Lund University

Linear and logistic regression

FMSN30

Project 2

Authors:

El-Tayeb Bayomi

Nora Lüpkes

May 13, 2020

Contents

Introduction

For the second project in the course linear and logistic regression, given in VT 2020 by Anna Lindgren, we will continue our analysis on data acquired from the Swedish Meteorological and Hydrological Institute. This is a continuation on the first project in the course

The data contains columns with 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 second project is to explore the possibilities of logistic regression compared to linear regression for our weather-data. Primarily, we want to evaluate the probability of having low amounts of precipitation dependent on different variables.

1 The null model

a) We want to estimate the probability of low precipitation (i.e. rain) during a month. We define low precipitation as a total monthly precipitation that lies below 25 mm. We create a new variable **lowrain** by writing the following in R:

```
weather$lowrain <- as.numeric(weather$rain < 25)
```

Now we proceed to calculate the proportion of the months that have low precipitation to estimate the odds of having low rain. This was done by writing the following in R: `mean(weather$lowrain)`

b) We fit a general linear model with lowrain as our Y-variable with only an intercept β_0.

$$\text{Model 0: } Y(= lowrain) = 1 \tag{1}$$

The β-estimate was 0.2868, with a 95 % confidence interval of (-1.0429, -0.7805).

The odds estimate of having low rain is 0.4023136, with a 95 % confidence interval of (0.3524, 0.4581).

Using the estimated odds to estimate the probability of a month having low precipitation is $\approx$ 29% with a 95 % confidence interval of (0.2605739, 0.3142102).

The odds and probabilities relate to one another by the following expression:

$$\begin{aligned} \beta_0 &= \text{log-odds} \\ e^{\beta_0} &= \text{odds} \\ p = \frac{e^{x_i\beta}}{1+e^{x_i\beta}} &= \text{probability} \end{aligned} \tag{2}$$

2 Temperature...

a)

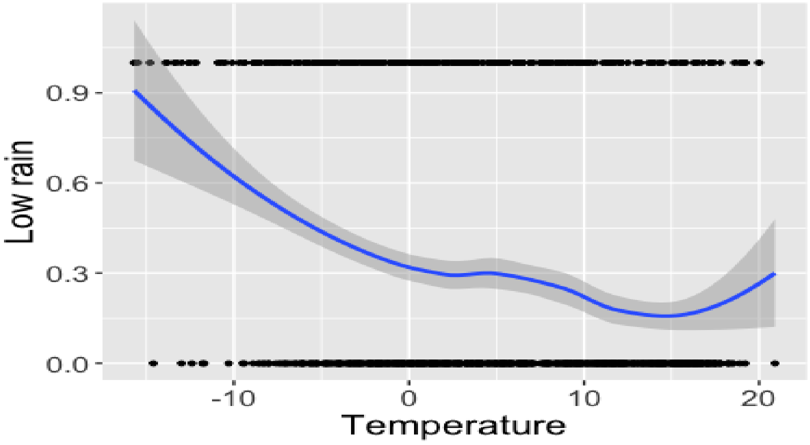


Figure 1: Moving average of lowrain vs temp

We plot lowrain against temp as in figure 1 together with the moving average of lowrain to get a rough estimate of the shape of the probability. It seems reasonable to first try using temperature as a covariate. Because physically, when the temperature decreases then it is less likely to rain; and when it is hotter it is more likely to rain.

b) Now we fit a general linear model using temperature as a covariate.

$$\text{Model 1: } Y(= lowrain) = temp \tag{3}$$

The β-estimates and their corresponding confidence intervals are: $\beta_0 = 0.5886$, (-0.7389, -0.4394) and $\beta_1 = -0.0739$, (-0.0921, -0.0559)

We can see from the confidence intervals above that the variable for the temperature covariate is statistically significant because it does not cover 0.

The odds ratio were $e^{\beta_0} = 0.5551$ with a 95 % confidence interval of (0.4776, 0.6443). $e^{\beta_1} = 0.9288$ with a 95 % confidence interval of (0.9119, 0.9455).

The odds decrease by 7% if the temperature increases by 1°C and increase by 7% if temp decreases by 1°C.

c) Using the model from **2 b)** the predicted probability for low rain when the temperature is -10°C $= 0.5374333 \approx 54$ % with a 95 % confidence interval of (0.2580971, 0.8167695).

For -9°C $= 0.5190288 \approx 52$ % with a 95 % confidence interval of (0.2553565, 0.7827011).

For 9°C $= 0.2855435 \approx 29$ % with a 95 % confidence interval of (0.05410484, 0.39013296).

For $+10$°C $= 0.2096192 \approx 21$ % with a 95 % confidence interval of (0.03032595 0.38891245).

d)

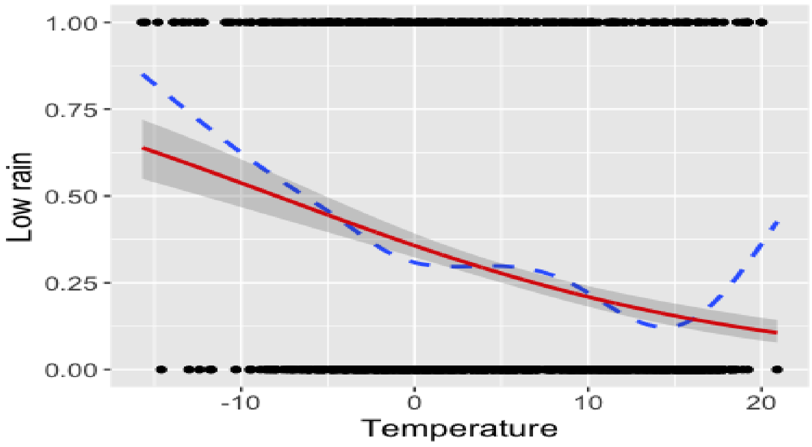


Figure 2: The predicted probabilities with 95 % confidence interval together with MA from 2 a)

The blue, dashed line in figure 2 represents the moving average and the red is the fitted line inside the 95% confidence interval. One can see that the is a larger margin for low temperatures then for high temperatures. So the model is a bit better for higher temperatures.

e)

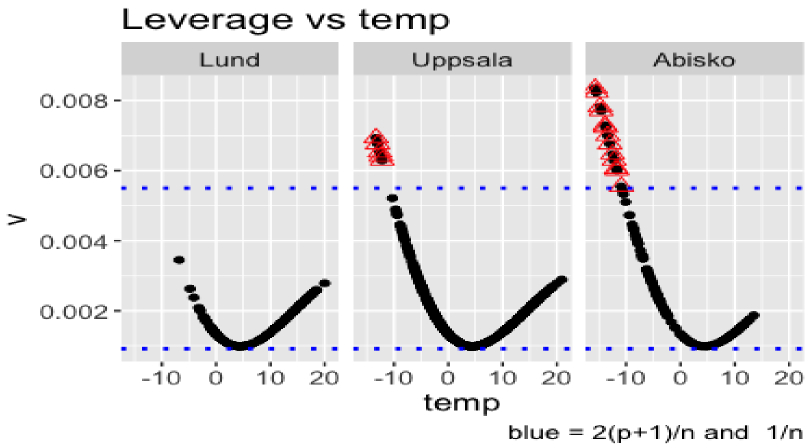


Figure 3: Leverages vs temperature, separated by location. Leverages bigger than $\frac{2(p+1)}{n}$ are highlighted in red triangles.

As one can see in figure 3, Abisko has the highest leverages. Uppsala has a few and Lund has none. This might be due to the fact that the more north the location is the more different the weather is to our model.

f)

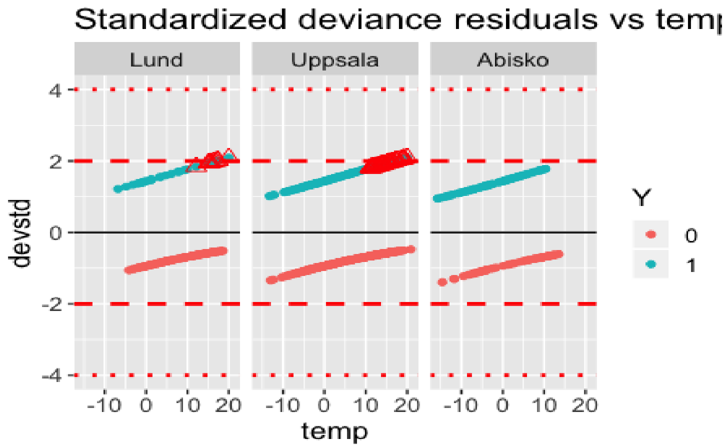


Figure 4: Standardized deviance residuals vs temperature, separated by location. Leverages bigger than $\frac{2(p+1)}{n}$ are highlighted in red triangles.

In figure 4, there are no alarmingly large residuals; even those with high leverages.

g)

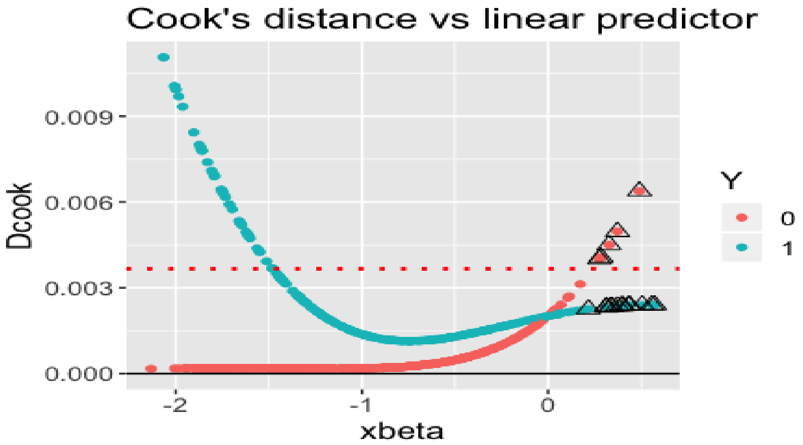


Figure 5: Cook's distance vs temperature, using different colors for low and not-low rain. Leverages bigger than $\frac{2(p+1)}{n}$ are highlighted in red triangles

The Cook's distance is not high when we have high leverages and low when we got low leverages. As seen in figure 5, for lowrain (red) a high Cook's D and high leverages coincide, but for not-low rain (blue) they do not.

3 ...or pressure...

a) We fit a logistic regression model with pressure as a covariate instead of temperature.

$$\text{Model 2: } Y(=lowrain) = I(\text{pressure} - 1012) \tag{4}$$

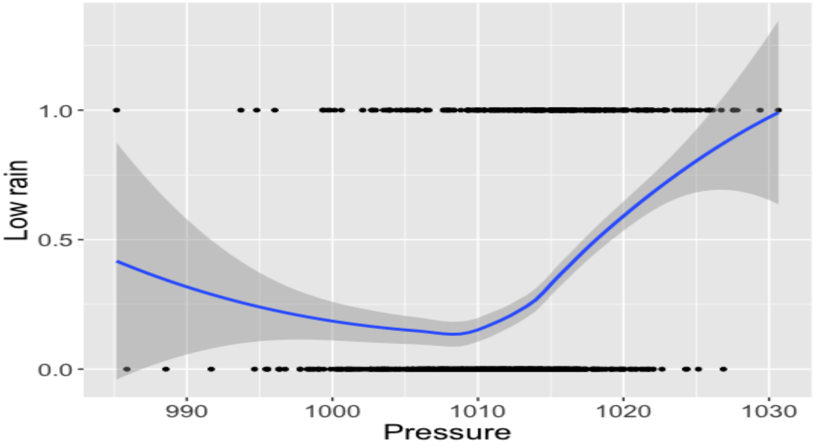


Figure 6: Moving average of low rain vs pressure

The estimate for β_0 is -1.0229 with a 95 % confidence interval of (-1.1692, -0.8807) and $\beta_1 = 0.1311$ with a 95 % confidence interval of (0.1035, 0.1600).

We can see from the confidence intervals above that the variable for the temperature covariate is statistically significant because it does not cover 0.

The odds estimate for e^{β_0} is 0.3595 with a 95 % confidence interval of (0.3105, 0.4144)

The odds estimate for e^{β_1} is 1.1401 with a 95 % confidence interval of (1.1091, 1.1735)

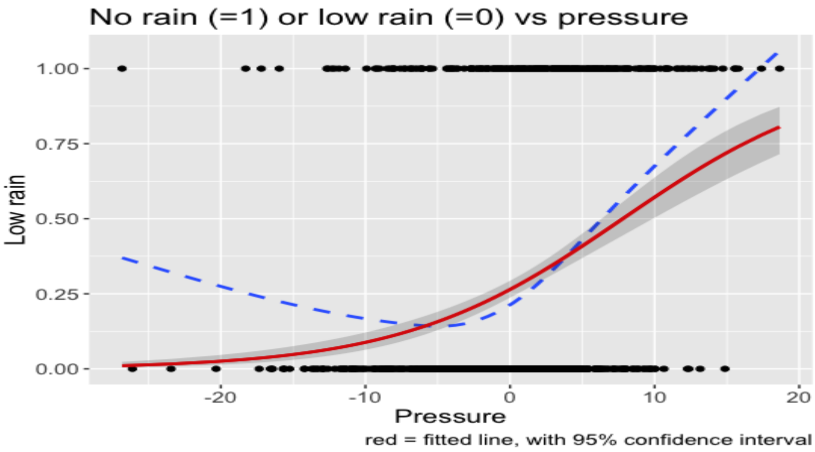


Figure 7: The predicted probabilities with 95 % confidence interval together with MA from 3 a)

What we can see in figure 7 is when the pressure is low, then there's a low probability for rain. When it increases the probability for rain does so too.

b)

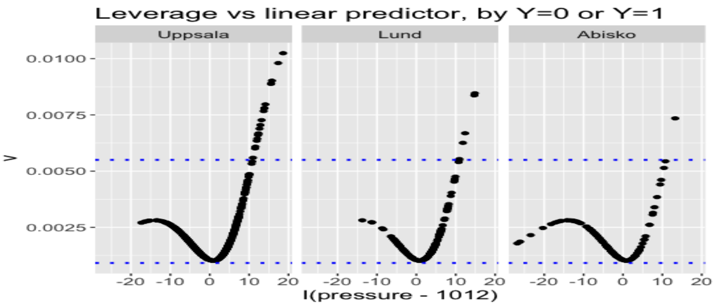


Figure 8: Leverage vs air pressure

As seen in figure 8, Uppsala has some really high leverages. The other location also have some leverages, but not as drastic. This is not necessarily a bad thing, it just means that those large leverages had a big effect on our generalized linear model (therefore the name, high leverages).

c)

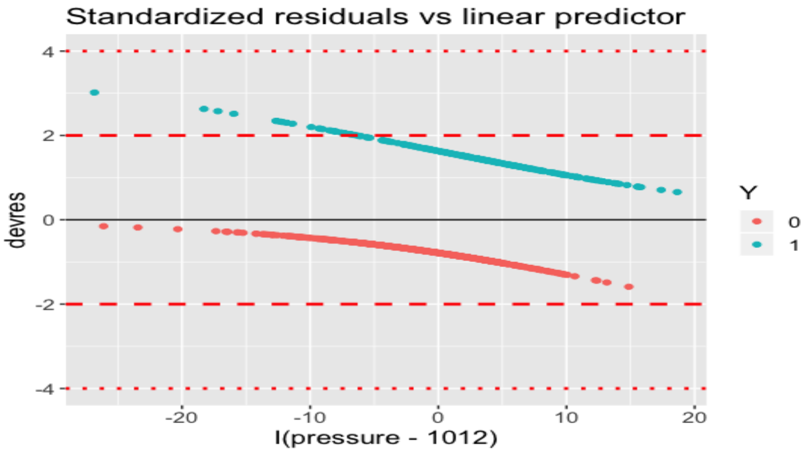


Figure 9: Standardized deviance residuals vs air pressure

There are no alarmingly large residuals in figure 9.

d)

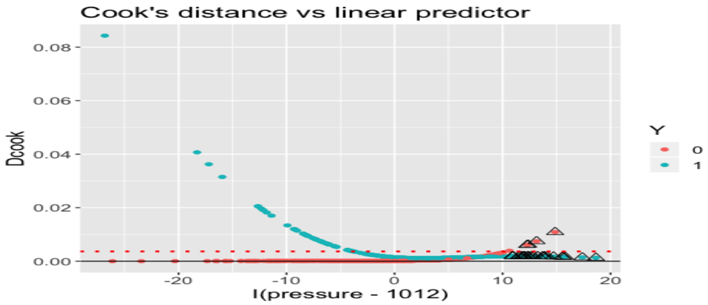


Figure 10: Cook's distance vs pressure, with different colors for low and not-low rain.

The data-points that had high leverages are highlighted in red triangles in figure 10. These do not seem to have high Cook's distances. So the points with high leverages did not have a large influence on the estimates. However we can see that there seems to be one point that looks like an outlier that had a large influence on the estimate.

e)

Comparing figure 5 and figure 10 we can see that the vast majority of datapoints have a smaller Cook's distance for the model with pressure as a covariate. A large value of Cook's distance indicates an influential observation so that means we have fewer influential points when we used pressure vs when we used temperature as a covariate.

f) and g)

Modelnumber	AIC	BIC	$R^2_{Cox-Snell}$	$R^2_{Nagelkerke}$
Model 0	1309.773	1314.768	0	0
Model 1	1242.604	1252.594	6.1	8.8
Model 2	1211.875	1221.864	8.7	12.5

Table 1: AIC, BIC, $R^2_{Cox-Snell}$ and $R^2_{Nagelkerke}$ values for different models

From table 1 we can see that the AIC and BIC values decrease when we compare the null-model (model 0) and model 1 (low rain vs temperature). This means that the variability in pressure describes the probability of having low rain the best. From the same table we can see that the R^2-values increase the most for the model using the pressure as a response variable; compared to just an intersect or temperature.

4 . . . or both with location?

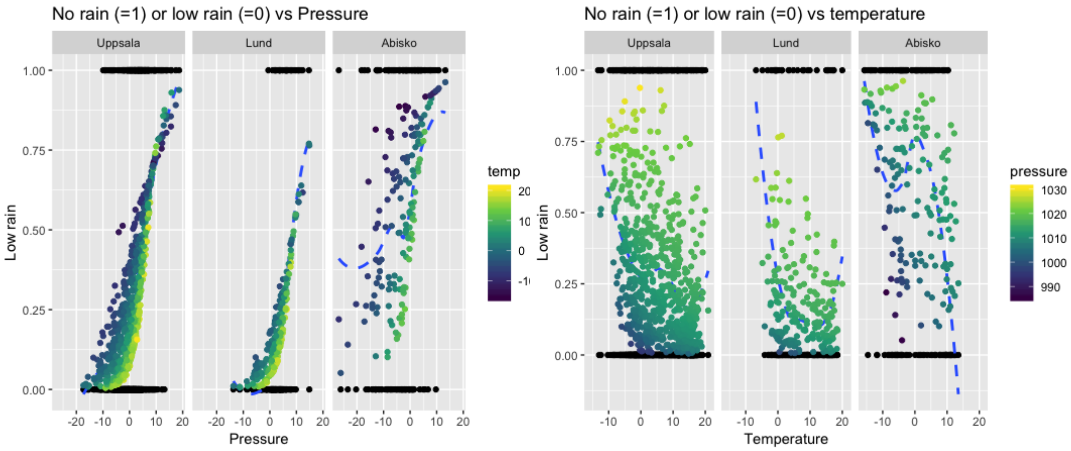
a)

We fit a logistic regression model with both temperature, pressure, an interaction between them, as well as the location variable (without interaction). The reference category was set to *Uppsala* after entering the command: `table(weather$location, weather$lowrain)`, due to Uppsala having the most data points having low rain.

$$\text{Model 3: } lowrain = temp * I(pressure - 1012) + location \tag{5}$$

The test that was used to determine the improvement of model 3 compared to previous models can be seen in section **4 f) and g)**. In summary, model 3 was better comparing AIC and BIC values.

b)



(a)

(b)

Figure 11: Low rain against temperature/pressure, in subplots by location

The temperature variable adds the largest variability in the probability of low rain, which can be seen from figure 11b. The location that the temperature add the most extra information, in addition to pressure is Lund.

c) From figure 11a we can see that for pressure follows the shape of the S-curve of the predicted probability of low rain; in contrast to how figure 11b seems to not follow the predictions. That means that if you'd want to predict the probability of low rain in Lund, then pressure is a more suitable variable.

d)

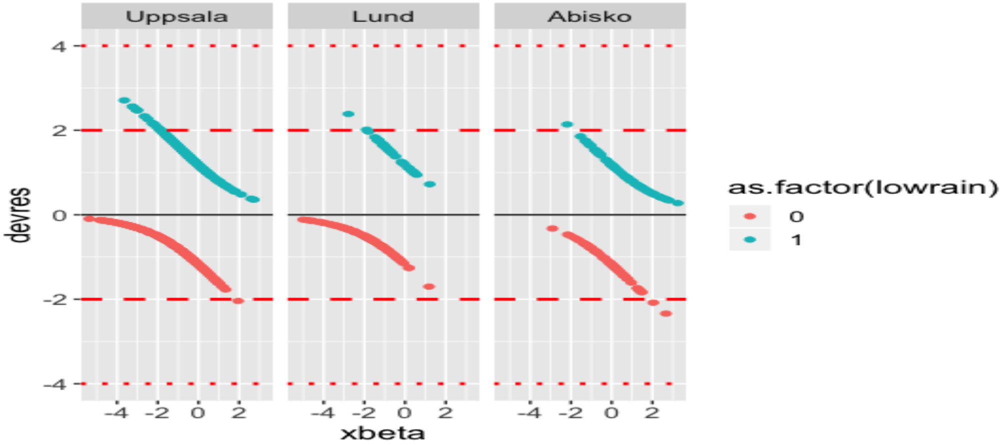
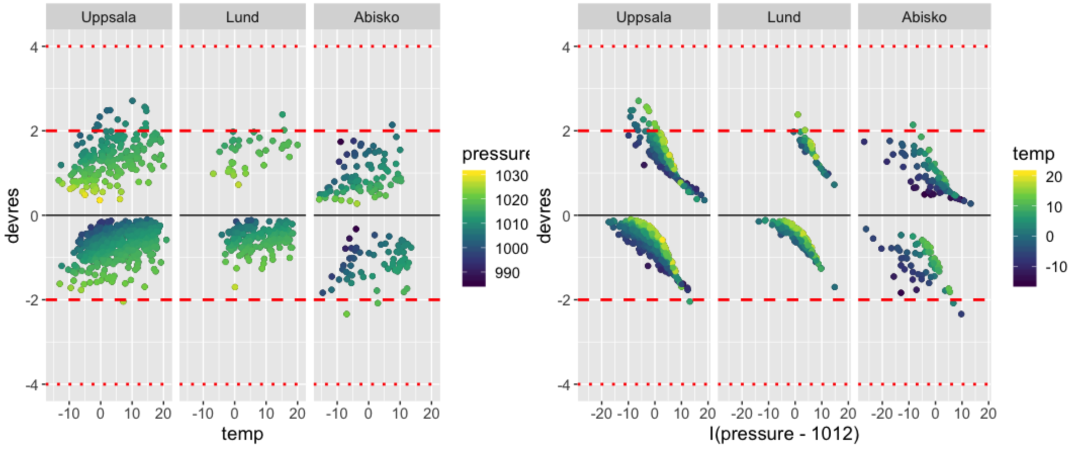


Figure 12: Standardized deviance residuals vs $x\beta$ for model 3.

According to figure 12 we see that there does not seem to be any specific points that cause problems with the standardized deviance residuals. Comparing these residuals when we only used temperature, we can see that it looks different from section 3 c) (model 3). Model 4 and model 3 look much more similar. This could be because pressure is present in both model 3 & 4 and that it describes that the variability in similar fashion compared to just using temperature.



(a) (b)

Figure 13: Standardized deviance residuals vs temperature or pressure, with colors according to pressure/temperature

e)

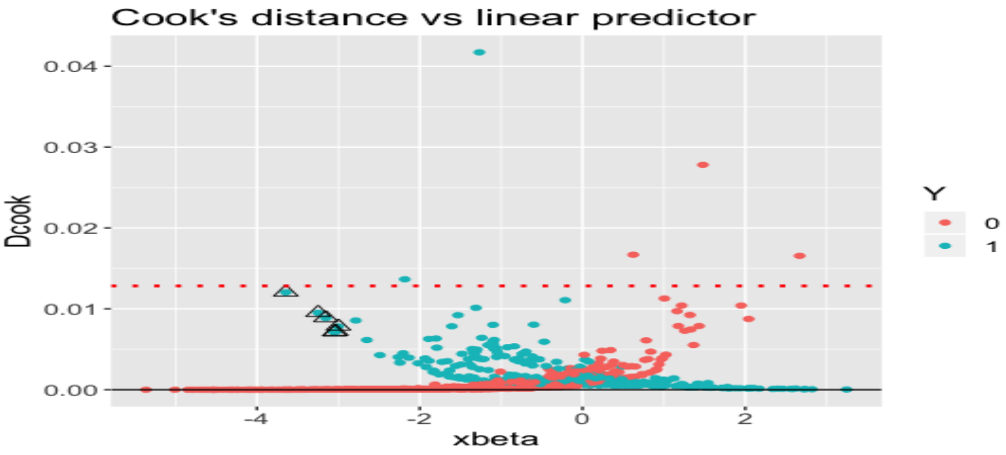


Figure 14: Cook's distance vs $x\beta$

Comparing Cook's distance for model 4 and model 3 we can see that the addition of location and an interaction changes which points that are influential on our predictions. The same concept applies to a comparison between model 4 and model 2.

In total however, there seems to be fewer points that have a single big influence on our predictions; which is a good thing. One could then assume that our model is not being overfitted.

f) and g)

Modelnumber	AIC	BIC	$R^2_{Cox-Snell}$ in %	$R^2_{Nagelkerke}$ in %
Model 0	1309.773	1314.768	0	0
Model 1	1242.604	1252.594	6.1	8.8
Model 2	1211.875	1221.864	8.7	12.5
Model 3	1017.724	1047.693	24.2	34.6

Table 2: AIC, BIC, $R^2_{Cox-Snell}$ and $R^2_{Nagelkerke}$ values for different models

From table 2 we can see that the AIC and BIC values decrease significantly when we compare model 4 with the model 0 (the null-model), model 1 (low rain vs temperature) and model 2 (low rain vs pressure). This means that the adding an interaction term and the categorical variable *location* yields much better results. This can be seen from the same table, where the AIC- and BIC-values are the lowest for model 3.

We can also see that the pseudo R^2-values increase the most for model 4 as well. A model with a larger pseudo R^2 is ”better”, but does not compensate for using more covariates (which can lead to overfitting)!

5 Goodness-of fit

In this section we will talk more about the quality of model 1, 2 and 3 (from **2b)**, **3a)**, and **4a)**)

a) We start with calculating the confusion matrices for all models. Confusion matrix for model 1: $\begin{pmatrix} 762 & 16 \\ 284 & 29 \end{pmatrix}$

Confusion matrix for model 2: $\begin{pmatrix} 753 & 25 \\ 260 & 53 \end{pmatrix}$

Confusion matrix for model 3: $\begin{pmatrix} 709 & 69 \\ 176 & 137 \end{pmatrix}$

With the values of the confusion matrices the sensitivity, specificity, accuracy, and precision for all models can be calculated which can be found in table 3.

Model	Sensitivity	specificity	accuracy	precision
1	0.09265176	0.9794344	0.7250229	0.6444444
2	0.1693291	0.9678663	0.7387718	0.6794872
3	0.4376997	0.9113111	0.7754354	0.8083485

Table 3: Sensitivity, specificity, accuracy and precision for all three models

If once compares the different models, model 3 has by far the highest sensitivity without having a bad specificity. Also accuracy and precision are the biggest for model 3, so it clearly outperfoms the others.

b) The ROC curves for all three models can be found in figure 15. As one can see, model 3 is better than the other two. Model 2 is a bit better than model 1, but for low specificity values the differences vanish.

AUC for model 1: 0.653985. The confidence interval is (0.6176712, 0.6902987). AUC for model 2: 0.6984. The confidence interval is (0.661908, 0.7349891). AUC for model 3: 0.8143433. The confidence interval is (0.7873179, 0.8413688).

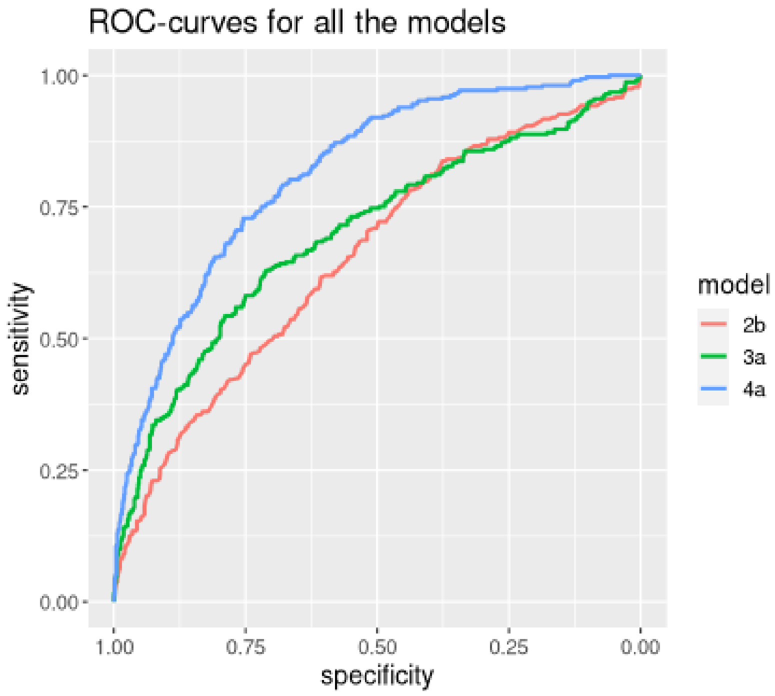


Figure 15: ROC curves for all models

This in inline with the ROC-plot where model 3 also gave the best results. When testing model 1 and 2 together with `roc.test` in R, the following results were obtained: AUC of model 1 is 0.6539850, whereas the AUC of model 2 is 0.6984485 The corresponding p-value is 0.1261. This is inline with the ROC curves, too. Model 2 is slightly better, but the difference is not that huge.

c) The optimal threshold for model 1 is 0.299. It gives the following values:

$$\text{Sensitivity} : 0.5942492$$
$$\text{Specificity} : 0.6182519$$

This gives an accuracy value of 0.6113657 and precision value of 0.3850932.

The optimal threshold for model 2 is 0.3. With that, the following values can be obtained:

$$\text{Sensitivity} : 0.6517572$$
$$\text{Specificity} : 0.659383$$

Accuracy then becomes 0.6571952 and precision 0.434968.

The optimal threshold for model 3 is 0.285. That gives the values below:

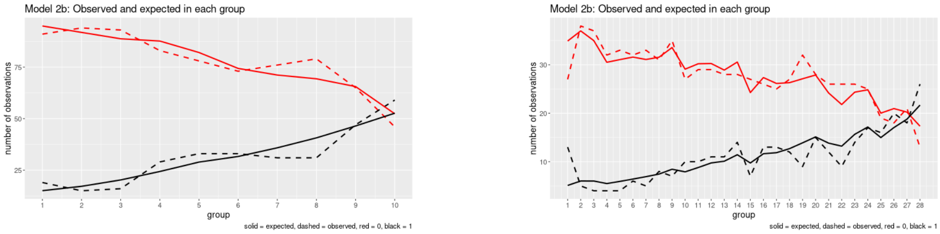
$$\text{Sensitivity} : 0.7316294$$
$$\text{Specificity} : 0.7326478$$

So accuracy and precision become 0.7323556 and 0.4882729 respectively.

For all three models, the precision went down noticeably. Accuracy did not change that drastically. In order to get sensitivity and specificity closer to each other, all models now have a lower specificity.

d)

For model 1, the optimal group number proved to be 28. It gives a mean of 5.1 and a p-value of 0.2973. If one looks at the plot 16b, one would have not guessed that this is the preferred number. Figure 16b looks a bit nicer. But for 10 groups the mean is at 15.05. The p-value is still fine, it is 0.213

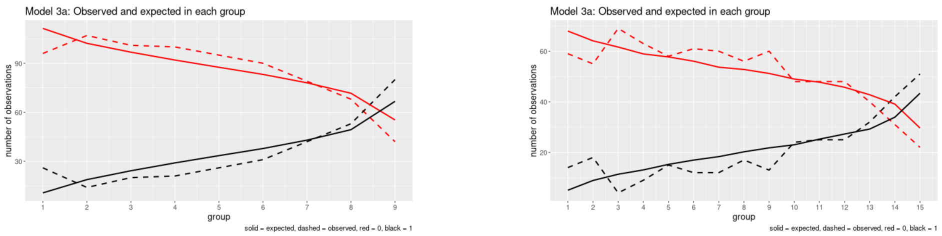


(a) 10 groups

(b) 28 groups

Figure 16: different groups for model 1

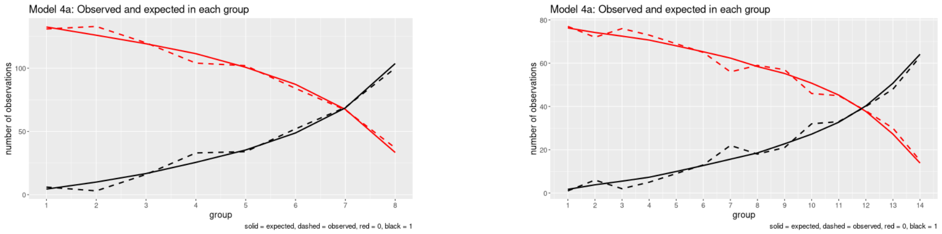
For model 2, 15 proved to be a good group number. It gives a mean of 5 and a p-value of $8.672e^{-07}$. The corresponding plot can be found



(a) 9 groups

(b) 15 groups

Figure 17: different groups for model 2



(a) 8 groups

(b) 14 groups

Figure 18: Different groups for model 3

in figure 17b. Again even though a group number of 9 gives a worse mean (12.7) and relatively good p-value of $2.587e^{-08}$, the plot, as seen in figure 17a looks nicer.

Model 3's ideal group number is 8. It gives a p-value of 0.1496 and a mean of 4.429233. If we change the group number to 14, the mean decreases to 1.71 and the p-value of 0.5978 shows that this was significantly bad. The corresponding plots can be in figure 18a and 18b. Here, the ones for the best group number also gives the best plot.

Conclusion

After looking at all the different models and testing them in various ways, one can conclude the tests pointed mostly in the same direction. AIC and BIC had their lowest values for model 3 and also in part 5, ROC and AUC prefered model 3.

One can also conclude that model 1 does not sufficiently describe the data. So it is necessary to use pressure as a covariate variable. The usage of the `step()` function in 4c) even showed that pressure is the most important variable of the model.

In the end, model 3 from section 4a) proved to be the best models through almost all tests. One could try to improve the way to deal with outliers or other problematic observation. But this is left for further studies; outside of the scope of this project.