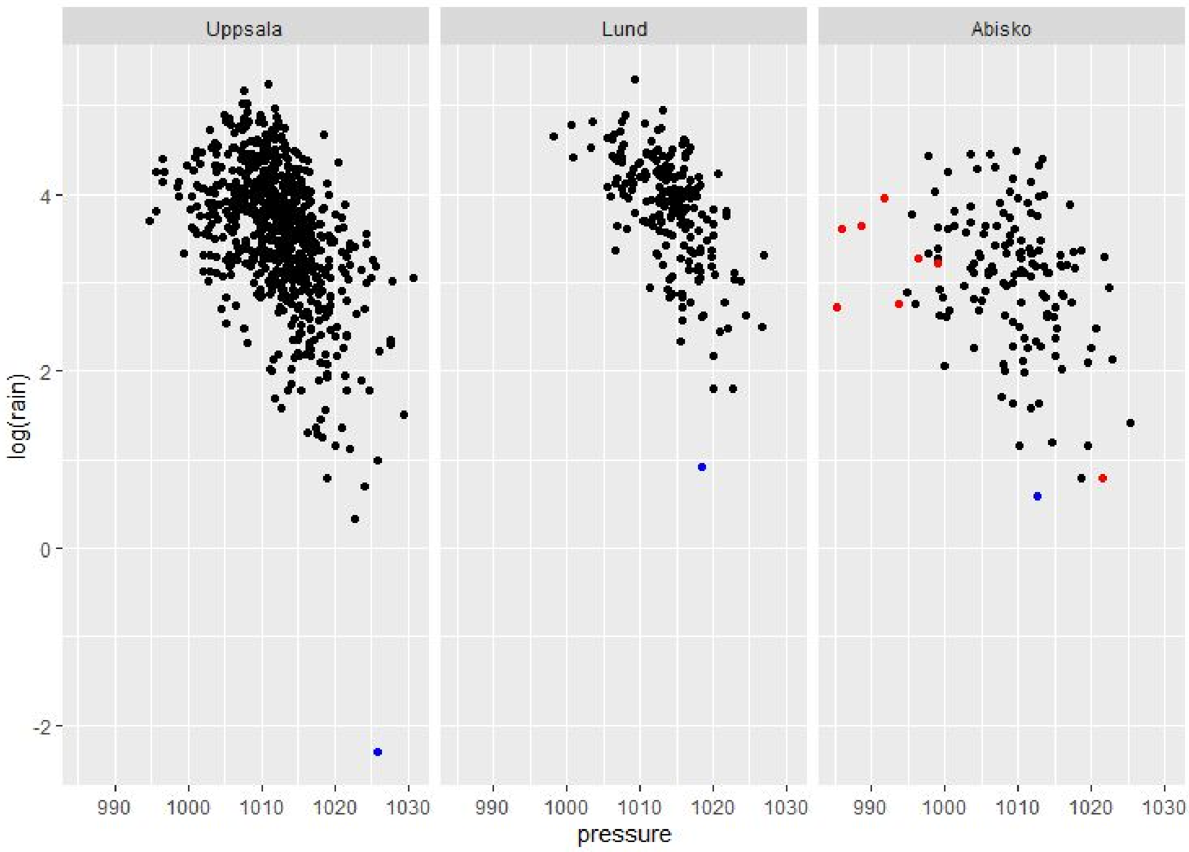
3.C



previously problematic residuals are no longer problematic, large residuals for uppsala.

3.D


Uppsala
Lund
Abisko
log(rain)
4
2
0
-2
-10
0
10
20
-10
0
10
20
-10
0
10
20
temp



3.G

```
> (AIC(logLinRefit))
[1] 1906.725
> (AIC(refit1b))
[1] 2330.778
> (AIC(refit2d))
[1] 2091.793
> (AIC(refit2h))
[1] 2045.335
>
> summary(logLinRefit)$r.squared
[1] 0.4014628
> summary(refit1b)$r.squared
[1] 0.1090147
> summary(refit2d)$r.squared
[1] 0.2863261
> summary(refit2h)$r.squared
```

```
[1] 0.3174707
>
> summary(logLinRefit)$adj.r.squared
[1] 0.3986918
> summary(refit1b)$adj.r.squared
[1] 0.1081927
> summary(refit2d)$adj.r.squared
[1] 0.2850081
> summary(refit2h)$adj.r.squared
[1] 0.3155783
>
> (BIC(logLinRefit))
[1] 1941.657
> (BIC(refit1b))
[1] 2345.749
> (BIC(refit2d))
[1] 2111.754
> (BIC(refit2h))
[1] 2070.286
```

All measurements agree that the model we fitted in 3.n is the better one.

3.H

Anova analysis shows that the addition of the three way interaction term is not significant.

3.I

We end up with our previous model.

3.J

Same here! First variable to be selected is pressure.