CASE CAMPY

Applied statistics and statistical software
02441

Kasper Einarson s134604
Ran Wang s111503
Linards Kalnins s124612

Technical University of Denmark

DTU compute

14-01-2015

1 Summary

Contents

1 Summary **1**

2 Introduction **3**

3 Data **3**

4 Analysis **7**
4.1 Regional differences 14

5 Discussion and future work **19**

6 Conclusion **19**

7 Appendix **20**

2 Introduction

Campylobactor has during the last years been the leading cause of enteric infections in Denmark. One strategy to reduce the number of infections among humans is to reduce the numbers of infected broilers. To obtain knowledge about the broilers tendency to get infections data has been recorded from many broil farms in the period of 1998 to 2008. The data contains recordings from the climate including variables like temperature, humidity, sun hours and precipitation as well as the numbers of infected broilers. The purpose of this project is to explain the change in numbers of positive tested broilers due to climate data. We will investigate which climate variables have the highest influence using multiple linear regression. Furthermore regional differences between the number of positive broilers are also being considered.

3 Data

During the years 1998 to 2008 broilers have been tested for campylobactor infections. The data from recordings are summarized in table 1. Here we see the parameters tested consist of what time (year and week) the test was carried out, the temperature (average temperature and maximum temperature) together with the average weekly relative humidity (RHum) and hours of sunlight the given week(sunHours). Also the precipitation each week (precip) was recorded. The columns "total" and "pos" are continuous (cont) variables describing how many broils was slaughtered the given week (total) and how many of those were tested positive in campylobactor (pos).

Variable	year	week	AveTemp	MaxTemp	RHum
Type	factor	factor	continuous	continuous	continuous
Mean	-	-	8.717	14.79	77.41
Range	[1998;2008]	[1;53]	[-5.4;21]	[0.60;29.70]	[0;98]
Observations	537	537	537	537	505
Variable	sunHours	precip	total	pos	
Type	continuous	continuous	cont	cont	
Mean	32.36	15.26	74.23	25	
Range	[0;103]	[0;75]	[5;130]	[0;85]	
Observations	463	537	537	537	

Table 1: Summarizing data of tested broiler farms in the period of 1998 - 2008

Figure 1 shows the amount of broilers slaughtered each year between 1998 and 2008. Compared to the rest of the years, 2008 have only very few observation and one should not conclude any tendency from this plot concerning that matter. From this plot it is hard to conclude that any of the years between 1998 and 2005 should be significantly different from one another. Both the numbers of total slaughters broilers seems somewhat constant and the number of slaugh-

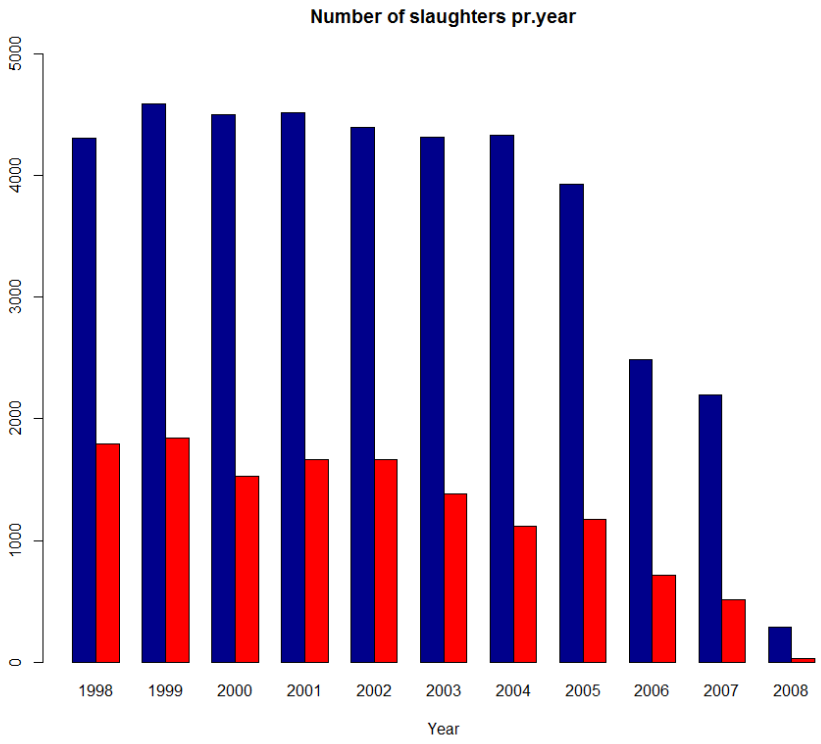

Figure 1: barplot of slaughtered broilers from each year Blue is total amount of broilers slaughtered Red is the amount of slaughtered broilers tested positive. Notice that 2008 have very few observation and should be not be taken into consideration when concluding tendencies on this figure

tered broilers tested positive. It does seems like the number of total slaughters are decreasing a little bit from year 2001 to 2007 with a very pronounced different from year 2006 and 2007. Also the amount of positive tested broilers seems generally to have decreased from 2001 to 2007. To investigate further the tendency of which regions there was the highest proportion of positive tested broilers figure, Figure 2 gives an overview.

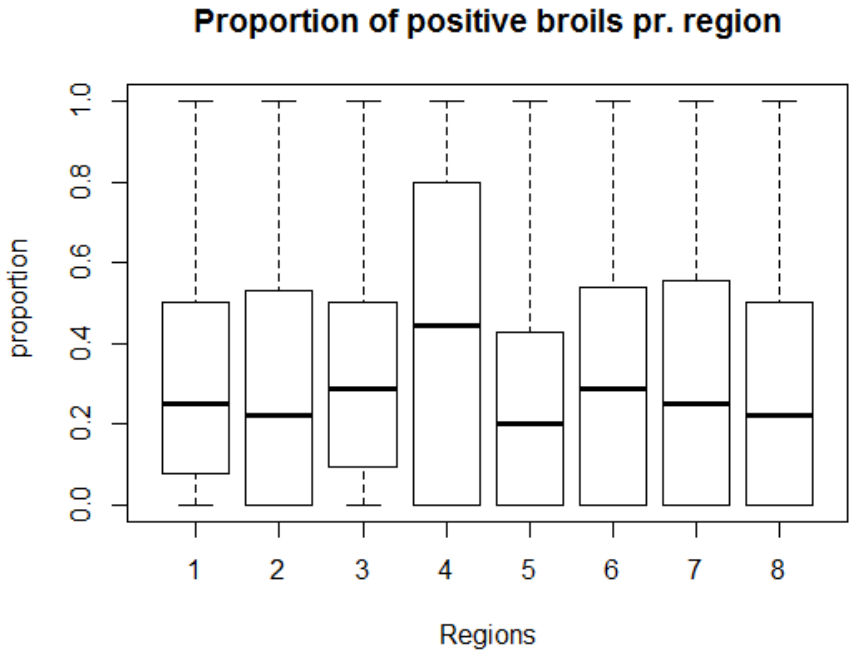

Figure 2: Boxplot of the proportion of positive tested broilers among the total amount of broilers given each year

In this figure see that the mean of positive tested broilers is highest in region 4 and lowest in region 5. Region 4 have a proportion of about 40 procent of slaughtered broilers were positive whereas the other regions lies between 20 - 30 procent.
Since we want to analyse whether the climate influences these effects we make a plot taking all the variables into account and look at them relative to the proportion variable (proportion between positive and total amount of slaughtered broils).The plot of this is shown in Figure 3.

In this plot we see that the both the average temperature and the maximum temperature seem to have an impact on the amount of positive tested broilers. From the figure it is hard to determine whether there is some tendency provided by the humidity, sunHours or precip. We note that two points from the humidity

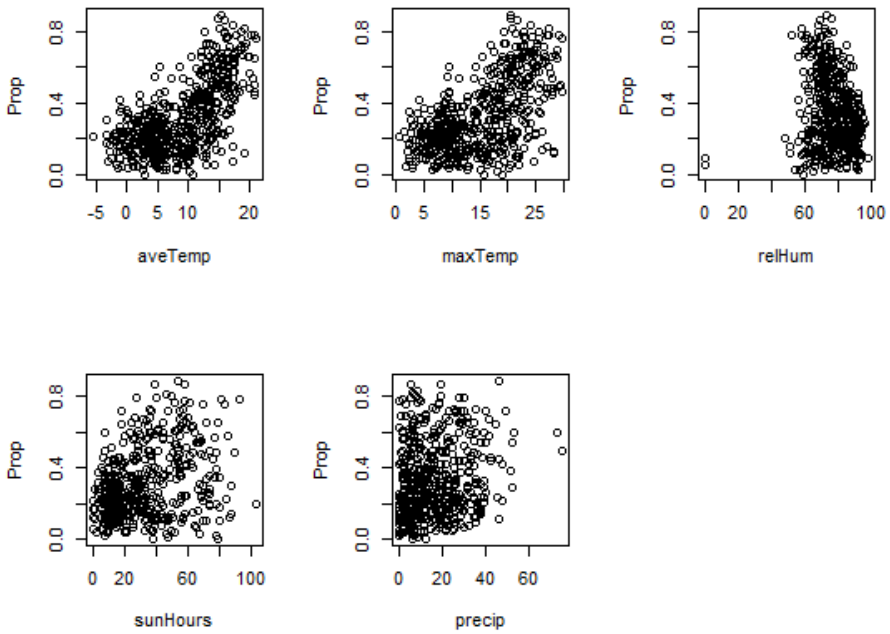

Figure 3: plot showing the different variables and some tendency of their interactions.

is 0 which makes no physical sense. This could be due an observation error and the two datapoints are excluded from the dataset. Further more, some of the observations in humidity (relHum) and sun hours (sunHours) per week were not signed any value forcing us to delete the rows involving a relHum or sunHours with oberservation "NA". We also noticed 4 observation in week 20,21,22 and 23 in 2007 with 0 sun hours. Given this is in late spring and also the average temperatures were actually fairly high, we consider this very unlikely and exclude them from the dataset.
This cuts our dataset down from a total of 537 (as seen in table 1) to 427 observations.

4 Analysis

Now that we have corrected the dataset to only the relevant observations, we preceed with the analysis. From figure 4 we see a comparison of the explanatory variables with respect to the proportion of positive tested broilers to look for correlations.

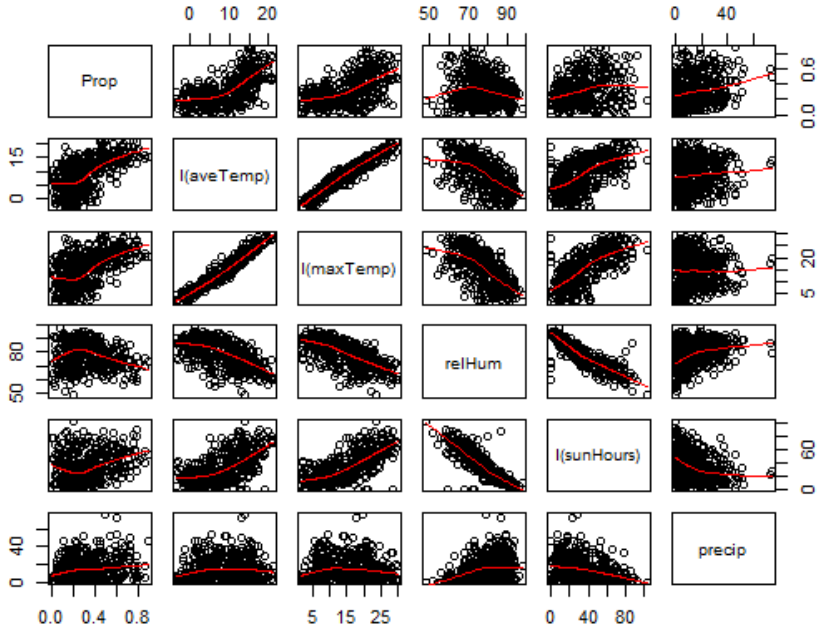

Figure 4: pairs plot of interactions between the explanatory variables and the proportion of positive tested broilers

It is seen how the average temperature and the maximum temperature seems to have a pronounced impact on the proportion, while the humidity and sun hours have a rather unclear tendency concerning this matter. The precipitation seems to have a positive linear tendency with the proportion. It is clear that there seem to be interactions between the explanatory variables as well, and we therefore wish to carry out a multiple linear regression. To investigate further which explanatory variables are most important to the proportion of broilers that were tested positive, we make a tree model shown in figure 5

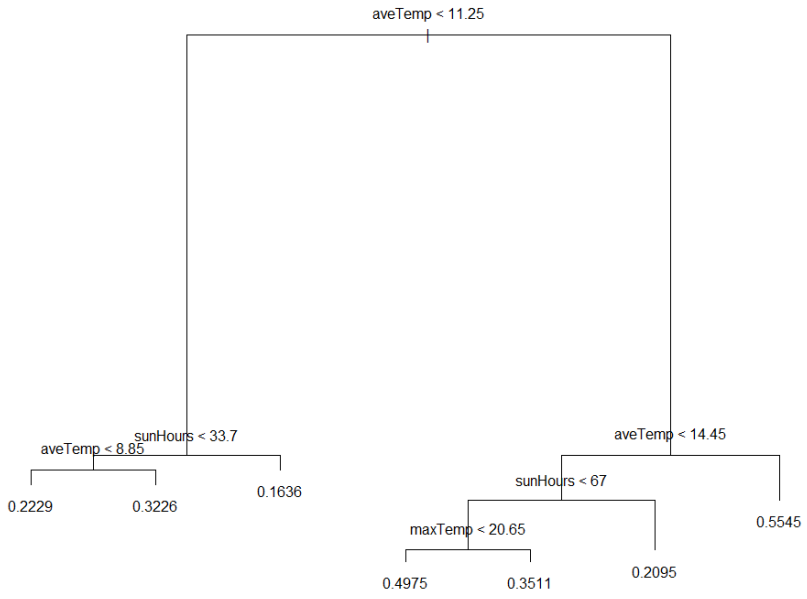

Figure 5: Tree model of the different explanatory variables as a function of the proportion.

From figure 5 we see that the average temperature is the far most important factor affection the proportion with a value of 11.25 separating the high average temperatures from the low. By investigating the left hand side of the tree model, we see that sunhours also have noticeable effect on the proportion of positive tested broilers. For sunhours below 33.7 hours pr week the average temperature above 8.85 matters. On the right hand side we see that average temperatures between 11.25 and 14.45 matters to the proportion . For sunhours below 67 in this interval of average temperature, the max temperature is also important. What we can see from this is, that there is a complex system of interactions between most of the variables and we therefore use multiple linear regression

to analyse the data. We see this complex tendency between the variables and saw on figure 4 that the relation between some of the explanatory variables and the proportion variable were humped and rather unclear. Because of this we start our analysis with the use of a generalized additive model(GAM) as seen in figure 6.

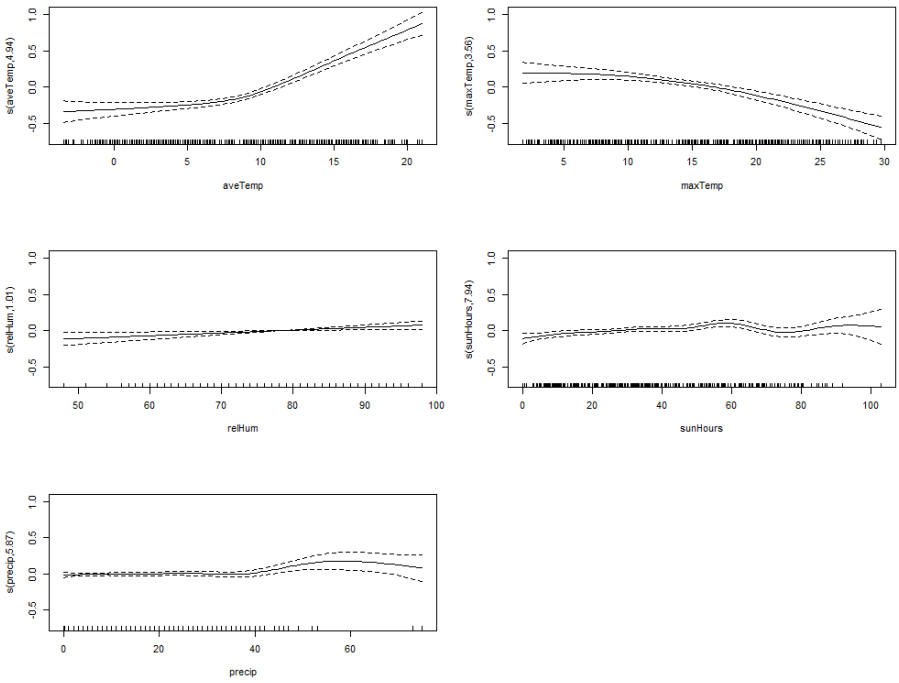

Figure 6: GAM plot of all the explanatory variables with respect to the proportion variable

We here see that average temperature and maximum temperature have a pronounced change in values and therefore those are of biggest interest. From the GAM model we choose what our maximal model should look shown in table 5

The model is reduced using step and manually remove variables that have a p value higher than 0.05. The minimized model is shown in table 3.

We check our final model by checking the plot of it. As seen in figure 7 the residuals somewhat follows a horizontal line meaning the residuals are normally distributet. The QQ plot shows the data following the QQ line and the variance seems constant.

	Estimate	Std Error	t-value	Pr(>|t|)
(Intercept)	1.120e+00	5.600e-01	2.000	0.04618 *
aveTemp	2.051e-01	1.067e-01	1.923	0.05515 .
maxTemp	-1.321e-01	9.921e-02	-1.332	0.18370
relHum	-7.606e-03	5.868e-03	-1.296	0.19567
sunHours	-5.401e-03	1.064e-02	-0.508	0.61189
precip	-4.696e-02	1.590e-02	-2.953	0.00333 **
I(aveTemp^2)	3.783e-03	2.395e-03	1.580	0.11495
I(maxTemp^2)	-7.279e-04	2.124e-03	-0.343	0.73202
I(sunHours^2)	-4.618e-05	4.733e-05	-0.976	0.32982
aveTemp:maxTemp	-1.845e-03	4.344e-03	-0.425	0.67117
aveTemp:relHum	-2.147e-03	1.072e-03	-2.002	0.04596 *
aveTemp:sunHours	5.313e-04	-1.566	0.11822	-8.318e-04
aveTemp:precip	-4.218e-04	4.627e-04	-0.911	0.36258
maxTemp:relHum	1.505e-03	1.007e-03	1.494	0.13582
maxTemp:sunHours	8.194e-04	5.329e-04	1.538	0.12490
maxTemp:precip	4.713e-04	4.641e-04	1.015	0.31047
relHum:sunHours	2.122e-05	1.112e-04	0.191	0.84882
relHum:precip	4.792e-04	1.640e-04	2.921	0.00368 **
sunHours:precip	1.986e-04	6.715e-05	2.958	0.00328 **

Table 2: Our maximum linear model

	Estimate	Std Error	t-value	Pr(>|t|
(Intercept)	1.425e+00	2.505e-01	5.686	2.45e-08 ***
aveTemp	1.504e-01	3.773e-02	3.985	7.97e-05 ***
maxTemp	-1.318e-01	2.857e-02	-4.613	5.28e-06 ***
relHum	-1.075e-02	2.916e-03	-3.686	0.000258 ***
sunHours	-3.143e-03	9.500e-04	-3.309	0.001019 **
precip	-4.110e-02	1.095e-02	-3.755	0.000198 ***
I(aveTemp^2)	1.158e-03	2.304e-04	5.026	7.46e-07 ***
aveTemp:relHum	-1.674e-03	4.583e-04	-3.652	0.000293 ***
maxTemp:relHum	1.488e-03	3.656e-04	4.070	5.62e-05 ***
relHum:precip	4.380e-04	1.185e-04	3.696	0.000248 ***
sunHours:precip	2.220e-04	5.076e-05	4.374	1.54e-05 ***

Table 3: Our minimized linear model

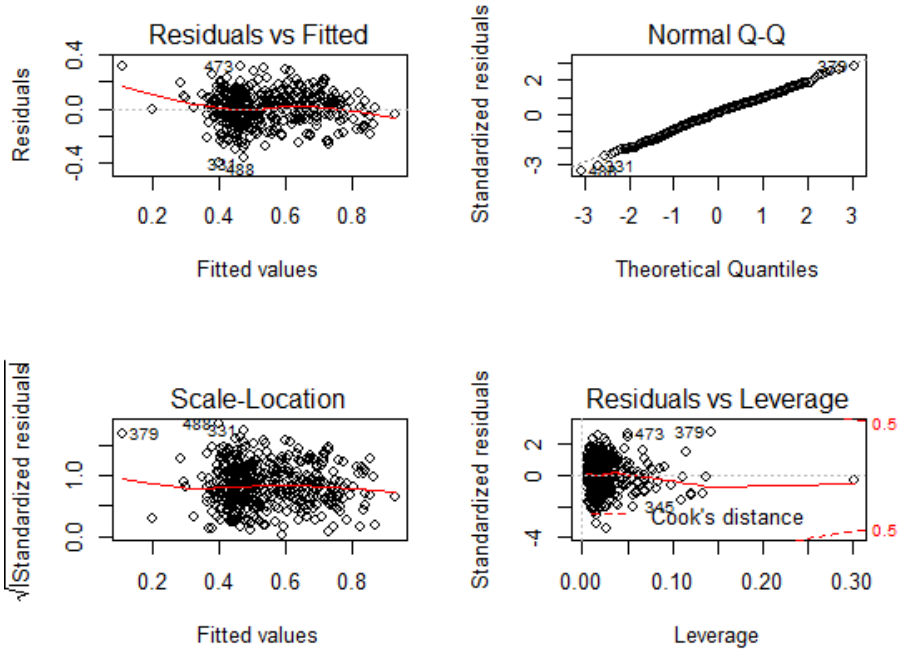

Figure 7: plot of the final model. The residuals are normally distributet and the variance is constant

Since the minimized model comes from the maximal model, an ANOVA test is made between the maximal model and the minimum model to check whether there is a significant difference.The null hypothesis is, that there is no difference between the two models. This ANOVA test is shown in 4 and gives us a p value on 0.2463. We cannot reject the null hypothesis and therefore we will carry on the analysis using the minimized model.

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>|t|)
1	408	5.7594				
2	416	5.9052	-8	-0.14579	1.291	0.2463

Table 4: ANOVA test of minimized model and maximum model. There is no significant difference between the model, which makes us happy.

In figure 8 the minimized model is plotted with the proportion of positive broilers on y axis and the explanatory variables on the x axis. A 95 procent prediction interval and a 95 procent confidence interval is added.

From this we see that most of the data lies within the 95 procent prediction interval from the model. Although it seems as though the model have some

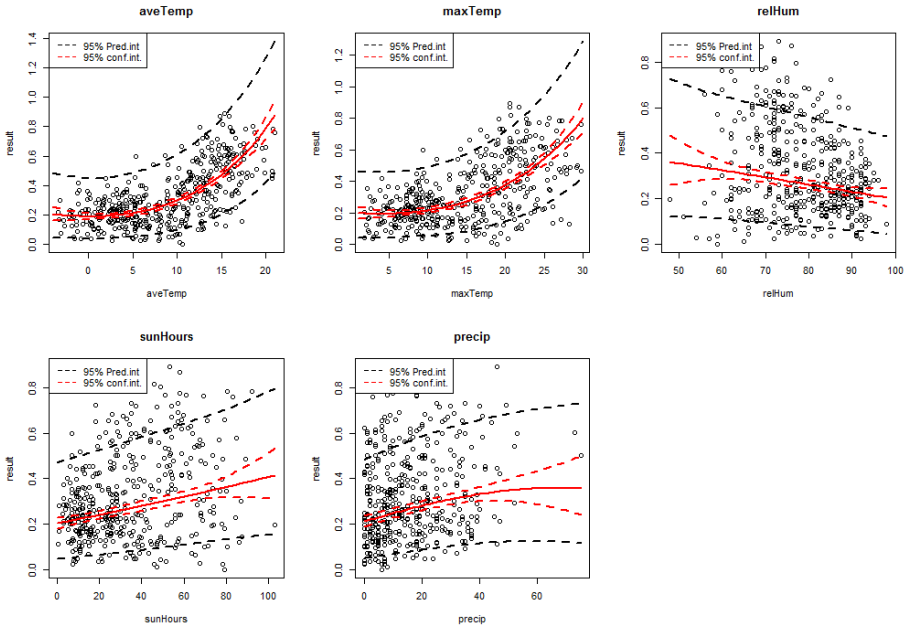

Figure 8: The model is plotted along with 95 procent prediction interval (black dashed lines) and the confidence interval (red dashed lines)

trouble explaining some of the datapoints concerning the relation between sun-hours, precipitation and humidity but still most of the points lies within the 95 procent prediction interval. It is clearly seen that the average and maximum temperature is explained well by the model leaving only a few points outside the prediction interval. From these two explanatory variables you also see the biggest correlation. Notice that this doesn't necessarily mean it's the biggest influence on positive broilers since hidden interaction between the explanatory variables should be taken into account.
In order to reveal some of the interactions with the temperature and the other explanatory variables we make a contour plot of the maximum value and the average as shown in 9

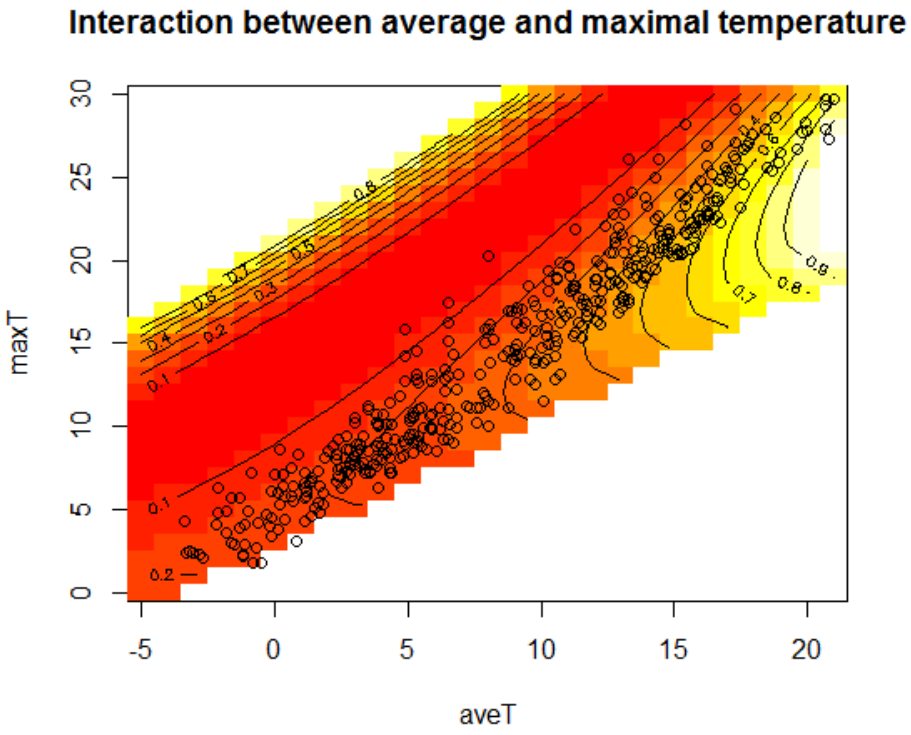

Figure 9: Contour plot of maximum temperature on the y axis and the average temperature on the x axis. The colors indicate red = low proportion of positive broilers while white = high proportion of positive broilers.

The colors in the plot in figure 9 is given as in red is low proportions of positive broilers and white is high proportions of positive broilers. Since the colors range from very red to very white, it is clear that the temperature have a strong

relation to the proportion. From the figure, we see a not completely homogeneous linear tendency between the two explanatory variables. At some values of average temperature the proportion actually gets lower as the maximum temperature increases. It still seems like the average temperature have a strong clear relation to the proportion meaning higher average temperatures giving higher proportion of positive tested broilers. The highest proportion is found when the average temperature and the maximum temperatures are not too far from each other and generally with high average temperatures. Low proportion amounts are found when the average temperature is low and the maximum temperature is around the same and also in general when the maximum temperature is much higher than the average temperatures. This mean that great amplitude in max temperature actually lower the proportion of positive tested broilers.

4.1 Regional differences

We want to investigate any regional differences in the proportion of slaughtered broilers that were tested positive.
We first try to make an ANCOVA model with the year and region as factors and the average temperature as continuous variable however doing do shows the plot in figure10

From figure 10 we see that the residuals are not normally distributed and therefore one of the assumptions doing an ANCOVA is violated. Therefore further investigation concerning regional differences cannot be done this way.

Instead we use the Pearsons Chi-square test. To do so the numbers of positive and negative tested broilers are summarized from each region as shown in table 5.

Region	POS	NEG
1	1594	3343
2	1541	3166
3	2531	4783
4	1197	1367
5	1557	4213
6	1159	2244
7	667	1262
8	996	2159

Table 5: Showing the numbers of positive and negative tested broilers in each region

To test whether there is a difference, Pearson's chi-squared test is used. It takes the values and compare them to theoretical expected values generated from the chi square distribution. Out null hypotheses is, that there is no difference between the groups and we test it with $\alpha = 0.05$ criteria. The test yields

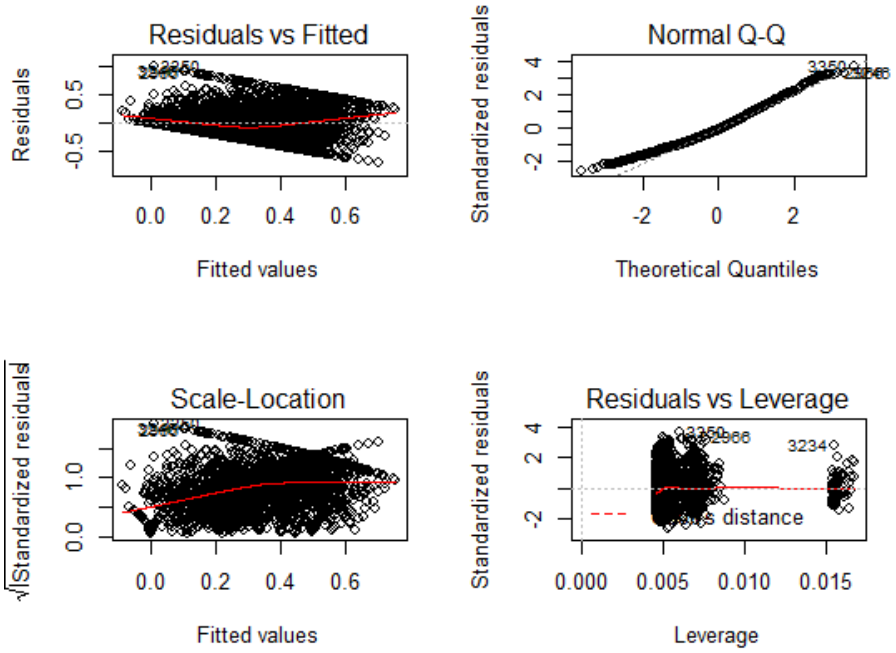

Figure 10: plot of residuals with an ANCOVA model between year, region and average temperature

following result:

Pearson's Chi-squared test
data: RegionNP
X-squared = 329.604, df = 7, p-value < 2.2e-16

From this we see that the p value is significantly low ($p < 0.05$) and we therefore reject the null hypothesis and conclude that there is a difference between the regions and the amount of positive tested broilers. By using a mosaicplot in R we see the residuals and by that which regions produced more positive broilers than expected and which region produced less positive broilers than expected. This plot is show in figure 11

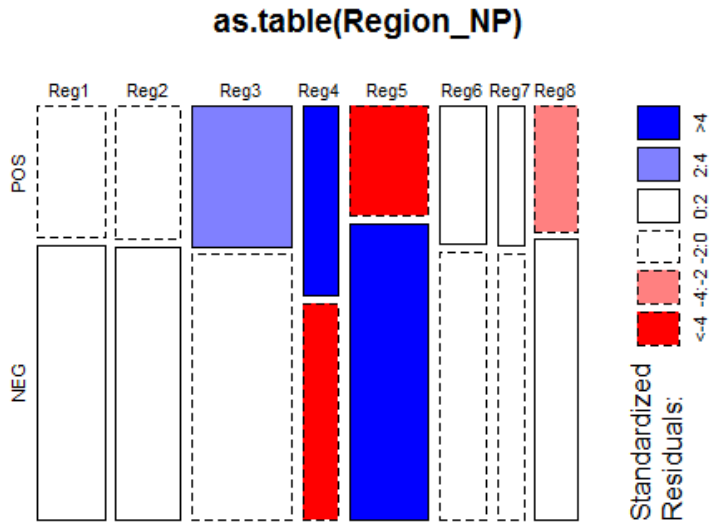

Figure 11: Mosaic plot of the residuals. It is seen that region 5 and 8 have a lower number of positive broilers compared to expected values, and region 3 and 4 have higher number of positive broilers compared to expected values

From this plot it is clear, that in region 5 the number of positive broilers are significantly lower (with more than 4 standard deviations) than the expected amount. The number of negative broilers are in this region is corresponding higher than expected. In region 8 the number of positive tested broilers is also lower than expected from the chi squared test but within 2 to 4 standard

deviations. Region 4 has a significantly higher amount of positive tested broilers with more than 4 standard deviations away from expected values. Region 3 is also seen to have more positive broilers than expected with values between 2 to 4 standard deviation away from expected. An overview of the residuals concerning the proportion of positive tested broilers is given in figure12

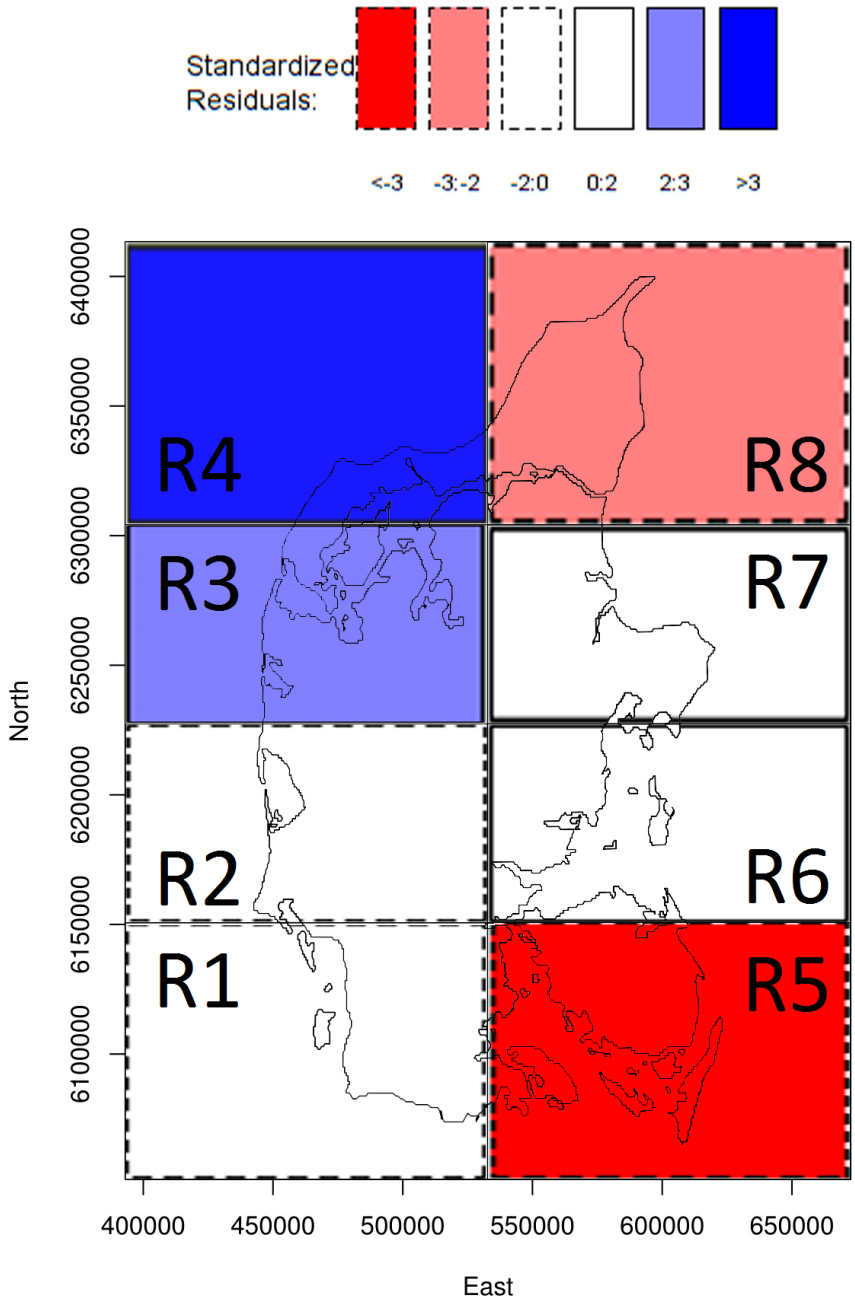

Figure 12: Map of regions and the residual colors plotted. Blue means that there was more positive broilers than expected and red means there were less positive broilers than expected in the particular region

5 Discussion and future work

We used the GAM model to choose which variables should be included and whether we should fit the model with linear terms or quadratic terms. From the GAM plot alone it was not perfectly clear which transformation of the explanatory variables should be used. We chose to use quadratic terms in average temperature, maximum temperature and sunhours since those variables showed curves on the GAM plot (especially average and maximum temperature). Also we only included second order interaction since interactions of higher order would be very hard to explain physically. It would have been interesting to see if an even better combination of quadratic terms and second order terms exists. Instead of using a quadratic term concerning the average temperature we could also have chosen a piecewise linear function with a breakpoint on 11.25 (Value taken from the tree model).
From our plot of the prediction interval we could see, that it is fairly likely that there is a model fitting our data in a better way concerning sunhours,humidity and preciption. But since we had the feeling that the temperatures are the absolute most significant variable concerning the proportion we concluded that our model was usable but not perfect. During our test of regional differences we were not able to produce a proper model using ANCOVA. This should be possible but goes beyond our scope for this project. Instead we used a Chi-Squared test and counted the positive and negative tested broilers. However we discovered that there were 33749 observations when we split it in positive and negative from each region but 39862 observation in total. This means some of the observation (a not completely insignificant amount) has not been assigned a location and is therefore missing during the test. We have seen through the Chi Squared test that there is a difference in the proportion of positive tested broilers within the regions. Therefore it could have been extra interesting to go back and try to look at the ANCOVA model including regions and climate (and try to make it work) since different region might have different climate.

6 Conclusion

In this project we have made a model,using multiple regression, trying to explain data from broiler farms and climate recorded in the years 1998-2008. From the model we can conclude that all climatic variables are significant to the amount of proportions but especially the weekly average temperature have a clear and pronounced influence on the amount of positive broilers. Changing the weekly average temperature from 0-20 degrees changes the proportion of positive broilers to around 80-85 procent. We can also conclude that having drastic changes in temperatures decreases the proportion of positive tested broilers.
We saw there was a regional difference especially between north west Jutland and Funen where Funen less proportion than expected and north west Jutland showed a higher proportion than expected.

7 Appendix