Longitudinal Data Analysis of Asthma and Wheezing in Children

Christian Dehlendorff

Kongens Lyngby 2007
IMM-M.Sc-2007-6

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

Summary

This thesis deals with statistical modelling of asthma and wheezing symptoms in childhood. The main purpose of the thesis is to search for and explore patterns in the occurrence of wheezing in order to understand the development of asthma in childhood.

Asthma has grown to be the most common chronic paediatric illness. The Copenhagen Study of Asthma in Childhood (COPSAC) maintains and collects data from a cohort of high-risk children. The objective of COPSAC is primarily to investigate into the causes of increasing asthma prevalence in society and identify methods for reducing the symptoms and discomforts hereby. Symptom diaries from COPSAC have been analysed in this thesis to find patterns in the occurrence of wheezing.

The thesis deals with a variety of statistical methods with emphasis on longitudinal data analysis. The applied methods include latent class regression, linear and non-linear mixed effects models, generalized estimating equations and logistic regression. The aim of the analysis has been to find sub-groups (or clusters) of children with the same longitudinal development of wheezing symptoms in order to gain understanding of the dynamics involved in the occurrence of wheezing. The analysis has been performed on different response scales and time-scales to investigate the consistency of the results. Furthermore, a comparison between sub-groups and the subsequent asthma diagnosis at the age of 5 years has been done.

Besides the analysis of sub-groups, a number of issues have been addressed in the thesis, including: analysis of seasonal variations in the occurrence of wheezing, analysis of the impact of medication use, and analysis of risk factors with respect to occurrence of wheezing and later diagnosis of asthma. Together these analysis provides

insight into the development of asthma in childhood.

Results

The thesis shows that the children in the cohort can be subdivided into three groups according to their symptom patterns. The three groups are, 1: Children with a high level of symptoms and an increasing symptom-rate until the age of 3 years, 2: Children with a medium level of symptoms initially and a decreasing rate and 3: Children with a low initial symptom-rate and a decreasing or constant rate. Analysis shows that the first group corresponds to the asthmatic group and the low and middle group to the non-asthmatic group. The agreement between group and diagnosis is satisfactory. The results coincide well with important, but sparse, results from literature.

The risk-factor analysis shows that the congenital resistance measured at the age of 1 month is related to the risk of being diagnosed as asthmatic at the age of 5 years. Wheezing and asthma is much more frequent in children with a low congenital resistance compared to children with an above average congenital resistance. No significant risk-factors besides age, season and diagnosis were found for the week to week symptoms. The seasonal component shows that one period corresponds to one year and that the risk of symptoms is higher in winter compared to summer. The seasonal component measures symptoms unrelated to asthma, since the seasonal component is seen to be common for all children.

The thesis shows that accurate predictions of the asthma diagnosis can be obtained by finding patterns in the yearly symptom-rates, since asthmatic and non-asthmatic children have different symptom characteristics for the relation between symptoms and age.

Resumé

Dette eksamensprojekt omhandler statistisk modellering af astma og hvæse symptomer i barndommen. Hovedformålet med eksamensprojektet er at søge efter og undersøge mønstre i forekomsten af hvæse symptomer med henblik på at forstå astma i barndommen.

Astma er vokset til at være den mest almindelige kroniske pædiatriske lidelse. COPenhagen Study of Asthma in Childhood (COPSAC) varetager and indsamler data fra en kohorte med højrisiko børn. Målet med COPSAC er primært at undersøge årsagerne til den stigende astma prævalens og identificere metoder til at reducere symptomerne og generne ved dem. Symptomdagbøger fra COPSAC er i dette projekt analyseret for at finde mønstre i forekomsten af hvæse symptomer.

Projektet omfatter forskellige statistiske metoder med hovedvægt på den aldersmæssige data analyse. De anvendte metoder inkluderer latent class regression, lineær og ikke-lineær mixed effects modeller, generalized estimation equations and logistisk regression. Målet med analyserne er at finde undergrupper (eller clustre) af børn med samme aldersmæssige udvikling i symptomer for at forstå dynamikken i forekomsten af disse. Analysen blev udført på forskellige skalaer for såvel tid som respons for at undersøge konsistensen af de fundne resultater. Endvidere blev en sammenligning af undergrupperne og den efterfølgende astma diagnose udført.

Udover analysen af undergrupper er et antal andre områder blevet undersøgt i projektet, disse inkluderer: Analyse af sæsonmæssige variationer i forekomsten af astma og hvæse symptomer, analyse af effekten af medicinering samt en analyse af risikofaktorer med hensyn til forekomsten af hvæse symptomer og senere også astma diagnosen. Sammen bibringer disse analyser indsigt i udviklingen af astma i barndommen.

Resultater

Projektet viser, at børnene i kohorten kan inddeles i tre grupper efter deres symptommønstre. De tre grupper er, 1: Børn med en høj forekomst af symptomer som er stigende indtil tre års alderen, 2: Børn med en middelhøj forekomst af symptomer i de første leveår men med en aftagende forekomst og 3: Børn med en lav start forekomst, der er faldende eller konstant. Analysen viser, at den første gruppe svarer til den astmatiske gruppe, samt at den lave og den mellemste gruppe svarer til den ikke-astmatiske gruppe. Overensstemmelsen mellem grupper og diagnoser er tilfredsstillende. Resultaterne passer endvidere godt med vigtige, men sparsomme, resultater fra litteraturen.

Risikofaktoranalysen viser, at den medfødte følsomhed ved 1 måneds alderen er relateret til risikoen for at have astma ved 5 års alderen. Hvæse og astma symptomer er mere hyppige blandt børn med en høj medfødt følsomhed sammenlignet med børn med en gennemsnitlig følsomhed. Ingen andre risikofaktorer udover alder, sæson og diagnose blev fundet for risikoudviklingen i uge til uge symptomer. Sæsonkomponenten viste, at en periode svarer til et år, samt at risikoen for symptomer er højere om vinteren sammenlignet med sommeren. Sæsonkomponenten måler symptomer, der ikke er relateret til astma, da sæsonkomponenten er fælles for alle børn.

Projektet viser, at præcise prædiktioner af astma diagnosen kan opnås ved at finde mønstre i årlige symptom-rater. Astmatisk og ikke-astmatiske børn har forskellige mønstre for sammenhængen mellem symptomer og alder.

Preface

This thesis was prepared at Informatics and Mathematical Modelling (IMM), the Technical University of Denmark in partial fulfillment of the requirements for acquiring the master degree in engineering. The thesis was supervised by Klaus Kaae Andersen at IMM and co-supervised by Per M. Bruun Brockhoff, IMM, and Hans Bisgaard from the Copenhagen Studies on Asthma in Childhood at Danish Pediatric Asthma Center.

The thesis deals with finding patterns in the longitudinal development in asthma and wheezing symptoms in childhood. The main result is the identification of three prototypes of symptoms and their connection to asthma.

Lyngby, January 2007

Christian Dehlendorff

Acknowledgements

I thank my supervisor Klaus Kaae Andersen for his commitment and enthusiasm in my work this thesis. Klaus has given many inputs, ideas and has been a great help in structuring the thesis and interpreting the results. Furthermore, I would like to thank Klaus for giving me the opportunity to continue my studies as a PhD-student at IMM. Per M. Bruun Brockhoff one of my co-supervisors has been a good support and has contributed with valuable inputs along the project.

I would also like to thank Hans Bisgaard from the COPenhagen Study of Asthma in Childhood (COPSAC) for collecting the data and for allowing me to analyzing them. Hans has been helpful in understanding and interpreting the biological and medical parts of my work as co-supervisor on the project. Furthermore, Malene Starup Stage and the rest of the COPSAC-group are to be thanked for helping with the data exchange.

Finally, my girlfriend Maiken needs a special thanks for not only being supportive and patient but also for giving useful comments and guidance on the biological and medical parts of the project.

Contents

Chapter 1

Introduction

The data analyzed in this project originates from the COPenhagen Study of Asthma in Childhood (COPSAC) [4], which is a longitudinal cohort for children having mothers with asthma. Since asthma to some extend is heritable, the children are thought as being at high risk of having or getting asthma. The data divides in three parts: Risk-factors, symptom diaries and diagnoses. The children are followed from birth and in this thesis to the age of 5 years.

The risk-factors are accessed at birth or at the age of 3 years and were analyzed in the preparatory thesis [13]. In the thesis a number of significant risk-factors and confounders were found with respect to the congenital lung-function in terms of the forced expiratory volume (FEV) and the dosis to give an 15 % decrease in the partial pressure of oxygen in the capillaries (PD15 PtcO2). The first measurement describes the lung-function in terms of size, whereas the other describes the resistance or responsiveness to the provocation.

FEV was seen to be correlated with the length at birth and the age of measurement, which leads to a corrected measure. This measure corrects the measurement for age and length. FEV was furthermore seen to be positive correlated with the gestational age and negatively with the mother smoking in the third trimester. PD15 PtcO2 was seen to be positively correlated with mutations in the fillagrin gene, which imply that having the gene-mutation increases the resistance at birth. The fillagrin gene-mutation has been shown to be related to eczema (atopic dermatitis) in the study by Palmer et al. [29].

At the age of 3 years the lung-function was evaluated by means of the specific resistance in the airways (sRAW) measured before and after a bronchodilator treatment, which opens the airways. The analysis showed that use of allergy quilt was correlated with sRAW and that a high pre bronchodilator sRAW gave the largest relative reductions after treatment.

For the diary data some analysis has been done in the preparatory thesis [13], which consisted of a population study of the prevalence and incidence and an analysis of the individuals in a mixed effects model. The population study will be reanalyzed in Chapter 2, where episodes lasting 3 days or longer are included. For the individual symptoms the redefinition of an episode leads to a new analysis of the mixed effects model, which is analyzed on two different scales. In the preparatory thesis all episodes were analyzed, now only episodes lasting 3 days or more are analyzed, since the third day with symptoms requires a visit at the COPSAC facilities. This imply that the third day with symptoms is an indicator of a more severe episode and the response is therefore seen to be different.

The diary data consists for each child of records of periods with symptoms and records of periods where the diary has been kept and validated. A day outside the recorded periods contains no information on whether the child has an episode at the day in question. The population study is based on aggregated episodes over children, whereas the individual symptom study is based on aggregated episodes over a year or a week for each child, i.e. contrasting between the number of sick children and the number of sick-days for a child. Furthermore, the individual time-series are considered on a week to week basis to analyze these changes.

An asthma diagnosis at the age of 5 years is established. The diagnosis is based on the symptom-picture in the fifth year of life and was not know until after the analysis in Chapter 2 and 3 had been carried out. Finally records of relevant medication have been kept, i.e. medication related to treatment of asthma and wheezing symptoms. Both use of medication and the risk-factors can be used as explanatory variables in the different model for the individual symptoms.

CHAPTER 2

Temporal development in population aggregated symptoms

Contents

2.1 Introduction

In this chapter the symptom-data from the COPSAC cohort is analyzed on a population level. Two measures are central in the analysis of the cohort, namely the prevalence and the incidence. The definition of an episode is a recorded incident lasting at least 3 days, since the children are required to consult the COPSAC clinic with episodes lasting 3 days or longer. This should hopefully filter the mildest episodes away and thereby help to make the distinction between mild and more severe episodes. It is furthermore clear that the third day with symptoms is special, due to the design of the study and a difference between episode lasting 1 and 2 days and episode lasting at least 3 days may be present. By excluding the short episodes, the remaining episodes may be more homogeneous, i.e. reflecting the same types of symptoms.

2.2 Descriptive analysis

The data consists of records of wheezing episodes for the COPSAC children [4]. For each episode, a startdate and a finishdate is given as well as the COPSAC number and the date of birth for the child. Furthermore, a table with records of start and end dates of periods, where the diary has been kept and validated, is included. The diaries are not complete, since they have not been kept at all time. The type of data is illustrated in Figure 2.1, i.e. showing the temporal connection between kept diary and recorded symptoms. It is clear that symptoms can only occur in periods, where the diary has been kept and validated.

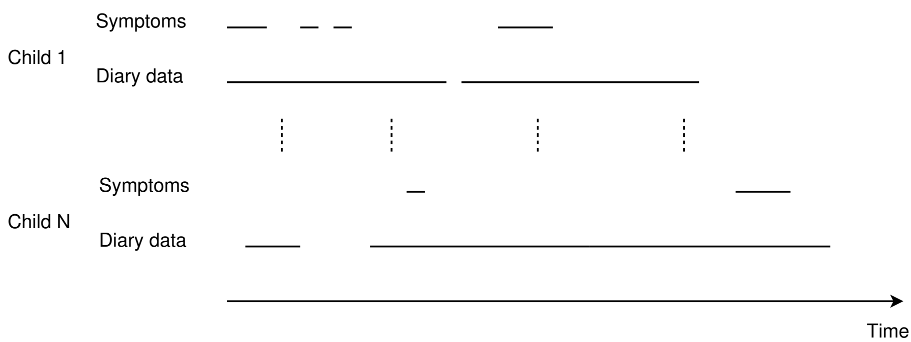

Figure 2.1: Illustration of diary data. Horizontal lines indicates, where the dairy has been kept and when symptoms have been present.

The size of the study population over time is shown in Figure 2.2, which shows that the number of participants younger than 5 years peaks around the end of December 2001 with 341 participants. The population is seen to have more than 100 participants from approximately January 1st 2000 to January 1st 2006. The overall percentage of boys is 49 %. The severe fall in participation in summer 2003 is caused by a slip in the diaries for some of the children at the age of three years. The slip is present since the study was prolonged from 3 to 6 years, which imply that some children lack diary for the period corresponding to the shift from the first 3 years to the last 3 years.

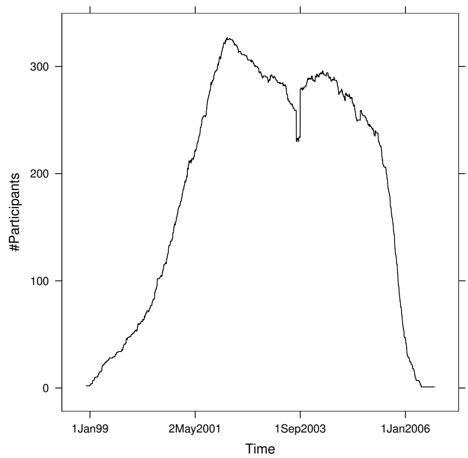

Figure 2.2: The number of participants over time

2.2.1 Prevalence

The prevalence, i.e. the number of children wheezing a given day, is shown in the left part of Figure 2.3. The prevalence is seen to vary over the year the same way for all years, eg the prevalence is always higher in the winter compared to the summer. The difference in levels in the middle compared to the ends is mostly caused by fewer participants in these periods. The normalized prevalence, the prevalence at day t divided by the number of participants at day t, is plotted in the right part of Figure 2.3. From this figure the season effect is even more clear and shows that the intensity of the prevalence is varying from year to year.

Analysis of the seasonal effect the analysis should be reduced to the middle five years from June 1999 to June 2005 to have a reasonable number of children in the study. This will reduce the uncertainty of the prevalence, since the uncertainty of the prevalence is inverse proportional to the number of children, which can be seen from the fluctuations in the beginning and end of the considered time range.

Yearly trend

The yearly trend in the time-series of the prevalence can be investigated by the autocorrelation function, Madsen p. 102 [25]

$$ACF(k) = Cov(y_t, y_{t-k})/Cov(y_t, y_t) \tag{2.1}$$

For a given lag/time-difference k, the autocorrelation-function measures the correlation between measurements taken k timesteps apart. A 95 % confidence interval is $\pm\sqrt{1/N}$, where N is the number of time-points.

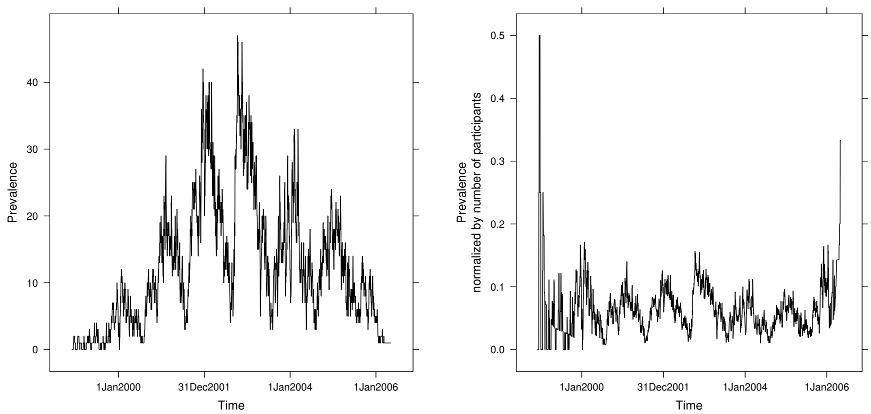

Figure 2.3: The prevalence over time. Left: the prevalence, right: the relative prevalence

From Figure 2.4 the yearly trend is clear, furthermore the prevalence is strongly persistent from day to day, i.e. the correlation decays slowly. The correlation for days 30 days apart is $ACF(30) = 0.53$, and for 1, 2 and 5 days apart 0.96, 0.92 and 0.80, respectively. From the upper right part of the figure it is seen that the correlation for days one year apart is 0.41 and 0.29 for days two years apart. Hence the seasonal trend is strong and persistent even over long time, which imply that one year tends to look like the other.

In the considered time range, the correlation is seen to be significant both the short term and the long term correlation. The long term correlation is related to the yearly patterns, whereas the short term correlation is related to a carry-over effect. The prevalence at day t is seen to be the prevalence at day $t-1$ plus the number of new children with symptoms and minus the number of children exiting an incident. The correlation is very persistent and one method to stabilize the correlation could be the considered the differense, i.e. $\nabla Y_t = Y_t - Y_{t-1}$. The difference correspond to the number of new children with symptoms (the incidence) minus the number of children exiting an incident. In section 2.2.2 the incidence is analyzed, which should be less correlated compared to the prevalence.

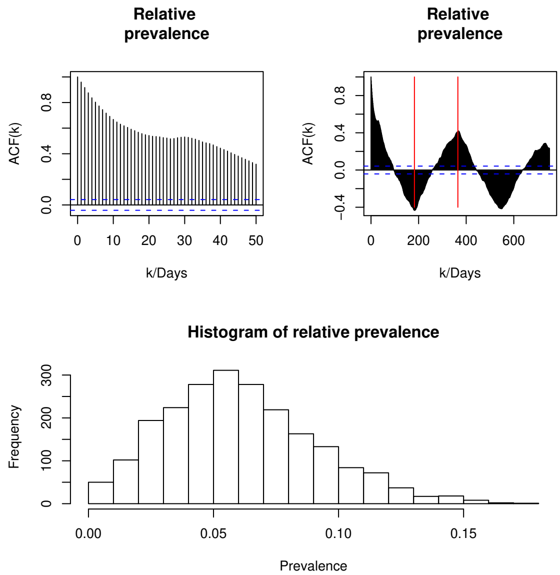

Figure 2.4: Autocorrelation (top) for the prevalence (the red vertical lines correspond to $\frac{1}{2}$ year and 1 year difference) and histogram for normalized prevalence (bottom)

2.2.2 Incidence

As mentioned the incidence may an interesting measure and given as the number of new wheezing episodes each day. The temporal development of the relative incidence is shown in Figure 2.5 and a yearly variation is seen to be present. The pattern is weaker compared to the prevalence. It is also seen that the relative incidence is much higher in the beginning, which is caused by the large impact an individual has in this parts since the number of children is sparse. The incidence is lower compared to the prevalence, which can be explained by the fact that the new cases are included in the prevalence as well as the old cases. This imply that the prevalence at day t is at least as large as the corresponding incidence.

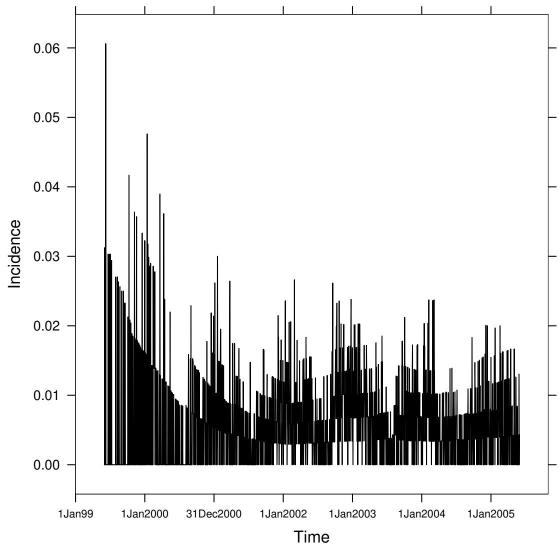

Figure 2.5: Incidence over time

From Figure 2.6 it is seen that the correlation is much weaker compared to the prevalence, eg $ACF(1) = 0.13$ and $ACF(31) = 0.12$. The difference in correlation can be explained by the fact that the prevalence can be formulated as the dynamic process:

$$\text{prevalence}_t = (1 - \lambda_t) \cdot \text{prevalence}_{t-1} + \text{incidence}_t \tag{2.2}$$

where λ_t is the rate by which wheezing children comes out of an episode at day t. It is seen that there is a carry-over effect in the prevalence, since children having an episode at day i are likely to have an episode at day $i + 1$.

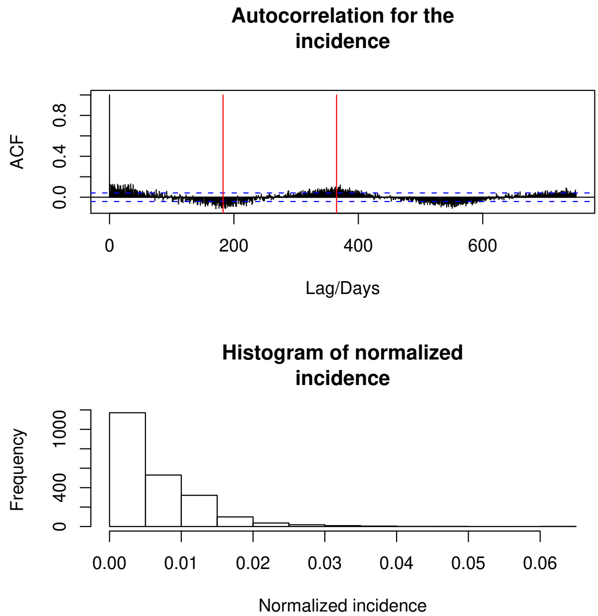

Figure 2.6: Autocorrelation (top)for the relative incidence (the red vertical lines correspond to $\frac{1}{2}$ year and 1 year difference) and histogram for the normalized incidence (bottom)

2.3 Seasonal effect in symptoms

In the following section an analysis of the prevalence and incidence, respectively, will be carried out. The analysis is done by fitting a generalized linear model to the two measures. Both the prevalence and incidence are characterized by being counts in the non-relative version, hence a poisson distribution with an intensity parameter may be fitted. The objective for the model is to describe the seasonal variations seen in the previous sections, which can be utilized in removing the seasonal component in the individual timeseries of wheezing episodes.

2.3.1 Prevalence

The prevalence over time as analyzed in section 2.2.1 may be modelled by a generalized linear model. The prevalence at a given day is assumed to be poisson-distributed with parameter μ_t. The poisson distribution probability function is defined as

$$f(y_t;\mu_t) = \frac{\mu_t^{y_t} e^{-\mu_t}}{y_t!} \tag{2.3}$$

The number of participants varies over time, which imply that the relative prevalence is modelled as noted previous. The link-function is the canonical link for a poisson-model $g(\mu_t) = \log(\mu_t)$ and in the following the quasi-poisson distribution will be used implying that the variance function is changed from $V[Y_t] = \mu_t$ to $V[Y_t] = \phi\mu_t$, where ϕ is the overdispersion parameter, see Wood p. 59-75 [42]. The overdispersion parameter is needed in cases where the variance is too large in the data compared to the theroretical variance. The model for the mean of the relative prevalence is

$$\log\left(\frac{\hat{\mu}_t}{n_t}\right) = \mathbf{X}\boldsymbol{\beta} \Leftrightarrow \log\left(\hat{\mu}_t\right) = \log\left(n_t\right) + \mathbf{X}\boldsymbol{\beta} \tag{2.4}$$

where $\log(n_t)$ is an offset (or baseline) for the relative prevalence and μ_t is the mean value in the poisson distribution. As explanatory variable only the season (time) is considered for now. It was seen from previous analysis (section 2.2.1 and 2.2.2) that a time-trend is present, which is a yearly pattern with a peak at in the winter and a minimum at summer.

An initial generalized additive model, see Wood p. 121-140 [42], given as

$$\log\left(\hat{\mu}_t\right) - \log\left(n_t\right) = s_1(t) \tag{2.5}$$

is fitted and the corresponding smoother for the time variable is shown in Figure 2.7. $s_1(t)$ is a smoothed function estimated with the data in order to describe the curvature in the relation between $\log(\mu)$ and t. The smoothed function is based on a thin plate regression spline: The default for the procedure, see Wood p. 154-160 and 226 [42]. The number of degrees of freedom for the smoother is chosen such that the generalized cross validation score is minimized, see Wood p. 178 [42]. The generalized cross validation score is given as

$$\mathrm{GCV} = \frac{nD(\hat{\boldsymbol{\beta}})}{(n - \mathrm{tr}(\mathbf{A}))^2}$$

where $\mathbf{A}$ is the influence matrix (hat matrix) and $D(\hat{\beta})$ the deviance for the model with parameters $\hat{\boldsymbol{\beta}}$, see Wood p. 70 [42]. The deviance is defined as

$$D = 2 \cdot (l(\text{full model}) - l(\text{current model}))$$

where l is the likelihood and the full model corresponds to the saturated model, i.e. a model with as many parameters as observations. The hat matrix describes the relation between the observations and the fitted mean value, i.e. $\hat{\boldsymbol{\mu}} = \mathbf{A}\mathbf{y}$.

The deviance is 5894 on $2189 - 8.63 - 1$ degrees of freedom. The overdispersion, Wood p. 71 [42], is estimated to be

$$\hat{\phi} = D/(n - p) = 5894/2179.37 = 2.7$$

This imply that the variance is 2.7 times as large as it should be if the data really is poisson distributed. The quasi-poisson distribution is seen to be necessary if the deviance can not be reduced be a different (better) description of the prevalence.

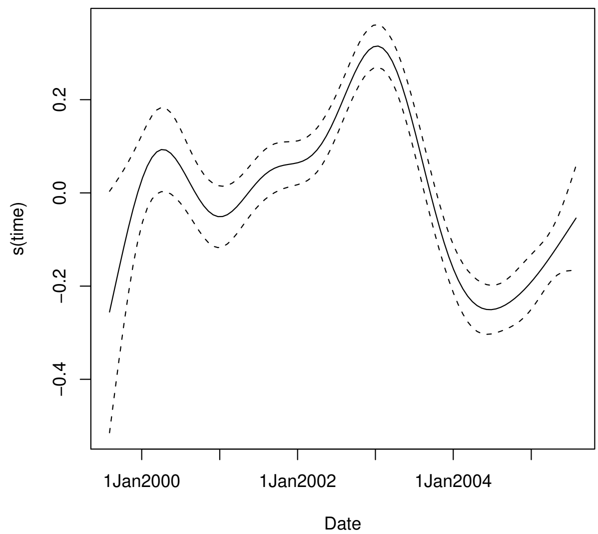

Figure 2.7: Smoothed curve for $s(t)$ in generalized linear model

From Figure 2.7 it is seen that there is a seasonal variation over the year, which was also seen from Figure 2.3. The oscillations could be described by the periodic function: $\cos(t)$. It is furthermore clear that the period should be one year, the amplitude may vary from year to year and peak at different times from year to year and finally the prevalence may have a different baseline for each year. This can be modelled by using one of the relations for the trigonometric functions and parameterize $s(t)$ as

$$\begin{aligned} s(t) =& a_{\text{year}} + A_{\text{year}} \cdot \cos\left(\frac{t \cdot 2\pi}{365} + \theta_{\text{year}}\right) \\ =& a_{\text{year}} + A_{\text{year}} \cdot \cos\left(\frac{t \cdot 2\pi}{365}\right) \cdot \cos(\theta_{\text{year}}) \\ & - A_{\text{year}} \cdot \sin\left(\frac{t \cdot 2\pi}{365}\right) \cdot \sin(\theta_{\text{year}}) \end{aligned} \tag{2.6}$$

where A_{year} is the amplitude, θ_{year} the phase (a forward shift of the maximum prevalence) and a_{year} the baseline prevalence for a given year. Including $\sin(t)$ as well as $\cos(t)$ makes it possible to estimate the phaseshift, θ_{year} and the amplitude and still maintain a linear predictor, which imply that the generalized linear model framework can be used. Updating the model in (2.5) with the parametric function for $s(t)$ and furthermore including the fraction of boys and the median age at each day yields

Model 1:

$$
\begin{aligned}
\log(\mu_t) = & \log(n_t) + \alpha + a_i + \beta_1 \cdot \cos\left(\frac{t \cdot 2\pi}{365}\right) + \beta_2 \cdot \sin\left(\frac{t \cdot 2\pi}{365}\right) \\
& + \beta_{3i} \cdot \cos\left(\frac{t \cdot 2\pi}{365}\right) + \beta_{4i} \cdot \sin\left(\frac{t \cdot 2\pi}{365}\right) \\
& + \beta_5 \cdot \text{sexfrac}_t + \beta_6 \cdot + \text{median age}_t \\
& i \in \{2000, 2001, 2002, 2003, 2004, 2005\}
\end{aligned}
\tag{2.7}
$$

The model has a separate prevalence baseline rate, amplitude and phase for each year. The summary for the model is given in Table 2.1, from which it is seen that the estimated overdispersion, $\hat{\phi}_1 = 1.23$, is rather low. The overdispersion is estimated in the procedure since the quasi-binomial distribution is used as model-family, see Wood p. 74-76 [42]. A test for the overdispersion is found by finding the probability $p = Pr(X > \hat{\phi}_1)$, which is χ^2_{n-p} distributed, this gives a p-value around 1. It is noticed that the non-parametric model uses 9.63 degrees of freedom, whereas the parametric model uses 23. The plot in the first row in Figure 2.9 shows that severe

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−4.09	0.57	−7.13	<0.0001
year2000	−0.47	0.10	−4.55	<0.0001
year2001	−0.28	0.08	−3.53	0.0004
year2002	−0.12	0.09	−1.43	0.1523
year2003	−0.39	0.10	−3.76	0.0002
year2004	−0.66	0.14	−4.88	<0.0001
year2005	−0.85	0.18	−4.72	<0.0001
cosi	0.37	0.08	4.89	<0.0001
sinus	0.63	0.13	4.89	<0.0001
MedianAge	0.00	0.00	0.77	0.4442
sexfrac	3.15	1.21	2.61	0.0092
year2000:cosi	0.10	0.09	1.05	0.2951
year2001:cosi	0.03	0.07	0.42	0.6749
year2002:cosi	0.11	0.08	1.44	0.1511
year2003:cosi	−0.03	0.08	−0.35	0.7236
year2004:cosi	0.05	0.08	0.60	0.5485
year2005:cosi	0.00	0.08	0.02	0.9854
year2000:sinus	−0.31	0.14	−2.15	0.0313
year2001:sinus	−0.45	0.13	−3.50	0.0005
year2002:sinus	−0.61	0.13	−4.69	<0.0001
year2003:sinus	−0.30	0.13	−2.32	0.0202
year2004:sinus	−0.40	0.14	−2.91	0.0036
year2005:sinus	−0.24	0.17	−1.41	0.1587

Table 2.1: Model 1 (2.7), Dispersion: 1.23

dis-continuities occurs at the shift from one year to another. Changing the years to go from June 1st to May 31st and removing the insignificant term median age yields a new model with an estimated dispersion of $\hat{\phi}_2 = 1.21$. The fit has a residual deviance of 2733 on 2170 degrees of freedom. A comparison with Model 1 shows that the deviance is reduced by 73 with a reduction in the model complexity by 4 parameters, 1 for the median age and the remaining three arise since the number of unique years are reduced from 7 (1999-2005) to 6 (0-5). The updated model is seen to be a better description of the prevalence. Model 2 is given as

$$\begin{aligned}\log(\mu_t) = &\log(n_t) + \alpha + a_i + \beta_1 \cdot \cos\left(\frac{t \cdot 2\pi}{365}\right) + \beta_2 \cdot \sin\left(\frac{t \cdot 2\pi}{365}\right) + \\ &\beta_{3i} \cdot \cos\left(\frac{t \cdot 2\pi}{365}\right) + \beta_{4i} \cdot \sin\left(\frac{t \cdot 2\pi}{365}\right) + \beta_5 \cdot \text{sexfrac}_t \\ &i \in \{1, 2, 3, 4, 5\}\end{aligned} \tag{2.8}$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−3.79	0.46	−8.22	<0.0001
yearshift1	−0.06	0.05	−1.08	0.2813
yearshift2	0.06	0.04	1.62	0.1054
yearshift3	0.20	0.04	5.17	<0.0001
yearshift4	−0.19	0.04	−4.54	<0.0001
yearshift5	−0.24	0.05	−5.06	<0.0001
cosi	0.52	0.05	11.23	<0.0001
sinus	0.28	0.06	5.02	<0.0001
sexfrac	1.93	0.95	2.02	0.0430
yearshift1:cosi	−0.20	0.05	−3.69	0.0002
yearshift2:cosi	−0.11	0.05	−2.19	0.0287
yearshift3:cosi	0.02	0.05	0.48	0.6302
yearshift4:cosi	−0.17	0.05	−3.24	0.0012
yearshift5:cosi	−0.15	0.05	−3.00	0.0027
yearshift1:sinus	0.04	0.08	0.48	0.6304
yearshift2:sinus	−0.06	0.06	−1.00	0.3181
yearshift3:sinus	−0.28	0.06	−4.93	<0.0001
yearshift4:sinus	−0.16	0.06	−2.73	0.0064
yearshift5:sinus	−0.08	0.06	−1.25	0.2122

Table 2.2: Model 2 (2.8): Shifted year and without median age, Dispersion: 1.21

From the estimates of β_1, β_{3i}, β_2 and β_{4i} it is possible to estimate the yearly amplitudes and phases. This gives eg for year 2 two equations to solve for θ_2 and A_2, the phase and amplitude for year 2. The equations are described previously and gives for years 2 the following set of equations

$$\begin{aligned}\hat{A}_2 \cdot \cos(\hat{\theta}_2) &= \hat{c}_2 = 0.52 + (-0.11) \\ -\hat{A}_2 \cdot \sin(\hat{\theta}_2) &= \hat{d}_2 = 0.28 + (-0.06) \\ \hat{A}_2 \geq 0 &\wedge \hat{\theta}_2 \in [-\pi; \pi]\end{aligned} \tag{2.9}$$

c_{year} corresponds to the estimated coefficient for cosine and d_{year} to sine, i.e. the sum of the reference and the contrast corresponding to the relevant year. The solution to the equations is

$$\begin{aligned}
\hat{\theta}_2 &= \arctan(-\hat{d}_2/\hat{c}_2) = \arctan(-(0.22)/(0.41)) \\
\hat{A}_2 &= \hat{c}_2/\cos(\hat{\theta}_2) = 0.41/\cos(\hat{\theta}_2) \\
\hat{A}_2 &\geq 0 \wedge \hat{\theta}_2 \in [-\pi;\pi]
\end{aligned} \tag{2.10}$$

the latter can be simplified, since (see Figure 2.8)

$$\theta_2 = \arctan(-\hat{d}_2/\hat{c}_2) = -\arctan(\hat{d}_2/\hat{c}_2)$$

where $-\theta_2$ is the angle in a right-angled triangel with cathetes $\hat{c}_2$ (adjacent) and $\hat{d}_2$ (opposite) (the lower triangle in Figure 2.8). This imply that

$$\cos(-\theta_2) = \cos(\theta_2) = \frac{\hat{c}_2}{\sqrt{\hat{c}_2^2 + \hat{d}_2^2}}$$

and hence that

$$\hat{A}_2 = \frac{\hat{c}_2^2}{\hat{c}_2^2/\sqrt{\hat{c}_2^2 + \hat{d}_2^2}} = \sqrt{\hat{c}_2^2 + \hat{d}_2^2} \tag{2.11}$$

The standard errors of the phase shifts can be calculated by means of the law of error propagation, Conradsen p. 69 [8]. The variance for a stochastic quantity, Z given by a function, f, of N stochastic variables $(X_1, \ldots, X_N)$ is approximately

$$\begin{aligned}
s_Z^2 \approx & \left(\frac{\partial f}{\partial X_1}(\bar{x}_1, \ldots, \bar{x}_N)\right)^2 s_{x_1}^2 + \cdots + \left(\frac{\partial f}{\partial X_N}(\bar{x}_1, \ldots, \bar{x}_N)\right)^2 s_{x_N}^2 \\
& +2 \cdot \left(\frac{\partial f}{\partial X_1}(\bar{x}_1, \ldots, \bar{x}_N)\right)\left(\frac{\partial f}{\partial X_2}(\bar{x}_1, \ldots, \bar{x}_N)\right) s_{x_1 x_2} + \ldots \\
& +2 \cdot \left(\frac{\partial f}{\partial X_{N-1}}(\bar{x}_1, \ldots, \bar{x}_N)\right)\left(\frac{\partial f}{\partial X_N}(\bar{x}_1, \ldots, \bar{x}_N)\right) s_{x_{N-1} x_N}
\end{aligned} \tag{2.12}$$

where s_{x_1} is the variance of X_1, s_{x_1,x_2} the covariance of X_1 and X_2 and $\bar{x}_1$ the mean of X_1. For the phaseshift this gives

$$\begin{aligned}
s_{\theta_i}^2 \approx & \left(s_d^2 + s_{d_i}^2 + 2 s_{dd_i}\right) \frac{1}{\bar{c}_i^2 \cdot n_i^2} \\
& + \left(s_c^2 + s_{c_i}^2 + 2 s_{cc_i}\right) \frac{\bar{d}_i^2}{\bar{c}_i^4 \cdot n_i^2} \\
& - 2 \cdot \left(s_{dc} + s_{d_i c_i} + s_{dc_i} + s_{d_i c}\right) \frac{\bar{d}_i}{\bar{c}_i^3 \cdot n_i^2} \bar{c}_i^2
\end{aligned} \tag{2.13}$$

where $n_i = \left(1 + \frac{\bar{d}_i^2}{\bar{c}_i^2}\right)$, $\bar{c}_i = c + c_i$, $\bar{d}_i = d + d_i$. d is the estimate for sine for the reference year and d_i the additional contribution for the i'th year. For the reference year all parts with a subscript i is zero, which imply that $\bar{c}_i = c$ for $i = 0$.

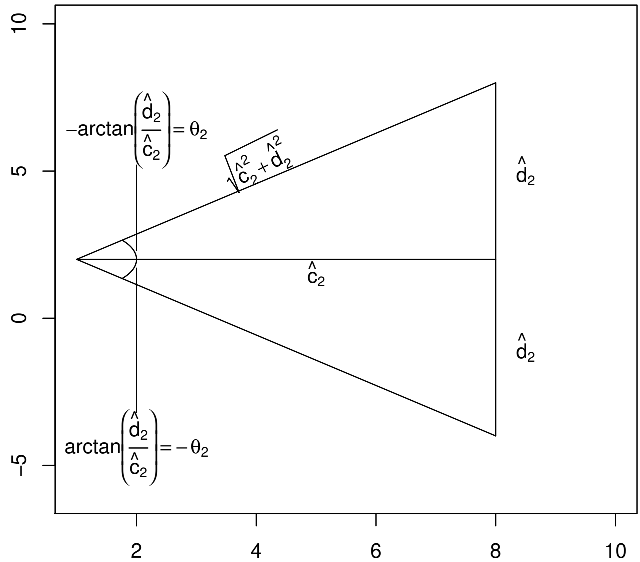

Figure 2.8: Illustration of connection between phase, amplitude and parameter estimates for periodic terms

The uncertainties can likewise be estimated for the amplitudes. This gives

$$
\begin{aligned}
s_{A_i}^2 \approx & (s_{c_1}^2 + s_c^2 + 2 \cdot s_{cc_1}) \cdot \frac{\bar{c}^2}{N} + (s_{d_1}^2 + s_d^2 + 2 \cdot s_{dd_1}) \cdot \frac{\bar{d}^2}{N} \\
& + 2 \cdot (s_{cd} + s_{c_1 d} + s_{cd_1} + s_{c_1 d_1}) \cdot \frac{\bar{c} \cdot \bar{d}}{N}
\end{aligned}
\tag{2.14}
$$

where $\bar{c}_i = c + c_i$, $\bar{d}_i = d + d_i$ and $N = \bar{c}^2 + \bar{d}^2$.

For the estimates, a table of estimated phases and amplitudes is shown in Table 2.3. It is seen that the prevalence peaks varies by $0 - (-0.8) = 0.8$ days. Furthermore, it is seen that the estimated amplitudes, the size of the yearly variations, are between 0.37 and 0.59. This gives rate-ratios of 2.1 and 3.24 for comparison of winter rates against summer-rates. From the second row of Figure 2.9 it is seen that the predicted prevalence seems to be nice and smooth, the only discontinuity is at the point, where the number of participants fall markedly. The fit to the observed prevalence is seen to be good.

From Table 2.3 it is seen that the modelling of individual phase-shifts and the ampli-

Year	c_{year}	d_{year}	A_{year}	s_A	θ_{year}	delay (days)	$s_{\theta_{year}}$
0	0.52	0.28	0.59	0.04	−0.49	28.46	0.08
1	0.32	0.31	0.45	0.02	−0.77	44.91	0.06
2	0.41	0.22	0.47	0.02	−0.49	28.62	0.04
3	0.54	−0.00	0.54	0.02	0.01	−0.49	0.03
4	0.35	0.12	0.37	0.02	−0.32	18.54	0.06
5	0.36	0.20	0.42	0.02	−0.50	29.27	0.05

Table 2.3: Interpretation for prevalence model

tudes seems to be necessary, since significant variations from year to year are seen. For a model with the same amplitude and phase for each year, the deviance increases by 265 on 10 degrees of freedom and the estimated overdispersion is $\hat{\phi} = 1.33$. The test for the increase in deviance is highly significant, since the deviance increases by 265 on 10 degrees of freedom, yielding p<0.0001. Furthermore it is seen from row 2 and 3 in Figure 2.9 that Model 2 is performing better around year 3 compared to Model 3. Model 3 can be formulated as

$$\begin{aligned}\log(\mu_t) = &\log(n_t) + \alpha + a_i + \beta_1 \cdot \cos\left(\frac{t \cdot 2\pi}{365}\right) + \\ &\beta_2 \cdot \cos\left(\frac{t \cdot 2\pi}{365}\right) + \beta_5 \cdot \text{sexfrac}_t \\ &i \in \{1, 2, 3, 4, 5\}\end{aligned} \tag{2.15}$$

The model is constrained to estimate a common seasonal part for all years. The difference in the fits for the two models is small, however Model 2 is statistically better compared to Model 3.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−2.68	0.29	−9.37	<0.0001
cosi	0.42	0.01	43.49	<0.0001
sinus	0.15	0.01	16.20	<0.0001
sexfrac	−0.21	0.58	−0.36	0.7194
yearshift1	−0.05	0.04	−1.36	0.1751
yearshift2	0.01	0.03	0.37	0.7135
yearshift3	0.17	0.03	5.15	<0.0001
yearshift4	−0.25	0.03	−7.20	<0.0001
yearshift5	−0.26	0.04	−7.07	<0.0001

Table 2.4: Model 3 (2.15), Dispersion: 1.33

A problem with Model 2 is that the residuals are correlated, Figure 2.10 shows that especially the short term correlation is present. It is seen that even the long term correlation is present but is greatly reduced. This makes the inference wrong, since the uncertainty on the parameters is too small, see Diggle et al. p. 59-73 [15]. The

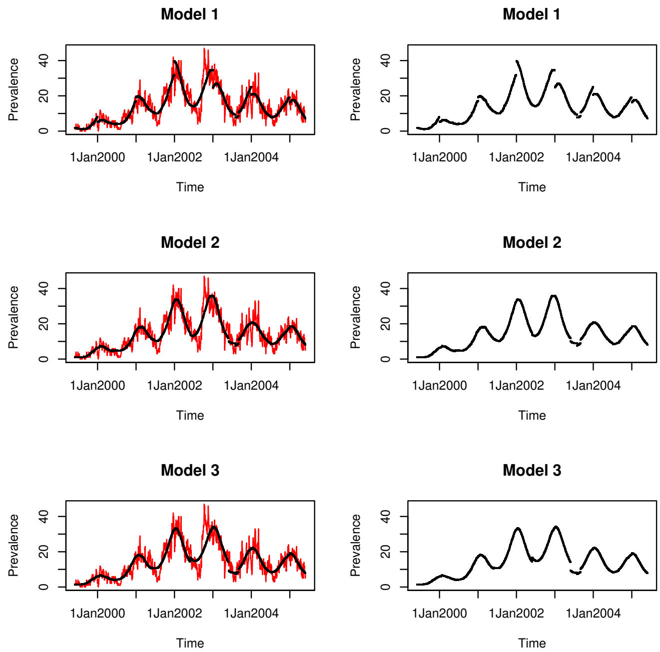

Figure 2.9: Predicted prevalence (black) and observed (red) as function of time. Model 1: The years follow the calendar years, Model 2: The years runs from June to June, Model 3: Same as Model 2, but with the same phase-shift

short term correlation can be explained by the carry over effect, children being sick at day i is more likely to be sick at day $i+1$.

From Figure 2.10 it is also seen that the deviance residuals, see Olsson p. 57 [28], seem to coincide as being standard gaussian random variables. This imply that no outliers seem to be present in the data, since they would be situated away from the straight line. The correlation is however a problem, which may be dealt with by considering the incidence, since the incidence was seen to have a much smaller short term correlation.

Interpretation

The interpretation is based on Model 2 (Table 2.2 and Table 2.3), which has the best fit of the models considered. Model 2 has a baseline level, α, corresponding to year 0 and a cohort with only females of $\alpha = -3.79$, which gives a normalized prevalence baseline of $e^{-3.79} = 0.02$. For a cohort with only boys the baseline is $e^{-3.79+1\cdot 1.93} = 0.16$, both is way out of the data range since the fraction of boys is

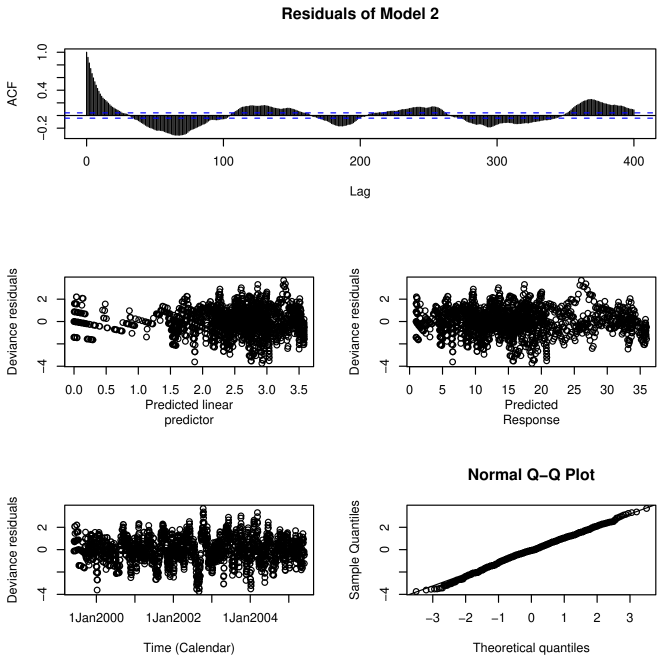

Figure 2.10: Top: Autocorrelation function for residuals from Model 2, middle left: Deviance residuals against predicted linear predictor, middle right: Deviance residuals against predicted response, bottom left: Deviance residuals against calendar time and bottom right: QQ-plot of deviance residuals

between 38 % and 75 %. However it illustrates that boys tend to have more episodes compared to girls. Considering a cohort with the same number of girls and boys, the baseline prevalence becomes $e^{-3.79+0.5\cdot 1.93} = 0.06$. Each percent point more boys gives an increase in prevalence by a factor $e^{0.01\cdot 1.93} = 1.02$.

For each year a separate rate-ratio is estimated by the model, the first year is reference and has rate $e^{a_0} = e^{-3.79} = 0.02$. For year $i \neq 0$ the rate-ratio can be calculated as e^{a_i}, which gives 0.94, 1.06, 1.22, 0.83, 0.79 for year 1, 2, 3, 4 and 5 respectively compared to the reference year.

The periodic part of the model shows that the amplitude varies between 0.37 and 0.59, which imply ratios between $e^{-0.37} = 0.69$ and $e^{0.37} = 1.45$ respectively $e^{-0.59} = 0.56$ and $e^{0.59} = 1.8$ for these two extremes. This shows that the relative prevalence is 2-3 times higher in the winter compared to the summer.

2.3.2 Incidence

In the following the incidence will be analyzed and modelled. The incidence is the number of new episodes at day t. The short term correlation is expected to be lower compared to the prevalence, since there is no carry over effect in the incidence. On the contrary children having an incidence at day i definitely do not have an incidence the next day, since two incidences are separated by at least 3 days: 1 day without symptoms after the first episode and two days corresponding to the two first days with symptoms. The lower short term correlation will hopefully imply that no correlation in the residuals is present after taking the long term correlation into account contrary to the results for the prevalence.

The initial model considered for the incidence corresponds to the final model for the prevalence. The model has different phase, amplitude and baseline level for each year and the incidence is assumed to depend on the fractions of boys of the children. Model 1 for the incidence therefore becomes

$$\begin{aligned}\log(\mu_t) = &\log(n_t) + \alpha + a_i + \beta_1 \cdot \cos\left(\frac{t \cdot 2\pi}{365}\right) + \beta_2 \cdot \sin\left(\frac{t \cdot 2\pi}{365}\right) + \\ &\beta_{3i} \cdot \cos\left(\frac{t \cdot 2\pi}{365}\right) + \beta_{4i} \cdot \sin\left(\frac{t \cdot 2\pi}{365}\right) + \beta_5 \cdot \text{sexfrac}_t \\ &i \in \{1, 2, 3, 4, 5\}\end{aligned} \tag{2.16}$$

where the years are defined the same way as for the prevalence model. The residual deviance for the model is 2449 on 2170 degrees of freedom, which gives a χ^2-test on one degree of freedom with the p-value, $p = P(X > D/\mathrm{df}) = 0.29$. This shows that the fit is adequate, i.e. that the dispersion can be assumed to be 1.

The summary for the model is given in Table 2.5 and shows that the estimated over dispersion in the quasi-poisson distribution is 1.03, which imply that the variance is close to the mean as the test indicated. A reduction to a model without the fraction of boys gives an insignificant increase in residual deviance of 0.01 on 1 degrees of freedom implying a χ^2-test with a p-value of 0.92.

The updated summary is shown in Table 2.6, which shows that further reduction is not possible, although a large subset of the parameter estimates for the periodic terms are insignificant. A test against a model without year-specific cosine and sine parts (Model 3) gives $p = 0.01$. In the middle left panel in Figure 2.11, the observed and fitted incidence are plotted in the same plot, in the right the fitted incidence alone. The figure shows that there are discontinuities, which is seen to be present for both model 2 and 3. However, these are not severe and the problem is not solved by complicating the model by re-including the fraction of boys. The discontinuities arise, since the model do not force the curve to be continuous in the year-shifts and furthermore at the time-point with a high number of missing diary-data.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−5.06	1.37	−3.70	0.0002
yearshift1	−0.06	0.17	−0.35	0.7254
yearshift2	0.08	0.12	0.74	0.4618
yearshift3	0.23	0.12	2.00	0.0453
yearshift4	0.03	0.12	0.26	0.7985
yearshift5	−0.08	0.14	−0.53	0.5974
cosi	0.44	0.14	3.13	0.0018
sinus	0.31	0.17	1.82	0.0689
sexfrac	−0.33	2.83	−0.12	0.9075
yearshift1:cosi	−0.05	0.16	−0.31	0.7544
yearshift2:cosi	−0.11	0.15	−0.75	0.4528
yearshift3:cosi	0.07	0.15	0.46	0.6468
yearshift4:cosi	−0.05	0.15	−0.34	0.7365
yearshift5:cosi	−0.13	0.15	−0.86	0.3886
yearshift1:sinus	−0.05	0.24	−0.20	0.8377
yearshift2:sinus	−0.06	0.17	−0.34	0.7334
yearshift3:sinus	−0.28	0.17	−1.59	0.1110
yearshift4:sinus	−0.23	0.18	−1.31	0.1908
yearshift5:sinus	−0.19	0.18	−1.06	0.2874

Table 2.5: Model 1 (2.16), Dispersion: 1.03

The resulting model for the incidence (Model 2) therefore becomes

$$\begin{aligned} \log(\mu_t) = {} & \log(n_t) + \alpha + a_i + \beta_1 \cdot \cos\left(\frac{t \cdot 2\pi}{365}\right) + \beta_2 \cdot \sin\left(\frac{t \cdot 2\pi}{365}\right) \\ & + \beta_{3i} \cdot \cos\left(\frac{t \cdot 2\pi}{365}\right) + \beta_{4i} \cdot \sin\left(\frac{t \cdot 2\pi}{365}\right) \\ & i \in \{1, 2, 3, 4, 5\} \end{aligned} \tag{2.17}$$

with the summary in Table 2.6 from which it is seen that year 0, June 1999-June 2000, has the highest relative baseline incidence. The periodic terms can be translated into a phase/delay and an amplitude as for the prevalence, which is shown in Table 2.8.

For the prevalence is was seen that the residuals were autocorrelated, which gives incorrect model inference. The residuals from the incidence model are uncorrelated as seen from Figure 2.12, which imply that the inference in this case is correct with respect to the correlation assumption. It was seen for the prevalence, that even if the long term correlation was removed, substantial short term correlation was still present in the residuals. This short term correlation may be a result of the carry over effect in the prevalence, children having an episode at day i is also likely to have one at day $i + 1$, whereas for the incidence a child having an incidence at day i the child is certain not to have a new incidence at day $i + 1$.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−5.2186	0.1039	−50.2429	<0.0001
yearshift1	−0.0724	0.1206	−0.6000	0.5485
yearshift2	0.0815	0.1115	0.7304	0.4653
yearshift3	0.2292	0.1107	2.0699	0.0386
yearshift4	0.0257	0.1127	0.2284	0.8194
yearshift5	−0.0852	0.1141	−0.7469	0.4552
cosi	0.4345	0.1374	3.1626	0.0016
sinus	0.2937	0.1361	2.1585	0.0310
yearshift1:cosi	−0.0478	0.1602	−0.2981	0.7656
yearshift2:cosi	−0.1108	0.1485	−0.7462	0.4556
yearshift3:cosi	0.0716	0.1474	0.4860	0.6270
yearshift4:cosi	−0.0488	0.1503	−0.3248	0.7454
yearshift5:cosi	−0.1309	0.1525	−0.8581	0.3909
yearshift1:sinus	−0.0282	0.1583	−0.1784	0.8585
yearshift2:sinus	−0.0481	0.1468	−0.3274	0.7434
yearshift3:sinus	−0.2655	0.1457	−1.8222	0.0686
yearshift4:sinus	−0.2193	0.1485	−1.4767	0.1399
yearshift5:sinus	−0.1827	0.1513	−1.2077	0.2273

Table 2.6: Model 2 as model 1 but without the fraction of boys., Dispersion: 1.03

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−5.1691	0.0892	−57.9252	<0.0001
yearshift1	−0.0937	0.1040	−0.9009	0.3678
yearshift2	0.0316	0.0967	0.3265	0.7440
yearshift3	0.2013	0.0960	2.0963	0.0362
yearshift4	−0.0290	0.0981	−0.2954	0.7677
yearshift5	−0.1492	0.1003	−1.4873	0.1371
cosi	0.3884	0.0270	14.3765	<0.0001
sinus	0.1383	0.0264	5.2376	<0.0001

Table 2.7: Model 3: Same seasonal part for all years, Dispersion: 1.03

Interpretation

The incidence is seen to vary during a year and has its maximum around the beginning of the calendar year. The baseline incidence (incidence when the periodic part is 0) is $e^{-5.22} = 0.01$. For each year a separate incidence rate is estimated by the model, the first year (year 0) is reference and has the incidence rate corresponding to the baseline. For year i ($i \neq 0$) the incidence rate can be calculated as the product of e^{a_0} and e^{a_i}, which gives 0.005, 0.006, 0.007, 0.006, 0.005 for year 1, 2, 3, 4 and 5 respectively.

The periodic part of the model shows that the amplitude varies between 0.32 and 0.52, which imply ratios between $e^{-0.32} = 0.72$ and $e^{0.32} = 1.38$ respectively $e^{-0.52} = 0.59$ and $e^{0.52} = 1.69$ for the two extremes. This imply that the relative incidence is 2-3 times higher in winter compared to summer as seen for the model for the prevalence.

Year	c_{year}	d_{year}	A_{year}	s_A	θ_{year}	delay (days)	$s_{\theta_{year}}$
0	0.43	0.29	0.52	0.11	-0.59	34.53	0.21
1	0.39	0.27	0.47	0.07	-0.60	34.95	0.14
2	0.32	0.25	0.41	0.04	-0.65	37.71	0.11
3	0.51	0.03	0.51	0.05	-0.06	3.24	0.10
4	0.39	0.07	0.39	0.06	-0.19	11.08	0.15
5	0.30	0.11	0.32	0.06	-0.35	20.37	0.19

Table 2.8: Interpretation for incidence model

2.3.3 Discussion

The analysis of the prevalence and the incidence shows that adequate parametric models can be found. For the prevalence model, the residuals are seen not to be white noise as required, since they are autocorrelated. Residuals from the incidence model shows that the residuals are uncorrelated, i.e. that both the long term and short term correlation is removed by the model. Both models show that the proportion of children having symptoms is highest in the winter and the rates are 2-3 times higher intensity in the winter compared to the summer. The models indicate that some of the symptoms are related to seasonal patterns, i.e. cold or flu. However, there may as well be parts of the variations caused be asthma.

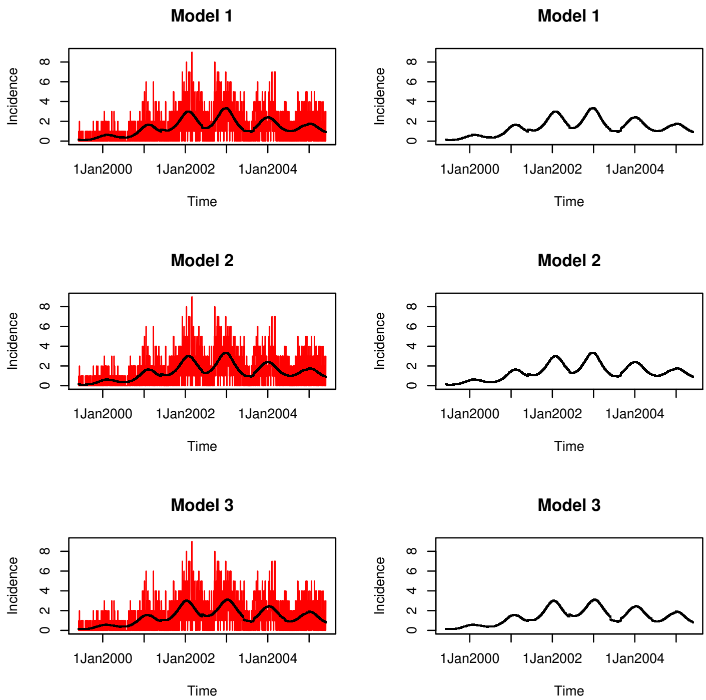

Figure 2.11: Predicted (black) and observed (red) incidence as function of time. Model 1: Season and fraction of boys, Model 2: Season, Model 3: Same as Model 2, but with the same phase-shift and amplitude for all years

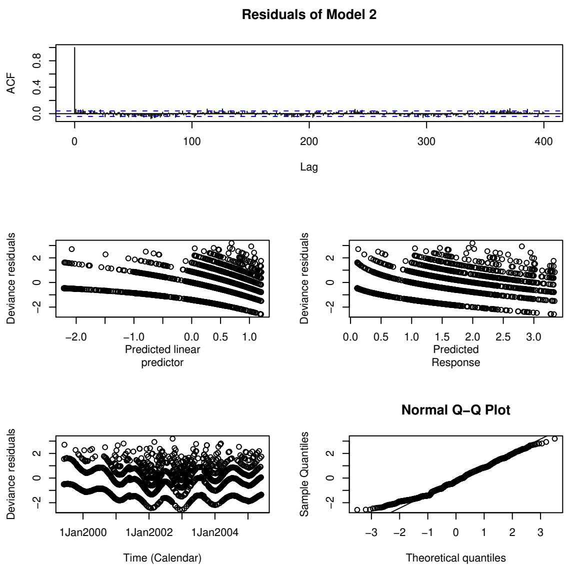

Figure 2.12: Top: Autocorrelation function for residuals from Model 2, middle left: Deviance residuals against predicted linear predictor, middle right: Deviance residuals against predicted response, bottom left: Deviance residuals against calendar time and bottom right: QQ-plot of deviance residuals

CHAPTER 3

Yearly aggregated symptoms

Contents

3.1 Introduction

This chapter contains an analysis of the temporal development of yearly wheezing symptoms. Based on the diaries on wheezing symptoms an yearly aggregation within each child is considered as the response to be modelled. This should hopefully average out the seasonal pattern seen in Chapter 2, since one season was seen to last one year.

The analysis is carried out by first analyzing the symptom-rates as being gaussian. This is done in order to simplify the initial analysis. The gaussian response can be analyzed by a generalized additive mixed effects model (GAMM) to analyze curvature for the age-related part of the model. Mixed effects models (MM) can be applied if the curvature can be parameterized adequate, which furthermore may give a model which is easier to interpret compared to the GAMM. MM may give clusters of identical children, however latent class regression (LCR) will also be applied to automate the clustering.

The gaussian approach may be adequate, however a more natural way to model the symptom-rates is to assume that the number of symptoms is poisson distributed. Again the GAMM can be used, but the parameterization must be done by a generalized linear mixed effects model (GLMM). LCR can be applied to cluster the children on the poisson scale in order to find possible subgroups of children. Both the gaussian and the poisson approach will be analyzed in the following chapter and finally compared in order to find possible similarities and differences.

3.2 Mixed effects model for wheezing intensity

The data considered consists of the number of wheezing episodes and days at risk per year for each child (the number of days the diary has been both kept and validated each year) as described in the population study in Chapter 2. The aggregation is done such that, an episode is recorded if the episode starts in the given interval, eg between the age of 1 and 2 years of life. In the following the relative count is modelled in order to analyze the longitudinal development in the wheezing symptoms. Mixed effects models will be used to model heterogeneity between children and to account for correlation between observations taken on the same child. Furthermore significant risk-factors for the congenital resistance and lung-function [13] are used as explanatory variables.

3.2.1 Yearly aggregated episodes

Since the number of days at risk varies from year to year, due to the fact that the children do not have complete diaries, the proportion of days with symptoms, i.e. the symptom rate, is considered as response. The observations are defined as $y_{ij} = n_{\text{episodes},ij}/n_{\text{days},ij}$, where $n_{\text{episodes},ij}$ is the number of episodes in year j for child i. The observations are bound to the interval $[0;1]$, which for a gaussian response naturally leads to a root-arcsine transformation of the proportions, see Olsson p. 92 [28]. The root-arcsine transformation transforms the observations to $\tilde{y} = \arcsin(\sqrt{y})$, which stretches the observations to $[0;\pi/2]$

The transformed relative count is shown in Figure 3.1, which shows that the variations in the longitudinal development from child to child are large. It is furthermore seen that most children have a transformed rate below 0.2 and that a small subset have much higher transformed rate between 0.3 and 0.4 in a short period of time often as their last observation. The number of episodes per year is summarized in Table 3.2, which shows that the majority of children has almost no symptoms.

	Quantiles					
Age	0 %	25 %	50 %	75 %	100 %	Mean (SD)
1 year	0.00	0.00	1.08	3.10	29.92	1.93 (2.56)
2 year	0.00	1.00	2.00	4.39	33.18	3.31 (4.20)
3 year	0.00	0.00	1.00	3.00	34.76	2.43 (3.73)
4 year	0.00	0.00	0.00	3.00	19.21	2.07 (3.38)
5 year	0.00	0.00	0.00	2.00	52.14	2.07 (4.92)

Table 3.1: Summary for the number of episodes per year, i.e. the corrected number: $n_{\text{episodes}}/n_{\text{days}} \cdot 365$ days/year

3.2.2 Missing data

The number of days at risk in each year should be 365 if all observations in a given year are present. This is however not the case in the ends of the considered age-range, i.e. the diaries do not start at birth and some of the children are dropouts, which influence the end. However, if the number of days in mean is sufficiently high this should not be a severe problem. The mean number of days per year is 335, 352, 344, 294 and 301 for first, second, third, fourth and fifth year of life, which shows that the

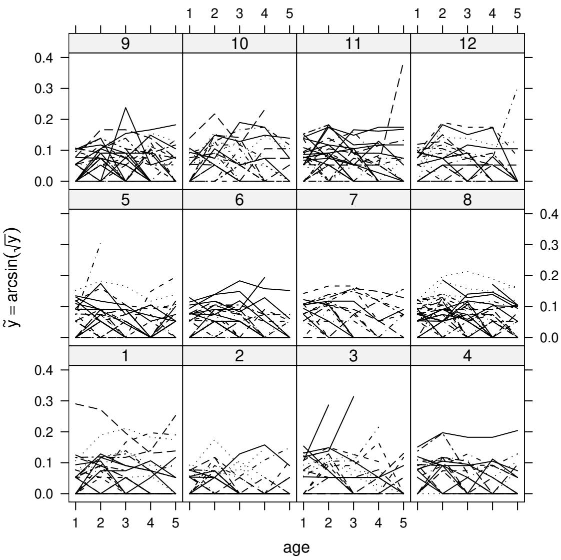

Figure 3.1: Longitudinal development of root-arcsine transformed relative count grouped by birth month

right tail is affected the most. However on average around 300 days at risk is seen for all years, which imply that the impact is not too big.

There are three types of participation status: Active (normal), resting and passive, where the difference between resting and passive is that the passive group actively has quit the study. In the upper part of Figure 3.2 histograms of the age at dropout for the resting and passive groups are shown. It is seen that most children drops out before the age of 3 years and that the highest number of children dropping out is seen in the first quarter of the fourth year of life. The latter may explain the lower number of mean days at risk in the fourth year of life, since these observation may lower the average considerable. The same pattern is seen for the active group for the age at the end of each diary-sequences ending before the age of 5 years. A large number of children have a break in the diary around the age of 3 years (last quarter of third year of life and first of the fourth year of life). This is caused by the fact that the study was initial only meant to include the first three years of life, which later in the study was changed to six years.

From the bottom right part of Figure 3.2 it is seen that diary-sequences in general end more frequently in March to August for the active children, whereas the non-active's sequences ends most frequent from March to September. Obviously the analysis of the yearly aggregated data will be biased if the children systematic are having less diary coverage in the some part of the years. The argument for using yearly aggregated

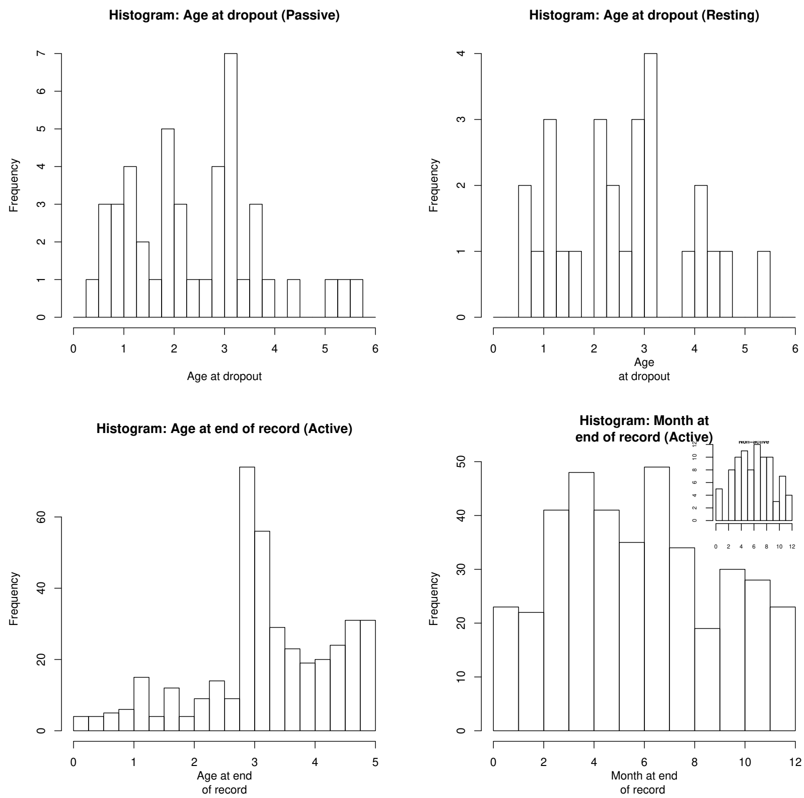

Figure 3.2: Upper left: Histogram of dropout age for passive group, upper right: Histogram of dropout age (last recorded observation) for resting, bottom left: histogram for age at end of record for the active group, i.e. the age at the last observation in each period where diary has been kept (more than 50 % of the individuals have more than 1 period with recorded diary), bottom right: The month at the last observation in each period for the active children (1 corresponds to January) with the corresponding histogram for non-active children in the upper right corner.

data is that it should cancel out the seasonal patterns, this will however not be the case if the missing data has a pattern. It is seen that the dropouts and end dates are sufficiently even distributed over the year.

The left part of Figure 3.3 shows that the monthly coverage is seen to be different over the year. The figure indicates how many times a given month is covered by a sequence, which should indicate if some months in general are seen less frequent than others. August is seen to be the month with fewest hits, which should not give too big problems, since most episodes are recorded in the winter, see the lower left corner of Figure 3.3. It is furthermore seen that the dropout rate and the month coverage is linked, such that the dropout-rate is highest in months, where the coverage is low. The right part of Figure 3.3 shows that the lengths of adjacent episodes in more than 70 % of the cases are longer than one year and that pauses of more than 4 years are seen. 221 children have more than one diary-sequence, whereas 169 children only have one.

Type	**Mean** (SD)	**Range**
Length of pause (all)	2.1 (1.2)	0 - 4.7
Length of pause (active)	2.1 (1.3)	0 - 4.7
Length of pause (non-active)	2.1 (1.1)	0.5 - 3.4
Number of sequences (all)	2 (1.2)	1 - 7
Number of sequences (active)	2 (1.2)	1 - 7
Number of sequences (non-active)	1 (0.6)	1 - 3
Age at last record (all)	5 (1.4)	0.3 - 7.9
Age at last record (active)	5 (1.1)	0.4 - 7.9
Age at last record (non-active)	3 (1.2)	0.3 - 5.7

Table 3.2: Mean, standard deviation (SD) and range for length of pause (years), number of sequences and age of last record (years). The latter corresponds to the dropout age for non-active children.

The analysis of the dropouts shows that children are more likely to dropout/end a sequence in the period from March to August. However, the analysis of the prevalence and the incidence, Chapter 2, showed that the prevalence and incidence was much higher in the winter compared to the summer. Even though Figure 3.3 shows that the number of episodes are lowest in the months with the lowest coverage by diary data, the analysis of the prevalence and the incidence shows that the relative number of episodes is significant higher in the winter compared to the summer. The analysis of the prevalence/incidence was based on more than 200 individuals in the majority of the considered time-interval, which should make the analysis sufficiently accurate.

One could analyze the missing data further, which however do not seem to be necessary in this case. One method for analyzing the missingness is to account for the additional randomness introduced by missing data as described by Borgan et al. [6]. This can be done by having a set of models: one for the outcome in question and one for the missingness, which gives the joint probability of the observed data and the missing data, see eg Albert [1]. The likelihood for the data and the missingness can then be maximized with respect to the model parameters by use of the EM-algorithm, see Dempster et al. [14]. However in the analysis of the symptoms, it is seen that the missingness mostly is relates to dropouts and longer pauses, which makes the missingness more static than dynamic. Dropouts and missing data will not be analyzed further in this thesis, which imply that imputing or modelling the missingness will be left for further studies.

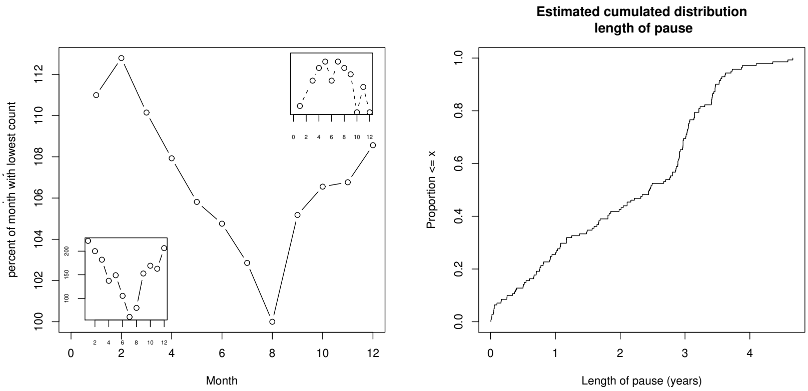

Figure 3.3: Left: The number of times a given months has been covered with a diary-sequence as percent of the lowest count , eg a sequence from March 3rd 2001 to April 1st 2003 imply that March is increased by 3 instances. In the upper right corner of the left plot the dropout frequency on months is shown and in the lower left corner the frequency of started episodes. Right: The histogram of the length of the pauses, i.e. the time between to adjacent sequences of diary for the same individual.

3.2.3 Generalized additive mixed effects model

The average longitudinal development can be examined with a generalized additive mixed effects model, which can be formulated as

$$\begin{aligned}\tilde{y}_{ij} = &\arcsin\left(\sqrt{n_{\text{cough}_{ij}}/n_{\text{days}_{ij}}}\right) = b_{0i} + s_1(\text{age}_{ij}) + s_2(\log_{10}(pd_i)) + \\ &s_3(\text{fev}_i) + s_4(\text{daycare}_i) + \varepsilon_{ij} \quad i = 1, \ldots, m \quad j = 1, \ldots, n_i \\ &b_{0i} \sim \mathcal{N}(0, \sigma_0^2) \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)\end{aligned} \tag{3.1}$$

where s_i is a smoothed function describing the curvature for $\tilde{y}$ against the i'th variable and b_{0i} is an individual intercept for each child uncorrelated with the ε_{ij}'s. b_{0i} represents the random component, i.e. the heterogeneity between the children. The GAMM is more elaborately described in Wood p. 316-318 [42] and is in this initial analysis only accounting for different baseline symptom-rates and Chapter 2.

pd corresponds to the congenital PD15 PtcO2 measurement, which is the interpolated dose to give a 15 % decrease in PtcO2 (partial pressure). The PD15 is interpolated from a base level and 5 medication levels, which in the preparatory thesis [13] was seen to be rather right skewed. A log-transformation of the variable is therefore applied to give a more adequate measure. PD15 PtcO2 is a measure of the congenital resistance, i.e. a high PD15 PtcO2 value implies that the child is very resistant to a medication provoking the airways.

Day-care start is the child's age in days at day care start and fev is the corrected

$FEV_{0.5}$ at the age of around 1 month. The correction of $FEV_{0.5}$ is done with respect to age of measurement and length at birth as described in the preparatory thesis [13], since the $FEV_{0.5}$ is highly correlated with both age and length at birth. The $FEV_{0.5}$ should have been measured at the age of 1 month but some children were measured later than that, which implied that these children had a higher $FEV_{0.5}$ than expected.

The smoothed functions from the GAMM, $s_1, \ldots, s_4$, are shown in Figure 3.4. The figure shows that the smoothed function for age probably can be parameterized with a second order polynomial, whereas the other three variables can be parameterized with linear functions. It is furthermore seen that the FEV probably has insignificant influence on the symptom-rate, since 0 is seen to be included in the confidence intervals at all measured values.

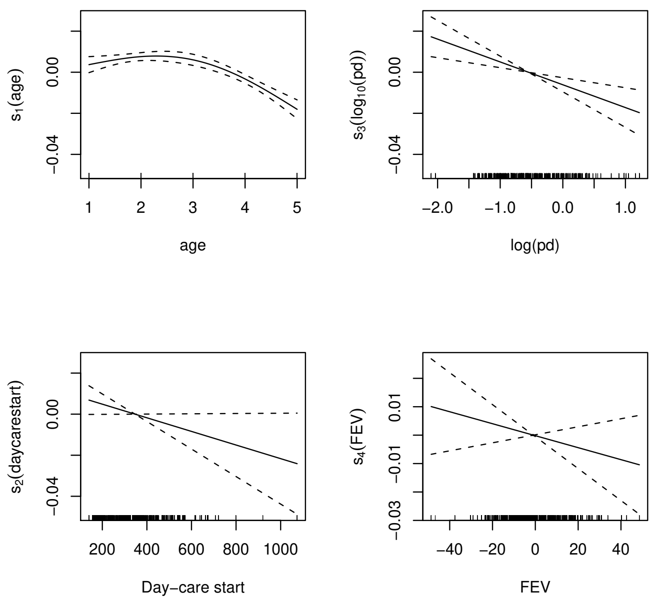

Figure 3.4: Smoothed function for age, log(PD PtcO$_2$), day-care start and FEV$_{0.5}$ vs. $\tilde{y}$. —: fitted line, - - -: confidence bands.

3.2.4 Mixed effects model

A parametric mixed effects model can be fitted to the data, which should give a model, which is easier to interpret compared to the GAMM. As seen from Figure 3.4 the parameterization of the smoothed for to age could be a second order polynomial, whereas the remaining three risk-factors are seen to be linear curves. The model with a random intercept, slope and curvature for age becomes

$$\begin{aligned}\tilde{y}_{ij} =&\beta_0 + b_{0i} + (\beta_1 + b_{1i}) \cdot \text{age}_{ij} + (\beta_2 + b_{2i}) \cdot \text{age}_{ij}^2 + \\ &\beta_3 \cdot \log_{10}(pd_i) + \beta_4 \cdot \text{daycare}_{\text{start},i} + \\ &\beta_5 \cdot \text{smoking.3rd}_i + \beta_6 \cdot \text{gender}_i + \beta_7 \cdot \text{fev}_i + \varepsilon_{ij}\end{aligned} \tag{3.2}$$
$$i = 1, \ldots, m \quad j = 1, \ldots, n_i \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}), \quad b_{0i} \sim \mathcal{N}(0, \sigma_0^2)$$
$$b_{1i} \sim \mathcal{N}(0, \sigma_1^2), \quad b_{2i} \sim \mathcal{N}(0, \sigma_2^2), \quad \mathbf{b_i} = \begin{bmatrix} b_{0i} & b_{1i} & b_{2i} \end{bmatrix}^T \sim \mathbf{MVN}(\mathbf{0}, \mathbf{G})$$

where $\mathbf{G}$ is a positive definite covariance matrix with no imposed structure a priori. This imply that the random components may be correlated, which is likely to be the case for a second order polynomial.

In order to analyze the arcsine-root transformed symptoms rates appropiate, the observations should be weighted with n_{days} [11]. This is necessary since the proportions are most precisely estimated when the number of days is large. Consider for instance a child having 2 episodes on 10 days compared to a child having 60 episodes during 300 days. This gives the same proportions, but the uncertainty is much higher for the proportion based on 10 days, since one episode more increase the proportion by 50 % for the proportion based on 10 days, whereas the proportion based on 300 days will increase by 1.67 %. Hence the variance is inverse proportional with the days at risk implying $V[e_{ij}] = \sigma^2 / n_{\text{days}}$

The observations within each individual are seen to be correlated due to the random components. This is seen by first separating the mean and covariance structure as (see Diggle et al. p. 83[15])

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e} \tag{3.3}$$

with

$$e_{ij} = \mathbf{d}_{ij}' \mathbf{B_i} + Z_{ij} \tag{3.4}$$

where Z_{ij} are independent identical distributed gaussian variables with zero mean and variance σ^2. B_i is the random component for individual i, the intercept, the slope and the intercept and has the covariance matrix G. $\mathbf{d}_i$ is the design matrix for the random component, which for a given j is $d_{ij} = \begin{bmatrix} 1 & t_j & t_j^2 \end{bmatrix}$.

The covariance for two observations taken on the same individual is then given by

$$\begin{aligned}\text{Cov}[e_j, e_k] = \text{Cov}[\mathbf{d}_{ij}' \mathbf{B_i} + Z_j, \mathbf{d}_{ik}' \mathbf{B_j} + Z_k] = \\ \text{Cov}[\mathbf{d}_{ij}' \mathbf{B_i}, \mathbf{d}_{ik}' \mathbf{B_k}] = \mathbf{d}_{ij}' \text{Cov}[\mathbf{B_i}, \mathbf{B_i}] \mathbf{d}_{ik} = \mathbf{d}_{ij}' \mathbf{G} \mathbf{d}_{ik}\end{aligned} \tag{3.5}$$

whereas the variance for a single observation is given as

$$\mathbf{d}_{ij}' \mathbf{G} \mathbf{d}_{ij} + \sigma^2 \tag{3.6}$$

In the simple case where $\mathbf{G}$ is a diagonal matrix, i.e. uncorrelated random components, the correlation between measurements taken on the same individual at time k and j

is

$$\rho = \frac{\sigma_0^2 + t_k t_j \sigma_1^2 + t_k^2 t_j^2 \sigma_2^2}{\sqrt{\left(\sigma_0^2 + t_k^2\sigma_1^2 + t_k^4\sigma_2^2 + \sigma^2\right)\left(\sigma_0^2 + t_j^2\sigma_1^2 + t_j^4\sigma_2^2 + \sigma^2\right)}} \tag{3.7}$$

which imply that the correlation is a function of (t_k, t_j). Furthermore the model is subject-specific rather than population averaged (see Zeger et al. [46]), which imply that the estimated fixed effects are effects for an individual and can not be interpreted as the effect for the population.

Diagnostics

In Figure 3.5 diagnostic plots are shown. The figure shows that the normality assumption for the residuals seems to be correct. Furthermore, it is seen that no pattern is present in the residuals plotted against each of the covariates, which shows that the model is adequate. The weighing with the number of days is important, which can be seen from the lower right plot with a QQ-plot for the residuals from an unweighted version of the model, where deviation for normality is seen in the two tails.

It is seen that the Pearson residuals, which in this case correspond to normalized residuals, since no within individual correlation is modelled (Pinheiro and Bates p. 239 [32]), is below 3.08 in absolute value. The Pearson residuals should be compared to a standard gaussian distribution, which gives $p = 0.001$ for a value of 3.08, which in a Bonferroni outlier test (p-value adjusted by the number of observations) gives $p_{\text{adj}} > 1$. Thus, there do not seem to be any outliers in the data, i.e. 3 is not an extreme observation in a large dataset.

Furthermore the random components should be examined to check the normality assumption. This is done with a QQ-plot for each of the 3 components BLUPs (best unbiased linear predictors) [37], which is seen in Figure 3.6. The QQ-plots show that the random components seems to coincide well with gaussianity, although some lower tail issues may be present for b_{0i}.

Estimation

The model is estimated by means of maximum likelihood (ML) for testing the fixed effects, since this allow likelihood ratio tests for the fixed effects. Using the restricted maximum likelihood (REML) technique gives more correct estimates of the random components, but mixed effects models with different fixed effect structure can not be compared with respect to their restricted likelihoods (Pinheiro and Bates p. 75-76 [32]). For the random components REML-estimates can be compared to ML-estimates to insure that the ML-estimates are not underestimated.

The summary for the fixed effects is given in Table 3.3, which shows that the age variables are highly significant, the congenital resistance, PD15 $PtcO_2$ is significant as well as the age of daycare start. Reducing the model by removing the factors

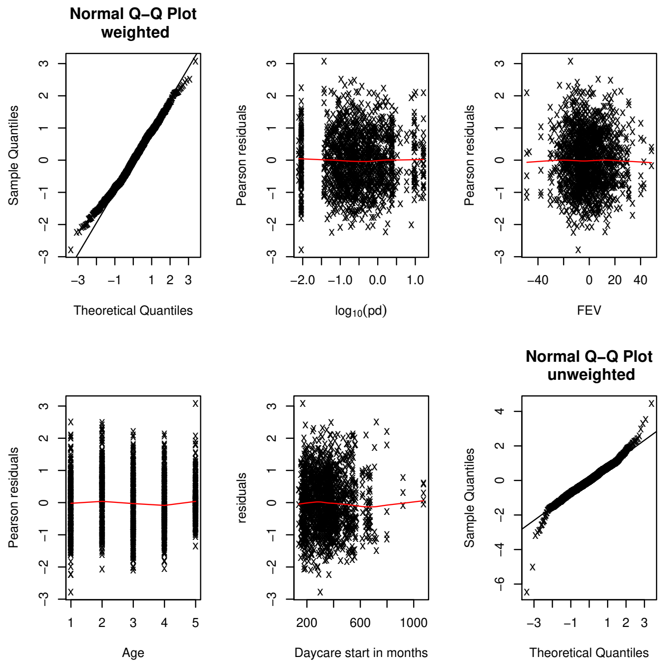

Figure 3.5: Residual diagnostics

gender, exposure to smoking in the third trimester and the congenital corrected FEV leads to an insignificant decrease in the likelihood (p=0.0865).

	Value	Std.Error	DF	t-value	p-value
(Intercept)	0.0528	0.0081	1097	6.5207	<0.0001
age	0.0172	0.0035	1097	4.9219	<0.0001
(age^2)	-0.0037	0.0006	1097	-6.2676	<0.0001
(log10(pd))	-0.0105	0.0029	308	-3.5638	0.0004
(daycare.start/30)	-0.0012	0.0005	308	-2.4281	0.0158
genderMale	0.0004	0.0044	308	0.0998	0.9206
(fev/100)	-0.0134	0.0160	308	-0.8343	0.4047
factor(smoking3rd)1	0.0062	0.0065	308	0.9674	0.3341

Table 3.3: Summary for fixed effects

The variance-covariance matrix for the random components for the reduced model is given in Table 3.4, which shows that the variance for the curvature is low. However the maximum likelihood estimate of σ_2 is $\hat{\sigma}_2 = 0.02 \cdot 10^{-3}$, which should be compared to the estimate of $\hat{\beta}_2 = -0.0038$. A test for the decrease in the likelihood gives $p = 0.02$ for a likelihood ratio of 7.6 on 3 degrees of freedom (the variance of the

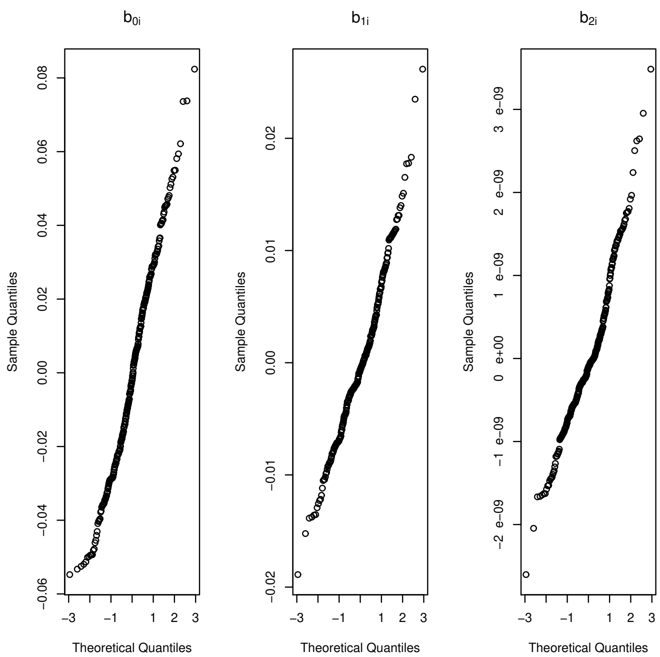

Figure 3.6: QQ-plots of the random components

random curvature and the covariance with the intercept and the slope).

	Intercept	age	(age^2)	Intercept	age	(age^2)
Intercept	1.4927	-0.1947	$-1.1494 \cdot 10^{-6}$	1.4000	-0.1807	$-0.0058 \cdot 10^{-6}$
age		0.1110	$-0.0647 \cdot 10^{-6}$		0.1119	$-0.0008 \cdot 10^{-6}$
(age^2)			$0.2284 \cdot 10^{-6}$			$0.0027 \cdot 10^{-6}$
Intercept	1.0000	-0.4784	-0.0020	1.0000	-0.4565	<0.0001
age		1.0000	-0.0004		1.0000	<0.0001
(age^2)			1.0000			1.0000

Table 3.4: Variance-covariance matrix for the random components multiplied by 10^3 above horisontal line and corresponding correlation matrix below the line (the first tre columns are for the ML-estimation and the last 3 for the REML estimation)

Applying the same test for REML estimations of the two model gives $p \approx 1$, which leads to the opposite conclusion. This is mainly due to the fact that the estimated variance in the REML estimation is $\hat{\sigma}_2 = 1.64 \cdot 10^{-6}$. The REML estimate of the random component is much smaller compared to the ML-estimate. Since the REML method is a better method to test the variance components the individual curvature is removed with basis in the REML-test (see Diggle et al. p. 69 [15] for a discussion of REML vs. ML).

The REML estimation of the random components shows that the random second order term is insignificant and furthermore uncorrelated with the intercept and the first order term. The correlation matrix for the fixed effects is shown in Table 3.5 (based on REML estimation), it shows that substantial negative correlations between the first and second order parameters for age are present.

From Figure 3.6 it is seen that the normality assumption for the random components seems to be fullfilled. The QQ-plots are seen to be linear with some small deviations, though not severe.

	(Intercept)	age	(age^2)	(log10(pd))	(daycare.start/30)
(Intercept)	1.00	−0.57	0.51	0.14	−0.72
age	−0.57	1.00	−0.96	0.01	−0.00
(age^2)	0.51	−0.96	1.00	−0.01	0.00
(log10(pd))	0.14	0.01	−0.01	1.00	0.09
(daycare.start/30)	−0.72	−0.00	0.00	0.09	1.00

Table 3.5: Correlation-matrix for fixed effects in model with random intercept, slope and curvature

An updated model without the random component corresponding to the second order parameter can be formulated as

$$\begin{gathered}
\tilde{y}_{ij} = \beta_0 + b_{0i} + (\beta_1 + b_{1i}) \cdot \text{age}_{ij} + \beta_2 \cdot \text{age}_{ij}^2 + \\
\beta_3 \cdot \log_{10}(pd_i) + \beta_4 \cdot \text{daycare}_{\text{start},i} + \varepsilon_{ij} \\
i = 1, \ldots, m \quad j = 1, \ldots, n_i, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{\Lambda}) \\
\boldsymbol{\Lambda} = \text{diag}(1/\text{ndays}_{i1}, 1/\text{ndays}_{i2}, \ldots, 1/\text{ndays}_{in_i}) \\
b_{0i} \sim \mathcal{N}(0, \sigma_0^2), \quad b_{1i} \sim \mathcal{N}(0, \sigma_1^2), \quad \mathbf{b_i} = \begin{bmatrix} b_{0i} & b_{1i} \end{bmatrix}^T \sim \mathbf{MVN}(\mathbf{0}, \mathbf{G})
\end{gathered} \tag{3.8}$$

A problem with the model, which may lead to incorrect analysis, is seen in the plot of the residuals vs. daycare start in Figure 3.5. The figure shows that a small number of children have skipped the nursery and are first starting in kinder-garden. This imply that these children start at the age of 3 compared to 75 % of the total cohort starting before the age of 1.1 and 90 % before 1.5 age of years.

Transformed daycare start

The covariate day care start is seen to be skewed and furthermore, there seems to be some pattern in the residuals for this covariate. In the following day-care start is replaced by $\log(\text{daycare}_{\text{start}})$ to reduce the skewness of day-care start. This leads to the model

$$\begin{aligned}\tilde{y}_{ij} = \beta_0 + b_{0i} + (\beta_1 + b_{1i}) \cdot \text{age}_{ij} + \beta_2 \cdot \text{age}_{ij}^2 + \\ \beta_3 \cdot \log_{10}(pd_i) + \beta_4 \cdot \log(\text{daycare}_{\text{start},i}) + \varepsilon_{ij}\end{aligned} \tag{3.9}$$

$$i = 1, \ldots, m \quad j = 1, \ldots, n_i, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{\Lambda})$$

$$\boldsymbol{\Lambda} = \text{diag}(1/\text{ndays}_{i1}, 1/\text{ndays}_{i2}, \ldots, 1/\text{ndays}_{in_i})$$

$$b_{0i} \sim \mathcal{N}(0, \sigma_0^2), \quad b_{1i} \sim \mathcal{N}(0, \sigma_1^2), \quad \mathbf{b_i} = \begin{bmatrix} b_{0i} & b_{1i} \end{bmatrix}^T \sim \mathbf{MVN}(\mathbf{0}, \mathbf{G})$$

Since the model in (3.8) has the same number of parameters as the model in (3.9), the models can be compared by the either Akaike's Information Criterion, AIC $= -2 \cdot \text{log-likelihood} + 2 \cdot n_{par}$, (see Michael Crawley p. 208 [10]) or the Bayesian Information Criterion, BIC $= -2 \cdot \text{log-likelihood} + \log(n_{obs}) \cdot n_{par}$.

Throughout the thesis, BIC will be used since AIC gives too many parameter, Ripley p. 34-35 [36] or Hurvich and Tsai's discussion of model selection [20]. In this particular example it really does not matter, since the number of both the observations and parameters are the same in the two models. This correspond to comparing the likelihood and choosing the model having the highest.

Model (3.8) has a BIC of -4583.71 compared to model (3.9)'s -4583.95, which shows that the updated model is marginally better in terms of minimizing the BIC, i.e. maximizing the likelihood.

The residuals from the updated model are shown in Figure 3.7, which shows that the improvement from the model with day-care start as a linear effect is very small. The p-value for the estimate of the transformed variable is now 0.0045 compared to 0.0053 for the model with the untransformed variable. From a parsimonic point of view the model without the log-transformed day care start is preferred and hence is therefore the model preferred from this point on.

3.2.5 Interpretation

The fixed effects for the model with the untransformed day care start as in (3.8) are shown in Table 3.6. The table shows that the remaining fixed effects are highly significant and further reduction leads to significant loss of information.

	Value	Std.Error	DF	t-value	p-value
(Intercept)	0.0555	0.0073	1097	7.5897	<0.0001
age	0.0172	0.0035	1097	4.9683	<0.0001
(age^2)	-0.0038	0.0006	1097	-6.5325	<0.0001
(log10(pd))	-0.0108	0.0029	311	-3.7233	0.0002
(daycare.start/30)	-0.0013	0.0005	311	-2.8103	0.0053

Table 3.6: Summary for fixed effects for mixed effects model with linear relation between the symptom rate and the day-care start.

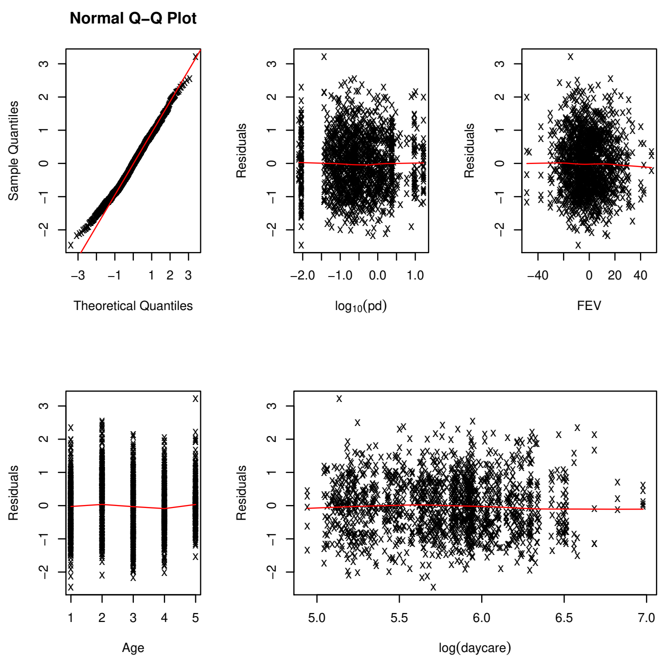

Figure 3.7: Residual diagnostics for re-fitted model (3.8)

Based on the estimated model the effect of the risk-factors PD $PtcO_2$ and age at day care start can be investigated. Since the estimation is based on the root-arcsine scale the effect of PD $PtcO_2$ will depend on the age and childs age at day care start and the effect of day care start will depend on the age and the PD $PtcO_2$. The effects are seen not to be additive nor multiplicative on the original scale. The effect of PD $PtcO_2$ for different ages and different times of day care start can be investigated by considering the fraction

$$1 - \frac{\hat{y}(\text{PD PtcO}_2 = 10 \cdot x)}{\hat{y}(\text{PD PtcO}_2 = x)} \tag{3.10}$$

for a given age and age of day care start. The expression correspond to the relative reduction for an increase by a factor 10 in PD15 PTcO2. This gives a rather complex expression, which for a child with average intercept and slope ($b_0 = b_1 = 0$) can be expressed as

$$1 - \frac{\hat{y}(\text{PD PtcO}_2 = 10 \cdot x)}{\hat{y}(\text{PD PtcO}_2 = x)} = 1 - \frac{\sin^2(c + \beta_3 \cdot (1 + \log_{10}(x)))}{\sin^2(c + \beta_3 \cdot \log_{10}(x))} \tag{3.11}$$
$$c = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{age}^2 + \beta_4 \cdot \text{median}(\text{daycare}_{\text{start}})$$

The effect of an individual increasing its PD15 level by a factor 10 is a function of both age, day care start and the reference PD15 level. A contour plot for the effect is shown in Figure 3.8 for day care start kept at the median value (11.3 months), which shows that the reductions are largest at the age of 5 years with reductions between 37 and 60 %. Obviously the interpretation is a little awkward, since the PD15 PtcO2 can not be increased for an individual, it however shows in which direction a population effect will be, see Zeger et al. [46].

It is furthermore seen in the upper part of the PD15 $PtcO_2$'s, this correspond to comparing PD15 $PtcO_2 = 10$ with PD15 $PtcO_2 = 1$ at the age of 5 years that the reductions are above 50 %. The high reduction are caused by comparing children being very resistant at birth (97 % quantile) with children not being so resistant (77 % quantile). It is seen that an increase in PD15 PtcO2 decreases the symptom-rate.

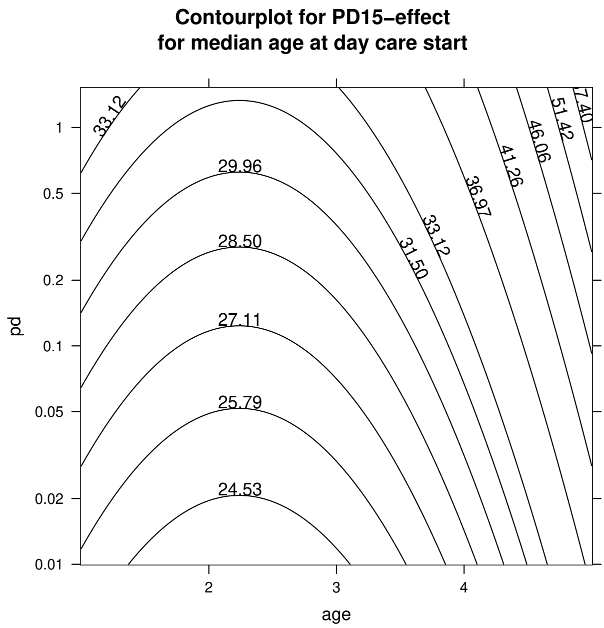

Figure 3.8: Contour plot of (3.11), each contour gives the reduction in percent of $\hat{y}(\text{pd}, \text{age})$ when pd = PD15 $PtcO_2$ is increased by a factor 10 (note the contours are not equally spaced)

Instead of considering the contour plot in Figure 3.8 the estimated reductions as function of the denominator PD15 $PtcO_2$ for each year of life is shown in the upper part of Figure 3.9. It is seen that the effect of an increase in PD15 $PtcO_2$ is highest for the age of 5.

For day care start a similar analysis can be done by keeping the PD15 $PtcO_2$-value

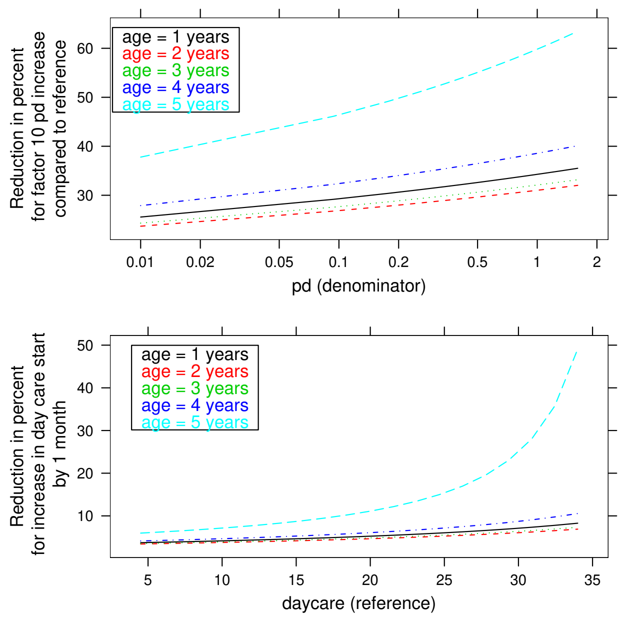

Figure 3.9: Top: Reduction for increasing PD15 by a factor 10, bottom: reduction for increasing daycare start by 1 month for each of the five considered years of life

fixed at the median value (0.26). A plot of

$$1 - \frac{\hat{y}(\text{daycare}_{\text{start}} = 1 + x)}{\hat{y}(\text{daycare}_{\text{start}} = x)} = 1 - \frac{\sin^2(d + \beta_4 \cdot (\text{daycare}_{\text{start}} + 1))}{\sin^2(d + \beta_4 \cdot \text{daycare}_{\text{start}})} \qquad (3.12)$$
$$d = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{age}^2 + \beta_3 \cdot \text{median}(\log_{10}(x))$$

is shown in the lower part of Figure 3.9. The figure shows that the age of 5 years is different compared to the other years of life, since it gives much higher reductions for a late day care start compared to the four other years of life. This which could be caused by the relative few observations in this part of the day-care range (see Table 3.7). Age 1 to 4 give reductions between 5 and 10 % for starting in day care start one month later. In general the reductions are seen to be highest at the age 5 years, i.e. the symptoms-rates at the age of 5 years are more sensitive to the PD15 PtcO2 measurements and the age of day-care starts compared to the first years. This imply that the children in this model have increased benefit as they grow older.

Quantiles in day-care start	0.00	0.05	0.25	0.50	0.75	0.95	1.00
Number of children	5	6	8	11	14	21	36

Table 3.7: Quantile for day-care start in months at the age of 5 years

Age at maximum symptom-rate	Freq	Maximum symptom-rate	Freq
1	18	1	176
2	103	2	85
3	131	3	32
4	50	4	8
5	10	5	8
6	2	6	4

Table 3.8: Distribution of the children's estimated age at maximum and fitted max proportion

3.2.6 Predictions

Based on the model found in (3.8) predictions of the longitudinal development can be estimated. To be able to compare children across different levels of PD15 and different ages at day care start, the predictions are based on all children having the median-value of PD15 and the median age at day care start. Furthermore, the proportion/rate is considered, which corresponds to the number of episodes per day.

In Figure 3.10 the predictions of the longitudinal development of the symptom-rate are shown, in which the children are grouped by the estimated age of their maximum symptoms-rates, age_{max}. This divides the children into a group with decreasing symptom-rate, a group with an initial increase until the age of 2 years and then a decline, a group topping between the age of 2 and 3 years, one group tops between 3 and 4, one between 4 and 5, and finally a group which has its maximum after the age of 5 years. The last group is seen to have an increasing symptom-rate in the considered age interval.

The plot for the 75 % quantile of PD15 PtcO2 and day-care start (Figure 3.10 right part) gives essentially the same results, the curves are shifted downwards due to the negative correlation between PD15 PtcO2 and the symptoms rate and between day-care start and the symptom rate. It is seen from both types of predictions that there seems to be a large group of children with more or less no symptoms. These children are seen to be distributed over the intervals $(1, 2]$ and $(2, 3]$ in Figure 3.10.

The distribution of the age at maximum is given in Table 3.8, which shows that the majority of the children have their maximum before the age of 3 years. The interpretation of a maximum before the age of 1 years is that these children keep getting better, i.e. having a lower symptom-rate. The opposite effect is seen for children with a maximum after the age of 5 years, which imply that the children keep having more symptoms. It is furthermore seen that the majority of the children have

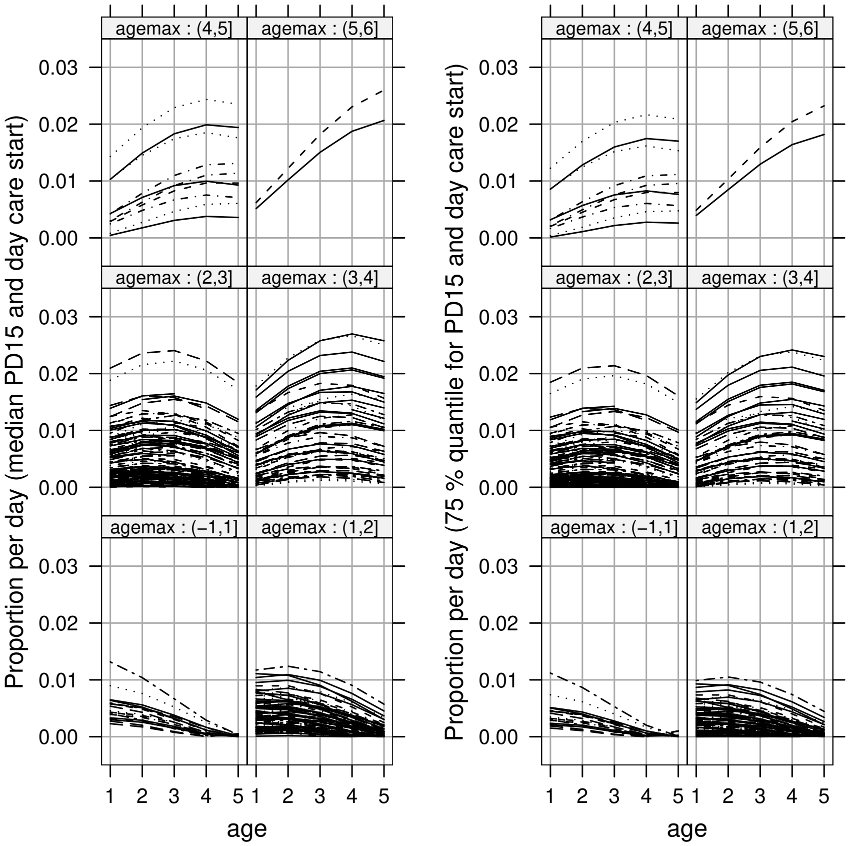

Figure 3.10: Left: predicted proportions for median PD15 and median age at day care start, right predicted proportions for 75 % quantile for PD15 and day care start. Both plots are grouped by the age of the estimated maximum proportion of days with episodes

close to no symptoms.

Unscaled predictions

In the upper part of Figure 3.11 the predicted proportions for all children with their respective values of PD15 PtcO2 and age of day-care start are shown. This gives a little more spread in some of the groups compared to the predictions based on quantiles of PD15 and day care start. It is noted that a group of children with a low maximum proportion is seen in both group $(1, 2]$ and $(2, 3]$.

There seems to be indications of a clustering of the children: one group has a low

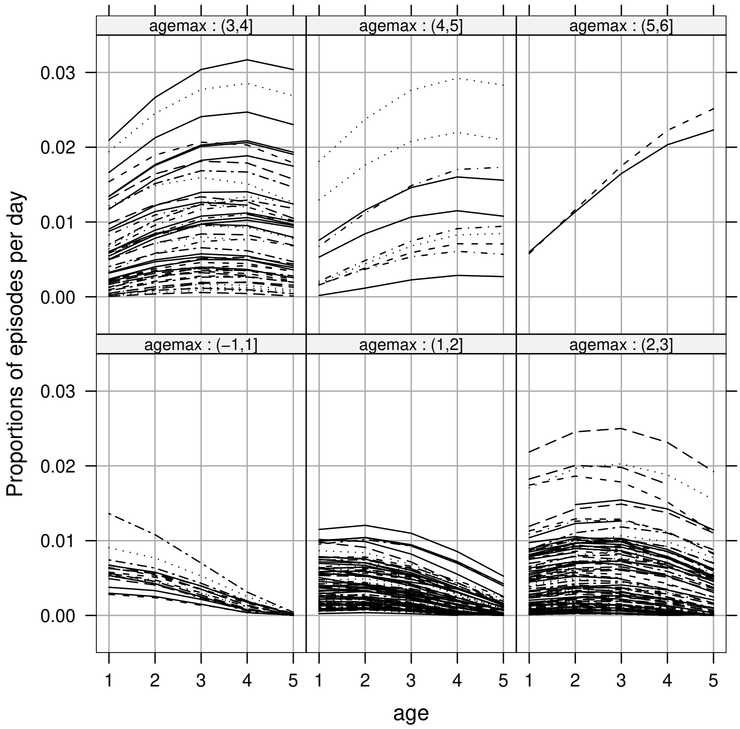

Figure 3.11: Top: Predicted proportions for observed values of PD15 and age at day care start grouped by the estimated age of their maximum symptom-rate.

maximum but can have their maximum at different ages, one starts at a medium level, but keeps getting fewer symptoms, one group starts at a medium level but does not have any improvement in the symptom-rate and finally one group starts at a medium or high level and keep getting more symptoms.

3.3 Latent class models for gaussian response

In the following section Latent Class Regression (LCR) is considered, which is a model-class using the data to find clusters of observations or individuals. LCR may therefore give the possibility to identify groups of children having the same type of temporal development of symptom-rates. The model class is more complex than a standard general linear model, but may be easier to interpret compared to a mixed effects models in terms grouping the children. The analysis of the mixed effects model indicated that there may be sub groups of children, it was however not clear how the children should be grouped.

3.3.1 Existing LCR litterature

The Latent Class Regression has been studied in many research fields. Erichsen et al. [16] and Poulsen et al. [33] used LCR and random coefficients on principal components analysis in a sensory analysis to identify latent segments within a given population and to decribe characteristics of these segments.

Garrett et al. [17] discuss latent class analysis in an epidemiologic study in order to identify similar individuals with respect to a questionnaire on mental health. The prevalence for each of the found groups and questions were estimated, which showed that the population could be divided into three groups, normals, a group with moderate depression and a group with severe depression.

The studies illustrate the main purpose of using latent class regression, namely that some underlying grouping of the population is present, but the grouping is hidden or latent. Consider a study of the muscle growth-curves for children at the age of 13-18 years exposed to different types of treatment. Furthermore, assume that one subgroup is girls and the other is boys, but the gender for some reason is unknown. The muscle growth-curves will probably differ from one group to another, which may be picked up by the Latent Class Regression.

A short simulation case is analyzed in the following to illustrate the model-class, the theory behind the estimation will be presented in section 3.3.2. Assume a phenomena measured in the variable y is measured from a population with a underlying grouping, which is unknown. In this simulation study two groups are assumed and the variable x is a covariate (see upper left part of Figure 3.12), which is correlated with y through the following expression

$$y = \begin{cases} 0.6 \cdot x - 0.1 \cdot x^2 + e_1 & \text{group 1} \\ 0.2 \cdot x + e_2 & \text{group 2} \end{cases} \tag{3.13}$$

where $V[e_1] = 4 \cdot V[e_2] = 1$. The upper right part of Figure 3.12 shows that the optimal number of clusters is 2. The predicted clusters agree perfectly with the grouping (known since it is a simulation) and the estimated parameters are shown in Table 3.3.1. It is seen that the agreement between estimated and true parameters is good. The standard errors for the parameters are not shown, but the intercepts are both insignificant as well as the curvature in the second cluster, as required from the model formulation.

	Group1	Group 2
(Intercept)	−0.19	0.22
x	0.76	0.14
(x^2)	−0.12	0.00
σ	0.94	0.53

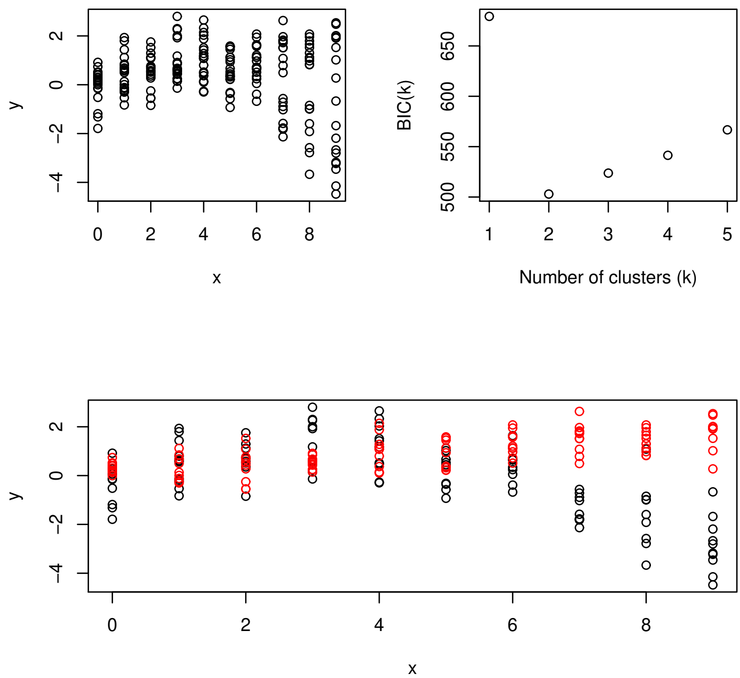

Figure 3.12: Upper left: x vs. y without grouping, upper right: BIC (Bayesian Information Criteria) vs. the number of clusters and bottom: x vs. y with observations colored according to their predicted cluster

3.3.2 Model specification

The basis for the estimation of a mixture model in the following section is the model structure found in section 3.2. The model had the following structure for the fixed effects

$$\begin{aligned}\tilde{y}_{ij} =& \beta_0 + \beta_1 \cdot \mathrm{age}_{ij} + \beta_2 \cdot \mathrm{age}_{ij}^2 + \beta_3 \cdot \log_{10}(\mathrm{pd}_i) \\ &+ \beta_4 \cdot \mathrm{daycare}_{\mathrm{start},i} + \varepsilon_{ij} \\ & i = 1, \ldots, m \quad j = 1, \ldots, n_i \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{\Lambda}) \\ & \mathbf{\Lambda} = \mathrm{diag}(1/\mathrm{ndays}_{i1}, 1/\mathrm{ndays}_{i2}, \ldots, 1/\mathrm{ndays}_{in_i})\end{aligned} \tag{3.14}$$

Model formulation for latent class regression

Based on the $\sqrt{n_{\mathrm{days}_{ij}}}$ weighted observations, $\tilde{Y}$ and design matrix, X, a LCR can be formulated. The background for the model formulation of the latent class regression is found in Leisch's vignette for the R-package *FlexMix* [23] and the following model formulation is with some small deviations entirely based on that. Letting $\tilde{y}_{ij}$ be observation j for individual i and x_{ij} be the row-vector from the design matrix corresponding to observation $\tilde{y}_{ij}$. The conditional density for $\tilde{y}_{ij}$ is

$$h(\tilde{y}_{ij}|x_{ij}, \psi) = \sum_{k=1}^{K} \pi_k f(\tilde{y}_{ij}|x_{ij}, \theta_k) \tag{3.15}$$

where π_k is the prior probability of cluster k, θ_k is the parameters for cluster k, $\theta_k = (\beta_k, \sigma_k^2)$, f the density function for the k'th cluster and $\psi = (\pi_1, \ldots, \pi_K, \theta_1, \ldots, \theta_K)$ the whole set of parameters. The prior probabilities are the probabilities for any observation to belong to one of the clusters, i.e. before using the information in that particularly observation. The prior probabilities are non-negative and sum to 1, i.e. the k clusters are the entire event space.

$$\pi_k \geq 0 \quad \wedge \quad \sum_{k=1}^{K} \pi_k = 1 \tag{3.16}$$

Parameter estimation can be done with the E-M algorithm, see Dempster et al. [14], which alternates iteratively between an Estimation step and a Maximization step. The E step is to estimate the posterior probabilities for individual i in cluster k with n_i observations

$$\hat{p}_{ik} = \frac{\pi_k \prod_{j=1}^{n_i} f(\tilde{y}_{ij}|x_{ij}, \theta_k)}{\sum_{k'=1}^{K} \pi_{k'} \prod_{j=1}^{n_i} f(\tilde{y}_{ij}|x_{ij}, \theta_{k'})} \tag{3.17}$$

and then to estimate the prior probabilities, the overall probability of cluster k given the posteriors. The prior probabilities are estimated by

$$\hat{\pi}_k = \frac{1}{m} \sum_{i=1}^{m} \hat{p}_{ik} \tag{3.18}$$

The maximization step is to maximize the log-likelihood with respect to the parameters for each cluster. In the maximization the posterior probabilities are used as weights, which leads to the M-step

$$\hat{\theta}_k = \arg\max_{\theta_k} \sum_{i=1}^{m} \sum_{j=1}^{m_i} \hat{p}_{ik} \cdot \log f(\tilde{y}_{ij}|x_{ij}, \theta_k) \tag{3.19}$$

This imply that observation i has weight $\hat{p}_{ik}$ in the estimation of cluster k. The EM-algorithm runs iteratively between the two steps until the likelihood stops improving more than a specified threshold. On top of the iterative EM-algorithm, the LCR is started at different initial parameter values to insure convergence to the maximum likelihood solution.

3.3.3 LCR applied to root-arcsine symptomrate

In the following a LCR will be estimated based on the linear model described in (3.14). To insure that the number of clusters is chosen in an optimal way, the Bayesian Information Criterion is calculated for models with $k \in 1, 2, \ldots, 5$ clusters, from which the optimal number of clusters can be found by maximizing the BIC. Furthermore, to obtain information on the uncertainty of the BIC, the BIC is determined in m estimations for each cluster size, where the i'th estimation is based on the dataset not containing individual i, i.e. a jackknife strategy.

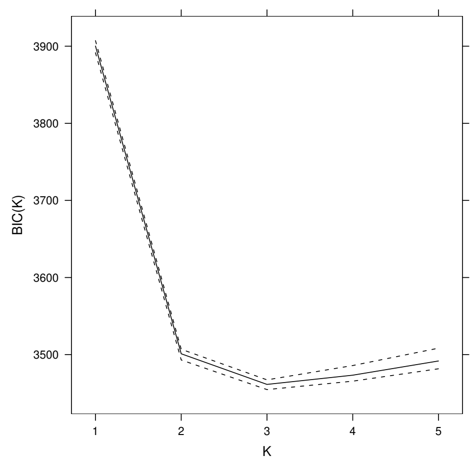

Figure 3.13: BIC as function of the number of clusters, K. With 90 % confidence bands based on jackknife strategy

It is seen from Figure 3.13 that the optimal number of clusters is $K^\star = 3$. The corresponding estimates are given in Table 3.9, which could be compared to the fixed

Parameter	Cluster 1	Cluster 2	Cluster 3	Mixed effects
Number of individuals	171	86	59	316
Number of observations	753	392	278	1423
Intercept	0.0768	0.0358	0.0489	0.0553
age	0.0122	-0.0043	0.0570	0.0173
age^2	-0.0036	0.0000	-0.0087	-0.0038
log10(pd)	-0.0046	-0.0013	-0.0135	-0.0108
daycare.start/30	-0.0020	-0.0006	-0.0012	-0.0013
σ	0.7521	0.4915	0.7182	0.6292

Table 3.9: Parameters for the clusters in the optimal mixture model with the estiamtes from the mixed effects model as reference

effects from the mixed effects model considered in last column in the table. It is seen that the intercepts in the three clusters are located around the estimated average intercept in the mixed effects model. Although numerical similarities, the mixed effects parameters are subject specific and the cluster estimates are cluster averages, which imply that the interpretation is not quite the same. However, it is seen that congenital resistance and age at day-care start is pointing in the same direction as the mixed effects model. The slopes are seen to be both positive (two clusters) and negative for one cluster compared to the positive slope for an average individual in the mixed effects model. It is furthermore seen that the residual variance is compared to the mixed effects estimate in one cluster and higher in the two others.

The three clusters are characterized by

- Cluster 1
  - highest intercept, a postive median slope and curvature
  - estimate for PD15 PtcO2 lies between the two other clusters
  - day care start effect parameter lowest of the 3 groups
  - estimated maximum at $\text{age}_1^\star = 1.7$ years
- Cluster 2
  - lowest intercept, slope (different sign compared to the other clusters) and absolute curvature
  - highest estimate for PD15 and day care start
  - estimated standard deviation is 60 % of the standard deviation in the other two groups
  - estimated maximum at $\text{age}_2^\star = -68.2$ years
- Cluster 3
  - median intercept, highest slope and absolute curvature

– lowest estimate of PD15 and median estimate for day care start

– estimated maximum at $\text{age}_3^\star = 3.3$ years

Predictions for median values of PD15 PtcO2 and day-care start are shown in Figure 3.14. The figure shows that three distinct groups are found, corresponding to the groups described above. The grouping yields one group with many symptoms and two groups with a decreasing level of symptoms of which one of them has essentially no symptoms.

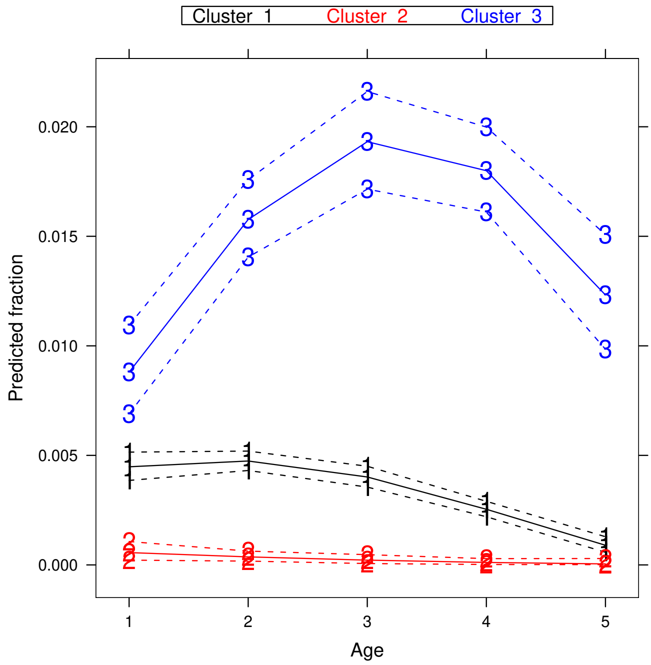

Figure 3.14: Predictions for (3.24) for each of the three clusters with confidence bands. PD15 and day care start are kept fixed at median level in all clusters

From the jackknife estimation it is possible to estimate the optimal number of clusters in each iteration (when omitting individual i from the estimation), this gives $K^\star = 3$, 315 times and $K^\star = 4$, 1 time. It is seen that the optimal number of clusters is not affected of the individual observations and only in the direction of expanding with an additional cluster in one case.

The individuals with divergent assignment from the complete analysis compared to the jack-knife analysis have posterior probabilities given in Table 3.10 for the three clusters considered, which shows that the 2 observations with the lowest maximum posterior have posteriors close to 0.5. These two individuals are seen to be borderline cases, which obviously make them sensitive to even small changes in the parameters in the model.

	Post 1	Post 2	Post 3
1	0.48	0.52	0.00
2	0.51	0.49	0.00
3	0.00	0.19	0.81

Table 3.10: Posteriors for individuals with different cluster assignment in overall estimation and jack-knife assignment

id	pd	daycare.start	age	ncough	ndays	cluster	jack.knife
171	11.15	250	1	0	353	1	2
171	11.15	250	2	2	365	1	2
171	11.15	250	3	1	365	1	2
171	11.15	250	4	0	365	1	2
171	11.15	250	5	0	365	1	2
352	0.20	428	1	1	333	2	1
352	0.20	428	2	2	365	2	1
352	0.20	428	3	0	365	2	1
352	0.20	428	4	0	365	2	1
352	0.20	428	5	0	365	2	1
376	0.28	158	1	1	347	3	2
376	0.28	158	2	1	365	3	2
376	0.28	158	3	7	359	3	2
376	0.28	158	4	7	328	3	2
376	0.28	158	5	3	282	3	2

Table 3.11: Summary for observations with divergent classifications from jack-knife compared to full estimation

The three individuals observations are tabulated in Table 3.11, which shows that the mis-classifications correspond to a child from the high group being misclassified as a child in the low group, a child in the low group being classified as a medium and a child from the medium classified as low. Mainly the first child is problematic, since the penalty from going from the high to the low should be much higher compared to shifting between the low and the medium, since they are much more similar. The child being classified as high and low for the full estimation and the jack-knife, respectively, is characterized by having a rather low day-care start age (1 % quantile) and an average PD15 PtcO2 measurement, see Table 3.11. The child is furthermore seen to start at a low level, since the first two years have only 1 episode each. The high group is characterized by starting at a high level, which imply that leaving an observation in the lower part of group will tend to give an even higher estimated proportion and thereby making it more unlikely to belong to the high cluster for children in the low end of the group.

Normality assumption

Before considering the model further the residuals are examined for normality. The QQ-plot for the residuals from each of the three clusters (soft assignment, all observations included with weight p_{ik}) and for the combined residuals (hard assignment, only those observations in cluster k, which are assigned to cluster k are considered).

It is seen that the residuals seem to be gaussian, whereas the individual plots have some departures from normality for the lower values (in particularly for the cluster with the lowest intercept). As a whole the residuals do not seem too bad, i.e. from the upper left part of the QQ-plot.

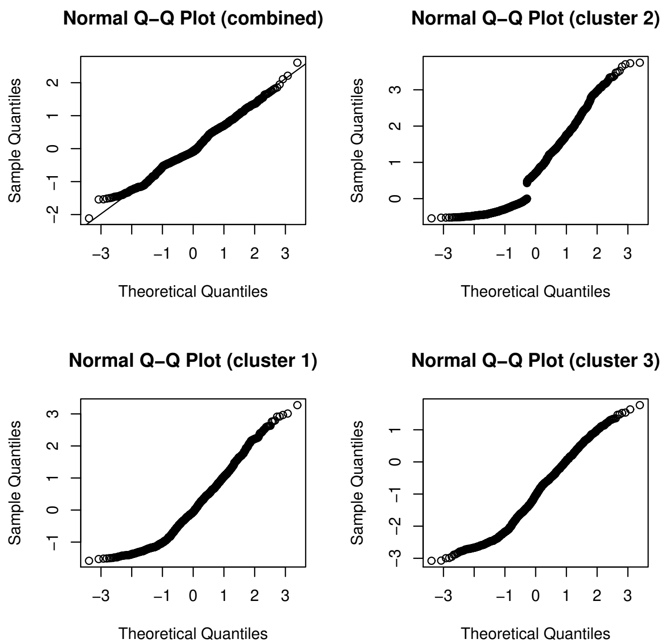

Figure 3.15: Residual examination for the three clusters in the optimal Latent Class Regression

Cluster characteristics

Based on the final posterior probabilities from the EM-algorithm the children can be divided into three groups. Each child is assigned to group k if $\hat{p}_{ik} \geq \hat{p}_{ik'}$ $\forall k'$, hence the children are assigned to the most likely group. The sizes of the groups can be seen from Table 3.9, which shows that group 1 has more than twice as many individuals as the other two groups.

In Figure 3.16 box plots of log(PD15 $PtcO_2$) and day care start are shown for each cluster. It is seen that the three clusters seem to have more or less the same distribution of the two variables. Bartlett's test for variance homogeneity in the three groups gives $p = 0.08$ and $p = 0.01$ for log(PD15 $PtcO_2$) and day care start, respectively. The latter test is mainly due to the skewed distributions of day care start as noted in section 3.2.4 p. 37, where especially cluster 3 has some old starters, testing the corresponding log-transformed variable gives $p = 0.66$.

To check if a cluster effect is present for log(PD15 $PtcO_2$) and log(day care start) an one-way ANOVA can be applied. The analysis can be carried out by the model (Petrucelli et al. p. 542-545 [30]) for PD15 PtcO2

$$\log(\text{PD15 PtcO}_{2,ij}) = \mu + \alpha_j + e_{ij} \tag{3.20}$$

with the hypotheses

$$\begin{aligned} H_0 &: \alpha_1 = \alpha_2 = \alpha_3 \\ H_1 &: \exists(i,j) \mid \alpha_i \neq \alpha_j \end{aligned} \tag{3.21}$$

Testing the cluster effect for the variables gives $p = 0.05$ and $p = 0.20$ for PD15 PtcO2 and log day-care start. This shows that the mean levels are not significantly different for the three clusters, hence the differences in the corresponding parameters are related to effects rather than differences in the group means.

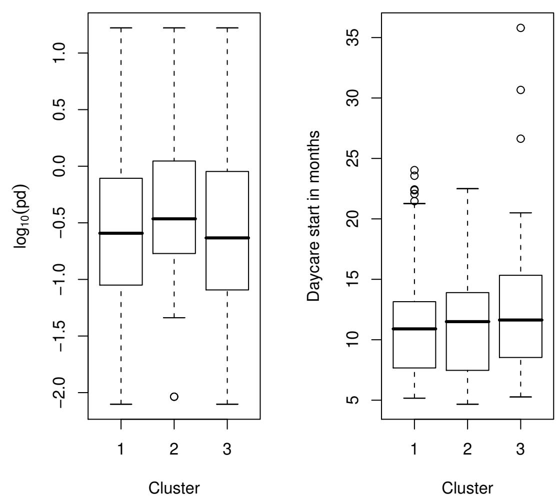

Figure 3.16: Box plots of PD15 and day care starting age for the three clusters

One can furthermore consider the mixed effects model from section 3.2 equation (3.8), where two random components were included. The BLUP (best unbiased linear prediction) estimates of the random components [37] can be compared to the clusters, which gives the possibility to see if a connection between the two models is present.

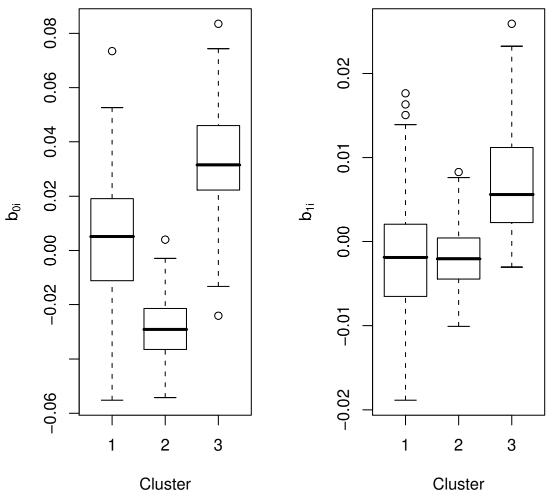

Figure 3.17: Box-plots for b_{0i} and b_{1i} obtained from mixed effects model grouped by clusters obtained from Latent Class Regression

From Figure 3.17 it is clear that the b_{0i}'s differ from cluster to cluster. Testing as in (3.21) but with log(PD15 $PtcO_2$) replaced by the BLUP-estimates gives $p<0.0001$ and $p<0.0001$ for b_{0i} and b_{1i}, respectively. It is seen that for b_{0i} all clusters are different, whereas for b_{1i} cluster 3 differ from the other two and that cluster 1 and 2 seem to have the same mean slope ($p = 0.9468$).

It follows that there is a connection between the clusters and the predictions of the random parameters: Cluster 3 has the highest b_{0i}'s and b_{1i}'s, whereas the difference between the two other clusters is that cluster 2 has the lowest b_{0i}'s. For the LCR an extra feature is that the estimates for day care start and PD15 PtcO2 differ from cluster to cluster compared to the fixed estimates in the mixed effects model.

Parameter comparison

The posterior probabilities can be used as weights in a general linear model, such that the uncertainty on the parameters can be obtained. Having the uncertainty implies that testing parameters from different clusters can be done with a Wald-test (see Wood p. 110 [42])

$$\chi^2 = \frac{(\hat{\beta}_{ik} - \hat{\beta}_{ik'})^2}{\sigma^2(\hat{\beta}_{ik}) + \sigma^2(\hat{\beta}_{ik'})} \sim \chi^2(1) \tag{3.22}$$

Where the correlation between parameters for different clusters are assumed absent, which imply that the variance of the differences become the sum of the parameter variances. Testing all pairwise comparison would be troublesome, since the probability of making type I errors would blow up, however calculating the χ-values for all comparisons will give some insight information on how different the clusters really are.

Furthermore, since all observations are weighted in all three clusters the standard error for the estimates are smaller compared to a hard clustering (only using observations classified as low in the estimation for the low group etc.). The uncertainty on the parameters decrease with the number of observations, since the dispersion matrix for the parameter-vector is (see for example Conradsen p. 114 [9])

$$D(\hat{\beta}) = \sigma^2(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x})^{-1} \tag{3.23}$$

where Σ is a diagonal matrix with the posteriors in the diagonal and $\mathbf{x}$ the design matrix. Since Σ has non-negative weights and is a diagonal matrix, the parenthesis in (3.23) will increase for an increasing number of observations, which imply that the uncertainty decreases. The χ-scores will therefore be estimated based on a model with hard clustering but with the posteriors corresponding to the included individuals. This gives higher standard errors for the parameters, which should reflect the actual amount of information at hand and hence give more reasonable comparisons.

The χ-scores are shown in Table 3.12, which shows that the estimated intercept for the high group (group 3) is seen to have a large standard error making it insignificantly different from the middle and low group (critical $\chi^2(1)$ is 3.84 at a 5 % level). The other parameter comparisons show that the low and middle group have the same effect of PD15 PtcO2 and that the middle and high group have the same effect of starting late in day-care. The main differences between the groups are seen to be the age-related development in symptoms, which is seen to be highly significantly different from cluster to cluster.

Parameter	1-2	1-3	2-3
Intercept	17.0006	3.3781	1.4742
age	6.6438	23.7358	49.0353
age^2	11.3402	11.8021	37.4737
log10(pd)	0.8553	5.3323	9.4150
daycare.start/30	7.4530	0.8270	2.6331

Table 3.12: χ-scores for parameter comparisons across clusters

Predictions from mixture model

Based on the mixture model with the optimal number of clusters, $K^{\star}$, predictions can be made. For each of the $K^{\star}$ clusters the median age of day care start and the median PD15 value are used to predict one line per cluster.

From Figure 3.18 it is seen that the three clusters have different temporal shapes. Cluster 2 keeps a constant low level, cluster 1 starts at a medium level has a minor increase towards the age of 2 years and then a decline. Cluster 3 starts higher than the other two more than doubles the fraction from age 1 to 3 and then declines back to a level higher than the starting point. The model differentiates between three types of children: Children with many symptoms, which probably are the asthmatic children, a middle group, which may consists of non-asthmatic children or perhaps asthmatic children and a group with no symptoms and hence probably non-asthmatic children.

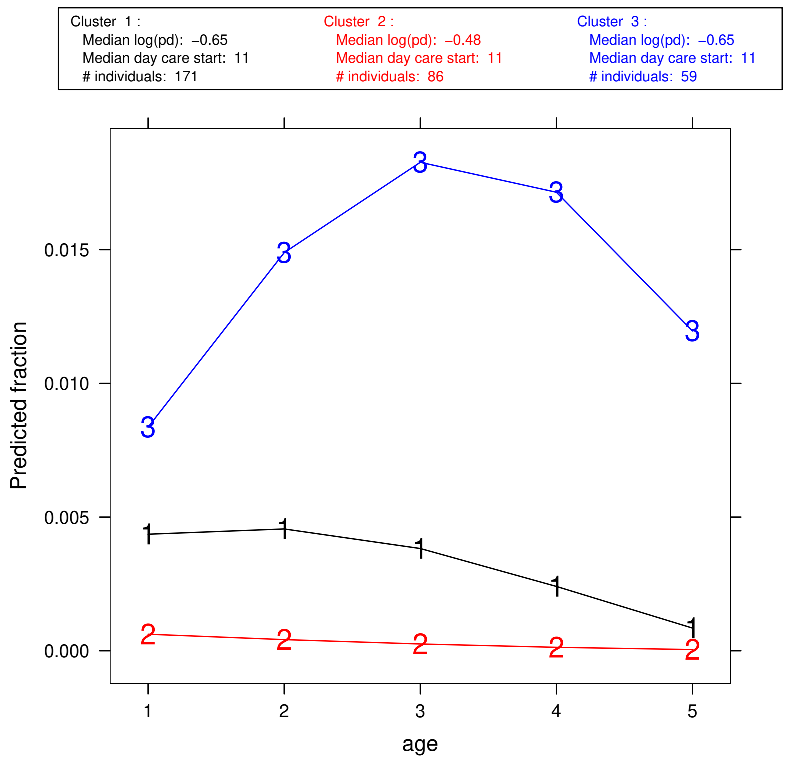

Figure 3.18: Prediction for each of the clusters at the median PD15 and age of day care start

3.3.4 Mixed effects model revisited

With the starting point in the mixed effects model from section 3.2 formulated in equation (3.8) and the clusters obtained from the LCR, a transformation from a mixed effects model to a purely deterministic model is sought. This can be done by introducing a *cluster* variable, which is a grouping variable giving the clusters from the LCR. By introducing the *cluster* and its interaction with all other variables, a new model can be formulated

$$\begin{aligned}\tilde{y}_{ijk} =& \beta_0 + \alpha_k + \beta_{1k} \cdot \text{age}_{ijk} + \beta_{2k} \cdot \text{age}^2_{ijk} + \beta_{3k} \cdot \log_{10}(pd_i) \\ &+ \beta_{4k} \cdot \log(\text{daycare}_{\text{start},i}) + \varepsilon_{ijk} \\ i = 1,&\ldots,m \quad j = 1,\ldots,n_i \quad k = 1,\ldots,3 \quad \boldsymbol{\varepsilon}_k \sim \mathcal{N}(\mathbf{0}, \sigma_k^2 \mathbf{G})\end{aligned} \tag{3.24}$$

where $\alpha_k + \beta_0$ is the mean level for cluster k, β_{ik} is the estimated for parameter i in cluster k and σ_k^2 is the variance for cluster k and $\mathbf{G}$ is a matrix describing the within subject correlation. The correlation structure is assumed to be exponential, which imply that observations taken on the same child are correlated by the amount

$$\rho(y_{ijk}, y_{ij'k}) = e^{-d/r} \tag{3.25}$$

where d is the distance in age and r the range (to be estimated). The corresponding semi-variogram then becomes $\gamma(d) = \sigma^2(1 - e^{-d/r})$. The separate variances for each cluster are included, since cluster 2 has lower variance compared to the two other clusters.

The model uses 19 degrees of freedom compared to 9 for the original mixed effects model, which for the likelihood ratio 632.1543 gives a test value of p<0.0001. This shows that the complication of the model leads to a highly significant increase in the log-likelihood.

The estimated semivariogram with the theoretical semivariogram appended is shown in Figure 3.19. It is seen that the fit is not too bad, but has a tendency to underestimate for high values of d. However the deviations are not too severe and changing the correlation-structure does not solve the problem nor change the parameter estimates significantly. Accounting for some type of correlation within each individuals observation is important, since each observation will contribute with too much information otherwise and hence give too small standard errors on the parameter estimates and thereby making it too easy to obtain significance. The estimated range of the correlation is 0.40 years, which shows that the correlation is small. The estimated correlation for observations 1 year apart is $e^{-1/0.41} = 0.09$. A likelihood ratio test against not having a correlation structure gives p=0.0084, which shows that the correlation is significant, although small.

A summary for the model is given in Table 3.13, which shows that eg the effect of day care start differ from cluster to cluster. This is seen to be case for all interactions terms and the model can not be reduced any further. The estimated ratios in standard deviations for cluster 2 and 3 compared to cluster 1 are 0.6472 and 0.9217, respectively.

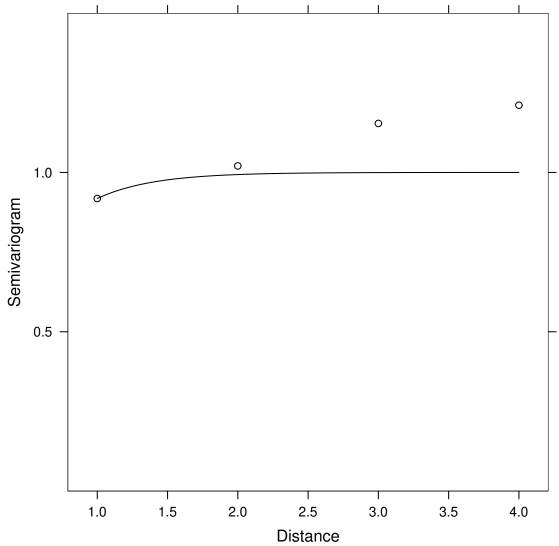

Figure 3.19: Semivariogram for residuals from model in (3.24) with assumed exponential correlation for the within-individual error. o: Sample semivariogram; —: fitted exponential semivariogram (range=0.40)

Comparing to a model without cluster specific variance, i.e. $\sigma_1 = \sigma_2 = \sigma_3$, gives a test of p<0.0001, which shows that the clusters have significant different variance levels.

Residual checking

The residuals from the model are examined by means of a QQ-plot to check the normality assumption. It is seen from Figure 3.20 that the normality assumption seems to be fulfilled with some minor deviations in the upper tail. It is also seen that the residuals have no pattern when plotting it against the covariates, whereas when plotting residuals against clusters, cluster 1 is seen to have the residuals skewed towards large positive values. This could indicate that some of the children in the middle group may indeed belong to the high. Furthermore, the difference in variance in group 2 is apparent in the lower right part of Figure 3.20.

Predictions

Based on the model in (3.24) predictions with confidence bands can be estimated. For simplicity the correlation within individual is neglected since it is low. In Figure 3.21 predictions and corresponding confidence bands are shown. The confidence bands are based on confidence bands on the root-arcsine scale and then scaled back to the

	Value	Std.Error	t-value	p-value
(Intercept)	0.0362	0.0073	4.9849	<0.0001
factor(cluster)1	0.0410	0.0109	3.7620	0.0002
factor(cluster)3	0.0162	0.0140	1.1573	0.2473
age	-0.0049	0.0048	-1.0120	0.3117
(age^2)	0.0001	0.0008	0.1448	0.8849
(log10(pd))	-0.0010	0.0022	-0.4600	0.6456
(daycare.start/30)	-0.0008	0.0003	-2.2855	0.0224
factor(cluster)1:age	0.0187	0.0072	2.5881	0.0097
factor(cluster)3:age	0.0649	0.0095	6.8631	<0.0001
factor(cluster)1:(age^2)	-0.0039	0.0012	-3.2653	0.0011
factor(cluster)3:(age^2)	-0.0094	0.0016	-5.9908	<0.0001
factor(cluster)1:(log10(pd))	-0.0035	0.0031	-1.1496	0.2505
factor(cluster)3:(log10(pd))	-0.0103	0.0037	-2.7907	0.0053
factor(cluster)1:(daycare.start/30)	-0.0013	0.0005	-2.6383	0.0084
factor(cluster)3:(daycare.start/30)	-0.0007	0.0005	-1.3272	0.1847

Table 3.13: Summary for model in (3.24)

rate scale, which imply that the bands are not entirely correct. They do however give some insight information on the uncertainty of the predicted fractions in the three clusters.

Figure 3.21 shows that the three clusters are different, the two clusters with the lowest levels seem to approach each other at the age of 5 years. It is seen that 257 children have few episodes (either a consistent low level or a median level at start and then a decline) and 59 have a consistent high level of episodes. The LCR gives some rather clear prototypes of children compared to the mixed effects model, where a grouping of the children was not immediately apparent. However, it has been shown that the estimated BLUPs from the mixed effects model and the cluster are connected and the heterogeneity related to group differences.

3.3.5 Existing literature

In the pediatric asthma field the study by Martinez et al. [26] is the main reference. In the study, the authors operates with two measurements times: 3 years and 6 years of life and two states: Wheezing and no wheezing. This gives 4 possible combinations and hence groups of children; no wheezing: no wheezing at age 3 and 6, transient wheezing; wheezing at age 3 but not at age 6, late onset wheezing; no wheezing at age 3 but wheezing at age 6 and persistent wheezing; wheezing at both age 3 and 6.

The analysis above for 3 groups was carried out, since 3 clusters was the optimal number of clusters based on data. One could do a similar analysis for 4 clusters to analyze the types of temporal development this would lead to. The LCR analysis is therefore done for a 4 cluster model, which gives the predictions shown in Figure 3.22.

It is seen from Figure 3.22 that the results from Martinez et al. [26] seem not to be reflected in the model. In the 4 cluster model three clusters ends at a low level and

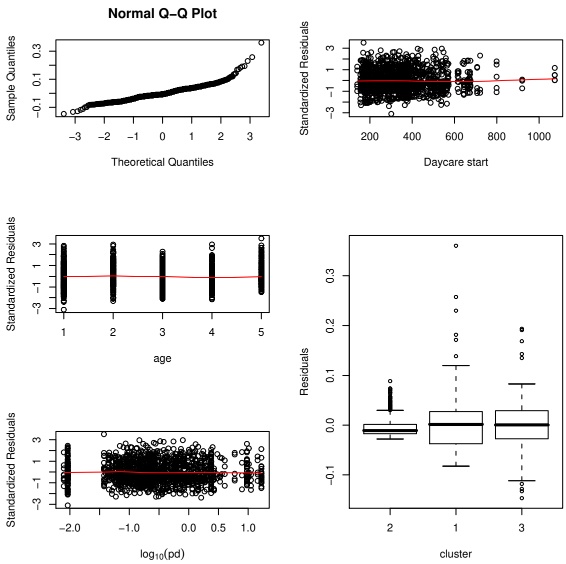

Figure 3.20: QQ-plot for linear model with cluster as factor

one ends at a high level. Two clusters starts at a high level of which one rapidly declines to the lowest level at the age of 5 years, whereas the other increases until the age of 3 and then declines to a level above the starting level but still at the highest level of all clusters. Another cluster starts at a median level then increases until the age of 2-3 years and then declines to a level a little above the two clusters with the lowest end-level. The last cluster starts and maintains a constant low level. To reflect the scheme by Martinez et al. [26], here for age 1 and 5, the group with a median starting level should have increased to a higher level at the end corresponding to late onset wheezing.

In Table 3.14 comparisons between cluster assignments in the $k = 3$ and $k = 4$ models are done. It is seen that the middle curve in $k = 3$ is a mixture of the three lowest curves in $k = 4$, the highest in $k = 3$ is composed of the middle and the highest curves in $k = 4$ and the lowest curve in $k = 3$ of the lowest curve and the curve starting high but ending low in $k = 4$. The estimated parameters compared to the estimates obtained from the mixed effects model are shown in Table 3.15, which shows that the most obvious difference is the fact that the high onset, low final level group has a positive curvature compared to the others having a negative curvature.

The model with $k = 4$ gives some extra information with respect to types of children, since a group with a high onset and an improvement in the symptoms is seen. However the results do not conform with the results found by Martinez et al. [26] and the model with $k = 3$ was seen to be statistically better. This imply that the 4 cluster model

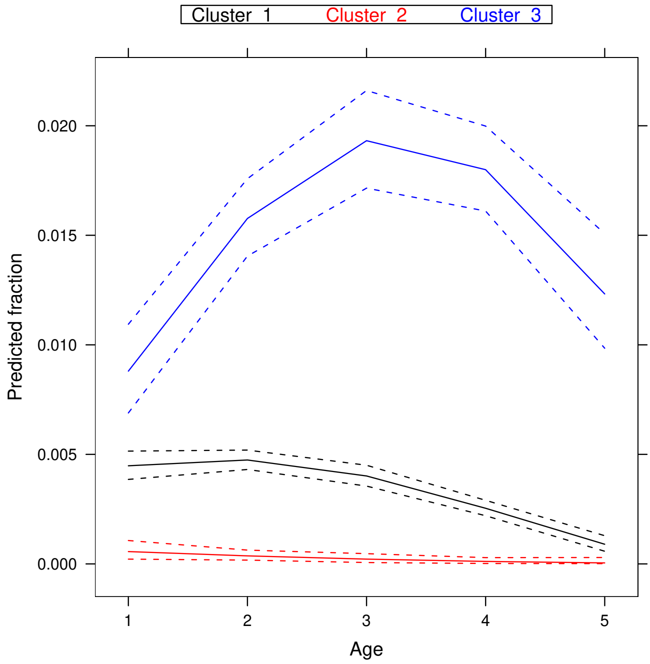

Figure 3.21: Predictions for (3.24) for each of the three clusters with confidence bands. PD15 and day care start are kept fixed at median level in all clusters

does not give significant additional information in the description of the symptom-rates.

	Middle	High-high	Low-low	High-low
Middle	137	0	7	27
Low	0	0	74	12
High	6	53	0	0

Table 3.14: Cross tabulation for clusters assignment for k=3 (rows) and k=4 (columns)

3.3.6 Cluster size analysis

In Figure 3.23 a comparison of the maximum likelihood solutions for $K \in \{1, 2, 5, 6\}$ is shown, see Figure 3.18 and Figure 3.22 for $K = 3$ and $K = 4$, respectively. It is seen that for $K = 1$ a rather flat curve with more or less no temporal development (there is a tendency to some curvature) is estimated, which corresponds to deflating all

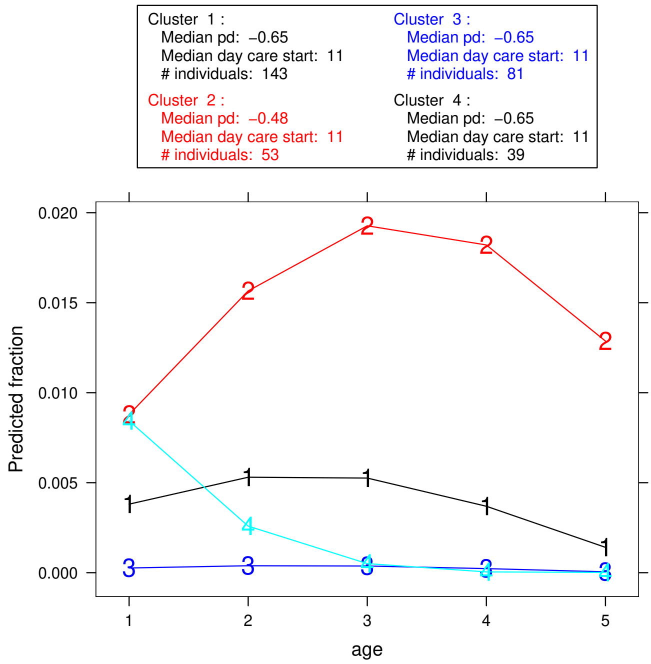

Figure 3.22: Fitted curves for 4 cluster model

Parameter	Middle	High-high	Low-low	High-low	Mixed effects
Number of individuals	143	53	81	39	316
Number of observations	634	248	369	172	1423
Intercept	0.0682	0.0546	0.0159	0.1342	0.0553
age	0.0283	0.0582	0.0092	-0.0601	0.0173
age^2	-0.0057	-0.0089	-0.0019	0.0063	-0.0038
$\log_{10}(pd)$	-0.0046	-0.0124	-0.0008	-0.0110	-0.0108
daycare start	-0.0028	-0.0015	-0.0006	0.0004	-0.0013
sigma^2	0.7427	0.7023	0.4874	0.5010	0.6292

Table 3.15: Parameters for the 4 cluster model

children to a no-symptom group. Expanding to a two cluster model gives a linearly decreasing curve and a parabola i.e. a symptom and a no-symptom group. The estimation in $K = 3$ gives a flat curve, a linearly decreasing curve and a parabola. Estimation for $K = 4$ gives a constant low group, an improving group, a group with no improvement (from and to a medium level) and a group worsening from a high level.

$K = 5$ gives two groups worsening but from different starting levels but at the same rate, the lines are close to being parallel. Two groups are improving, one with symptoms only at the first year of life and the other with symptoms declining through the whole interval, and finally a group at a constant low level. For $K = 6$, a group with a high starting and end level is seen (worsening group). The end-level is much higher compared to the other curves for $K < 6$. Furthermore, a group starts at a medium level with a worsening and then an improvement to a level below the starting level, a group is starting at the lowest level and increases a little. Finally the same three groups as the last three described for $K = 5$ are seen, i.e. two improving with different speed and one has a constant low level.

The comparison shows that there are some similarities for the clusters for different K's as one would expect. It is seen that most of the extra information when including additional clusters is related to children with few symptoms, whereas the high group is split into two as K is 5 or 6. This could explain why $K^\star = 3$ is the optimal number of clusters, since the clusters becomes more and more specialized to a certain group of children and hence do not attribute with enough extra information to justify the extra parameters.

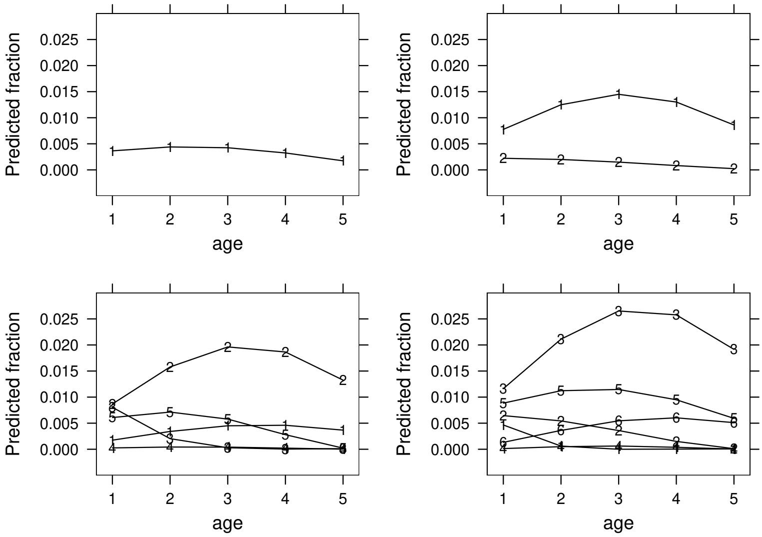

Figure 3.23: Comparison between different cluster sizes. Top-left: 1 cluster, top-right: 2 clusters, bottom-left: 5 clusters and bottom-right: 6 clusters (3 clusters: see Figure 3.21. 4 clusters: see Figure 3.22)

3.3.7 Consistency of grouping

Based on the parameters obtained from the optimal LCR model the likelihood for each child and year can be calculated. In the regression, the posterior probabilities were the posteriors for each individual and cluster (a cumulative probability of being in cluster k for all observations for individual i). In the following the posterior probabilities are allowed to vary within each individual's observations. The density for the j'th observation on individual i in cluster k is

$$f(\tilde{y}_{ijk}|x_{ijk}, \theta_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \cdot e^{-\frac{(\tilde{y}_{ijk} - x_{ijk}\beta_k)^2}{2\sigma_k^2}} \tag{3.26}$$

where x_{ijk} is the observation matrix and θ_k the parameter-vector given as $\theta_k = (\beta_1, \ldots, \beta_n, \sigma_k)$. From LCR the prior probabilities are estimated to be 0.52, 0.27 and 0.21 for cluster 1, 2 and 3. Hence the cluster corresponding to the middle group in Figure 3.18 is more than twice as likely a priori as the two other groups. The posterior probabilities can then be estimated by

$$\hat{p}_{ijk} = \frac{\pi_k f(\tilde{y}_{ijk}|x_{ijk}, \theta_k)}{\sum_{k'=1}^{3} \pi_{k'} f(\tilde{y}_{ijk'}|x_{ijk'}, \theta_{k'})} \tag{3.27}$$

Based on the posterior probabilities the cluster for observation j on individual i, c_{ij}, can be estimated by

$$\hat{c}_{ij} = \arg\max_k \ \hat{p}_{ijk} \tag{3.28}$$

hence assigning the observation to the most likely cluster. Comparing the c_i's (each individual belongs to one and only one cluster) with the c_{ij} gives the cross tabulation shown in Table 3.16, which shows that all groups are assigned the same way in above 50 % of the time. The children in the cluster corresponding to the group with the highest symptom rates (cluster 3) seem to be the most difficult to classify at each time-point.

	Type	A1C1	A1C2	A1C3
C1	count	532	166	55
	r-%	71	22	7
	c-%	66	40	27
C2	count	152	240	0
	r-%	39	61	0
	c-%	19	59	0
C3	count	123	4	151
	r-%	44	1	54
	c-%	15	1	73

Table 3.16: Cross-tabulation of c_i vs. c_{ij}. Columns correspond to c_{ij} and rows to c_i. Counts the actual counts, r% the row per cent and c% the column per cent.

Yearly grouping

Instead of considering all times points one could analyze, how early a cluster prediction can be done consistenly. This would give the possibility to classify children at a lower age than the age of 5. Using the predicted cluster at age j and comparing with the cluster found when using all observations in the clustering gives Table 3.17 and 3.18.

From the tables it is seen that age 1, 4 and 5 years are the worst-performing ages to predict the clusters, whereas the age of 2 and 3 seem to perform better. The overall aggrements are 59 %, 71 %, 70 %, 67 % and 57 % for the information in age 1, 2, 3, 4 and 5 years, respectively. The agreement is seen not to be too high, which can be explained by the relatively little amount of information used at each age. It is however seen that children classified overall as coming from the low group (group 2) is never classified as belonging to the high group in the yearly classifications. Furthermore, only 4 children from the high group is classified as belonging to the low grouping at the age of 1 years. It is important that the model gives consistent group and that the group-changes are limited to changing to the nearest group (i.e. from the low to the middle or the middle to the high), which is seen to be fulfilled.

	Type	A1C1	A1C2	A1C3	A2C1	A2C2	A2C3
C1	count	121	44	4	125	16	21
	r-%	72	26	2	77	10	13
	c-%	61	42	40	71	24	34
C2	count	30	56	0	32	52	0
	r-%	35	65	0	38	62	0
	c-%	15	54	0	18	76	0
C3	count	47	4	6	18	0	40
	r-%	82	7	11	31	0	69
	c-%	24	4	60	10	0	66

Table 3.17: Cross-tabulation of c_i vs. c_{ij}. Columns correspond to c_{ij} for the age of 1 and 2 years respectively and rows to c_i. *count* shows the actual number for the given cell, whereas r-% shows the row percent (within age) for the given cell and c-% the column percent. (A's correspond to Age and C's to cluster)

Grouping on cumulated information

Another possibility is to use cumulated probabilities for observations up to the age of j' years, i.e. use $(\tilde{y}_{ijk}, x_{ijk})$ for $j = 1, \ldots, j'$ and then base the clustering on these new probabilities. This can be done as for the posterior probabilities in (3.17) but

	Type	A3C1	A3C2	A3C3	A4C1	A4C2	A4C3	A5C1	A5C2	A5C3
	count	108	42	6	90	45	8	88	19	16
C1	r-%	69	27	4	63	31	6	72	15	13
	c-%	72	42	14	70	44	19	56	54	33
	count	22	58	0	18	58	0	50	16	0
C2	r-%	28	72	0	24	76	0	76	24	0
	c-%	15	58	0	14	56	0	32	46	0
	count	19	0	37	21	0	35	18	0	33
C3	r-%	34	0	66	38	0	62	35	0	65
	c-%	13	0	86	16	0	81	12	0	67

Table 3.18: Cross-tabulation of c_i vs. c_{ij}. Columns correspond to c_{ij} for the age of 3, 4 and 5 years respectively and rows to c_i. *count* shows the actual number for the given cell, whereas r-% shows the row percent(within age) for the given cell and c-% the column percent. (A's correspond to Age and C's to cluster)

with a different product term

$$\hat{p}_{ik}(j') = \frac{\pi_k \prod_{j=1}^{j'} f(\tilde{y}_{ij}|x_{ij}, \theta_k)}{\sum_{k'=1}^{K} \pi_{k'} \prod_{j=1}^{j'} f(\tilde{y}_{ij}|x_{ij}, \theta_{k'})} \tag{3.29}$$

Based on the new posterior probabilities $\hat{p}_{ik}(j')$, clustering can be done for the age of 2-5 years, since the age of 1 years is covered by the approach leading to Table 3.17. The procedure utilizes the information at hand at the age of j' years, rather than just using the local information and neglecting the past.

It is seen from the Tables 3.19 and 3.20 that the agreement between using all information and cumulated information increases for increasing age as expected, to total agreement for the age of 5 as required. The overall aggrement is 75 %, 88 %, 93 % and 100 % for the cumulative information from age 1 to age 2, 3, 4 and 5 years respectively.

It is seen that going from $j' = 1$ to $j' = 2$ improves the agreement the most, whereas the steps from $j' = 3$ to $j' = 4$ and $j' = 4$ to $j' = 5$ give smaller increases. This can be explained by considering the curves in Figure 3.21, which shows that the 3 groups are close at the age of 1 years and then evolve in separate directions, which are quite distinct at the age of 2 years. This imply that knowing the symptom history for the first two years of life, reduces the need for additional information, i.e. the significance of the information in the third, fourth and fifth years of life is small when relative to the first two years of life.

	Type	A2C1	A2C2	A2C3	A3C1	A3C2	A3C3
	count	126	15	21	140	8	8
C1	r-%	78	9	13	90	5	5
	c-%	76	19	34	88	11	14
	count	21	63	0	13	67	0
C2	r-%	25	75	0	16	84	0
	c-%	13	81	0	8	89	0
	count	18	0	40	7	0	49
C3	r-%	31	0	69	12	0	88
	c-%	11	0	66	4	0	86

Table 3.19: Cross-tabulation of c_i vs. $c_{i(1:j)}$. Columns correspond to $c_{i(1:j)}$ for the age of 2 and 3 years respectively and rows to c_i. *count* shows the actual number for the given combination, whereas r-% shows the row percent (within age) for the given cell and c-% the column percent. The table is based on information from the age of 1 to 2 and 3 years respectively (A's correspond to Age and C's to cluster)

	Type	A4C1	A4C2	A4C3	A5C1	A5C2	A5C3
	count	137	3	3	123	0	0
C1	r-%	96	2	2	100	0	0
	c-%	91	4	5	100	0	0
	count	9	67	0	0	66	0
C2	r-%	12	88	0	0	100	0
	c-%	6	96	0	0	100	0
	count	4	0	52	0	0	51
C3	r-%	7	0	93	0	0	100
	c-%	3	0	95	0	0	100

Table 3.20: Cross-tabulation of c_i vs. $c_{i(1:j)}$. Columns correspond to $c_{i(1:j)}$ for the age of 4 and 5 years respectively and rows to c_i. *count* shows the actual number for the given combination, whereas r-% shows the row percent (within age) for the given cell and c-% the column percent. The table is based on information from the age of 1 to 4 and 5 years respectively (A's correspond to Age and C's to cluster)

3.4 Generalized additive mixed model

In the following sections yearly aggregated episodes and corresponding days at risk are considered on a poisson scale. A generalized additive mixed effects model is fitted to grasp the curvature for the relation between age and the link. Previous modelling of the yearly aggregated data was based on the assumption that the relative number of wheezing episodes was gaussian, which can be questionable. Another possibility would be to model the data as binomial, i.e. model the probability of an episode a given days. In that analysis, which will not be done, the number of days at risk would correspond to the number of trials and the number of episodes to the number of successes. However since the number of days at risk is large and the proportion under 50 % the poisson distribution should not be affected by the theoretical upper limit for the number of episodes.

3.4.1 Model formulation

A perhaps more reasonable approach, than considering the relative number of episodes to be gaussian as in section 3.2, would be to assume that the number of episodes is poisson distributed. Y_{ij} is the number of episodes for individual i at age j, $n_{\text{days}_{ij}}$ is the number of exposure days and Y_{ij} is assumed to come from a poisson distribution with mean $E[Y_{ij}] = \mu_{ij}$ and variance $V[Y_{ij}] = \mu_{ij}$. However, in the following the quasi-poisson distribution is used giving $V[Y_{ij}] = \phi \cdot \mu_{ij}$, where ϕ is the overdispersion, see Wood p. 74 [42]. The canonical link function is used, which for a poisson distribution is a log-link: $\eta = \log(\mu)$.

A spline to estimate the curvature for age is fitted in order to model the symptom-rate. To account for the difference in the number of exposure days an offset of $\log(n_{\text{days}_{ij}})$ is included, which gives the model

$$\begin{aligned}\eta_{ij} = \log(\mu_{ij}) &= \log(n_{\text{days}_{ij}}) + \alpha + b_{0i} + s_1(\text{age}_{ij}) \\ &+ s_2(\log(\text{pd})) + s_3(\text{daycare}_{\text{start}}) \\ b_{0i} &\sim \mathcal{N}(0, \sigma_0^2)\end{aligned} \tag{3.30}$$

where b_{0i} is an individual baseline (for individual i) coming from a gaussian distribution with standard deviation σ_0. $s_1(\text{age})$ is the spline fitting the curvature for age, s_2 for PD15 PtcO2, etc. By including the offset, $\log(n_{\text{days}_{ij}})$, the rate rather than the absolute number of symptoms is modelled, since $\log(\mu_{ij}/n_{\text{days}_{ij}}) = \log(\mu_{ij}) - \log(n_{\text{days}_{ij}})$.

3.4.2 Results

From Figure 3.24 it is seen that the link-function for the yearly aggregated wheezing symptoms with respect to age probably is the right part of a parabola. The figure confirms the results from the mixed effects model based on the gaussian assumption, where a second order polynomial was used to model the temporal development. For PD15 and daycare start linear relations are seen, which shows that these variables can be included linearly. This leads to the same conclusion as for the gaussian model. A parabolic relation between age and the linear predictor seems appropriate, which change the model for the yearly aggregated symptoms to a parametric model given by

$$\begin{aligned}\eta_{ij} = \log(\mu_{ij}) = &\log(n_{\text{days}_{ij}}) + \alpha + b_{0i} + \beta_1 \cdot \text{age}_{ij} + \beta_2 \cdot \text{age}_{ij}^2 \\ &+ \beta_3 \cdot \log(\text{pd}_i) + \beta_4 \cdot \text{daycare}_{\text{start},i} \\ \mathbf{b_0} &\sim \mathcal{N}(0, \sigma_0^2 \mathbf{I})\end{aligned} \tag{3.31}$$

The scale parameter for the quasi-poisson distribution and the random effects will be dealt with in section 3.5 when modelling the age effects as well as the other effect with parametric models, since the non-parametric modelling is applied to be able to analyze the curvature.

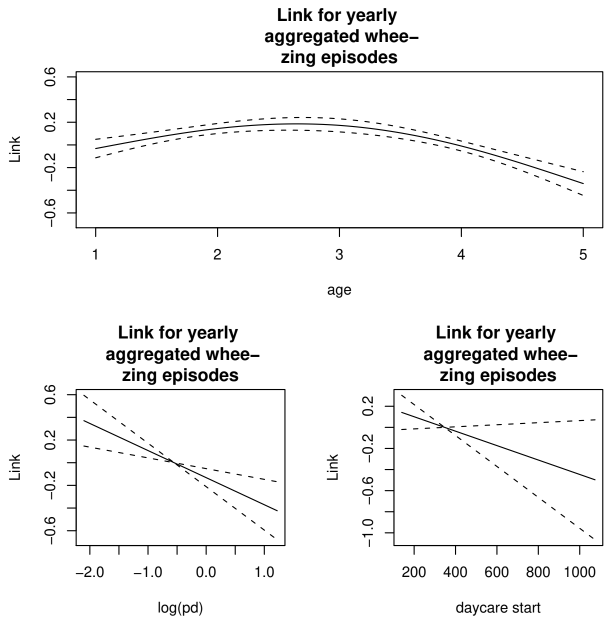

Figure 3.24: Generalized additive mixed effects model. Link functions in the left part, response in the right (scaled) for yearly aggregated episodes

3.5 Parametric modelling for poisson response

In section 3.4 a non-parametric method was used in order to find appropriate parametric functions for the relation between the linear predictor, η, and age, pd and day-care start. It was seen that for the yearly aggregated data a parabola could replace the spline to obtain a parametric representation. First only the age is considered as explanatory variable to analyze the amount of randomness to be incorporated in the model. This gives the initial model

$$\eta_{ij} = \log(\mu_{ij}) = \log(n_{\text{days}_{ij}}) + \alpha + b_{0i} + \beta_1 \cdot \text{age}_{ij} + \beta_2 \cdot \text{age}_{ij}^2 \tag{3.32}$$
$$\mathbf{b_0} \sim \mathcal{N}(0, \sigma_0^2 \mathbf{I})$$

which is a generalized linear mixed effects model (GLMM) with poisson response and gaussian random effects. The model is seen to be equal to the age part of the GAMM with exception of the smoothed function is replaced by a second order polynomial. It is assumed that the observation, Y_{ij}, comes from a quasi poisson distribution having $E[Y_{ij}|b_{0i}] = \mu_{ij}$ and $V[Y_{ij}|b_{0i}] = \phi \cdot \mu_{ij}$.

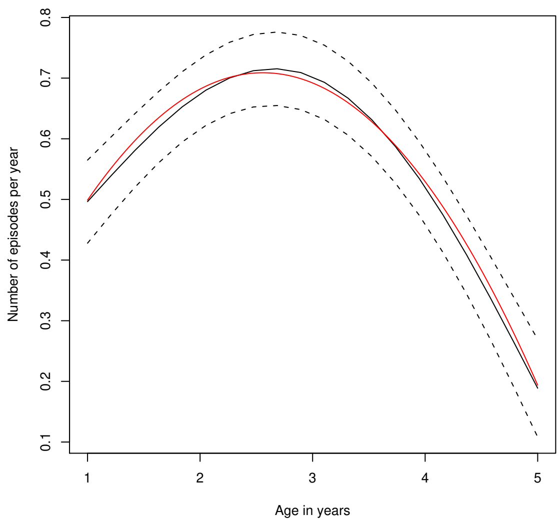

Figure 3.25: Comparison between non-parametric and parametric model for yearly aggregated wheezing symptoms (black: non-parametric model, red: parametric model)

Estimating the model gives a deviance of 2984 on 1565 degrees of freedom, which shows that the quasi-poisson distribution should be used due to the over-dispersion ($\hat{\phi} = 1.91$). The random effect, β_{0i} has variance $\hat{\sigma}_0^2 = 0.7^2$, which is highly significant since a model without the individual baselines has deviance 4678 on 1566 degrees of freedom. The children are seen to be heterogeneous, which the latent class regression also showed.

In Figure 3.25 the non-parametric model and the parametric model from (3.32) are plotted. The figure shows that the parametric model seems to give the same mean link-function as the non-parametric, which imply that the parameterization is seen to be adequate.

Different levels of randomness are now considered, which is shown in Table 3.21. It is seen that having individual intercept, first and second order parameters for age is significantly better compared to only having random intercept and slope, which again is significantly better compared to only having random intercept.

	Df	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
Random intercept	4	3013.58	−1492.08			
and slope	6	2772.19	−1364.02	256.11	2	<0.0001
and quadratic term	9	2767.28	−1350.53	26.98	3	<0.0001

Table 3.21: Comparison of models for yearly aggregated wheezing episodes

The most adequate model of the considered is

$$
\begin{aligned}
\hat{\eta}_{ij} = & \log(n_{\text{days}_{ij}}) + \alpha + b_{0i} + \beta_1 \cdot \text{age}_{ij} + b_{1i} \cdot \text{age}_{ij} + \\
& \beta_2 \cdot \text{age}_{ij}^2 + b_{2i} \cdot \text{age}_{ij}^2 \\
& b_{0i} \sim \mathcal{N}(0, \sigma_0^2) \quad b_{1i} \sim \mathcal{N}(0, \sigma_1^2) \quad b_{2i} \sim \mathcal{N}(0, \sigma_2^2) \\
& \begin{bmatrix} \mathbf{b_0} & \mathbf{b_1} & \mathbf{b_2} \end{bmatrix}^T \sim \mathbf{MVN}(\mathbf{0}, \mathbf{G})
\end{aligned}
\tag{3.33}
$$

This imply that each child have a separate intercept, slope and curvature and that these estimates may be correlated as described by the matrix **G**. The corresponding summary for the fixed effects is given in the upper part of Table 3.22, whereas the estimates of the random components and correlations are shown in lower part of Table 3.22. The random components are estimated as variances and covariances, the individual parameters can be estimated by means of BLUP-estimation (best linear unbiased predictor) [37]. It is seen that considerable correlations between the random components are present, which will be dealt with later.

	Fixed effects		
	Estimate	Std.error	t-value
(Intercept)	-5.6675	0.0113	-502.7725
age	0.4421	0.0071	62.4684
age^2	-0.1122	0.0002	-492.4423

	Random components		
	Variance-estimate	Correlation	
(Intercept)	0.99	(Intercept)	age
age	0.53	-0.61	
age^2	0.02	0.47	-0.89
residuals	0.97		

Table 3.22: Top: Summary for fixed effects in initial GLMM with random intercept, slope and curvature. Bottom: Random components for initial GLMM with random intercept, slope and curvature

3.5.1 Risk-factors

Including the $\log(PD15PtcO_2)$ and the time of day-care start in the model (the significant risk-factors in the gaussian model) gives an increase in the log-likelihood of 8.7 for 2 extra degrees of freedom, which imply that the congenital responsiveness and day-care start are significant (p= 0.0002). The relation between residuals and the variables from Figure 3.26 are seen to be linear. The updated model therefore becomes

$$\begin{aligned}\hat{\eta}_{ij} = \log(n_{\text{days}_{ij}}) + \alpha + b_{0i} + (\beta_1 + b_{1i}) \cdot \text{age}_{ij} + \\ (\beta_2 + b_{2i}) \cdot \text{age}_{ij}^2 + \beta_3 \cdot \log_{10}(\text{PD15 PtcO}_2)_i + \\ \beta_4 \cdot \text{daycare}_{start,i} \\ b_{0i} \sim \mathcal{N}(0, \sigma_0^2) \quad b_{1i} \sim \mathcal{N}(0, \sigma_1^2) \quad b_{2i} \sim \mathcal{N}(0, \sigma_2^2) \\ \begin{bmatrix}\mathbf{b_0} & \mathbf{b_1} & \mathbf{b_2}\end{bmatrix}^T \sim \mathbf{MVN}(\mathbf{0}, \mathbf{G})\end{aligned} \tag{3.34}$$

The fixed effects and random components are summarized in Table 3.23. The estimated coefficient for $\log_{10}(\text{PD15PtcO}_2)$ is -0.24, which imply that the rate-ratio for an individual increasing it's $PD15PtcO_2$ with a factor 10 is $e^{-0.24} = 0.7866$ (for high $PD15PtcO_2$ divided with low $PD15PtcO_2$). The rate-ratio corresponds to the effect of changing the PD15 PtcO2 with all other variables kept fixed and corresponds to a decrease in the symptom-rate by 21 %. For the day-care start variable the estimated coefficient is -0.0296, which imply that the ratio for a child starting in day-care one month later is $e^{-0.0296} = 0.9708$, which is a decrease by 3 %.

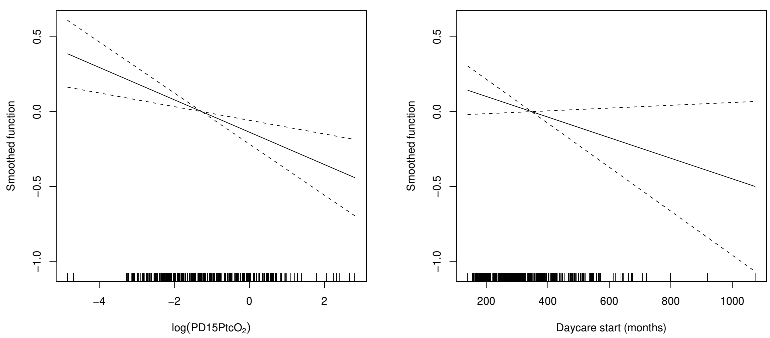

Figure 3.26: Smoothed curves for log(PD15 PtcO_2) and day-care start

The estimated second order polynomial is seen to be highly significant, since both the first and second order parameters are significant. Furthermore, the random components are significant, i.e. a test for $\sigma_2^2 = 0$ gives a likelihood ratio of 23.08 on 11 degrees of freedom, which gives p<0.0001. The heterogeneity of the slopes is seen to be large, eg $\sigma_1 > \beta_1$. This could be a result of an underlying grouping as analyzed in section 3.3 for the gaussian model or may reflect that the children are heterogeneous.

The random components are seen to be heavily correlated, which may be reduced by introducing orthogonal polynomials to replace age and age^2. Orthogonal polynomials

	Fixed effects		
	Estimate	Std.error	t-value
(Intercept)	-5.5305	0.0295	-187.5948
age	0.5129	0.0074	69.5582
agekvar	-0.1248	0.0002	-521.7878
I(log10(pd))	-0.2400	0.0048	-49.5281
I(daycare.start/30)	-0.0296	0.0001	-229.1517

	Random components		
	Variance-estimate	Correlation	
(Intercept)	0.84		
age	0.41	-0.55	
agekvar	0.01	0.37	-0.86
residuals	0.98		

Table 3.23: Top: Summary for fixed effects in updated GLMM with random intercept, slope and curvature. Bottom: Random components for updated GLMM with random intercept, slope and curvature

of order 2, see Wood p. 305 [42], are defined such that the inner-product of the vectors containing the transformed first and second order for any two observations is zero. Re-fitting the model with orthogonal polynomials gives an updated summary shown in Table 3.24. It is seen that the correlations for the random components are greatly reduced, but not entirely.

Using centering of the age variable (Wood p. 305 [42]) gives almost the same results as using orthogonal polynomials, which shows that the correlation between the random components is high. Since introducing the orthogonal polynomials and the centering do not give a large reduction of the correlation, the original model without orthogonal polynomials and centering is kept for simplicity.

	Fixed effects		
	Estimate	Std.error	t-value
(Intercept)	-5.3290	0.0216	-246.3834
I(log10(pd))	-0.2400	0.0048	-49.5281
I(daycare.start/30)	-0.0296	0.0001	-229.1517
poly(age, 2)1	-12.3812	2.7100	-4.5687
poly(age, 2)2	-8.2143	1.0359	-7.9295

	Random components		
	Variance-estimate	Correlation	
(Intercept)	0.86		
poly(age, 2)1	386.11	0.52	
poly(age, 2)2	58.04	-0.07	0.37
residuals	0.98		

Table 3.24: Top: Summary for fixed effects in updated GLMM with random intercept, slope and curvature. Bottom: Random components for updated GLMM with random intercept, slope and curvature

Diagnostics

In Figure 3.27 diagnostics based on the deviance residuals are shown. It is seen that the quantile plot seems to be sufficiently linear. The deviance residuals plotted against the age is seen to be skewed toward the negative side, however not giving severe problems. The QQ-plot shows that no outliers seem to be rpesent and it is furthermore seen that the variance of the deviance residuals seem to be the same for different ages.

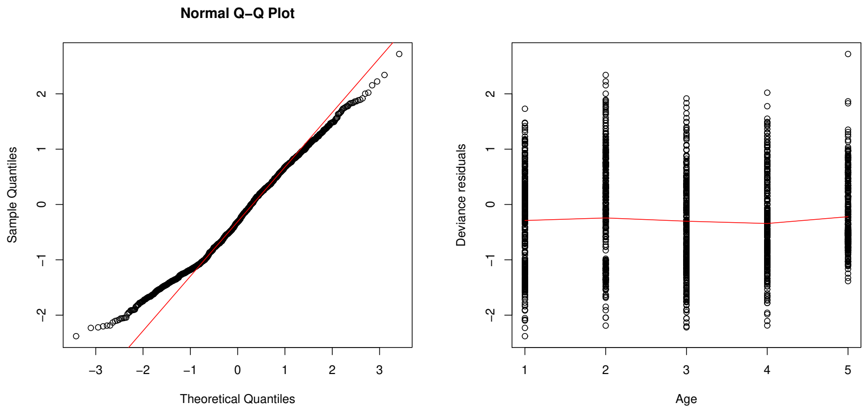

Figure 3.27: Diagnostics plot for model in (3.33). Left: Quantile plot of deviance residuals compared to N(0,1) quantiles. Right: Deviance residuals against age

Random component analysis

In Figure 3.28 QQ-plots of the estimated individual parameters (the Best Linear Unbiased Predictors as described by Robinson [37]) are shown. From the figure it is seen that the intercept has a skewed distribution, whereas the other two parameters seems to be closer to normality. A Shapiro-Wilk test for normality [39] gives p=0.0001, p<0.0001 and p<0.0001 for b_{0i}, b_{1i} and b_{2i} respectively, hence there is strong evidence against the normality assumption for the individual starting levels. A this stage the model is kept as it is, since the deviations do not seem too severe from the QQ-plots.

3.5.2 Prediction

Figure 3.29 shows the predictions for the model with three random components in (3.33). The curves for the individuals are seen to vary from increasing curves to decreasing curves. It is seen that a large group has more or less linear curves, $\beta_2 +$

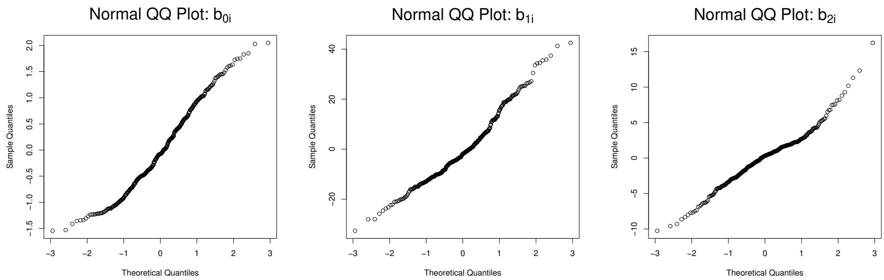

Figure 3.28: QQ-plots of individual parameters for model with yearly aggregated data

$b_{2i} \approx 0$. As seen for the gaussian model the children divides in roughly four groups: No symptoms, some symptoms but declining, many symptoms and an increasing symptom-rate and some symptoms and increasing rate. However, a clear grouping is not seen, which makes grouping based on the random components difficult.

As for the mixed effects model with gaussian errors the grouping can be examined by means of Latent Class Regression, which may give the possibility to find groups and compare them to the results from the gaussian case.

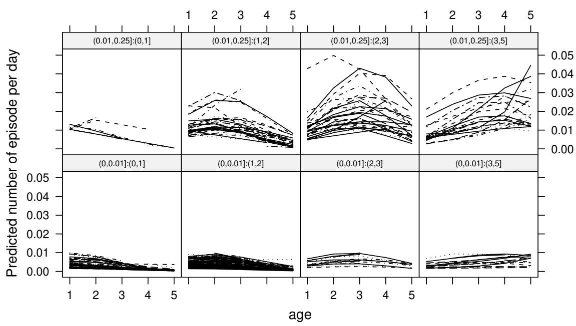

Figure 3.29: Predictions from model with individual parameters 0, 1. and 2. order parameters. The panels are different combinations of highest rate and age of highest rate, eg $(0, 0.1] : (1, 2]$ contains prediction for children with a maximum rate between 0 and 0.1 which occur between the age of 1 and 2 years of life.

3.6 Latent class regression for poisson response

In this section LCR is considered for the poisson response as described in section 3.5. The basis for the LCR is the model found in section 3.5.1 and the goal is to find groups of children having different between groups but same within groups longitudinal development in episode rate. The episodes are considered to be poisson distributed with an offset corresponding to $\log(n_{\text{days}})$ and the canonical link as link function ($\log(\mu)$). This leads to the formulation for cluster k

$$\begin{aligned}\hat{\eta}_{ijk} = \log(\hat{\mu}_{ijk}) = \log(n_{\text{days}_{ij}}) + \beta_{0k} + \beta_{1k} \cdot \text{age}_{ij} + \beta_{2k} \cdot \text{age}_{ij}^2 \\ +\beta_{3k} \cdot \log_{10}(pd_i) + \beta_{4k} \cdot (\text{daycare}_{\text{start},i})\end{aligned} \tag{3.35}$$

where μ_{ijk} is the intensity for the episodes and $\eta_{ijk} = g(\mu_{ijk}) = \log(\mu_{ijk})$ the linear predictor to be modelled. As for the gaussian latent class regression each individuals observations contribute to each cluster with the amount p_{ik} (posterior probability) and each cluster has a prior probability in each iteration which is π_k and is the probability a priori to be in cluster k. The estimation method follows the procedure outlined in section 3.3.2 with the density function though being different.

3.6.1 Model complexity

The optimal number of clusters can be found by means of BIC as seen for the gaussian case. For the gaussian case $K^\star$ was found to be 3, so for comparison reasons a similar number for a poisson model would be appreciated. In Figure 3.30 BIC as function of the number of clusters is shown, which shows that the optimal number of clusters is $K^\star = 5$. It is however seen that the biggest improvement in BIC is from a 1 cluster model to a 2 cluster model and an elbow is seen at $K = 4$. A 3 cluster model is seen not to be quite as good as the five cluster model, but is however better than a 2 cluster model.

In the gaussian case $K^\star$ was found to be 3, which imply that this model is of special interest for comparison purposes. This imply that both the 5 cluster situation and the 3 cluster model will be considered in the following. The three cluster model is easier to interpret since it has fewer trajectories, whereas the five cluster model is statistical better.

3.6.2 Five cluster model

For $K = 5$ the parameter estimates for each of the 5 clusters are shown in Table 3.25 and the corresponding fitted values are shown in Figure 3.31. The figure shows that the clusters (the one with the highest levels) are not quite the same as for the gaussian case in Figure 3.23. However some similarities are seen: A group with many symptoms and a group with no symptoms. It is furthermore seen that one of the clusters is rather small, which makes it highly dependent on the particularly individuals in the cluster.

To evaluate the level of overdispersion, a standard generalized linear model is fitted to each cluster with weights corresponding to the posterior probabilities, p_{ik}. The

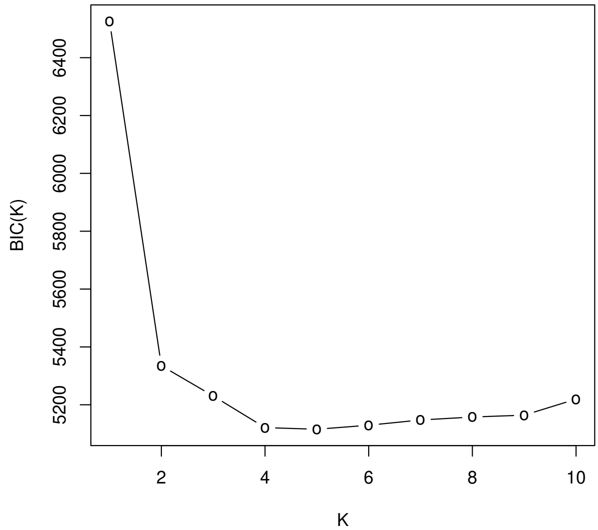

Figure 3.30: BIC for different K's

Parameter	Cluster 1	Cluster2	Cluster 3	Cluster 4	Cluster 5	GLMM
# of ind	75	34	16	57	134	316
# of obs	319	164	75	266	599	1423
Intercept	-5.3729	-6.2062	-4.7711	-5.6050	-6.2278	-5.5188
age	1.1267	0.0984	0.8113	0.9291	0.3652	0.5126
age^2	-0.3299	0.0247	-0.1153	-0.1661	-0.1145	-0.1248
log10(pd)	-0.1125	-0.4021	-0.1907	-0.2729	-0.4281	-0.2395
daycare	-0.0308	0.0164	-0.0289	-0.0008	-0.0403	-0.0308

Table 3.25: Parameters for the clusters in the optimal poisson mixture model ($K = 5$)

relevant degrees of freedom for the residuals for cluster k is then

$$df_k = \sum_{i=1}^{m} n_i \cdot p_{ik} - \text{df(model)} \tag{3.36}$$

where n_i is the number of observations for individual i and df(model) $= 5$. Using the number of observations in the full observation set would be highly misleading, since most observations contributes to only one cluster, i.e. has posteriors around 1 for one cluster and 0 for the rest. The 25 % quantile for the largest posterior is 0.72, the smallest is 0.3 and the median is 0.91.

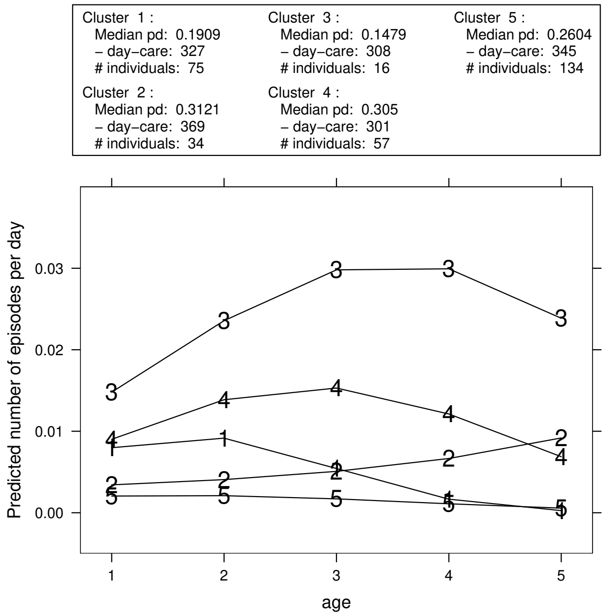

Figure 3.31: Predictions of the number of episodes per day for $K = 5$ for poisson LCR

The residual deviances are

$$D = \{400.79, 298.22, 130.83, 351.51, 601.28\}$$

with an estimated number of degrees based on the posterior probabilities

$$df = \{317.34, 191.61, 80.94, 258.12, 549.99\}$$

This gives estimated overdispersions of

$$\hat{\phi} = \{1.26, 1.56, 1.62, 1.36, 1.09\}$$

which do not seem too bad. The cluster having the largest overdispersion is the cluster with the fewest observations, which shows that the inference in this cluster might be questionable if the over-dispersion is not taking into account. The overdispersion is seen to be highest in the clusters with the fewest individuals.

The above estimation is based on the assumption that the observations within an individual are uncorrelated, which is questionable to be true. One could use Generalized Estimation Equations (GEE) to estimate both the (possible) over-dispersion i.e. the scale parameter and the correlation. The GEE-approach is to solve the GEE (see

Olsson p. 165 [28])

$$\sum_{i=1}^{m} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = 0 \tag{3.37}$$

where $\mathbf{Y}_i$ is the observations, $\boldsymbol{\mu}_i$ the corresponding mean values, $\mathbf{V}_i$ the covariance matrix for $\mathbf{Y}_i$ and $\boldsymbol{\beta}$ the parameters to be estimated. The $\mathbf{V}$ matrix is left unspecified to allow any type of covariance, since the number of correlation elements is moderat (10).

	gee/glm1	gee/glm2	gee/glm3	gee/glm4	gee/glm5
β_0	0.874	0.939	9.390	1.053	0.957
β_1	0.936	0.913	28.909	0.975	0.997
β_2	0.912	0.961	28.677	0.919	0.985
β_3	0.794	0.802	9.953	0.779	0.847
β_4	0.811	1.043	25.295	0.796	0.876

Table 3.26: Standard errors for parameter estimates for GEE divided by standard error for GLM

The estimated standard errors for the generalized linear model and from the GEE-approach are compared in Table 3.26, which shows that the difference is largest for the parameter corresponding to the $\log_{10}(\text{pd})$ measurement and day-care start. It is furthermore seen that the differences are extremely high for cluster 3, which shows that the parameter estimates in the GEE-model is very uncertain. The gee estimates are consistently as the number of individuals goes to infinity, Diggle p. 139-140 [15], even for a wrong specified correlation structure. However, for cluster 3 the number of individuals is small, which imply that the nice properties can not be guaranteed. The estimated scale parameters (over-dispersion) are

$$\hat{\phi} = \{1.39, 1.72, 59.26, 1.3, 1.29\}$$

with estimated standard errors

$$\hat{\sigma}_{\phi} = \{0.3204, 0.2703, 4.7299, 0.0839, 0.1177\}$$

Since the estimated scale-parameters are asymptotically normally distributed with mean 1, see Yan et al. [44], a Wald test can be performed on the quantity

$$\chi^2 = \left(\frac{\hat{\phi} - 1}{\sigma_{\hat{\phi}}}\right)^2 \sim \chi^2(1) \tag{3.38}$$

For the 5 cluster model, this gives the test-statistics

$$\chi^2 = 1.5, 7.03, 151.74, 12.58, 5.98$$

which shows that $\hat{\phi}_1$ is insignificant, whereas the others are significantly larger than 1 (critical value is 3.84).

It is furthermore seen that for cluster 3, the gee-estimation tends to give questionable results, hence giving a very high scale-parameter and much higher standard errors for the parameters (including the scale-parameter). This is probably caused by the small number of individuals in this particular group.

For the correlation matrix, 3 significantly correlations are estimated (for the upper triangle in the correlation matrix), which are distributed as 0, 2, 0, 0 and 1 on cluster $1, \ldots, 4$ and 5. The number of correlation parameters is 10 for each cluster, which shows that there do not seem to be much evidence for correlation between observations on the same individual. This was also the conclusion in section 3.3.4 for the gaussian three cluster case.

Finally the residuals can be inspected, this is done by taking the residuals from the ordinary generalized linear model, since the correlation in the gee-model was moderate. The deviance residuals are used, which imply that the sample quantiles of the residuals can be compared to quantiles in a standard gaussian distribution in a QQ-plot for outlier detection (see Olsson p. 57 [28]). From Figure 3.32 it is seen that deviance residuals seem to follow a standard gaussian distribution. For cluster 5 some deviations from normality is seen, but they do not seem to be severe. However, since normality is an approximation and none of the observations seems to be outliers, the residuals are found adequate.

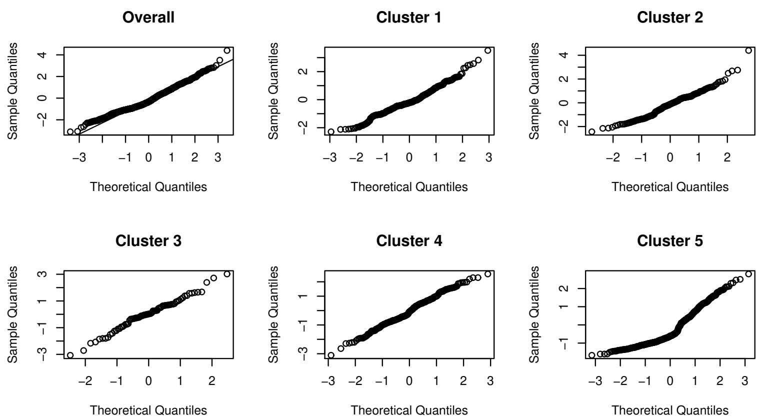

Figure 3.32: QQ-plot for deviance residuals for $K = 5$

3.6.3 Three cluster model

For the 3 cluster model the parameter estimates for each of the 3 clusters are shown in Table 3.27 and the corresponding predictions in Figure 3.33. The figure shows that the longitudinal development in the three clusters seems to coincide well with the longitudinal development in Figure 3.18 (the gaussian 3 cluster model) but with some differences. The cluster of children with the lowest level of symptoms starts higher in the poisson LCR compared to the gaussian LCR. The shapes of the curves and the corresponding interpretations seem to be similar, however.

Parameter	Cluster 1	Cluster 2	Cluster 3
# of ind	105	55	156
# of obs	469	250	704
(Intercept)	-4.5559	-5.0362	-5.5561
age	0.3849	0.8787	0.3226
(age^2)	-0.0841	-0.1329	-0.1321
(log10(pd))	-0.2701	-0.2579	-0.4378
(daycare.start/30)	-0.0777	-0.0335	-0.0654

Table 3.27: Parameters for the clusters in the 3 cluster model poisson mixture model

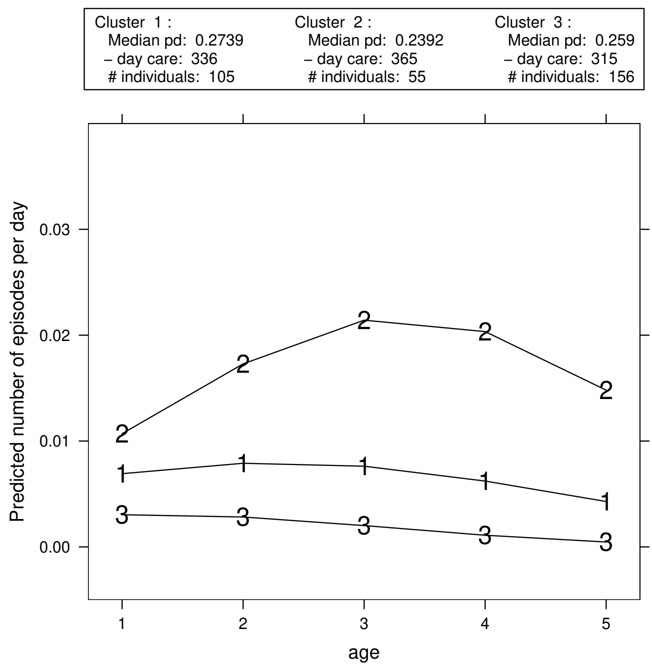

Figure 3.33: Predictions of the number of episodes per day for $K = 3$ for poisson LCR

It is seen that the group with the highest amount of episodes is the smallest, whereas the group with the fewest is the largest. In the gaussian case the middle group was the largest group and the two other groups half the size of it, which shows that the two methods do not lead to the same grouping. For the three cluster model the residuals deviances are

$$D = \{857.28, 431.69, 803.92\}$$

with an estimated number of degrees based on the posterior probabilities

$$df = \{495.94, 234.03, 678.03\}$$

This gives an estimated overdispersions of

$$\hat{\phi} = \{1.73, 1.84, 1.19\}$$

The over-dispersion parameters, ϕ, are in general seen to be a little higher compared to the 5 cluster model, which reflects that more heterogeneous children are placed in the same cluster for $K = 3$ compared to $K = 5$. However the difference is small and the heterogeneity are probably close to be the same.

	gee/glm1	gee/glm2	gee/glm3
β_0	0.656	0.632	1.047
β_1	0.567	0.570	1.181
β_2	0.491	0.578	1.416
β_3	0.664	0.876	1.136
β_4	0.854	0.674	1.409

Table 3.28: Standard errors for parameter estimates for GEE divided by the standard error from GLM

As for the five cluster model, the model can be examined by means of GEE, this leads to the comparison between the estimated standard errors for a standard generalized linear model and from the GEE-approach in Table 3.28. The table shows that the differences are largest for the parameter corresponding to the curvature and that the standard errors in the gee model are higher compared to the GLM-model for the first two clusters. The estimated scale parameters (over-dispersion) are

$$\hat{\phi} = \{1.85, 1.82, 1.42\}$$

with the standard errors

$$\hat{\sigma}_{\phi} = \{0.2018, 0.2004, 0.1668\}$$

The Wald-test statistic for testing $\hat{\phi}_k = 1$ therefore becomes

$$\chi^2 = 17.73, 16.92, 6.2$$

which shows that all scale-parameters are significantly larger than 1, since the critical value for a χ^2-distribution on 1 degree of freedom at a 5 % level is 3.84.

For the correlation structure 9 significantly elements are estimated, which are distributed as 4, 4 and 1 on cluster $1, 2, 3$, where the number of correlation elements is

10 for each cluster.This imply that there do seem to be some evidence for correlation between observations on the same individual.

For cluster 1 the significant elements are: 1:5, 2:4, 2:5 and 3:5 with corresponding estimates of -0.31, -0.29, -0.34 and -0.28. For cluster 2: 1:2, 1:3, 3:4 and 4:5 with corresponding estimates of 0.47, 0.41, 0.24 and 0.41 and for cluster 3: 1:2 with corresponding estimate of 0.13. It is hard to interpret how the correlation structure should be formulated, since most of the significant correlation parameters are small and distributed over many different combinations for the three clusters. Furthermore, the significant estimates are negative in cluster 1 and positive in cluster 2 and 3. It probably reflects that the correlation may be zero or at least of little significance.

Finally the residuals can be inspected, this is done by taking the residuals from the ordinary generalized linear model, since the correlations in the GEE-model were seen to be somewhat unstructured and small. The deviance residuals are used, which imply that a QQ-plot for the sample quantiles of the residuals can be used, where the reference distribution is a standard gaussian distribution (see Olsson p. 57 [28]). From Figure 3.32 it is seen that the deviance residuals seem to follow a standard gaussian distribution. For cluster 3 some deviations from a linear QQ-plot are seen, but they do not seem to be severe. The QQ-plot serves more as an analytic tool for outlier detection, since the outliers will be situated away from the straight line, which is seen not to be the case.

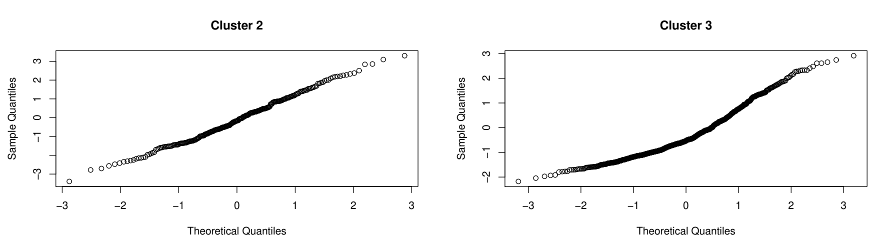

Figure 3.34: QQ-plot for deviance residuals for $K = 3$

3.6.4 GEE-model

The three models, corresponding to the 3 groups found for $K = 3$, are now combined in a total model. Interactions between clusters obtained from the LCR and the effects in the model are included to model cluster differences in the parameters. This leads to the model

$$\begin{aligned} \hat{\eta}_{ij} = \log(\hat{\mu}_{ij}) = \log(n_{\text{days}_{ij}}) + \beta_{0k} + \beta_{1k} \cdot \text{age}_{ij} + \beta_{2k} \cdot \text{age}_{ij}^2 \\ + \beta_{3k} \cdot \log_{10}(\text{pd}_i) + \beta_{4k} \cdot (\text{daycare}_{\text{start},i}) \\ k = 1, 2, 3 \end{aligned} \tag{3.39}$$

where β_{32} for instance is the parameter for $\log_{10}(pd)$ for the individuals in cluster 2. The correlation is assumed to be unstructured in the GEE-estimation. Combining in a total model gives one scale parameter, ϕ, which seems reasonable since the individual dispersion parameters were seen to be insignificantly different.

Fitting the model gives an estimated over-dispersion of $\hat{\phi} = 1.57$, which shows that the over-dispersion is at the same level as for the individual models. The parameter-estimates are shown in Table 3.29 with corresponding rate-ratios appended for each parameter, the latter corresponds to the effect of increasing the related variable one unit.

It is seen that the group with the highest number of episodes has high estimates of the first order variable, i.e. a high initial increase in the relative number of episodes. As the age increases the second order term begins to dominate and cancel most of the high slope out, which gives the decline from age 3 to 5 years for the high group. The two other groups are seen to have the same slope but the low group has a more negative curvature-parameter. It is furthermore seen that the low group has a lower intercept compared to the other groups.

The three groups are seen not to start from the same level at the artificial starting point, age = 0, when neglecting the contribution to the intercept coming from $\log_{10}$(pd) and day-care start. The cluster with the fewest symptoms starts significantly lower than the group with the middle level of symptoms. It is seen that the group with the fewest symptoms benefits the most of having a high congenital resistance and starting late in day-care. The estimated cluster difference for $\log_{10}$(pd) is seen to be insignificant, whereas the interaction for day-care start is seen to be significant. It is furthermore seen that the estimate of the interaction between cluster and curvature is insignificant.

Removing the interaction between cluster and pd and the curvature gives a new model with an esimated over-dispersion of $\hat{\phi} = 1.57$ for a model with 4 parameters fewer. The heterogeneity is seen to be the same after the reduction, i.e. the reduction do not introduce additional heterogeneity. The updated model is shown in Table 3.30, which shows that the children as a group get a reduction in the starting level of $1 - e^{-0.30} = 26$ % from increasing their PD15 PtcO2 level with a factor 10.

	estimate	san.se	wald	p	e^{β}
(Intercept)	-5.1116	0.2072	608.8829	<0.0001	0.0060
cluster1	0.5272	0.2663	3.9207	0.0477	1.6943
cluster3	-0.4371	0.3433	1.6211	0.2029	0.6459
age	0.9324	0.1129	68.1931	<0.0001	2.5405
(age^2)	-0.1408	0.0183	58.9708	<0.0001	0.8687
(log10(pd))	-0.2653	0.0658	16.2696	<0.0001	0.7669
(daycare.start/30)	-0.0339	0.0077	19.4274	<0.0001	0.9666
cluster1:age	-0.5205	0.1734	9.0094	0.0027	0.5942
cluster3:age	-0.5284	0.2327	5.1564	0.0232	0.5896
cluster1:(age^2)	0.0516	0.0303	2.8910	0.0891	1.0529
cluster3:(age^2)	-0.0042	0.0433	0.0095	0.9225	0.9958
cluster1:(log10(pd))	-0.0398	0.0763	0.2724	0.6017	0.9609
cluster3:(log10(pd))	-0.1433	0.0903	2.5177	0.1126	0.8665
cluster1:(daycare.start/30)	-0.0435	0.0110	15.5958	<0.0001	0.9575
cluster3:(daycare.start/30)	-0.0406	0.0143	8.0644	0.0045	0.9602

Table 3.29: Summary for combined gee model. San.se are the robust standard errors, which account for the correlation within an individuals observations as described by Prentice [34] or Diggle et al. p. 347 [15].

	estimate	san.se	wald	p	e^{β}
(Intercept)	-5.0006	0.1845	734.8688	<0.0001	0.0067
cluster1	0.2322	0.2016	1.3270	0.2493	1.2614
cluster3	-0.3638	0.2413	2.2740	0.1316	0.6950
age	0.8175	0.0879	86.4747	<0.0001	2.2649
(age^2)	-0.1214	0.0139	76.3842	<0.0001	0.8857
(log10(pd))	-0.3042	0.0358	72.3842	<0.0001	0.7377
(daycare.start/30)	-0.0341	0.0078	19.2797	<0.0001	0.9665
cluster1:age	-0.2283	0.0492	21.5468	<0.0001	0.7959
cluster3:age	-0.5286	0.0613	74.3160	<0.0001	0.5894
cluster1:(daycare.start/30)	-0.0435	0.0111	15.4328	<0.0001	0.9574
cluster3:(daycare.start/30)	-0.0408	0.0144	8.0496	0.0046	0.9600

Table 3.30: Summary for updated combined GEE model

For the day-care start the rate-ratios for an increase of 1 month are

$$\left(e^{-0.08}, e^{-0.03}, e^{-0.07}\right) = (0.93, 0.97, 0.93)$$

for cluster 1, 2 and 3, respectively. The rate-ratios show that cluster 1 and 3, which corresponds to the children with the two lowest levels of symptoms, have the highest benefit of starting later in day-care.

For the longitudinal development it was seen that the curvature was the same in the three clusters, whereas the slopes are different. The estimated slopes are

$$\hat{\beta}_{11} = 0.5892, \quad \hat{\beta}_{12} = 0.8175, \quad \hat{\beta}_{13} = 0.2889$$
$$e^{\hat{\beta}_{11}} = 1.8025, \quad e^{\hat{\beta}_{12}} = 2.2649, \quad e^{\hat{\beta}_{13}} = 1.3350$$

which shows that cluster 2 has the highest increase in the number of symptoms (if the curvature is neglected), which gives the high initial increase for the high group. The clusters have a maximum predicted rate at the age of

$$\hat{\text{age}}^{\star}_{k=1} = -\frac{\hat{\beta}_{11}}{2\cdot\hat{\beta}_2} = -\frac{0.5892}{2\cdot(-0.1214)} = 2.4271$$
$$\hat{\text{age}}^{\star}_{k=2} = -\frac{\hat{\beta}_{12}}{2\cdot\hat{\beta}_2} = -\frac{0.8175}{2\cdot(-0.1214)} = 3.3677$$
$$\hat{\text{age}}^{\star}_{k=3} = -\frac{\hat{\beta}_{13}}{2\cdot\hat{\beta}_2} = -\frac{0.2889}{2\cdot(-0.1214)} = 1.1901$$

It is seen that the cluster with the highest level symptoms tops the latest (age=3.37 years), then the middle group and the lowest group tops first. The uncertainty on the age of the maximum can be found by using the law of error propagation (see Conradsen, p. 69 [8]), which for the mapping $Y = f(X, Z)$ states the uncertainty for Y around the means $\bar{x}$ and $\bar{z}$, i.e. the mean of Y, as

$$\begin{aligned} s_y^2 =& s_x^2\left(\frac{\partial Y}{\partial X}(\bar{x},\bar{z})\right)^2 + s_z^2\left(\frac{\partial Y}{\partial Z}(\bar{x},\bar{z})\right)^2 \\ &+ 2\cdot s_{xz}\left(\frac{\partial Y}{\partial X}(\bar{x},\bar{z})\right)\left(\frac{\partial Y}{\partial Z}(\bar{x},\bar{z})\right) \end{aligned} \tag{3.40}$$

where s_x^2 is the estimated variance of X and s_{xz} is the covariance between X and Z. For the maximum for a parabola the mapping is $Y = \frac{-X}{2\cdot Z}$, which gives

$$\begin{aligned} s_y^2 =& s_x^2\cdot\left(\frac{-1}{2\cdot\bar{z}}\right)^2 + s_z^2\cdot\left(\frac{\bar{x}}{2\cdot\bar{z}^2}\right)^2 + 2\cdot s_{xz}\cdot\left(\frac{-1}{2\cdot\bar{z}}\right)\left(\frac{\bar{x}}{2\cdot\bar{z}^2}\right) \\ =& s_x^2\cdot\left(\frac{1}{4\cdot\bar{z}^2}\right) + s_z^2\cdot\left(\frac{\bar{x}^2}{4\cdot\bar{z}^4}\right) - 2\cdot s_{xz}\cdot\left(\frac{\bar{x}}{4\cdot\bar{z}^3}\right) \end{aligned} \tag{3.41}$$

The predicted age of maximum with corresponding variances can then be estimated to be (by setting $\bar{x} = \hat{\beta}_{1i}$ and $\bar{z} = \hat{\beta}_2$)

$$\hat{\text{age}}^{\star}_{k=1} = 2.4271 \quad s^2_{\hat{\text{age}}^{\star}_{k=1}} = 0.0227$$
$$\hat{\text{age}}^{\star}_{k=2} = 3.3677 \quad s^2_{\hat{\text{age}}^{\star}_{k=2}} = 0.0184$$
$$\hat{\text{age}}^{\star}_{k=3} = 1.1901 \quad s^2_{\hat{\text{age}}^{\star}_{k=3}} = 0.0693$$

The approximately 95 % confidence intervals for age at maximum rate can be estimated to

$$[2.4271 \pm 1.96\cdot 0.1506] = [2.1319, 2.7223]$$
$$[3.3677 \pm 1.96\cdot 0.1358] = [3.1015, 3.6339]$$
$$[1.1901 \pm 1.96\cdot 0.2632] = [0.6742, 1.7060]$$

Since the estimated mean vector and the covariance-matrix for the multivariate normal distribution desribing the slope in the k'th cluster and the overall curvature is known, simulation can be used to produce the confidence intervals as well ($Z = [X, Y]$). The multivariate gaussian distribution is defined as

$$f(\mathbf{z}) = \frac{1}{2\pi|\Sigma|^{1/2}} e^{-\frac{(\mathbf{z}-\boldsymbol{\mu})w^T \Sigma^{-1}(\mathbf{z}-\boldsymbol{\mu})}{2}} \tag{3.42}$$

where the mean is $\hat{\boldsymbol{\mu}} = [\hat{\beta_{1k}}, \hat{\beta}_2]$ and the estimated covariance matrix is Σ_k, which consists of the variances of the slope and the curvature in the diagonal and the covariance in the two off-diagonal elements. Simulating 100.000 samples for each cluster, computing the maximum age as in (3.6.4) and then finding the 2.5 % and 97.5 % quantiles gives the confidence intervals

$$\begin{gathered}[2.1182, 2.7227] \\ [3.1149, 3.6644] \\ [0.5836, 1.6527]\end{gathered}$$

which is seen to coincide well with the results found using the law of error propagation as expected. It is seen that none of the intervals are close to overlap, which shows that the 3 groups have their maximum-rates at significant different ages. This imply that one of characteristic is a different age of maximum.

The starting points at age = 1 are

$$\begin{aligned}
\hat{\beta}_{01} + \hat{\beta}_{11} + \hat{\beta}_2 &= -4.3006, \quad e^{\hat{\beta}_{01}+\hat{\beta}_{11}+\hat{\beta}_2} = 0.01356 \\
\hat{\beta}_{02} + \hat{\beta}_{12} + \hat{\beta}_2 &= -4.3045, \quad e^{\hat{\beta}_{02}+\hat{\beta}_{12}+\hat{\beta}_2} = 0.01351 \\
\hat{\beta}_{03} + \hat{\beta}_{13} + \hat{\beta}_2 &= -5.1969, \quad e^{\hat{\beta}_{03}+\hat{\beta}_{13}+\hat{\beta}_2} = 0.00553
\end{aligned}$$

where the right part corresponds to the estimated number of symptoms per days in the first year of life for a child with a pd value of 1 and a day-care start at the age of 0 months. It is seen that the middle and high groups start at similar levels and this imply that the difference between the groups is the longitudinal development from this point on. The low group is seen to start lower compared to the two other groups and the low level is seen to be kept throughout the considered interval.

3.6.5 Poisson model for gaussian LCR clusters

Based on the clusters found in section 3.3, a model as in (3.39) can be estimated. The advantage of doing this is that the clusters in the gaussian model were identified different compared to the grouping from the poisson response, i.e. may pick up different symptoms patterns. For the poisson response the majority of the children ends in the cluster with the fewest symptoms, whereas the middle cluster is the largest for the gaussian model. The two strategies lead to different groupings, which can be compared on the same scale by fitting a poisson model for the gaussian clusters. Using the gaussian clusters but on a poisson scale gives a model, which will be much easier to interpret and compare to the original gaussian LCR.

Estimating the model gives the summary in Table 3.31, which shows that the curvature and the effect of the congenital resistance are close to be the same for the three clusters. The curvature for the low cluster is seen to be significant, however for comparison reasons the parameter is omitted. The parameter is either way not highly significant different from the two other clusters.

	estimate	san.se	wald	p	e^{β}
(Intercept)	-5.1250	0.1944	695.2497	<0.0001	0.0059
clusterlow	-0.4139	0.4935	0.7033	0.4017	0.6611
clustermiddle	0.2985	0.2619	1.2982	0.2545	1.3478
age	0.8659	0.1015	72.7766	<0.0001	2.3772
(age^2)	-0.1328	0.0165	64.8695	<0.0001	0.8757
(log10(pd))	-0.1839	0.0570	10.4107	0.0013	0.8321
(daycare.start/30)	-0.0245	0.0070	12.0589	0.0005	0.9758
clusterlow:age	-1.2487	0.3871	10.4036	0.0013	0.2869
clustermiddle:age	-0.5777	0.1687	11.7272	0.0006	0.5612
clusterlow:(age^2)	0.1489	0.0711	4.3847	0.0363	1.1605
clustermiddle:(age^2)	0.0479	0.0309	2.4025	0.1211	1.0491
clusterlow:(log10(pd))	0.1020	0.1054	0.9355	0.3334	1.1074
clustermiddle:(log10(pd))	0.0867	0.0742	1.3645	0.2428	1.0906
clusterlow:(daycare.start/30)	-0.0316	0.0153	4.2632	0.0389	0.9689
clustermiddle:(daycare.start/30)	-0.0218	0.0114	3.6341	0.0566	0.9785

Table 3.31: Summary for GEE model with clusters from gaussian LCR. San.se indicates robust standard errors.

The model shows that the groups with the low and middle level of symptoms have a greater benefit of starting later in day-care with a reduction of 4 % respectively 2 % compared to the group with a high level. The effect of the congenital PD15 PtcO2 is seen to the same for all 3 groups, but indicates that the benefit increases with the level of episodes. The reduction in symptom intensity for an increase in PD15 PtcO2 is 14 %, whereas the effects for the each cluster before removing the interaction between cluster and PD15 PtcO2 were 8 %, 10 % and 17 % for the group with a low, middle and high level of symptoms, respectively.

The results from using the clusters obtained from the gaussian LCR seem more plausible, since one would expect that the children coming from the group with the most symptoms with a high congenital resistance would start at a lower symptom level compared to the rest of the group. It is seen that compared to the results obtained when using a poisson LCR the effect of PD15 PtcO2 has the opposite effect, i.e. the effect of PD15 PtcO2 is highest in the group with the fewest symptoms in the poisson LCR.

	estimate	san.se	wald	p	e^{β}
(Intercept)	-4.9062	0.1748	787.4544	<0.0001	0.0074
clusterlow	-1.2609	0.3180	15.7198	<0.0001	0.2834
clustermiddle	-0.0680	0.1932	0.1239	0.7248	0.9343
age	0.7161	0.0920	60.6285	<0.0001	2.0465
(age^2)	-0.1080	0.0147	54.2153	<0.0001	0.8976
(log10(pd))	-0.1456	0.0377	14.9414	0.0001	0.8645
(daycare.start/30)	-0.0246	0.0071	12.1345	0.0005	0.9757
clusterlow:age	-0.4685	0.1084	18.6889	<0.0001	0.6260
clustermiddle:age	-0.3075	0.0491	39.2153	<0.0001	0.7353
clusterlow:(daycare.start/30)	-0.0319	0.0151	4.4334	0.0352	0.9686
clustermiddle:(daycare.start/30)	-0.0220	0.0116	3.5880	0.0582	0.9783

Table 3.32: Summary for updated GEE model with clusters from gaussian LCR

3.6.6 Diagnostics

The model having poisson response and clusters based on the gaussian LCR grouping has an estimated scale-parameter of $\hat{\phi} = 1.88$, which is seen to be higher than the model based on poisson grouping. Deviance residuals for the model are shown in a QQ-plot in Figure 3.35, which shows that there seems to be a little too many large positive residuals. The high residuals are seen to be related to the age of 2 and 5 years and the middle cluster.

	high	low	middle	high	low	middle	high	low	middle
$\lvert\text{Residuals}\rvert > x$		2.57			3.00			4.00	
1	3.00	1.00	3.00	2.00	0.00	2.00	0.00	0.00	0.00
2	4.00	1.00	7.00	2.00	0.00	4.00	0.00	0.00	1.00
3	3.00	0.00	3.00	0.00	0.00	0.00	0.00	0.00	0.00
4	1.00	1.00	3.00	0.00	0.00	1.00	0.00	0.00	0.00
5	2.00	3.00	9.00	0.00	1.00	7.00	0.00	0.00	1.00

Table 3.33: Distribution of residuals larger than x on group and age for two different

The residual analysis in Table 3.33 shows that observations in the middle group are causing most of the high residuals (> 3), whereas the low group have a low number of high residuals and is furthermore seen to have a smaller variance than the other groups. The large positive residuals correspond to children having a too low estimated symptom intensity, since the deviance residuals are defined as

$$r_D = \text{sign}(y_i - \hat{\mu}_i)\sqrt{w_i d(y_i, \hat{\mu}_i)} \tag{3.43}$$

where $d(y_i, \hat{\mu}_i)$ is the unit deviance of observation i and w_i is the weight of observation

i (Wood p. 74 [42]). Tthe link between the deviance and the unit-deviance is

$$D = \sum_i w_i d(y_i, \hat{\mu}_i) \tag{3.44}$$

The sign-part of the definition of the deviance residuals shows that a positive residual corresponds to $y_i > \hat{\mu}_i$, which imply that the observed intensity is higher than the fitted. This can occur if eg an individual from the middle group at some time-point have too many symptoms and hence approaching the high group. A Bonferroni type of test shows that a residual with a value of 4 gives an adjusted p-value of around 9 %, which shows that the residuals are not too bad.

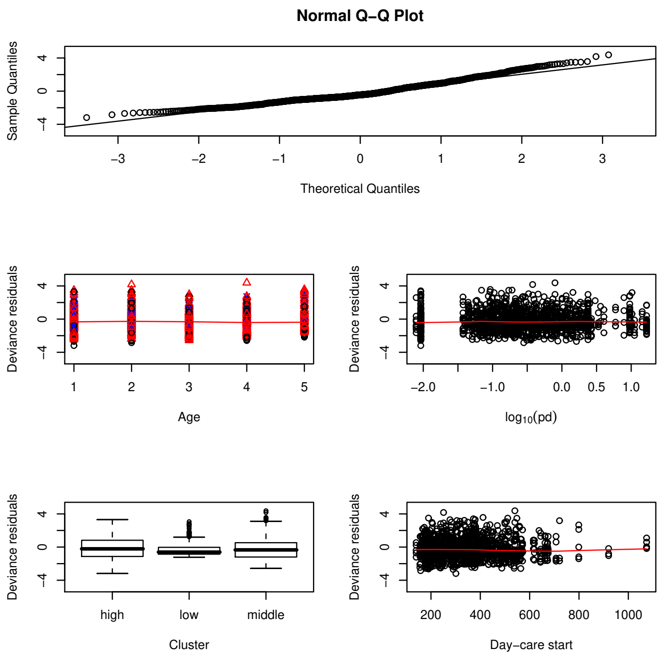

Figure 3.35: QQ-plot for deviance residuals along with deviance residuals against the explaining the variables. In deviance vs. age the observations are colored according to their group: black (circle) = high, blue (x) = low group, red (triangle) = middle group

3.7 Comparing LCR models

In Figure 3.36, a comparison between the three methods based on LCR is shown. It is seen that the poisson and gaussian arcsine-root methods based on gaussian LCR clustering are seen to be close to equal for the low and high group, whereas the middle group has a higher poisson model. The model based on the poisson LCR clustering is seen to give predictions some what higher than the two other models, but is not far away either. It is seen that the estimated curves from the three methods are close to parallel and the differences are small.

It is seen that the data can be described almost equally good with a poisson response model based on the gaussian LCR grouping as for the gaussian response. The advantage of the poisson model is that the effects become easier to interpret, since the effects are multiplicative on the response scale. Hence if the grouping based on the gaussian LCR is proven to be the best grouping, the groups can be used as grouping variable in a model with poisson response. This gives almost the same model, however the effects are multiplicative and hence easier to interpret. With respect to the deviance the poisson model based on gaussian clustering is poorer model compared to the poisson model from the poisson clustering. This have to be the case, since latent class regression is optimized on that scale.

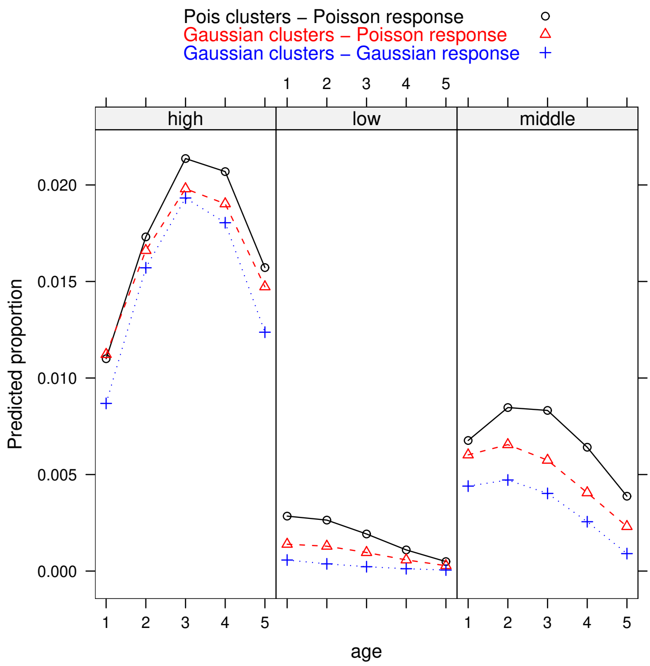

Figure 3.36: Comparison between model obtained for poisson model when clusters are determined in a gaussian LCR and poisson LCR, respectively, and with the gaussian response model determined by gaussian LCR

3.8 Discussion

In the previous sections the yearly aggregated symptoms were analyzed by different statistical methods. The methods lead to more or less the same results, which will be summarized in the following.

The first method considered was a mixed effects model (MM) based on the relative number of episodes, $n_{\text{episodes}}/n_{\text{days at risk}}$. The model showed that the children were heterogeneous and thus needed to be modelled with individual longitudinal developments. The analysis showed that the congenital responsiveness, PD15 PtcO2, and the age at day care start had significant influence on the level of symptoms. The congenital responsiveness was seen to give reductions for age 1-4 of 20-35 %, whereas the age of 5 years showed reductions of 40-60 % for a factor 10 increase in PD15 PtcO2. For day care start reductions between 5 and 10 % for an increase of 1 month were seen for the age of 1-4 years and 5-50 % for the age of 5 years.

The MM showed that the longitudinal development could be divided in different groups according to the pattern of the longitudinal development: No symptoms at all, some initial symptoms and then a decline, some symptoms at all ages, many symptoms at all ages, many symptoms initially and increasing level. The optimal way to identify subgroups or clusters of children was however not clear.

With the basis in MM, latent class regression (LCR) was applied to find groups of children having similar parametric characteristics. This led to 3 subgroups of children: A group with no symptoms, a group with some symptoms initially and then a decline to fewer symptoms and a group with many symptoms with an increase to the age of 3 and then a decline back to around, but above, the starting level. 50 children were assigned to the high group, 171 to the middle and 86 to the low group.

A general linear model based on the grouping obtained from the LCR showed that the individuals in the group with the fewest symptoms benefited the least from having a high congenital PD15 PtcO2 and the group with the most symptoms benefited the most. The effect of day start was seen to give the highest reductions for individuals in the middle group and the high group.

It was furthermore shown that the LCR was able to identify the group for each child with reasonable accuracy at the age of 3 years. An early classification is interesting, since this may give the possibility for more direct treatment of the children with many symptoms. Obviously borderline cases with children having high probabilities of belonging to more than one group will occur, which will involve some level of medical judgment or additional data to give a better estimate of the group (a wait and see strategy).

Since the analysis of the symptoms in the MM and LCR were based on the arcsine-root transformations, interpretation of the effects was seen to be troublesome. The effects were neither additive nor multiplicative on the untransformed scale. This implied that the effect of eg PD15 PtcO2 should be evaluated conditional on the value of the other variables in the model and the current level of PD15 PtcO2. To avoid the awkward interpretation, a non-gaussian approach was analyzed. The method modelled the relative number of symptoms with a generalized linear model based on the assumption of poisson distributed symptom intensity.

On the poisson scale both the congenital PD15 PtcO2 and the day-care start were seen to be significant in a generalized linear mixed effects model. A factor 10 increase in PD15 PtcO2 gave a reduction of the intensity by 21 %, whereas starting 1 month later in day-care gave a reduction of 3 % for an individual. The longitudinal development was seen to lead to the same type of curve as the mixed effects model based on arcsine-root transformed values. The difference between the model with gaussian response and the model with poisson response was that the latter model is much easier to interpret, since the effects became multiplicative on the response scale.

As for the gaussian response LCR was applied to find latent groups in the population. The analysis showed that the optimal number of groups was 5, however since the decision criterion for 3 clusters was low as well and mostly for comparison reasons, the 3 cluster model was analyzed more thoroughly. This led to similar longitudinal development types, but with a different number of children in each cluster. Most children were now assigned to the group with the fewest symptoms (increased from 86 to 156), the group with the most symptoms was reduced from 59 children to 55 and finally the middle group was reduced from 171 to 105 children. It was seen that the size of the high group was to being unaffected by the change of scale.

The LCR with poisson response showed that the benefit of a high congenital PD15 PtcO2 was the same for all clusters, whereas the benefit of starting late in day-care was largest in the group with the fewest symptoms. The effect of the congenital PD15 PtcO2 was seen to be a reduction of 26 % for an increase in PD15 PtcO2 with a factor 10, whereas the day-care start was seen to give reductions of 3-7 % for increases in age at day-care start by 1 month. The highest reductions were seen for the groups with the fewest symptoms.

The clusters obtained from the gaussian LCR were furthermore used as grouping variable in a GEE-model with poisson response, which showed that the effect of the congenital PD15 PtcO2 was the same for all clusters and gave reductions of 14 % for a factor 10 increase. The model indicated that if an effect should be present the group with the most symptoms would benefit the most and the group with the fewest symptoms the least. The parameters for the curvature were seen to be the same for all groups but the slopes were significantly different.

Since the cluster-difference of the PD15 PtcO2 was seen to be insignificant for 2 out of 3 models, it may be concluded that the difference probably is insignificant. The two models based on gaussian LCR grouping both indicated that the group with the most symptoms benefited the most of having a large PD15 PtcO2, which seemed most plausible, whereas the poisson clusters indicated that the group with the fewest symptoms benefited the most, although a trend was indicated.

Finally comparing the gaussian and poisson models showed that the differences were small. The gaussian model classified most children to the middle group, whereas the poisson model classified most to the low group. It was furthermore seen that the high group had almost the same size for the two different models. The poisson model had obviously advantages over the gaussian model, since the effects were multiplicative on the response scale, which gave a nice interpretation. In Chapter 4 the two model are compared with respect to the asthma-diagnoses, which may lead to the conclusion that one of the method is to be preferred in this perspective.

Chapter 4

Asthma diagnoses

Contents

4.1 Introduction

In the following chapter the groups found in the LCR having gaussian and poisson response, respectively, are compared to the diagnosis established by the COPSAC-group [4]. The asthma-diagnosis is given if the child went to the 5 year visit, has been treated with Bricanyl (Terbutalin)[1] in the fifth year of life, has been treated with Spirocort[2] in the fifth year of life, has been having symptoms in the fifth year of life and therefore the symptom-diary for the fifth year of life is also required.

The children are diagnosed as either having asthma, not having asthma or not having a diagnosis due to lack of information. In the LCR modelling three subgroups were found, namely the low (no or very few symptoms), the middle (few and a decreasing number of symptoms) and the high group (a consistenly high level of symptoms, which peaks around the age of 3 years). Comparing the groups with the diagnoses will possibly give two subgroups of children in the middle group, which are diagnosed as either non-asthmatic or asthmatic. However if either one of the groups are small or empty the middle group may be interpreted as the asthmatic or non-asthmatic group depending on which group is empty. This is under the assumption that the high group and low groups correspond to the asthmatic and non-asthmatic groups, respectively.

The comparison between the two classifications can be summarized in a cross-tabulation table of dimension 3x2 (or 2x2 if the middle group is neglected or merged with either the low or the high group). These types of tables are usually analyzed by means of the two concepts: Sensitivity and specificity (see eg Kirkwood [22]). Sensitivity is defined as the proportion of true positives (asthma) correctly classified, which equals the proportion of the doctor-diagnosed asthma children classified as belonging to the high group. Specificity is the proportion of true negatives (non-asthma) classified correctly, hence the proportion of doctor-diagnosed non-asthma children classified as belonging to the low group. Finally the overall agreement (accuracy) can be used as a combined measure of the accordance of the different grouping methods and the diagnoses.

		Observed		
		Non-asthma	Asthma	Total
Predicted	Non-asthma	a	b	r_1
	Asthma	c	d	r_2
Total		c_1	c_2	n

Table 4.1: Contigency table for predicted asthma status vs. observed (doctor-diagnosed)

In Table 4.1 the possible outcomes of comparing predicted vs. observed/doctor-

[1] a short acting β_2-agonist, relaxes the bronchial smooth muscles, see p. 112 in "Kompendium i Farmakologi", Christophersen et al. [7]

[2] an inhalation steroid (budesonid) for anti inflammatory treatment of asthma [7]

diagnosed asthma are summarized. The table covers a 2x2 situation, in a 3x2 situation an additional row will occur. In the 2x2 table a number of measures are interesting as mentioned above, which will be defined in the following. The sensitivity is given as d/c_2, the specificities likewise as a/c_1 and the overall agreement as $(a + d)/n$.

For the doctor-diagnosed classification 47 children are classified as having asthma, 249 children as not having asthma and 115 children have no classification. For the grouping based on a gaussian LCR the distribution on low, middle and high is 86, 171 and 59 children, whereas the grouping based on poisson LCR gives 156, 105 and 55 children, respectively. A total of 245 children have both a diagnosis and a group from the LCR analysis.

4.2 Validation on all five years

Table 4.2 shows the cross-tabulations between the doctor-diagnosed asthma-status and the groups from gaussian and poisson LCR, respectively. It is seen that the sensitivities for the two methods are 78 % and 67 % for the gaussian and poisson grouping, respectively. The specificities are 34 % and 60 % and the overall agreements are 42 % and 61 %. It is seen that the gaussian approach classify more true positives, but fewer true negatives compared to the poisson model. Furthermore both methods classify a small number (but close to being the same) of non-asthmatic children in the high, but no asthmatic children as belonging to the low group. It is seen that the low group in both models only contains non-asthmatic children.

	Gaussian		Poisson	
LCR-group	Non-asthma	Asthma	Non-asthma	Asthma
Low	69	0	120	0
Middle	114	10	64	15
High	17	35	16	30

Table 4.2: Cross-tables of subgroups from gaussian LCR vs. doctor diagnosed asthma-status and poisson LCR vs. doctor diagnosed asthma-status, respectively

It can be argued that classifying a non-asthmatic child in the middle group is not really a misclassification, since for both methods the majority of the middle group is non-asthmatic children, 92 % and 81 % for the gaussian and poisson model, respectively. Furthermore, the middle and low group are seen to be similar, in particularly at the age of 5 years. Merging the two groups, low and middle, gives updated specificities of 91 % and 92 %, whereas the sensitivities stays the same. The overall agreements are 89 % and 87 %, hence more than doubled for the gaussian model.

Classifying the middle group as belonging to the high group (asthmatic group) gives sensitivities of 100 % and 100 % and specificities of 34 % and 60 %. The overall agreements are 47 % and 67 %. It follows from the two ways of classifying the middle group that assigning the middle group to the low group improves the specificity significantly, whereas assigning it to the high group improves the sensitivity to perfect agreement. However the overall agreement is best when merging the low and middle group and these groups are also seen to be similar at the age of 5 years.

Neglecting the children classified as belonging to the middle group gives sensitivities of 100 %, 100 % and specificities of 80 % and 88 %. However this implies that a large number of children is left out (124 and 79), hence either using the middle group as an unique group or as a part of the low group seems more adequate.

The methods are seen not to give any false-negatives, when the middle group is kept as a unique group or merged to the high group. This is typically wanted, since it is often better to maintain the suspicion of an illness rather than falsely claiming the child to be healthy. However even with the middle and low group merged the sensitivity and specificity are high. The grouping based on the poisson response is seen to find the most non-asthmatic children, whereas the gaussian grouping finds the most asthmatic children. The poisson model is additionally seen to give a smaller middle group, hence giving a more clear initial diagnosis before merging.

For the grouping with the low and middle group merged together as the non-asthmatic group, the predicted positive value [40] is now calculated. The predicted positive value is the percentage of the children classified in the LCR as having asthma, who also have a doctor-diagnosed asthma-diagnosis (d/r_2 in Table 4.1). This imply that the positive predicted values become 35/52 % = 67% and 65 % for the gaussian and poisson model, respectively. The negative predictive values (a/r_1) are likewise $183/193 = 95\%$ and 92 % for the gaussian and poisson model, respectively. It is seen that the gaussian model is performing better in terms of both the positive predictive value and the negative, which is seen in the overall agreements as well. For the grouping with the middle and high group merged, the negative predictive values are 100 % for both the gaussian and the poisson model, whereas the positive predictive values are 26% and 36 %. The latter imply that this grouping is rather poor, since most of the positive indeed are false.

In Figure 4.1 boxplots of the posterior probabilities are shown. It is seen that the asthmatic group has low posterior probabilities of belonging to the low group for both the gaussian and poisson model. The middle group is seen mostly to be related to the non-asthmatic group for the gaussian model, whereas the poisson model in median has a higher posterior probability of the middle group. The high group is likewise seen to be related to the asthmatic group, however it is seen that some of the children in the non-asthmatic group have high posteriors for belonging to the high group. The latter is also seen from Table 4.2, where the high group is seen to consists of around 33 % non-asthmatic children. It is seen that in both models the asthmatic children are very unlikely to belong to the low group (posteriors are zero), whereas for the other groups the posteriors for asthmatic and non-asthmatic children become less separated.

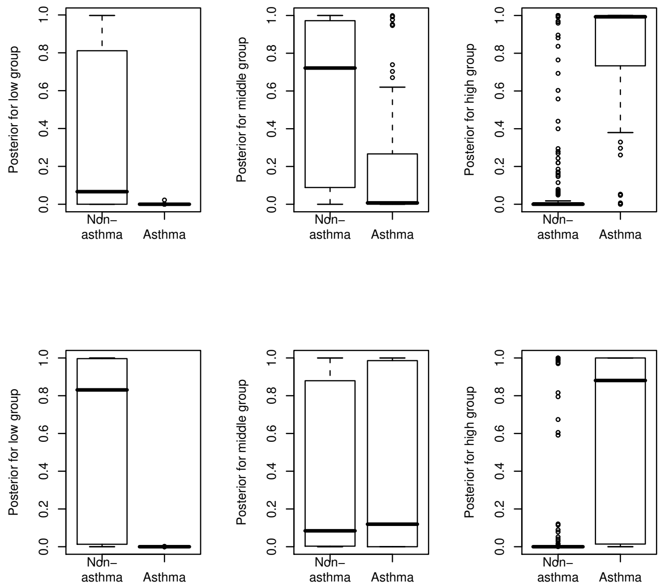

Figure 4.1: Top: Boxplots for posterior probabilities for low (left), middle and high group (right) in the gaussian model, grouped by the asthma-diagnosis. Bottom: Boxplots for posterior probabilities for low (left), middle and high group (right) in the poisson model, grouped by the asthma-diagnosis

4.3 Yearly classifications

As seen in section 3.3.7 the models give the possibility to estimate the posterior probabilities based on a subset of the data, eg the second year of life (local information) or the first two years of life (cumulated information). This imply that the diagnoses provided by the COPSAC-group can be evaluated on subsets of the diaries, which again imply that the models may give an earlier predicted diagnosis or at least some indication of how the diagnosis at the age of 5 years and the model for the symptoms at previous age are related.

The basis for the posterior probabilities is a parametric model and the prior probabilities of being in the low, middle and high group, respectively. Each observation has a probability of belonging to each of the three groups [23], which is given as

$$\hat{p}_{ijk} = \frac{\pi_k f(y_{ijk}|x_{ijk}, \theta_k)}{\sum_{k'=1}^{3} \pi_{k'} f(y_{ijk'}|x_{ijk'}, \theta_{k'})} \tag{4.1}$$

where $f(y_{ijk}|x_{ijk},\theta_k)$ is the density for the observation in group k and π_k is the prior probability of group k. The probabilities in (4.1) are seen to vary within each individuals observations, which imply that they reflect severity of the individual year in terms of symptoms.

Instead of using only the local information at the age of j, the past information from the age of 1 years to the age of j years can be included to estimate the probabilities at the age of j years. The posteriors are found by the joint probabilities [23], given by

$$\hat{p}_{ik}(j') = \frac{\pi_k \prod_{j=1}^{j'} f(y_{ij}|x_{ij},\theta_k)}{\sum_{k'=1}^{K} \pi_{k'} \prod_{j=1}^{j'} f(y_{ij}|x_{ij},\theta_{k'})} \tag{4.2}$$

The probabilities in (4.1) gives yearly classifications (local), whereas the probabilities in (4.2) gives classifications based on cumulated information, namely the episode-rates in the first j' years of life. The probability in (4.2) is an approximation, since the joint probability is the product of the marginal probabilities only if the observations are uncorrelated. However as shown in Chapter 3, the correlation is seen to be low.

Density-functions for the two types of response are needed in order to calculate the posterior probabilities. The density-functions need two parameters for each observation, namely the observation and the predicted response ($\hat{\mu}$). This holds for both the model with gaussian response and the model with poisson response. With the observation, y (arcsin($\sqrt{n_{\text{episode}}/n_{\text{days}}}$), and n_{episode} for the gaussian and poisson case, respectively), and the predicted response $\hat{\mu}$ the likelihood can be calculated as

$$f(y_{ijk}|\hat{\mu}_{ijk})_{gaus} = \frac{1}{\sqrt{2\pi\hat{\sigma}_k^2}} \cdot e^{-\frac{(y_{ijk}-\hat{\mu}_{ijk})^2}{2\cdot\hat{\sigma}_k^2}} \tag{4.3}$$

$$f(y_{ijk}|\hat{\mu}_{ijk})_{pois} = \frac{\hat{\mu}_{ijk}^{y_{ijk}} \cdot e^{-\hat{\mu}_{ijk}}}{y_{ijk}!} \tag{4.4}$$

where $\hat{\sigma}_k^2$ is the residual variance for the k'th group. The analysis in section 3.3.4 showed that the variances were different over clusters in the gaussian model, the low group had a smaller residual variance. The predicted mean values are described in section 3.3.4 and 3.6.4.

Three types of groupings are now considered, 1: The middle group is considered as a special group (Table 4.3), 2: Children in the middle group are assigned to the low group (Table 4.4) and 3: The children in the middle group are assigned to the high group (Table 4.5). For each of these three groupings 4x5 sensitivities, specificities and overall agreements are computed corresponding to 2 grouping methods (gaussian and poisson), 2 types of information (local and cumulated) and 5 different ages.

4.3.1 No merging

Table 4.3 shows that the overall agreement is seen to be highest for the grouping based on the poisson LCR with a maximum of 67 % for the local information at the age of 4 years. The sensitivity is clearly highest for the gaussian grouping, whereas the specificity is highest for poisson response. Grouping based on a poisson LCR gives a markedly better specificity for the cost of a little lower sensitivity, which imply that the overall agreement becomes higher for this method.

Group	Type	Age 1	Age 2	Age 3	Age 4	Age 5
Gaussian	Local	n: 242 sen: 7 % spe: 36 % all: 31 %	n: 237 sen: 56 % spe: 29 % all: 34 %	n: 235 sen: 56 % spe: 43 % all: 45 %	n: 233 sen: 64 % spe: 48 % all: 51 %	n: 221 sen: 77 % spe: 19 % all: 31 %
	Cumulated	n: 242 sen: 7 % spe: 36 % all: 31 %	n: 237 sen: 60 % spe: 32 % all: 38 %	n: 235 sen: 71 % spe: 33 % all: 40 %	n: 233 sen: 73 % spe: 32 % all: 40 %	n: 221 sen: 77 % spe: 35 % all: 43 %
Poisson	Local	n: 242 sen: 20 % spe: 71 % all: 62 %	n: 237 sen: 53 % spe: 55 % all: 54 %	n: 235 sen: 44 % spe: 68 % all: 63 %	n: 233 sen: 56 % spe: 69 % all: 67 %	n: 221 sen: 61 % spe: 67 % all: 66 %
	Cumulated	n: 242 sen: 20 % spe: 71 % all: 62 %	n: 237 sen: 56 % spe: 57 % all: 57 %	n: 235 sen: 58 % spe: 58 % all: 58 %	n: 233 sen: 58 % spe: 60 % all: 60 %	n: 221 sen: 66 % spe: 59 % all: 60 %

Table 4.3: **n**umber of observations, **sen**sitivity, **spe**citivity and overall agreement for the gaussian and poisson grouping, local and cumulated information at different ages. The cumulated information corresponds to the information from the age of 1 year to the age in the column, whereas the local information correspond to using only the observation corresponding to the age in the column. An assignment of a child in the middle group is a mis-classification for both the asthmatic and non-asthmatic group.

It is seen that the difference between using local and cumulated information is moderate. The cumulated information strategy tends to give higher sensitivities but lower specificities compared to the local information strategy. The specificities are seen to be constant or decreasing as the children get older, whereas the sensitivities are increasing, which is particularly apparent from the age of 1 to the age of 2.

4.3.2 Middle and low group merged

With the middle group classified as belonging to the low group, the specificities for the two grouping methods increase and become close to identical (Table 4.4). The overall agreements are increased compared to having the middle group as an unique group, due to the increase in the specificity on the same sensitivity. Obviously the merging of the low and middle group greatly improves the specificity for the gaussian grouping, since the middle group is the largest group of the 3 and mostly consists of non-asthmatic children. The poisson grouping is however also seen to be improved significantly with increases by 20-30 %-points.

Group	Type	Age 1	Age 2	Age 3	Age 4	Age 5
Gaussian	Local	n: 242 sen: 7 % spe: 97 % all: 81 %	n: 237 sen: 56 % spe: 89 % all: 82 %	n: 235 sen: 56 % spe: 92 % all: 85 %	n: 233 sen: 64 % spe: 94 % all: 88 %	n: 221 sen: 77 % spe: 92 % all: 89 %
	Cumulated	n: 242 sen: 7 % spe: 97 % all: 81 %	n: 237 sen: 60 % spe: 88 % all: 83 %	n: 235 sen: 71 % spe: 89 % all: 86 %	n: 233 sen: 73 % spe: 90 % all: 87 %	n: 221 sen: 77 % spe: 91 % all: 88 %
Poisson	Local	n: 242 sen: 20 % spe: 92 % all: 79 %	n: 237 sen: 53 % spe: 89 % all: 82 %	n: 235 sen: 44 % spe: 93 % all: 84 %	n: 233 sen: 56 % spe: 95 % all: 88 %	n: 221 sen: 61 % spe: 93 % all: 86 %
	Cumulated	n: 242 sen: 20 % spe: 92 % all: 79 %	n: 237 sen: 56 % spe: 90 % all: 83 %	n: 235 sen: 58 % spe: 91 % all: 84 %	n: 233 sen: 58 % spe: 92 % all: 85 %	n: 221 sen: 66 % spe: 92 % all: 86 %

Table 4.4: **n**umber of observations, **sen**sitivity, **spe**citivity and over**all** agreement for the gaussian and poisson grouping, local and cumulated information at different ages when classifying the middle group as being low.

Table 4.4 shows, that the highest overall agreements are found at the age of 4 and 5 and that the local information is just as good as the cumulated information in terms of total agreement. The cumulated information strategy tends to give a higher sensitivity at the age of 4-5 years, but a lower specificity compared to the local information. It is seen that the specificity is high at the age of 1 and keeps a high level, whereas the sensitivity is improving as the children get older (same as for the analysis with no groups merged).

4.3.3 Middle and high group merged

Table 4.5 shows that merging the middle group with the high group gives a poorer overall agreement compared to merging the middle and the low groups. The reduced overall agreement is a results of an increase in sensitivity (now close to 100 % at the age of 2-5 years for the gaussian grouping and at the age of 5 years for the poisson grouping) and a severe reduction in the specificity from above 90 % to below 50 % for the gaussian grouping and below 70 % for the poisson grouping.

The age related patterns for the specificity and sensitivity are seen to be the same for the merging of the middle and high groups as seen for the analysis with no merging and the analysis with the middle and low groups merged. The main difference is that sensitivities start at higher levels, whereas the specificities are the same as for the analysis without merging of groups.

4.3.4 2 cluster model

In previous analysis it was shown that collapsing the two lowest groups gives the best ability to predict the diagnosis correct. It is therefore interesting to analyze the 2

Group	Type	Age 1	Age 2	Age 3	Age 4	Age 5
Gaussian	Local	n: 242 sen: 86 % spe: 36 % all: 45 %	n: 237 sen: 98 % spe: 29 % all: 42 %	n: 235 sen: 96 % spe: 43 % all: 53 %	n: 233 sen: 98 % spe: 48 % all: 58 %	n: 221 sen: 100 % spe: 19 % all: 35 %
	Cumulated	n: 242 sen: 86 % spe: 36 % all: 45 %	n: 237 sen: 96 % spe: 32 % all: 44 %	n: 235 sen: 96 % spe: 33 % all: 45 %	n: 233 sen: 98 % spe: 32 % all: 45 %	n: 221 sen: 100 % spe: 35 % all: 48 %
Poisson	Local	n: 242 sen: 45 % spe: 71 % all: 66 %	n: 237 sen: 84 % spe: 55 % all: 60 %	n: 235 sen: 82 % spe: 68 % all: 71 %	n: 233 sen: 87 % spe: 69 % all: 73 %	n: 221 sen: 98 % spe: 67 % all: 73 %
	Cumulated	n: 242 sen: 45 % spe: 71 % all: 66 %	n: 237 sen: 80 % spe: 57 % all: 62 %	n: 235 sen: 87 % spe: 58 % all: 63 %	n: 233 sen: 91 % spe: 60 % all: 66 %	n: 221 sen: 100 % spe: 59 % all: 67 %

Table 4.5: **n**umber of observations, **sen**sitivity, **spe**citivity and over**all** agreement for the gaussian and poisson grouping, local and cumulated information at different ages when classifying the middle group as being high.

cluster model with respect to sensitivity, specificity and overall agreement. The two cluster model was deselected by means of the bayesian information criteria, which imply that an optimal model was not found with respect to the ability to diagnose the children correct, but to describe the symptoms at different ages. It can however be the case that the 2 cluster model is better to predict the asthma-diagnosis.

In the following the 2 cluster models are considered, namely one with a gaussian response for the arcsine-root transformed values and one with poisson response. A comparison of the predictions from the model is shown in Figure 4.2, which shows that the models are very similar, eg the two lines for the low groups are parallel. The poisson model gives higher predicted rates compared to the gaussian model, which was seen to be the case in the 3 cluster models as well (Figure 3.36).

Since only two groups are present and it therefore is relatively easy to identify which group should be the asthmatic, an analysis of the performance in terms of sensitivity, specificity and overall agreement can be done. In the previous analysis of the LCR children were assigned to the most likely group, which for the two cluster model imply that individual i is assigned to group j if $p_{ij} > 0.5$. It is however simple in the two cluster group to evaluate upon this criteria, i.e. trying different threshold values for the posterior probabilities for the high group.

In Figure 4.3 the sensitivity, specificity and overall agreement for the two LCR's is plotted as function of the cut-point for the posterior probability of the high group, i.e. an individual is classified as belonging to the high group if the posterior $p_{i,\text{high}}$ is larger than p_{cut}. The figure shows that the sensitivity at most is 60 % for $p_{\text{cut}} > 0$, which is seen to be rather bad compared to the 3 cluster model. The specificity and the overall agreement are increasing as the high group is forced to be more and more unlikely, which is caused by the over-weight of non-asthmatic children, i.e. a 200 to 45 ratio. It is seen that the two cluster models are rather bad at finding the

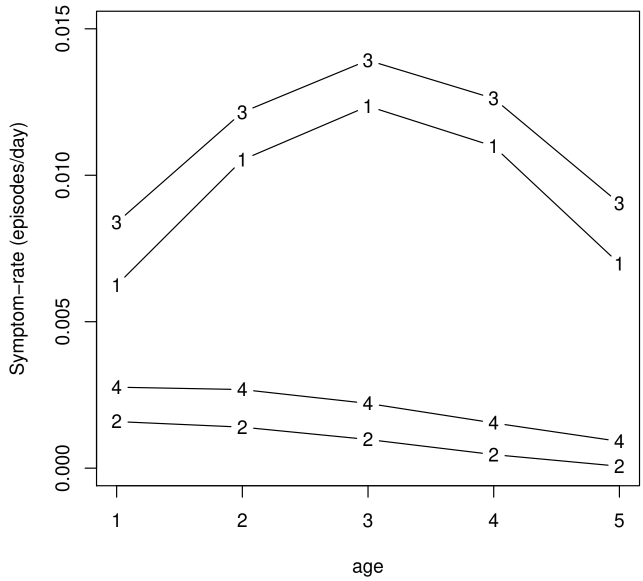

Figure 4.2: 2 cluster model prediction for gaussian and poisson model. Line 1: High cluster for gaussian LCR, 2: Low cluster for gaussian LCR, 3: High cluster for poisson LCR and 4: Low cluster for poisson LCR

asthmatic children compared to the three cluster models, where the sensitivity was 77 % for the best gaussian model. So maximizing the overall-agreement implies that all children are classified as non-asthmatic. The figure also shows that the poisson model performs better compared to the gaussian model. As such the models are seen to be poorer compared to the three cluster models, which imply that the three cluster model do predict the asthma diagnosis better.

Cutpoint in 3 cluster model

As for the two cluster models a cut-point analysis can be done for the three cluster models. This can be done by considering the posterior probability for the high group, since the best performance was seen when collapsing the two other groups. Using the same classification-method as for two cluster analysis, i.e. an individual is classified as belonging to the high group if $p_{i,\text{high}} > p_{\text{cut}}$, gives the ability to do the same analysis as for the 2 cluster model.

In Figure 4.4 the sensitivity, specificity and overall agreement for the gaussian and poisson three cluster models are shown. It is seen for the gaussian model that a

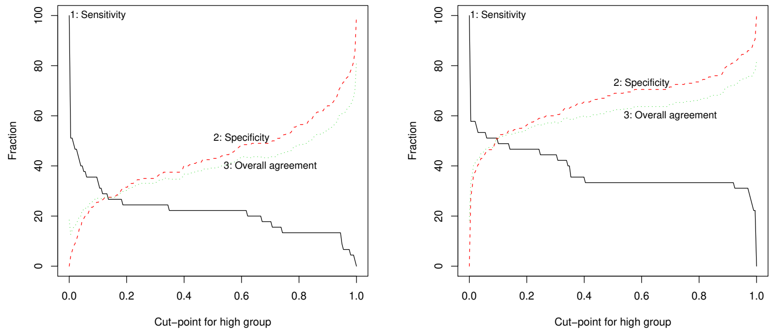

Figure 4.3: Sensitivity, specificity and overall agreement for gaussian two cluster model (left) and poisson two cluster model (right) for various cut-points for the high group. A cut-point, p_{cut}, imply that an individual is classified as belonging to the high group if $p_{i,\text{high}} > p_{\text{cut}}$. In both plots 1 corresponds to sensitivity, 2 to specificity and 3 to overall agreement.

reasonable good model can be obtained for a cut-off value at 0.26, which gives a sensitivity of 87 %, a specificity of 89 % and an overall agreement of 89 % compared to 77 %, 91 % and 86 % for classifying to the most likely group and then merge the low and middle group. The low cut-point may be a result of the weighing with the prior probabilities, where the middle group a priori is twice as likely compared to the low and high group for the gaussian model.

For the poisson model the sensitivity is quickly reduced to 67 %, which is the same value as for the analysis in Table 4.4. If a significant improvement in the sensitivity is wanted the cut-point should be around 0.01, which leads to a sensitivity of 78 %, a specificity of 83 % and an overall agreement of 82 %. This should be compared to 66 %, 92 % and 86 % for the sensitivity, specificity and overall agreement in Table 4.4.

In Figure 4.4 a cut-point analysis is shown for the cumulated posterior for the first two and three years of life, respectively for the high group in the gaussian model. It is seen that with 3 years of information a sensitivity, specificity and overall agreement of around 80 % at the same time is obtainable for a threshold value around 0.22. For information contained in the first two years of life, the performance is poorer compared to the first three years, the sensitivity falls to 60 for an overall agreement of 80 %.

4.4 Existing literature

In section 3.3.5 the results from the gaussian 4 cluster LCR was compared to the results from Martinez et al. [26]. The comparison showed that the agreement was

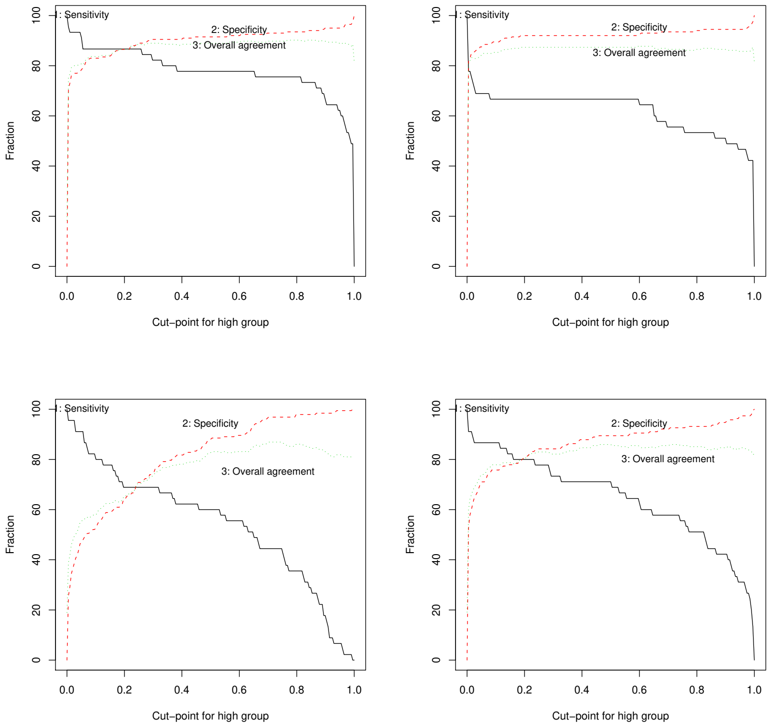

Figure 4.4: Sensitivity, specificity and overall agreement for gaussian three cluster model (upper left), poisson three cluster model (upper right) and for gaussian model for cumulated information until the age of 2 (bottom left) and 3 years (bottom right) for various cut-points for the high group. A cut-point, p_{cut}, imply that an individual is classified as belonging to the high group if $p_{i,\text{high}} > p_{\text{cut}}$. In both plots 1 corresponds to sensitivity, 2 to specificity and 3 to overall agreement.

low, i.e. the group corresponding to late onset wheezing was missing. In the following the age of 3 years and the age of 5 years are considered to approximate the procedure suggested by Martinez et al. [26], who considered the age of 3 and 6 years, respectively, and mainly the types of changes in the wheezing status from the age of 3 years to the age of 6 years.

From the gaussian and the poisson model groups at the age of 3 and 5 years can be estimated either as a local information or as a cumulated information as shown in section 4.2 and 4.3. Cross-tabulations over the group-assignments at the age of 3 and 5 years are shown in Table 4.6, which shows that the cumulated grouping has fewer shifts from the age of 3 to the age of 5 compared to the local information. This is however expected, since the cumulated information at the age of 3 years is used at

			Age 5					
			Local			Cumulated		
	Method	Group	Low	Middle	High	Low	Middle	High
Age 3	Gaussian	Low	17	54	6	54	8	0
		Middle	17	85	16	12	104	7
		High	1	14	26	0	8	43
	Poisson	Low	97	25	13	97	19	1
		Middle	34	25	8	19	50	6
		High	2	13	19	0	8	36

Table 4.6: Cross-tabulation of grouping from LCR between the age of 3 years (rows) and the age of 5 years (columns). The numbers correspond to the number of children for a given combination

the age of 5 as well.

Denoting the low and the middle group the non-asthmatics and the high group the asthmatics, a table for non-wheezers, transient early wheezers, late-onset wheezers and persistent wheezers can be made as seen in Martinez et al [26]. The result is seen in Table 4.7 and comparing the 4 proportions can be done by a $2 \times c$ χ^2-test as described at p. 93 in Kirkwood [22] with the following test-quantity

$$\chi^2 = \frac{N^2 \cdot \left(\sum_{i=1}^{c} R_{1i}^2/n_i - \left(\sum_{i=1}^{c} R_{1i} \right)^2 /N \right)}{\left(\sum_{i=1}^{c} R_{1i} \right) \left(N - \sum_{i=1}^{c} R_{1i} \right)} \tag{4.5}$$

and is χ^2 distributed with c-1 degrees of freedom. R_{1i} corresponds to the size of the i'th group size in the LCR grouping, n_i to the sum of the size of i'th group in the LCR and the size of the i'th group in Martinez et al. and N to the total number of children in the groups, i.e. the sum of the n_i's. The null-hypothesis is that the two sample distributions come from the same distribution. The tests show that it is highly unlikely to believe that the proportions come from the same distribution.

Group	Method	No	Transient	Late	Persistent	$P(>\chi^2)$
Gaussian	Local	73	6	9	11	<0.0001
	Cumulated	75	3	3	18	<0.0001
Poisson	Local	77	6	9	8	<0.0001
	Cumulated	78	3	3	15	<0.0001
Martinez et al		52	20	15	14	

Table 4.7: Asthma status shift (in %) based on LCR. In the last row the corresponding distribution for Martinez et al. [26] is shown and in the last column χ^2 goodness-of-fit test for LCR shift vs. Martinez et al.

The study by Martinez et al. is however based on a cohort with children regardless of their mothers asthma status, which is the main inclusion criteria for the COPSAC children, i.e. only children with asthmatic mothers are included in the COPSAC study.

Hence adjusting Martinez et al. proportion such that only children with asthmatic is considered mothers seems to be a better approach. The updated table is shown in Table 4.8, which shows that the proportions still differ significantly. Mainly the non-wheezing group in for the LCR method is too large compared to Martinez et al. but also the late-onset group is seen to be too small. This may be caused by a difference in horizon, i.e. that the children in Martinez et al. have had a year longer to develop asthma. It is seen that the local information is better in terms of identifying both the transient wheezers and the late onset wheezers. This shows that the local information is more important than the general symptom-picture for the groups, which changes pattern in the first 5 years of life.

Group	Method	No	Transient	Late	Persistent	$P(>\chi^2)$
Gaussian	Local	73	6	9	11	<0.0001
	Cumulated	75	3	3	18	<0.0001
Poisson	Local	77	6	9	8	<0.0001
	Cumulated	78	3	3	15	<0.0001
Martinez et al. adj		33	18	22	27	

Table 4.8: Asthma status shift (in %) based on LCR. In the last row the corresponding distribution for Martinez et al. [26] is shown and in the last column χ^2 goodness-of-fit test for LCR shift vs. Martinez et al. adjusted such that only children with asthmatic mothers are considered

4.5 Modelling with diagnosis

The models with gaussian and poisson response, respectively, could be reconsidered with a new grouping variable, namely the diagnosis. Diagnosis can be thought as a way of dividing the cohort in to two subgroups, which then can be analyzed to see the longitudinal development, the effect of PD15 PtcO2 and day care start in the two groups.

If a connection between group and diagnosis is present, one would expect the estimated models to be close to identical. Since the agreement was seen to rather good in Table 4.3 and 4.4, it is likely that the asthma group will be similar to the high group and the non-asthmatic comparable with the low and middle groups.

4.5.1 Gaussian model

The gaussian model with the arcsine-root transformed relative number of episodes is considered first. To make the analysis simple, the final model with the 3 groups considered in (3.24) at p. 57, but with untransformed age at day-care start, is used as the starting model, i.e.

$$\begin{aligned}\tilde{y}_{ijk} =& \beta_0 + \alpha_k + \beta_{1k} \cdot \text{age}_{ij} + \beta_{2k} \cdot \text{age}_{ij}^2 + \beta_{3k} \cdot \log_{10}(\text{pd}_i) \\ &+ \beta_{4k} \cdot \text{daycare}_{\text{start},i} + \varepsilon_{ijk} \\ i = 1, \ldots, m \quad & j = 1, \ldots, n_i \quad k = 1, 2 \quad \boldsymbol{\varepsilon}_k \sim \mathcal{N}(\mathbf{0}, \sigma_k^2 \mathbf{G})\end{aligned} \tag{4.6}$$

where k now has two levels asthma and non-asthma. The model is estimated by means of generalized least squares and the corresponding summary shows that a reduction to

$$\begin{aligned}\tilde{y}_{ijk} =& \beta_0 + \alpha_k + \beta_{1k} \cdot \text{age}_{ij} + \beta_{2k} \cdot \text{age}_{ij}^2 \\ &+ \beta_3 \cdot \log_{10}(\text{pd}_i) + \varepsilon_{ijk} \\ i = 1, \ldots, m \quad & j = 1, \ldots, n_i \quad k = 1, \ldots, 2 \quad \boldsymbol{\varepsilon}_k \sim \mathcal{N}(\mathbf{0}, \sigma_k^2 \mathbf{G})\end{aligned} \tag{4.7}$$

is possible. The reduction leads to a likelihood-ratio test of 5.38 on 3 degrees of freedom, which is seen to be an insignificant decrease in the likelihood ($p = 0.15$).

The summary is shown in Table 4.9, which shows that the asthmatic group has a higher slope, a slightly more negative curvature and the same intercept as the non-asthmatic group. The model estimates are seen to be similar to the analysis with the model with 3 groups, although some deviations are seen. The main similarity is the temporal development for the asthmatic/high group, whereas the non-asthmatic group is seen to be a mixture of the low and middle group (see curve 1 and 2 in Figure 4.5). The residual variance in the two groups is seen to be close to identical, the asthmatic group has a variance being 1.12 times higher compared to the non-asthmatic group. A test for equal variances gives p=0.29, which shows that the variances can be assumed to be equal.

	Value	Std.Error	t-value	p-value
(Intercept)	0.0472	0.0069	6.8279	<0.0001
diagnosisAsthma	-0.0060	0.0166	-0.3610	0.7182
age	0.0077	0.0053	1.4640	0.1435
(age^2)	-0.0026	0.0009	-2.9673	0.0031
(log10(pd))	-0.0047	0.0018	-2.6040	0.0093
diagnosisAsthma:age	0.0411	0.0126	3.2559	0.0012
diagnosisAsthma:(age^2)	-0.0042	0.0021	-2.0395	0.0416

Table 4.9: Summary for gaussian model with diagnosis as grouping variable, which can be compared to Table 3.9 at p. 49 or Table 3.13 at p. 59.

Figure 4.5 shows that the asthmatic group's curve (curve 2) is different compared to the dashed curves (the models considered in Figure 3.36). It is characterized by an initial increase to the age of 3-4 years and a decrease from the age of 4 to 5 years, where the LCR models top at the age of 2-3 years. The non-asthmatic group is seen to be a mixture of the low and the middle group (curve d and c), hence starting at a lower level compared to the asthmatic group and having a decreasing rate.

Fitting the same model structure (the model in (4.7)) with the groups from the gaussian LCR regression gives the possibility to compare the two grouping methods with the Bayesian Information criteria ($-2 \cdot \text{log-lik} + n_{\text{par}} \cdot \log(n_{\text{obs}})$). The grouping based on the diagnosis has a BIC of -3698.64, whereas the grouping based on the gaussian LCR has a BIC of -4143.85. This shows that the inclusion of the middle group, which essentially corresponds to dividing the low group into two subgroups gives significantly extra information in describing the symptoms. This confirms the results found in the LCR, i.e. that three clusters are optimal in order to describe the symptoms.

4.5.2 Poisson model

Likewise a model with poisson response can be estimated with the diagnosis-variable as the grouping variable. The estimation is done with GEE as seen in section 3.6.4 with an unstructured correlation for the observations within individual. Modelling is based on the model in (3.39) at p. 84, which is

$$
\begin{aligned}
\hat{\eta}_{ij} = \log(\hat{\mu}_{ij}) &= \log(n_{\text{days}_{ij}}) + \beta_{0k} + \beta_{1k} \cdot \text{age}_{ij} + \beta_{2k} \cdot \text{age}_{ij}^2 \\
&+ \beta_{3k} \cdot \log_{10}(pd_i) + \beta_{4k} \cdot (\text{daycare}_{\text{start},i}) \\
k &= \{\text{non-asthma}, \text{asthma}\} \quad Corr(Y_{ij}, Y_{ik}) = \alpha_{jk}
\end{aligned}
\tag{4.8}
$$

The model can be reduced to a model without PD15 PtcO2 and day-care start and the same curvature for both groups. The summary corresponding to the reduced model is shown in Table 4.10 and the predictions are shown in Figure 4.5 (curve 3 and 4). The predictions are seen to be higher compared to gaussian model with diagnosis as grouping variable. However the asthma group is seen to be closer to the curves from the LCR compared to the gaussian model.

	estimate	san.se	wald	p
(Intercept)	-5.6755	0.1220	2164.5702	<0.0001
diagnosisAsthma	0.3384	0.1789	3.5799	0.0585
age	0.4407	0.0868	25.7431	<0.0001
(age^2)	-0.1147	0.0153	56.0579	<0.0001
diagnosisAsthma:age	0.3615	0.0533	45.9998	<0.0001

Table 4.10: Parameter summary for GEE model with diagnosis as grouping variable, san.se corresponds to robust standard errors for the estimates.

4.5.3 Comparison of models for yearly symptom rate

In Figure 4.5 predictions for the 5 different models, the two diagnosis models and the three LCR models, for the symptom rate are shown. It is seen that the 5 models all have a high group (curve 2, 4, 5, 8 and 11), which has an initial increase to the age of 3-4 years and a decrease from that point on. For children not in the high group two groups are seen for the LCR approach, whereas the diagnosis only has one group. It is seen that curve 1 and 3 for non-asthmatic children for the gaussian and poisson response, respectively, are located between the corresponding low and middle curves for the LCR, i.e. curve 1 between curve d and c and curve 2 between curve a and 9.

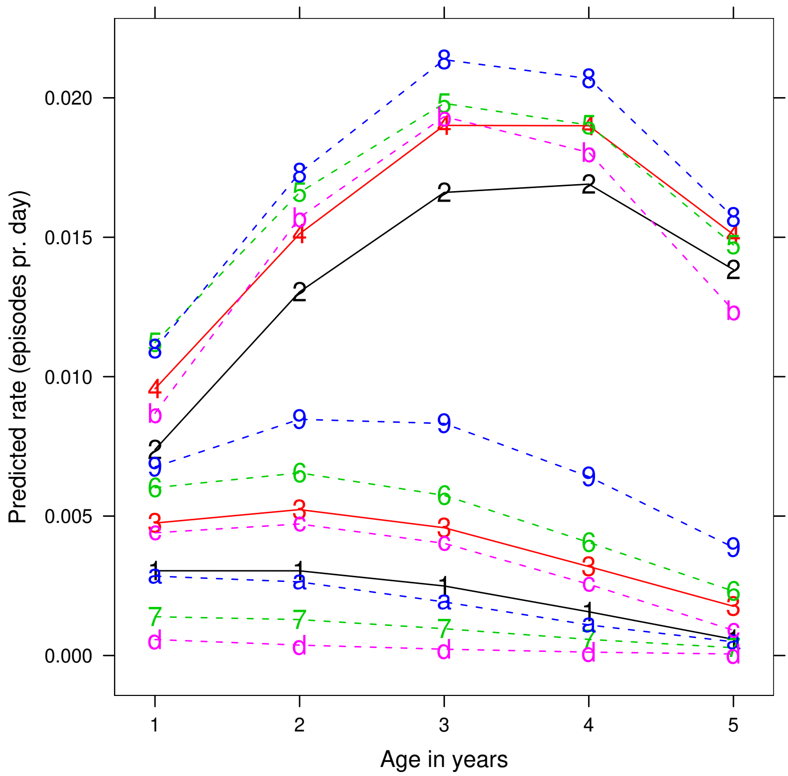

Figure 4.5: Comparison between gaussian model based on diagnosis (1 and 2, solid black lines), poisson model based on diagnosis (3 and 4, solid red lines), poisson model based on gaussian grouping in LCR (5-7, dashed green lines), poisson model based on poisson grouping (8, 9 and a, dashed blue lines) and gaussian model for gaussian grouping (b-d, dashed pink lines)

4.6 Mixed effects models and diagnosis

In the analysis of the random effects in both the gaussian and poisson model, the question of a precise grouping of the estimates was not easily addressed. However with the diagnosis at hand, the estimates can be analyzed by grouping by diagnosis. The LCR models showed that the estimates differed for the three clusters, i.e. that grouping based on the diagnosis may be possible.

4.6.1 Mixed effects vs. diagnosis

In equation (3.8) in section 3.2.4 a mixed effects model given as

$$\begin{gathered}\tilde{y}_{ij} = \beta_0 + b_{0i} + (\beta_1 + b_{1i}) \cdot \mathrm{age}_{ij} + (\beta_2 + b_{2i}) \cdot \mathrm{age}_{ij}^2 + \\ \beta_3 \cdot \log_{10}(\mathrm{pd}_i) + \beta_4 \cdot \mathrm{daycare}_{\mathrm{start},i} + \varepsilon_{ij} \\ i = 1, \ldots, m \quad j = 1, \ldots, n_i, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{\Lambda}) \\ \boldsymbol{\Lambda} = \mathrm{diag}(1/\mathrm{ndays}_{i1}, 1/\mathrm{ndays}_{i2}, \ldots, 1/\mathrm{ndays}_{in_i}) \\ b_{0i} \sim \mathcal{N}(0, \sigma_0^2), \quad b_{1i} \sim \mathcal{N}(0, \sigma_1^2), \quad \mathbf{b_i} = \begin{bmatrix} b_{0i} & b_{1i} \end{bmatrix}^T \sim \mathbf{MVN}(\mathbf{0}, \mathbf{G})\end{gathered} \tag{4.9}$$

was fitted. The BLUP-estimates [37] from this model can be compared to the diagnosis to analyze if patterns are present. The BLUP-estimates is plotted against the diagnosis for each child in Figure 4.6, which shows that the asthmatic group tends to have a higher baseline rate, but the most pronounced tendency is that the children in the asthmatic group have steeper slopes. Corresponding Wilcoxon-test, see Petrucelli et al. p. 657 [30], for the group-differences in mean gives $p<0.0001$ and $p<0.0001$ for the intercept and the slope, respectively. This confirms the results from the gaussian LCR, which showed that the group with the steepest slopes were most related to the high group.

In section 3.5.1 a generalized linear mixed effects model was considered, where the response was assumed to be poisson distributed. The model was formulated as

$$\begin{aligned}\hat{\eta}_{ij} = \log(n_{\mathrm{days}_{ij}}) + \alpha + b_{0i} + (\beta_1 + b_{1i} \cdot \mathrm{age}_{ij} + \\ (\beta_2 + b_{2i}) \cdot \mathrm{age}_{ij}^2 + \beta_3 \cdot \log(\mathrm{PD15\ PtcO_2})_i + \\ \beta_4 \cdot \mathrm{daycare}_{start,i} \\ b_{0i} \sim \mathcal{N}(0, \sigma_0^2) \quad b_{1i} \sim \mathcal{N}(0, \sigma_1^2) \quad b_{2i} \sim \mathcal{N}(0, \sigma_2^2) \\ \begin{bmatrix} \mathbf{b_0} & \mathbf{b_1} & \mathbf{b_2} \end{bmatrix}^T \sim \mathbf{MVN}(\mathbf{0}, \mathbf{G})\end{aligned} \tag{4.10}$$

The model has three random components from which BLUP-estimates are estimated. As for the linear mixed effects model, the BLUP-estimates are compared to the diagnoses, which is done in Figure 4.7.

It is seen that the parameter corresponding to the curvature is the same for the two groups, which coincide well with the results from the poisson LCR. The GEE model

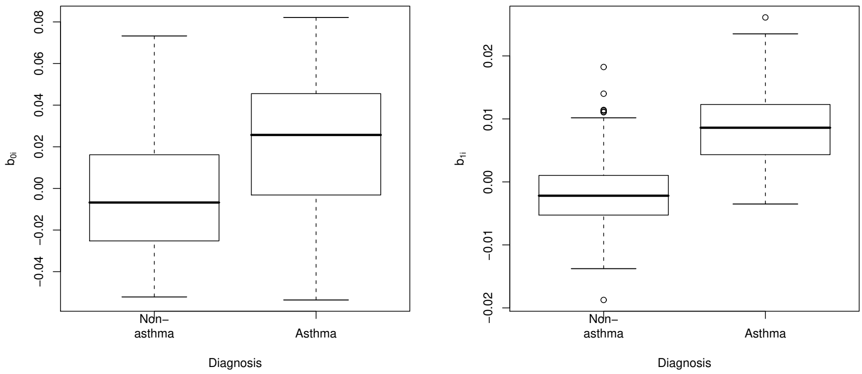

Figure 4.6: Boxplot of b_{0i} (left) and b_{1i} (right) vs. diagnosis obtained from the linear mixed effects model

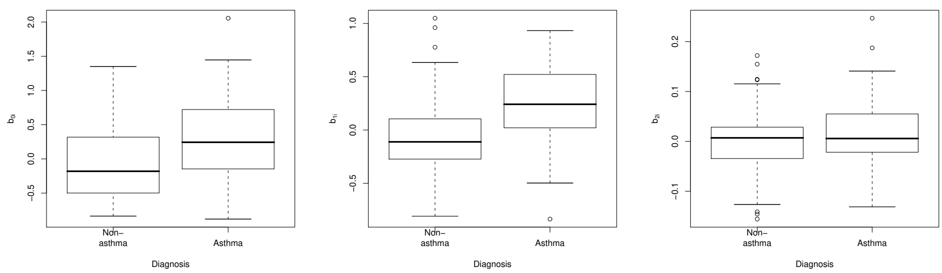

Figure 4.7: Boxplot of b_{0i} (left),b_{1i} (middle) and b_{2i} (right) vs. diagnosis from generalized linear mixed effects model

based on the poisson LCR (p. 85) was reduced to a model with the same curvature parameter for all three groups. Wilcoxon-tests for the three random components give p=0.0008, p<0.0001 and p=0.4236 for the intercept, slope and curvature, respectively. The Wilcoxon-tests show that the asthmatic group has higher intercept and slope compared to the non-asthmatic group, which was seen for the gaussian case as well. It is seen that the different approaches lead to the same results. However LCR is seen to be a more useful tool in order to identify the groups of children compared to the mixed effects models without knowledge of the diagnoses.

4.6.2 BLUP estimates as predictor

As seen from Figure 4.6 and 4.7 BLUP-estimates of the slope and intercept in the mixed effects model are closely linked to the diagnoses. Using these two measures as the cutpoints for an asthma prediction are shown in Figure 4.8, which shows that setting the b_1-criteria to $\hat{b}_{1i} > 0.004 \rightarrow$ asthma leads to an overall agreement of 85 % on a sensitivity and specificity of 80 % and 86 %, respectively. For b_0 the performance is lower and it is furthermore seen that the gaussian model is better than the poisson model. The lower performance for b_0 is expected, since the boxplots showed that the b_{0i} are overlapping for the asthmatic and non-asthmatic children more than seen for b_{1i}. The gaussian estimates are seen to be better separated compared to the estimates for the poisson model, which explain the better performance in the gaussian separation analysis.

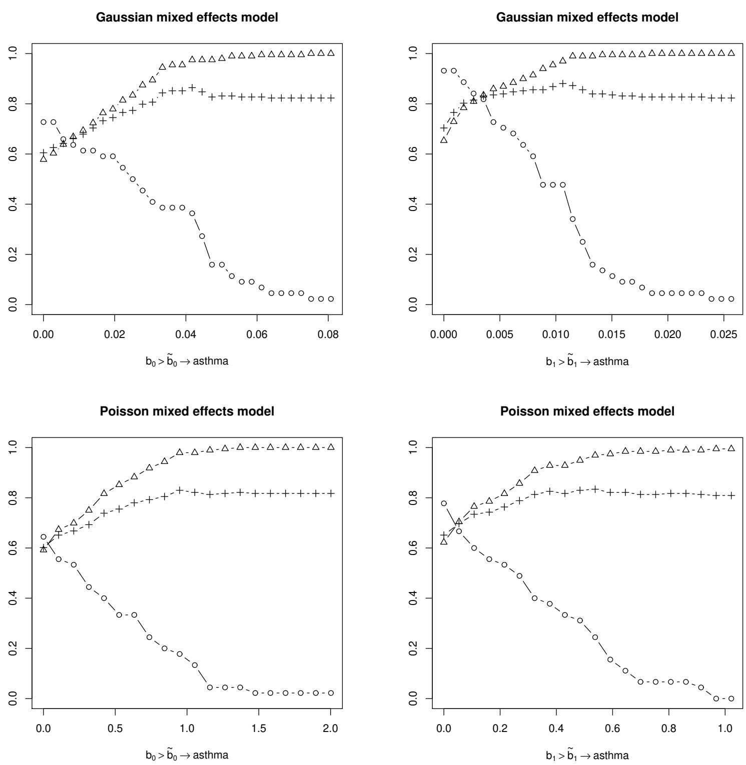

Figure 4.8: Cut-points for gaussian (top row) and poisson (bottom row) mixed effects model. Each plot shows sensitivity: o, specificity: Δ and overall agreement: +

By using both the intercept and the slope as a combined cut-off for the gaussian model the performance may be improved. This can be obtained by replacing the 1-parameter cut-offs by the criteria $b_{0i} > \tilde{b}_0 \vee b_{1i} > \tilde{b}_1 \rightarrow$ asthma. This imply that if at least one the BLUP-estimates are higher than their respective threshold-value the child is classified as asthmatic. The sensitivity, specificity and overall agreements are shown in Figure 4.9, which shows that using both parameters are seen not to improve the sensitivity or the overall agreement significant. It is possible to obtain sensitivity, specificity and overall agreement above 86 % with the threshold values $b_{0i} > 0.0362 \vee b_{1i} > 0.0075 \rightarrow$ asthma. It is seen that this improves the sensitivity on the same specificity as only using the b_{1i} values, however the difference is small and entirely related to improving the sensitivity.

Compared to the LCR-approach it is seen that the method is better, i.e. compared to the gaussian model with the middle and low groups merged. However, the improved version with a cut-point for the posterior probability of the high group in the gaussian model is better compared to the cutting based on the mixed effects. For the cutting on posterior probabilities the sensitivity is 87 % for a specificity of 89 %, which gives an overall agreement of 89 %, which is seen to be higher compared to cutting in the mixed effects.

Another difference in the two methods is that the LCR (without cut-point) is able to find the groups without knowing the asthma status. This is however not the case in the mixed effects case, since no apparent grouping in seen in the estimates. It was furthermore seen that a strategy based on a subset of the age-range was easy to incorporate in LCR, which for the mixed effects model would imply that a new model should be estimated.

4.7 Diagnosis and five cluster grouping

In section 3.6.2 the optimal poisson LCR was estimated, which had 5 clusters. The clusters from that model (see Figure 4.10) can be compared to the diagnoses as for the 3 cluster models, which has been done in Table 4.11. The table contains the cross-tabulation between the 5 clusters and the diagnosis, which shows that group 1 and 5 clearly contains non-asthmatic children and group 3 asthmatic children. Using these definitions and defining group 2 and 4 as misclassification for both the asthmatic and non-asthmatic group gives a specificity of 76 % with a sensitivity of 31 % and an overall agreement of 68 %.

Group 2 and 4 are difficult to interpret, however if they are regarded as asthmatic children, the sensitivity becomes 100 % with a specificity of 76 % and an overall agreement of 80 %. This is a higher sensitivity compared to the three cluster models, but lower overall agreement. Another possibility is to assign cluster 2 and 4 to the non-asthmatic groups (1 and 5), which increase the specificity to 100 % for a sensitivity of 31 % and an overall agreement of 87 %. If group 2 is classified as asthmatic and group 4 as non-asthmatic the specificity becomes 88 % for a sensitivity of 49 % and an overall agreement of 81 %

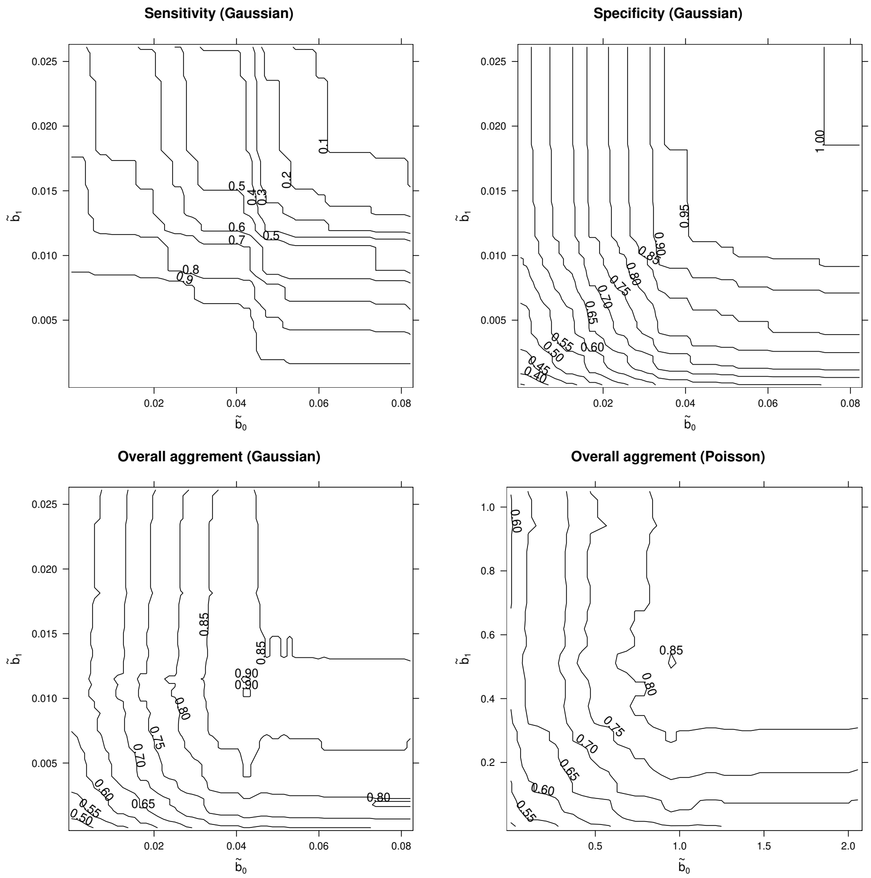

Figure 4.9: Contour-plots for sensitivity (upper left), specificity (upper right, overall agreement (lower left) for gaussian model and overall agreement for poisson model (lower right).

The five cluster model is seen not to be markedly better in terms of predicting the diagnoses even though it uses more parameters. The five clusters are seen to be more specialized, since two groups with symptoms at the age of 1 and 5 years are seen, one with some symptoms and one with many symptoms. Furthermore, one group with late onset symptoms, one group with early transient symptoms and one group with no symptom are seen.

As mentioned in section 4.4, Martinez et al. [26] operate with 4 prototypes of children, which the five clusters can be compared to. Group 3 and 4 can be interpreted as persistent wheezers, since they both keep a high level respectively medium level, group 2 is classified as late onset wheezers, group 1 as early transient wheezers and finally group 5 as children not wheezing at all. The distribution of the children in

Group	Non-asthma	Asthma
5	104	0
1	48	0
4	25	23
2	22	8
3	1	14

Table 4.11: Cross-tables of subgroups from 5 cluster poisson LCR vs. doctor diagnosed asthma-status

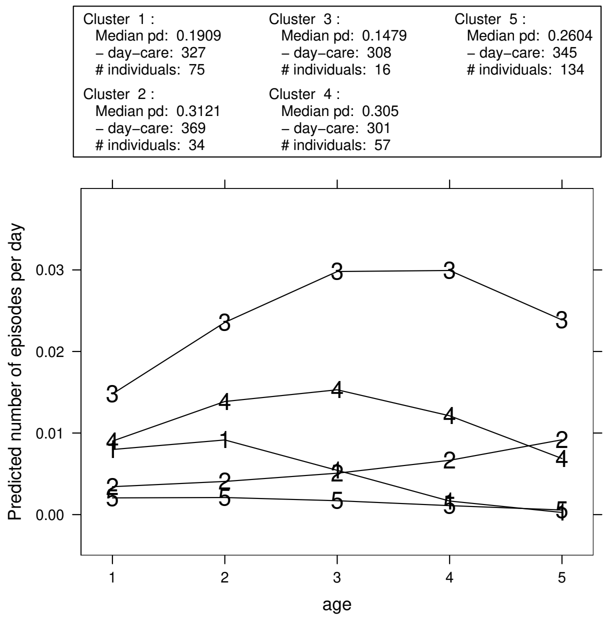

Figure 4.10: 5 cluster model

the five cluster model can be compared to the distribution found by Martinez et al, which is shown in Table 4.12. Comparing the 4 proportions can be done by a $2 \times c$, χ^2-test as described at p. 93 in Kirkwood [22] and section 4.4.

From the table it is seen that too few children have no symptoms, whereas too many have persistent symptoms. However some similarities are seen, namely that the non-wheezing group is the largest group and the transient early wheezing group the second largest. The inclusion criteria may influence the distribution, since the COPSAC children are included only if their mothers have asthma, whereas the children in Martinez et al. do not have this inclusion criteria [26]. However both the late-onset wheezers and the persistent wheezers in Martinez et al. had a significant odds-ratio for maternal asthma, 95 % confidence intervals are $[1.4; 5.5]$ and $[2.1; 7.9]$, respectively, compared to the non-wheezers. So including only the newborns with asthmatic mothers may increase the number of wheezers compared to a population representing newborns in general as seen in section 4.4.

Method	No	Transient	Late	Persistent	$P(>\chi^2)$
5 cluster	42	24	11	23	$0.3202{\cdot}10^{-4}$
Martinez et al	52	20	15	13	
Pearson residuals	-2.22	1.46	-1.6	3.72	

Table 4.12: Asthma status shift (in %) based on 5 cluster poisson LCR. In the middle row the corresponding distribution for Martinez et al. [26] is shown, the last row the pearson residuals and in the last column χ^2 goodness-of-fit test for 5 cluster LCR shift vs. Martinez et al.

If only the children with asthmatic mothers are used from the study by Martinez et al. [26] the number of children in each of the four group is 27, 15, 18 and 22. The updated table for comparison of the 5 cluster model and Martinez et al. is shown in Table 4.13, which shows that there still is a significant difference at a 5 % level between the two populations although the p-value is 100 times larger. It is seen that the corrected distribution for Martinez et al. has a lower proportion non wheezers compared to the proportion for all children and a twice as large proportion of persistent wheezers compared to the original distribution. The persistent wheezing groups are seen to be of equal sizes, whereas the late onset group is smaller for the 5 cluster model compared to the corrected Martinez et al. proportion and is contributing by 70 % of the total test-statistic.

It is seen that the late-onset group in the five cluster model is significant lower compared to Martinez et al. The Pearson residuals are $N(0,1)$ distributed, which imply that the residual for the late-onset group is highly significant. The non-wheezing group is seen to be too large, however not significantly ($p = 0.11$). The intersections in Martinez et al. are the age of 3 and the age of 6 years, whereas the analysis in the LCR is based on the first 5 years of life, which imply that the late-onset group may increase in the sixth year of life and thereby reducing the non-wheezing group. However the deviations between the two studies are seen not to be large for the remaining three prototypes.

Method	No	Transient	Late	Persistent	$P(>\chi^2)$
5 cluster	42	24	11	23	0.0304
Martinez et al. corrected	33	18	22	27	
Pearson residuals	1.2	0.92	-2.5	-0.62	

Table 4.13: Asthma status shift (in %) based on 5 cluster poisson LCR. In the middle row the corresponding distribution for Martinez et al. [26] for children with asthmatic mothers is shown, the last row the pearson residuals and in the last column χ^2 goodness-of-fit test for 5 cluster LCR shift vs. Martinez et al. for children with asthmatic mothers

4.8 Medication

All previous analysis has been done without taking the symptom related medication into account. This may influence the inference, since correct medication decreases the amount of symptoms. In the following use of medication will be analyzed with respect to the symptoms, the diagnosis and the longitudinal development.

Through the first 3 years of life the children are randomly assigned to either inhaled corticosteroid (budesonide) or placebo [4]. The children are treated for 14 days from the third day with symptoms. Each child is assigned to one and only one treatment, however if an asthma diagnosis is given the trial is stopped.

Aside from the treatment in the nested trial, the children have been treated with spirocort (budesonide) and prednisolon. Both drugs are glucocorticoids, but spirocort has local effect and is inhaled, whereas prednisolon is a systemic drug. Prednisolon is given only at severe acute asthma episodes. Budesonide is given for each episode the first 3 years of life as described in the previous paragraph and the general treatment with spirocort the first 6 years of life is roughly as follows: If 5 episodes have occurred the last 6 months a three month trial medication with spirocort is initiated. If relapse after the trial period is observed, a 6 months period is started otherwise the treatment is stopped, if relapse after the 6 months period is observed the treatment length is increased to 12 months.

It is seen that the length of the treatment is increased as long as the treatments have an effect, i.e. that relapse is seen when stopping the treatment. All medication is seen to be related to the level of symptoms, since it is given according to the symptoms.

4.8.1 Longitudinal development in medication amount

In Figure 4.11 boxplots of the relation between the number of days with the initial medication (placebo/budesonide) and the asthma diagnoses is shown. For the placebo treatment it is seen that the non-asthmatic group has fewer days with medication compared to the asthmatic group. The asthmatic group is seen to have more days with placebo treatment in the second year of life compared to the first, whereas the third year is seen to be comparable with the non-asthmatic group, which can be explained by the fact that the asthmatic children typically will be excluded from the initial experiment and put on an active asthma treatment with spirocort.

For the active treatment in the initial trial the number of days is highest at the age of 2 for the non-asthmatic group and at the age of 1 for the asthmatic. The asthmatic group is seen only to receive budesonide in the first year of life, which could be explained by the fact that they receive spirocort treatment instead as seen from Figure 4.12.

Figure 4.12 shows the boxplots for spirocort and prednisolon the first 6 years of life, respectively. It is seen that the majority of the non-asthmatic children are seen not to be medicated with spirocort, whereas the level of spirocort treatment for the

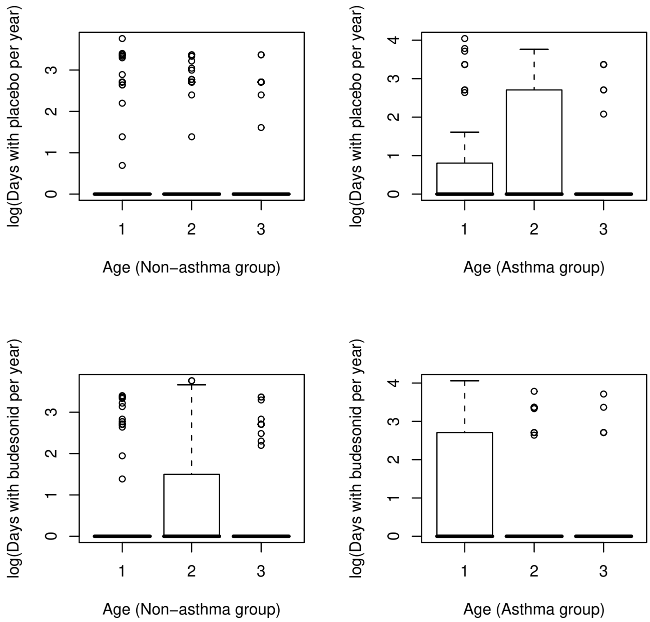

Figure 4.11: Boxplots of medication the first three years vs. asthma diagnosis. Medication is here either placebo or budesonide

asthmatic children is seen to be increasing. The prednisolon treatment level is seen to be similar across the two groups and it is furthermore seen to be sparse, which is explained by the fact that prednisolon is given only at acute severe asthma, i.e. under special circumstances.

At a yearly aggregated level the number of episodes and the number of days with eg spirocort will be positively correlated, which do not reveal the benefits of the treatment. However on a weekly basis the faster dynamics may reveal the benefits of the medication.

4.9 Risk-factors for asthma

In the previous sections the diagnosis and the clustering from LCR was analyzed. In the following section the risk-factors for the diagnosis are analyzed by means of logistic regression. The logistic regression is a model for the probability of getting the asthma-diagnosis given that the children have certain risk-factors, eg a mother smoking in the third trimester. A subset of the risk-factors obtained at birth and the age of 3 are considered to reduce the analysis, the same risk-factors will be analyzed

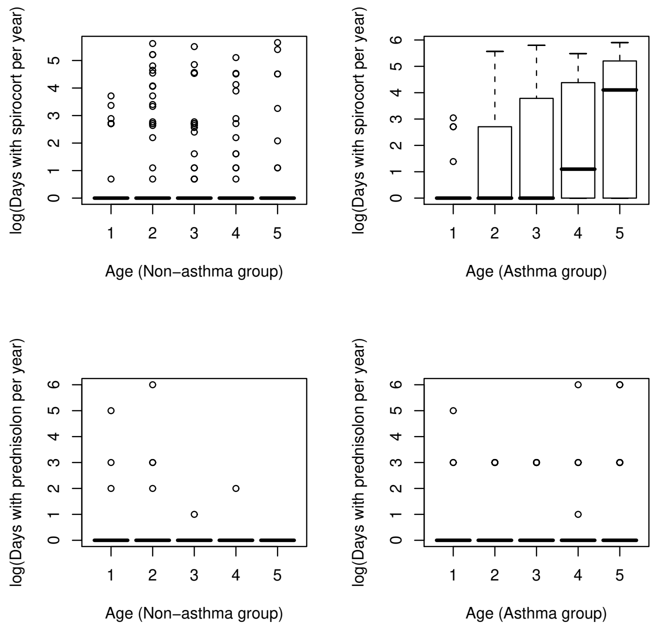

Figure 4.12: Boxplot of medication the first 6 years of life vs. asthma diagnosis. Medication types are spirocort and prednisolon.

in section 5.6. A description of the risk-factors and confounders is given in the preparatory thesis [13].

In Figure 4.13 a NA-plot from Harrels `Hmisc` package [21] is shown. It is seen that the variable with the most missing values is the sraw measurement with over 40 % missing values followed by the diagnosis with $\approx 25\%$ missing values in a dataset with 316 children. The remaining variables are seen to have less than 10 % missing values and furthermore most have none.

The logistic regression aims at modelling the probability of having asthma, hence the success in each of the Bernoulli trials is a positive asthma-diagnosis. The usual way to model a logistic regression is in a generalized linear model framework, which calls for the specification of a distribution, a link-function and a linear-predictor. Obviously the distribution is the binomial-distribution, which has the logit as the canonical link-function for the mean p, i.e. $f(p) = \log(p/(1-p))$, see Wood p. 61 [42]. The linear predictor consists of the risk-factors, which for simplicity is assumed additive (however with an interaction between the two rhinitis variables) on the logit-scale, hence multiplicative on the probability scale.

A few things are noted on the generalized linear model for a Bernoulli trial (binomial

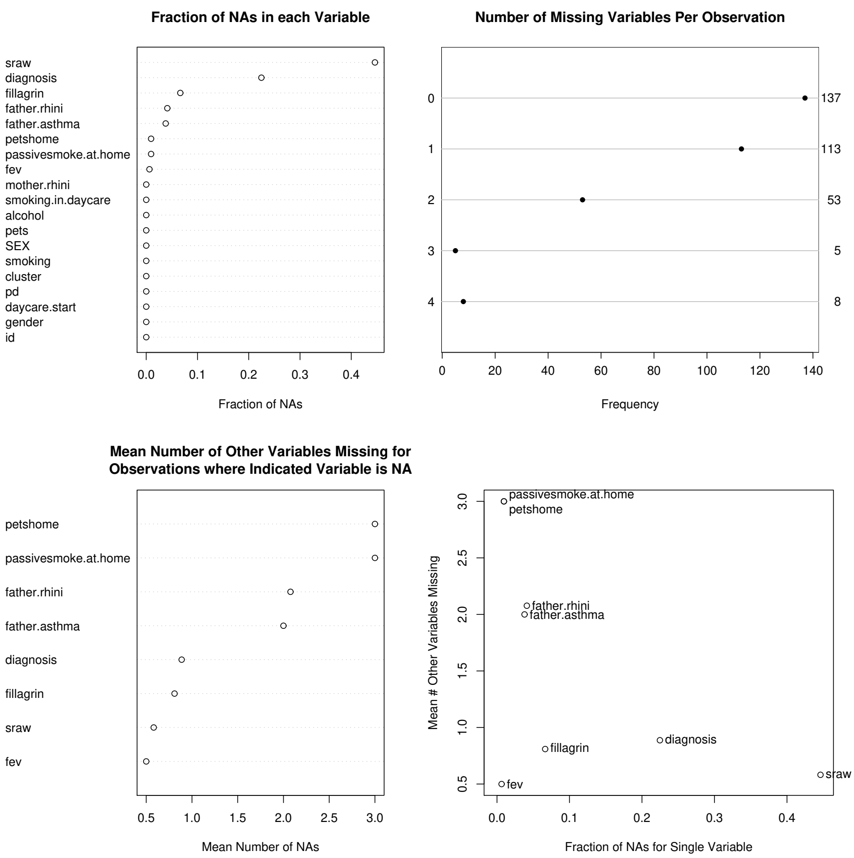

Figure 4.13: NA-plot for risk-factor dataframe. Top left: Fraction of missing values, top right: Frequency of number of missing values per observation, bottom left: Mean number of missing values given that a certain variable is missing and bottom right: Fraction of missing-values vs. mean of other variables missing

with $n = 1$), the mean value is modelled through the link-function, the variance is a known function of the mean value $V[Y] = p \cdot (1 - p)$ and over-dispersion may be present due to heterogeneity or an inadequate model, i.e. $V[Y] = \phi \cdot p \cdot (1 - p)$, where $\phi > 1$. The overdispersion can be modelled by quasi-likelihood, see Wood p. 74 [42], by including the scale-parameter ϕ.

The logistic regression model therefore becomes

$$\begin{aligned}\text{logit}\,(\hat{p}_i) = \hat{\eta}_i = {} & \beta_0 + \beta_1 \cdot \text{fillagrin} + \beta_2 \cdot \text{smoking}_{\text{3rd},i} + \beta_3 \cdot \text{smoking}_{\text{home},i} \\ & + \beta_4 \cdot \text{smoking}_{\text{daycare},i} + \beta_5 \cdot \text{gender}_i + \beta_6 \cdot \text{father asthma}_i \\ & + \beta_7 \cdot \text{pets}_{\text{3rd},i} + \beta_8 \cdot \text{alcohol}_{\text{3rd},i} + \beta_9 \cdot \text{smoking}_{\text{3rd},i} \\ & + \beta_{10} \cdot \text{sraw}_i + \beta_{11} \cdot \text{father rhinitis}_i + \beta_{12} \cdot \text{mother rhinitis}_i \\ & + \beta_{13} \cdot \log_{10}(\text{pd})_i + \beta_{14} \cdot \text{daycare start}_i + \beta_{15} \cdot \text{pets}_{\text{home},i}\end{aligned} \tag{4.11}$$

The measurement of the specific resistent in the airways variable, sraw, is the relative increase in sRAW after inhalation of terbutaline. sraw, amount of passive smoking in the home and exposure to furred pets in the home are all continuous and therefore needs to be analyzed with a generalized additive model to see if curvature is present. PD15 PtcO2 has previous been shown to be linear after a logarithmic transformation, whereas age at day-care start is linear.

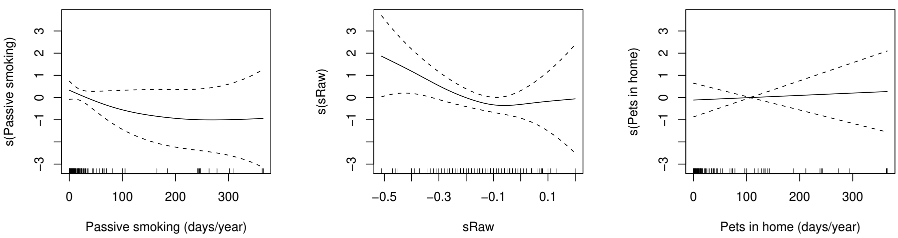

Figure 4.14: Smoothed function for continuous variables in logistic regression for asthma diagnosis

A model with smoothed functions for the three continuous variables is fitted in order to evaluate the curvature. The resulting smoothed functions are shown in Figure 4.14, which shows that neither of the variables seem to be significant and furthermore the relation to the linear predictor seems linear. Test for the smoothed functions, see Wood p. 194-195, can be done with a F-test as

$$\frac{\hat{\beta}_j^T \hat{\mathbf{V}}_{\hat{\beta}_j}^{r} \hat{\beta}_j / r}{\hat{\phi}/(n - \text{edf})} \sim F_{r,\text{edf}} \tag{4.12}$$

where $\hat{\beta}_j$ is the parameters corresponding to the smoothed function j, $\hat{\mathbf{V}}_{\hat{\beta}_j}^{r}$ the corresponding variance-covariance matrix, r the estimated degrees of freedom for smooth function j and edf the estimated degrees of freedom for the model. A F-test is used if the scale parameter is to be estimated otherwise a χ^2 test is used on the numerator without the scaling by r, which is a test with r degrees of freedom.

The F-tests gives p= 0.77, p= 0.19 and p= 0.07 for furred pets, passive smoking in the home and sraw, respectively. It is furthermore seen from Table 4.14 that the interaction can be removed as well as exposure to pets in the third trimester, fathers asthma, gender, fillagrin gene mutation, smoking in day-care and age at day care start. It is furthermore seen that the alcohol variable gives some numerical problems, since the groups corresponding to 2 units and 3 or more units per week are small (7 and 4 children). Collapsing the two groups to one joint group seems to be a reasonable approach, since the standard error of the group corresponding to 3 or more units is astronomic in size. It is also seen that the over-dispersion parameter is below 1, $\hat{\phi} = 0.81$, which shows that it is appropriate to estimate the model with a standard binomial distribution, i.e. without the scale parameter ϕ.

In the following the tests for the reductions are done on datasets of equal sizes, whereas the number of observations in general will be increasing in the summaries, due to fewer incomplete observations as the number of variables is decreased.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.8508	1.1518	-2.4752	0.0148
genderMale	-0.0451	0.5087	-0.0887	0.9295
fillagrin1	1.1567	0.9292	1.2448	0.2157
daycare.start	-0.0023	0.0021	-1.0970	0.2749
fev	-0.0468	0.0238	-1.9681	0.0514
(log10(pd))	-1.3192	0.4033	-3.2709	0.0014
smoking3rdYes	1.8211	0.8260	2.2046	0.0294
$\text{father}_{\text{ashtma}}$Yes	0.9732	0.7771	1.2524	0.2129
pets_{3rd}	-0.5223	1.1157	-0.4681	0.6406
alcohol1	0.4353	0.7401	0.5881	0.5576
alcohol2	2.9137	1.3883	2.0987	0.0380
alcohol3	18.3673	3563.8058	0.0052	0.9959
$\text{smoking}_{\text{daycare}}$1	0.6655	0.5593	1.1899	0.2365
father.rhiniYes	0.3744	1.0564	0.3544	0.7237
mother.rhiniYes	-0.3136	0.8313	-0.3772	0.7067
father.rhiniYes:mother.rhiniYes	1.0918	1.2586	0.8675	0.3875

Table 4.14: Summary for parametric terms in GAM model. n= 137

A model without the smoothed functions and corresponding variables and with the reductions outlined is estimated by means of a standard generalized linear model. The updated model gives an increase in the residual deviance of 14.88 on 12.35 degrees of freedom, which gives a F-test with a p-value of p=0.14, which shows that the reduction leads to an insignificant increase in the residual deviance.

It is furthermore seen that the number of observations are increased by more than 100, since the sRAW is excluded, which had more than 40 % missing values. The extra observations are appreciated for variables with few observations in a given category, eg the alcohol variable doubles the number of children with a mother drinking at least one unit of alcohol per week in the third trimester from 23 to 51. This gives smaller standard error and a much more accurate estimates of the effects in the small groups, since the estimates will rely on more children.

The updated model summary is shown in Table 4.15, which shows that for instance

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.7563	0.4986	-5.5276	<0.0001
fev	-0.0127	0.0140	-0.9132	0.3611
(log10(pd))	-0.7497	0.2563	-2.9254	0.0034
smoking3rdYes	0.8754	0.4941	1.7719	0.0764
alcohol1	0.0867	0.5126	0.1691	0.8657
alcohol>1	0.3104	0.8534	0.3638	0.7160
father.rhiniYes	0.5068	0.3759	1.3484	0.1775
mother.rhiniYes	0.4056	0.4631	0.8759	0.3811

Table 4.15: Summary for reduced model without smoothed functions and only a subset of the initial risk-factors, n= 235

smoking in the third trimester increases the risk, $OR = 2.4$, of getting the asthma diagnosis, however not significantly (p=0.08). The summary shows that mothers and fathers rhinitis history are insignificant as well as drinking alcohol and smoking cigarettes in the third trimester and the congenital FEV. Reducing the model gives an increase in the deviance of 6.19 on 6 degrees of freedom, which gives a F-test with a p-value of p=0.40. It is seen that the only significant risk-factor is the congenital responsiveness, which shows that including more risk-factors to the LCR probably will not give a better grouping.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.0324	0.2526	-8.0467	<0.0001
(log10(pd))	-0.8205	0.2389	-3.4346	0.0006

Table 4.16: Summary for final model for risk of asthma, n= 245

The model shows that congenital responsiveness is a predictor for the risk of getting the asthma diagnosis at the age of 5 years. It is seen from Figure 4.15 that the risk of asthma is about 40 % for the children with the lowest resistance and drops to below 20 % at the median pd (a 34.1 times increase in PD15 PtcO2). In general the odds-ratio for an increase by a factor 10 in the PD15 PtcO2 gives an odds-ratio for high versus low of

$$\text{OR} = e^{-0.82} = 0.44$$

The ratio between the lowest PD15 PtcO2-value and the highest is 2110.63, which corresponds to an odds-ratio of

$$\text{OR} = e^{3.32\cdot(-0.82)} = 0.07$$

hence the risk at the highest level is 7 % of the risk at the lowest level.

It is seen that even the lowest PD15 PtcO2 does not lead to a predicted risk above 50 %, i.e. the PD15 PtcO2 may need additional inspection to find the relevant cutting point and will probably not be good at finding the children with asthma. In

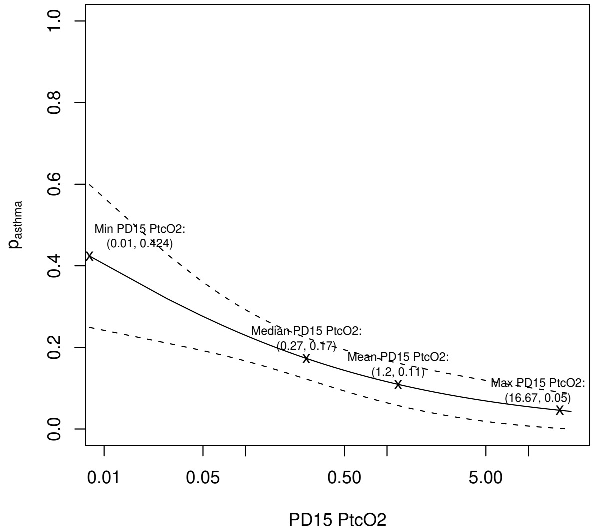

Figure 4.15: Prediction and prediction interval for risk of getting asthma diagnosis at the age of 5 years as function of PD15 PtcO2

Figure 4.16 it is seen that the overall agreement tops for a PD15 PtcO2 value of around 0.05 with a value of 81 %. This is lower compared to the LCR, where the overall agreement was around 90 % for the best models. It is furthermore seen that the model has a good specificity for a low cutting point and that the sensitivity is improved only by decreasing the overall agreement (and the specificity) significantly.

The model is seen not to be sensitive to the asthma diagnosis, which is expected due to low fitted risks. Compared to the LCR the sensitivity for the logistic regression is more than 40 %-points lower on a lower overall agreement, however it is comparable with the results obtained by only using the information in the first year of life in the LCR. This shows that using the full 5 year range in a LCR (or even just the two first years) gives a better model in terms of identifying the asthmatic children, which is should be a natural consequence of including more information in the decision criteria: LCR has the PD15 PtcO2 as one of the variables. However as a single measure at the birth, PD15 PtcO2 gives good predictions of the asthma status at the age of 5 years. The main reason for LCR's improved performance in predicting the diagnosis is that the longitudinal patterns are used.

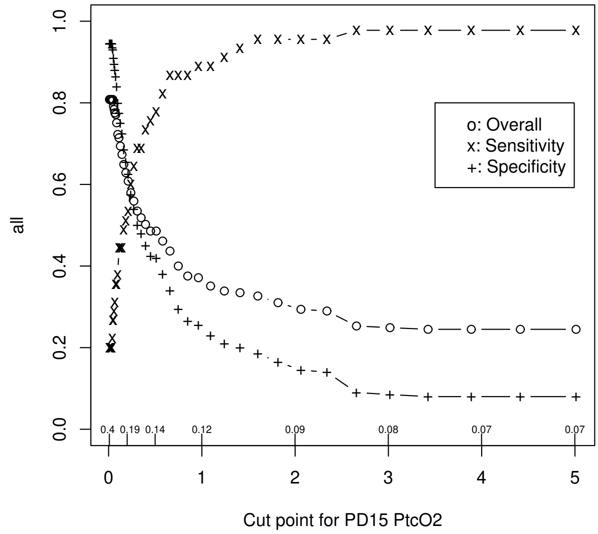

Figure 4.16: Overall agreement, sensitivity and specificity as function of the cut-point for PD15 PtcO2, i.e. PD15 $\mathrm{PtcO2}_i <$ PD15 $\mathrm{PtcO2}_{\mathrm{cut}} \rightarrow$ asthma diagnosis for child i. Above PD15 PtcO2 the corresponding probabilities of asthma, eg a PD15 PtcO2 of 1 gives a 12 % risk of asthma.

4.10 Discussion

In this chapter the diagnoses were compared to the groups found in the LCR based on both gaussian and poisson response on the first five years of life. It was seen that the gaussian approach gave the highest sensitivity and that the poisson response gave the highest specificity. However merging the middle and low group gave nearly the same specificity for the two methods with the same sensitivities.

The grouping methods were seen to give a fairly good classification at the age of two-three years. However the best results are obtained at the age of 5 years, which might be related to the fact that the diagnoses are established based on the medication and symptoms in the fifth year of life. The possibility for an early classification is interesting, since it may give some insight information on exactly how the asthmatic children's symptoms are developing. Furthermore, it was seen that the classification was good at the age of 2-3 years, which imply that future studies of childhood asthma might be reduced from considering children the first five years of life to the first 3 years. This is obviously desirable due to costs, participation percent and the waiting time for results.

It was seen that using the diagnosis as grouping variable gave the same estimated temporal development for the asthmatic and high group of children, whereas the non-asthmatic children had a curve which was a combination of the low and middle group.

In section 4.8 medication patterns for various combination of group and diagnosis were analyzed. It was seen that the misclassifications were related to either an increase (asthma-middle group) in the symptom-rate in the fifth year of life or an decrease (non-asthma-high). The grouping by LCR was seen to be more related to the entire symptom-pattern, which gave some dissimilarities compared to the diagnosis established at the age of 5 years. However the overall-agreement was good and the dissimilarities may be explained by the fact that there might be more than 2-3 groups as discussed by Martinez et al. [26]. The latter was seen to be indicated in the analysis of the 5 cluster model, where the corrected results by Martinez et al. were seen to be close to results found for the five cluster model.

Finally a logistic regression for the asthma diagnosis at the age of 5 years was analyzed. The regression showed that the only significant riskfactor was the congenital PD15 PtcO2 measurement, i.e. the dose to give a 15 % decrease in the PtcO2 (partial pressure). It was seen that the most resistant children have highly reduced risks of getting the asthma diagnosis, however no children had a fitted risk above 50 %. It was furthermore seen that finding a good cutting point for classification of asthma was possible, the performance was lower compared to the latent class regression. The logistic model was seen not to be particular sensitive to asthma, eg the highest overall agreement of 81 % was obtained for a specificity around 95 % and a sensitivity of 20 %. To increase the sensitivity to above 50 % a deterioration of the specificity to below 66 % and the overall agreement to below 63 % was seen.

Analysis of different cut-points strategies showed that the mixed effects gaussian model was capable of giving quite precise prediction based on the individual slope and intercept. It was seen that using the cumulated posterior probability at the age of 3 years for the gaussian model led to a good model, which had a sensitivity, specificity and overall agreement above 80 %. Using the posterior for all three years increased the performance to above 87 %. The analysis showed that the gaussian LCR tended to favor the groups with few symptoms, i.e. that the children should be assigned to the asthmatic group for a posterior above 0.26.

The methods discussed in this chapter showed that good prediction of the asthma diagnosis can be made. It was furthermore seen that diagnoses could be established at the age of 3 years with an acceptable precision above 80 %. Analysis showed that the asthmatic and non-asthmatic groups had distinct temporal development in the symptom-rate, which enabled the good predictions.

CHAPTER 5

Weekly episodes

Contents

5.1 Introduction

In the following chapter an analysis of the individual weekly symptoms is considered. For each child a time-series from the first year of life to the fifth year of life with as many observations as days in the considered time-range is the basis for the analysis.

For each day three outcomes are possible: an episode has occurred (day 3 or later in an incident), an episode has not occurred (indicated with a 0), which corresponds to either the two first days with symptoms or a day without symptoms and the day can be without diary data (marked as not available, see section 3.2.2 for more information on missing data). The response is seen to be the prevalence of episodes lasting 3 days or longer. 3 days was the length of an episode, which required a visit at the facilities at COPSAC.

From previous analysis of the population in Chapter 2, it is known that the symptoms vary over the year. The yearly variations can be modelled with the inclusion of the periodic functions, sine and cosine, with a period of 365 days. From the analysis of the yearly aggregated symptoms it was shown that the temporal development of the symptoms was seen to have a form corresponding to a second order polynomial in the age.

The LCR-regression analysis in Chapter 3 showed that the children could be divided into three groups according to their symptoms development. This may be used in terms of evaluating the odds-ratios between the different groups, in particularly the (probably) increased odds for the group with a high level of symptoms. However, the diagnosis considered in Chapter 4 may be used instead, in order to evaluate differences between asthmatic and non-asthmatic children.

The observation set consists of approximately 600.000 observations distributed on around 400 individuals. This imply that data analysis is a computer extensive task, which limits the level of complexity in the models. A problem of more theoretical matter with using daily episodes is that since an episode is recorded only if it last 3 days or longer, the 3 days up to the episodes will always be 0 (where 0 corresponds to no symptoms), since the first 0 is the last day in a period without symptoms and the next 2 the two first days with symptoms in a period lasting 3 days or longer. This gives severe negative correlations, which are some what artificial and entirely caused by the way the study and dataset are designed.

To avoid this problem one can analyze weeks, where a week is defined to contain an episode if at least one day of an episode having lasted 3 days or longer is contained in the week. The weeks go from Monday to Sunday, which should be fairly arbitrary with respect to the analysis, the same results should be reached having weeks from eg Wednesday to Tuesday.

The relation between the symptoms, the episodes and the weekly symptoms is illustrated by a sequence of symptoms with corresponding episodes and weekly symptoms as shown in Table 5.1. Since a week is used, the issues with strange negative correlations may to some extend be avoided. This may give a better analysis in terms of being able to estimate the autocorrelation, which will be seriously affected in a day to day analysis.

Week no.	1	2	3	4	5	6
Symptoms	0000000	0000001	110111	100100-	0111111	- - - - - - -
Episodes	0000000	0000000	010001	100000-	0001111	- - - - - - -
Weeks	0	0	1	1	1	-

Table 5.1: Illustration of relation between symptoms, episodes and week

5.2 Initial model formulation

The weekly data can be seen as a sequence of bernoulli trials: Week j for child i has a probability, p_{ij}, of being a week with an episode and $1-p_{ij}$ of not being one. Since a Bernoulli trial is a special case of a binomial trial with $n = 1$, a generalized linear model can be used to model the probability. The individual bernoulli trial has the probability function

$$f(\text{symptom}) = p^{\text{symptom}}(1-p)^{1-\text{symptom}} \tag{5.1}$$

where p is the probability of having symptoms and symptom is 1 if symptoms are present and 0 if not. The mean of a binomial distribution is $p \cdot n = p$ and the variance of the response is $V[Y] = n \cdot p \cdot (1-p) = p \cdot (1-p)$.

In a generalized linear model framework one need to specify the distribution for the response, the link-function and the linear predictor. The distribution has been specified and the link-function is the canonical link (the logit) for the binomial distribution

$$g(\mu) = g(p) = \log\left(\frac{p}{1-p}\right) = \theta \tag{5.2}$$

with the inverse

$$g^{-1}(\theta) = p = \frac{e^{\theta}}{1+e^{\theta}} \tag{5.3}$$

the linear predictor, η, is linear in the explanatory variables, which gives model equations of the form

$$\hat{\eta}_{ij} = \mathbf{x}_{ij}\beta = \beta_0 + x_{ij1} \cdot \beta_1 + \cdots + x_{ijp} \cdot \beta_p \tag{5.4}$$

for the fitted value of the linear predictor. The subscript i is the individual i and j the j'th observation for individual i. $p/(1-p)$ is the called the odds, which can vary between 0 for $p = 0$ and infinity for $p = 1$. $\exp(\beta_1)$ is the odds-ratio (OR) for an increase in x_1 of 1 unit, which is seen from

$$OR = \frac{p_1/(1-p_1)}{p_0/(1-p_0)} = \frac{\exp(\beta_0 + (x_1+1) \cdot \beta_1 + \cdots + x_p \cdot \beta_p)}{\exp(\beta_0 + x_1 \cdot \beta_1 + \cdots + x_p \cdot \beta_p)} = \exp(\beta_1) \tag{5.5}$$

Odds ratios are typically used in the context of comparing two different populations, which is modelled by indicator variables, eg. $x_{ijk} = 1$ if individual i belongs to population 1 and 0, otherwise.

Since the dataset is longitudinal correlation between the observation for an individual may be present. The correlation can be utilized in the sense that lagged values can be used as predictors, which changes the modelling from

$$\text{logit}(P(Y_{ij} = 1|x_1, \ldots, x_p)) = \mathbf{x}_{ij}\beta \tag{5.6}$$

to

$$\text{logit}(P(y_{ij} = 1|x_1, \ldots, x_p, y_{i(j-1)}, \ldots, y_{i(j-m)})) = \mathbf{x}_{ij}\beta + \mathbf{y}_{ij'}\beta_y \tag{5.7}$$

hence the modelling becomes conditional on the history. Considering two neighboring weeks two things can happen: either the symptom state is the same or it shifts to the other state. The state diagram is illustrated in Figure 5.1, which contains four probabilities, each a function of the current state (0 or 1), the explanatory variables and possibly more history than current state. These types of models are called transition models, since the transition probabilities are modelled. The state diagram may be complicated further if the probabilities depend on more history than the current state, i.e. if the first order Markov property is not fulfilled.

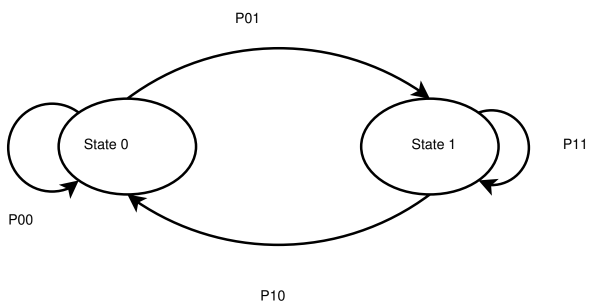

Figure 5.1: State diagram for week to week evolvement

In modelling the individual time-series an usual generalized linear model can not be used, since the observations within an individual are likely to be correlated. Three approaches are therefore possible: generalized mixed effects models (GLMM), generalized estimating equations (GEE) or transition models (hence regressing on old values of the response). Zeger et al. [46] have shown the connection between the estimates in GEE and in the GLMM. It is important to note that the interpretation of the estimates in the three approaches is different, since the first (GLMM) is subject specific, the middle (GEE) is population averaged and unconditional and the latter is marginal but conditional on the history. In the following only GEE and transition models are considered.

5.3 Lorelogram analysis

The association between different weeks can be investigated by the lorelogram (see Diggle et al. p. 52 [15] and Heagerty and Zeger [19]), which is defined as

$$\mathrm{LOR}(t_j, t_k) = \log(\gamma(y_j, y_k)) \tag{5.8}$$
$$\gamma(y_1, y_2) = \frac{P(y_1 = 1, y_2 = 1)P(y_1 = 0, y_2 = 0)}{P(y_1 = 1, y_2 = 0)P(y_1 = 0, y_2 = 1)}$$

The LOR can be estimated by considering 2 by 2 tables for each individual, where the 2 by 2 table is a contingency table for observation with $t_j - t_k$ time-points apart ($t_j > t_k$). γ is seen to be larger than 1 for a positive relation between weeks u time-points apart, values around 1 correspond to no relation and values below 1 to negative relation. The numerator is the probability of staying in the same state and the denominator is the probability of different states. For tables containing cells with zero counts a correction is applied, which imply replacing 0 with 0.5. This is done in order to calculate the log odds-ratio, which otherwise would be $\pm\infty$ depending on whether the zero-count appears in the denominator or the numerator.

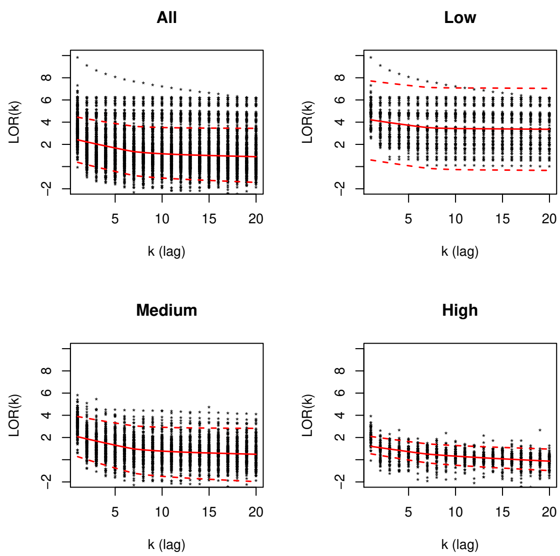

Figure 5.2: Lorelograms for week symptoms with fitted smoother for the mean (solid line) and lower and upper limit in a 95 % confidence interval (broken lines). The panels correspond to cluster found in section 3.3.3, where low is the group of children with the fewest symptoms and high the group with the most

The lorelograms for the weekly symptoms are shown in Figure 5.2, which shows that long-range correlations are present. The log odds-ratios are seen to be above 0 for $k \leq 9$. It is seen that the log odds-ratios decline the most from lag 1 to lag 7, where they flatten out for all three groups. From the lorelograms it is seen that the group with few symptoms have large odds-ratios, whereas the other two is somewhat lower. The high group is seen to go zero, whereas the medium group flattens out around 0.5 and the low group flattens out at 3.5.

The low group is characterized by the fact that the children have few symptoms, which imply that the response in this group mostly will be 0. The number of weeks overall in the low group with symptoms is 255 compared to 18553 weeks without ($\frac{255}{18553} = 0.01$). This imply that the probability of staying in the no-symptoms state is much higher compared any of the three other probabilities. The high odds-ratios for the low group are therefore caused by the fact that the no-symptoms weeks are highly correlated due to the sparse number of symptoms, which imply that $P(y_j = 0, y_k = 0)$ tends to dominate the γ-expression in (5.8). For the medium and high groups the ratio between weeks with symptoms and weeks without symptoms are 0.08 and 0.28, which shows that especially the high group is much more likely to shift states compared to the medium and low group. This is also seen from the odds-ratio, which is much lower compared to the low group.

The standard errors for the log odds-ratio can be approximated by the square-root of the inverse sum of the cell-numbers as illustrated by Bland and Altman [5],

$$\hat{\sigma}(LOR) = \sqrt{\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{1}{T_{ij}}} \tag{5.9}$$

where T_{ij} is the i'th row in the j'th column in the contingency-table from which the odds-ratio is calculated. A 95 % confidence interval is $\widehat{LOR} \pm 1.98 \cdot \hat{\sigma}(LOR)$. From Figure 5.2 it is seen that especially the lower group and medium group have broad confidence limits, whereas the high group's confidence interval is seen to be quite narrow. The uncertainties of the low and the medium groups are mainly caused be the fact the children tends to stay in the no-symptoms state, i.e. that T_{ij} is small for all other combination, which makes the overall sum large. It is seen that the low and medium group have confidence intervals for the LOR that contains 0 from $k \approx 2$ and the intervals are broad, which shows that the many 0 counts have a severe impact on the estimation of the LOR.

In Figure 5.3 four overall lorelograms are shown, overall in the sence that the individual are neglected and one lorelogram is estimated instead of m. From Figure 5.3 it is seen that the confidence bands are broadest for the low group, which is caused by the fact that 3 out of four table entries are small. The two other groups are seen to have much tighter bands. The difference between Figure 5.2 and Figure 5.3 illustrates the difference between the individual LOR's and the mean LOR's: The mean LOR's are more accurately estimated ($\hat{\sigma}(\overline{\mathrm{LOR}}) \approx \frac{\hat{\sigma}(\mathrm{LOR})}{\sqrt{n}}$). For the low group the impact of the high probability of being in the no symptoms state is reduced, since data from all individuals in the low group is used and more shifts therefore are seen.

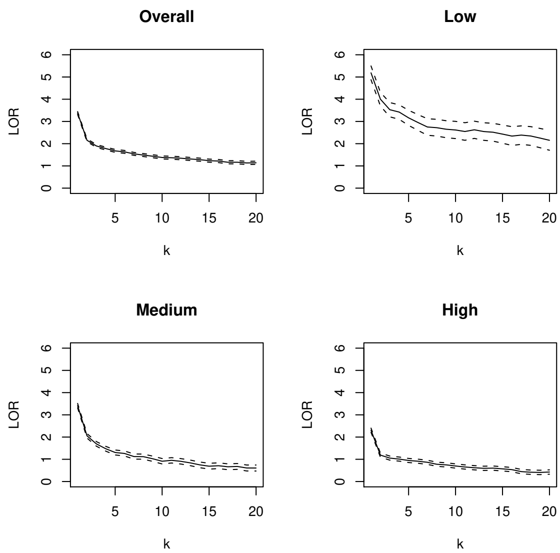

Figure 5.3: Lorelograms estimating overall correlation neglecting individuals: all clusters (upper left), low group (upper right), medium (lower left) and high group (lower right)

The analysis of the lorelograms shows that the group with few symptoms are highly correlated from week to week, which is caused by the fact that the children have no-symptoms in 99 % of the time. The high group is seen to have a significant correlation up to lags equal to 7 weeks, however the elbow is situated at $k = 2$, which indicates that the correlation is most pronounced for weeks up to two weeks apart. The medium group is seen to look like the low group, but has lower LORs but the same order of uncertainty caused by the many no-symptoms observations. A transition model should therefore contain group specific parameters with respect to the regression on lagged values in order to model the different transitions.

5.4 Marginal model

Based on the initial model statements and the analysis of the lorelogram models for the probability of having a week with an episode can be proposed. The response is bernoulli trials, which can be modelled with a binomial distribution having $n = 1$. The link-function is the logit of the probability of success (having an episode).

As explanatory variables $\cos(\text{week}/52 \cdot 2\pi)$ and $\sin(\text{week}/52 \cdot 2\pi)$ are included to account for the seasonal variations (as described in Chapter 2), age and age^2 are included to describe the temporal development related to age (as described in Chapter 3), the groups found in the gaussian LCR and finally the risk-factor PD15PtcO2 are included to describe the differences between individuals.

The probability of symptoms, p, is modelled through its link-function and the model becomes

$$\begin{aligned}\hat{\eta}_{ij} =& \beta_{0k} + \beta_{1k} \cdot \log_{10}(\mathrm{pd}_i) + \beta_{2k} \cdot \mathrm{age}_{ij} + \beta_{3k} \cdot \mathrm{age}_{ij}^2 \\ &+ \beta_{4k} \cdot \cos\left(\frac{\mathrm{week} \cdot 2\pi}{52}\right) + \beta_{5k} \cdot \sin\left(\frac{\mathrm{week} \cdot 2\pi}{52}\right) \\ &+ \beta_{6k} \cdot \mathrm{gender}_i \\ \eta =& \log\left(\frac{p}{1-p}\right)\end{aligned} \tag{5.10}$$

In the model, β_{1k} is the estimate of the parameter for PD15PtcO2 for the k'th group, β_{2k} the estimate for age for the k'th group, etc. In the basic generalized linear model, the observations are assumed to be being mutually independent, which from the lorelograms were seen to be a doubtful assumption. However as an initial analysis the model is estimated with the identity matrix as the assumed variance-covariance matrix corresponding to mutually independence. The season component is simplied to having the same amplitude and phaseshift for each year in order to keep the complexity down.

The estimated model is shown in Table 5.2, which shows that the seasonal part is the same for the three groups. The high group is seen to have the highest intercept and the highest benefit of a high PD15PtcO2 as seen in the previous analysis as well.

Figure 5.4 shows that the residuals from the generalized linear model are correlated within each individual. It is seen that the correlation is autoregressive and that adjacent weeks are correlated by 50 %.

5.4.1 GEE model

The analysis of the symptoms as being mutually independent within individual may however be wrong since correlation within individual is present. The correlation can be dealt with by using generalized estimation equations, where correlation is introduced as a working correlation. A correlation corresponding to an AR(1)-structure is assumed, since the lorelograms are exponentially decaying. The basis for the estimation is the model in (5.10), but with an AR(1) correlation for the within individual correlation.

$$\begin{aligned}\hat{\eta}_{ij} =& \beta_{0k} + \beta_{1k} \cdot \log_{10}(\mathrm{pd}_i) + \beta_{2k} \cdot \mathrm{age}_{ij} + \beta_{3k} \cdot \mathrm{age}_{ij}^2 \\ &+ \beta_{4k} \cdot \cos\left(\frac{\mathrm{week} \cdot 2\pi}{52}\right) + \beta_{5k} \cdot \sin\left(\frac{\mathrm{week} \cdot 2\pi}{52}\right) \\ &+ \beta_6 \cdot \mathrm{gender}_i \\ Corr(y_{ij}, y_{ik}) =& \rho^{|\mathrm{week}_j - \mathrm{week}_k|}\end{aligned} \tag{5.11}$$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.4033	0.1708	-19.9202	<0.0001
clusterMedium	0.8144	0.1827	4.4575	<0.0001
clusterHigh	0.8890	0.1889	4.7074	<0.0001
(log10(pd))	0.1731	0.0941	1.8391	0.0659
genderMale	-0.4628	0.1274	-3.6335	0.0003
age	-0.7672	0.1647	-4.6574	<0.0001
(age^2)	0.1404	0.0333	4.2220	<0.0001
cosine	0.6709	0.0972	6.9004	<0.0001
sine	0.2107	0.0927	2.2739	0.0230
clusterMedium:(log10(pd))	-0.2795	0.0981	-2.8491	0.0044
clusterHigh:(log10(pd))	-0.4650	0.0980	-4.7424	<0.0001
clusterMedium:genderMale	0.4413	0.1341	3.2901	0.0010
clusterHigh:genderMale	0.7557	0.1345	5.6178	<0.0001
clusterMedium:age	1.0113	0.1755	5.7630	<0.0001
clusterHigh:age	1.7115	0.1774	9.6492	<0.0001
clusterMedium:(age^2)	-0.2437	0.0358	-6.8147	<0.0001
clusterHigh:(age^2)	-0.3272	0.0357	-9.1704	<0.0001
clusterMedium:cosine	-0.0948	0.1021	-0.9284	0.3532
clusterHigh:cosine	-0.1887	0.1021	-1.8489	0.0645
clusterMedium:sine	-0.0338	0.0975	-0.3464	0.7290
clusterHigh:sine	-0.0809	0.0975	-0.8297	0.4067

Table 5.2: Summary of estimated parameters of simple generalized linear model

The summary for the parameters in the linear predictor is shown in Table 5.3, which shows that the standard errors for the parameters are larger compared to the model assuming independent observations within subject (Table 5.2).

The standard errors are increased, since the analysis of the mutually independent observations is incorrect, i.e. too much information is used from each observation. GEE accounts for the fact that the observations within each child are correlated and hence that observation j to some degree can be explained by the value of observation $j-1$. GEE gives robust standard errors in contrast to the naive standard errors obtained in the generalized linear model. Furthermore, GEE gives consistent estimates even with an incorrect correlation structure if the number of individuals is high, Diggle et al. p. 140 [15].

The procedure estimates a scale parameter of 0.9989, which shows that the model is adequate with respect to the expected variance-function (there are no excess heterogeneity among the observations). It is seen that the seasonal parts of the model for the three clusters are insignificantly different. Furthermore, the cluster difference for genders is seen to be insignificant. The model can be reduced to having the same seasonal risk and gender risk.

The updated summary is shown in Table 5.4, which shows that the model can not be reduced any further. The seasonal part of the model is seen to be the same for all children, which imply that it can be interpreted as a common background variable

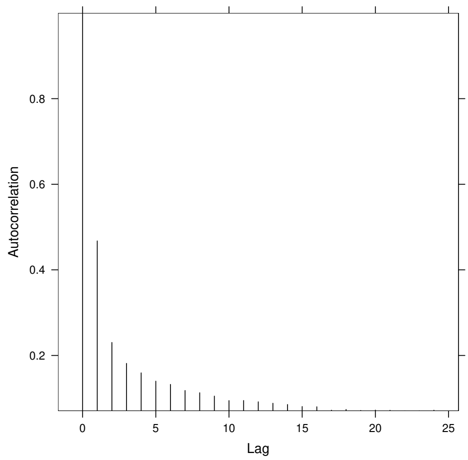

Figure 5.4: Autocorrelation for deviance residuals from initial GLM model, autocorrelation estimated only on observations belonging to the same individual.

that may explain symptoms related purely to the season. The model shows that a child coming from the high level group has the same odds-ratio of getting symptoms due to the season as a child coming from the low level group, whereas the difference in risk for the three groups is seen to be related to the longitudinal development.

Seasonal part

The seasonal part in Table 5.4 can be converted to a pure cosine-function with an amplitude and phaseshift using the relation

$$\begin{aligned} A \cdot \cos\left(\frac{\text{week} \cdot 2\pi}{52} + \theta\right) =& A \cdot \cos\left(\frac{\text{week} \cdot 2\pi}{52}\right) \cdot \cos(\theta) \\ & - A \cdot \sin\left(\frac{\text{week} \cdot 2\pi}{52}\right) \cdot \sin(\theta) \end{aligned} \tag{5.12}$$

This gives the following two equations to solve for θ and A

$$\begin{aligned} A \cdot \cos(\theta) &= \beta_4 \\ -A \cdot \sin(\theta) &= \beta_5 \end{aligned} \tag{5.13}$$

	Estimate	Std.err	Wald	p(>W)
(Intercept)	-3.4772	0.3813	83.1764	<0.0001
clusterMedium	0.8553	0.4031	4.5023	0.0338
clusterHigh	0.9239	0.4195	4.8516	0.0276
(log10(pd))	0.1737	0.1093	2.5263	0.1120
genderMale	-0.4764	0.3520	1.8318	0.1759
age	-0.6597	0.4173	2.4990	0.1139
(age^2)	0.1204	0.1221	0.9720	0.3242
cosine	0.6424	0.1524	17.7746	<0.0001
sine	0.2054	0.1563	1.7267	0.1888
clusterMedium:(log10(pd))	-0.2800	0.1327	4.4502	0.0349
clusterHigh:(log10(pd))	-0.4624	0.1528	9.1561	0.0025
clusterMedium:genderMale	0.4603	0.3698	1.5492	0.2133
clusterHigh:genderMale	0.7713	0.3931	3.8507	0.0497
clusterMedium:age	0.9196	0.4356	4.4564	0.0348
clusterHigh:age	1.6356	0.4350	14.1355	0.0002
clusterMedium:(age^2)	-0.2243	0.1265	3.1451	0.0762
clusterHigh:(age^2)	-0.3123	0.1247	6.2748	0.0122
clusterMedium:cosine	-0.0762	0.1603	0.2263	0.6342
clusterHigh:cosine	-0.1485	0.1602	0.8590	0.3540
clusterMedium:sine	-0.0167	0.1634	0.0104	0.9186
clusterHigh:sine	-0.0791	0.1669	0.2243	0.6358

Table 5.3: Summary of estimated parameters of GEE

	Estimate	Std.err	Wald	p(>W)
(Intercept)	-3.6894	0.2513	215.5298	<0.0001
clusterMedium	1.0663	0.2770	14.8125	0.0001
clusterHigh	1.2855	0.3008	18.2659	<0.0001
(log10(pd))	0.1509	0.1356	1.2399	0.2655
age	-0.6511	0.4212	2.3897	0.1221
(age^2)	0.1174	0.1222	0.9223	0.3369
cosine	0.5347	0.0343	242.4774	<0.0001
sine	0.1593	0.0366	18.8914	<0.0001
clusterMedium:(log10(pd))	-0.2572	0.1551	2.7498	0.0973
clusterHigh:(log10(pd))	-0.4263	0.1743	5.9791	0.0145
clusterMedium:age	0.9117	0.4392	4.3096	0.0379
clusterHigh:age	1.6246	0.4388	13.7084	0.0002
clusterMedium:(age^2)	-0.2211	0.1266	3.0531	0.0806
clusterHigh:(age^2)	-0.3081	0.1248	6.1007	0.0135

Table 5.4: Summary of estimated parameters of updated GEE

with respect to $A \geq 0 \quad \theta \in [-\pi, \pi]$. This leads to the solution

$$\hat{\theta} = \arctan\left(-\frac{\beta_5}{\beta_4}\right) = -0.29 = -16.82 \text{ days}$$
$$\hat{A} = \sqrt{\beta_4^2 + \beta_5^2} = 0.56 \tag{5.14}$$

It is seen that the seasonal part of the odds of symptoms tops at January 18th, which is seen to be within the range of the prevalence and incidence peaks considered in Chapter 2. The log odds-ratio from summer to winter is $2 \cdot A$, which gives an odds-ratio of $e^{2 \cdot A} = 3.05$. It is therefore seen to be 3 times as likely to have symptoms during winter compared to the summer regardless the level of symptoms.

In Figure 5.5 the ratio corresponding to the seasons effect on the odds is plotted as function of the calender year in weeks. A value below 1 imply that the odds is drawn down by the season and a value above the opposite. The season effect is negative (decreasing the odds) from week 15 to 41, which corresponds to the period from the beginning of March to the beginning of September and is obviously positive from week 41 to 15. The seasonal part shows that the risk of an episode is increased in the winter.

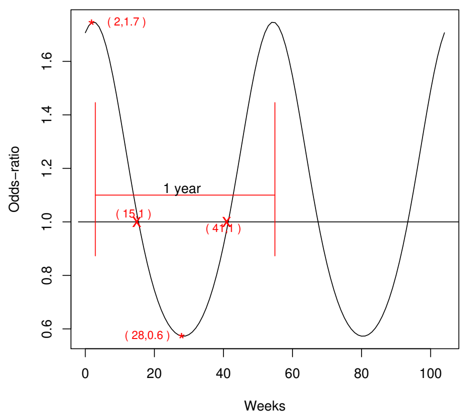

Figure 5.5: Odds-ratio related to the season as function of the week number (week 1 corresponds to first week of January)

Age part

The contribution to the odds coming from the age variables can be examined the same way as for the season. However the age variables depends on which group the child comes from, which imply that 3 curves for the age effect on the odds of getting a symptom are obtained from the model. Since the linear predictor is a second order polynomial with respect to age the effect on the odds becomes the right part of the product in

$$\left(\frac{p}{1-p}\right)_k = e^{\text{rest of model}} \cdot e^{\beta_{1k} \cdot \text{age}_j + \beta_{2k} \cdot \text{age}_j^2} \tag{5.15}$$

The amount $e^{\beta_{1k}\cdot\text{age}+\beta_{2k}\cdot\text{age}^2}$ is interpreted as the odds-ratio for comparing the current age (age = age$_j$) against age = 0 at the same time of year (since a contribution from the season would be added otherwise).

From Figure 5.6 it is seen that the high group has an odds-ratio, which is more than 1.5 from the age of 1 years and tops with a value of 3.5 at the age of 3 years. The low group has odds-ratios from age below 1, which correspond to a decrease in the odds compared to birth. Consistently decreasing effect is seen for the medium group for age $\geq$ 2 years after an initial increase the first year of life. The curvature for especially the low group is questionable since the standard error for the estimate is high. However the effect corresponding to age is seen to be small for the low and medium group and stable, whereas the high group both vary more over age and is larger than 1, which imply that the risk of symptoms is increasing to the age of 3 years and in general more likely compared to the starting level.

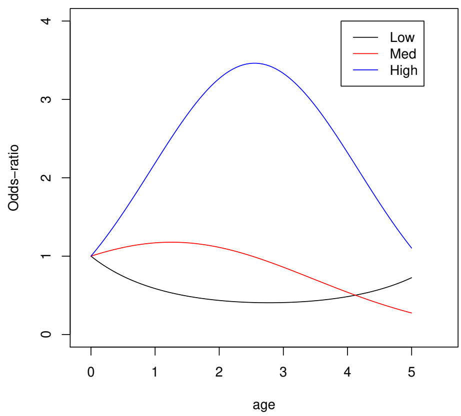

Figure 5.6: Odds-ratio of age = age$_j$ compared to age = 0

Correlation and residuals

The estimated correlation parameter is 0.5841 with a standard error of 0.1249. This imply that observations on the same individual are correlated with by amount

$$Corr(y_{ij}, y_{ik}) = \rho^{|\text{week}_{ij}-\text{week}_{ik}|} = 0.5841^{|\text{week}_{ij}-\text{week}_{ik}|} \qquad (5.16)$$

The correlation is seen to decay rapidly as the time-difference increases. This was seen to be the case in the lorelograms as well, i.e. that the elbow was seen for $k = 2$. The estimated correlation is seen to be 0.3412 for observation taken on the same individual two weeks apart.

In Figure 5.7 the qq-plot for the pearson residuals is plotted, which shows that the pearson residuals seems normal but coming from two different distributions corresponding to whether the response was 0 or 1. This is caused by the way the pearson residuals behave in the bernoulli trial. The Pearson residuals are defined as

$$r_i = \frac{y_i - \hat{y}_i}{\sqrt{\hat{V}[\hat{y}_i]}} \tag{5.17}$$

and since $y_i \in \{0, 1\}$ and $\hat{y}_i \in [0, 1]$ the sign of the residuals is seen to be dependent of the value of y_i. For $y_i = 0$ the residuals are restricted to negative values and for $y_i = 1$ to positive. The QQ-plot serves as a outlier detection tool and shows that no outliers seem present.

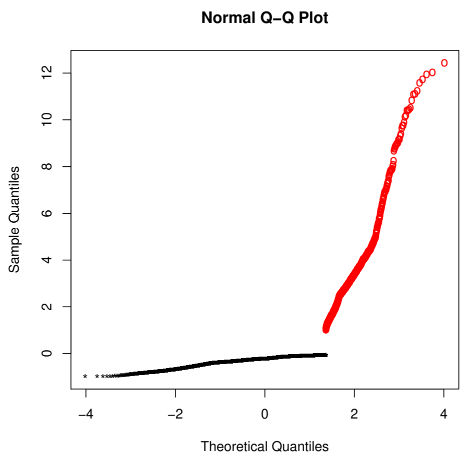

Figure 5.7: QQ-plot of Pearson residuals. The black points correspond to the weeks with no symptoms and the red to the weeks with symptoms

Diagnosis

In the GEE-model the groups obtained from the gaussian LCR were used, in the following the diagnoses for the asthma status are used instead as grouping factor. This imply that the initial GEE-model must be reestimated to evaluate the differences between the two groups in terms of parameters estimates. The initial model is

$$\begin{aligned}\hat{\eta}_{ij} =& \beta_{0k} + \beta_{1k} \cdot \log_{10}(\text{pd}_i) + \beta_{2k} \cdot \text{age}_{ij} + \beta_{3k} \cdot \text{age}_{ij}^2 \\ &+ \beta_{4k} \cdot \cos\left(\frac{\text{week} \cdot 2\pi}{52}\right) + \beta_{5k} \cdot \sin\left(\frac{\text{week} \cdot 2\pi}{52}\right) \\ &+ \beta_6 \cdot \text{gender}_i \\ \eta =& \log\left(\frac{p}{1-p}\right) \qquad Corr(y_{ij}, y_{ik}) = \rho^{|\text{week}_j - \text{week}_k|}\end{aligned} \tag{5.18}$$

where $k \in \{\text{Non-asthma}, \text{Asthma}\}$. The model is estimated and the summary in Table 5.5 shows that the model can be reduced to the model in Table 5.6.

	estimate	san.se	wald	p
(Intercept)	-2.8410	0.1488	364.7708	<0.0001
diagnosisAsthma	0.5736	0.2737	4.3918	0.0361
(log10(pd))	-0.1542	0.1189	1.6819	0.1947
genderMale	-0.1328	0.1549	0.7345	0.3914
age	0.2717	0.1507	3.2497	0.0714
(age^2)	-0.1102	0.0393	7.8650	0.0050
cosine	0.5592	0.0455	150.9258	<0.0001
sine	0.2193	0.0509	18.5835	<0.0001
diagnosisAsthma:(log10(pd))	0.0328	0.1818	0.0327	0.8566
diagnosisAsthma:genderMale	0.2968	0.2634	1.2697	0.2598
diagnosisAsthma:age	0.5105	0.1996	6.5423	0.0105
diagnosisAsthma:(age^2)	-0.0299	0.0472	0.4006	0.5268
diagnosisAsthma:cosine	-0.0944	0.0760	1.5425	0.2142
diagnosisAsthma:sine	-0.1281	0.0803	2.5409	0.1109

Table 5.5: Initial gee with grouping by diagnosis

Furthermore the scale-parameters corresponding to the over-dispersions are estimated to 1.0073 and 1.0069 in the full-model and the reduced model, respectively. The reduced model is seen to be adequate and the updated summary is shown in Table 5.6. Compared to the model with the grouping obtained from the gaussian LCR, the new model estimates a common curvature parameter, whereas the previous model had a highly significant different estimate for the high group. The model with the asthma diagnosis is expanded in the following to see differences between asthmatic and non-asthmatic children.

5.5 Medication

Based on the reduced GEE model the medication variables as considered in section 4.8 can be analyzed. The model considered until now is estimating the temporal patterns, the seasonal pattern and the difference between children having different diagnosis

	estimate	san.se	wald	p
(Intercept)	-2.9376	0.1145	658.5110	<0.0001
(log10(pd))	-0.1380	0.0901	2.3455	0.1256
diagnosisAsthma	0.8018	0.2119	14.3193	0.0002
age	0.3393	0.0928	13.3582	0.0003
(age^2)	-0.1250	0.0245	26.0260	<0.0001
cosine	0.5144	0.0373	189.8584	<0.0001
sine	0.1593	0.0398	16.0382	<0.0001
diagnosisAsthma:age	0.3672	0.0760	23.3566	<0.0001

Table 5.6: Reduced gee with grouping by diagnosis

and is given by

$$\begin{aligned}\hat{\eta}_{ij} =& \beta_{0k} + \beta_{1k} \cdot \log_{10}(\mathrm{pd}_i) + \beta_{2k} \cdot \mathrm{age}_{ij} + \beta_3 \cdot \mathrm{age}_{ij}^2 \\ &+ \beta_4 \cdot \cos\left(\frac{\mathrm{week} \cdot 2\pi}{52}\right) + \beta_5 \cdot \sin\left(\frac{\mathrm{week} \cdot 2\pi}{52}\right) \\ Corr(y_{ij}, y_{ik}) &= \rho^{|\mathrm{week}_j - \mathrm{week}_k|}\end{aligned} \tag{5.19}$$

where $k \in \{\text{Non-asthma}, \text{Asthma}\}$. The model is seen to have a common seasonal part, which can be interpreted as a risk purely related to seasonal variations. In this model the medication variables can be included. The fast dynamics may reveal the positive effect of the medication, which was seen not to be the case for the yearly aggregated symptoms. The level of medication and the level of symptoms is positive correlated in the yearly aggregated analysis, which may change with the faster dynamics.

In Table 5.7 cross-tabulations of medication types and the episodes are shown. It is seen that placebo and budesonide are positively correlated with having symptoms, which is caused by the fact that the medication in the initial trial is symptom initiated: A two week treatment period is applied for each symptom period. The odds-ratios are 23.08 and 19.51 for receiving placebo and budesonide vs. not receiving this treatment, respectively. Prednisolon is given only in weeks with symptoms, since it is used against acute severe asthma symptoms and therefore given in weeks with symptoms. For weeks with spirocort treatment 69 % of the weeks are symptom free and the odds-ratio is 7.25

One can furthermore analyze the more long term implications of medication, namely given that a child has been treated with drug k in the previous week, which is the probability of having an episode this week. In Table 5.8 lagged values of the medication data is tabulated against the episodes. It is seen that the proportion of weeks without episodes given that the child was medicated the previous week is increased compared to the direct tabulation in Table 5.7.

One needs to evaluate whether the one week lagged values of medication are related to the episodes or if the shifts from Table 5.7 to Table 5.8 is a result of the natural shifts from week to week. To analyze the results, a cross-tabulation of the lagged

Type	Medication	Episode (Y_{ij}) 0	1	total
	0	45063	3712	48775
placebo		(0.92)	(0.08)	(1.00)
	1	435	827	1262
		(0.34)	(0.66)	(1.00)
	0	44947	3663	48610
budesonide		(0.92)	(0.08)	(1.00)
	1	551	876	1427
		(0.39)	(0.61)	(1.00)
	0	62370	3880	66250
spirocort		(0.94)	(0.06)	(1.00)
	1	5004	2258	7262
		(0.69)	(0.31)	(1.00)
	0	67374	6066	73440
prednisolon		(0.92)	(0.08)	(1.00)
	1	0	72	72
		(0.00)	(1.00)	(1.00)

Table 5.7: Cross-tabulation of medication types and episodes, i.e. a (0,1) combination for a given medication type and the episodes imply that the drug have not been used and that the child had an episode that week. The two first medication types are only given i nthe first three years of life, whereas the other two has been given in the whole period, but primarily the latter part of the period. The table has counts and (row-percents).

medication and lagged episode and current episode is shown in Table 5.9. It is seen that the distribution of the current symptoms is similar regardless of the medication the previous week, i.e. the lagged medication seems not to be related to the symptoms.

Obviously one need to access whether the correspondence between medication and episode is different for asthmatic children and non-asthmatic children. This can be done by expanding Table 5.7 and 5.9 with the diagnosis. Table 5.10 and 5.11 correspond to the expansion with diagnosis. It is seen from Table 5.10 that the asthmatic group is more likely to have an episode regardless of the medication status. Furthermore, it is seen that a week with some kind of medication has a higher proportion of symptoms-weeks. This may be caused by the close relationship between symptoms and medication, i.e. that only children with symptoms receives medication.

Table 5.11 shows the lagged medication for the two types of asthma status and given the symptoms status the previous week. It is seen that the non-asthmatic group has a high probability of keeping the no-symptoms status, which is decreased if medication has been given the previous week. The asthmatic group is seen to have a higher proportion of symptom-weeks, however the response on medication is seen to be an increase in the proportion of symptoms.

Type	Medication_{ij-1}	Episode (Y_{ij}) 0	1	total
placebo	0	44381	4046	48427
		(0.92)	(0.08)	(1.00)
	1	764	491	1255
		(0.61)	(0.39)	(1.00)
budesonide	0	44235	4025	48260
		(0.92)	(0.08)	(1.00)
	1	910	512	1422
		(0.64)	(0.36)	(1.00)
spirocort	0	61859	4075	65934
		(0.94)	(0.06)	(1.00)
	1	5160	2061	7221
		(0.71)	(0.29)	(1.00)
prednisolon	0	66994	6090	73084
		(0.92)	(0.08)	(1.00)
	1	25	46	71
		(0.35)	(0.65)	(1.00)

Table 5.8: Cross-tabulation of one week lagged medication types and episodes, i.e. a (0,1) combination for a given medication type and the episodes imply that the drug have not been used the previous week and that the child had an episode that week. The two first medication types are only given in the first three years of life, whereas the other two has been given in the whole period, but primarily the latter part of the period. The table has counts and (row-percents).

Type	Medication$_{ij-1}$	Y_{ij-1}	Episode (Y_{ij}) 0	1	total
	0	0	42768	1929	44697
placebo			(0.96)	(0.04)	(1.00)
	1	0	382	52	434
			(0.88)	(0.12)	(1.00)
	0	1	1577	2111	3688
			(0.43)	(0.57)	(1.00)
	1	1	381	439	820
			(0.46)	(0.54)	(1.00)
	0	0	42657	1925	44582
budesonide			(0.96)	(0.04)	(1.00)
	1	0	493	56	549
			(0.90)	(0.10)	(1.00)
	0	1	1542	2094	3636
			(0.42)	(0.58)	(1.00)
	1	1	416	456	872
			(0.48)	(0.52)	(1.00)
	0	0	59967	1984	61951
spirocort			(0.97)	(0.03)	(1.00)
	1	0	4218	751	4969
			(0.85)	(0.15)	(1.00)
	0	1	1772	2082	3854
			(0.46)	(0.54)	(1.00)
	1	1	934	1308	2242
			(0.42)	(0.58)	(1.00)
	0	0	64185	2735	66920
prednisolon			(0.96)	(0.04)	(1.00)
	1	0	0	0	0
			-	-	-
	0	1	2681	3344	6025
			(0.44)	(0.56)	(1.00)
	1	1	25	46	71
			(0.35)	(0.65)	(1.00)

Table 5.9: Cross-tabulation of one week lagged medication types and lagged and current episodes, i.e. a (0,0,1) combination for a given medication type and the episodes imply that the drug have not been used the previous week, that no symptoms was recorded the previous week and that the child had an episode that week. The two first medication types are only given in the first three years of life, whereas the other two has been given in the whole period, but primarily the latter part of the period. The table has counts and (row-percents).

Diagnosis	Type	Medication	Episode (Y_{ij}) 0	1	total
Non-asthma	placebo	0	29942	1677	31619
			(0.95)	(0.05)	(1.00)
		1	281	444	725
			(0.39)	(0.61)	(1.00)
	budesonide	0	29894	1596	31490
			(0.95)	(0.05)	(1.00)
		1	329	525	854
			(0.39)	(0.61)	(1.00)
	spirocort	0	45861	2235	48096
			(0.95)	(0.05)	(1.00)
		1	1121	444	1565
			(0.72)	(0.28)	(1.00)
	prednisolon	0	46982	2660	49642
			(0.95)	(0.05)	(1.00)
		1	0	19	19
			(0.00)	(1.00)	(1.00)
Asthma	placebo	0	5375	1398	6773
			(0.79)	(0.21)	(1.00)
		1	64	211	275
			(0.23)	(0.77)	(1.00)
	budesonide	0	5335	1419	6754
			(0.79)	(0.21)	(1.00)
		1	104	190	294
			(0.35)	(0.65)	(1.00)
	spirocort	0	5180	907	6087
			(0.85)	(0.15)	(1.00)
		1	3614	1674	5288
			(0.68)	(0.32)	(1.00)
	prednisolon	0	8794	2534	11328
			(0.78)	(0.22)	(1.00)
		1	0	47	47
			(0.00)	(1.00)	(1.00)

Table 5.10: Cross-tabulation of medication types and episodes for the two types of asthma status, i.e. a (0,1) combination for a given medication type and the episodes imply that the drug have not been used and that the child had an episode that week. The two first medication types are only given in the first three years of life, whereas the other two has been given in the whole period, but primarily the latter part of the period. The table has counts and (row-percents).

Diagnosis	Type	Medication_{ij-1}	Y_{ij-1}	Episode (Y_{ij}) 0	1	total
Non-asthma	placebo	0	0	(0.97)	(0.03)	(1.00)
		1	0	(0.88)	(0.12)	(1.00)
		0	1	(0.45)	(0.55)	(1.00)
		1	1	(0.54)	(0.46)	(1.00)
	budesonide	0	0	(0.97)	(0.03)	(1.00)
		1	0	(0.91)	(0.09)	(1.00)
		0	1	(0.46)	(0.54)	(1.00)
		1	1	(0.50)	(0.50)	(1.00)
	spirocort	0	0	(0.97)	(0.03)	(1.00)
		1	0	(0.88)	(0.12)	(1.00)
		0	1	(0.48)	(0.52)	(1.00)
		1	1	(0.46)	(0.54)	(1.00)
	prednisolon	0	0	(0.97)	(0.03)	(1.00)
		1	0	(NaN)	(NaN)	(NaN)
		0	1	(0.48)	(0.52)	(1.00)
		1	1	(0.50)	(0.50)	(1.00)
Asthma	placebo	0	0	(0.89)	(0.11)	(1.00)
		1	0	(0.81)	(0.19)	(1.00)
		0	1	(0.38)	(0.62)	(1.00)
		1	1	(0.29)	(0.71)	(1.00)
	budesonide	0	0	(0.89)	(0.11)	(1.00)
		1	0	(0.87)	(0.13)	(1.00)
		0	1	(0.37)	(0.63)	(1.00)
		1	1	(0.38)	(0.62)	(1.00)
	spirocort	0	0	(0.91)	(0.09)	(1.00)
		1	0	(0.84)	(0.16)	(1.00)
		0	1	(0.37)	(0.63)	(1.00)
		1	1	(0.41)	(0.59)	(1.00)
	prednisolon	0	0	(0.88)	(0.12)	(1.00)
		1	0	(NaN)	(NaN)	(NaN)
		0	1	(0.40)	(0.60)	(1.00)
		1	1	(0.32)	(0.68)	(1.00)

Table 5.11: Cross-tabulation of one week lagged medication types and lagged and current episodes for the two types of asthma status, i.e. a (0,0,1) combination for a given medication type and the episodes imply that the drug have not been used the previous week, that no symptoms was recorded the previous week and that the child had an episode that week. The two first medication types are only given in the first three years of life, whereas the other two has been given in the whole period, but primarily the latter part of the period. The table has only (row-percents).

5.5.1 Initial GEE estimation medication

As an initial analysis two models are considered: One for the nested trial with budesonide/placebo treatment and one for the prednisolon and spirocort treatment. GEE is applied to account for the correlation along the individual time-series. The lorelograms (section 5.3) indicates that an AR(1) correlation structure might be adequate for describing the correlation.

Initial trial medication

The model for the probability of having an episode for the medication types in the initial trial can based on the previous analysis be formulated as

$$\begin{aligned}\hat{\eta}_{ij} =& \beta_{0k} + \beta_{1k}\cdot\log_{10}(\mathrm{pd}_i) + \beta_{2k}\cdot\mathrm{age}_{ij} + \beta_3\cdot\mathrm{age}_{ij}^2\\ &+ \beta_4\cdot\cos\left(\frac{\mathrm{week}\cdot 2\pi}{52}\right) + \beta_5\cdot\sin\left(\frac{\mathrm{week}\cdot 2\pi}{52}\right)\\ &+ \beta_{6k}\cdot\mathrm{budesonide}_{ij} + \beta_{7k}\cdot\mathrm{placebo}_{ij}\\ \eta =& \log\left(\frac{p}{1-p}\right) \qquad Corr(y_{ij}, y_{ik}) = \rho^{|\mathrm{week}_j - \mathrm{week}_k|}\end{aligned} \tag{5.20}$$

where $k \in \{\text{Non-asthma}, \text{Asthma}\}$. It is seen that each group has an individual parameter for both placebo, budesonide and age. Furthermore, the groups start at different levels to model the possible different baseline risks.

	estimate	san.se	wald	p
(Intercept)	-3.9941	0.1353	871.3815	<0.0001
(log10(pd))	-0.1778	0.1004	3.1365	0.0766
diagnosisAsthma	0.6157	0.2400	6.5789	0.0103
age	1.5544	0.1837	71.5790	<0.0001
(age^2)	-0.5352	0.0617	75.2056	<0.0001
cosine	0.4839	0.0390	153.8472	<0.0001
sine	0.1409	0.0420	11.2328	0.0008
budesonid1	2.4343	0.1770	189.1358	<0.0001
placebo1	2.1706	0.1658	171.3045	<0.0001
diagnosisAsthma:age	0.6094	0.1069	32.5203	<0.0001
diagnosisAsthma:budesonid1	-0.6934	0.2374	8.5282	0.0035
diagnosisAsthma:placebo1	-0.0999	0.2598	0.1478	0.7006

Table 5.12: Summary for GEE-model with initial trial medication predictor

The summary for the model is shown in Table 5.12, which shows that the effect of placebo is the same for non-asthmatic and asthmatic children. The model can therefore be reduced to a model with one parameter for placebo treatment, which is shown in Table 5.13.

The updated summary shows that placebo treatment increases the risk of having an episode in the same week, which is expected since the treatment is initiated in weeks with symptoms. For the active treatment, budesonide, a difference is seen between the two groups. The odds-ratio is twice as large for the non-asthmatic group compared to the asthmatic group. However, both groups have an increased risk of symptoms in week with active medication, which is explained the same way as for the placebo treatment. The summary indicates that asthmatic children respond more positive on the treatment compared to the non-asthmatic children, since the odds-ratio is lower.

The analysis of the initial medication highlights an important problem, namely that using the medication in week j as predictor for the probability of symptoms in the same week gives a non-causal model. Medication in week j is given due to symptoms in week j or $j-1$, since the trial medication is given in two weeks periods from the third day with symptoms. This imply that lagged values of medication may be more appropiate, which correspond to analyzing the effect of the treatment given the previous week.

	estimate	san.se	wald	p	OR
(Intercept)	-3.9895	0.1343	882.0777	<0.0001	0.0185
(log10(pd))	-0.1787	0.1003	3.1746	0.0748	0.8364
diagnosisAsthma	0.5932	0.2319	6.5427	0.0105	1.8097
age	1.5569	0.1827	72.6383	<0.0001	4.7441
(age^2)	-0.5366	0.0613	76.6032	<0.0001	0.5847
cosine	0.4844	0.0390	153.9708	<0.0001	1.6232
sine	0.1410	0.0420	11.2964	0.0008	1.1514
budesonid1	2.4319	0.1755	192.1019	<0.0001	11.3807
placebo1	2.1468	0.1321	264.0398	<0.0001	8.5571
diagnosisAsthma:age	0.6167	0.1080	32.6193	<0.0001	1.8529
diagnosisAsthma:budesonid1	-0.6875	0.2344	8.5992	0.0034	0.5028

Table 5.13: Summary for GEE-model with initial trial medication predictor reduced to only one parameter for placebo treatment

Prednisolon and spirocort

As for the initial analysis of the medication, a model for the use of prednisolon and spirocort can be formulated as

$$\begin{aligned} \hat{\eta}_{ij} =& \beta_{0k} + \beta_{1k} \cdot \log_{10}(\text{pd}_i) + \beta_{2k} \cdot \text{age}_{ij} + \beta_3 \cdot \text{age}^2_{ij} \\ &+ \beta_4 \cdot \cos\left(\frac{\text{week} \cdot 2\pi}{52}\right) + \beta_5 \cdot \sin\left(\frac{\text{week} \cdot 2\pi}{52}\right) \\ &+ \beta_{8k} \cdot \text{spirocort}_{ij} + \beta_{9k} \cdot \text{prednisolon}_{ij} \\ Corr&(y_{ij}, y_{ik}) = \rho^{|\text{week}_j - \text{week}_k|} \end{aligned} \tag{5.21}$$

where $k \in \{\text{Non-asthma}, \text{Asthma}\}$. However since Table 5.7 shows that the observed probability of having symptoms in a week with prednisolon treatment is 1, the parameters corresponding to prednisolon will be estimated to values $\rightarrow \infty$, since this correspond to $\mu = p \rightarrow 1$. The estimates are odd, but are indeed reflecting the data.

	estimate	san.se	wald	p
(Intercept)	-2.8670	0.1056	737.5000	<0.0001
(log10(pd))	-0.1089	0.0743	2.1520	0.1424
diagnosisAsthma	0.8471	0.1882	20.2500	<0.0001
age	0.0852	0.0928	0.8433	0.3585
(age^2)	-0.0705	0.0241	8.5740	0.0034
cosine	0.5199	0.0392	176.1000	<0.0001
sine	0.1411	0.0438	10.3900	0.0013
spirocort1	1.8110	0.1815	99.6100	<0.0001
prednisolon1	$4.5{\cdot}10^{15}$	$1373{\cdot}10^{4}$	$1.0{\cdot}10^{17}$	<0.0001
diagnosisAsthma:age	0.2193	0.0760	8.3150	0.0039
diagnosisAsthma:spirocort1	-0.9343	0.2177	18.4300	<0.0001
diagnosisAsthma:prednisolon1	$-55240{\cdot}10^{4}$	$1964{\cdot}10^{4}$	791.0000	<0.0001

Table 5.14: Summary for GEE-model with spirocort and prednisolon as part of the linear predictor

The summary for the model with both spirocort and prednisolon in shown in Table 5.14, which shows that the estimates for prednisolon are large as expected. It is furthermore seen that the odds-ratio for spirocort treatment for both groups are above 1 as seen for placebo and budesonide treatment. Reducing the model by removing the prednisolon treatment gives almost the same estimates for the remaining parameters as seen from Table 5.15.

	estimate	san.se	wald	p	OR
(Intercept)	-2.8903	0.1054	751.9751	<0.0001	0.0556
(log10(pd))	-0.1116	0.0723	2.3795	0.1229	0.8944
diagnosisAsthma	0.8609	0.1848	21.6953	<0.0001	2.3653
age	0.1125	0.0916	1.5095	0.2192	1.1191
(age^2)	-0.0764	0.0241	10.1014	0.0015	0.9264
cosine	0.5204	0.0382	185.5444	<0.0001	1.6826
sine	0.1392	0.0426	10.6540	0.0011	1.1493
spirocort1	1.8664	0.1814	105.8586	<0.0001	6.4648
diagnosisAsthma:age	0.2132	0.0743	8.2271	0.0041	1.2376
diagnosisAsthma:spirocort1	-0.9649	0.2178	19.6312	<0.0001	0.3810

Table 5.15: Summary for GEE-model with spirocort as part of the linear predictor

5.5.2 Lagged medication effects

From the models summarized in Table 5.13 and 5.15 it is seen that the effect of the medication has an awkward interpretation. Using the medication status the previous week could be a way of avoiding the close link between symptoms and medication.

For the initial medication both budesonide and placebo lead to decrease in risk for the non-asthmatic group, whereas the asthmatic group only benefits from treatment with budesonide. However none of the estimates are significant, which imply that receiving the initial treatment the previous week has no effect on the risk in the current week. For spirocort treatment the risk is increased for the asthmatic and non-asthmatic groups. It is furthermore seen that the difference between the groups is insignificant, which shows that the medication gives no additional description of the risk of an episode.

	estimate	san.se	wald	p	OR
(Intercept)	-3.6624	0.1431	655.3924	<0.0001	0.0257
(log10(pd))	-0.1830	0.1038	3.1086	0.0779	0.8327
diagnosisAsthma	0.4898	0.2386	4.2120	0.0401	1.6320
age	1.8532	0.1901	95.0393	<0.0001	6.3805
(age^2)	-0.6740	0.0657	105.1928	<0.0001	0.5097
cosine	0.5852	0.0437	179.3367	<0.0001	1.7953
sine	0.2016	0.0476	17.9502	<0.0001	1.2234
budelag1	-0.5993	0.5010	1.4308	0.2316	0.5492
placebolag1	-0.6142	0.4801	1.6362	0.2008	0.5411
diagnosisAsthma:age	0.5798	0.1195	23.5430	<0.0001	1.7856
diagnosisAsthma:budelag1	0.4713	0.5945	0.6285	0.4279	1.6021
diagnosisAsthma:placebolag1	0.9667	0.5198	3.4595	0.0629	2.6294

Table 5.16: Summary for GEE-model with lagged values of budesonide and placebo as part of the linear predictor

	estimate	san.se	wald	p	OR
(Intercept)	-2.9171	0.1124	673.9138	<0.0001	0.0541
(log10(pd))	-0.1324	0.0832	2.5318	0.1116	0.8759
diagnosisAsthma	0.8459	0.1998	17.9263	<0.0001	2.3302
age	0.2548	0.0918	7.7091	0.0055	1.2902
(age^2)	-0.1078	0.0243	19.6598	<0.0001	0.8978
cosine	0.5140	0.0376	186.8388	<0.0001	1.6719
sine	0.1502	0.0398	14.2076	0.0002	1.1621
spirolag1	0.7632	0.2759	7.6539	0.0057	2.1451
diagnosisAsthma:age	0.3249	0.0768	17.9071	<0.0001	1.3838
diagnosisAsthma:spirolag1	-0.5247	0.3044	2.9717	0.0847	0.5918

Table 5.17: Summary for GEE-model with lagged spirocort as part of the linear predictor

5.6 Risk-factors for weekly episodes

In the following risk-factors recorded at and prior to the birth of the COPSAC children and risk-factors recorded at the age of 3 years are analyzed with the basis in the model formulated in (5.18). The model is changed to having a common seasonal part for the two groups: Asthma and non-asthma. In the analysis of the risk-factors and confounders the model with season, age and diagnosis correction is used as the basis, i.e. the model is expanded with the risk-factors. This is done in order to analyze the risk-factors in a framework where background variations have been eliminated.

A model with a common season part and a part related to age, which differ from asthmatic to non-asthmatic children, and the congenital PD15 PtcO2 as explanatory variables was previous seen to be an adequate description of the background variation. This model is therefore used as the basis for the following analysis of risk-factors and confounders and is formulated as

$$\begin{aligned}\eta_{ij} =& \mu_{ij} + \beta_{0k} + \beta_{1k} \cdot \mathrm{age}_{ij} + \beta_2 \cdot \mathrm{age}_{ij}^2 \\ &+ \beta_3 \cdot \cos(\mathrm{week}_{ij} \cdot 2\pi/52) + \beta_4 \cdot \sin(\mathrm{week}_{ij} \cdot 2\pi/52) \\ &+ \beta_5 \cdot \log_{10}(\mathrm{pd}_i) + \text{risk-factors} \\ Corr(y_{ij}, y_{ij'}) =& \rho^{|\mathrm{week}_{ij} - \mathrm{week}_{ij'}|} \quad \eta = \log\left(\frac{p}{1-p}\right)\end{aligned} \tag{5.22}$$

where $k \in \{\text{non-asthma, asthma}\}$. The correlation between observations for the same individual is autoregressive of order 1, which is seen to be a reasonable structure from the lorelograms in section 5.3. For a more elaborate description of the missing risk-factor see section 4.9.

As an initial model, risk-factors and their interaction with the diagnosis to model differences in the two groups can be considered. This gives a risk-factor part of the model, which is given as

$$\begin{aligned}\text{risk-factors} =& \beta_{6k} \cdot \mathrm{fillagrin}_i + \beta_{7k} \cdot \mathrm{smoking}_{\mathrm{3rd},i} + \beta_{8k} \cdot \mathrm{smoking}_{\mathrm{home},i} \\ &+ \beta_{9k} \cdot \mathrm{smoking}_{\mathrm{daycare},i} + \beta_{10k} \cdot \mathrm{gender}_i \\ &+ \beta_{11k} \cdot \text{father asthma}_i + \beta_{12k} \cdot \mathrm{pets}_{\mathrm{3rd},i} + \beta_{13k} \cdot \mathrm{alcohol}_{\mathrm{3rd},i} \\ &+ \beta_{14k} \cdot \mathrm{smoking}_{\mathrm{3rd},i} + \beta_{15k} \cdot \mathrm{pets}_{\mathrm{home},i}\end{aligned} \tag{5.23}$$

where all variables are indicator variable, i.e. indicating whether the child is boy or girl, whether the mother has smoked in the third trimester or not, etc. The variable alcohol has four levels, 0, 1, 2 and ≥ 3 corresponding to the alcohol intake in the third trimester of the pregnancy in units. The variable smoking in the home is divided into four levels: 0, $]0, 100]$, $]100, 200]$ and > 200 days per year.

The summary of the estimated model is shown in Table 5.18, which shows that all interactions between diagnosis and risk-factors, besides alcohol, are insignificant. However the significant interaction seems rather odd, since an alcohol intake of two

	estimate	san.se	wald	p	OR
(Intercept)	-2.1696	0.3143	47.6528	<0.0001	0.1142
cosine	0.5057	0.0387	170.6358	<0.0001	1.6582
sine	0.1634	0.0422	15.0078	0.0001	1.1775
diagnAsthma	-0.0382	0.6578	0.0034	0.9537	0.9625
age	0.3210	0.1030	9.7176	0.0018	1.3785
(age^2)	-0.1229	0.0271	20.5683	<0.0001	0.8843
log10(pd)	-0.1493	0.1159	1.6588	0.1978	0.8613
SEXMale	-0.1500	0.1586	0.8950	0.3441	0.8607
$smoking_{3rd}$Yes	-0.1354	0.2843	0.2268	0.6339	0.8734
$father_{ashtma}$Yes	0.3208	0.2037	2.4801	0.1153	1.3782
$pets_{3rd}$	-0.2493	0.2795	0.7952	0.3725	0.7794
alcohol1	0.0079	0.2522	0.0010	0.9749	1.0080
alcohol2	-0.0422	0.4638	0.0083	0.9275	0.9587
alcohol3	-0.6240	1.0745	0.3373	0.5614	0.5358
fillagrin1	0.4233	0.2871	2.1747	0.1403	1.5271
$smoking_{home}$1<100	-0.7348	0.3000	5.9992	0.0143	0.4796
$smoking_{home}$1[100,200]	-0.6807	0.5041	1.8236	0.1769	0.5063
$smoking_{home}$1>200	-0.6100	0.3913	2.4305	0.1190	0.5433
$smoking_{daycare}$1	-0.1662	0.1776	0.8752	0.3495	0.8469
$pets_{home}$	0.0005	0.0009	0.2771	0.5986	1.0005
diagnAsthma:age	0.3653	0.0837	19.0422	<0.0001	1.4409
diagnAsthma:log10(pd)	0.0249	0.2115	0.0139	0.9061	1.0253
diagnAsthma:SEXMale	0.4704	0.3043	2.3892	0.1222	1.6006
diagnAsthma:$smoking_{3rd}$Yes	0.2737	0.5314	0.2653	0.6065	1.3148
diagnAsthma:$father_{ashtma}$Yes	-0.1096	0.3097	0.1253	0.7233	0.8962
diagnAsthma:$pets_{3rd}$	0.4846	0.6408	0.5719	0.4495	1.6235
diagnAsthma:alcohol1	-0.3241	0.4241	0.5839	0.4448	0.7232
diagnAsthma:alcohol2	2.1413	0.5046	18.0097	<0.0001	8.5103
diagnAsthma:alcohol3	0.9375	1.3057	0.5156	0.4727	2.5537
diagnAsthma:fillagrin1	0.1069	0.5082	0.0442	0.8334	1.1128
diagnAsthma:$smoking_{home}$1<100	0.4930	0.5451	0.8177	0.3658	1.6371
diagnAsthma:$smoking_{home}$1[100,200]	1.1466	0.7505	2.3342	0.1266	3.1475
diagnAsthma:$smoking_{home}$1>200	-0.7807	0.7378	1.1197	0.2900	0.4581
diagnAsthma:$smoking_{daycare}$1	0.2131	0.3125	0.4653	0.4952	1.2376
diagnAsthma:$pets_{home}$	-0.0010	0.0021	0.2050	0.6507	0.9990

Table 5.18: Summary for initial risk-factor analysis (5.23)

units in the third trimester increases the risk of getting an episode more than for 3 or more units. Asthmatic children with a mother drinking two units per week have an increased risk: $\hat{OR} = 8.51$ for comparison with non-asthmatic children having mothers with had no alcohol intake in the third trimester. The model can however be reduced, such that only the interaction between diagnosis and alcohol is kept of the interactions between diagnosis and risk-factors.

Estimating the updated model gives the summary in Table 5.19, which shows that an alcohol intake of two units still increase the odds significantly. It is furthermore

	estimate	san.se	wald	p	OR
(Intercept)	-2.4918	0.2950	71.3533	<0.0001	0.0828
diagnAsthma	0.7134	0.2616	7.4368	0.0064	2.0409
age	0.3283	0.1032	10.1182	0.0015	1.3887
(age^2)	-0.1246	0.0272	21.0658	<0.0001	0.8828
cosine	0.5046	0.0391	166.2368	<0.0001	1.6563
sine	0.1635	0.0422	15.0263	0.0001	1.1776
log10(pd)	-0.1533	0.0991	2.3950	0.1217	0.8579
SEXMale	0.0067	0.1343	0.0025	0.9605	1.0067
$\text{smoking}_{\text{3rd}}$Yes	0.1112	0.2164	0.2640	0.6074	1.1176
$\text{father}_{\text{ashtma}}$Yes	0.1620	0.1645	0.9694	0.3248	1.1758
pets_{3rd}	-0.1091	0.1393	0.6136	0.4334	0.8967
alcohol1	-0.0094	0.2522	0.0014	0.9701	0.9906
alcohol2	-0.0478	0.4356	0.0120	0.9127	0.9534
alcohol3	-0.6167	1.0176	0.3673	0.5445	0.5397
fillagrin1	0.4027	0.2372	2.8826	0.0895	1.4959
$\text{smoking}_{\text{home}}$1<100	-0.4668	0.2575	3.2857	0.0699	0.6270
$\text{smoking}_{\text{home}}$1[100,200]	-0.2491	0.4054	0.3775	0.5389	0.7795
$\text{smoking}_{\text{home}}$1>200	-0.5253	0.3747	1.9653	0.1610	0.5914
$\text{smoking}_{\text{daycare}}$1	-0.1391	0.1496	0.8637	0.3527	0.8702
diagnAsthma:age	0.3621	0.0838	18.6680	<0.0001	1.4363
diagnAsthma:alcohol1	-0.1085	0.3494	0.0964	0.7562	0.8972
diagnAsthma:alcohol2	1.3496	0.5344	6.3771	0.0116	3.8558
diagnAsthma:alcohol3	-0.2686	1.0487	0.0656	0.7978	0.7644

Table 5.19: Summary for reduced model for risk-factor analysis

seen that the estimates for ≥ 3 units are very uncertain, which gives the χ^2-statistic,

$$\chi^2 = \frac{(\beta_{p1} - \beta_{p2})^2}{\sigma^2(\beta_{p1} - \beta_{p2})} = \frac{(\beta_{p1} - \beta_{p2})^2}{\sigma^2_{\beta_{p1}} + \sigma^2_{\beta_{p2}} - 2 \cdot \sigma_{p1,p2}} \tag{5.24}$$

for comparing the ≥ 3 unit group with the 2 unit group-estimate. The statistics have the values 2.07 and 0.26 for the asthmatic and non-asthmatic group, respectively. The statistics should be compared to a χ^2-distribution with one degree of freedom, which has the critical 10 % quantile 2.71. It is therefore possible to join the two levels in a common level ≥ 2.

It is furthermore seen from Table 5.19 that the risk-factors: Pets in third trimester, smoking in the third trimester, smoking in the day-care and gender, can be removed from the model, since the estimates are insignificant. The resulting model for the risk-factors is

$$\begin{aligned} \text{risk-factors} =& \beta_6 \cdot \text{fillagrin}_i + \beta_{8k} \cdot \text{smoking}_{\text{home},i} \\ &+ \beta_{13k} \cdot \text{alcohol}_{\text{3rd},i} \end{aligned} \tag{5.25}$$

The summary for the reduced model (5.25) is shown in Table 5.20, which indicates that drinking 2 or more units of alcohol per week do increase the probability of an episode for the asthmatic children, however not significantly. The model can therefore be reduced, to a model without the diagnosis-specific effect of alcohol.

	estimate	san.se	wald	p	OR
(Intercept)	-2.5217	0.2793	81.4940	<0.0001	0.0803
diagnAsthma	0.7491	0.2500	8.9778	0.0027	2.1151
age	0.3262	0.1029	10.0483	0.0015	1.3856
(age^2)	-0.1241	0.0272	20.8205	<0.0001	0.8833
cosine	0.5027	0.0391	165.6231	<0.0001	1.6532
sine	0.1619	0.0419	14.8993	0.0001	1.1757
log10(pd)	-0.1523	0.0956	2.5377	0.1112	0.8587
alcohol1	0.0266	0.2606	0.0104	0.9187	1.0270
alcohol>2	-0.2361	0.4159	0.3222	0.5703	0.7897
fillagrin1	0.3514	0.2463	2.0359	0.1536	1.4211
$smoking_{home}$1<100	-0.4693	0.2384	3.8761	0.0490	0.6254
$smoking_{home}$1[100,200]	-0.3077	0.3865	0.6337	0.4260	0.7352
$smoking_{home}$1>200	-0.5124	0.3379	2.2988	0.1295	0.5991
diagnAsthma:age	0.3601	0.0831	18.7559	<0.0001	1.4335
diagnAsthma:alcohol1	-0.0894	0.3464	0.0666	0.7964	0.9145
diagnAsthma:alcohol>2	0.5733	0.8113	0.4993	0.4798	1.7741

Table 5.20: Summary for reduced model for risk-factor analysis after collapsing the two highest levels in the variable alcohol intake in the third trimester

	estimate	san.se	wald	p	OR
(Intercept)	-2.5313	0.2771	83.4575	<0.0001	0.0796
diagnAsthma	0.7642	0.2384	10.2772	0.0013	2.1473
age	0.3258	0.1027	10.0523	0.0015	1.3851
(age^2)	-0.1240	0.0272	20.7820	<0.0001	0.8834
cosine	0.5025	0.0390	165.9336	<0.0001	1.6528
sine	0.1619	0.0419	14.9176	0.0001	1.1757
log10(pd)	-0.1514	0.0949	2.5442	0.1107	0.8595
alcohol1	-0.0032	0.1885	0.0003	0.9863	0.9968
alcohol>2	0.0804	0.4219	0.0363	0.8488	1.0838
fillagrin1	0.3622	0.2447	2.1918	0.1388	1.4365
$smoking_{home}$1<100	-0.4740	0.2399	3.9030	0.0482	0.6225
$smoking_{home}$1[100,200]	-0.3180	0.3764	0.7137	0.3982	0.7276
$smoking_{home}$1>200	-0.4489	0.3229	1.9329	0.1644	0.6383
diagnAsthma:age	0.3599	0.0831	18.7352	<0.0001	1.4332

Table 5.21: Summary for reduced model for risk-factor analysis after collapsing the two highest levels in the variable alcohol intake in the third trimester and removing the corresponding interaction with diagnosis as well as some of the other risk-factors

After removing the diagnosis-specific effect of alcohol, the summary in Table 5.21 shows that essentially none of the risk-factors besides diagnosis influence the risk of an episode. The smoking in the home variable is seen not to be significant at more than on a 5 % level for one of the categories. The analysis shows that the probability of an episode is determined by the season, diagnosis and the different age-progress for each of the corresponding groups.

Finally the specific resistance at the age of 3 years is considered, since it has around 40 % missing values and therefore would lead to a seriously affected dataset in the previous analyses. The basis for the estimation is the model in (5.22), where the risk-factor now is the interaction between diagnosis and sRAW. sRAW was previously seen to be sufficiently linear in relation to the probability of having asthma, see Figure 4.14, which results in that the variable is treated as such in the following analysis.

The model is estimated and the summary shown in Table 5.22. From the table it is seen that the relative increase in the specific resistance at the age of 3 years does not have significant influence on the risk of an episode. It is however seen that the asthmatic group has a negative estimate (although insignificant), which shows that the tendency for this group is a reduction in the risk if the relative increase is high. This imply that the asthmatic children which benefit the most of the bronchodilator treatment have a higher risk of an episode compared to other asthmatic children. This is not surprising, since a large relative decrease in the specific airway resistance is seen for children with a narrow airway, who respond well on the airway-relaxationing treatment. These children are seen to be more likely of having asthma [27].

	estimate	san.se	wald	p	OR
(Intercept)	-2.7819	0.1789	241.6729	<0.0001	0.0619
diagnAsthma	0.6703	0.3492	3.6842	0.0549	1.9548
age	0.3452	0.1239	7.7702	0.0053	1.4123
(age^2)	-0.1308	0.0334	15.3461	<0.0001	0.8774
cosine	0.5050	0.0429	138.3702	<0.0001	1.6570
sine	0.1284	0.0499	6.6167	0.0101	1.1370
sraw	0.0067	0.7082	0.0000	0.9924	1.0067
diagnAsthma:age	0.4116	0.1065	14.9272	0.0001	1.5093
diagnAsthma:sraw	-0.3306	1.0380	0.1015	0.7501	0.7185

Table 5.22: Summary for reduced model for risk-factor analysis after collapsing the two highest levels in the variable alcohol intake in the third trimester and removing the corresponding interaction with diagnosis as well as some of the other risk-factors and including the sraw measurement

5.7 Transition model

Another approach to analyze the longitudinal evolvement of the symptoms is to regress on historic observations of the symptoms. This leads to a model, which describes the transition from a week to the following week. Diggle et al. [15] formulates the q'th order transition model as

$$\text{logit}(P(Y_{ij}|\mathcal{H}_{ij})) = \mathbf{x}_{ij}\beta + \sum_{r=1}^{q} \alpha_r y_{i,(j-r)} \tag{5.26}$$
$$\mathcal{H}_{ij} : Y_{i,(j-1)} = y_{i,(j-1)}, \ldots, Y_{i,(j-q)} = y_{i,(j-q)}$$

This imply that the interpretation of the parameters is conditional on the history. In the following the analysis with $q = 1$ is considered to keep things simple, this corresponds to assuming the system to be Markovian of first order. Diggle et al. p. 131 [15] describes the model by the transition matrix

$$\begin{array}{ccc} & & y_{i,j-1} \\ & & \begin{array}{cc} 0 & 1 \end{array} \\ y_{i,(j)} & \begin{array}{c} 0 \\ 1 \end{array} & \begin{array}{cc} \frac{1}{1+\exp(\mathbf{x}_{ij}\beta)} & \frac{1}{1+\exp(\mathbf{x}_{ij}\beta+\alpha_1)} \\ \frac{\exp(\mathbf{x}_{ij}\beta)}{1+\exp(\mathbf{x}_{ij}\beta)} & \frac{\exp(\mathbf{x}_{ij}\beta+\alpha_1)}{1+\exp(\mathbf{x}_{ij}\beta+\alpha_1)} \end{array} \end{array} \tag{5.27}$$

From the analysis of the lorelograms it was seen that the correlation pattern was different for the three groups, which should be incorporated in the model (however here only the diagnosis is used, i.e. asthma/non-asthma). The probability of staying in the no symptoms state was seen to much higher for the non-asthmatic group, since the children in this group had no symptoms most of the weeks.

		Non-asthma		Asthma	
		Previous week			
		No	Yes	No	Yes
Current	No	97	48	88	40
	Yes	3	52	12	60

Table 5.23: Contingency tables corresponding to the two groups in percent (each column sums to 100 %). Columns correspond to $y_{i,(j-1)}$ (lagged value/previous state) and rows to $y_{i,j}$ (observed/current state)

In Table 5.23 the evolvement for the both groups is shown, it is seen that the non-asthmatic is more likely to stay in the no symptoms state given that they had no symptoms in the previous week ($P(y_{ij} = 0|y_{i,(j-1)} = 0)$) compared to the asthmatic group. The probability of staying in the symptom state, $P(y_{ij} = 1|y_{i,(j-1)} = 1)$, is seen to be more similar for the two groups. It is seen that $P(y_{ij} = 1|y_{i,(j-1)} = 1) >$ 52 %, which shows that having symptoms in week $j-1$ in over 52 % of the cases will lead to an episode in week j.

5.7.1 Modeling

With the consideration about the group dependent transition, the model in (5.10) is expanded with the interaction between group and the state the week before, which gives the new model

$$\begin{aligned}\hat{\eta}_{ij} =& \beta_{0k} + \beta_{1k} \cdot \text{age}_{ij} + \beta_{2k} \cdot \text{age}_{ij}^2 \\ &+ \beta_{3k} \cdot \cos\left(\frac{\text{week} \cdot 2\pi}{52}\right) + \beta_{4k} \cdot \sin\left(\frac{\text{week} \cdot 2\pi}{52}\right) \\ &+ \alpha_{1k} \cdot y_{i,(j-1)} \\ \eta =& \log\left(\frac{p}{1-p}\right)\end{aligned} \tag{5.28}$$

This essentially reduces the data-set by 1 observation per period the diary has been kept per child, which is 394 observations on 357 individuals from a dataset with 61036 observations. Furthermore the probability to be modelled is changed from $P(Y_{ij} = 1)$ to $P(Y_{ij} = 1|Y_{i,(j-1)} = y_{i,(j-1)})$, i.e. the interpretation will depend on the value of $Y_{i,(j-1)}$. The model is estimated by standard generalized linear model techniques, since the correlation is explained by the regression on old values. The autocorrelation function for the residuals is shown in Figure 5.8, which clear shows that no correlation is present.

The summary for the estimation of the model in (5.28) is shown in Table 5.24, which shows that the seasonal part can be reduced to a common structure for both groups. Testing for the same seasonal part for both groups gives an increase in the deviance of 6.22 on 6 degrees of freedom, which shows that the reduction leads to an insignificant increase in the deviance.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.3656	0.0705	-47.7598	<0.0001
diagnosisAsthma	0.8662	0.1169	7.4115	<0.0001
age	0.0384	0.0682	0.5627	0.5736
(age^2)	-0.0466	0.0139	-3.3475	0.0008
cosine	0.4047	0.0340	11.8918	<0.0001
sine	0.1111	0.0329	3.3733	0.0007
lag11	3.5130	0.0488	71.9590	<0.0001
diagnosisAsthma:age	0.4455	0.1059	4.2085	<0.0001
diagnosisAsthma:(age^2)	-0.0422	0.0208	-2.0299	0.0424
diagnosisAsthma:cosine	-0.0968	0.0503	-1.9262	0.0541
diagnosisAsthma:sine	-0.0805	0.0492	-1.6359	0.1019
diagnosisAsthma:lag11	-1.1503	0.0719	-15.9900	<0.0001

Table 5.24: Summary transition model with group dependent transition

Updating the model with a mutual seasonal part gives the summary in Table 5.25, which shows that the amplitude for the seasonal part, $A \cdot \cos(t \cdot \pi/365 - \theta)$, is 0.37

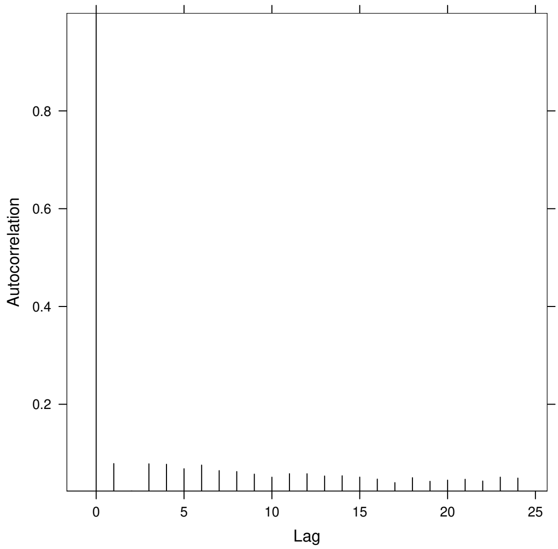

Figure 5.8: Autocorrelation for residuals in transition model. The autocorrelation is estimated by considering the correlation between observations taken on the same individual.

and the phaseshift is -0.2 or -11.85 days, which is seen to be 7 days before compared to the marginal unconditional model, see equation (5.4.1).

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.3580	0.0703	-47.7392	<0.0001
diagnosisAsthma	0.8474	0.1167	7.2600	<0.0001
age	0.0390	0.0682	0.5716	0.5676
(age^2)	-0.0466	0.0139	-3.3455	0.0008
lag11	3.5197	0.0487	72.2236	<0.0001
cosine	0.3599	0.0250	14.3955	<0.0001
sine	0.0745	0.0245	3.0447	0.0023
diagnosisAsthma:age	0.4502	0.1059	4.2506	<0.0001
diagnosisAsthma:(age^2)	-0.0431	0.0208	-2.0724	0.0382
diagnosisAsthma:lag11	-1.1615	0.0718	-16.1693	<0.0001

Table 5.25: Summary transition model with group dependent transition with mutual seasonal part

From the summary it is furthermore seen that the odds-ratios for the lagged variables (the ratio between the odds coming form a week with symptoms and coming form a week without symptoms) are largest for the non-asthmatic group and smallest for the asthmatic group. The odds-ratios are 33.77 and 10.57 for respectively the non-asthmatic and asthmatic group. This coincide well with Table 5.23, where the corresponding odds-ratios can be estimated to

$$OR_{\text{non-asthma}} = \frac{P(Y_{ij} = 1|Y_{i,(j-1)} = 1)/P(Y_{ij} = 0|Y_{i,(j-1)} = 1)}{P(Y_{ij} = 1|Y_{i,(j-1)} = 0)/P(Y_{ij} = 0|Y_{i,(j-1)} = 0)} = \frac{0.52/0.48}{0.03/0.97} = 35.03$$

$$OR_{\text{asthma}} = \frac{0.60/0.40}{0.12/0.88} = 11.00$$

The large estimated odds-ratio for the non-asthmatic group is caused by the low probability of staying in the high symptom state and the high probability of staying in the non-symptom state. For both groups it is seen that a week with symptoms is much more likely after a week with symptoms compared to a week without. It is seen that the empirical odds-ratios are close to identical when comparing to the model estimates. The analysis shows that the episodes tend to influence adjacent weeks, i.e. either lasting more than a week or goes across the week-shift. In Table 5.26 the estimated transition probabilities are shown, it is seen that they are close to the results found in Table 5.23.

		Non-asthma		Asthma	
		Previous week			
		No	Yes	No	Yes
Current	No	0.97	0.53	0.86	0.38
	Yes	0.03	0.47	0.14	0.62

Table 5.26: Predicted transition probabilities at the age of 3 years. The probabilities correspond to the time of year, where the risk is average (cosine and sine cancels out), i.e. around April or October

The longitudinal development is seen to be similar to the results found in the diagnosis-analysis: The non-asthmatic group starts lowest, has a negative slope and weak curvature and the asthmatic group is seen to have the highest onset a positive slope and a negative curvature.

5.7.2 Expansion

In the analysis above only the lagged response of order 1 was used, which corresponds to assuming that the first order Markov condition is fulfilled. This imply that all historic information relevant for the current state is contained in the previous. The lorelograms showed that correlations might be present beyond week to week.

One method to investigate the level of the transition model is to consider for example the probabilities $P(Y_{ij}|Y_{i,(j-1)} = y_{i,(j-1)}, Y_{i,(j-2)} = y_{i,(j-2)})$ and $P(Y_{ij}|Y_{i,(j-1)} = y_{i,(j-1)}, Y_{i,(j-2)} = y_{i,(j-2)}, Y_{i,(j-3)} = y_{i,(j-3)})$, which correspond to a second and third order model, respectively. For the second order model four probabilities are to be estimated, namely

$$
\begin{aligned}
&P(Y_{ij} = 1|Y_{i,(j-1)} = 0, Y_{i,(j-2)} = 0)\\
&P(Y_{ij} = 1|Y_{i,(j-1)} = 1, Y_{i,(j-2)} = 0)\\
&P(Y_{ij} = 1|Y_{i,(j-1)} = 0, Y_{i,(j-2)} = 1)\\
&P(Y_{ij} = 1|Y_{i,(j-1)} = 1, Y_{i,(j-2)} = 1)
\end{aligned}
$$

The model can be specified by including an interaction term between $y_{i,(j-1)}$ and $y_{i,(j-2)}$. For the third order model the number of probabilities to be modelled is 8 (2 times the number for $q = 2$). The model is seen to be a modification of the model in (5.26), which is seen not to allow the interaction.

In Table 5.27 transition probabilities based on second order information for the three groups are shown. It is seen that the two combinations with no symptoms in week $j - 1$ gives nearly the same transition probabilities regardless of the symptoms in week $j - 2$.

Asthma status	Current state	Week $j-2$ / $j-1$	No: No	No: Yes	Yes: No	Yes: Yes
Non-asthma	No		97	45	91	50
	Yes		3	55	9	50
Asthma	No		89	36	80	42
	Yes		11	64	20	58

Table 5.27: Transition probabilities for second order information in %, i.e. a column within a box sums to 100 %.

Table 5.27 indicates that if week $j - 1$ is a no symptom week then the odds increase when $j - 2$ is changed from no symptoms to symptoms. If the state in week $j - 1$ is the symptom state then the odds decreases for a change from no symptoms to symptoms in week $j - 2$. This shows that an interaction between the state in week $j - 1$ and $j - 2$ is likely to be present.

An model corresponding to the second order transition probabilities can be formulated as

$$
\begin{aligned}
\hat{\eta}_{ij} =& \beta_{0k} + \beta_{1k} \cdot \text{age}_{ij} + \beta_{2k} \cdot \text{age}_{ij}^2 \\
&+ \beta_{3k} \cdot \cos\left(\frac{\text{week} \cdot 2\pi}{52}\right) + \beta_{4k} \cdot \sin\left(\frac{\text{week} \cdot 2\pi}{52}\right) \\
&+ \alpha_{1k} \cdot y_{i,(j-1)} + \alpha_{2k} \cdot y_{i,(j-2)} \\
&+ \alpha_{3k} \cdot (y_{i,(j-1)} \times y_{i,(j-2)})
\end{aligned}
\tag{5.29}
$$

In Table 5.28 the summary for the model is shown. It is seen that all estimates corresponding to $y_{i,(j-2)}$ are significant, which indicates that the effect of the 2 weeks

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.4039	0.0711	-47.8512	<0.0001
diagnosisAsthma	0.8689	0.1179	7.3716	<0.0001
age	0.0093	0.0687	0.1349	0.8927
(age^2)	-0.0394	0.0140	-2.8126	0.0049
lag11	3.7066	0.0644	57.5271	<0.0001
lag21	1.2054	0.1022	11.7887	<0.0001
cosine	0.3495	0.0251	13.9100	<0.0001
sine	0.0613	0.0245	2.4962	0.0126
lag11:lag21	-1.4294	0.1291	-11.0764	<0.0001
diagnosisAsthma:age	0.4247	0.1069	3.9744	<0.0001
diagnosisAsthma:(age^2)	-0.0412	0.0210	-1.9686	0.0490
diagnosisAsthma:lag11	-1.0732	0.0994	-10.7963	<0.0001
diagnosisAsthma:lag21	-0.6036	0.1350	-4.4720	<0.0001
diagnosisAsthma:lag11:lag21	0.5278	0.1776	2.9711	0.0030

Table 5.28: Summary transition model with group dependent transition and second order information

lagged values is present. The estimates show that the asthmatic group is significant different from the non-asthmatic group.

Odds-ratios for $y_{i,(j-1)} = 0, y_{i,(j-2)} = 1$ against $y_{i,(j-1)} = 0, y_{i,(j-2)} = 0$, $y_{i,(j-1)} = 1, y_{i,(j-2)} = 1$ against $y_{i,(j-1)} = 1, y_{i,(j-2)} = 0$ and $y_{i,(j-1)} = 1, y_{i,(j-2)} = 1$ against $y_{i,(j-1)} = 0, y_{i,(j-2)} = 0$ can be estimated to evaluate the effect of 2 weeks lagged values against 1 week lagged values for each of the three groups. This corresponds to $e^{\alpha_{2k}}$, $e^{\alpha_{3k}+\alpha_{2k}}$ and $e^{\alpha_{3k}+\alpha_{2k}+\alpha_{1k}}$. The odds-ratios are summarized in Table 5.29, which shows the three comparisons

		Odds-ratio	
Comparison	Parameters	Non-asthma	Asthma
$\frac{\text{Odds}(y_{i,(j-1)}=0,y_{i,(j-2)}=1)}{\text{Odds}(y_{i,(j-1)}=0,y_{i,(j-2)}=0)}$	α_2	3.34	1.83
$\frac{\text{Odds}(y_{i,(j-1)}=1,y_{i,(j-2)}=1)}{\text{Odds}(y_{i,(j-1)}=1,y_{i,(j-2)}=0)}$	$\alpha_3 + \alpha_2$	0.80	0.74
$\frac{\text{Odds}(y_{i,(j-1)}=1,y_{i,(j-2)}=1)}{\text{Odds}(y_{i,(j-1)}=0,y_{i,(j-2)}=0)}$	$\alpha_3 + \alpha_2 + \alpha_1$	32.54	10.31

Table 5.29: Odds-ratios for different comparisons for 2 level transition model

It is seen that the effect of having an episode two weeks ago and no episode in the previous week increases the odds with between 234 % and 83 % compared to 3971 % and 1292 % for having symptoms in the previous week but not to week ago. The odds of an episode decreases by 20 % and 26 % when having an episode in both the previous week and the week before that compared to only having symptoms in the previous week. It is seen that having had symptoms in two consecutive weeks decreases the probability of a new episode compared to having had only 1 consecutive week. Finally it is seen that the effect of two consecutive week with symptoms is a little lower than just having symptoms in the previous week, however much higher compared to just having symptoms two weeks ago.

Asthma status	Current state	Week $j-2$ / $j-1$	No: No	No: Yes	Yes: No	Yes: Yes
Non-asthma	No		98	51	93	56
	Yes		2	49	7	44
Asthma	No		88	34	80	41
	Yes		12	66	20	59

Table 5.30: Estimated transition probabilities from transition model for second order information in %, i.e. a column within a box sums to 100 %.

The analysis shows that for the asthmatic and the non-asthmatic group the odds of an episode after two consecutive weeks with symptoms is smaller compared to having only one week with symptoms. This could be explained by short episodes, hence if the episode last no more than 7-14 days, the risk will decrease the longer the current episode has lasted. It is however seen that having had an episode two week ago increases the risk of an episode by 234 % and 83 % for having had symptoms in week $j-2$ but not in week $j-1$ and 3154 % and 931 % for having had symptoms in both weeks for the non-asthmatic group and the asthmatic group, respectively.

5.8 Discussion

In this chapter weekly episodes were considered in order to analyze a faster dynamic than yearly aggregated symptoms. It was seen that the symptoms typically lasted 1-2 weeks, which could be seen from the lorelograms having an elbow at $k = 2$ and from the modelling of the transition probabilities. The transition probabilities were seen to indicate that a new episode was less likely after two weeks with symptoms compared to only one week.

The risk of an episode was seen to be related to the longitudinal pattern as seen for the yearly aggregated symptoms, i.e. the non-asthmatic children had a decreasing risk as they got older and the asthmatic children started at a higher level and had an increasing risk until the age of 3 years. It was furthermore seen that the risk of an episode was 2-3 times higher in the winter compared to the summer. The seasonal risk was seen to be the same for the asthmatic and non-asthmatic children, which indicated that the seasonal part was an indicator of symptoms not related to asthma.

The analysis showed that the non-asthmatic children were likely to be in the non-symptom state, which gave them high odds-ratios of staying in the same state. The odds-ratios of an episode in the following week were 33 and 11 for a week with symptoms against a week with non symptoms for the non-asthmatic and asthmatic groups, respectively. The high odds-ratio for the non-asthmatic group was caused by the high probability of staying in the no-symptom state, whereas the probabilities of an new episode given the previous week was with symptoms are more equal. The asthmatic children were more likely of getting an episode compared to the non-asthmatic children, the odds-ratios were $(0.12/0.88)/(0.03/0.97) = 4.41$ and $(0.60/0.40)/(0.52/0.48) = 1.38$ for no symptoms and symptoms the previous week, respectively. This showed that the main difference was seen for the weeks preceded by no symptoms.

The analysis of the risk-factors and medication showed that none of the considered risk-factors increased the risk of an episode after correction for age, diagnosis and season. Medication was seen to be difficult to model due to the tight connection between symptoms and medication and lagged values were needed in order to obtain a causal system.

The analysis of the weekly episodes did not contribute by additional information on the risk of symptoms. It was seen that the analysis confirmed the results found previous and gave a model combining the model-elements from Chapter 2, 3 and 4.

Chapter 6

Conclusion

The work in this thesis has shown that yearly symptom rates can be subdivided into three sub-categories. Latent class regression was applied to year aggregated symptoms rates in order to find unique patterns in the longitudinal development. The identified groups are characterized by having different longitudinal development, which separate the groups well as early as the age of 2 years. Using different assumptions for the response, i.e. gaussian rates and poisson counts, yield some minor differences, which mainly leads to slight changes in grouping. However, the identified groups for the two methods were seen to have almost the same longitudinal development. The grouping was based on symptoms rates in the first five years of life and was estimated prior to the knowledge of the asthma diagnosis at the age of 5.

Prior to the analysis of subgroups mixed effects models were considered to analyze the heterogeneity of the longitudinal developments. The analysis by means of mixed effects model revealed a highly significant heterogeneity in the longitudinal patterns. The analysis of the mixed effects models was carried out for two different response scales, but led to the same conclusion, namely that the children differed mainly on their starting level and the first order parameter for age. This implied that 3-4 different patterns were seen according to the age-related trend.

The predictions of individual parameters from the mixed effects models were seen to differ for the three groups found later in the latent class regression, both in starting level and development. The groups were seen to start at different levels and the group with most symptoms had increasing initial symptom rate compared to a decreasing or constant symptom rate for the other groups. The heterogeneity between individuals was seen to insignificant within the three groups, which implied that the initial heterogeneity in the mixed effects model was caused by the differences between the three groups.

The identified groups were compared to the asthma diagnosis at the age of 5 years, which showed that the agreement was good. The two groups with the lowest symptoms rates were classified as non-asthmatic and the last as asthmatic, which gave the possibility to estimate sensitivity, specificity and overall agreements. Sensitivity was higher for the gaussian model, whereas the specificity was marginal better for the poisson model. It was seen that using the first 3 years of observations gave convincing results, indicating that the episodes from the first years of the childrens life were most important in predicting asthma.

In further analysis of the symptom patterns and the ability to predict the diagnosis, one could analyze whether the patterns could be identified at a lower age. The patterns leading to the three groups were based on the symptoms of the first five years of life and since the diagnosis is based on the symptoms in the fifth year of life, this implied that the grouping was partially based on the same information as the diagnoses were determined on. The parameters used in an early identification in this thesis is estimated from all five years, i.e. a classification at the age of three years of life was obtained by calculation a likelihood on the first three years of life for a model based on all five years of life. However, the pattern recognition was initial proposed in order to find subgroups of children, i.e. finding unique longitudinal patterns among the children and led to the three different groups. Prediction of the asthma-diagnosis was not the original aim of the analysis of symptom patterns, however the analysis showed that the patterns found were in accordance with the diagnosis. Since the symptom-patterns are quite distinct at the age of 3 years, predictions based only on the first three years of life may be possible, which will lead to a prediction model separated from the information on which the diagnosis was determined.

Analysis of the mixed effects models showed that the heterogeneity could be described by the diagnosis. The asthmatic children had a significant higher starting level compared to the non-asthmatic children and had a significant higher initial slope as well. Heterogeneity was seen to be related to the differences between asthmatic and non-asthmatic children or to differences between the groups found in latent class regression and not to heterogeneity between children with the same diagnosis or from the same group.

Comparison of the longitudinal development for the asthmatic group showed that it was similar to the pattern seen for the high symptom group and the non-asthmatic group was a mixture of the low and middle groups. It was seen that in describing the symptoms a three cluster model was optimal for the gaussian model, whereas the poisson model had five clusters as optimum. The five cluster model gave patterns, which were seen to be in agreement with the results found by Martinez et al. [26] after adjusting their results by the prevalence of maternal asthma. The comparison showed that fewer late onset wheezers and more non-wheezers were identified in the COPSAC-group compared to the study by Martinez et al, which might be explained by the fact that the children in Martinez et al. received their second classification at the age of six, which gave them a year longer to development asthma.

Risk-factors for the risk of getting the asthma-diagnosis at the age of 5 years were analyzed by means of logistic regression. The analysis showed that a high congenital resistance reduced the risk of asthma significant. The risk of asthma for the lowest resistance was 40 % and was seen to decline to around 10 % for the mean resistance. Congenital resistance was seen to have a inconsistent effect in the cluster models, i.e. having positive, negative or no effect depending on the model. The longitudinal medication pattern at a yearly aggregated level were analyzed and was seen to be closely related to both the diagnosis and the symptom rate. It was seen that in years with many symptoms a high level of medication was given. However since the medication was given according to the symptoms, the medication could not be used as predictor for the symptoms without establishing a non-causal relation.

In the first part of the thesis the population was analyzed at a daily basis in order to determine the seasonal variations. The seasonal risk was analyzed in both a prevalence model and a incidence model, this led to models for the daily percentages for the total number of children with symptoms and the number of new children with symptoms. The analysis showed that a season lasted approximately one year and that the prevalence as well as the incidence were 2-3 times higher in the winter compared to the summer. The seasonal variations were seen to be adequately described by a periodic function with a period of a year.

In the last part of the thesis weekly episodes were discussed as a method for analyzing the fast dynamics. A logistic regression for the probability of having an episode in a given week was applied. Results from the yearly aggregated symptoms were reused in order to account for the age effect, i.e. that asthmatic and non-asthmatic children had different risk due to age. The parametric model for seasonal effect was included in order to evaluate the risk for different seasons. It was seen for both the prevalence and incidence models and the model for the day to day model that the risk was 2-3 times higher in the winter compared to the winter and that the risk related to season was the same for asthmatic and non-asthmatic children. The seasonal effect may therefore be interpreted at being un-related to the asthma symptoms and instead related to more general wheezing variations, i.e. cold and flu. The risk of an episode was seen to have the same pattern for the asthmatic and non-asthmatic children as seen in the analysis of the yearly symptoms rates.

The analysis of the weekly episodes showed that none of the risk-factors were significant and modelling of the medication use was seen to give the same problems as for the yearly aggregated symptoms. The medication status in the week considered was seen to increase the risk of symptoms, which was caused by the causality of symptoms and medication. Furthermore, lagged medication information was seen have an insignificant impact on the risk of symptoms. The significant differences in the risk for an episode were seen to be related to the season and to a difference in the age-related risk between asthmatic and non-asthmatic children.

A transition model for the transition probabilities for the weekly data was analyzed, which showed that having had an episode the previous week increased the odds of an episode markedly by more than 11 times. The transitions in the non-asthmatic group were seen to be difficult to model, since this group had very few symptoms and thereby few shifts in symptom status. Furthermore, the analysis showed that two consecutive weeks with symptoms decreased the odds of symptoms compared to having had symptoms the previous week, which indicated that the length of the periods with symptoms typically were around 1-2 weeks. The transition model analysis may be extended to analyze risk-factors on week to week dynamics, but in the thesis the transition model did not give additional insight into the weekly dynamics.

The main result in the thesis was the identification of three distinct symptom patterns. The connection to the diagnosis showed that the patterns could be used for predicting the asthma diagnosis. Future work will show if the prediction can be based on a shorter time-range than all five years of life, eg the first three years, in order to separate the information used for prediction and the information used to determine the diagnosis. If an adequate model can be established, the model can be used in statistic process control for the symptoms, i.e. if parents reports the wheezing symptoms continuously then the model can estimate the probability of belonging to the high level symptom group in an automated way.

Appendix A

Conclusion from preparatory thesis

A.1 Conclusion

In chapter 3 and 4 the congenital lung-function was analyzed to find risk-factors or confounders describing FEV and PD15 PtcO2. The analysis showed that FEV was described by the childrens length at birth and their age of measurement, which accounted for 38 % of the variation in the congenital FEV. Furthermore a model with a general size measure used instead of length at birth was seen to give the same degree of explanation as the length, whereas using the body mass index gave a poorer model in terms of describing the FEV.

A model for age of measurement and length at birth corrected FEV showed that smoking in the third trimester significantly decreased the corrected FEV. The model also showed an increasing corrected FEV for increasing gestational age. It was seen that the risk-factor smoking was insignificant for the uncorrected FEV, which was solved with the modelling of the corrected FEV.

For the PD15 PtcO2 measurement at the age of approximately one month, the only certain risk-factor found was the gene mutation variable. Having at least one of the two gene mutation increased the resistance against the provocation, hence children having at least one of the mutations are less sensitive. The model for PD15 PtcO2 was capable of describing 4-9 % of the variation depending on how complex the model was allowed to be. The initial modelling had a significant interaction between the mothers asthma status in the first trimester and having contractions in the second, which seemed rather unlikely to be true. A model restricted to work on third trimester variables, when second and first were available, proposed that the gene variable and use of paracetamol both influenced the resistance significantly.

The congenital lung-function was seen to depend on rather few of the risk-factors and confounders recorded. For FEV length at birth, age of measurement and gestational age as confounders and smoking in the third trimester as a risk-factor and for PD15 PtcO2 the gene mutation risk-factor were the only estimated effects.

At the age of three years the specific airway resistance for pre (base) and post bronchidilator treatment was considered (chapter 5). It was seen that the base measurement showed a difference between boys and girls as the only significant effect and it explained 4 % of the variation in specific airway resistance. The boys had a 11 % higher specific airway resistance compared to girls. The corrected FEV was negatively correlated with the pre-bronchodilator measurement, but it was questionable if the effect was significant. If significant it would imply that children with a relatively good lung-function at the age of 1 month have a lower resistance in the airways. However, the corrected FEV was only significant in the model with the variable home-type included (different lung-function for children living in terrace houses compared to children living in house or apartment)

The variations in post bronchidilator measurements of the specific airway resistance were describing 38 % of the pre-bronchodilator measurements. A high pre-bronchodilator resistance gave an estimated low ratio between the post and the base measurements, which implied that children with a high pre-bronchodilator resistance benefited the most of bronchodilator treatment. The model identified that use of allergy quilt decreased the specific airway resistance by up to 9 %. The results for the pre-bronchodilator measurements influence on the post measurements were confirmed for the modelling of the relative increase in the specific airway resistance.

In chapter 6 models for the prevalence and the incidence were analyzed. This gave two models, which modelled the seasonal pattern in the two time-series. The model for the prevalence showed that even after removing the seasonal effect correlation between neighboring days was still present, which was shown not to be the case for the incidence. Having modelled the incidence and prevalence in a parametric model, the incidence or prevalence can be used as confounders for the modelling of the individual time-series of wheezing symptoms. The model showed that the peak of the prevalence and incidence varied from year to year and that the incidence peaked before the prevalence.

The individual wheezing symptoms were analyzed in the latter part of chapter 6. The analysis showed that dividing the children into two groups, one where the children had many symptoms and one where the symptoms was rather low, gave a good model for describing the longitudinal progress of the wheezing symptoms. It was seen that the two groups both had a decreasing number of symptoms as the children got older. The decrease was seen to be the same for the two groups and low (two days per year). Even though the model accounted for difference in variance for the two groups and was taking care of the correlation within the individuals the residuals still had a pattern, which implied that the model tended to underestimate the symptoms.

Finally the relation between the two groups of children, defined by the amount of symptoms, and the congenital PD15 PtcO2-measurements was examined. The analysis showed that the group with few symptoms was more likely to have a high PD15 PtcO2, which implied that children with a low congenital sensitivity tended to be classified in the group with few symptoms. The mean levels of PD15 PtcO2 were significant different for the two groups and were highest in the low group.

The work in this report points forward to the analysis of patterns in the wheezing data. Primarily chapter 6 is the basis for the extension of the work presented in this report. The parametric models of the incidence and prevalence are important to be able to remove confounders in the individual time-series, such that the interesting patterns becomes visible. The confounder-list may be extended with data on for example influenza-prevalence, which hopefully can remove more of the population related variation from the individual time-series.

Bibliography

[1] P.S. Albert. A transitional model for longitudinal binary data subject to nonignorable missing data. *Biometrics*, 56(2):602–608, 2000.

[2] D. Bates and M. Maechler. *Matrix: A Matrix package for R*, 2006. R package version 0.995-19.

[3] D. Bates and D. Sarkar. *lme4: Linear mixed-effects models using S4 classes*, 2006. R package version 0.995-2.

[4] H. Bisgaard. The Copenhagen Prospective Study on Asthma in Childhood (COPSAC): design, rationale, and baseline data from a longitudinal birth cohort study. *Annals of Allergy, Asthma & Immunology*, 93, October 2004.

[5] J.M. Bland and D.G. Altman. Statistics notes. The odds ratio. *BMJ*, 2000.

[6] Ø. Borgan, R.L. Fiaccone, R. Hendersen, and M.L. Barreto. Dynamic analysis of recurrent event data with missing observations, with application to infant diarrhoea in brazil. *Scandinavian Journal of Statistics*, 2006. In press.

[7] B. Christophersen, J. Heisterberg, T. Senderovitz, and R. Rabøl. *Kompendium i Farmakologi*. FADL's Forlag, 2nd edition, 2002.

[8] K. Conradsen. *En Introduktion til Statistik*, volume 1. Informathics and Mathematical modelling, 1999.

[9] K. Conradsen. *En introduktion til statistik*, volume 2. Informathics and Mathematical modelling, 6th edition, 2002.

[10] M.J. Crawley. *Statistics An Introduction using R*. John Wiley & Sons, Ltd, 2005.

[11] J.H. Curtiss. On transformations used in the analysis of variance. *The Annals of Mathematical Statistics*, 14(2):107–122, 1943.

[12] D.B. Dahl et al. *xtable: Export tables to LaTeX or HTML*, 2006. R package version 1.3-2.

[13] C. Dehlendorff. A review of statistical methods applied to asthma and wheezing symptoms in children. Preparatory Thesis, June 2006.

[14] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society*, Series B(39):1–38, 1977.

[15] P.J. Diggle, P.J. Heagerty, K.Y. Liang, and S.L. Zeger. *Analysis of Longitudinal Data*. Oxford University Press, 2002.

[16] L. Erichsen and P.B. Brockhoff. An application of latent class random coefficient regression. *Journal of Applied Mathematics and Decision Sciences*, 8(4):247–260, 2004.

[17] E.S. Garrett and S.L. Zeger. Latent class model diagnosis. *Biometrics*, 56:1055–1067, 2000.

[18] F.E. Jr Harrell. *Design: Design Package*, 2005. R package version 2.0-12.

[19] P.J. Heagerty and S.L. Zeger. Lorelogram: A regression approach to exploring dependence in longitudinal categorical responses. *Journal of the American Statistical Association*, 93(441):150–162, March 1998.

[20] C.M. Hurvich and C.L. Tsai. Regression and time series model selection in small samples. *Biometrica*, 76(2):297–307, June 1989.

[21] Frank E Harrell Jr and with contributions from many other users. *Hmisc: Harrell Miscellaneous*, 2006. R package version 3.0-12.

[22] B.R. Kirkwood. *Esentials of Medical Statistics*. Blackwell Science Ltd, 1988.

[23] F. Leisch. Flexmix: A general framework for finite mixture models and latent class regression in r. *Journal of Statistical Software*, 11(8), October 2004.

[24] M. Lesnoff and R. Lancelot. *aod: Analysis of Overdispersed Data*, 2006. R package version 1.1-10.

[25] H. Madsen. *Tidsrækkeanalyse*. IMM, DTU, 3rd edition, 1998.

[26] F.D. Martinez, A.L. Wright, L.M. Taussig, C.J. Holberg, M. Halonen, and W.J. Morgan. Asthma and Wheezing in the first six years of life. *The New England Journal of Medicine*, 332(3), January 1995.

[27] K.G. Nielsen and H. Bisgaard. Discriminative capacity of bronchodilator response measured with three different lung function techniques in asthmatic and healthy children aged 2 to 5 years. *American Journal of Respiratory and Critical Care Medicine*, 164(4):554–559, August 2001.

[28] U. Olsson. *Generalized Linear Models, An Applied Approach*. Studentlitteratur, 2002.

[29] C.N.A. Palmer, A.D. Irvine, A. Terron-Kwiatkowski, Y. Zhao, H. Liao, S.P. Lee, A. David R Goudie, Sandilands, L.E. Campbell, F.J.D. Smith, G.M. O'Regan, R.M. Watson, J.E. Cecil, S.J. Bale, J.G. Compton, J.J. DiGiovanna, P. Fleckman, S. Lewis-Jones, G. Arseculeratne, A. Sergeant, C.S. Munro, B.E. Houate, K. McElreavey, L.B. Halkjaer, H. Bisgaard, S. Mukhopadhyay, and W.H.I. McLean. Common loss-of-function variants of the epidermal barrier protein filaggrin are a major pre-disposing factor for atopic dermatitis. *Nature Genetics*, pages 441 – 446, March 2006.

[30] J.D. Petrucelli, B. Nandram, and M. Chen. *Applied Statistics for Engineers and Scientists*. Prentice-Hall, Inc., 1st edition, 1999.

[31] J. Pinheiro, D. Bates, S. DebRoy, and D. Sarkar. *nlme: Linear and nonlinear mixed effects models*, 2006. R package version 3.1-75.

[32] J.C. Pinheiro and D.M. Bates. *Mixed Effects Models in S and S-PLUS*. Springer, 2000.

[33] C.S. Poulsen, P.B. Brockhoff, and L. Erichsen. Heterogeneity in consumer preference data - a combined approach. *Food Quality and Preference*, 8(5):409–417, 1997.

[34] R.L. Prentice. Correlated binary regression with covariates specific to each binary observation. *Biometrics*, 44:1033–1048, 1988.

[35] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0.

[36] B.D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

[37] G.K. Robinson. That blup is a good thing: The estimation of random effects. *Statistical Science*, 6(1):15–32, 1991.

[38] D. Sarkar. *lattice: Lattice Graphics*, 2006. R package version 0.13-10.

[39] S.S. Shapiro and M.B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3):591–611, 1965.

[40] A. Steckelberg, A. Balgenorth, J. Berger, and I. Mühlhauser. Explaining computation of predictive values: 2 x 2 table versus frequency tree. a randomized controlled trial. *BMC Medical Education*, 4(13), August 2004.

[41] T. Therneau, T. Lumley, K. Halvorsen, and K. Hornik. *date: Functions for handling dates*, 2006. R package version 1.2-19. S original by Terry Therneau, R port by Thomas Lumley, Kjetil Halvorsen, and Kurt Hornik.

[42] S.N. Wood. *Generalized Additive Models - An Introduction with R*. Chapman & Hall/CRC, 1st edition, 2006.

[43] J. Yan. geepack: Yet another package for generalized estimating equations. *R-News*, 2(3):12–14, 2002.

[44] J. Yan and J. Fine. Estimating equations for association structures. *Statistics in Medicine*, 23:859–874, 2004.

[45] J. Yan and J.P. Fine. Estimating equations for association structures. *Statistics in Medicine*, 23:859–880, 2004.

[46] S.L. Zeger, K.Y. Liang, and B.P.S. Albert. Model for longitudinal data: A generalized estimation equation approach. *Biometrics*, 44:1049–60, 1988.