Lund University

Linear and logistic regression

FMSN30/MASM22

Project 1

Authors:
El-Tayeb Bayomi
Nora Lüpkes

May 6, 2020

Contents

Introduction

For the first project in the course *linear and logistic regression*, given in VT 2020 by Anna Lindgren, we will analyze data acquired from the Swedish Meteorological and Hydrological Institute.
The data contains columns such as rain in mm, temperature in °C, pressure in hPa, and locations for each row which is a categorical variable. These are the units that will be used throughout the report.

The goal of this project is to develop and examine what type of model will be the most suitable to model how much it will rain.

The project is split into three parts where which each part the model gets more complex in order to fit better.

1 Precipitation as a function of temperature

For the first part, we will fit a model using only rain and temperature and ignore all other information.

In figure 1, we can see the raw, untouched data.

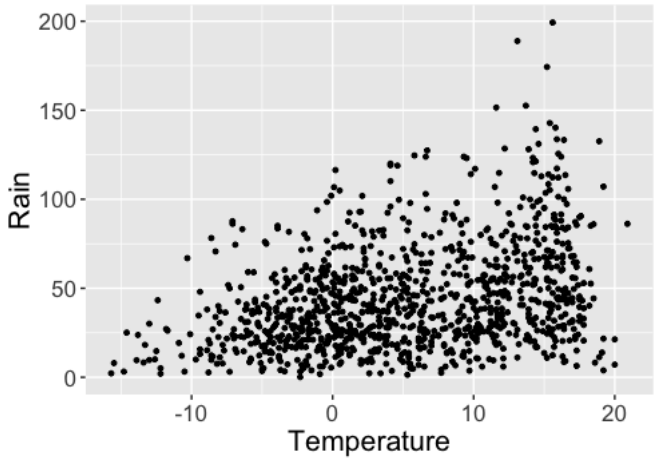


Figure 1: Untransformed data, rain vs. temperature

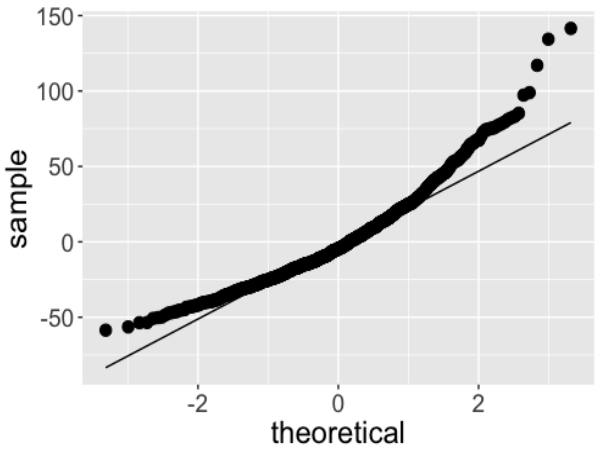
One can already guess that this dataset may not be modeled sufficiently by a simple linear model.

Without any transformations, we fit a linear model with the following code in R:
`Model 1 <- lm(rain ~ temp, data = weather)` .
We get the linear model given by equation (1).

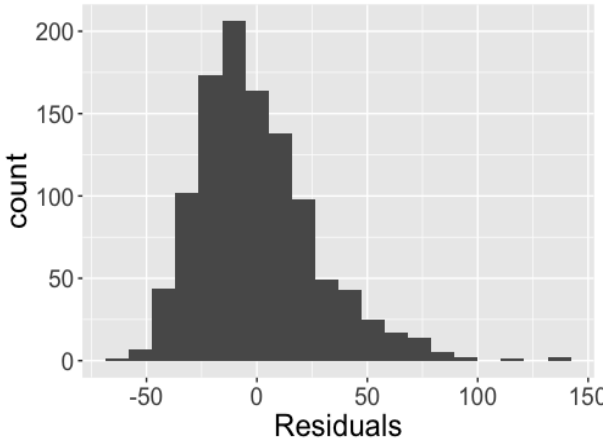
$$y = \beta_0 + \beta_1 x_1 = 37.49983 + 1.301149 x_1 \tag{1}$$

The 95 % confidence intervals for the variables, are $\beta_0 = (35.5462, 39.4533)$ and $\beta_1 = (1.0933, 1.5089)$

For the basic residual analysis, we look at the QQ-plot, seen in figure 2a and the histogram, seen in figure 2b.



(a) QQ-plot of first model



(b) Histogram of first model

Figure 2: Residual analysis of model 1

Clearly, the untransformed model is horrible. The QQ-plot is not Normal distributed. So the first transformation we do is the following:

```
Model 2 <- lm(log(rain) ~ temp, data = weather)
```

Its beta estimates:

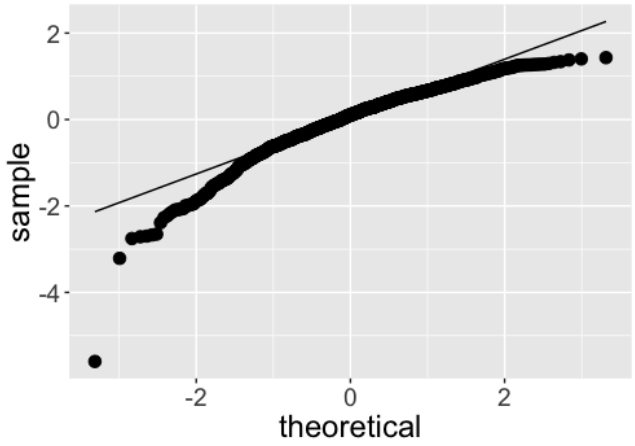
$$\log(y) = \beta_0 + \beta_1 x_1 = 3.371602 + 0.03337392x$$
$$\Rightarrow \quad y = \exp(\beta_0) \cdot \exp(\beta_1 x_1) = 29.12513 \cdot 1.033937^{x_1}$$

According to the model given by Equation (2), when we increase the tem-

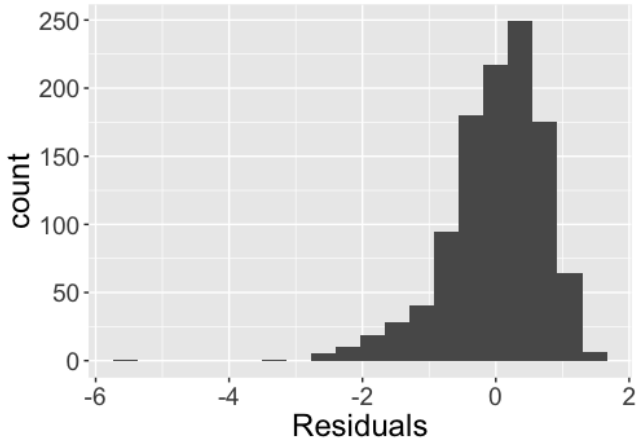
perature by $1°C$, the precipitation changes by 3%.

The 95% confidence intervals are (3.31871385, 3.42448920) for β_0 and (0.02774783, 0.03900001) for β_1.

To judge its quality, we look at the QQ-plot again, seen in figure 3a, and the histogram, seen in figure 3b.



(a) QQ-plot of second model



(b) Histogram of second model

Figure 3: Residual analysis of model 2

Although both are not ideal, the second model looks slightly better (but still not optimal). It did fix the left wing behavior, but it overdid it which results

in a right wing behavior. One can conclude that the transformation was a step in the right direction, but there is more that can be done to improve the results.

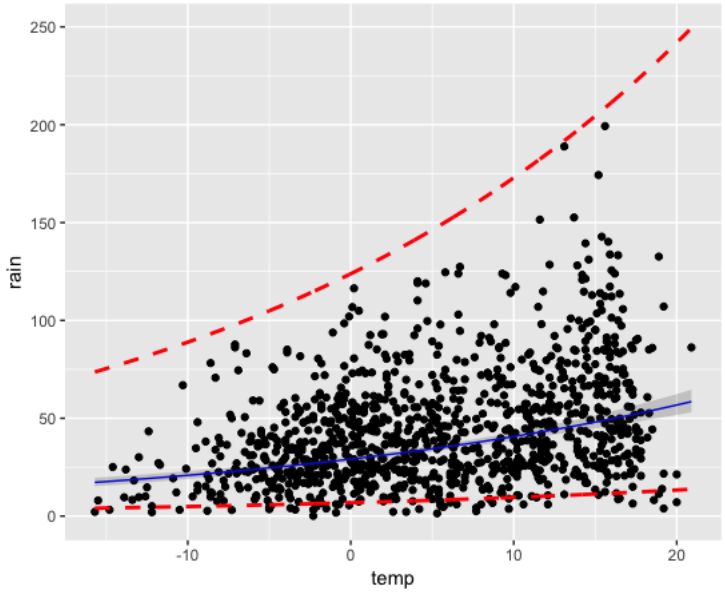


Figure 4: Prediction and confidence intervals for log-transformed model on untransformed data.

For months where the average temperature is 5°C, the total monthly amount of precipitation is expected to decrease by 6% and to be inside the prediction interval (8.096201, 146.2836).

2 Precipitation as a function of temperature and more

For the second part, we will add the variable pressure and see whether that gives us a better model than part 1.

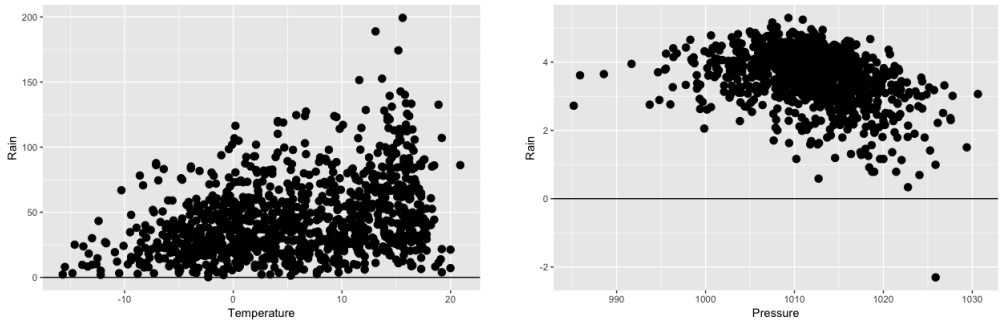
2.1 Temperature again

To see if the temperature has a significant effect on the amount of precipitation, we look at it's p-value and compare it to 0.05:

$$2.2e^{-16} << 0.05 \tag{2}$$

So temperature has a significant effect.

2.2 Temperature and pressure



(a) Temperature vs. rain (b) Rain vs. pressure

Figure 5: Nonlinear parts of data set

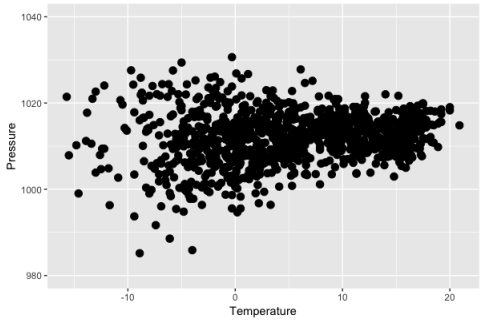


Figure 6: Pressure vs. temperature

Neither data shown in figure 5a nor figure 5b look linear.

The data shown in figure 6 does look like there may be a strong linear relationship. So to take care of that, we form the following model:
`Model 3 <- lm(log(rain) ~ temp + pressure, data = weather)`

This gives the following expression

$$log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 = 61.9026 + 0.0404 x_1 - 0.0578 x_2 \qquad (3)$$

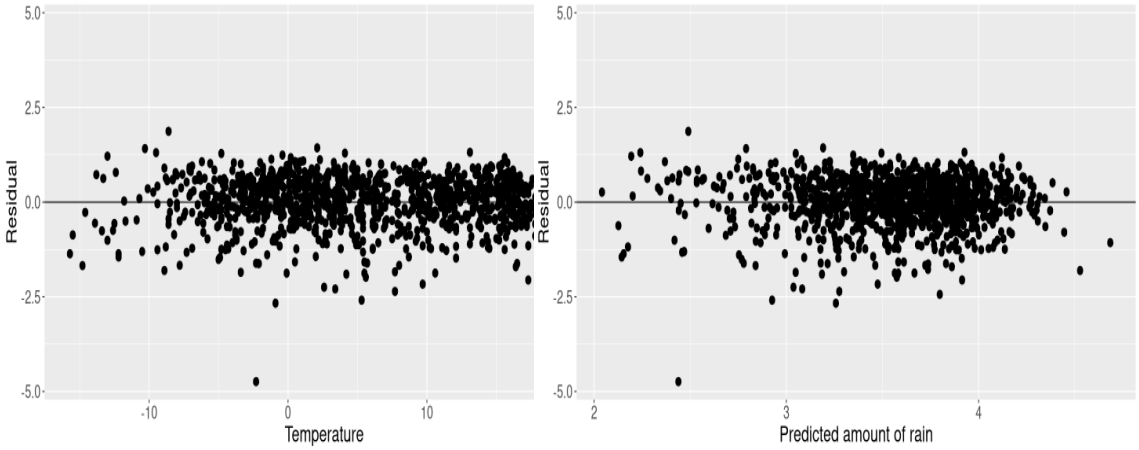
The 95 % confidence intervals are $\beta_0 = (54.9210, 68.8841)$, $\beta_1 = (0.0353, 0.0455)$ and $\beta_2 = (-0.0647, -0.0509)$

We look at the p-value : $2e^{-16} << 0.05$.
Also 0 is not inside the confidence intervals, so both temp and pressure are statistically significant. The figures 7a, 7b, and 8a show the residuals plotted versus temperature, the anticipated amount of rain and pressure for `model 3`.

The QQ-plot of `model 3` (seen in figure 8b) looks a bit better than the QQ-plot for the second model shown in 3a, but it is still not as good as we would like.

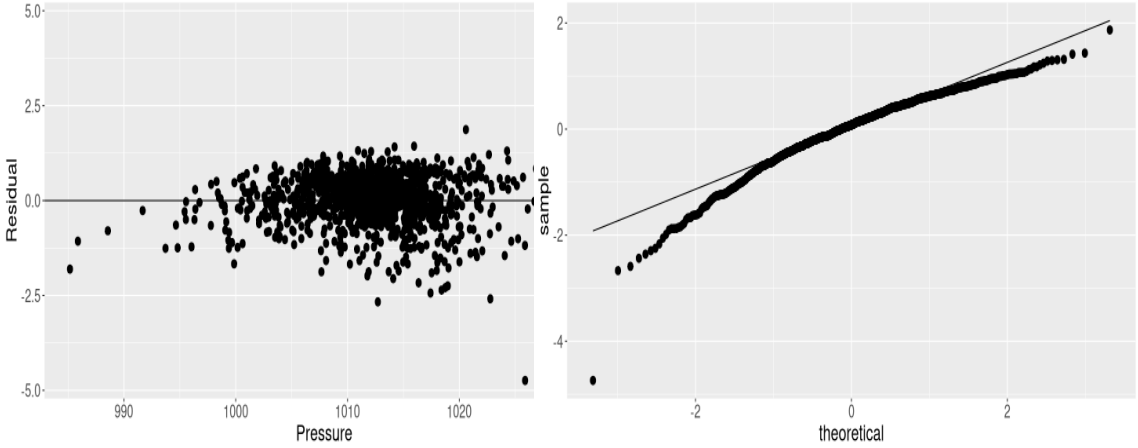
According to the model given by Equation (3), when we increase the temperature by 1°C, the precipitation changes by 4%. The first model given by Equation (2) had a decrease.

Changing the same model, changing the pressure by 20 hPa gives a 69 % decrease in precipitation.



(a) Residuals vs. temperature (b) Residuals vs. amount of rain

Figure 7: Residual analysis I



(a) Residuals vs. pressure (b) Normal QQ-plot for residuals

Figure 8: Residual analysis II

For months when the average temperature is 5 ° C and the pressure is 1000 hPa, the total monthly amount of precipitation is estimated to be 61.12348 with the prediction interval of (17.11899, 218.24188).

Increasing the pressure by 20 hPa, the total monthly amount of precipitation is estimated to be 18.33645 instead with the prediction interval of

(5.14136 65.39617).
One can see that there is quite a difference between the estimation for 1000 hPa and 1020 hPa.
After the results from 2f) above, this is not surprising as
61.12348·0.31 = 18.94828. So the results are in line.

2.3 Temperature and pressure with interaction

Changing the way that temperature and pressure interact we get a new model:

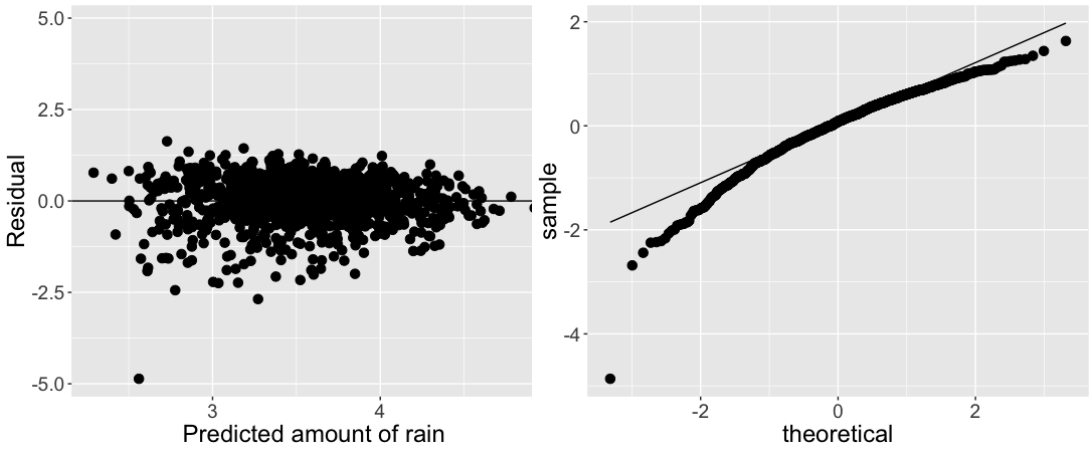
```
model 4 <- lm(log(rain) ~ temp * pressure, data = weather)
```

$$\begin{aligned} log(y) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_1 \beta_2 x_1 x_2 \\ &= 61.041369 + 3.271798 x_1 - 0.057009 x_2 - 0.003191 x_1 x_2 \end{aligned} \tag{4}$$

The 95 % confidence intervals are $\beta_0 = (54.1924, 67.8903)$, $\beta_1 = (2.3273, 4.2162)$, $\beta_2 = (-0.0637, -0.0502)$ and $\beta_{1,2} = (-0.0041, -0.0022)$

We look at the p-value : $3e^{-11} << 0.05$.
Also 0 is not inside the confidence intervals, so the interaction term between temp and pressure is statistically significant.

When we compare the different β-values from the model without the interaction term with the model with interaction, we can see that it changed substantially. This might be because we normalized the pressure by subtracting 1012.



(a) Residuals vs fitted y-values. (b) Normal QQ-plot for residuals

Figure 9: Residuals analysis

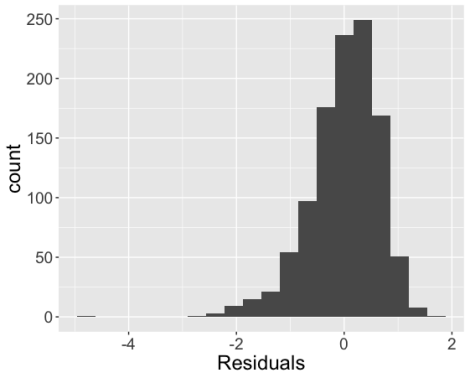


Figure 10: Histogram of 2 j) model.

Right now, we have the following model:

$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

If we increase the temperature, x_1, by 1°C, we get:

$$\log(y) = \beta_0 + \beta_1 (x_1 + 1) + \beta_2 x_2 + \beta_3 (x_1 + 1) x_2$$

So for the change we subtract the equations and get:

$$\log(y') - \log(y) = 0 + \beta_1 + 0 + \beta_3 x_2$$

If we plug in a pressure of 1000 hPa, we get that the rain will increase by 8.4%. If we plug in the higher pressure, the rain increases very slightly by 1.7%.

If the average temperature is -10°C and the average pressure is 1000 hPa, then the estimated precipitation is 25.12955 with the confidence interval (6.990012, 90.34240). If the average pressure is 1020 hPa, then the estimated precipitation is 21.36500 with confidence interval (5.982919, 76.29442).

Now, if we look at a positive temperature of 10°C on average and the average pressure is 1000 hPa, then the estimated precipitation is 126.48684 with a confidence interval of (35.214375, 454.32929). If we deal with an average pressure of 1020 hPa, the estimated precipitation is 15.21159 with a confidence interval of (4.245308, 54.50546).

2.4 Temperature, pressure, and location

After using the `summary()` function in R it was decided that *Uppsala* should be chosen as a reference variable for the new categorical variable; it has the most observations (738 observations compared to Lund´s 207 and Abisko´s 146). We form the following model :

```
Model 4 <- lm(log(rain) ~ temp * pressure + location, data = weather)
```

This gives the following expression where α is the categorical variables for both Lund and Abisko respectively:

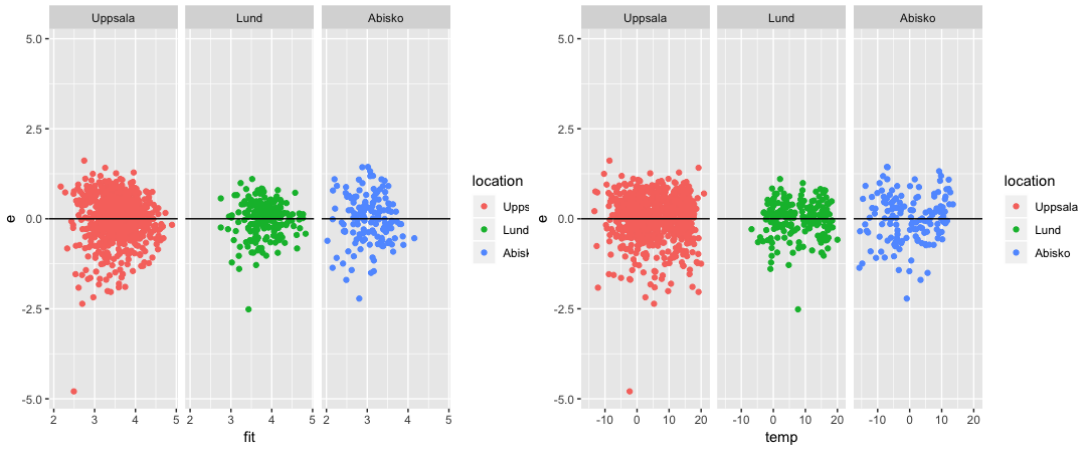
$$\begin{aligned} log(y) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_1\beta_2 x_1 x_2 + \alpha_4 x_4 + \alpha_5 x_5 \\ &= 70.041775 + 3.015263 x_1 - 0.065860 x_2 - \\ &\quad 0.002945 x_1 x_2 + 0.340686 x_4 - 0.511277 x_5 \end{aligned} \tag{5}$$

The confidence intervals for the different variables above are:

$\beta_0 = (63.4419, 76.6416)$, $\beta_1 = (2.1136, 3.9168)$, $\beta_2 = (-0.0723, -0.0593)$, $\beta_{1,2} = (-0.0038, -0.0020)$, $\alpha_1 = (0.2451, 0.4361)$ and $\alpha_2 = (-0.6259, -0.3965)$

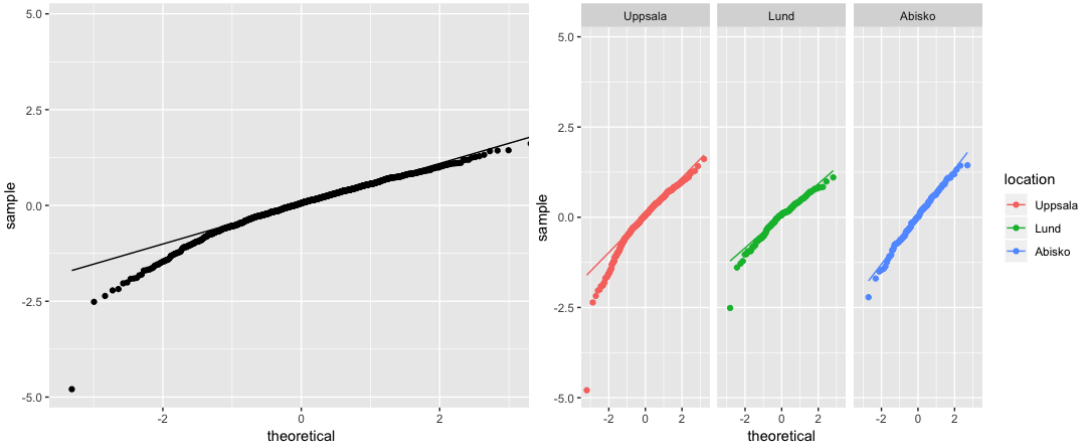
If one compares (5) and (6), one can see that the parameters change a bit but not too drastically.

In figure 11, the residuals are shown.



(a) Residuals vs fitted y-values (b) Residuals vs temperature

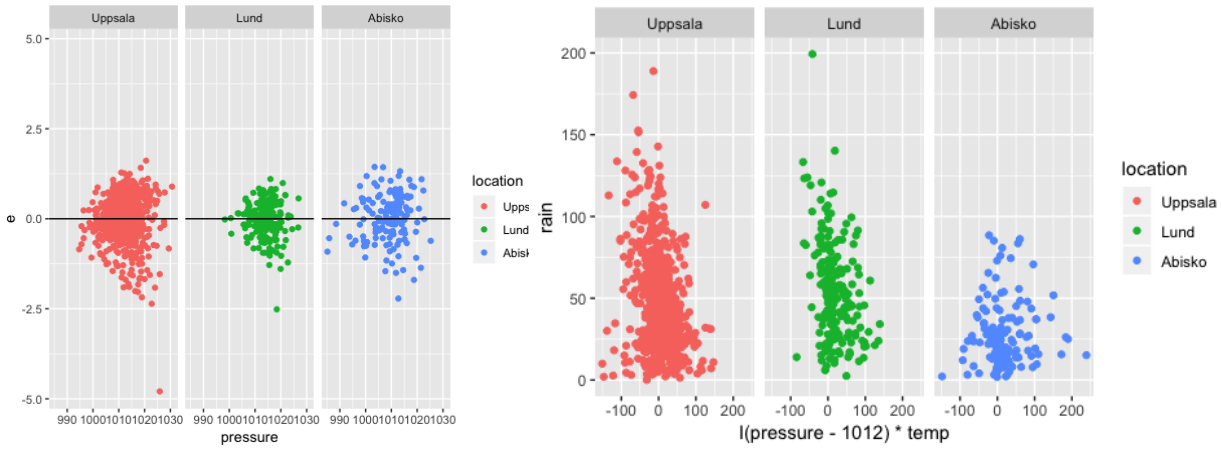
Figure 11: Residuals for each location



(a) QQ-plot for all residuals. (b) QQ-plots for different locations

Figure 12: QQ-plots for model 4

As seen in figure 12a, the QQ-plot got a lot better at the end. There is not much deviation from the desired line. The beginning of the tail is still off. To investigate further we separated the different locations as seen in figure 12b and it becomes clear that Uppsala causes the most problems. Lund and Abisko can be modeled relatively fine.



(a) Residuals vs pressure

(b) Rain vs pressure and temperature for different locations

Figure 13: Residuals for model 4

The location with the highest expected value of rain is Lund as can be seen from Figure 13b.

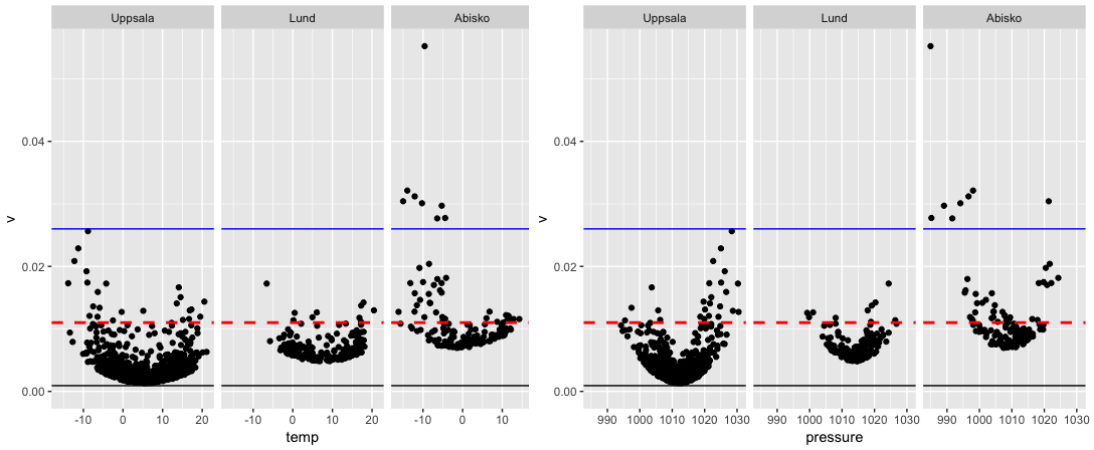
3 Which variables are needed?

For the third and last section, we will have a look at outliers that might have corrupted the models so far. Apart from that, we will determine how many variables are needed and which are not needed for our model.

3.1 Outliers and influential observations

In the following section, we identify outliers and problematic points to make our model even better.

Using model 4 from 2 n), we calculate the leverages and plot them against temperature and against pressure below. The blue line is at 0.026 on the y-axis.

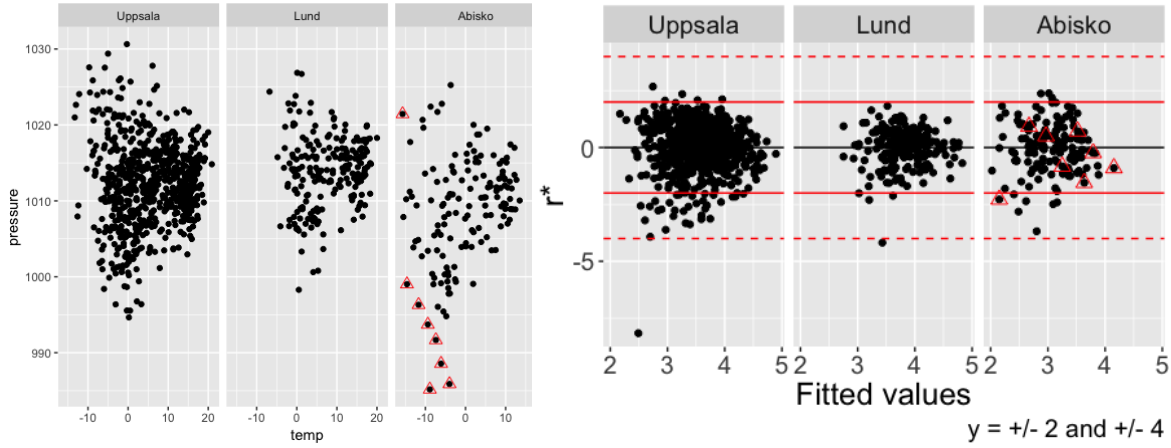


(a) Leverage vs temperature (b) Leverage vs pressure

Figure 14: Leverages for different locations

Uppsala has the largest amount of observations. That means that the models fits better for the weather recorded in Uppsala and has difficulties fitting the Weather in Lund and Abisko. That is the reason why their leverages are significantly higher.

When plotting the datapoints that exceed a leverage value of 0.026 in temperature vs pressure we get figure 15.



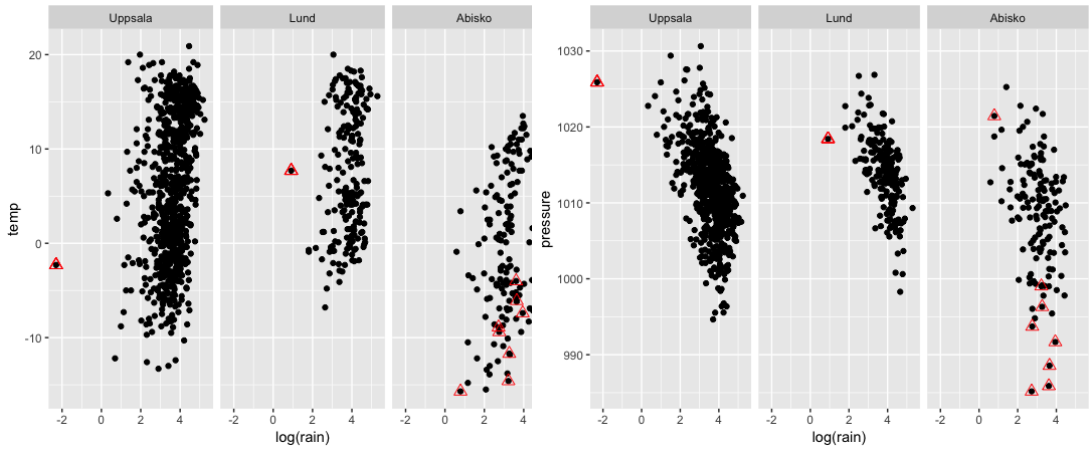
(a) temp vs. pressure for different locations

(b) Studentized residuals vs. fitted values for different locations

Figure 15: Data with highlighted problematic leverages for different locations

One can see in figure 14, that both for temperature and pressure Abisko has very high leverages. If we translate that back to the original data seen in figure 15a, the problematic samples are the one with little rain and little pressure.

Next, we look at the studentized residuals seen in figure 16.



(a) log(rain) vs. temp for different locations

(b) log(rain) vs. pressure for different locations

Figure 16: High leverage and studentized residuals highlighted for various data parts

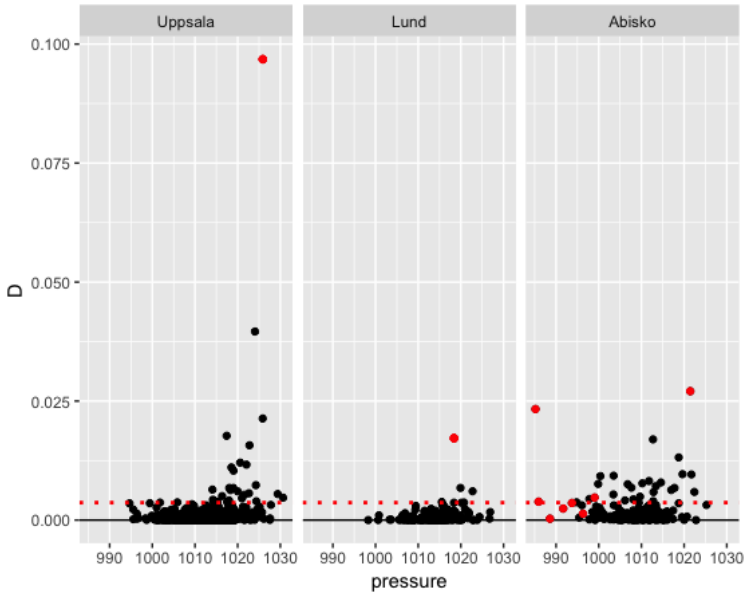
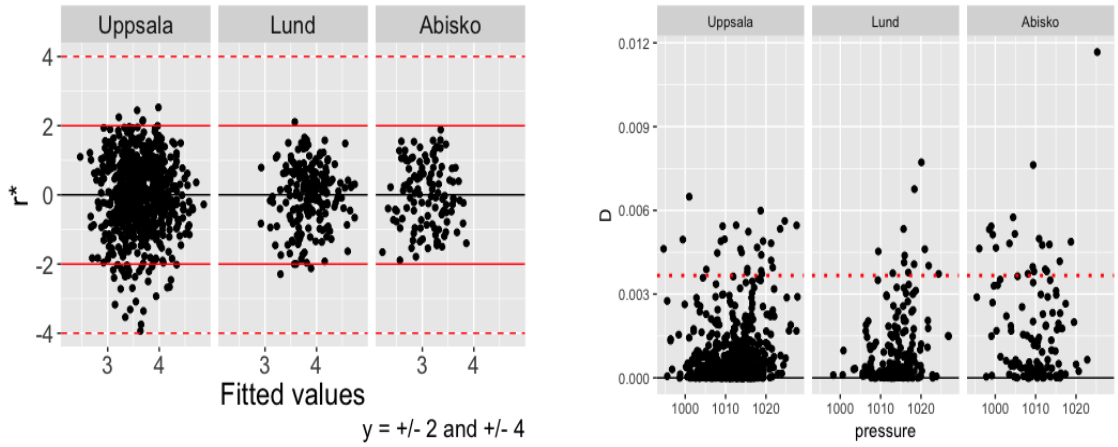


Figure 17: High leverage-values and studentized residuals highlighted for different locations in relation to Cook's D

In figure 17 one can see the problematic points marked in red. It is noteworthy that not all of the points with high Cook's distance were problematic before and vice versa.

Therefore, we want to exclude all red points that are above the threshold marked by the red dotted line and refit the model. The model itself and the corresponding equations remain untouched, we just remove problematic observations.

The new studentized residuals and Cook's D after removing the troublesome observations can be seen in the figures 18a and 18b.



(a) Studentized residuals of trimmed dataset (b) Cook's D of trimmed dataset

Figure 18: Weather data without the problematic datapoints

Figure 18 shows that there are no troublesome studentized residuals anymore which is an improvement. Because we chose not to eliminate all points with high Cook's distance, there are still some high ones that we have already seen before.

3.2 Model comparisons

Modelnumber	AIC	BIC	R^2	$R^2_{adjusted}$
Model 1	10058.589	10073.507	0.1166	0.1157
Model 2	2147.606	2167.497	0.2886	0.2872
Model 3	2105.808	2130.671	0.3172	0.3152
Model 4	1985.251	2020.059	0.3924	0.3895

Table 1: AIC, BIC, R^2 and $R^2_{adjusted}$ values for different models

AIC and BIC are a single number score that can be used to determine which of multiple models is most likely to be the best model for our weather dataset. A lower AIC/BIC score is better. Unlike the AIC, the BIC penalizes the model more for its complexity, meaning that more complex models will have a worse (larger) score and will, in turn, be less likely to be selected. That's one reason for why BIC is larger than AIC for Model 4.

From table 1 we can see that Model 4 has the lowest value for AIC and BIC relative to the other models.

R^2 is the percentage of the response variables variations that are explained by our linear models; the higher the R^2, the better the model fits the data.

The $R^2_{adjusted}$ is a modified version of R^2 that has been adjusted for the number of predictors in the model. $R^2_{adjusted}$ increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance.

Looking at table 1 we can see that Model 4 increases both R^2 and $R^2_{adjusted}$. This means that Model 4 is the best compared to the previous models.

By fitting a model with the three-way interaction $temp \cdot pressure \cdot location$ instead of just adding location we get 12 variables. This is far too much and we have overfitted our model. This can be seen when running the function summary() in R and we see that most variables that were added due to the interaction term *location was not statistically significant from zero. In fact we only end up with the original four variables that are deemed significant enough.

We now perform a backward elimination, using BIC as criterion in order to

reduce the number of parameters used in the overfitted model, instead of just looking at what's not significant from zero using the summary() function.

After using the step()-function we ended up with the following model (same as previous Model 4):

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_1 x_1 \beta_2 x_2 + \alpha_1 x_3 + \alpha_2 x_4 \\ &= 59.5653 + 3.2302 x_1 - 0.0554 x_2 - 0.0031 x_1 x_2 + 0.3110 x_3 - 0.5107 x_4 \end{aligned} \tag{6}$$

Moving on to performing a forward selection, using BIC as criterion, the first variable to be selected was temperature. This means that the variable that has the biggest effect on how much it should rain is deemed to be the temperature; which seems to be reasonable. The final model that we end up with after the forward selection has been run is the exact same model as when we performed the backwards elimination.

The final model after adding the categorical variable season, is:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_1 x_1 \beta_2 x_2 + \alpha_1 x_3 + \alpha_2 x_4 + \gamma_1 x_5 + \gamma_2 x_6 \\ &= 56.9922 + 2.9887 x_1 - 0.0527 x_2 - 0.0029 x_1 x_2 + 0.3138 x_3 \\ &\quad - 0.5531 x_4 - 0.3029 x_5 - 0.1453 x_6 \end{aligned} \tag{7}$$

Where β's are the coefficients for temperature and pressure, α's are the categorical variables for location and γ's are the categorical variables for season.

After running summary on this final model, the only thing that was not statistically significant were the seasons summer and autumn. What that means is that during summer/autumn, the rain depends more on other variables than on the current season. However when it's winter or spring the current season has a significant effect on how much it rains.

The R^2 and R^2_{adj} values are 0.4951128 and 0.4707996 respectively. That means that our model only explains less then 50% of the variability. This is not bad, but also not as good as we hoped in the beginning.

4 Conclusion

After all, we end up with a fairly good model which is presented in equation (7). It's the best regarding forward and backward elimination via BIC. One must keep in mind that there were still some points with high Cook's distance. In order to improve the model further and explain more than the achieved 50% explanation of variability, one could eliminate more of the problematic points and refit. But this is left for further studies; outside of the scope of this project.