Lund University

FMSN30

Linear and logistic regression - Project 2

Author:
Maria Gunnarsson
tfy15mgu@student.lu.se

Author:
Ebba Toreheim
tfy15eto@student.lu.se

May 13, 2020

LUND
UNIVERSITY

Introduction

In this project we will use the a weather data of precipitation in three different locations in Sweden. We will look into how the logistic regression of low precipitation as a function of different variables such as the temperature, air pressure and location. Low rainfall is defined as when the average precipitation is below 25 mm per month.

1 The null model

Before trying out different logistic regression models depending on different variable we want to have a base to compare with, a null model. We also want to understand the connection between the beta estimate, the odds and the probability. Therefore we start by estimating the odds and log-odds for a month to have low precipitation by using the mean of the binary variable `lowrain` which will give us the proportion p that a month will have low precipitation. The odds and log-odds are defined as:

$$\begin{cases} \text{odds} = \frac{\text{p}}{1-\text{p}} \\ \text{log-odds} = \ln \frac{\text{p}}{1-\text{p}} = \text{logit(p)} \end{cases}$$

and the calculated values can be found in table 1.

p	0.2868928
odds	0.4023136
log-odds	-0.9105233

Table 1: The probability, odds and log-odds for low precipitation.

After this a null logistic regression model is fitted called `model.0` and it is shown below in table 2 with its β_0-estimate, odds and probability together with their corresponding 95 % confidence intervals.

`model.0` : Y(=`lowrain`) = β_0

	Estimate	2.5 %	97.5 %
β_0	-0.9105	-1.042988	-0.78509
odds	0.4023136	0.3524002	0.4581728
$\hat{p}$	0.2868928	0.2605739	0.3142102

Table 2: The β_0-estimate, odds and probabilities with their corresponding 95 % confidence intervals for `model.0`.

Comparing the values from the model with the ones derived from the mean of the `lowrain` from the raw data one can conclude that the odds corresponds to the odds, the estimated probability

to the proportion of the months with low precipitation and the log-odds to the intercept estimate β_0.

2 Temperature...

Now we want to enhance our model and start with looking at `lowrain` with temperature (`temp`) as covariate. As a first approach 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. Looking at the figure it seems somewhat S-shaped and we can go further with the this approach and we fit a new model `model.1`. The β-estimates and corresponding confidence intervals are shown in table 3.

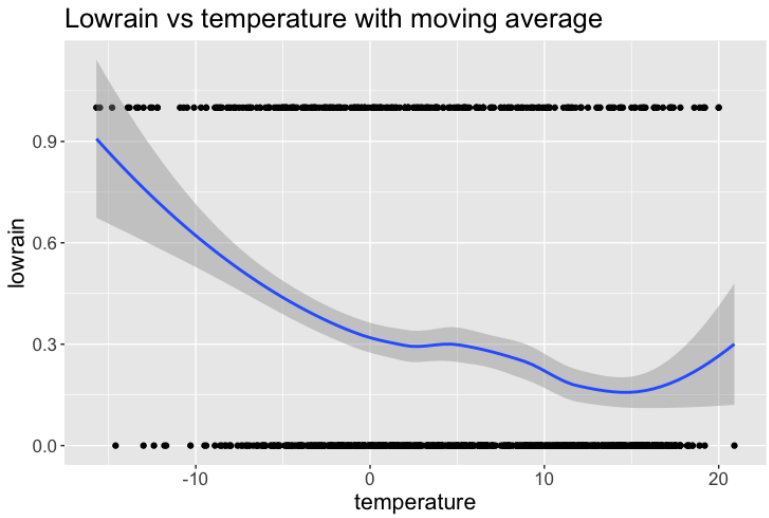


Figure 1: Variable `lowrain` plotted against `temp` with a moving average.

`model.1` : Y(=`lowrain`) $= \beta_0 + \beta_1 *$`temp`

	Estimate	2.5 %	97.5 %
β_0	-0.58860434	-0.73897322	-0.43949978
β_1	-0.07386182	-0.9218264	-0.05599002
odds	0.5551015	0.4776041	0.6443587
odds ratio	0.9288000	0.9119386	0.9455486

Table 3: The β_0-estimate, β_1-estimate odds and odds ratio with their corresponding 95 % confidence intervals for `model.1`

To see if β_1 has a significant effect on the model we conduct a Wald test with $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. We can reject H_0 if $|Z| = \frac{|\hat{\beta}_1 - 0|}{d(\hat{\beta}_1)} > \lambda_{\alpha/2}$ and in this case we have $|Z| = 8.006733$ and

$\lambda_{\alpha/2} = 1.959964$ which means that we can throw H_0 and β_1 has a significant impact. In figure 2 the predicted probabilities are plotted with their 95 % confidence intervals. The probability curve shown in red are very straight which suggests that the model is not perfect.

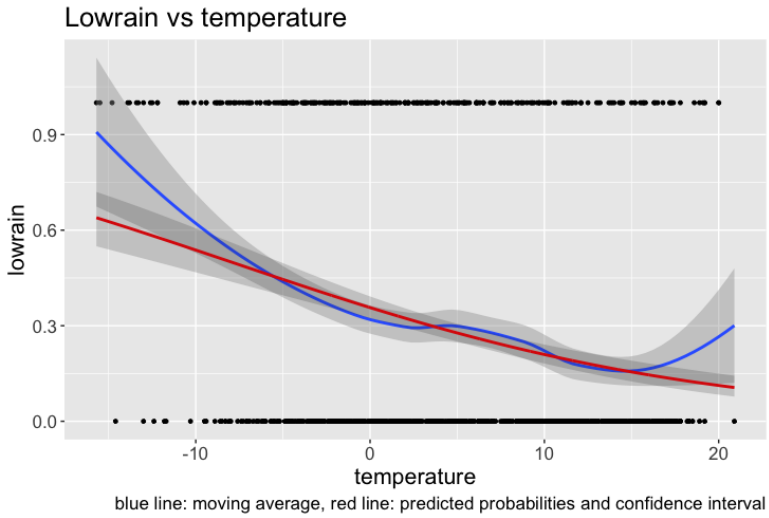


Figure 2: Predicted probablities with 95 % confidence interval for `model.1` (red line).

Looking at what will happen to the odds if we increase or decrease the temperature with 1 °C we get that an increase of temperature will give a decrease of the odds with 7.12 % and a decrease of temperature will give an increase in odds with 7.67 %. If we estimate the probabilities for low precipitation for the temperatures −10 °C, −9 °C, 9 °C and 10 °C we get the probabilities shown in table 4.

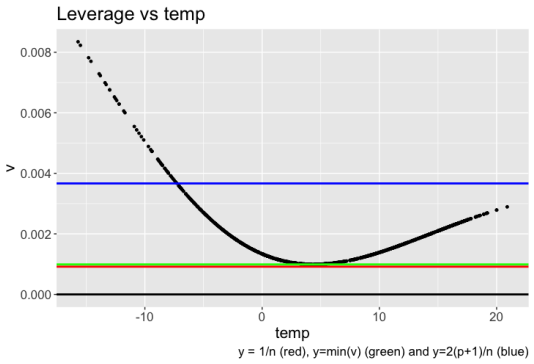
Temperature	Estimate	2.5 %	97.5 %
−10 °C	0.5374333	missing	missing
−9 °C	0.5190288	missing	missing
9 °C	0.2221189	missing	missing
10 °C	0.2096192	missing	missing

Table 4: The estimated probabilities for low precipitation for temperatures −10 °C, −9 °C, 9 °C and 10 °C together with their 95 % confidence intervals.

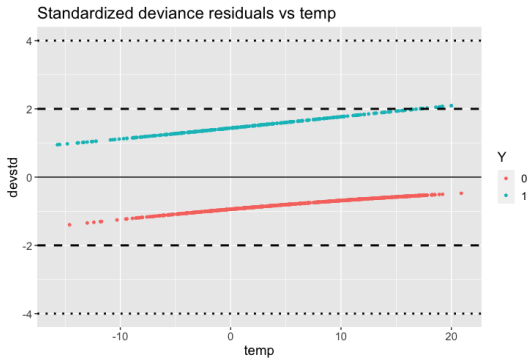
The difference in probability between −10 °C and −9 °C is approximately 0.0184 which corresponds to about a decrease of 3.5 % and between 9 °C and 10 °C the difference is approximately 0.0125 which corresponds to a decrease of 5.6 %. The difference in percentile decrease between these to temperature changes can be caused by the ....

Next we want to see if there are any data points with alarmingly large values regarding leverage,

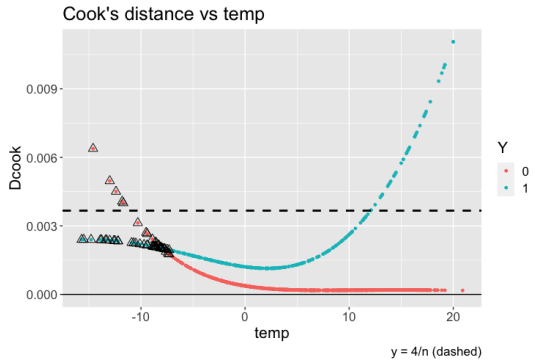
standardized deviance residuals and Cook's distance. From figure 3 you find the leverages in 3a, then standardized deviance residuals in 3b and Cook's distance in 3c.



(a) Leverages plotted against `temp`.



(b) Standardized deviance residuals.



(c) Cook's distance plotted against `temp`.

Figure 3: Three different variables for evaluation of `model.1` plotted against `temp`.

Looking more closely at the leverages in 3a you can see that from about -7 °C and down all points have a high leverage if we call a leverage high when it is larger than $2 * (p + 1)/n$ which for this case is $2 * 2/1091 = 0.003666$. We collect these observations with large leverage for later use. If we continue to figure 3b where we have the standardized deviance residuals we don't see any alarmingly large values for any observations. A few observations of low precipitations is located around the value 2 but there is no need to worry about that. In the last sub figure (3c) we see the Cook's distance of each observation and the collected observations with high leverage are marked. The observations with both high leverage and a large Cook's distance are all from the observations not belonging to the ones with low precipitation. There are however a lot of observations of low precipitation for the temperature of about 12 °C and higher with a big Cook's distance but since they don't seem to have a great leverage they are not too alarming.

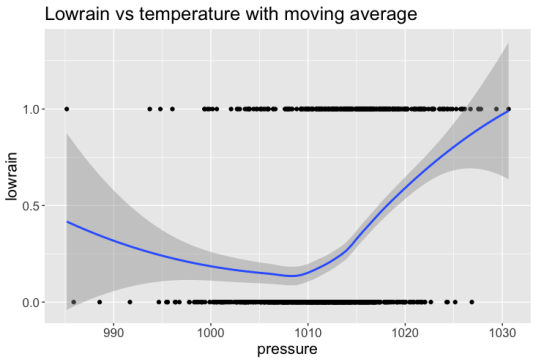
3 ... or pressure ...

We now want to see how the pressure behaves as a covariate to the variable `lowrain` and a new model, `model.2`, is fitted for the relation between `lowrain` and `pressure`. In table 5 the β-estimates, the odds and the odds ratio are displayed.

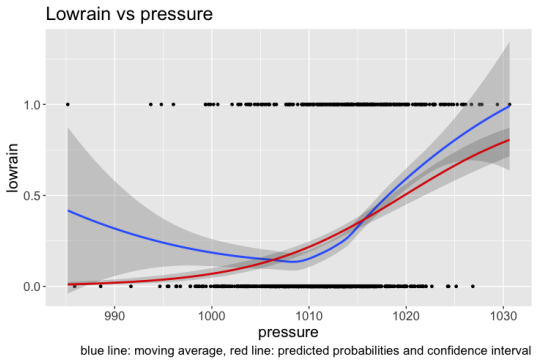
`model.2` : Y(=`lowrain`) $= \beta_0 + \beta_1 *$`pressure`			
	Estimate	2.5 %	97.5 %
β_0	-1.022943	-1.692893	-0.8807596
β_1	0.131191	0.1035733	0.1600229
odds	0.3595352	0.3105876	0.414468
odds ratio	1.1401855	1.1091271	1.173538

Table 5: The β_0-estimate, β_1-estimate odds and odds ratio with their corresponding 95 % confidence intervals for `model.2`

We also calculated the estimated probabilities and their confidence intervals and plotted them. As can be seen in figure 4b the predicted probabilities somewhat follows an S curve which calls for a probable relationship between air pressure and low precipitation. In figure 4a `lowrain` is plotted against `pressure` with an moving average without the predicted probabilities.



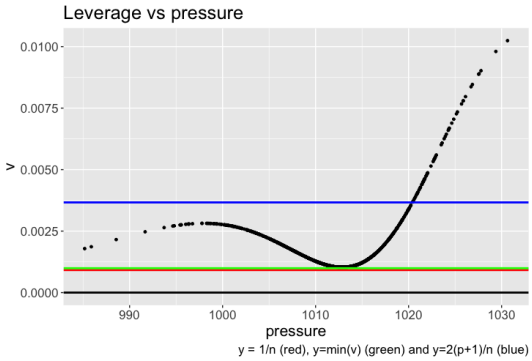
(a) Variable `lowrain` plotted against `pressure` with a moving average.



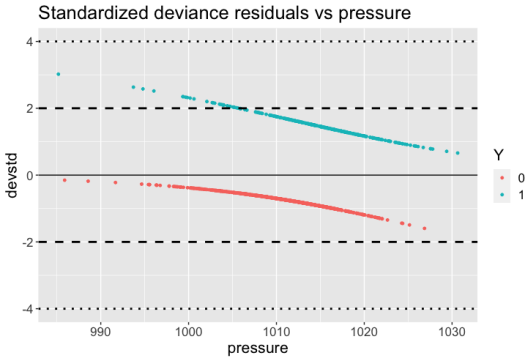
(b) Predicted probabilities with 95 % confidence interval for `model.2` (red line).

Figure 4: Connection between `lowrain` and `pressure`.

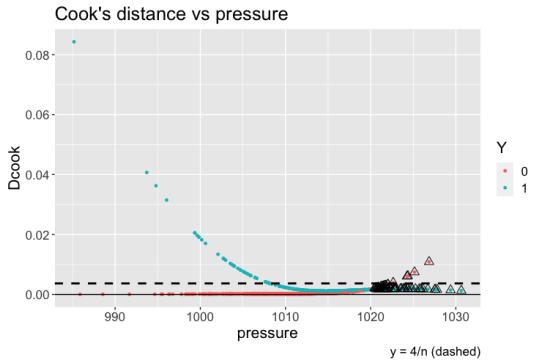
As for the case with `model.1` we want to see if any observation have alarmingly large leverage, standardized deviance residuals or Cook's distance. These plots can be seen in figure 5a, 5b and 5c.



(a) Leverages plotted against `pressure`.



(b) Standardized deviance residuals.



(c) Cook's distance plotted against `pressure`.

Figure 5: Three different variables for evaluation of `model.2` plotted against `pressure`.

In figure 5a we can see that all observations with an air pressure over 1020 hPa has a very high leverage as described in the same section for the temperature. Once again the these observations are collected as to be displayed in figure 5c. Comparing this plot to 3a we can see some similarities, the shape of the curve is somewhat alike however mirrored and the number of alarmingly observations are of the same number. When looking in figure 5b there are some observations of low precipitation with a standardized deviance residual with a value larger than 2 but none larger than 4 and there seems not to be anything to worry for here. In the comparison of standardized deviance residuals the same as for the comparison of leverage between the models stand. The values for the pressure model are a higher and the plots are mirrored with higher values for this model. Lastly looking at the plot of the Cook's distance in figure 5c there are a lot of large values for mostly observations of low precipitation but here, like for the observations in 3c, the observations with a large Cook's distance and a large leverage are all exclusively observations of not low precipitation.

Comparing the models further we have calculated the $R^2_{Cox-Snell}$, $R^2_{Nagelkerke}$, AIC and BIC for `model.1` and `model.2` and the results can be found in table 6. Comparing the values for the tests they all agree that the best model is `model.2`, it is very believable when looking at figure

2 and 4b the curve for the predicted probabilities in 4b are more S shaped such as a good model should in a situation as this.

Comparisons of `model.1` and `model.2`

Model	$R^2_{Cox-Snell}$	$R^2_{Nagelkerke}$	AIC	BIC
`model.1`	6.14	8.80	124260.4	125259.4
`model.2`	8.75	12.53	121187.5	122486.4

Table 6: $R^2_{Cox-Snell}$, $R^2_{Nagelkerke}$, AIC and BIC for `model.1` and `model.2`

4 ... or both with location

We now want to make a larger model where we use temperature, air pressure and location as covariates with an interaction between temperature and air pressure. We also make sure to use Uppsala as reference location since it has most observations. We call this model `model.3` and in table 7 some of its properties are displayed.

`model.3`: Y(=`lowrain`) $= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$

x_i	β_j	$\hat{\beta}_j$	std. Error	z value	Pr(>|z|)
Intercept	β_0	-0.900921	0.108357	-8.314	$<$ 2e-16
I(pressure - 1012)	β_1	0.195724	0.016533	11.838	$<$ 2e-16
temp	β_2	-0.092003	0.11586	-7.941	$<$ 2e.01-15
locationLund	β_3	-0.836368	0.225811	-3.704	0.000212
locationAbisko	β_4	1.697879	0.234816	7.231	4.81e-13
I(pressure - 1012):temp	β_5	0.009917	0.002127	4.664	3.11e-06

Table 7: The variables from `model.3` and some of there properties.

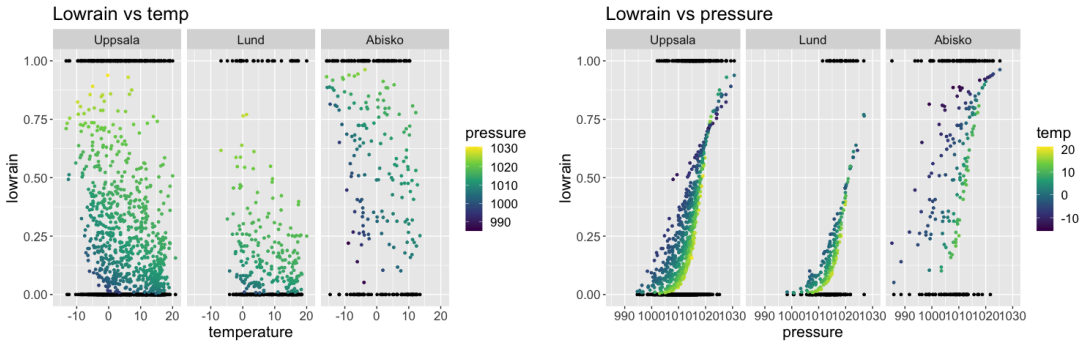
To see if this model is better than either of `model.1` or `model.2` we perform a partial likelihood ratio test where we compare the difference of the deviance with a $\chi^2_\alpha(k)$ with k as the difference of variables β_j between the models. The results can be seen in table 8 and it is clear that `model.3` is better than both the previous models.

models	Deviance difference	$\chi^2_\alpha(k)$
`model.1` - `model.3`	232.8803	9.487729
`model.2` - `model.3`	202.1507	9.487729

Table 8: Partial likelihood ratio tests for `model.3` compared with `model.1` and `model.2`.

Knowing that the `model.3` is the best model so far we go on calculating the predicted probabilities and plot these in two figures with different variables on the axes and in subplots separated

by location. In plot 6a we have that `lowrain` on the y-axis and `temp` on the x-axis and the predicted probabilities is added to the plot with different colors depending on the pressure. In figure 6b we have instead the `pressure` on the x-axis and the predicted probabilities have the color of temperature. Looking in these figures one can see that in Abisko the temperature add the most extra information given the pressure.



(a) `Lowrain` and `phat` against `temp`.

(b) `Lowrain` and `phat` against `pressure`.

Figure 6: Predicted probabilities `phat` from `model.3` and `lowrain` plotted against `temp` respectively `pressure` separately for each location.

Looking more into the location Lund you would expect from the figures in 6 that the pressure is the most useful variable. To see if this conlcusion is correct we fit a model for Lund with air pressure, temperature and an interaction term. This model is called `model.lund` and can be seen in table 9 with its β-estimates and e^{β} with 95 % confidence intervals. These estimates and their confidence interval supports our suggestion about the most significant variable to be `pressure`. Looking at the β-estimates for `temp` and the interaction term both are smaller than the estimate for air pressure. Also the the confidence intervals are very close to or includes zero which suggest that they are not very accurate nor significant.

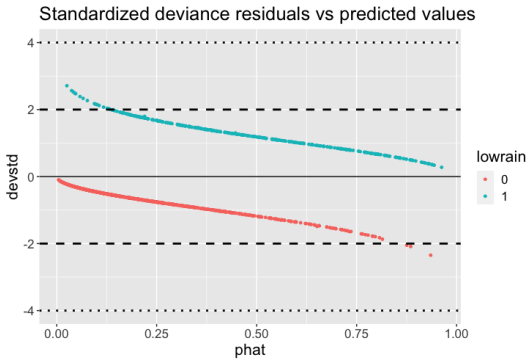
`model.lund`

	β_j	Estimate	2.5 %	97.5 %
Intercept	β_0	-1.95349878	-2.90102713	-1.19169302
I(pressure-1012)	β_1	0.27579223	0.15599432	0.42524011
temp	β_2	-0.15036051	-0.28713294	-0.02788756
I(pressure - 1012):temp	β_3	0.01792693	-0.00357026	0.04072037
Intercept	e^{β_0}	0.1417772	0.05486673	0.3037066
I(pressure-1012)	e^{β_1}	1.3175741	1.16881956	1.5299577
temp	e^{β_2}	0.8603977	0.75041197	0.9724977
I(pressure - 1012):temp	e^{β_3}	1.0180886	0.99643611	1.0415608

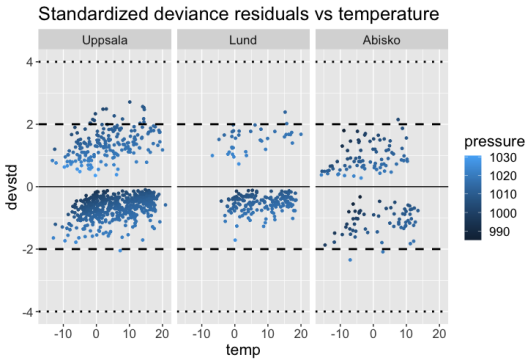
Table 9: The β-estimates and the e^{β}-estimates for `model.lund`.

To even reduce the model a backward eliminations using BIC criteria is conducted and the result is that for Lund the best model is a model with air pressure as only covariate.

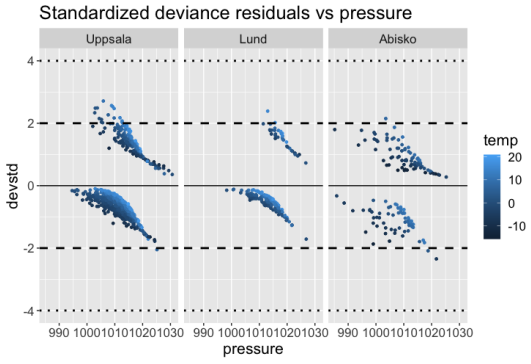
Further investigations of `model.3` is now done by calculating the standardized deviance residuals and then plot the first against the predicted values $\mathbf{x_i}\hat{\boldsymbol{\beta}}$ (figure 7a). Here we can see that some values for both `lowrain` = 0 and `lowrain` = 1 have an absolute value larger than two, but none that is alarmingly large. Secondly against temperature, now separate for each location and the temperature of the pressure (figure 7b). In the figure 7b there is a large dispersion of the residuals, they are not formed as two lines such as in figure 3b. We have also for `model.3` some observations with larger values than for `model.1`. Lastly the standardized error are plotted against the pressure, also separately for each location, but now with the color of temperature (figure 7c). In this plot we can see, as opposed to the plot against temperature, that the standardized deviance errors are more collected and not as dispersed. When comparing with 5b we can see that out largest value for `model.3` as smaller than the largest value for `model.2` which is good.



(a) Plotted against predicted values `phat`.



(b) Plotted against `temp`.

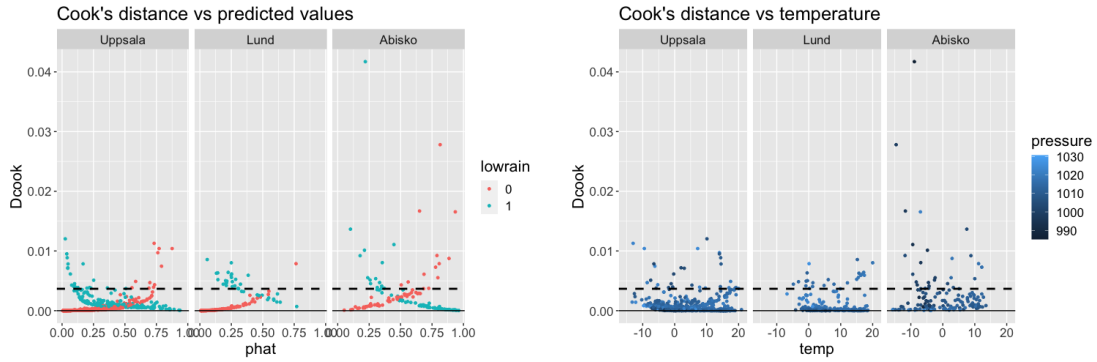


(c) Plotted against `pressure`.

Figure 7: Standardized deviance residuals plotted in three different ways.

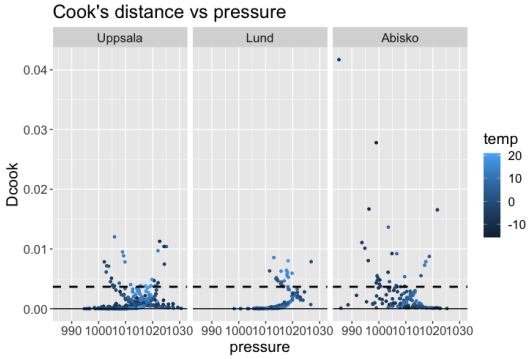
Next we calculate Cook's distance for `model.3` and plot it in figure 8a and we can see that with

a limit of 4/n with n = 1091 there are a lot of observation with a very high value. With that said the highest value is just over 0.04. Now we want to compare with the earlier models as we did with the standardized deviance residuals we have therefore plotted Cook's distance against temperature in 8b with the color of pressure and against pressure with the color of temperature in figure 8c. Comparing these with figure 3c for temperature and figure 5c we can see that for the temperature for `model.1` the largest Cook's distance is smaller that 0.01 where we for both Uppsala and Abisko have distances a lot larger that this so it would seem like they have not improved for the temperature. Now comparing the plots done against pressure the distances are a lot better where it in figure 5c the largest value is larger than 0.08 which means that looking at pressure the `model.3` has improved a lot.



(a) Plotted against predicted values `phat`.

(b) Plotted against `temp`.



(c) Plotted against `pressure`.

Figure 8: Cook's distance plotted in three different ways.

We again want to compare the models with respect to $R^2_{Cox-Snell}$, $R^2_{Nagelkerke}$, AIC and BIC. In table 10 the results for these quantities are collected and as the other tests suggested all these quantities say that `model.3` is the best model.

Comparisons of `model.1`, `model.2` and `model.3`

Model	$R^2_{Cox-Snell}$	$R^2_{Nagelkerke}$	AIC	BIC
`model.1`	6.14	8.80	124260.4	125259.4
`model.2`	8.75	12.53	121187.5	122486.4
`model.3`	24.18	34.63	101772.4	104769.3

Table 10: $R^2_{Cox-Snell}$, $R^2_{Nagelkerke}$, AIC and BIC for `model.1`, `model.2` and `model.3`.

5 Goodness of fit

Another way to measure the performance of a model is to look at its goodness of fit. In tables 11, 12 and 13 the confusion matrix as well as specificity, sensitivity, accuracy and precision are presented for each model. The specificity is calculated as the proportion of true negatives that have been correctly classified, i.e $\frac{TN}{TN+FP}$ and the model with the highest specificity is `model.1`. The sensitivity on the other hand is a measure the proportion of the true positives, i.e $\frac{TP}{FN+TP}$ and the by far best model in this sense is `model.3`. Continuing to the accuracy measure which is the overall proportion of correctly classified, i.e $\frac{TP+TN}{n}$. Here as well as for the sensitivity is `model.3` the best model. Lastly we have the precision which is a measure of the proportion of the predicted successes that are true successes, i.e $\frac{TP}{FP+TP}$. When it comes to precision it is closer between the three models and the one with the slightly higher precision is `model.2`. To make a conclusion here we can say that for specificity, precision and accuracy all three models are quite similar but for the measure of sensitivity `model.3` out-performs the rest.

Confusion matrix `model.1`

True \Predicted	0	1	total	correctly classified	
0	762	16	778	97.9 %	specificity
1	284	29	313	9.3 %	sensitivity
total	1046	45	1091	72.5 %	accuracy
				64.4 %	precision

Table 11: Confusion matrix, specificity, sensitivity, accuracy and precision for `model.1`.

Confusion matrix `model.2`

True \Predicted	0	1	total	correctly classified	
0	753	25	778	96.8 %	specificity
1	260	53	313	16.9 %	sensitivity
total	1013	78	1091	73.9 %	accuracy
				67.9 %	precision

Table 12: Confusion matrix, specificity, sensitivity, accuracy and precision for `model.2`.

Confusion matrix `model.3`

True \Predicted	0	1	total	correctly classified	
0	709	69	778	91.1 %	specificity
1	176	137	313	43.8 %	sensitivity
total	885	206	1091	77.5 %	accuracy
				66.5 %	precision

Table 13: Confusion matrix, specificity, sensitivity, accuracy and precision for `model.3`.

In figure 9 the ROC curves for all three models are shown. It is evident that for all point `model.3` is the best, however it has quite a lot to work on to become optimal.

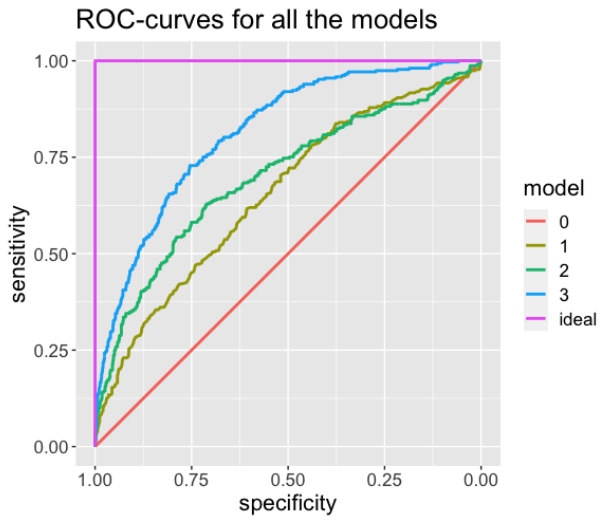


Figure 9: ROC-curves for `model.1`, `model.2` and `model.3`.

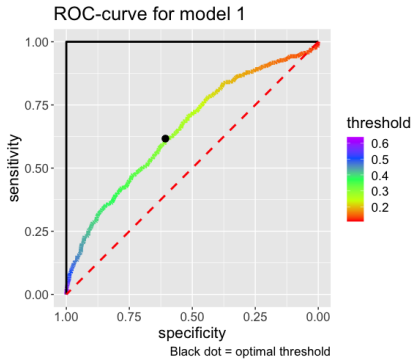
In table 14 the calculated AUC are shown with their corresponding 95 % confidence intervals. `Model.3` is clearly the best one with no overlap of confidence interval with either of the other models but since `model.1` and `model.2` have overlapping confidence intervals we want to test to see if they are significantly different. The result gives a z value of -1.5298 and a p-value of 0.1261 which clearly is larger than 0.025 and therefore `model.2` is slightly better than `model.1`.

model	AUC	2.5 %	97.5 %
1	0.653985	0.6176712	0.6902987
2	0.6984485	0.661908	0.7349891
3	0.8143433	0.7873179	0.8413688

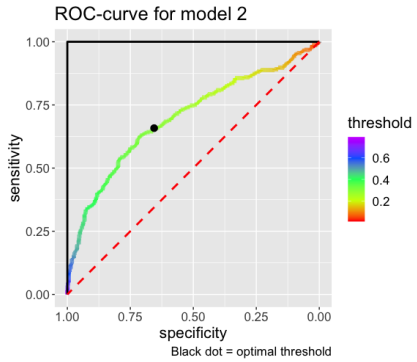
Table 14: The AUC for all models with their corresponding 95 % confidence intervals.

Now we want to optimize the models by changing the threshold so that the sensitivity and

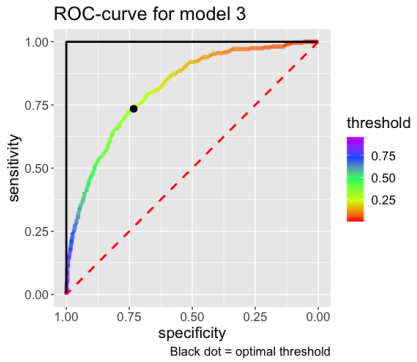
specificity is approximately equal and as large as possible. The results can be seen in figures 10a - 10c where the optimal threshold value is marked out for each of the models' ROC-curves. With these new thresholds we calculate new confusion matrices, sensitivities, specificities, accuracies and precisions, and those are displayed in tables 15, 16 and 17. Doing this comes with the price that we increase the number of correctly classified negatives but we drastically decrease the number of correctly classified true positives and depending on what the results are to be used for this could be troublesome.



(a) ROC-curve and threshold for `model.1`.



(b) ROC-curve and threshold for `model.2`.



(c) ROC-curve and threshold for `model.3`.

Figure 10: ROC-curves and optimal threshold values for the three models.

Confusion matrix `model.1`

True \Predicted	0	1	total	correctly classified	
0	777	1	778	99.9 %	specificity
1	308	5	313	1.6 %	sensitivity
total	1085	6	1091	71.7 %	accuracy
				83.3 %	precision

Table 15: Modified confusion matrix, specificity, sensitivity, accuracy and precision for `model.1`.

Confusion matrix `model.2`

True \Predicted	0	1	total	correctly classified	
0	776	2	778	99.7 %	specificity
1	301	12	313	3.8 %	sensitivity
total	1077	14	1091	72.2 %	accuracy
				85.7 %	precision

Table 16: Modified confusion matrix, specificity, sensitivity, accuracy and precision for `model.2`.

Confusion matrix `model.3`

True \Predicted	0	1	total	correctly classified	
0	764	14	778	98.2 %	specificity
1	254	59	313	18.8 %	sensitivity
total	1018	73	1091	75.4 %	accuracy
				80.8 %	precision

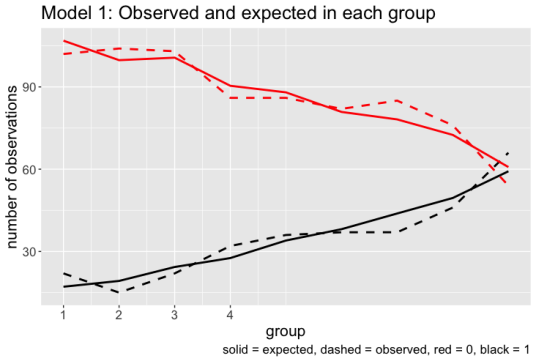
Table 17: Modified confusion matrix, specificity, sensitivity, accuracy and precision for `model.3`.

Another method to see the goodness of fit of a model is to perform a Hosmer-Lemeshow goodness of fit test. We will do that separately for the three models and try a handful of different number of groups with the minimum number set to p+2 and make sure that the smallest expected number wont go down much under the value 5. In table 18 the test results from the HL-tests are presented for all models with the number of groups, the lowest expected number, the χ^2 value as well as the p-value. The test is saying that H_0: "The model gives correct probabilities" and we reject this if $p < 0.025$ and if we can't reject the model this is not the same as to say it is correct. It just means that we can't prove that the model is wrong. With this we can see that the only model we can reject is `model.2` and that this is the only model we can prove can't predict the number of outcomes correctly.

g	smallest expected number	χ^2	p-value
	`model.1`		
3	60.7	0.59291	0.4413
4	41.9	2.0969	0.3505
9	17.13	7.7477	0.3554
12	13.15	8.9119	0.5405
15	9.7	15.948	0.2519
30	4.6	34.348	0.1897
	`model.2`		
3	53.67	10.584	0.001141
4	34.9	13.433	0.001211
9	10.7	39.817	1.364e-06
12	7.0	35.002	0.0001247
15	5.05	53.103	8.672e-07
	`model.3`		
7	5.5	5.3296	0.377
8	4.4	9.454	0.1496

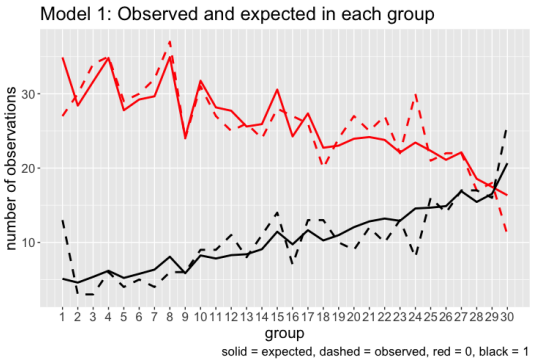
Table 18: The Hosmer-Lemeshow test for all the models displaying number of groups, the smallest expected value, the χ^2 value and the p-value. If $p > 0.025$ we cannot reject the model and this means that we can't prove that the model is wrong.

In figure 11, 12 and 13 a few plot for each model is displayed of the Hosmer-Lemeshow goodnes of fit test.



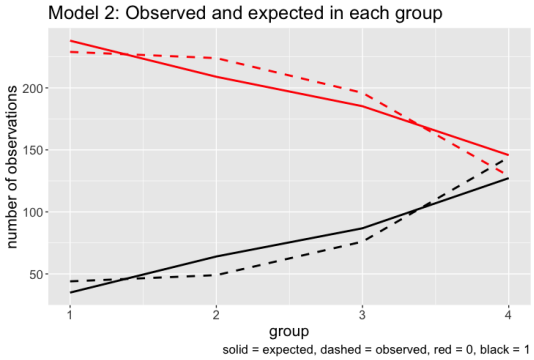
(a) HL-test with g=4.

(b) HL-test with g=12.

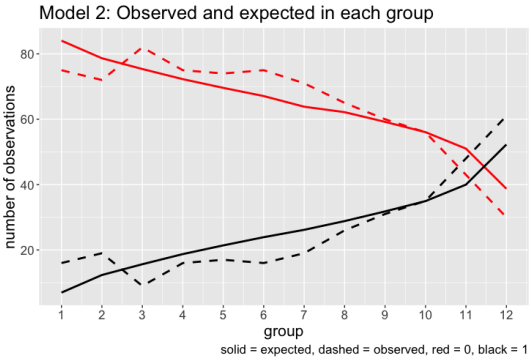


(c) HL-test with g=30.

Figure 11: HL-test for `model.1`.



(a) HL-test with g=4.



(b) HL-test with g=12.

Figure 12: HL-test for `model.2`.

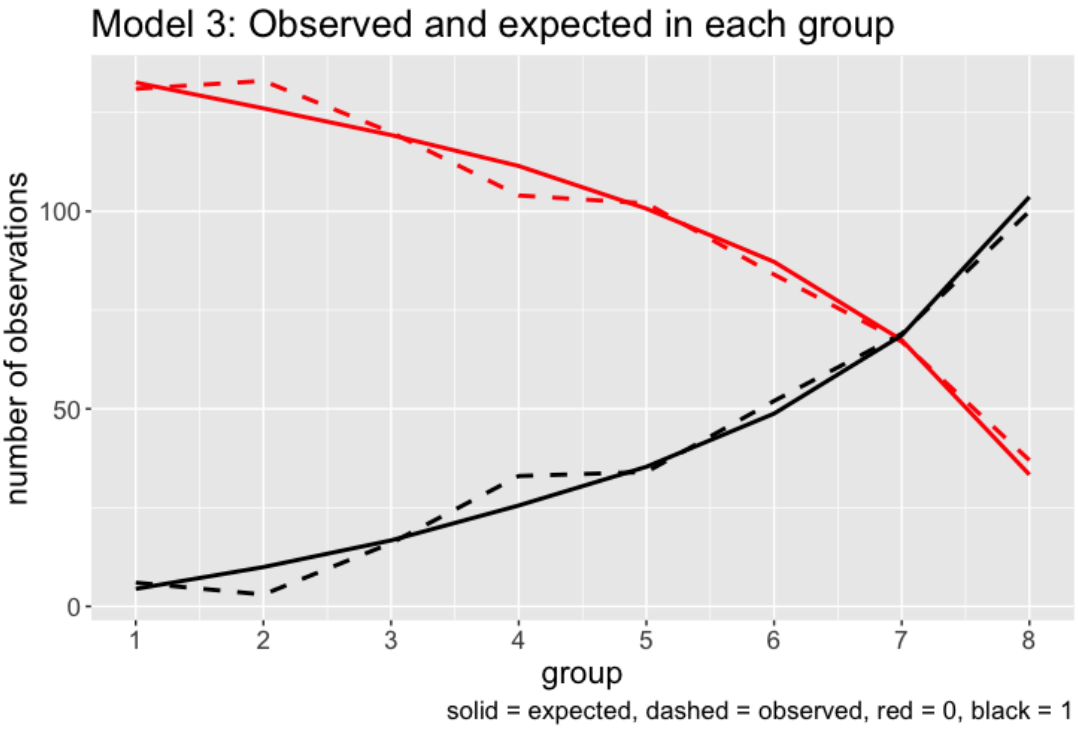


Figure 13: HL-test for `model.3` with g=8.

Final remarks

In conclusion of this project it is not always clear if a model is good or not. Sometimes you think a model is better than another but in the end this is a model to reject.