LUND
UNIVERSITY

Linear and logistic regression
Project 1

Joel Ahnvik, 960909-7853
Oscar Brink Bolin, 940724-2891

Introduction

In this project we have been given various weather data collected in Sweden from the Swedish Meteorological and Hydrological Institute. The goal has been to develop a suitable model for the total monthly average precipitation. The project has been divided into three parts. In each one of these parts we have been using techniques introduced in different parts of the course's programme to develop said model.

Part 1

As a start, we plotted precipitation against temperature (figure 1) and then fitted a linear model of how the precipitation varies depending on the temperature. After the residuals had been calculated and plotted in a Q-Q-plot (figure 2), we concluded that the model is good between the -1 and 1 quantiles, but that the residuals deviate a lot from the model outside this interval.

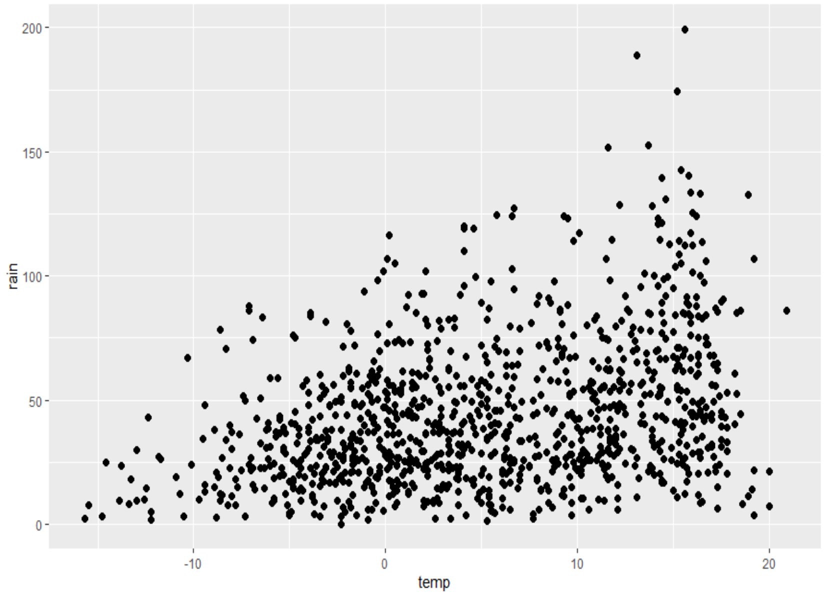


Figure 1: Plot of precipitation against temperature.

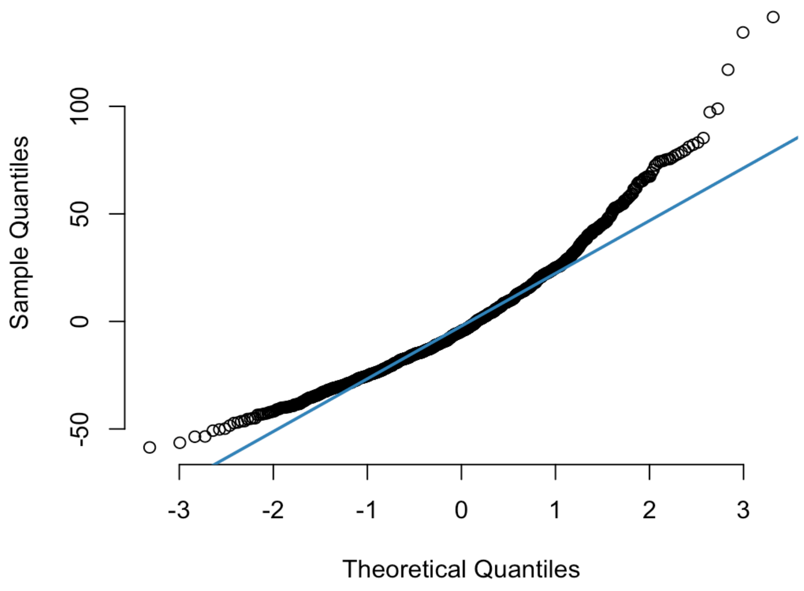


Figure 2: QQ-plot of residuals of the linear model from 1a).

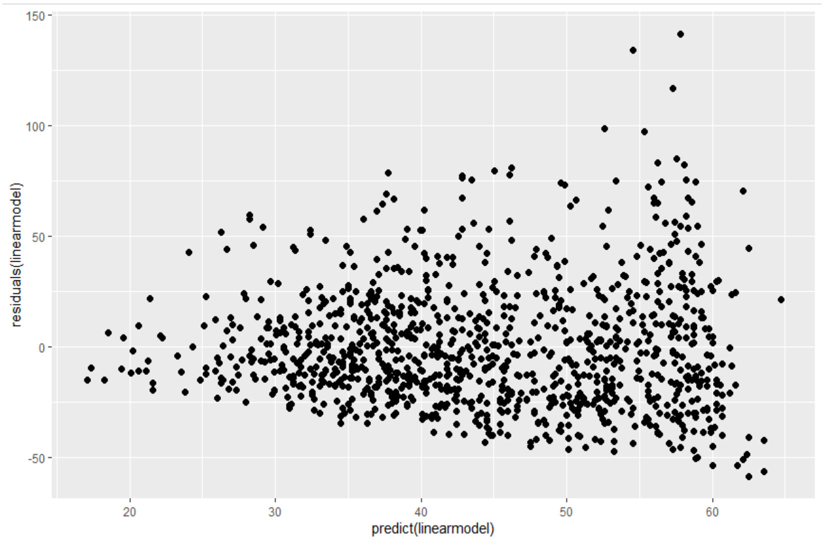


Figure 3: Plot of residuals against predicted values for the logarithmic model from 1a)

Consequently, we refitted the model by taking the logarithm of the observations of precipitation and performed the same plots as above. We concluded that the logarithmic model illustrated a more correct relationship between the variables due to less spread in the plot of observations (figure 4) and in the Q-Q-plot of residuals (figure 5). The same can be said about the plot of residuals against predictions(figure 6).

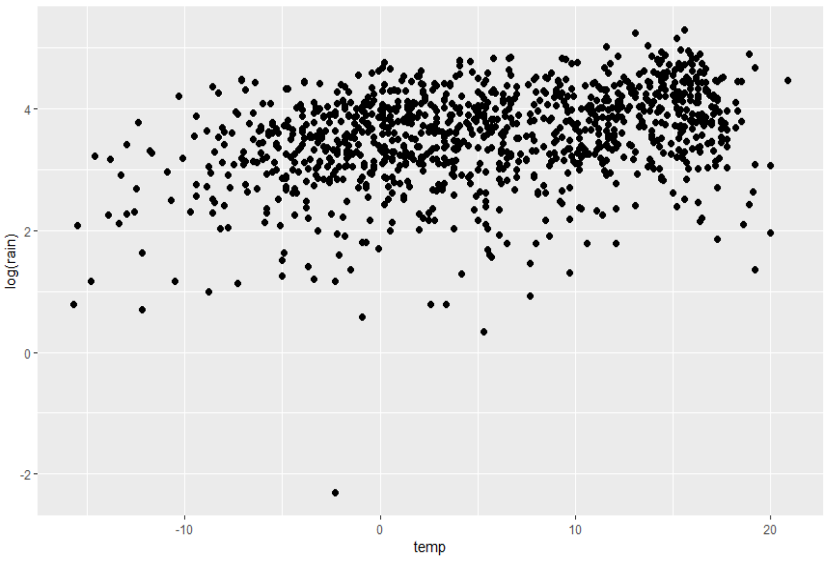


Figure 4: Plot of logarithmized precipitation against temperature.

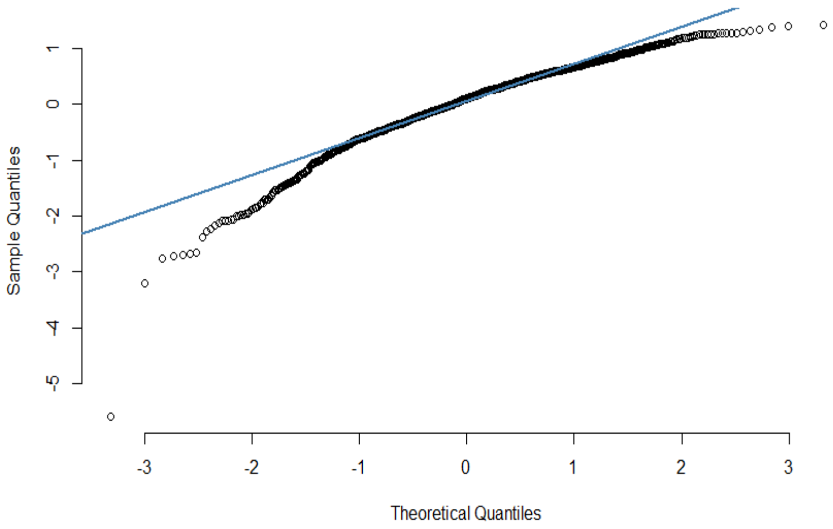


Figure 5: QQ-plot of residuals of the log model from 1b).

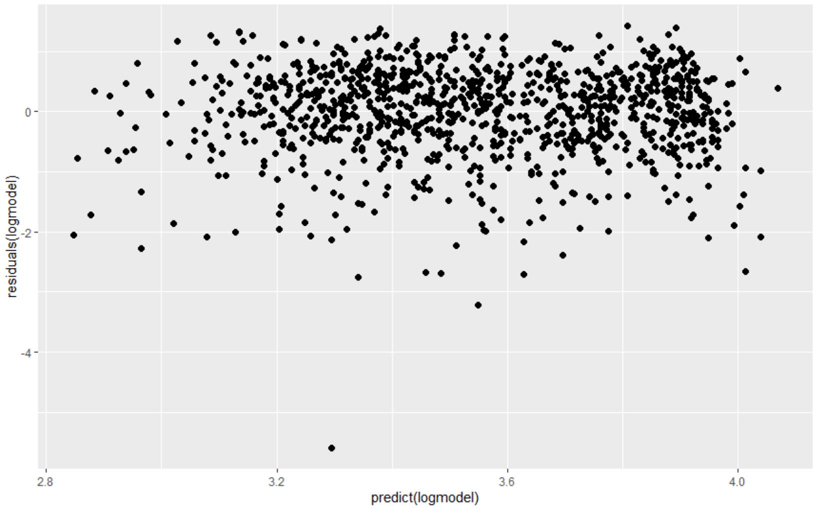


Figure 6: Plot of residuals against predicted values for the logarithmic model from 1b)

By calculating the B-estimates of the logarithmic model from 1b) we got information about how the precipitation changes when the temperature varies. More precisely, when the temperature increases by average 1°C the precipitation increases by 1,3011 mm.

In a comparison between the plot of confidence interval and prediction interval for the logarithmic model from 1b) (figure 7) and for the linear model from 1a) (figure 8), it is clear that the intervals follow the observations better for the logarithmic model. When the average temperature is 5°C, the precipitation is expected to vary between 8.096201 and 146.2836.

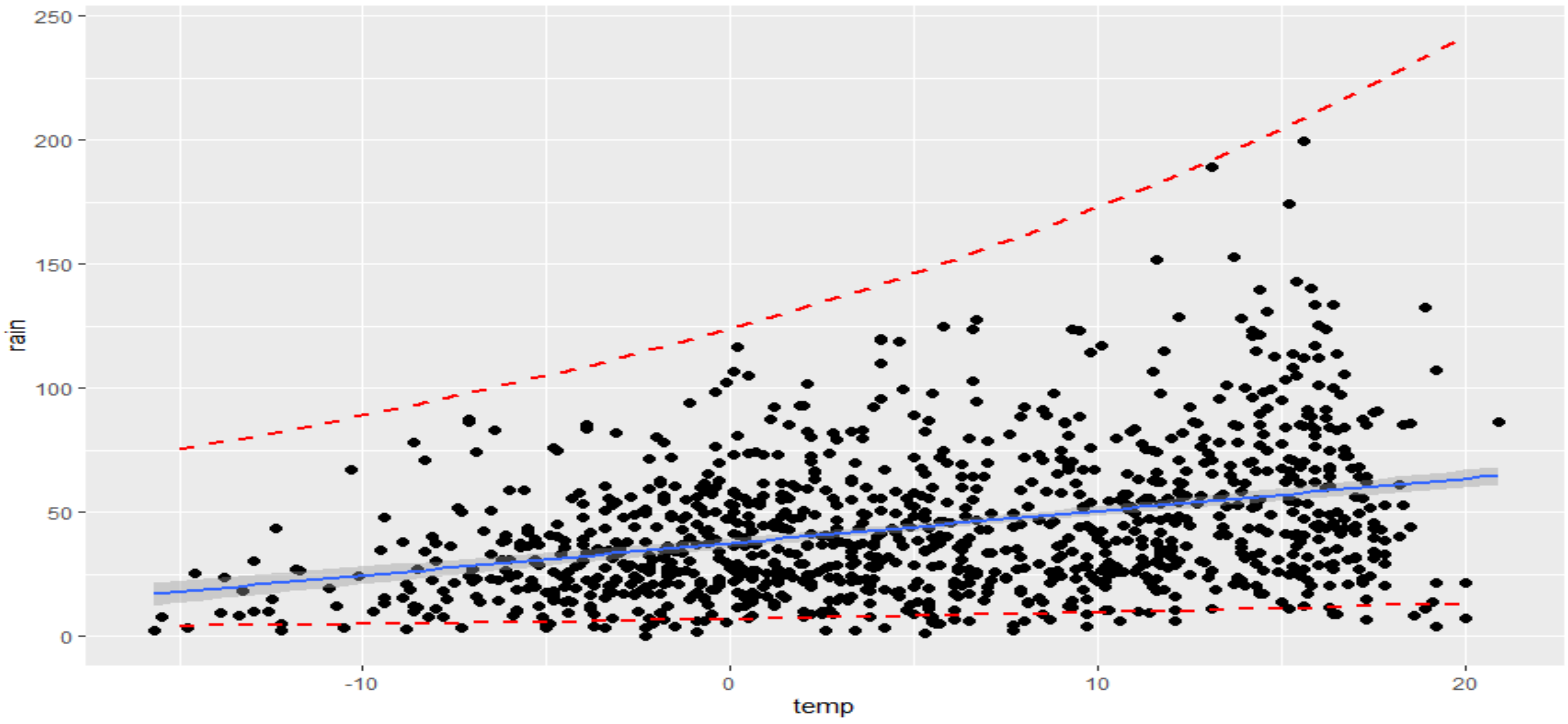


Figure 7: Plot of precipitation against temperature, with confidence interval and the prediction interval for logarithmic model from 1b).

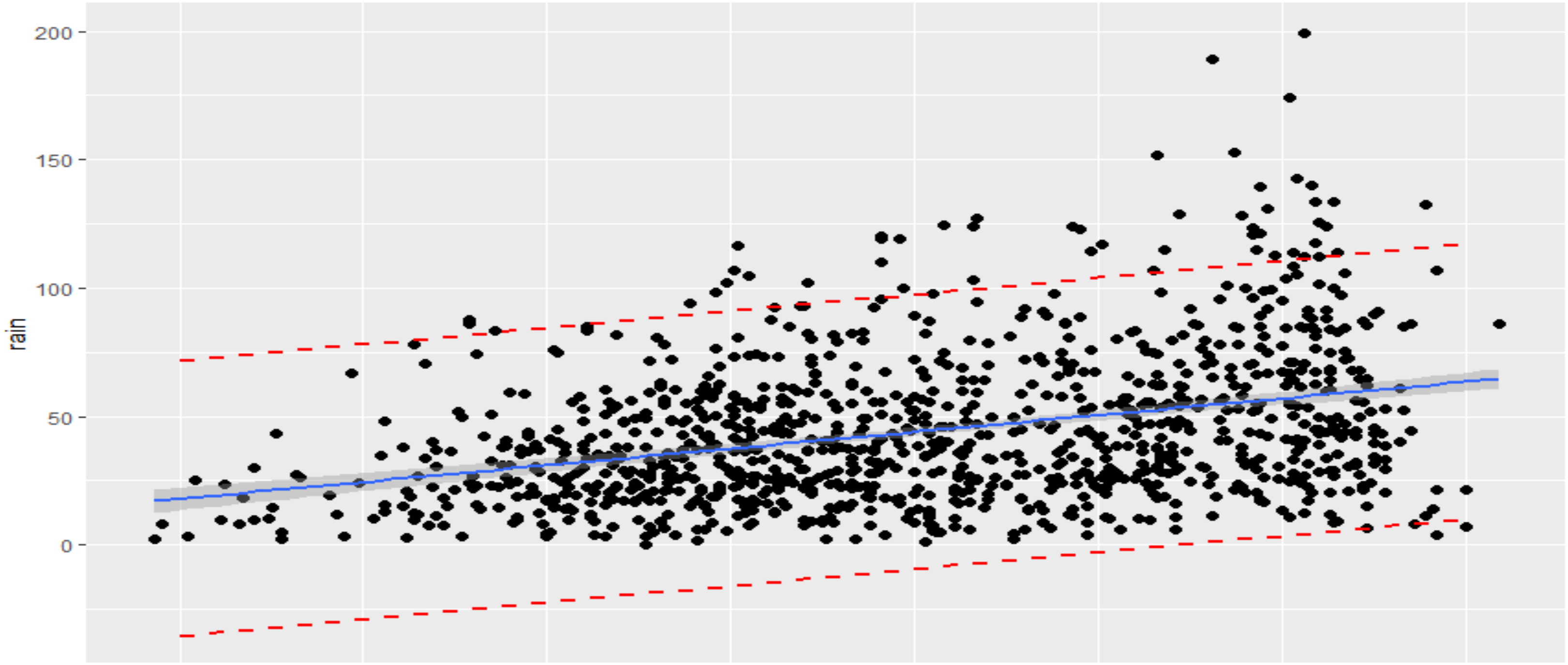


Figure 8: Plot of precipitation against temperature, with confidence interval and the prediction interval for logarithmic model from 1b).

Part 2

In this part we intend to examine whether the monthly average precipitation not only depends on the average temperature but also on the average air pressure. Firstly we did another residual analysis on the model from 1b) but this time we also divided the Q-Q-plot after location. This suggested that we can conclude the same as in part 1 about this model with the addition that for the largest part it is for measurements taken in Uppsala that the model deviates from (figure 9).

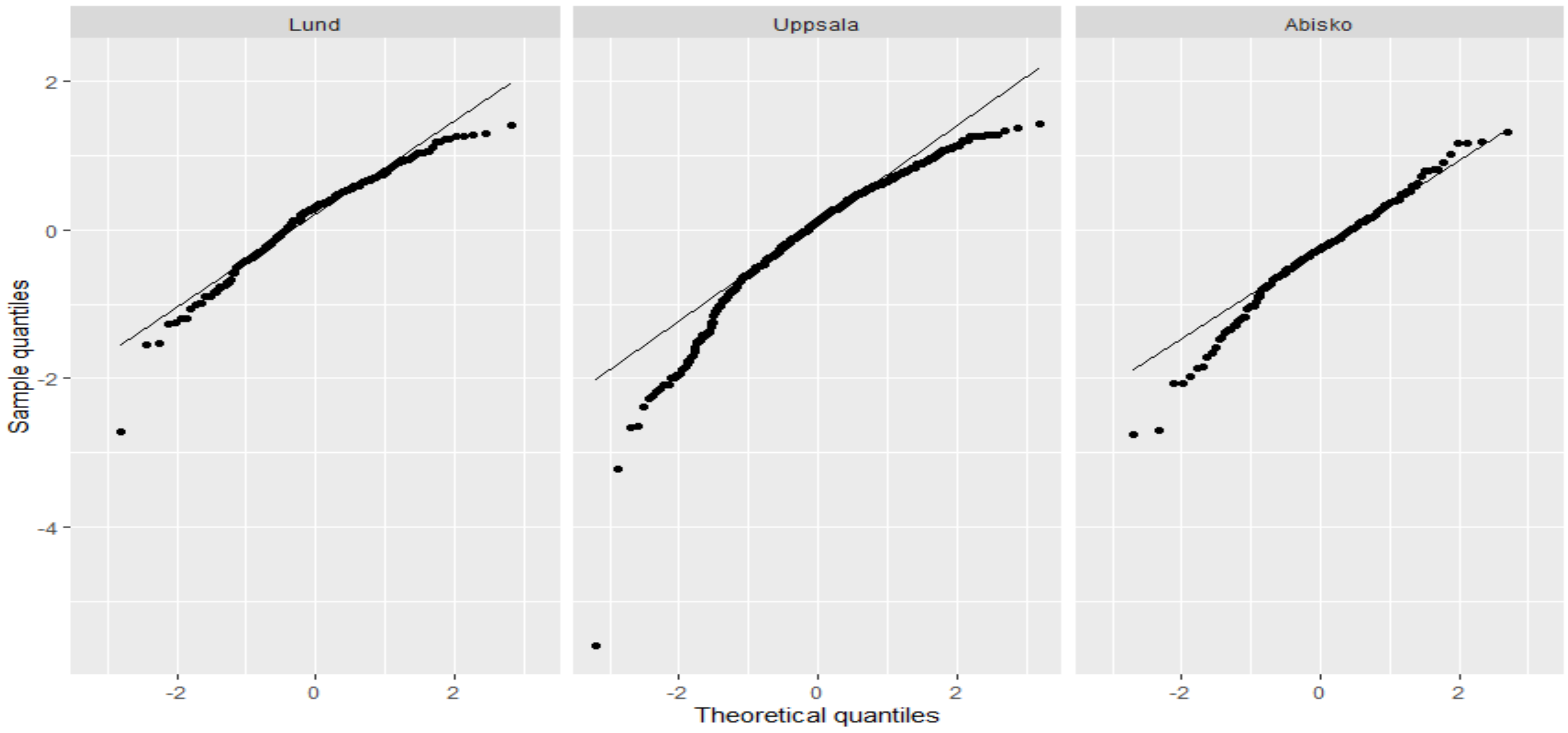


Figure 9: Q-Q-plot separately for each location.

Next thing that we did was to plot the variables rain, temperature, and pressure against each of the others. By this we concluded that there seems to be a linear relationship between precipitation and air pressure (figure 10) . More precisely the rain seems to increase as air pressure increases to about 1010 hPa. Any further increase in air pressure seems to result in decreasing rain. If we use the same transformation as in 1b) and plot log(rain) against air pressure this somewhat also speaks for the linear relationship between rain and air pressure (figure 11). Reflecting on this the linear relationship could also be that normal air pressure should lie slightly above 1000 hPa and therefore we have more data points for these air pressures. This could result in the observed relationship because of the increased chance of higher precipitation due to the fact that there simply are more observed days with air pressure around 1010 hPa. Further on it does not seem like there is a strong linear relationship between temperature and air pressure that will cause any major problems (figure 12).

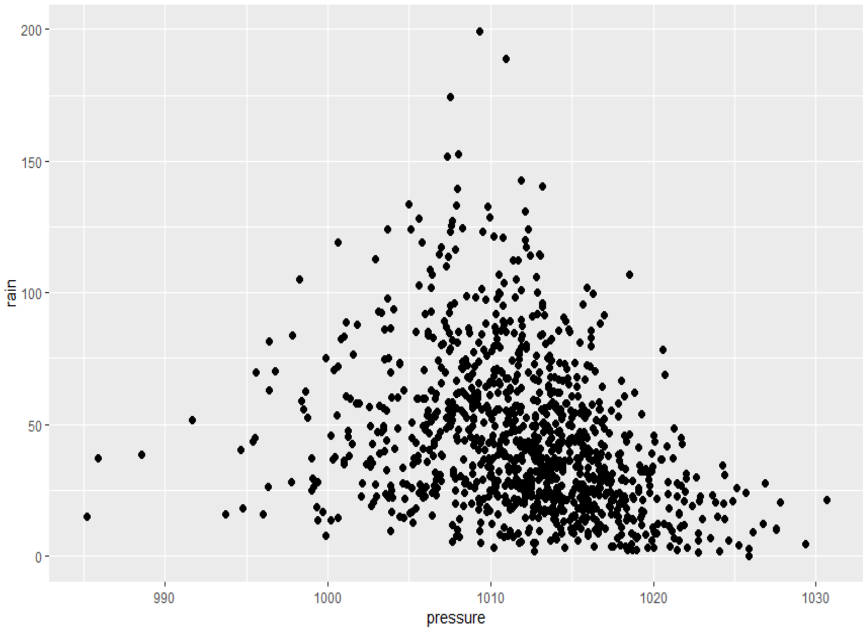


Figure 10: Plot of rain against pressure

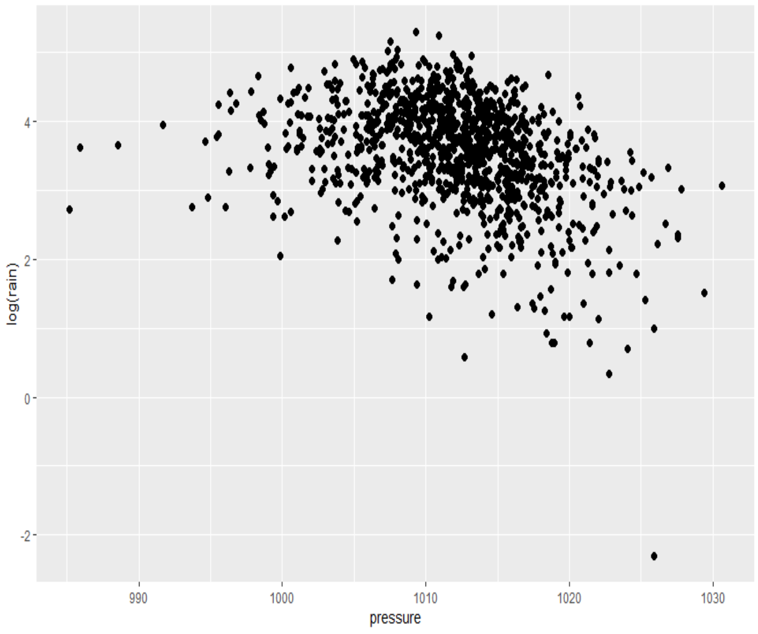


Figure 11: Plot of log(rain) against pressure

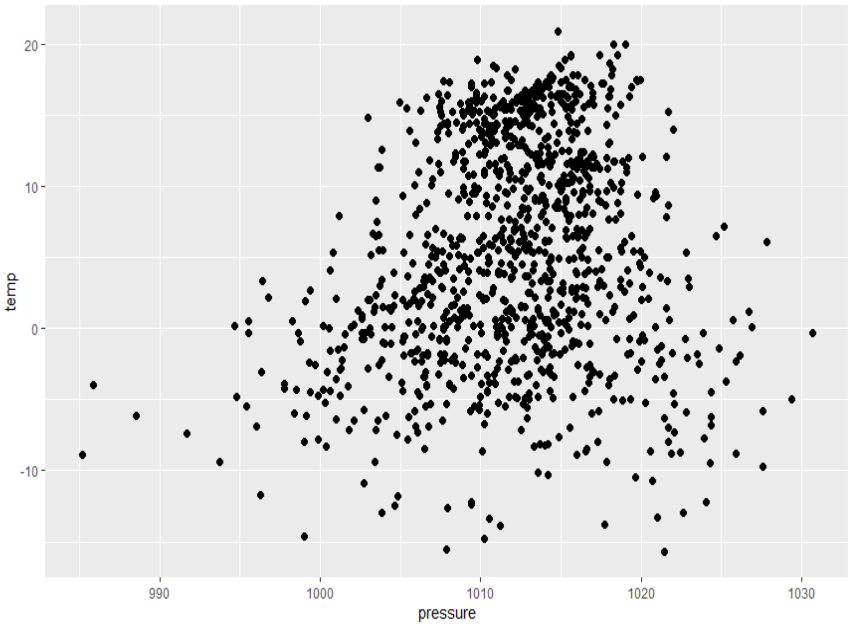


Figure 12: Plot of temperature against pressure

When fitting a new model, log(rain) as a function of both temperature and pressure the inclusion of both temperature and pressure was statistically significant. $\beta_0 = 61,903$ (intercept), $\beta_1 = 0.0404$ (temperature), $\beta_2 = -0.0579$ (air pressure). All these *β*-values had a p-value < 2.2*10^(-16) concluding that they are statistically significant. Their 95%-confidence interval was: $Confint_{Intercept}(54.921,\ 68,884)$, $Confint_{temperature}(0.0353,\ 0.0456)$ and $Confint_{air\ pressure}(-0.0648,\ -0.0510)$.

Now performing another residual analysis and making a new Q-Q-plot (figure 13), the Q-Q-plot suggests further improvement since we see less deviation at the tails of the plots. Plotting of residuals separately against temperature (figure 14) and air pressure (figure 15) indicates that there is no special relationship to be found. This also holds for when residuals are plotted against predicted values (figure 16).

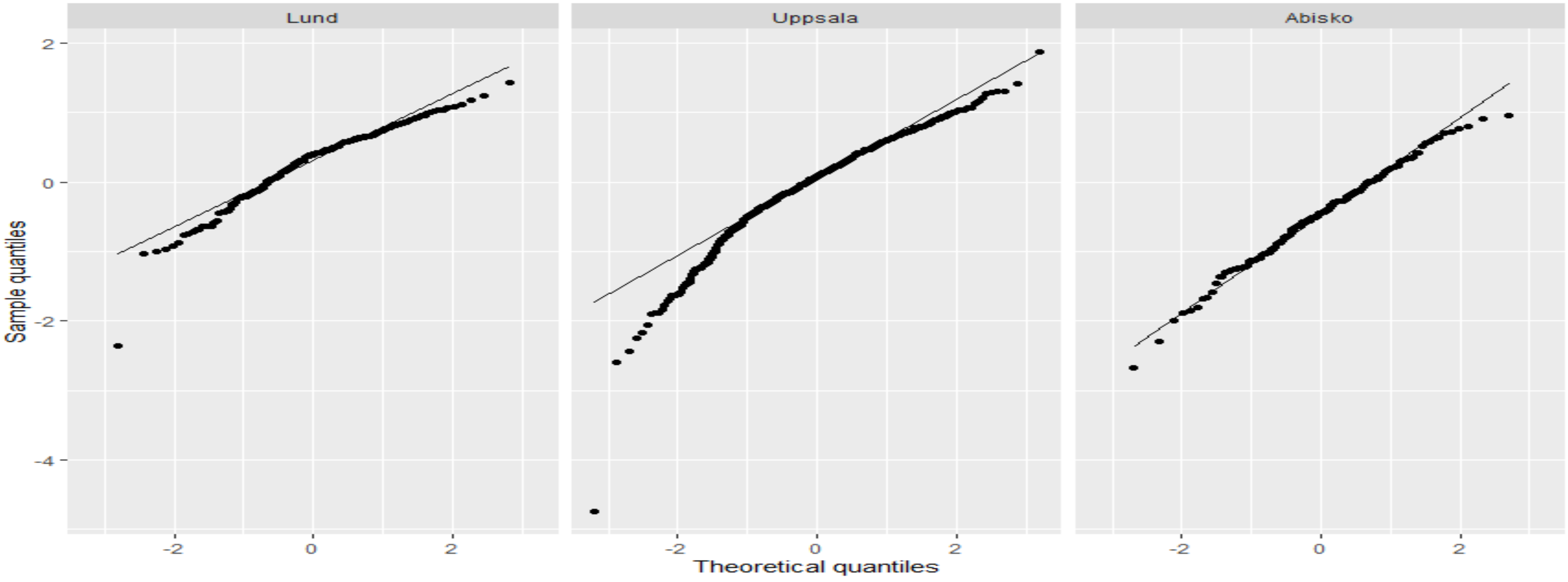


Figure 13: Q-Q-plot of model where log(rain) is a function of temperature and pressure, divided by location

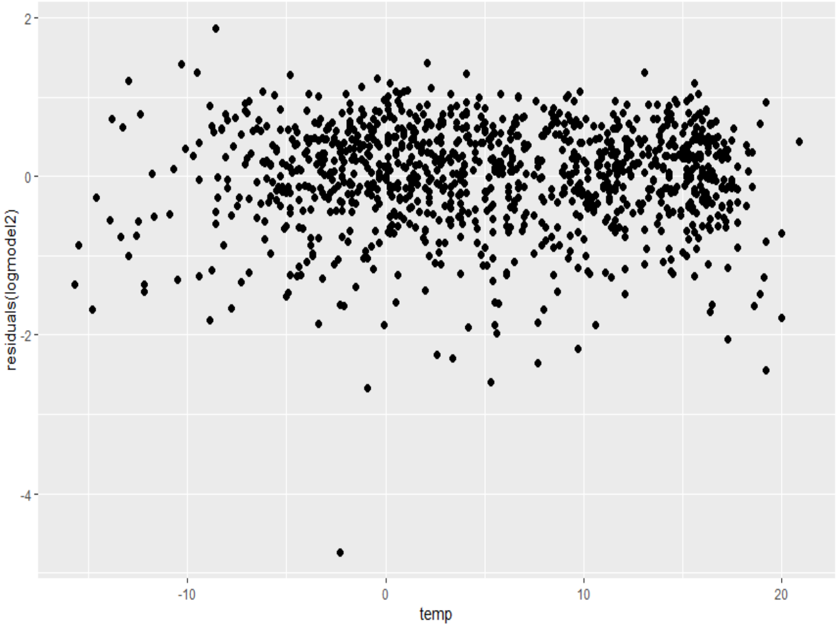


Figure 14: Plot of residuals of model where log(rain) is a function of temperature and pressure against temperature

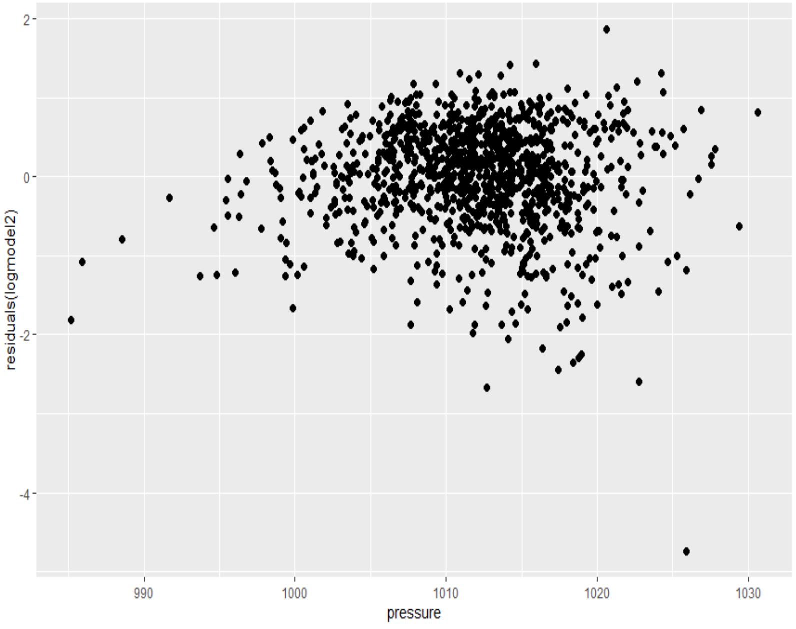


Figure x: Plot of residuals of model where log(rain) is a function of   temperature and pressure against pressure

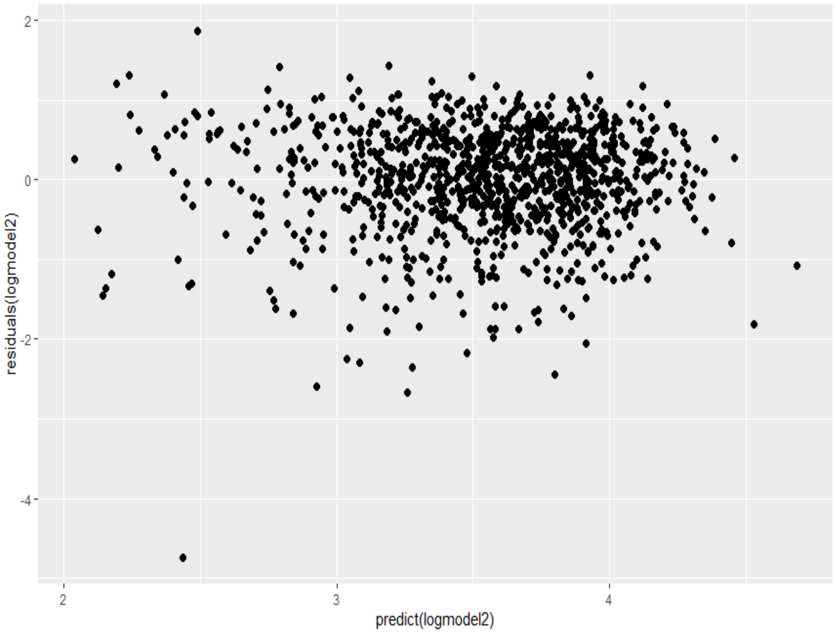


Figure 15: Plot of residuals of model where log(rain) is a function of temperature and pressure against predicted values of same model

According to this latest model (log(rain) as a function of both temperature and air pressure), when temperature increases with 1°C the precipitation increases with 1.534 mm, and when air pressure increases with 20 hPa the precipitation decreases with 38.2 mm (1.91*20 mm). The change in precipitation due to change in temperature compared to the previous model from 1d) where this increase was 1.301 mm is a significant increase in precipitation of roughly 18%. Further on for months where the temperature is 5°C prediction intervals of precipitation for months with average air pressure of 1000 hPa and 1020 hPa are: $Predint_{1000\ hPa}(18.795,\ 252.134)$, $Predint_{1020\ hPa}(5.915,\ 79.126)$. This with average precipitation of 68.839 mm for months with an average air pressure of 1000 hPa and 21.634 mm for months with an average air pressure of 1020 hPa. This is in line with the model that when air pressure increases the precipitation decreases which will result in that for months with higher average air pressure the average precipitation will be lower and the prediction intervals for months with different average air pressure will differ. Next we examined the effect of interaction between temperature and air pressure. Firstly we fit a model where log(rain) is a function of temperature, pressure and the interaction of temperature and pressure. Here we have the β-values as follows: $\beta_0 = 61{,}041$ (intercept), $\beta_1 = 3.272$ (temperature), $\beta_2 = -0.057$ (pressure), $\beta_3 = -0.003$ (temperature*pressure). The largest of the β-values' p-values is 3.06*10^-11 making the terms statistically significant. When refitting to this new model the impact of air pressure became significantly lower and the impact of temperature became significantly higher. Refitting the model onces again, this time replacing pressure with I(pressure - 1012) (because of 1012 hPa being the average air pressure), gives a new model where this time the impact of temperature is significantly lower than before. Once again making a residual analysis and a Q-Q-plot and comparing to earlier models shows no significant improvement when looking at the residuals and a marginal improvement when looking at the Q-Q-plot (figure 19). According to this model the precipitation decreases with 0.185*(pressure - 1012) as a function of pressure when the temperature rises. This means that if the pressure is 1000 hPa the precipitation will increase and if the pressure is 1020 hPa the precipitation will decrease. Compared to the model in part 1 the main difference is that that model does not take the effect of air pressure on precipitation into account. Further on we want to consider the effect on precipitation if temperature changes as a function of air pressure. Looking at confidence intervals of combinations of different temperature and air pressure (-10°C, 10°C, 1000 hPa, 1020 hPa) shows that for a given temperature a change in air pressure, or vice versa, makes a significant difference in average precipitation.

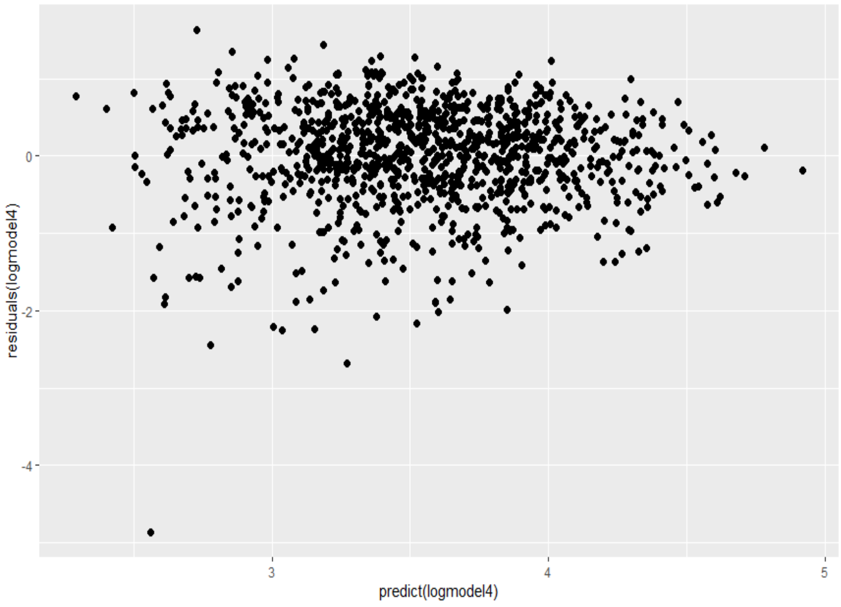


Figure 16: Plot of residuals of model where log(rain) is a function of function temperature, I(pressure-1012) and temperature interacting with I(pressure-1012) against predicted values of same model

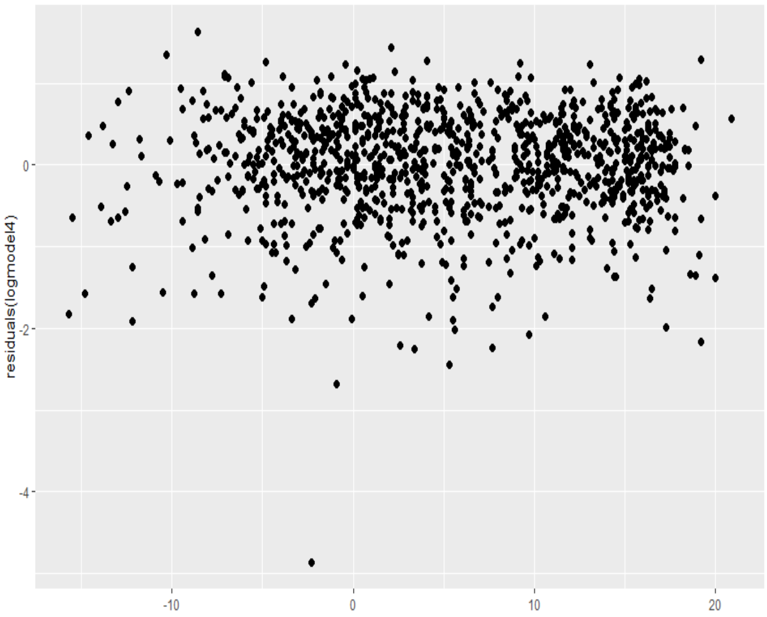


Figure 17: Plot of residuals of model where log(rain) is a of temperature, I(pressure-1012) and temperature interacting with I(pressure-1012) against temperature

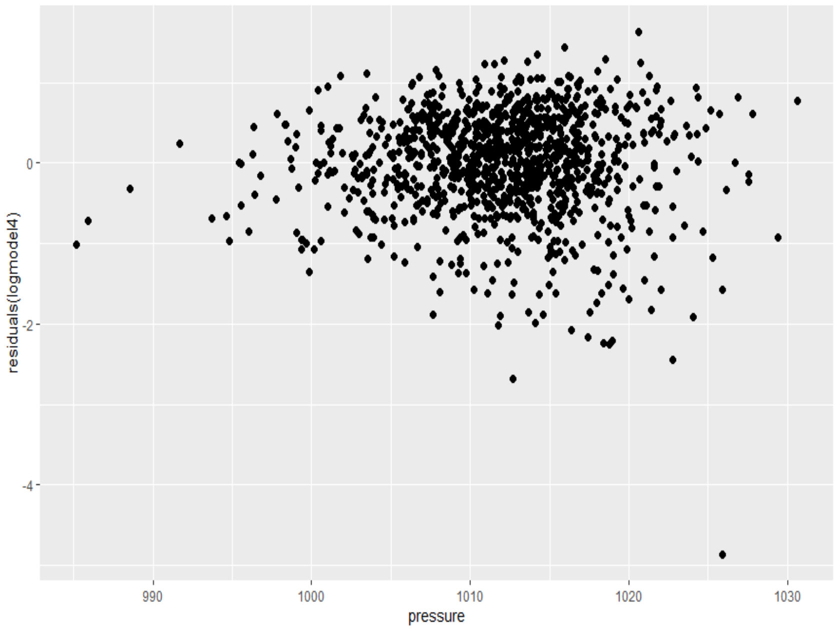


Figure 18: Plot of residuals of model where log(rain) is a function of temperature, I(pressure-1012) and temperature interacting with I(pressure-1012) against pressure

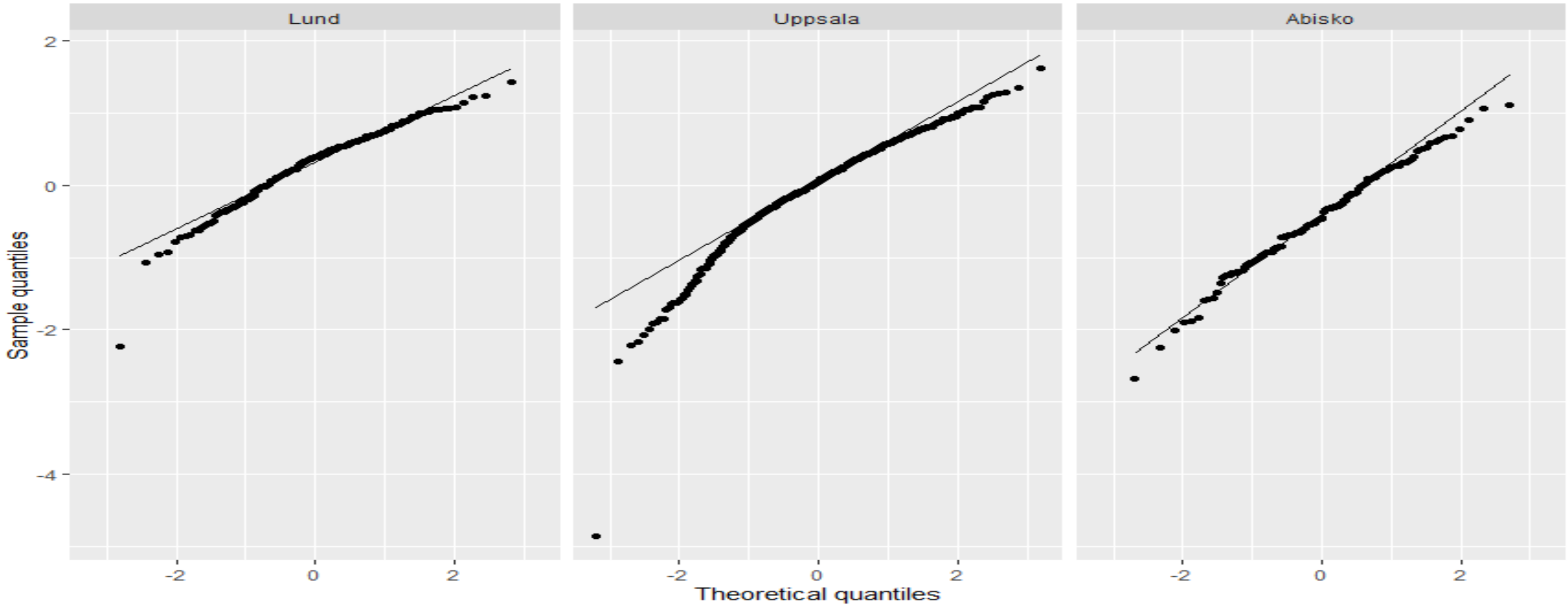


Figure 19: Q-Q-plot of model where log(rain) is a function of temperature, I(pressure-1012) and temperature interacting with I(pressure-1012), divided by location

Finally we examined the effect of the measurement stations location on the average precipitation as log(rain) as a function of temperature, pressure, temperature interacting with pressure and location. Fitting this new model gave the following new *β*-values: $\beta_0 = 70{,}041$ (intercept), $\beta_1 = 3{,}015$ (temperature), $\beta_2 = -0{,}066$ (pressure), $\beta_3 = 0.341$ (Lund), $\beta_4 = -0.511$ (Abisko), $\beta_0 = -0.003$ (temperature*pressure). All of which are statistically significant when looking at their p-values. The biggest and the significant change in *β*-values was the increase in the intercept from ~61 to ~70. A residual analysis showed no significant difference compared to before. However when comparing Q-Q-plots (figure 20) with earlier models we see a slightly better fit.

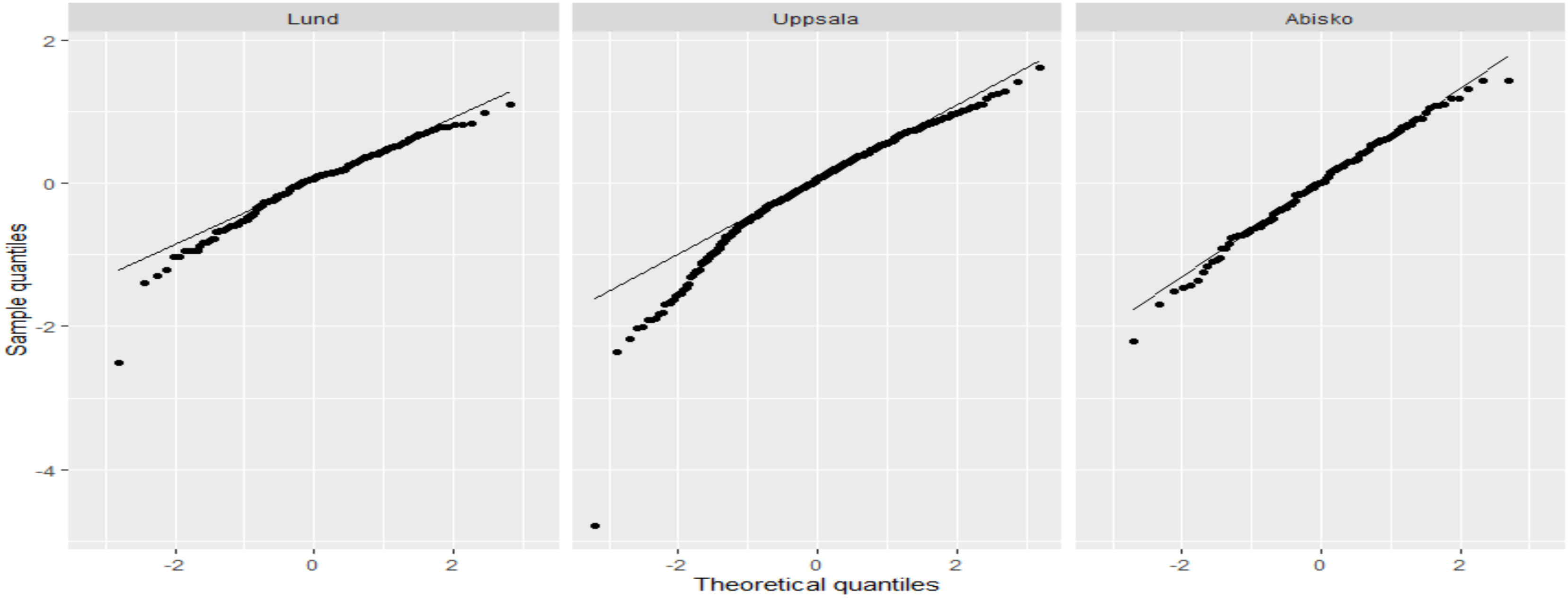


Figure 20: Q-Q-plot of model where log(rain) is a function of temperature, pressure, temperature interacting with pressure and location, divided by location

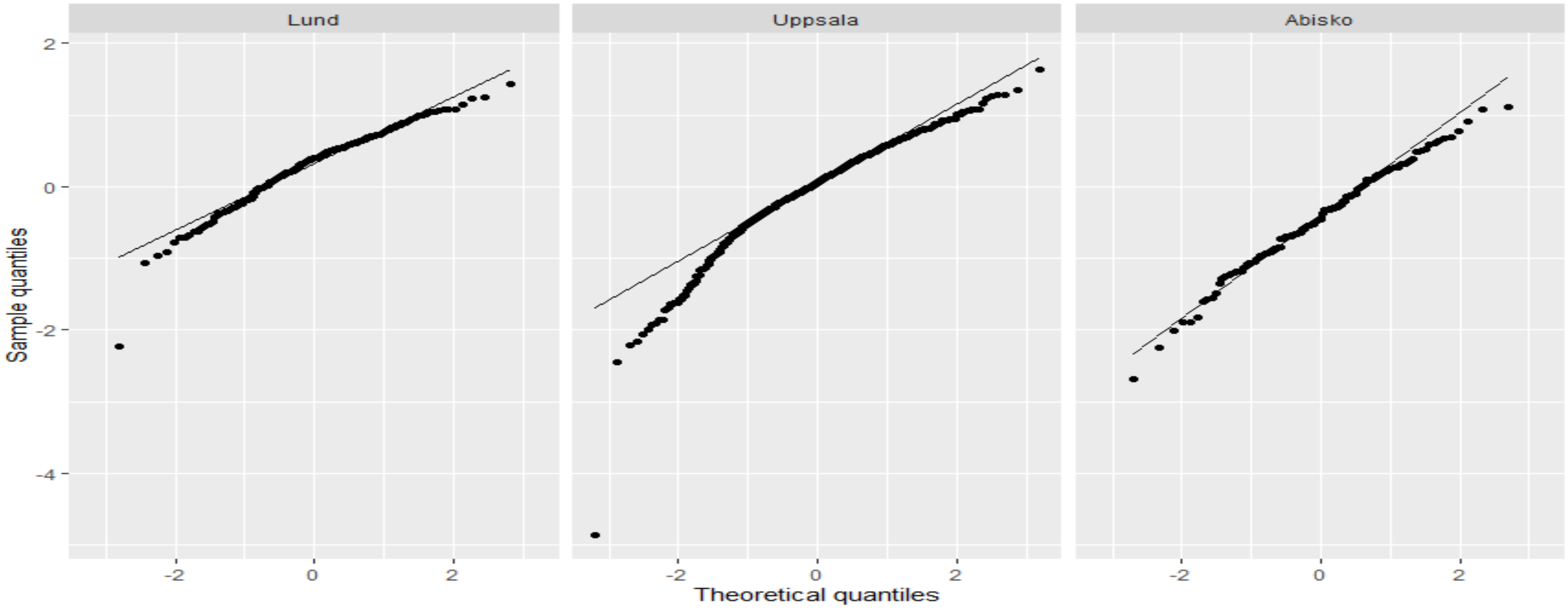


Figure 21: Q-Q-plot of model where log(rain) is a function of temperature, I(pressure-1012) and temperature interacting with I(pressure-1012), divided by location

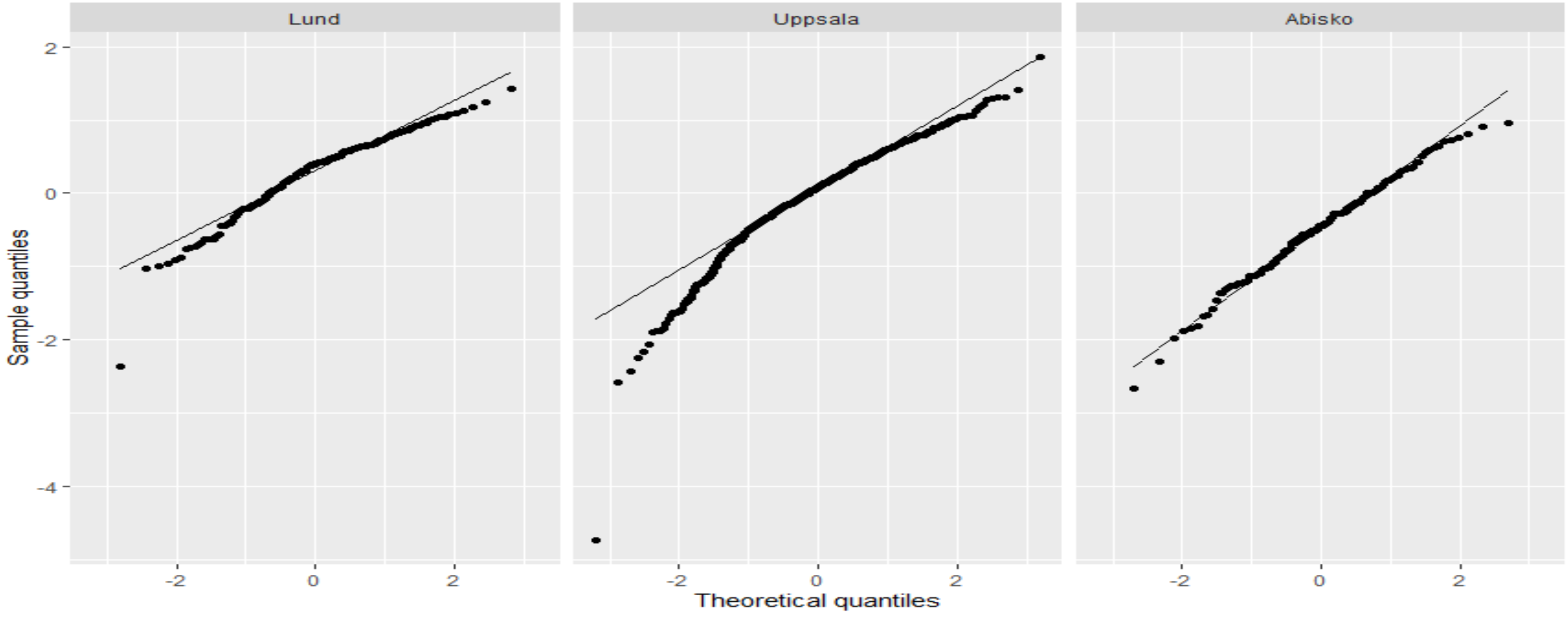


Figure 22: Q-Q-plot of model where log(rain) is a function of temperature and pressure, divided by location

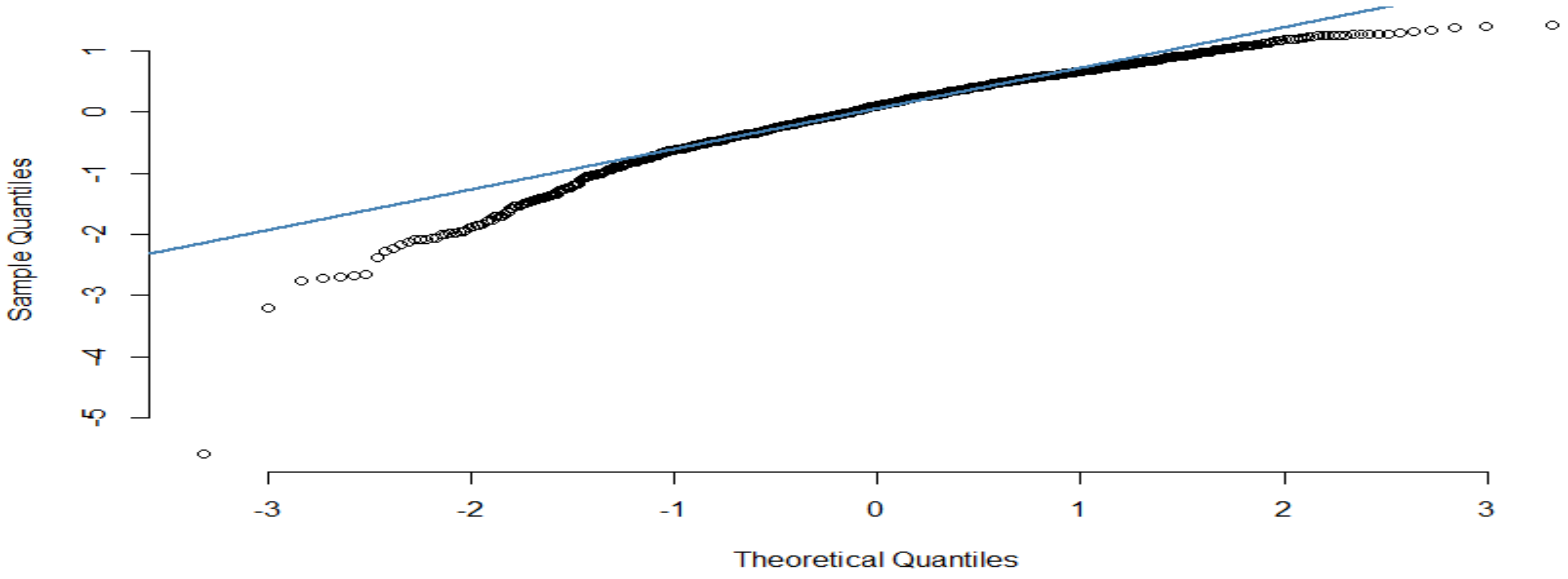


Figure 23: Q-Q-plot of model where log(rain) is a function of temperature

Part 3

In the third part, the goal was to find out which variables that are needed in the model for precipitation. The process started by plotting the leverages for Lund, Uppsala and Abisko against pressure and temperature separately. As both of the plots show, the lowest leverage of Uppsala is lower than those of Lund and Abisko. The reason for this is because of the significant larger amount of observations in Uppsala. A condition for a leverage v_{ii} is that $v_{ii} < 1/c$, where c is the number of identical x-values. Therefore, the large number of observations for Uppsala mean more identical values, which result in a generally lower leverage than those for Lund and Abisko.

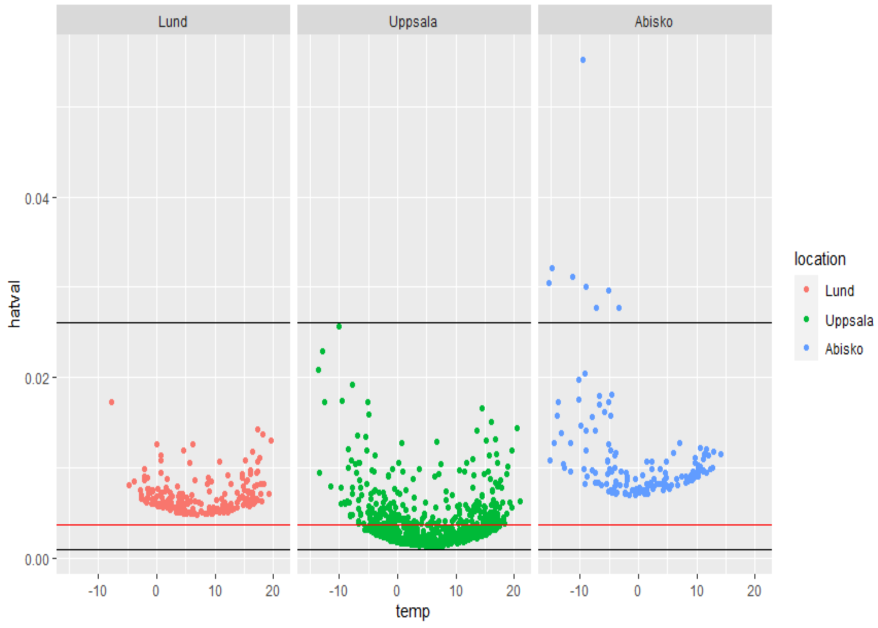


Figure 24: Plot of leverages against pressure

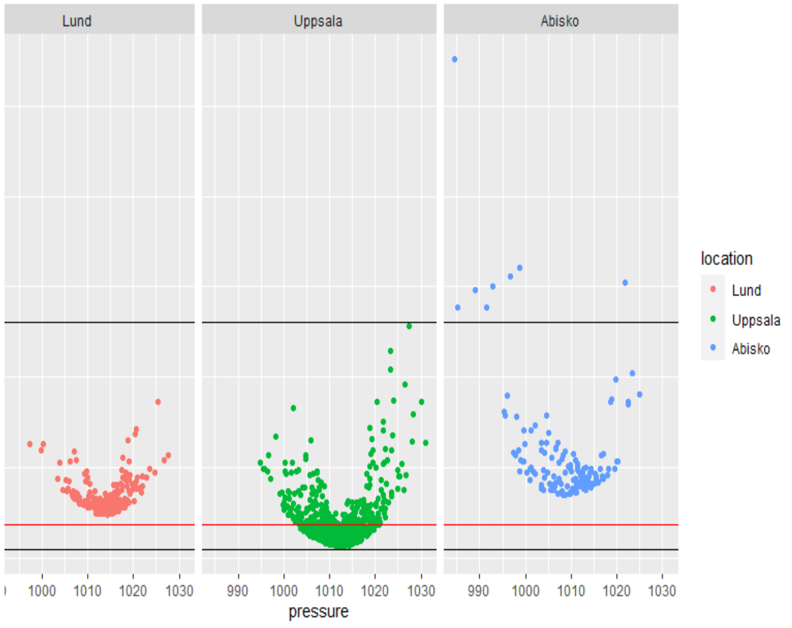


Figure 25: Plot of leverages against pressure

The black horizontal line in figure 24 and figure 25 that intersect the vertical axis at 0,026 helps demonstrate which leverages that are considered unusually high. The points above this limit at 0,026 are highlighted in figure b, where temperature has been plotted against pressure. As the graph illustrates, all of the leverages above 0,026 share the same location. This can be explained by the following condition: $1/n < v_{ii}$, where n is the number of observations. Abisko has the least amount of observations and therefore has the largest leverages.

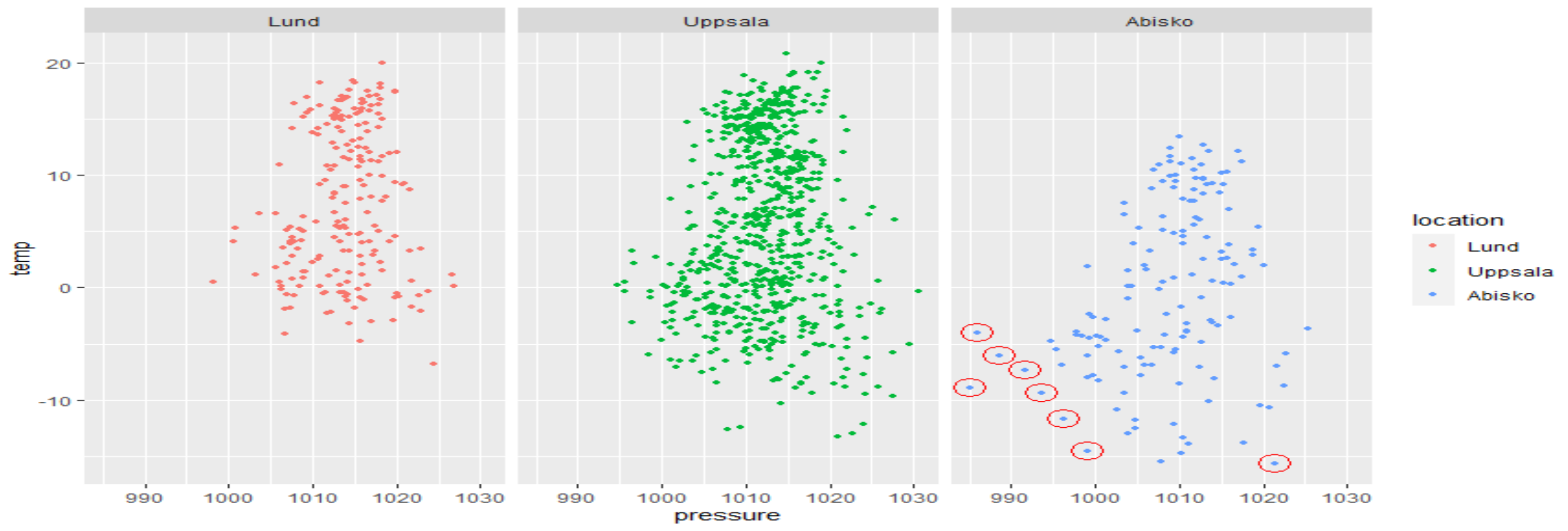


Figure 26: Plot of temperature against pressure separately for each location.

The studentized residuals were then calculated and plotted against the fitted values, as shown in figure 27. A residual is considered large if its absolute value is above 1,96, which means that not all of the highlighted observations residuals is considered problematic. However, there are other studentized residuals that lie beyond this interval. These observations are highlighted in blue in figure 27, while the observations with large leverages are highlighted in red.

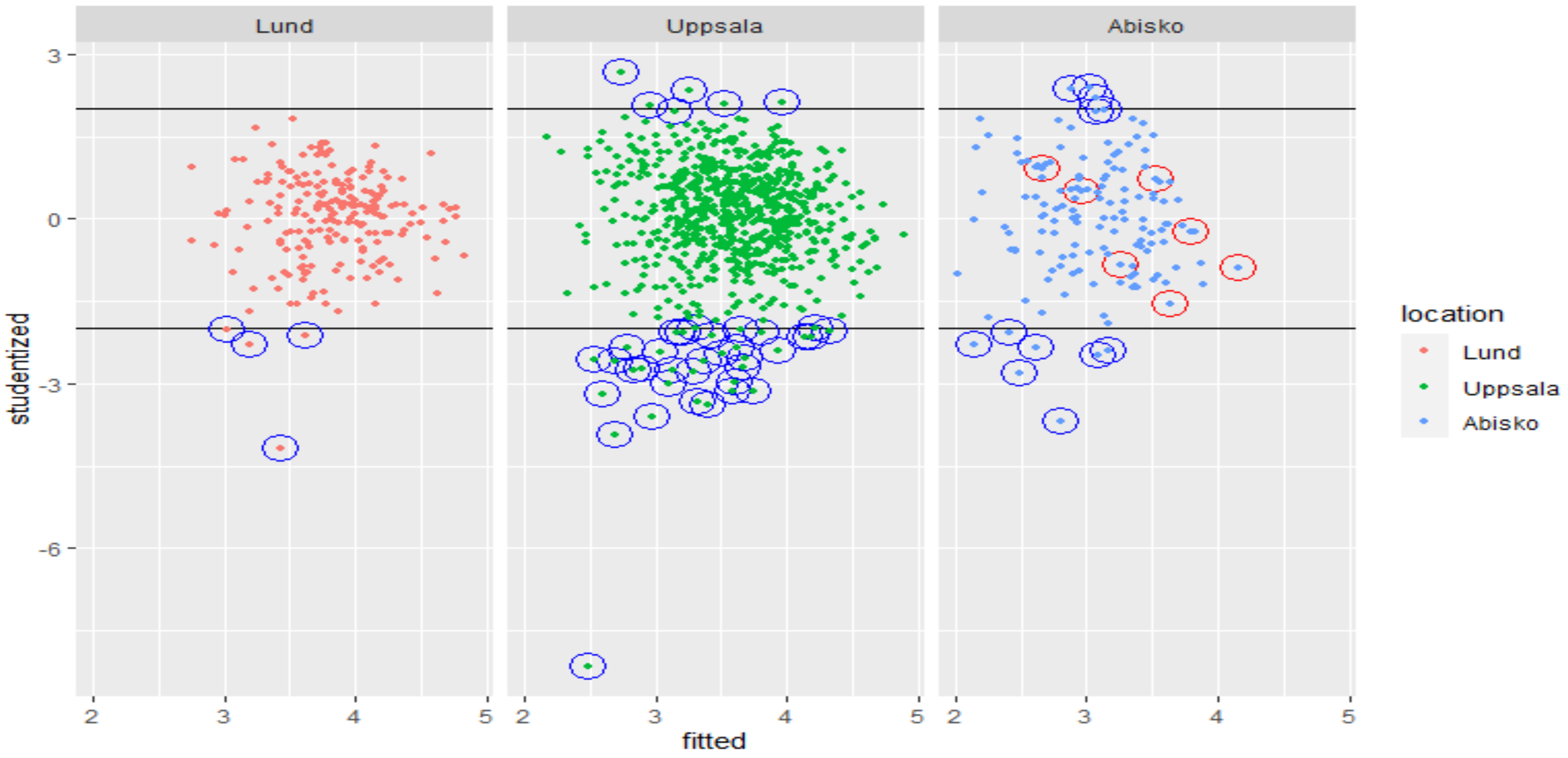
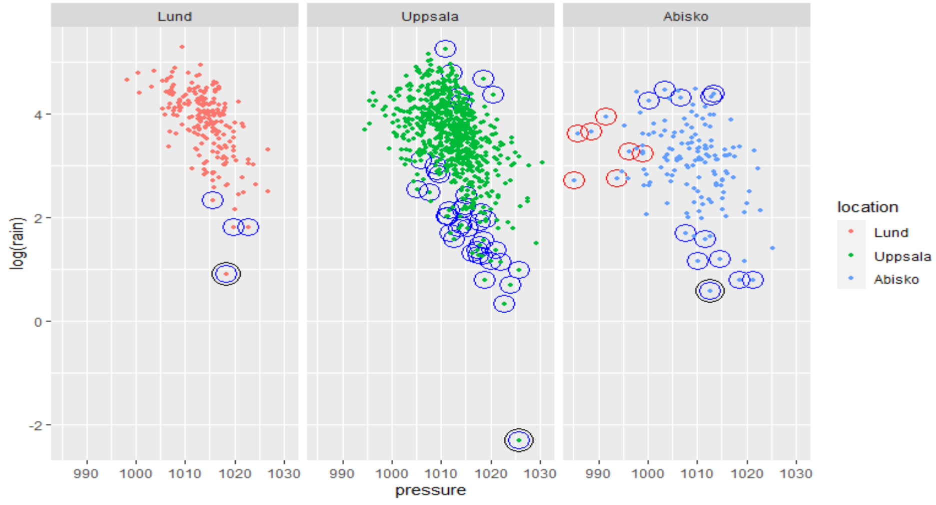


Figure 27: Plot of studentized residuals against fitted values separately for each location.

Then, the largest residual for each location was identified and highlighted together with the observations with the largest residuals in two different plots; logarithmized precipitation against pressure and logarithmized precipitation against temperature separately for each location. The observations with the largest residual are highlighted in black while the observations with large leverages are highlighted in red and observations with large residuals (absolute value greater than 1.96) highlighted in blue.



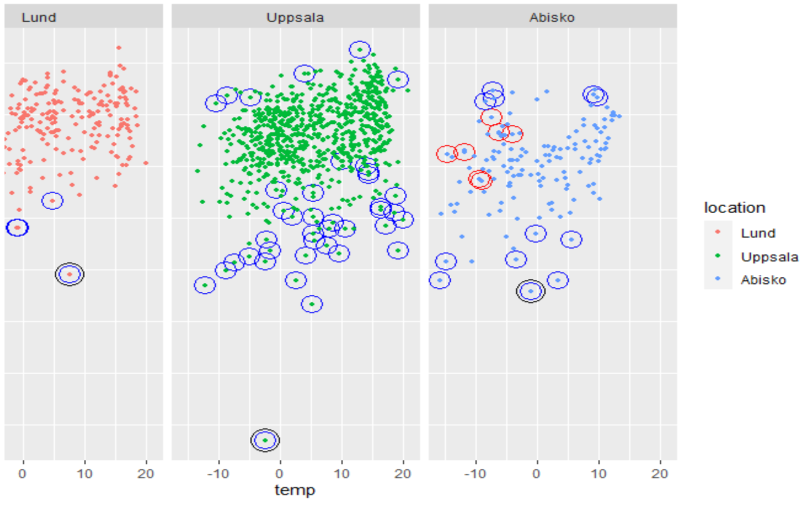


Figure 28: log(rain) against pressure

Figure 29: log(rain) against temperatur.

Afterwards, Cook's D is calculated and plotted in order to identify which of the observations highlighted in figure 28 that have an influence. The highlighted observations that lie beyond the horizontal line have a large Cook's D, thus a large influence on the measurements. As the plot illustrates, the observations with the largest residuals (highlighted in black) all lie beyond the horizontal line and have a large Cook's D, but the same does not apply for the rest of the highlighted observations. Some of them have a small Cook's D, for instance do all except three of the observations with high leverages (highlighted in red) lie below the line. However, of the observations with a large influence almost everyone had large residuals (highlighted in blue). In other words, the observations with large residuals have a larger influence than those with high leverage in general.

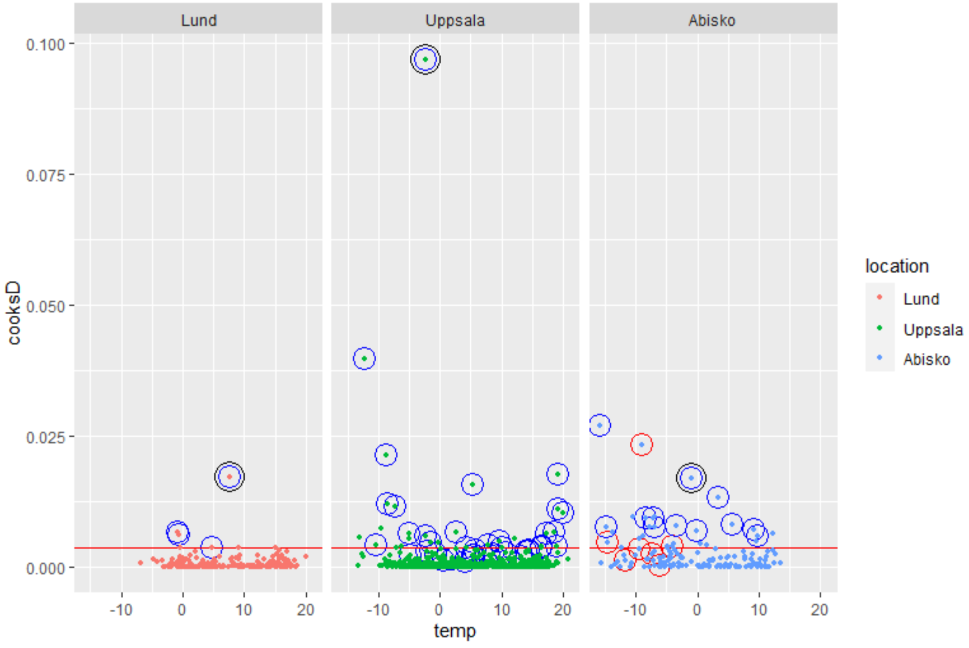


Figure 30: Plots of Cook's D against temperature

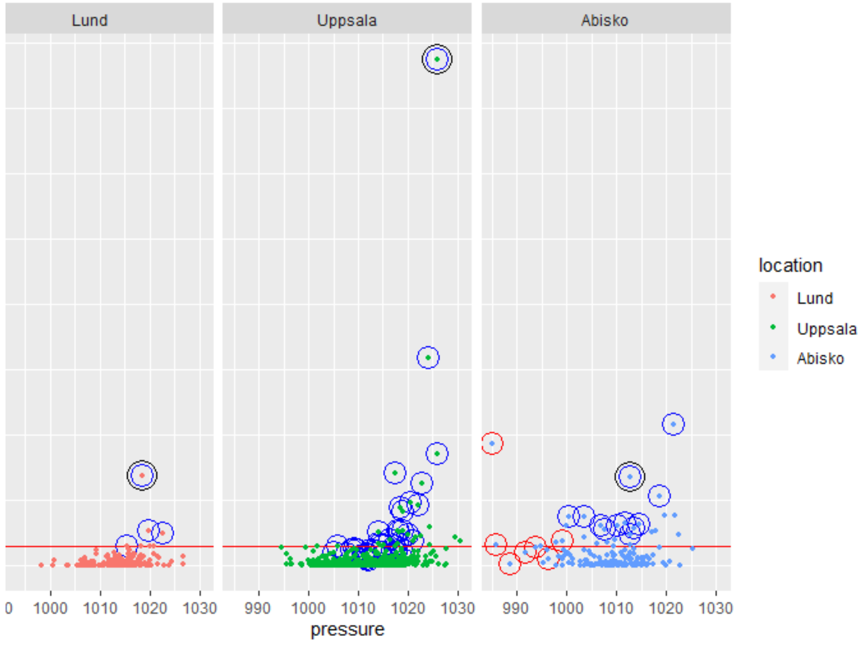


Figure 31: Plots of Cook's D against pressure

A new data set was then created in which the problematic observations with high influence were excluded. The figures from 32 to figure 34 illustrate the difference between the plots with the original data set and the ones with the modified data set.

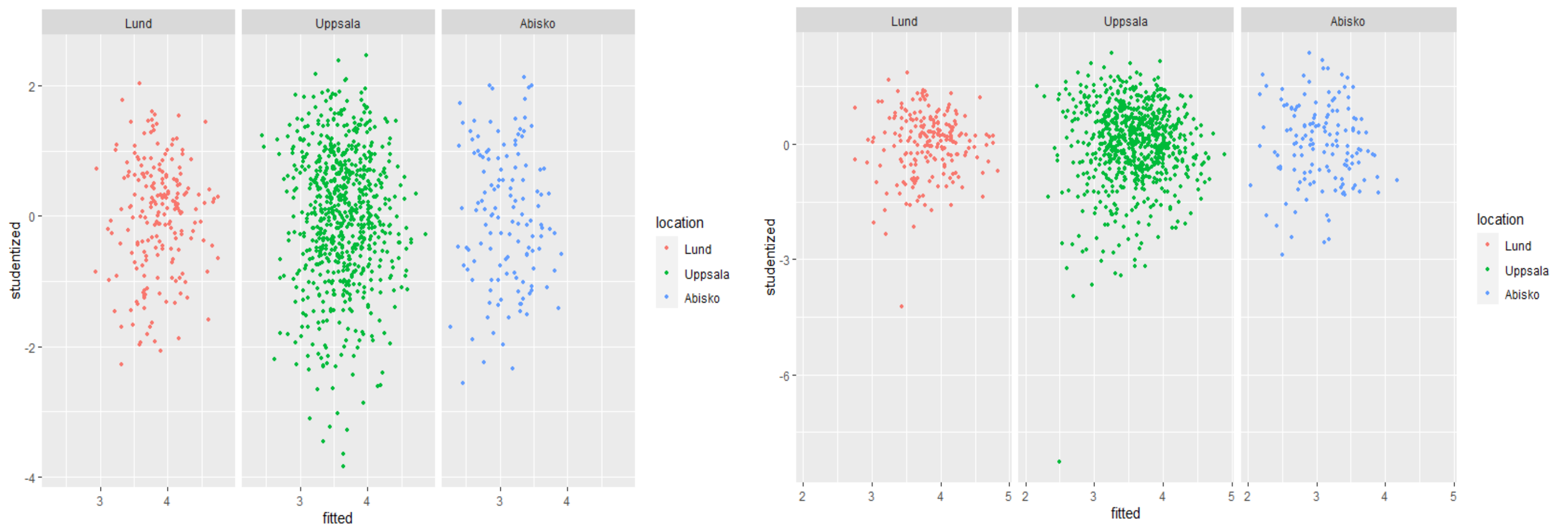


Figure 32: Plots of residuals against fitted values with the new data set (left) and original data set (right)

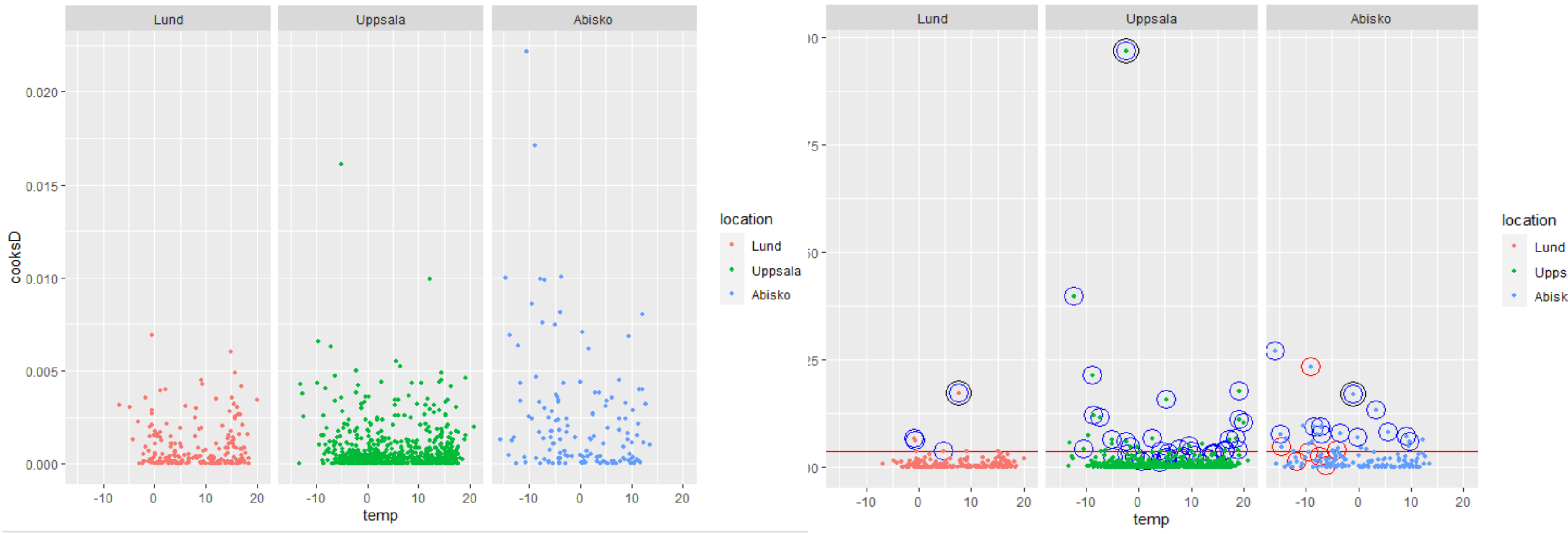


Figure 33: Plots of Cook's D against temp with new data set (left) and original data set (right)

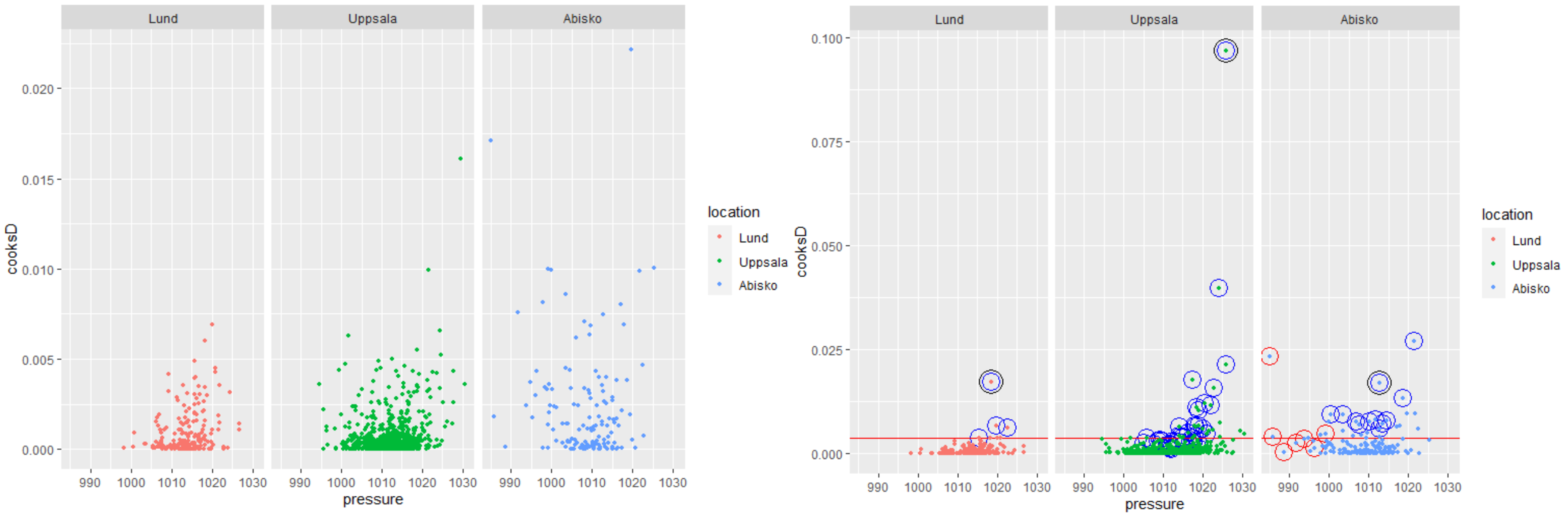


Figure 34: Plots of Cook's D against pressure with new data set (left) and original data set (right)

The models from 1b), 2d) and 2h) were then refitted with the reduced data set. R^2, R^2_{adj}, BIC, AIC was calculated and the result is presented in table 1. As illustrated, the model fitted with temperature*pressure is considered to be the best model due to its low BIC- and AIC-value, and its large R^2- and R^2_{adj}-value.

Model	R^2	R^2_{adj}	BIC	AIC
Pressure	0.1333252	0.1324998	2027.863	2012.988
Temperature	0.2907340	0.2893817	1823.966	1804.132
Temperature* Pressure	0.3334231	0.3315150	1765.621	1740.829

Table 1: R^2, R^2_{adj}, BIC, AIC for the models from 1b), 2d) and 2h).

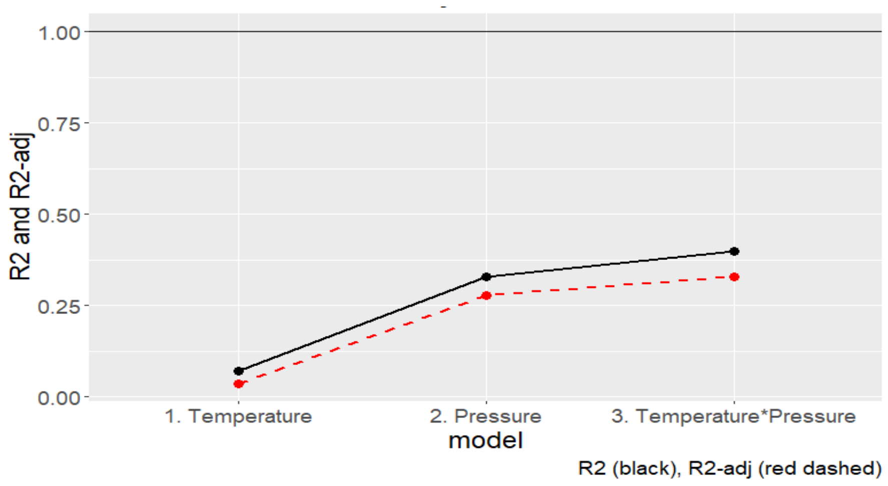


Figure 35 : Plot of R^2, R^2_{adj}

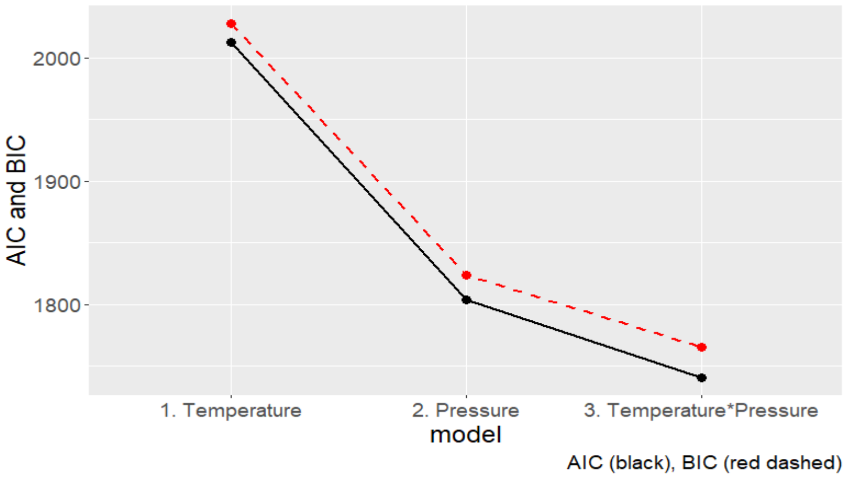


Figure 36: Plot of AIC and BIC

A new model is then fitted where the three-way interaction temp*pressure*location is introduced. Performing an anova test on this new model doesn't show any statistically significant improvement as the F-statistic is 0.711 and the P-value is 0.6508. In order to reduce the AIC value, a backward elimination and forward selection using BIC as criterion was performed. The result of the backward elimination is the same as the one from the forward elimination (table 2). In the forward selection, we started with log(rain) - 1 as the only intercept, and the first variable that was selected to reduce the AIC value the most was temperature. The AIC value was then reduced by 964.54.

Intercept	Temperature	Pressure	Location: Uppsala	Location: Abisko	Location: Lund
59.009420	3.328432	-0.054592	-0.314891	-0.804652	-0.003256

Table 2: Coefficients of the model after backward elimination and forward selection with BIC as criterion.

With the goal to check if there exists any seasonal difference, a new categorical variable was created where the months were grouped into fall, winter, spring and summer. Forward selection was then performed on a four-way interaction model with the variables temp*pressure*location*season. The result is presented in table 3, and as shown does it include seasonal variables. However, even if including these variables in the model result in better approximations of the precipitation, the model could get too complex to work effectively. The coefficients are also quite small and do not affect the approximation of precipitation that much.

Intercept	Temp	Pressure	Temp* Pressure	Location Uppsala	Location Abisko	Season Spring	Season Summer	Season Winter
56.35234	3.051205	-0.05181	-0.00299	-0.31611	-0.84581	-0.32949	-0.03417	-0.13696

Table 3: Coefficients of the model after forward selection with BIC as criterion.

The aim of the project was to fit a model that can predict the monthly average precipitation. When finally choosing our model we have opted to follow the principle of Occam's Razor which states that among several plausible explanations for a phenomenon, the simplest is best. Looking at what results were given from backward elimination and forward selection we see that even if those models include location and season these variables often are very small in comparison. Further on, earlier tests have shown that a model with rain dependent on temperature and pressure is good enough. With this in mind and looking at our results throughout the report we have chosen our final model, with the reduced data set, to be: log(rain) ~ temperature + pressure.