De transnationale Universiteit Limburg is een uniek samenwerkingsverband van twee universiteiten in twee landen:
de Universiteit Hasselt en Maastricht University

Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek
Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt

2011
2012

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

Flexible modeling of the cost evolution of pneumococcal infections

Promotor :
Prof. dr. Niel HENS

Promotor :
Prof. ADRIAAN BLOMMAERT

Abera Mulugeta Tohye

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

2011
2012

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

Flexible modeling of the cost evolution of pneumococcal infections

Promotor :
Prof. dr. Niel HENS

Promotor :
Prof. ADRIAAN BLOMMAERT

Abera Mulugeta Tohye

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

Maastricht University

universiteit hasselt
KNOWLEDGE IN ACTION

Certification

I declare that this thesis was written by me under the guidance and counsel of my supervisors.

……………………………………………… Date …………………………

Tohye Abera Mulugeta Student

This is to certify that this report was written by Tohye Abera Mulugeta under our supervision.

……………………………………………… Date …………………………

Prof. dr. HENS Niel Internal Supervisor

……………………………………………… Date …………………………

BLOMMAERT Adriaan External Supervisor

Acknowledgments

First of all I would like to thank to the almighty God, I could attain this Biostatistics ICP programme.

I would like to express my deepest thanks and respect to my internal supervisor professor Dr. Niel Hens, for his motivating advice, for sharing great ideas and for giving me the opportunity to work in a most inspiring environment. I was really motivated and happy when I am advised by him. I would like also to extend my deepest thanks to my external advisor Adriaan Blommaert, for his kindly continuous suggestions, encouragement and comments.

Special thanks also go to all my lecturers of Hasselt University. Last but not least, sincere thanks are extended to my families for their endless love and support.

Abstract

The aim of the study was to compare the evolution of medical costs over time for persons with a known positive isolate of streptococcus pneumoniae with the control group. The data set was used from the database from national alliance of Christian Sickness Fund (NACSF). We performed an age specific analysis using flexible modeling techniques. In this report, we present the flexible modeling techniques that take the correlation among measurements of the same subject in to account. We considered parametric and semiparametric models. First, mixed effect model with first-order fractional polynomial mean structure was fitted and, this was compared with the second-order fractional polynomial mean structure based on their AIC values. To generate fractional polynomial powers for the time variable, a MACRO was implemented. Once the fractional powers for the time variable were obtained, a mixed model was fitted using PROC MIXED statement in SAS. The fractional polynomial mixed model assumes the relationship between a covariate and a response is fully parametric. We extend this model to a semiparametric mixed effects models framework using penalized splines. One of the major benefits of the correspondence between penalized splines smoother and mixed model, is that software for mixed model analysis can be used for smoothing.

In conclusion, the health-care costs incurred by diagnosed pneumococcal patients are larger than those incurred by undiagnosed matched patients for all age groups.

Keywords: fractional polynomials, mixed effects model, penalized splines, Semiparametric mixed models, streptococcus pneumoniae

Contents

List of Figures

List of Tables

1. Introduction

1.1. Background

Streptococcus pneumoniae, or pneumococcus, is a type of bacteria that can attack different parts of the body and is a leading cause of illness and mortality among children worldwide and particularly in developing countries (Greenwood *et al,* 2007). It was estimated that 10.6 million children less than 5 years present with pneumococcal disease every year (Black *et al,* 2003). The most severe forms of the disease are the invasive diseases (IPD) that include meningitis, especially in infants and young children (an infection of the lining that covers the brain), especially in infants and young children, and bacteremia (blood stream infection). IPD affects the extreme ages of life, the young children and the elderly. The noninvasive diseases, comprising mainly pneumonia which is an infection of the lungs, and otitis (include middle-ear infection) are usually less severe, but considerably more common than IPD. These infections can be dangerous to very young children, the elderly and people with certain high risk health conditions. Otitis is mainly found in young children, and pneumonia affects all ages. Mortality and illness from IPD remain high nowadays, despite appropriate access to care and antibiotic treatment (Beutels *et al,* 2011). Although all age groups may be affected, the highest rate of pneumococcal disease occurs in young children and in the elderly population. In addition, persons with immune deficiencies are at an increased risk.

In some developing countries, for instance Southern India, 50% of infants has been occupied by Streptococcus pneumoniae by 2 months of age and 80% are carriers by the age of 6 months (Coles *et al,* 2001). A study in South Africa showed that the prevalence of carriage was 30%, 44%, 51% and 61% in children aged 6 weeks, 10 weeks, 14 weeks and 9 months, respectively (Mbelle *et al*, 1999).

The symptoms of pneumococcal pneumonia include high fever, cough, shortness of breath or chest pain and extreme tiredness while the symptoms of meningitis include stiff neck, high fever, headache, vomiting, extreme tiredness, and loss of appetite. Young children commonly develop middle ear infections when they have colds or other viral respiratory infections. The symptoms of Otitis include ear pain, fever, crying and runny nose.

Infections spread from one person to another the same way as cold spreads, by droplets passed through the air from coughing, sneezing and through touching unwashed hands. The disease occurs most often during the winter months. In this thesis, we focus on the medical costs incurred by people who acquire pneumococcal infections, which are so severe that they warrant diagnosis by a positive isolate.

The data considered in this thesis falls within the frame-work of continuous longitudinal data, and hence it was modeled by use of a mixed effects model which will be described in section 3.2. For the modeling of longitudinal data, mixed effects models are a widely used approach (Verbeke and Molenberghs, 2000). In this thesis we will focus on two modeling approaches; a mixed effect model with fractional polynomial mean structure to include the flexibility and a mixed model with splines as proposed by (Ruppert *et al*, 2003). First of all, the data was analyzed using linear mixed model where the mean structure of the cumulative costs was estimated using degree-one fractional polynomials. In order to compare for a possible improvement in fit, the mean structure was estimated again using second-degree fractional polynomials and then the two models were compared by use of the model selection criterion.

A major limitation of these methods is that the relationship of the longitudinal response to covariates is assumed fully parametric. To capture the irregular trends in the dataset, a more flexible model is needed. Several approaches can be used that allow flexibility in order to cope with the irregularities observed in the mean profiles. Thus, a more flexible model which is the semiparametric mixed model was performed. In recent years, this has placed a strong demand where flexible functional forms can be estimated from the data to capture possibly complicated relationships between longitudinal outcomes and covariates.

1.2. Objective of the study

The aim of the study was to compare the health-care costs between the two study groups. One is pneumococcal and the other is matched group.

1.3. Structure of the report

The report has six sections. In section 2 the data used in the analysis is described followed by section 3 that gives a general description of the methods used. Explanations of the results obtained from the methods applied to the data are given in section 4. And finally concluding remarks and discussion will be given in section 5. Sample statistical codes used for the analysis are included in the appendix.

2. Dataset

The dataset contains the total medical costs measured at fixed time points; they were followed over a monthly period. The database from national alliance of Christian Sickness Fund (NACSF) contains all resource use information of members of the largest sickness fund in Belgium. For this thesis, we considered medical costs of patients who have a positive culture taken between 1994 and 2004. The data set is used consisting of cases (members of NACSF) of whom a positive isolate of the pneumococcal was found, and controls (members of the NACSF) matched in terms of gender, age, municipality and social category. To study that evolution over time, the variable match was considered; which has two categories, 0 represents pneumococcal group and 1 the control group, the response variable is the total medical cost of an individual incurred throughout the study, and cumulative cost is the cumulative of the original costs. The dataset contains information on 1752 patients, from these 876 diagnosed and 876 undiagnosed. The patients were divided in two four age groups based on the expected differences in levels of differences in levels of severity of experienced pneumococcal disease. The lowest age group consists of medical cost measurements of 632 individuals over two groups: 316 individuals were measured in pneumococcal group while the remaining 316 comprise a control group and are displayed in Figure 1. The data represents 253 individuals in the pneumococcal group and the same number of persons in the control group under study with respect to the second age group. In the third age category, there are 113 numbers of individuals in the pneumococcal group and 113 in the control group, and the oldest age group consists of 194 individuals in the pneumococcal group and 194 in the control group. The lowest age category consists of individuals younger than five years of age and the second age group between 5 and 49 years, the third category contains individuals between 49 and 64 where as the oldest age group consists of patients aged 65 years or more. There are four data sets available for each of the age categories. Therefore age specific analysis is performed. Special interest lies in the oldest and youngest age groups, since these groups are known to have a highest risk of serious complications from their pneumococcal infections and from other unrelated illnesses.

From the descriptive plot presented in figure 1, it is virtually impossible to model using parametric techniques describing the different picks. Therefore, the original costs are transformed in to cumulative costs.

Let Y_{ij} be the cost for person i (this can be a diagnosed patient or a matched group) at month j, where j is negative before, and positive after diagnosis. The cumulative cost Z_{ij} for person i at month j is defined as

$$\mathrm{Z_{ij}} = \begin{cases} \sum_{k=0}^{j} \mathrm{Y_{ik}}, & \text{if } j \in \{0, 1, 2, \ldots\}; \\ \sum_{k=j}^{-1} \mathrm{Y_{ik}} & \text{if } j \in \{-1, -2, \ldots\}. \end{cases}$$

The cost Z_{ij} accumulates the original costs in two directions, from the time of diagnosis onwards. Before time point zero, the cumulative costs are decreasing until zero. From time zero onwards cumulative costs are increasing. Again, one can observe a difference between the pneumococcal group and the matched group, with higher cumulative costs in the pneumococcal group.

There are a few reasons why the cumulative costs are chosen to model rather than the original costs. First, cumulative costs summarize the total cost up to a particular point in time, which is highly relevant for practical interpretations. Next, the incremental process is highly irregular, in its occurrences over time (Figure 1). Cumulative costs smooth out these incremental costs. The number of incremental costs as well as the times of their occurrences is highly variable over patients. Therefore, the data are highly unbalanced over patients. Fractional polynomials can then be used to describe this cumulative process in a flexible way. In the plots the time span was reduced to 30 months before and after diagnosis. However, for most patients, much more information is available. In performing the analysis, the time span was reduced in each age group. This avoids mean and correlation structure selection being highly influenced by a few outliers that have observations far from time point zero and the SAS macro in statistical modeling did converge on the selection.

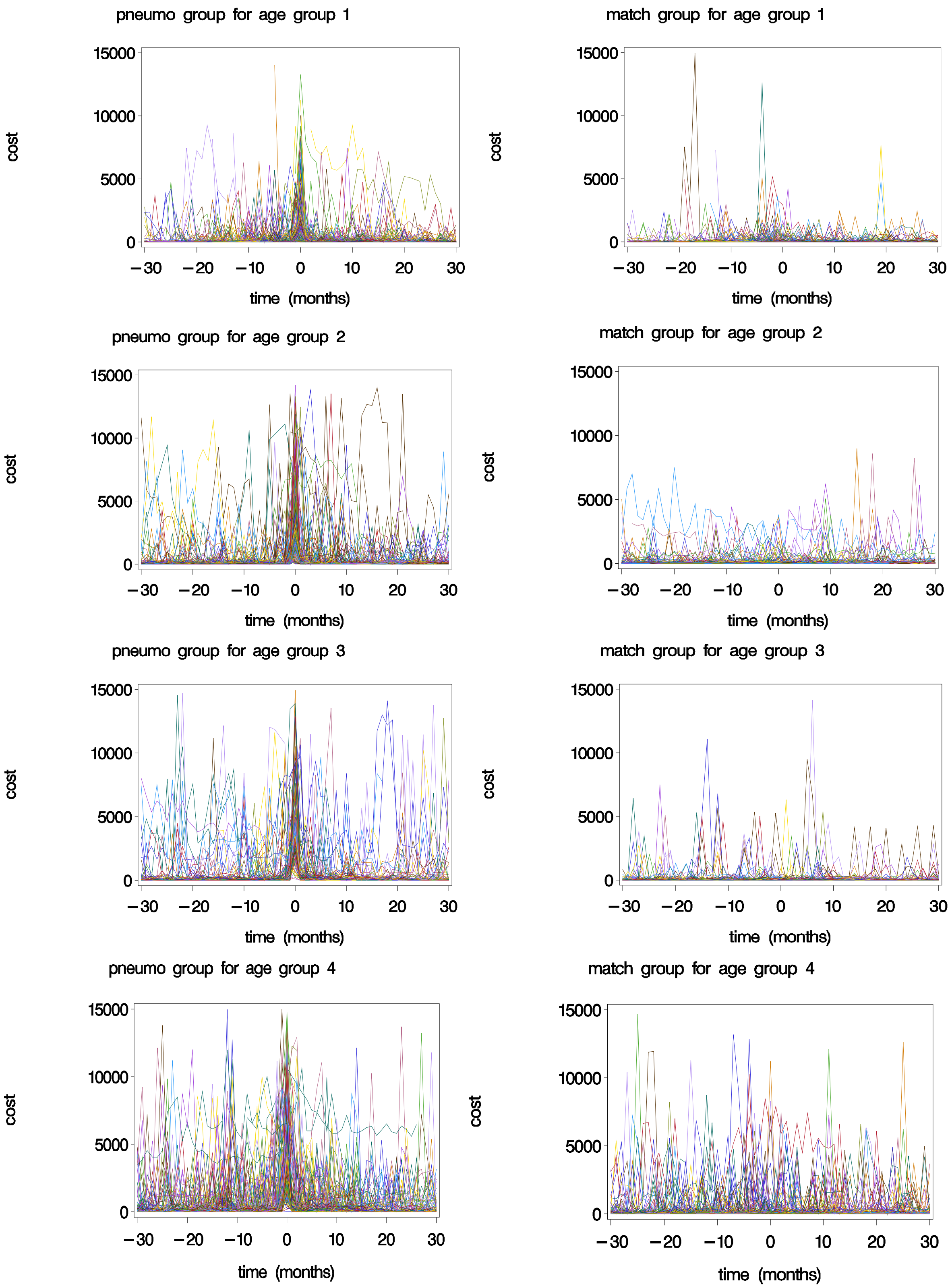


Figure 1: *Individual profiles in each age group for pneumococcal and matched patients based on the observed costs.*

Table 1: Minimum and maximum cluster sizes (subjects) for the different age groups

	Before		After	
Age group	Minimum	Maximum	Minimum	Maximum
1	1	34	1	97
2	1	34	1	97
3	1	120	1	70
4	1	132	1	84

Table two shows the minimum and the maximum cluster sizes for each age group. Individuals in age group one are measured up to 34 times monthly before diagnosis where as they are measured for a longer time after diagnosis, that is they are measured a maximum of 97 times. Similar numbers of measurements are found for the second age group as well. For those individuals who belong to the third age category, they are measured for a longer time before diagnosis compared to after diagnosis. Before the moment of diagnosis, the number of observations within subject is at most 132 while after diagnosis the maximum number of measurements is 84 for the oldest age category.

3. Statistical Methodology

3.1. Exploratory data analysis

In this report, an exploratory data analysis (EDA) was performed to explore the data. Before starting statistical modeling to obtain some useful insights in the structure of the data and an idea about a possible statistical model, descriptive statistics as well as graphical displays such as individual profiles and mean profiles by match group were used to address the research question.

3.2. Linear Mixed model

In this section, we briefly examine the general linear mixed-effects model. Longitudinal data arise frequently in many medical applications. They generally involve a collection of data at different time points for several subjects, and they are characterized by the dependence of repeated observations over time within the same subject. Statistical models taking the repeated nature of the data in to account will be presented. Since observations coming from the same subject tend to be more alike than observations from different subjects, they are said to be correlated. This correlation needs to be taken into account when analyzing longitudinal data. The general linear mixed model is a commonly used statistical tool to study the relationship between a normally distributed dependent variable and one or more independent variables. The name mixed model comes from the fact that the model contains both fixed-effects parameters and random-effect parameters. Linear mixed models provides a flexible framework to model longitudinal data parametrically (Laird and Ware, 1982).

A linear mixed effects model (LMM) then assumes that Y_i satisfies:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i$$

where $\mathrm{Y_i}$ is the n_i dimensional vector of measurements available for subject $i = 1, \dots, N$, X_i and Z_i are the $(n_i \times p)$ and $(n_i \times q)$ dimensional matrices of the predictor variables and the random effect variables respectively, β is a p dimensional vector of fixed effects, $\mathrm{b_i}$ is the q-dimensional vector of subject specific random effects and ε_i is an n_i dimensional vector of residual components. Finally, D is a general $q \times q$ covariance matrix for random effects and Σ_i corresponds to $(n_i \times n_i)$ covariance matrix for the error terms.

The distributional assumptions made by the mixed model are as follows:

$$\begin{cases} \boldsymbol{b}_i \sim N(\boldsymbol{0}, \mathrm{D}), \\ \boldsymbol{\varepsilon}_i \sim N(\boldsymbol{0}, \Sigma_\mathrm{i}), \\ \boldsymbol{b}_i \text{ and } \boldsymbol{\varepsilon}_i \text{ are independent.} \end{cases}$$

and they are independent. The fixed effects have population-average regression coefficients and the b_i*'s* (random effects) have subject specific regression coefficients.

3.3. Fractional Polynomial with Mixed Model

Fractional polynomials allow very flexible models, extending the classical polynomials. Fractional polynomials were proposed by Royston and Altman (1994) as a flexible parametric modeling approach. In addition to having the flexible properties of the classical polynomial models, the fractional polynomial models allow for non-integer powers and logarithmic functions and products of these (Aerts, 2006). Royston and Altman argued that in practice, fractional polynomials of order higher than 2 are rarely needed, since higher degrees do not mostly improve the model fit. The powers are selected so that conventional polynomials are a subset of the family. The powers of a best fractional model are selected from a suggested set of {-2, -1, -0.5, 0, 0.5, 1, 2, 3}. Thus all possible combinations of two powers from this set were considered and the combination that gave the model with the lowest Akaike information criteria (AIC) was chosen.

A fractional polynomial in x of degree m is defined as any function of the form

$$\phi_m(X; {}^{\beta}; p_1, p_2, \ldots, p_m) = {}^{\beta}{}_0 + \sum_{j=1}^{m} {}^{\beta}{}_j H_j(X),$$

where the degree m is a positive integer, p is a real-valued vector of powers with $p_1 \leq \cdots \leq p_m$ and ${}^{\beta}{}_0, {}^{\beta}{}_1, \ldots, {}^{\beta}{}_m$ are real-valued coefficients and $H_j(X)$ is a transformation function given by

$$H_j(X) = \begin{cases} \ln(X) & if\ p_j = 0\ and\ p_j \neq p_{j-1}, \\ X^{p_j} & if\ p_j \neq p_{j-1}, \\ H_{j-1}(X)\ln(X) & if\ \ p_j = p_{j-1}. \end{cases}$$

A great advantage of fractional polynomials over classical polynomials is that they provide a wide range of functional forms and their behavior near the extreme values is often more reasonable. Fractional polynomials of Royston and Altman are frequently used. Another advantage of fractional polynomials is that they are straightforward to fit using standard methods.

It should be noted at this point that the model discussed here has a restriction in the relationship between the response and the covariate is fitted parametrically. This does not make the flexibility of models like fractional polynomials optimal. In furthering our search, we also considered another model, the semi-parametric mixed model as an alternative which will be discussed in latter sections. We used a SAS MACRO to produce the yearning power of the time. The obtained powers are therefore used in the proc mixed procedure with both random and repeated statements.

3.4. Derivation of cumulative costs

Due to the highly skewed nature of the cumulative costs, logarithmic transformation is used and the model is fitted on this transformed response. Once the model is fitted on the logarithm of cumulative costs, we need to back transform for the original costs. In this case, we first transform to the cumulative costs. In order to compare the population-averaged marginal evolutions of the two groups on the original cost level, additional computations are needed. The marginal expected evolution of the cumulative costs measured at time point t_{ij} is given by

$$E[CS_{ij}] = E\left[E[CS_{ij} | b_i]\right] \neq \exp[{}^{\beta}{}_0 + {}^{\beta}{}_1 match + {}^{\beta}{}_2 t^p + {}^{\beta}{}_3 match * t^p]$$

Since we have random effects in the non-linear function, calculation of the above equation requires numerical averaging technique; we follow this procedure with 1000 draws. We fitted by randomly drawing 1000 realized values for the random effects b_i, taken from a bivariate normal distribution with mean vector zero and with covariance matrix equal to the fitted random-effects covariance matrix D given by: $var(b_i) = D = \begin{pmatrix} d11 & d12 \\ d21 & d22 \end{pmatrix}$.

The Cholesky decomposition of the covariance matrix (D), defined as the upper triangular matrix L such that $L`L = D$ and needed in the SAS code for drawing the 1000 random vectors b_i is given by: $L = \begin{pmatrix} l_{11} & l_{12} \\ 0 & l_{22} \end{pmatrix}$. For each of the 1000 realized random vectors b_i, the conditional expectation $\exp[{}^{\beta}{}_0 + b_{oi} + {}^{\beta}{}_1 match + {}^{\beta}{}_2 t^p + {}^{\beta}{}_3 match * t^p + b_{1i}]$ is computed, with fixed effects replaced by their fitted values. Once the logarithm of cumulative costs are transformed in to cumulative costs in this way, secondly interest also lies in the derivative of the cumulative cost function to back transform to the original costs. An estimate for unconditional mean at a given time point is then obtained from averaging the 1000 conditional means, i.e.,

$$\hat{E}[CS(t)] = \frac{1}{1000} \sum_{i=1}^{1000} \exp[{}^{\beta}{}_0 + b_{oi} + {}^{\beta}{}_1 match + {}^{\beta}{}_2 t^p + {}^{\beta}{}_3 match * t^p + b_{1i}] \, ,$$

where the random slope b_{1i} belongs to t^p.

3.5 Radial Basis function

There is a clear essence to be able to handle the nonlinear relationships revealed in figure 1 effectively through more flexible techniques such as splines. Although fractional polynomial terms can be used to handle nonlinearities, it should be kept in mind that their use can require a good deal of time. Thus, a more flexible approach which is situated within the mixed-model framework is through penalized splines (Ruppert *et al,* 2003). In this section we will look at some ways of freeing oneself of the restrictions of parametric models.

The basis function which we used is the radial basis. Ruppert *et al* (2003) defines a radial basis function by

$$|t - K_k|^p = r(|t - K_k|),$$

where $r(u) = u^p$ for some function *r*, and a degree *p*. This shows that the basis functions $|t - K_k|^p$ *(*$1 \leq k \leq K$*)* depend only on the distance ($|t - K_k|$) and the function r. Figure 2 displays the function $|t - k_k|$ for 10 equally spaced knots. This is not the only basis function.

Another set of basis functions with this property is the truncated lines basis with knots at $k_1, \ldots, k_K$. In principle, a change of basis does not change the fit. One of reasons for selecting one basis over another is ease of implementation (Ruppert *et al,* 2003). We considered radial basis function, the method of semiparametric mixed model which will be discussed in the next section, is easier to implement in the Glimmix procedure in SAS with radial basis as presented in the appendix. We can generalize this to a spline model of general degree.

According to Ruppert, different choices of the smoothing parameter lead to different estimated models. When a linear mixed model is used as a scatter plot smoother, one does not need to use any additional procedure in order to select the smoothing parameter. The amount of smoothing is determined by $\lambda = \sigma_\varepsilon^2 / \sigma_b^2$.

Linear radial basis

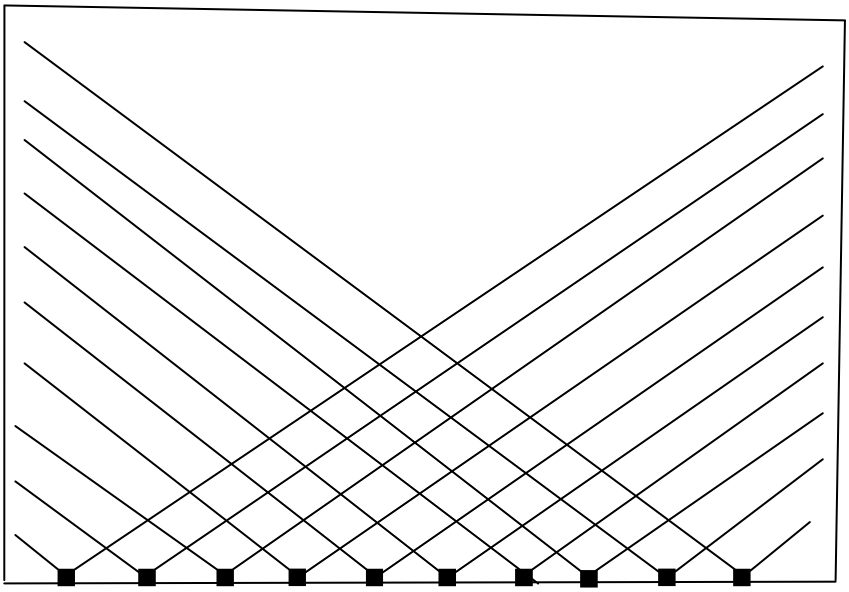


Figure 2*: linear radial basis functions where the positions of the knots are indicated by the black squares with 10 equally spaced knot points.*

3.6. Semiparametric mixed models

A simple and straightforward method to fit splines is by considering the coefficient of each knot as fixed effect, usually referred to as regression spline. However, this approach tends to over fit the data, leading to computational problems. This can be overcome by including splines in the mixed model framework, meaning that treating each knot point coefficient as random (Ruppert *et al,* 2003). As discussed in section 3.5 an appealing alternative to fractional polynomials is to model the irregular trends with a semiparametric smooth function, *f*(*t*), which can be estimated with penalized splines. Let Y_{ij} denote the response taken from subject $i, i = 1, \ldots, m$ at time t_{ij} ($j = 1, \ldots, n_i$). The model of interest can be expressed as $Y_{ij} = f(t_{ij}) + b_{0i} + \varepsilon_{ij}$ for a smooth function $f(.)$ and subject-specific random intercepts b_{0i} accounting for the clustered nature of the observations. The penalized spline representation, based on a linear radial basis, can be written as

$$Y_{ij} = \underbrace{\beta_0 + \beta_1 t_{ij} + \textstyle\sum_{k=1}^{K} b_k |t_{ij} - k_k|}_{f(t_{ij})} + b_{0i} + \varepsilon_{ij} = f(t_{ij}) + b_{0i} + \varepsilon_{ij}$$

Where $k_1, \ldots, k_K$are a set of distinct knots in the range of t_{ij}, $b_k \sim N(0, \sigma_b^2)$ and $b_{oi} \sim N(0, \sigma_{b_0}^2)$. The two sets of random effects b_k and b_{0i} are assumed to be independent. This enables us to write the above equation as a semiparametric mixed model similar to $\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i$where now

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}_1 \boldsymbol{1} & 0 & \ldots & 0 \\ \boldsymbol{Z}_2 & 0 & \boldsymbol{1} \ldots & 0 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ \boldsymbol{Z}_m & . & . & \boldsymbol{1}_m \end{bmatrix}, \quad \boldsymbol{Z}_i = \begin{bmatrix} |t_{i1} - k_1| & \cdots & t_{i1} - k_K \\ \vdots & \ddots & \vdots \\ |t_{in_i} - k_1| & \cdots & |t_{in_i} - k_K| \end{bmatrix}, \quad b = [b_1, \ldots, b_K, b_{01}, \ldots, b_{0m}]^T$$

and

$G = \begin{bmatrix} \sigma_b^2 \boldsymbol{I} & 0 \\ 0 & \sigma_{b0}^2 \boldsymbol{I} \end{bmatrix}$, the vector of fixed effects $\beta = (\beta_0, \beta_1)^T$ and the corresponding design matrix $\boldsymbol{X} = [1 \ \ t_{ij}]$. Fitting penalized splines by the mixed model approach has some appealing advantages, such as the automatic determination of the smoothing parameter and the flexibility with which the models can be extended. The method is usually implemented in the SAS procedure GLIMMIX.

The semiparametric model discussed above implies that the mean response for each group can be represented by an additive model of two components, a linear component and a smooth component.

We are interested in investigating whether there is a difference between the two study groups. In this modeling, we fit two separate curves which are assumed to be the fixed parts are different across the two groups but the non-parametric component, responsible for the smoothing, is identical. The penalized spline representation of the model in the two groups can be expressed as

$$Y_{ij} = \begin{cases} \beta_0 + \beta_1 t_{ij} + \sum_{k=1}^{K} b_k \left| t_{ij} - k_k \right| + b_{0i} + \varepsilon_{ij}, & \mathit{Pneumococcal\ group,} \\ (\beta_0 + \beta_{01}) + (\beta_1 + \beta_{11}) t_{ij} + \sum_{k=1}^{K} b_k \left| t_{ij} - k_k \right| + b_{0i} + \varepsilon_{ij}, & \mathit{Matched\ group.} \end{cases}$$

where b_{0i} is a subject-specific random effect, β_{01} is the difference in group specific intercepts, β_{11} is the difference in group specific slope and ε_{ij} are residuals. The covariance matrix for the random effects $b_1, \dots, b_K, b_{01}, \dots, b_{0n})$ is defined as

$$\boldsymbol{G} = \begin{bmatrix} \sigma_b^2 \boldsymbol{I}_K & 0 \\ 0 & \sigma_{bo}^2 \boldsymbol{I}_K \end{bmatrix},$$

where $\sigma_b^2 = var(b_k)$ and $\sigma_{b0}^2 = var(b_{0i})$.The model takes into account within and between-subject variability, as well as variability arising from smoothing.

3.7. Software

The main software used for data analysis was SAS, mostly PROC MIXED to fit linear mixed models. Moreover, important graphs were built using R package. The analysis of the spline modeling based on the radial basis function was performed using the GLIMMIX procedure in SAS.

4. Results

In this section we present the results obtained from the different techniques applied to address the research question. Based on the objective of the study, most of the exploratory results are presented for each of the groups. This is because it might be informative for the methods employed.

4.1. Exploratory data analysis

The table shown bellow represents descriptive statistics of costs at each time points for pneumococcal and the control group for the lowest age category. In descriptive statistics of the data; the mean, fifth, ninety fifth percentile values and the associated number of observations conditional on the particular group were presented. For the lowest age category, a total of 278 individuals are diagnosed with pneumococcal infection and their medical costs were compared with the control group. At the moment of diagnosis, cost differs among individuals. The cost of pneumococcal patients ranged between 0 euro and 18420.34 euro with mean cost of 1874.93 euro while that of the control group ranged between 0 euro and 2940.37 euro with a mean of 61.233 euro. Thus on average the medical cost for pneumococcal patients is larger than that of matched group (Table 2). The mean cost for pneumococcal patients are higher in every measurement time for before and after diagnosis. Only the results of before diagnosis are presented here, a full table showing before and after diagnosis is given in Appendix A (Table 8).

Table 2: Descriptive statistics of cost at each time point for the lowest age group

		Pneumococcal					Match			
Time points	n	Mean	Median	P_5	P_{95}	n	Mean	median	P_5	P_{95}
-30	51	245.422	16.86	0	2304.33	51	99.073	15.47	0	585.33
-25	61	262.088	15.12	0	341.83	61	64.338	23.95	0	150.8
-20	85	182.608	17.85	0	660.96	85	39.447	0	0	159.38
-15	119	113.076	20.45	0	670.73	119	72.353	15.74	0	243.49
-10	170	142.198	22.525	0	946.94	170	74.799	0	0	595.29
-5	248	207.914	22.74	0	735.93	248	57.897	17	0	136.06
0	278	1874.93	1245.08	15.3	6277.77	278	61.233	15.51	0	218.52

From table 2 it can be seen that except at the moment of diagnosis the fifth percentile is zero in every measurement time for diagnosed pneumococcal patients, for the control group it is zero in every measurement time, which indicates that large number of individuals have zero costs (incurred no medical costs). The 95th percentiles give additional information in that, patients diagnosed for pneumococcal infections incurred higher health-care costs than that of the control persons in all the time points in both directions before and after diagnosis. Results for the other age groups are not shown here, but in all age groups the pneumococcal patients incurred more costs than those who belong to the control group. This might indicate that group has an impact on the cost; that is it seems that there is group difference. The cost data are taken every month and the number of measurements taken per individual are not fixed therefore we have unbalanced data set.

Figure 1 depicts the individual profiles of the cost of individuals as a function of month at each age categories for each of the matched and pneumococcal groups. The profiles show substantial between as well as within subject variability. Given the individual profiles in Figure 1, it appears a suitable parametric function to describe the evolution may not be easily assumed. Therefore the cumulative costs are used. Figure 2 shows the profiles of the cumulative costs across age categories for pneumococcal as well as for the matched groups, which reflects the overall increasing trend of cumulative costs along measurement time, that means measurements are taken on a monthly basis. The cumulative costs look higher in pneumococcal patients than matched persons. Different individuals follow different evolutions and there seems to be individuals in the pneumococcal groups are different from those of the matched groups in all age categories. Notice that the cumulative costs increases as time increases in both directions, before and after diagnosis, this is expected as the cumulative costs are the sum of costs at some time points. Looking at the plot, the variability between individuals increases over time. Although these plots also give us the variability at a given time, it also gives about the correlation between measurements of the same subject. The profiles show substantial between as well as within-subject variability. The profiles suggest a need for a flexible model, which would be able to capture the functional dependence of cumulative costs on time.

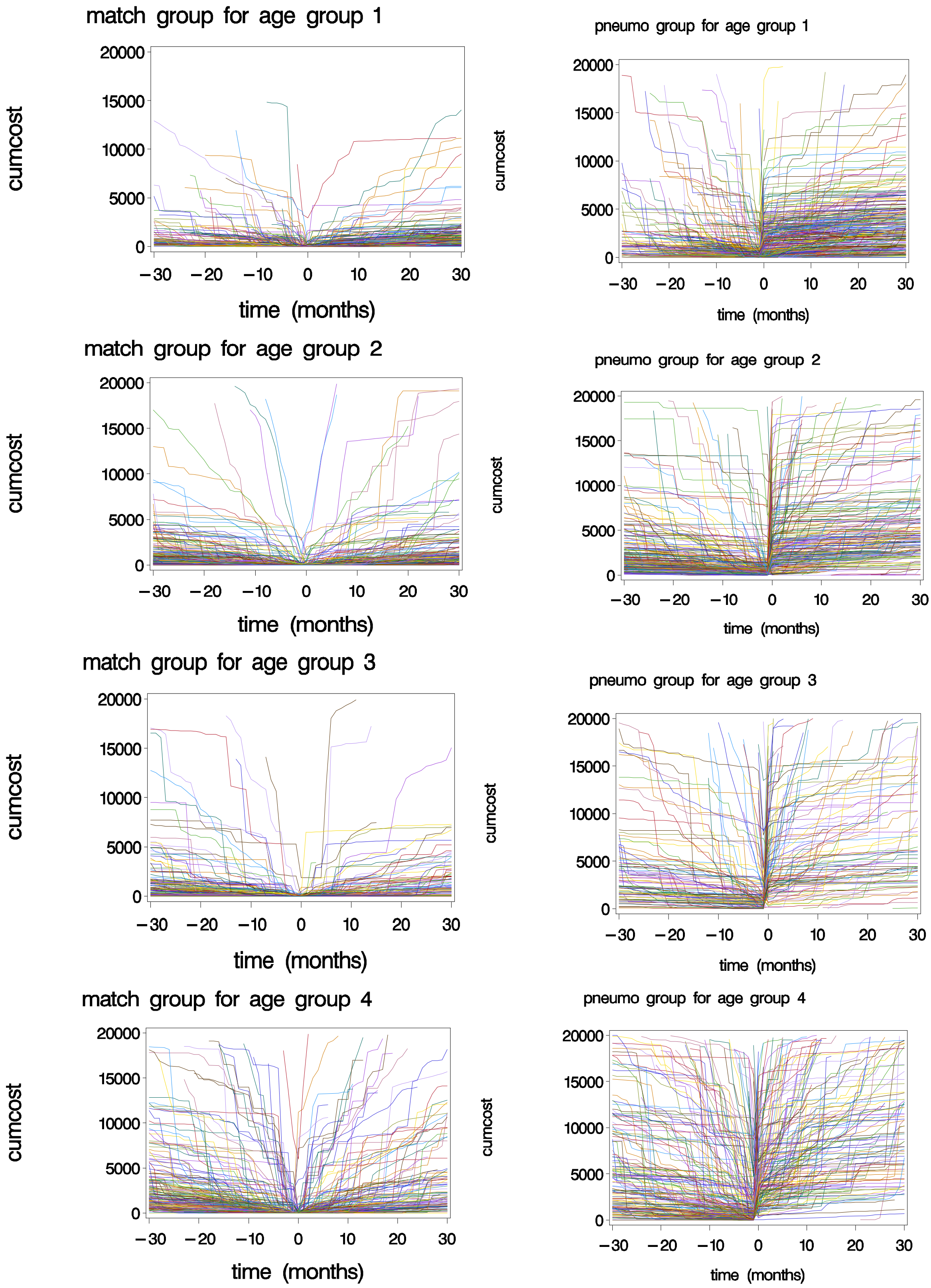


Figure 3*: Individual profiles in each age group for pneumococcal and matched patients based on the cumulative costs.*

As we can see from the plot, the linearity assumption is not reasonable for the data, and so the linear mixed model should be at least extended to the fractional polynomial mean structure. Based on the transformed costs a mixed model with fractional polynomial mean structure was used. We observe that the individual profile of the cumulative cost against time in both the pneumococcal and the control groups present a non linear trend. So in order to capture the flexibility, discovered by the profiles of the two groups, we generated fractional polynomial powers.

From the profiles, the variable group seems to have an effect that is there seems to be a difference between pneumococcal and matched groups. This has to be proven statistically that is we need to investigate formal statistical tests.

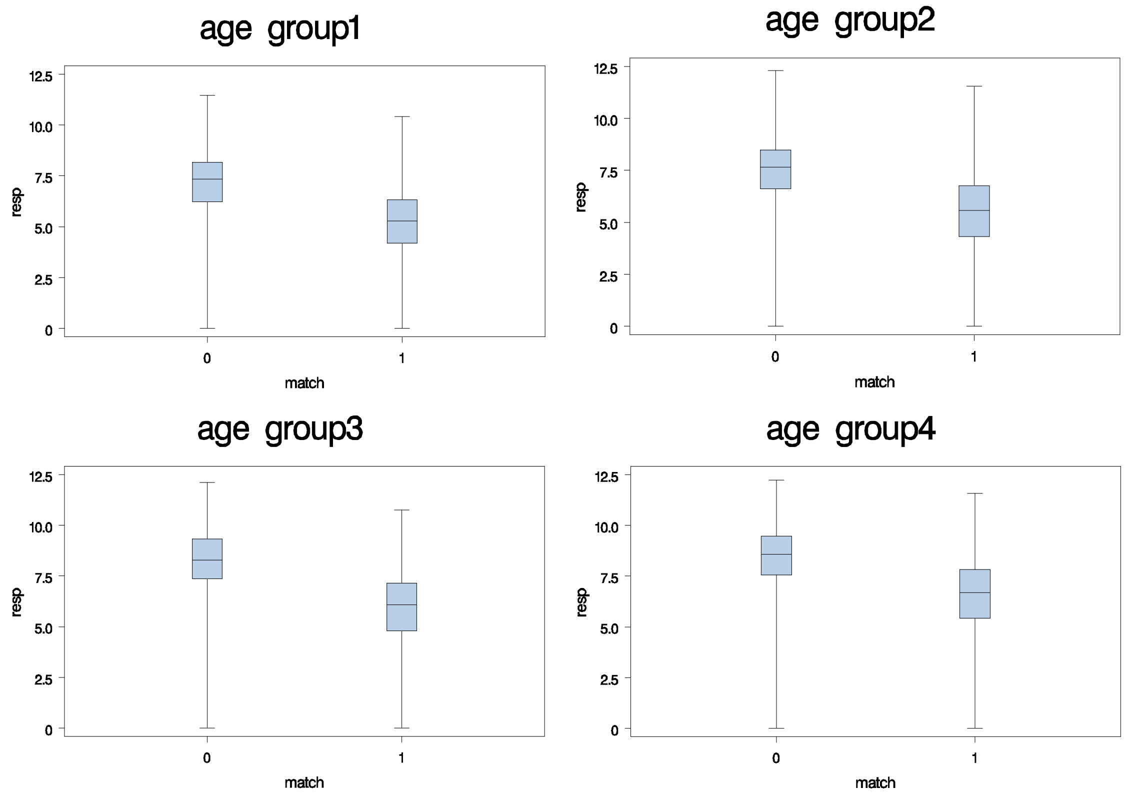


Figure 4*: Box-plots of the log cumulative cost with respect to each level of group*.

Box-plots of the cumulative costs on logarithmic scale with respect to each level of group (named match in the data set) are presented in Figure 4. The plots are used to identify whether the average cumulative cost is related to the match group. It is evident that there is some difference in the mean cumulative cost between pneumococcal and matched persons; the mean cumulative costs are slightly higher in pneumococcal patients than individuals who belong to the control group.

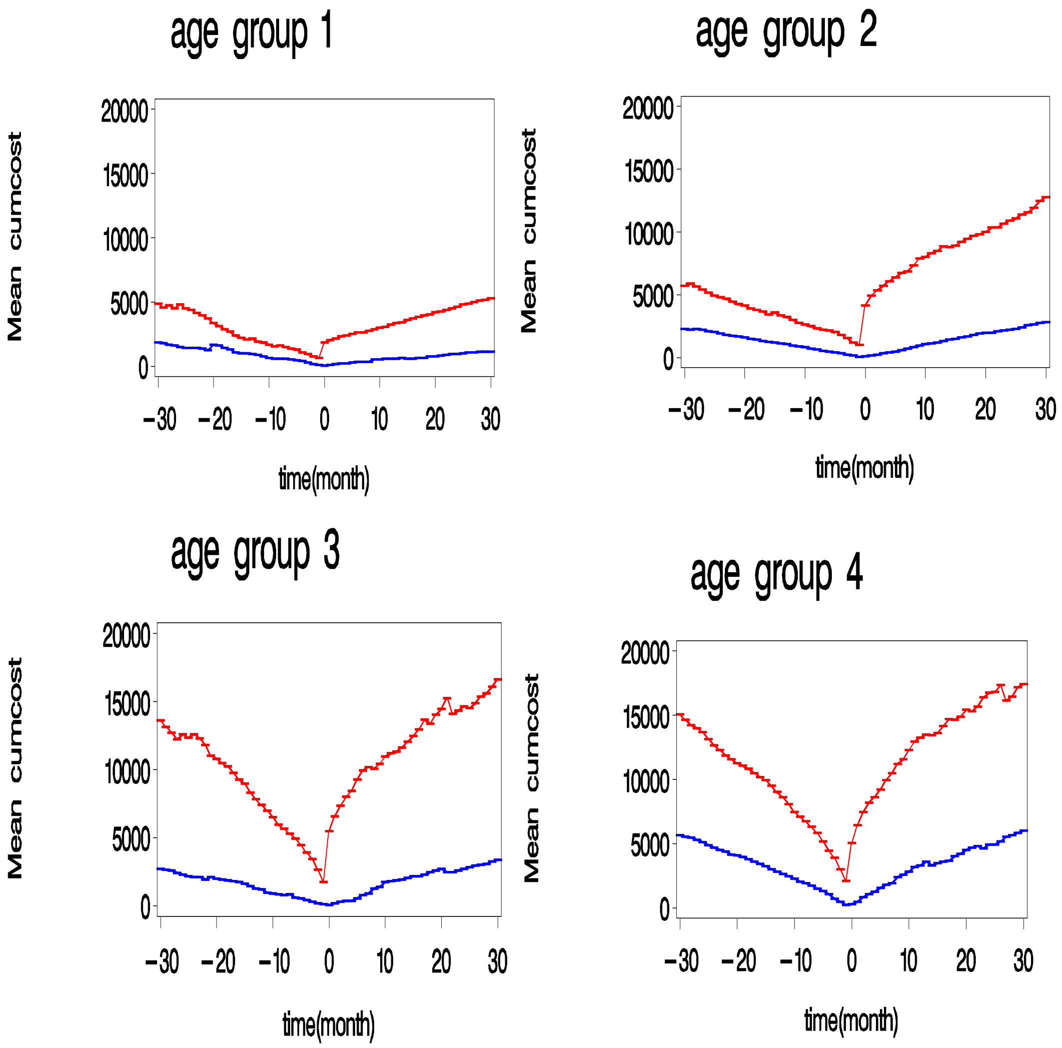


Figure 5*: mean profile of cumulative cost over time by match group in each age category.*

The average evolution describes how the profile for the population evolves over time. The results of this exploration will be helpful in order to choose a fixed effects structure for the mixed model. Besides plotting the response over time, it is also useful to include the two groups on the same graph to illustrate the relationship between the response (cumulative cost) and an explanatory variable (group) over time. The red and blue colors correspond with pneumococcal and control groups respectively. This allows us to make a rough comparison between the two groups. Looking at the figure, the cumulative costs of the pneumococcal group are consistently higher than the control group in all age categories from the beginning to the end of the study. Moreover the average profiles indicate an increase over time in both directions before and after the moment of diagnosis. Furthermore, there appears to be a relatively large difference between the pneumococcal and control groups. Note that a parametric modeling technique for these mean profiles might not be easily determined. Hence the need to use more flexible modeling, semi-parametric modeling techniques, is apparent.

Figure 6 depicts the cost over time under each of the age groups for pneumococcal and control groups. It gives further information in order to compare the costs of the four different age groups.

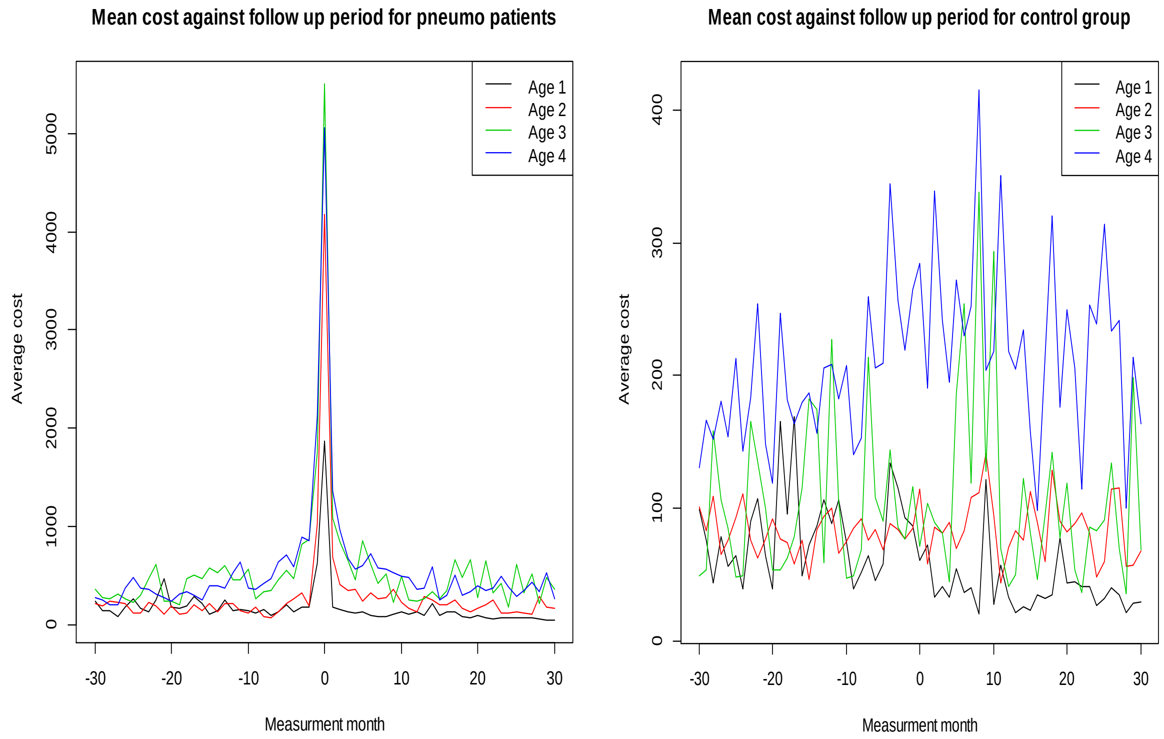


Figure 6*: mean cost against measurement time across age categories for pneumococcal and matched groups.*

The different age groups are represented in different colors. As it can be seen from the graph, after diagnosis, younger patients have the lowest cost whereas before diagnosis and some measurement times the second and third age groups have the lowest cost for the control group. For the control group, the highest age category incurred higher costs than any other groups throughout the measurement times before and after diagnosis. When we look at the plot of the pneumococcal patients, again after diagnosis the lowest age category incurred the lowest cost. Furthermore the oldest age category incurred highest costs but it is not regular throughout the measurement times. In general there is a slight difference in cost between the age categories for pneumococcal and control groups. In both groups, the oldest age group has the highest cost followed by the 3rd age group.

4.2. Covariance structure

One of the basic features of mixed models is the covariance structure such as simple, compound symmetry, AR(1) and unstructured. According to Molenberghs and Verbeke (2005), in the case of balanced data, i.e., when a fixed number of measurements are taken for all subjects and when measurements are taken at fixed time points, a useful covariance structure is the unstructured structure. Depending on the context and actual data at hand, other choices may be appropriate. A first-order autoregressive model assumes that the covariance between two measurements y_{ij} and y_{ik} from the same subject i is of the form $\sigma^2\rho^{|t_{ij}-t_{ik}|}$ for unknown parameters σ^2 and ρ. Another covariance structure is compound symmetry, which assumes that the correlation between observations is constant over time. For the choice of the covariance structure to the data at hand, we tried to compare different covariance structures based on the Akaike Information Criterion Molenberghs and Verbeke (2005) in order to select the best covariance structure. The unstructured covariance structure, ar(1) and compound symmetry leads to convergence problem. Hence the simple covariance structure was applied. Once the covariance structure is selected, and given that the mean profile revealed a non linear trend over time, we had to come out with the ideal power of the time variable. According to Royston and Altman, low order polynomials offer a limited family of shapes, and high order polynomials may fit poorly at the extreme values of the covariates. Fractional polynomials were therefore proposed as a solution by Royston and Altman (1994).

4.3. Modeling the results based on degree one fractional polynomial

The model described in Section 3.2 was fitted to the data. As a starting point, we fit a mixed model with the first-order fractional polynomial mean structure to the data. Moreover, second degree fractional polynomials were also taken in to account for possible improvements in fit.

Table 3: Selected fractional powers.

	Age1	Age2	Age3	Age4
Before	0	0	0	0
After	0	0.5	0.5	0.5

As shown in Figure 13 in the appendix, the cumulative cost data is not normally distributed. Therefore, we used a logarithmic transformation. Once the covariance structure is selected, the mean structure of the log cumulative costs is estimated using fractional polynomials. An age specific analysis was conducted for before and after diagnosis separately. The selected fractional powers for each age group for before and after diagnosis are presented in Table 3.

Table 4: Parameter estimates (S.E.) based on the degree-one fractional polynomial model

Effect	parameter	Age1 after		Effect	Age1 before	
		Estimate(S.E.)	p-value		Estimate	p-value
Intercept	β_0	2.5728 (0.1759)	<.0001	Intercept	2.0011(0.1295)	<.0001
Match	β_1	3.4573 (0.2488)	<.0001	Match	2.8642 (0.1832)	<.0001
t^{p_1}	β_2	1.1344 (0.0492)	<.0001	t^{p_1}	1.5357 (0.0447)	<.0001
$t^{p_1} *$match	β_3	-0.6143(0.0695)	<.0001	$t^{p_1} *$match	-0.7003 (0.0632)	<.0001
Covariances				**Covariances**		
Var(b_{1i})	d_{11}	9.0571 (0.5916)		Var(b_{1i})	4.4873(0.2813)	
cov(b_{1i}, b_{2i})	$d_{12} = d_{21}$	-2.2982 (0.1625)		cov(b_{1i}, b_{2i})	-1.0404(0.0842)	
Var(b_{2i})	d_{22}	0.6936(0.0477)		Var(b_{2i})	0.4697(0.0342)	
		Age2 after			Age2 before	
Intercept	β_0	1.9551(0.1908)	<.0001	Intercept	1.7619(0.1695)	<.0001
Match	β_1	5.2484(0.2699)	<.0001	Match	3.1333(0.2398)	<.0001
t^{p_1}	β_2	0.9904(0.0355)	<.0001	t^{p_1}	1.4486(0.0460)	<.0001
t^{p_1}*match	β_3	-0.7551(0.0502)	<.0001	$t^{p_1} *$match	-0.6682(0.0651)	<.0001
Covariances				**Covariances**		
Var(b_{1i})	d_{11}	8.4435 (0.6520)		Var(b_{1i})	6.6730 (0.4492)	
cov(b_{1i}, b_{2i})	$d_{12} = d_{21}$	-1.2527 (0.1125)		cov(b_{1i}, b_{2i})	-1.4559 (0.1112)	
Var(b_{2i})	d_{22}	0.2755(0.0226)		Var(b_{2i})	0.4770 (0.0328)	
		Age3 after			Age3 before	
Intercept	β_0	2.5064 (0.2174)	<.0001	Intercept	2.1972 (0.2226)	<.0001
Match	β_1	5.3744 (0.3075)	<.0001	Match	4.0793 (0.3148)	<.0001
t^{p_1}	β_2	0.9759 (0.0461)	<.0001	t^{p_1}	1.4371 (0.0596)	<.0001
$t^{p_1} *$match	β_3	-0.7202 (0.0652)	<.0001	$t^{p_1} *$match	-0.8186 (0.0838)	<.0001
Covariances				**Covariances**		
Var(b_{1i})	d_{11}	5.0388(0.504)		Var(b_{1i})	5.1428 (0.5100)	
cov(b_{1i}, b_{2i})	$d_{12} = d_{21}$	-0.7723(0.0933)		cov(b_{1i}, b_{2i})	-0.9239 (0.1183)	
Var(b_{2i})	d_{22}	0.2100(0.0234)		Var(b_{2i})	0.3539 (0.0361)	
		Age4 after			Age4 before	
Intercept	β_0	3.6544(0.1663)	<.0001	Intercept	3.2683 (0.1799)	<.0001
Match	β_1	4.3457 (0.2352)	<.0001	Match	2.8175 (0.2545)	<.0001
t^{p_1}	β_2	0.8824 (0.0319)	<.0001	t^{p_1}	1.3063 (0.0433)	<.0001
t^{p_1}*match	β_3	-0.5941 (0.0452)	<.0001	$t^{p_1} *$match	-0.5655 (0.0433)	<.0001
Covariances				**Covariances**		
Var(b_{1i})	d_{11}	4.9775(0.4366)		Var(b_{1i})	6.0505(0.4628)	
cov(b_{1i}, b_{2i})	$d_{12} = d_{21}$	-0.7095 (0.0766)		cov(b_{1i}, b_{2i})	-1.0997(0.1008)	
Var(b_{2i})	d_{22}	0.1669(0.0161)		Var(b_{2i})	0.3424 (0.0271)	

The likelihood ratio test was used under REML for the inference of variance components. We considered the random-intercept and slope model, that is, a mixed model where the subject specific effects are intercepts and slopes. We used likelihood ratio test to test if only random intercept model can be specified. Let us now take first the analysis of before diagnosis for the first age group. A model with only random intercept had minus twice log likelihood of 20987.2 while a model with both random terms included had a 17821.4. This will be compared with a 50:50 mixture of two chi-square distributions with 1 and 2 degree of freedom, respectively. The test statistic is 3165.8, yielding a highly significant result, implying the need for random slope. Similarly for each age group for before and after diagnosis, a model with a random intercept only is rejected. Therefore models with both random intercept and random slope were fitted across age groups.

Table 4 presents the parameter estimates and their corresponding standard errors of the fitted mixed model based on the degree one fractional polynomial mean structure for each age group. We obtained a significant main as well as interaction effects. The interaction effect can be interpreted as there is a difference in cumulative cost over time between the pneumococcal and the control group.

4.4. Second-degree Fractional Polynomials

The fractional polynomial powers based on the restricted set of {-2, -1, -0.5, 0, 0.5, 1, 2, 3} were selected. For degree two fractional polynomials, all possible combinations of two powers from this set were considered and the combination that gave the model with the lowest AIC was selected. The model with the smallest AIC value is chosen. The best fit fractional polynomials of degree 2 have the same powers (0, 0) for the lowest age group at time before the moment of diagnosis while after diagnosis the powers are (0.5, 0.5).

Table 5: Selected powers- degree two

	Age1	Age2	Age3	Age4
Before	0, 0	-2,0	0, 0	0, 0
After	0.5, 0.5	-1, 0	-1, 0	0, 0

A mixed model with degree two fractional polynomial mean structures were fitted with the selected powers from Table 5. The same procedures were followed for model building as in degree one fractional polynomial case. The models based on degree one and degree two fractional polynomial mean structures were compared based on the information criterion. Models based on degree two fractional polynomials had smaller AIC values. Thus this model showed an improvement over the former one. Table 6 presents the parameter estimates and their corresponding standard errors of the fitted models for each age group.

Table 6: Parameter estimates (S.E.) based on the degree-two fractional polynomial model with outliers

Effect	parameter	Age1 after		Effect	Age1 before	
		Estimate(S.E.)	p-value		Estimate	p-value
Intercept	β_0	0.7435 (0.1354)	<.0001	Intercept	2.0011(0.1295)	<.0001
Match	β_1	5.0546(0.1915)	<.0001	Match	2.8642 (0.1832)	<.0001
t^{p_1}	β_2	2.1226 (0.0443)	<.0001	t^{p_1}	1.5357 (0.0447)	<.0001
t^{p_2}	β_3	-0.3161 (0.0124)	<.0001	t^{p_1} *match	-0.7003 (0.0632)	<.0001
t^{p_1}*match	β_4	-1.5515 (0.0626)	<.0001	**Covariances**		
t^{p_2}*match	β_5	0.2612 (0.0175)	<.0001	Var(b_{1i})		
Covariances				cov(b_{1i}, b_{2i})		
Var(b_{1i})	d_{11}	4.3212(0.2610)		Var(b_{2i})	4.5045(0.2828)	
cov(b_{1i}, b_{2i})	$d_{12} = d_{21}$	-0.1882 (0.0133)		Intercept	-1.0448 (0.0846)	
Var(b_{2i})	d_{22}	0.0127 (0.0008)		Match	0.4719 (0.0344)	
		Age2 after			Age2 before	
Intercept	β_0	2.1577 (0.1825)	<.0001	Intercept	1.7619 (0.1699)	<.0001
Match	β_1	5.1118 (0.2581)	<.0001	Match	3.1332 (0.2403)	<.0001
t^{p_1}	β_2	1.3520 (0.0479)	<.0001	t^{p_1}	1.4486 (0.0461)	<.0001
t^{p_1} *match	β_3	-1.0355 (0.0678)	<.0001	t^{p_1} *match	-0.6682 (0.0652)	<.0001
Covariances				**Covariances**		
Var(b_{1i})	d_{11}	7.7337 (0.6033)		Var(b_{1i})	6.7024 (0.4521)	
cov(b_{1i}, b_{2i})	$d_{12} = d_{21}$	-1.6140 (0.1507)		cov(b_{1i}, b_{2i})	-1.4624 (0.1119)	
Var(b_{2i})	d_{22}	0.5112 (0.0428)		Var(b_{2i})	0.4792 (0.0330)	
		Age3 after			Age3 before	
Intercept	β_0	2.6843 (0.2202)	<.0001	Intercept	2.1972 (0.2236)	<.0001
Match	β_1	5.2218 (0.3113)	<.0001	Match	4.0793 (0.3163)	<.0001
t^{p_1}	β_2	1.3660 (0.0668)	<.0001	t^{p_1}	1.4371 (0.0596)	<.0001
t^{p_1} *match	β_3	-1.0093 (0.0944)	<.0001	t^{p_1} *match	-0.8186 (0.0842)	<.0001
Covariances				**Covariances**		
Var(b_{1i})	d_{11}	5.2119 (0.5353)		Var(b_{1i})	5.1927 (0.5173)	
cov(b_{1i}, b_{2i})	$d_{12} = d_{21}$	-1.1401 (0.1422)		cov(b_{1i}, b_{2i})	-0.9331 (0.1183)	
Var(b_{2i})	d_{22}	0.4549 (0.0502)		Var(b_{2i})	0.3574 (0.0366)	
		Age4 after			Age4 before	
Intercept	β_0	3.9205 (0.1501)	<.0001	Intercept	3.2683 (0.1805)	<.0001
Match	β_1	4.2359 (0.2123)	<.0001	Match	2.8173 (0.2552)	<.0001
t^{p_1}	β_2	1.1633 (0.0398)	<.0001	t^{p_1}	1.3063 (0.0435)	<.0001
t^{p_1} *match	β_3	-0.8136 (0.0562)	<.0001	t^{p_1} *match	-0.5655 (0.0615)	<.0001
Covariances				**Covariances**		
Var(b_{1i})	d_{11}	4.0780 (0.3397)		Var(b_{1i})	6.0859 (0.4668)	
cov(b_{1i}, b_{2i})	$d_{12} = d_{21}$	-0.7662 (0.0818)		cov(b_{1i}, b_{2i})	-1.1065 (0.1017)	
Var(b_{2i})	d_{22}	0.2663 (0.0243)		Var(b_{2i})	0.3445 (0.0273)	

The same covariance structure was used as in degree one fractional polynomial model, and a significant group effect was obtained. Since the interaction between group and time was found to be significant with p-value <0.0001, the marginal interpretation of the group would depend on the interaction term. The fitted model can be written as:

$$\ln(cumcost + 1) = {}^{\beta}{}_{0} + b_o + {}^{\beta}{}_{1}match + {}^{\beta}{}_{2}t^{p_1} + {}^{\beta}{}_{3}t^{p_2} + {}^{\beta}{}_{4}match * t^{p_1} +$$

$$ {}^{\beta}{}_{5}match * t^{p_2} + b_1 t^{p_1} + b_2 t^{p_2} + \varepsilon_{ij}$$

When fitting the model with both fractional powers, with one of the fractional powers is found to be insignificant across age categories, before and after diagnosis. Therefore the models are refitted again after exclusion of the insignificant time effect.

After simplification the following model is obtained:

$$\ln(cumcost + 1) = {}^{\beta}{}_{0} + b_o + {}^{\beta}{}_{1}match + {}^{\beta}{}_{2}t^{p_1} + {}^{\beta}{}_{3}match * t^{p_1} +$$

$$b_1 t^{p_1} + \varepsilon_{ij}$$

The addition of one to cumulative cost avoids zero values which would prevent the use of logarithms and negative power transformations. When the best fitting power is (0, 0) logarithm of time and logarithm of time square were used, but a model with the highest fractional power (logarithm of time square) was found to be insignificant and hence excluded from the model. Thus, logarithm of time was used together with the interaction. This model appeared to have smaller AIC value and better fit compared to a model where both logarithm of time and logarithm of time square were included. In the case of fractional power (-1, 0), a model was fitted with both fractional powers that is logarithms of time and inverse of time was incorporated as a main effect and their interactions with group variable. However, the covariate (1/time) was not significant and removed from the model. For the analysis of the first age group, after diagnosis, the selected fractional power was (0.5, 0.5), in this case square root of time, and a combination of logarithm of time and square root of time were used and as a fixed effect, however when both terms are included as random effects the model did not converge. Therefore a model with random intercept and random slope (square root of time) was fitted. Similarly, for (-2, 0), only logarithmic term is significant. In general, as in the previous section random intercept and slope model is fitted across age groups. In all age groups, there is a significant main effect as well as interaction effect. Before interpreting a model, we are going to deal with model diagnosis which will be the focus of section 4.5.

4.5. Model diagnostics

Prior to interpreting the model, for the possible outliers were investigated. In order to identify the potential outlying observations of the data, residual plots were examined. The Studentized residuals shown in the appendix confirmed that residuals outside of the interval -2 and 2 are observed, which is an indication that there are outliers. The key point is that whether they are influential or not? In this section the influence measures in mixed model will be investigated. Removing data points affects fixed effects as well as covariance parameter estimates. Update formulas for Leave-one-out estimates typically fail to account for changes in covariance parameters. Moreover in longitudinal studies one is often interested in multivariate influence rather than the impact of isolated points. Broadly defined, influence is understood as the ability of a single or multiple data points, through their presence or absence in the data, to alter important aspects of the analysis or yield different inferences.

When determining the influence of an observation on the analysis, we must determine whether this is influence on the fixed effects for a given value of the covariance parameters, influence on the covariance parameters, or influence on both. The estimates of the fixed effects depend on the estimates of the covariance parameters.

In linear mixed model the overall influence measure is the likelihood displacement. The likelihood displacement is a global summary measure expressing the joint influence of the influential observations on all parameters. If the global measure suggests that some observations are influential, then the next step is to determine the nature of that influence. The influential points can affect the estimates of fixed effects, the estimates of the precision of the fixed effects, the estimates of the covariance parameters and the estimates of the precision of the covariance parameters.

Cook’s distance was used to capture the change in the fixed effect parameter. Large values of cook’s distance indicate that the change in the parameter estimate is large relative to the variability of the estimate. For the analysis of before diagnosis for individuals in the lowest and highest age group, the overall influence diagnostic (RLD) are displayed in Figure 7. Other plots such as restricted likelihood distance (RLD) for after diagnosis, diagnostics for the fixed effects and covariance parameter influence diagnostics are shown in the appendix for each age group. For the lowest age group for the analysis of before diagnosis, results from restricted likelihood distance showed that clearly the influence of an individual with id number 1455 far exceeds that of other subjects. Moreover, individuals with id 768, 1442 and 1490 had somewhat the highest RLD. The fixed effect estimates are altered by the removal of these four individual’s observations; this is due to the fact that all these subjects have the largest cook’s distance statistics. Covariance ratio was used to capture the effect on the precision of the estimate. All had covariance ratio of less than one which shows that in the absence of these individual's observations the fixed effects parameters can be estimated much more precisely.

The model is refitted after exclusion of the influential data points from the analysis and updated estimates of all parameters are obtained. Though the p-values did not change, there is a change in parameter estimates. The estimates of the main and the interaction effects changed. As can be seen from the cook's distance and covariance ratio statistic for the covariance parameters in the appendix B, these four subjects exert influence on the estimates of the covariance parameters and their precision as well. The results based on the reduced data estimates are presented bellow which will be contrasted with the model of the full data point in order to determine how the absence of the observations changes the analysis. Figure 7 (right panel) shows for the highest age group, before diagnosis. Individuals with id 114, 947, 968, 1001, 1061, 1070 and 1071 had the highest RLD.

When we look at the Cook's distance value of the covariance parameters there existed an impact on the covariance parameter estimates. The covariance parameter estimates can also be assessed deleting influential individuals. These estimates are a bit altered by the removal of the observations. It is important to note that influence analyses are performed under the assumption that the chosen model is correct. Changing the model structure can alter the conclusions. For instance changing the covariance structure will affect the conclusions about which subjects are influential on the analysis.

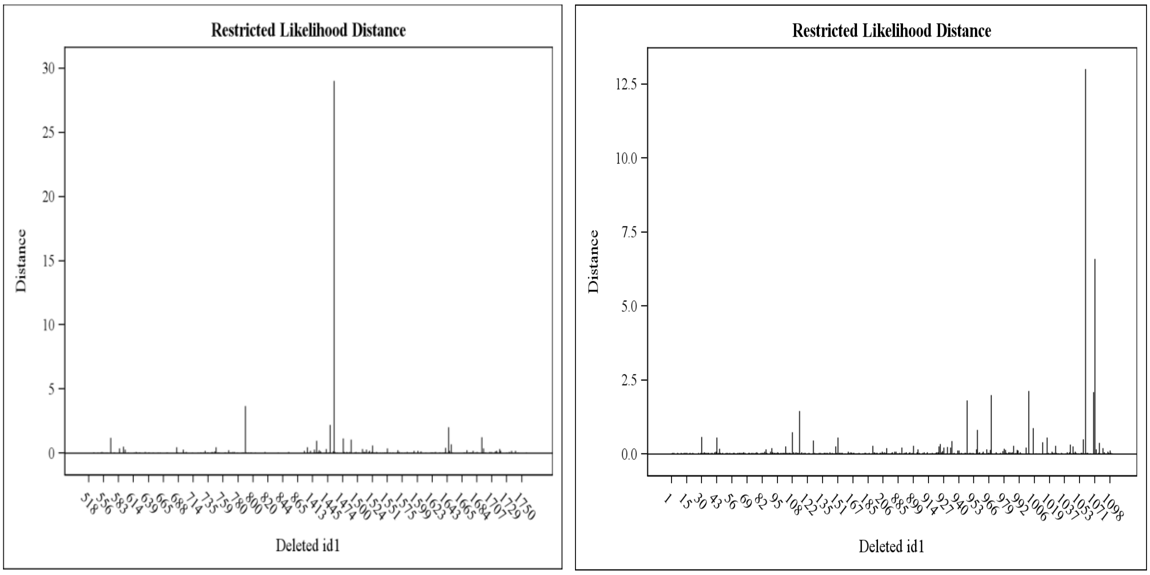


Figure 7*: Restricted likelihood distance (left panel for age group 1 and right panel for age group 4).*

We found substantial differences between the models with and without outliers. The difference in estimates for the intercepts, match group, slope and interactions were large. A similar remark holds for the random intercept variance (d_{11}), covariance (d_{12}) and for the variance of random slope (d_{22}).

For the data of the lowest age category, after diagnosis, 16 possible potential outlying individuals were discarded and the data was fitted again while before diagnosis 4 outlying individuals were not included in the analysis. The second age group consists of 8 outlying individuals before and 8 outliers after diagnosis. The parameter estimates together with their corresponding standard errors after excluding the outlying observations are presented in Table 7. Finally similar to our earlier results (with outliers), none of the analyses revealed an insignificant interaction as well as main effect.

Based on the above Table 7, the time variables (fractional powers) generated by the SAS MACRO and the group variable are significant. Since the time variable is present in the interaction, the interpretation has to be focused on the interactions. For the lowest age group, before diagnosis, we have a significant interaction effect (p<.0001) which can be interpreted as there is a significant difference in cumulative cost between pneumococcal and matched groups over time. The pneumococcal group had the higher cost than the control group which is also observed in the predicted mean evolutions of the two groups as shown in Figure 8. The negative estimate indicates that the cumulative cost was decreased for individuals who belong to the control group compared to those with pneumococcal infection. Meaning that having infected with the pneumococcal disease, the medical expenses increase over time in a higher rate than the undiagnosed persons. Similar interpretations hold for the interaction between the time variable and group variable for the other age groups, before and after diagnosis. In all age groups, before and after diagnosis, we found a significant interaction effect and the cumulative cost is higher in pneumococcal group than the matched group.

For instance for the lowest age group, before diagnosis, the estimated model after excluding outliers is given by:

$$E(logcumcost) = 2.0547 + 2.8534 * \text{match} + 1.5314 * t^{p_1} - 0.7089\text{match} * t^{p_1}$$

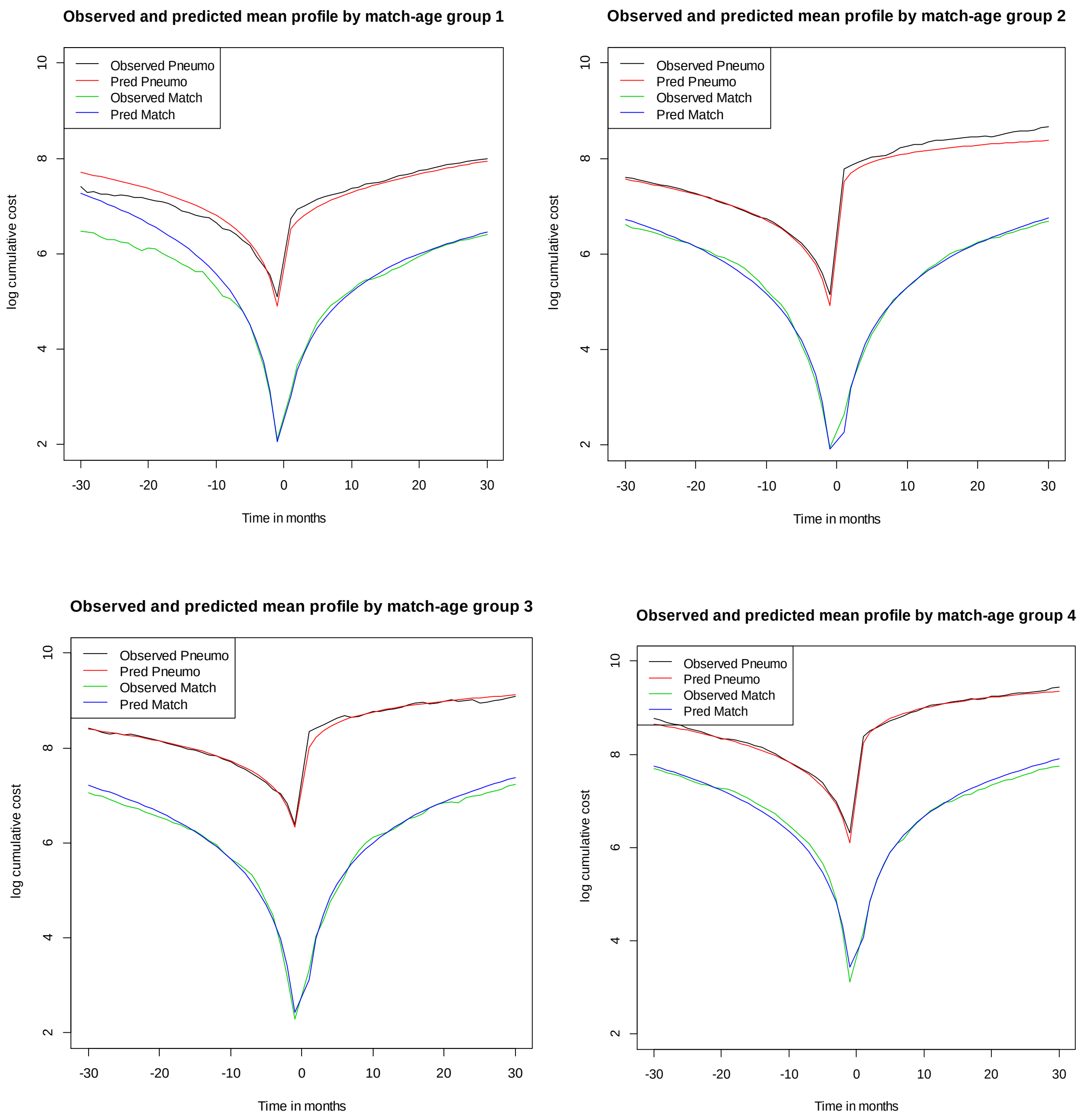
Table 7: Parameter estimates (S.E.) without outliers

Effect	parameter	Age1 after		Effect	Age1 before	
		Estimate(S.E.)	p-value		Estimate	p-value
Intercept	β_0	0.9551 (0.1206)	<.0001	Intercept	2.0547 (0.1249)	<.0001
Match	β_1	5.0762 (0.1699)	<.0001	Match	2.8534 (0.1763)	<.0001
t^{p_1}	β_2	2.0601 (0.0398)	<.0001	t^{p_1}	1.5314 (0.0441)	<.0001
t^{p_2}	β_3	-0.3108 (0.0111)	<.0001	$t^{p_1} *$match	-0.7089 (0.0622)	<.0001
t^{p_1}*match	β_4	-0.6019 (0.0559)	<.0001			
t^{p_2}*match	β_5	0.2681 (0.0156)	<.0001			
Covariances				**Covariances**		
Var(b_{1i})	d_{11}	3.3001(0.2005)		Var(b_{1i})	4.1468(0.2598)	
cov(b_{1i}, b_{2i})	$d_{12} = d_{21}$	-0.1371(0.0100)		cov(b_{1i}, b_{2i})	-0.9589 (0.0789)	
Var(b_{2i})	d_{22}	0.0094(0.0006)		Var(b_{2i})	0.4571 (0.0334)	
		Age2 after			Age2 before	
Intercept	β_0	2.2809(0.1504)	<.0001	Intercept	1.9161(0.1669)	<.0001
Match	β_1	5.2398 (0.2120)	<.0001	Match	3.0138(0.2345)	<.0001
t^{p_1}	β_2	1.3154 (0.0402)	<.0001	t^{p_1}	1.4145(0.0457)	<.0001
$t^{p_1} *$match	β_3	-1.0624(0.0566)	<.0001	$t^{p_1} *$match	-0.6405(0.0642)	<.0001
Covariances				**Covariances**		
Var(b_{1i})	d_{11}	5.1267(0.3591)		Var(b_{1i})	6.2841(0.4272)	
cov(b_{1i}, b_{2i})	$d_{12} = d_{21}$	-0.9646(0.0851)		cov(b_{1i}, b_{2i})	-1.3750(0.1064)	
Var(b_{2i})	d_{22}	0.3478(0.0259)		Var(b_{2i})	0.4592(0.0318)	
		Age3 after			Age3 before	
Intercept	β_0	3.1280(0.2042)	<.0001	Intercept	2.4348(0.2148)	<.0001
Match	β_1	4.8743 (0.2799)	<.0001	Match	3.9018 (0.3008)	<.0001
t^{p_1}	β_2	1.2492 (0.0575)	<.0001	t^{p_1}	1.4054 (0.0588)	<.0001
$t^{p_1} *$match	β_3	-0.9213(0.0787)	<.0001	$t^{p_1} *$match	-0.8004 (0.0824)	<.0001
Covariances				**Covariances**		
Var(b_{1i})	d_{11}	3.9425(0.4118)		Var(b_{1i})	4.5747(0.4614)	
cov(b_{1i}, b_{2i})	$d_{12} = d_{21}$	-0.8220(0.1008)		cov(b_{1i}, b_{2i})	-0.8377(0.1081)	
Var(b_{2i})	d_{22}	0.2967(0.0322)		Var(b_{2i})	0.3347(0.0345)	
		Age4 after			Age4 before	
Intercept	β_0	4.0737(0.1313)	<.0001	Intercept	3.4342(0.1757)	<.0001
Match	β_1	4.1595(0.1848)	<.0001	Match	2.6777 (0.2468)	<.0001
t^{p_1}	β_2	1.1249 (0.0351)	<.0001	t^{p_1}	1.2681(0.04306)	<.0001
$t^{p_1} *$match	β_3	-0.7951 (0.0493)	<.0001	$t^{p_1} *$match	-0.5215 (0.06049)	<.0001
Covariances				**Covariances**		
Var(b_{1i})	d_{11}	3.0550 (0.2333)		Var(b_{1i})	5.5925(0.4374)	
cov(b_{1i}, b_{2i})	$d_{12} = d_{21}$	-0.5122 (0.0527)		cov(b_{1i}, b_{2i})	-1.0384(0.0974)	
Var(b_{2i})	d_{22}	0.2027 (0.0168)		Var(b_{2i})	0.3286(0.0265)	

As shown at the bottom panel of Table 7, the covariance between the random intercept and slope in all age groups for before and after diagnosis is negative. Therefore the negative correlation between random intercept and slope point out that those starting with low costs, their slopes are vertical i.e. their cumulative costs increase at faster rates, while those starting with higher costs, their cumulative costs increase at lower rates.

Predictions based on degree two fractional polynomials

The Plot of the observed versus predicted values by match group based on the estimated model after exclusion of the outliers are shown in Figure 8. Since there is no formal test to test if our final model fits well we opted for an alternative which was to compare the observed mean profile and the predicted mean profile.


Observed and predicted mean profile by match-age group 1

Observed and predicted mean profile by match-age group 2

Observed and predicted mean profile by match-age group 3

Observed and predicted mean profile by match-age group 4

Figure 8: *observed and predicted mean profile in each match group.*

According to these evolutions, the mean profile of the observed and predicted values of the models was relatively well on a few time intervals near the moment of diagnosis. However, at time points long before and after diagnosis, the model was not fit the data well that might be due to the presence of excess missing values. There, is a much difference between the predicted and observed values of the cumulative costs at the extremes of the time points especially for the lowest age group, before diagnosis. Furthermore it appears that the cumulative costs of the pneumococcal group are higher than that of the control group in each of the age groups, we can say that individuals who are diagnosed with pneumococcal infection incurred a higher medical cost than the matched persons throughout the time intervals.

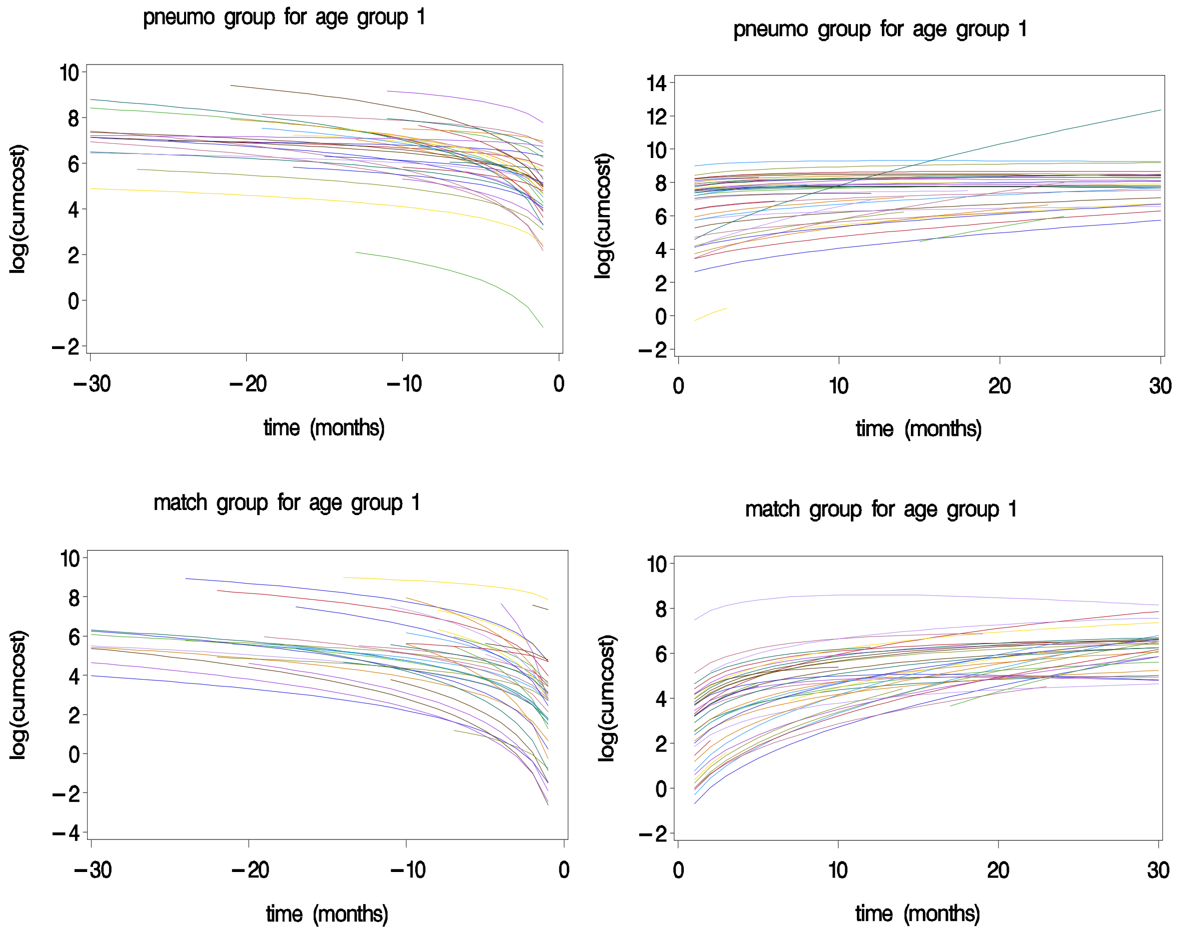


Figure 9*: prediction of individual profiles for pneumococcal and matched persons across age group.*

In order to assess how well the fitted model describes observed longitudinal profiles, the fitted individual profiles were plotted. Figure 9 shows the estimates of random effects corresponding to the subject-specific curves of the fitted fractional polynomial curves. The thinning of the data toward the latter study times suggests a lot of dropouts. This occurs especially for the first two age groups before diagnosis. For the sake of illustration here we presented the predicted profiles of the first age group, for the other age groups the profiles are shown in the Appendix B (Figure 16). Figure 9 shows predicted profiles of 40 randomly selected individuals from each group. When compared to the observed profiles, the predicted profiles seem to reflect the observed pattern. The observed profiles of log cumulative costs are presented in Appendix B (Figure 14).

Derivative plots

Once cumulative costs are modeled, to compare the evolutions of the original costs over time of patients with pneumococcal infection with that of the control group, derivation of the cumulative cost is needed as discussed in section 3.4. A graphical representation of the average evolutions in each of the age groups is presented here.

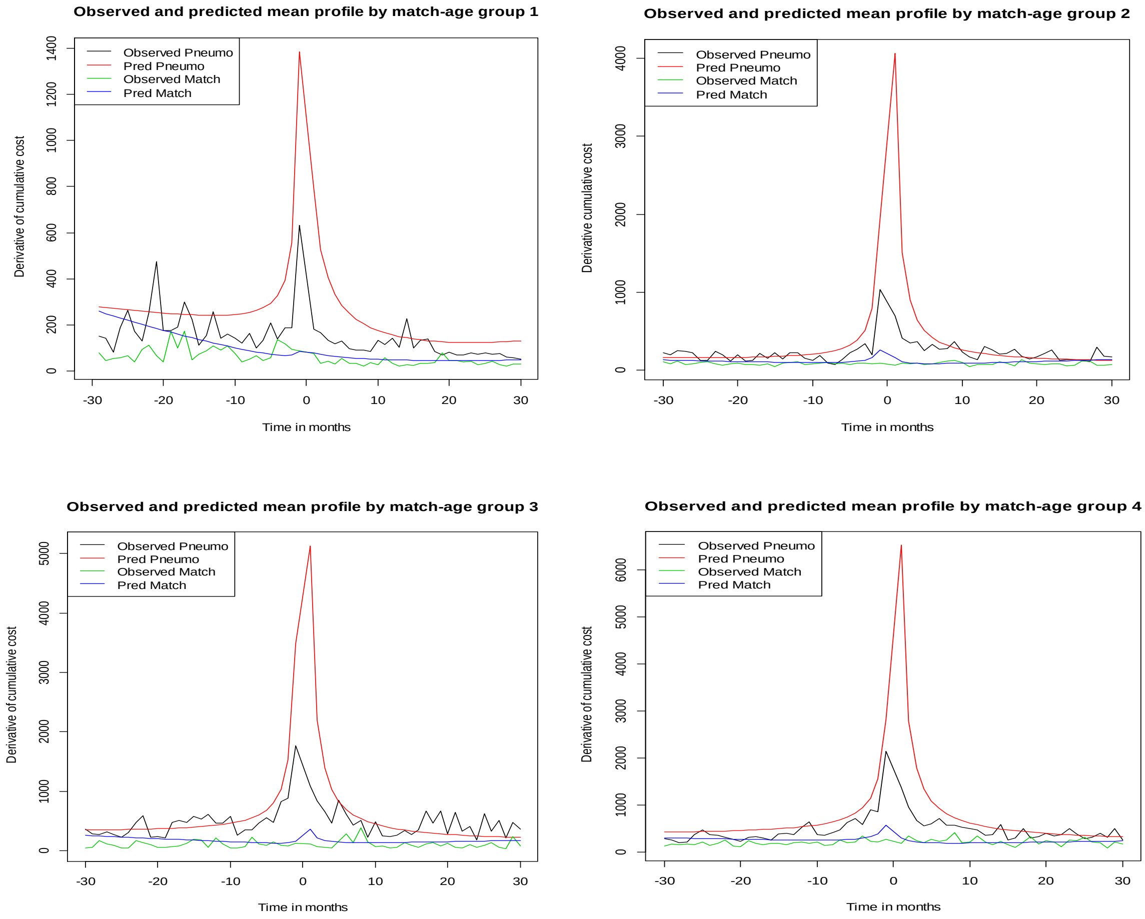


Figure 10*: derivatives of the observed and predicted mean profile in each match group.*

Before as well as after diagnosis, patients with pneumococcal disease tend to have higher costs in each of the age groups. The fitted curves do not appear to describe the observed mean evolution well.

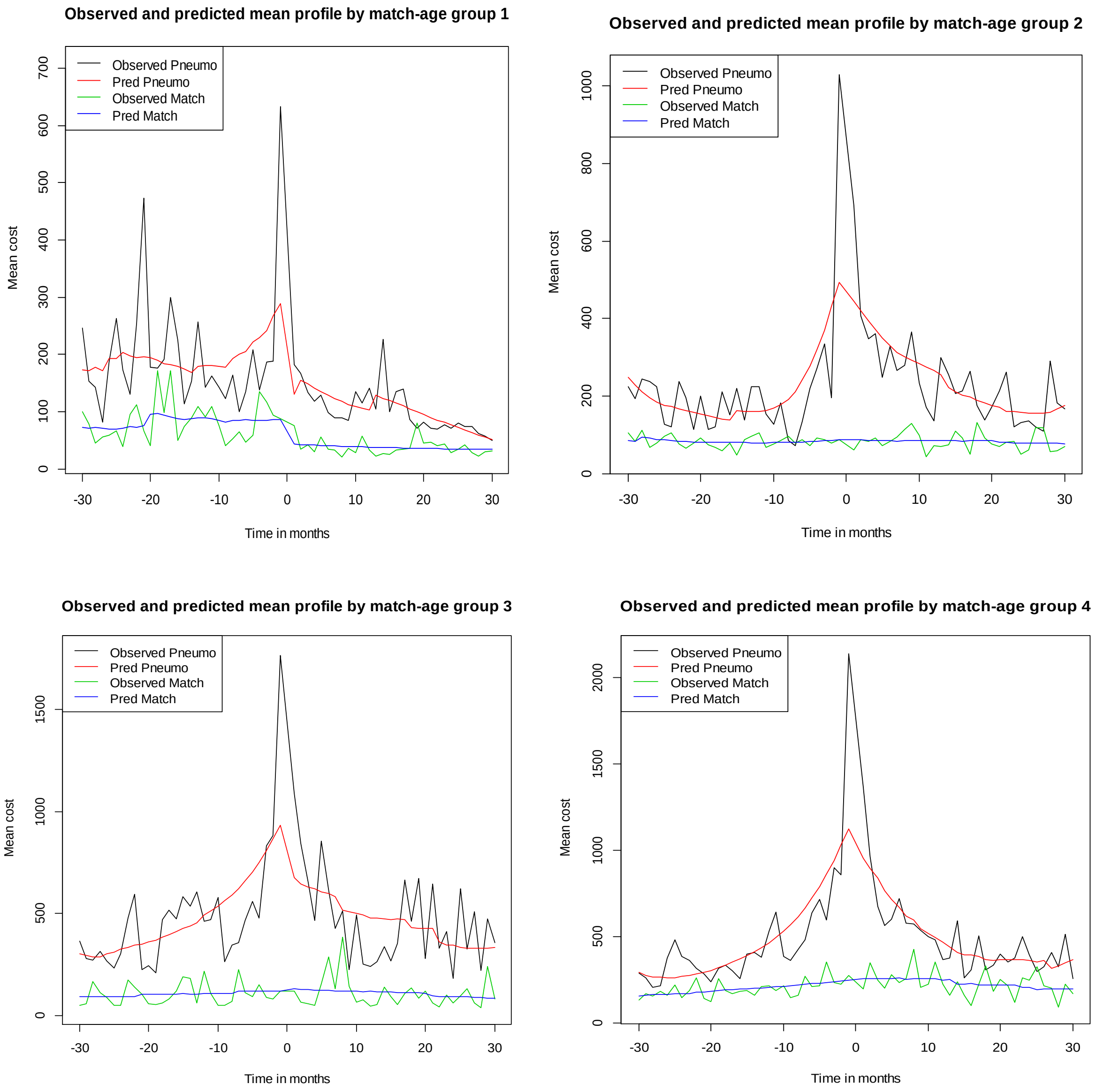
4.6. Semiparametric Mixed Models

In this section we presented the result of the mixed model formulation of penalized splines. Smoothing methods that use basis functions with penalization can be formulated as fits in a mixed model framework. One of the major benefits is that software for mixed model analysis can be used for smoothing. It has become a widely used tool for data analysis and inference and its incorporation into complex models and use in applications has increased.

All the models were fitted with the time scale in months, and in each of the data sets ten equally spaced knots were used to be able to fit the model with radial basis function. Comparisons between flexible parametric (based on fractional polynomial) and data driven flexible models (semiparametric mixed model) were made. The semiparametric mixed model was fitted on the original cost variable while mixed model with fractional polynomial mean structure was fitted using the cumulative costs. Therefore, the derivative plots shown in Figure 10 and the plots based on spline modeling displayed in Figure 11 were compared. For derivative estimation via penalized splines, it is recommended that higher degree polynomial basis functions be used to ensure that the resulting derivative estimates are smooth. Note that the degree of the spline used to estimate the cost function should exceed the order of derivative by at least one. Polynomial bases are not standard because of their numerical instability. We tried to implement the model by means of cubic radial basis function using IML procedure in SAS; however the model fails to converge. Therefore, it was fitted based on linear radial basis function.

The splines model parameterization described in Section 3.5 is now used and the results are presented here. Figure 11 shows the observed and predicted values for both the pneumococcal and the control group. When comparing the derivative plots (Figure 10) and this modeling approach, the fit of the spline model is trying to capture the irregularities in the profiles. Thus, the spline model fits the data better than the fractional polynomial models.

In red is the predicted mean profile for the pneumococcal group while in the blue we have plotted the predicted mean profile for the control group. The estimated mean cost for the pneumococcal group is higher than the control group in every time points across age categories. Moreover, the difference in estimated costs between pneumococcal and the control group is higher in spline modeling than the earlier approach. The estimated parameters of the semiparametric mixed model are given in the Appendix A (Table 9).


Observed and predicted mean profile by match-age group 1

Observed and predicted mean profile by match-age group 2

Observed and predicted mean profile by match-age group 3

Observed and predicted mean profile by match-age group 4

Figure 11*: Observed and predicted mean cost in each match group based on spline model.*

If someone is interested to see the estimated costs in each of the age groups, Figure 17 illustrates the predicted mean evolutions of cost over time under each of the age categories for pneumococcal and matched groups. The different age groups are represented by different colors. As it can be seen from the plot, after diagnosis, younger patients tend to have the lowest cost consistently in every measurement time followed by the second age group in both pneumococcal and the control group. The oldest age group tends to have the highest cost for those who belong to the matched group whereas in pneumococcal patients it fluctuates, it the first few measurement times the oldest age category incurred the highest cost.

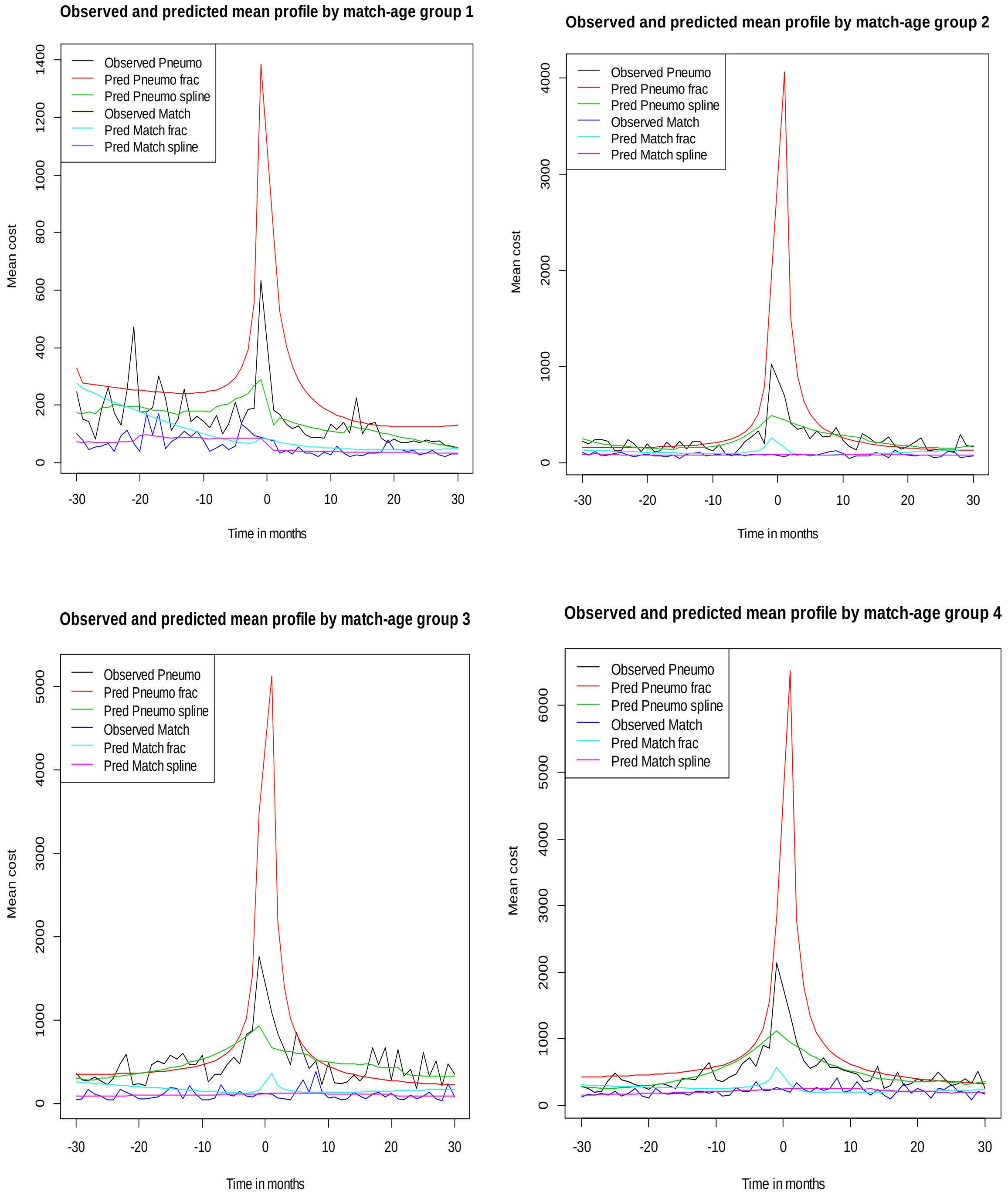


Figure 12: *Observed mean cost for both pneumococcal and control group together with the fitted fractional polynomial mixed model and the fitted semiparametric mixed model.*

In order to assess the goodness of fit between parametric and the semiparametric approaches, the predicted versus observed mean evolutions for both the pneumococcal and the matched group is plotted on the same graph. In Figure 12, the model fits obtained with the fractional polynomial mixed model and the semiparametric mixed model are compared. In the plots, it can be seen that the semiparametric mixed model is trying to capture the irregularities in the profile, while the fractional polynomial mixed model only focuses on the main trend.

5. Discussion and conclusion

In this study we compared health-care expenditures of 876 individuals who have had a positive isolate taken for Streptococcus pneumoniae at a known time with matched persons who have not. The objective of the study was to compare the health-care costs between the pneumococcal and the control groups. The data set used in this study composed of a two groups of individuals. One is pneumococcal and the other is control group. Measurements were taken on individuals on a monthly basis and because of the longitudinal nature of the data, observations within individuals are correlated. This correlation needs to be taken in to account in model building.

We have considered an application of flexible modeling techniques in project to achieve the objective of the study, flexible mixed effect model and penalized splines smoothing of longitudinal data. As we have seen from the plots in the exploratory data analysis, the individual profiles of the two groups are non linear. To fit a model that takes into account the trend of the individual profiles, a fractional polynomial was implemented to determine the accurate power of the time variable. To fit this model, we came out with a macro as presented in the appendix, which allowed generating the appropriate power for a fractional polynomial of order 2. To generate first-order fractional polynomial powers it can easily be modified. The macro gives the possibility of using mixed effects model. We fit a mixed effect model for each age group, before and after diagnosis. Once the polynomial powers were obtained in this way, the time variables were then used in the model and also their interactions with matched group using PROC MIXED in SAS. Both parametric and semi-parametric modeling approaches were applied to the data and compared their fitting abilities. The semiparametric modeling have placed a strong demand on developing semi-parametric regression methods for longitudinal data, where flexible functional forms can be estimated from the data to capture possibly complicated relationships between longitudinal outcomes and covariates. The semiparametric mixed model fitted in this approach is convenient; it can be applied by means of widely used available commercial software for mixed model. The semiparametric mixed model we have considered is general and it can be extended in several ways, for example a model assuming the smoothing level to vary in the group variable and also several possible scenarios depicting the evolution. Differences or similarities can be assumed in the linear part of the model, in the non-linear or in both of the model. Fractional polynomial was used to find a good fitting mean structure. A mixed effects model was fitted using fractional polynomial mean structure. An age specific analysis was performed since the risk of severe pneumococcal infection is different for different age groups.

The best model was selected from the two parametric models namely mixed effects model based on degree-one and degree-two fractional polynomial mean structures using Akaike Information Criteria (AIC). Likelihood ratio test was used to compare any two nested models such as a model with only random intercept as a null model and a model that comprises both random intercept and slope, however any two non-nested models was compared AIC. Thus, a model based on degree-two fractional polynomial mean structure and with both random intercept and slope was selected as the best model.

In line with the finding that the average evolution of cost depends on the time of interest, a spline approach was considered. In a semiparametric mixed model using penalized splines, a random intercept model was fitted. A random intercept model only assumes a shift in subject-specific profiles, a rather restrictive assumption. More complex models including subject specific random intercepts and slopes can be considered, however for this data set we are confronted with convergence problem when both random effects are included in the model.

The data was also fitted again after removing the potential outlying observations in order to study their effect on the model. There was a substantial difference between the parameter estimates of the mixed effects model based on the second-order fractional polynomial before and after excluding the outlying observations, however in both models the variable group and an interaction between groups with time effects were found to be significantly associated with the cumulative costs. Hence the two group effects are different in different time points. The plot of the predicted versus observed of the costs approved that the model did not fit the data well. Therefore a more flexible modeling, the semiparametric mixed model was also fitted. Use of semiparametric modeling technique provided better fit than the fractional polynomial mixed model.

Based on the analysis described in this report one can conclude that the average evolution of the costs depends on the time of measurement. According to the models we can also conclude that there exists a significant difference between pneumococcal and control groups over time. In both modeling approaches for all age groups, the health-care costs incurred by diagnosed pneumococcal patients are larger than those undiagnosed persons, before and after diagnosis. This is expected in the sense that individuals infected with the disease expend more health-care costs. If the pneumococcal episode were removed from an individual's health by vaccination, it would be helpful to consume more health-care resources. In both models the fitted curves do not appear to describe the mean evolution well. Finally, the obtained results can be used to inform policy on the budget impact of pneumococcal vaccination programs and for cost estimation. The data had missing observations. The missing mechanism was treated as missing at random (MAR). Under the likelihood approach this missing mechanism is ignorable (Molenbeghs and verbeke, 2005).

References

Aerts, M. (2006) Applied Data Modelling. Msc Biostatistics course notes, Hasselt University.

Black, R., Morris, S., Bryce, J. (2003) where and why are 10 million children dying every year?

Beutels, P., Blommaert, A., Hanquet, G., Bilcke, J., Thiery, N., Sabbe, M., and Verhaegen, J. (2011) *Cost-effectiveness of 10- and 13-valent pneumococcal conjugate vaccines in childhood.*

Coles, C., Kanungo, R., Rahmathullah, L., Thulasiraj, R., Katz, J., Santosham, M. (2001) Pneumococcal nasopharyngeal colonization in young South Indian infants.

Creemers, A., Aerts, M., Hens, N., Shkedy, Z., Frank, D., Smet and Beutels, P. (2011). Revealing age-specific past and future unrelated costs of pneumococcal infections by flexible generalized estimating equations. *Journal of applied statistics* **38**, 1533-1547

Durban, M., Harezlak, J., Wand, M.P., and Carroll R.J. (2004) Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine*, **00**, 1-24.

Lesaffre, E., Asefa, M., and Verbeke, G. (1999) Assessing the goodness-of-fit of the Lair and Ware model-an example:the jimma infant survival differential longitudinal study. *Statistics in Medicine* **18**, 835-854.

Eugene, D., and Therese, A., Stukel (2005) Influence analysis for linear mixed-effects models. *Statistics in Medicine* **24**, 893-909.

Greenwood, B., Weber. M., Mulholland, K. (2007) Childhood pneumonia-preventing the world’s biggest killer of children. *Bull World Health Organ*, **85**, 502-513.

Guo, W. (2002) Functional mixed effects models. *Biometrics*, **58**, 121-128.

Hastie, T.J., Tibshirani, R.J (1990) *Generalized Additive Models.* Chapman & Hall.

Huiqu, p., and Harvey, G. (1998). Multi-level repeated measures growth modeling using extended spline functions. *Statistics in Medicine* **17**, 2755-2770.

Maringwa, J., Faes, C., Geys, H., Molenberghs, G., Cadarso-Suarez, C., Pardo-Vazquez, J., Leboran, V., and Acuna, C. (2008) Application of Penalized Smoothing Splines in Analyzing Neuronal Data.

Laird, N.M., and Ware, J.H. (1982). Random effects models for longitudinal data. *Biometrics*, **38**, 963-974.

Mbelle, N., Huebner, R., Wasas, A., Kimura, A., Chang, I., and Klugman, K. (1999) Immunogenicity and impact on nasopharyngeal carriage of a nonavalent pneumococcal conjugate vaccine.

Molenberghs, G. & Verbeke, G. (2005) *Models for Discrete Longitudinal Data*. New York: Springer.

Ngo, L., Wand, M.P. (2004) Smoothing with mixed model software. *Journal of Statistical Software*, **9**, 1-56.

Ronald, Larry, M., and Pearson (1992) Case-deletion diagnostics for mixed models. *Technometrics*: **34**, 38-45.

Royston, P., Altman, D.G. (1994) Regression using fractional polynomials of continuous covariates: parsimonious parametric modeling. *Applied Statistics*, **43**, 429-467.

Ruppert, D.,Wand, M.P., Carroll, R.J. (2003) *Semiparametric Regression*. Cambridge University Press.

Schabenberger, O. (2004) Mixed Model Influence Diagnostics. Cary, NC: SAS Institute Inc.

Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data*. New York: springer.

Wood, Simon N. (2006) *Generalized Additive Models An Introduction with R.* Boca Raton: Chapman and Hall/CRC.

Appendix

Appendix A: Tables

Table 8: Descriptive statistics of cost for the lowest age group at each time point

		Pneumococcal				Match				
Time points	N	Mean	median	P_5	P_{95}	n	mean	median	P_5	P_{95}
-30	51	245.422	16.86	0	2304.33	51	99.073	15.47	0	585.33
-25	61	262.088	15.12	0	341.83	61	64.338	23.95	0	150.8
-20	85	182.608	17.85	0	660.96	85	39.447	0	0	159.38
-15	119	113.076	20.45	0	670.73	119	72.353	15.74	0	243.49
-10	170	142.198	22.525	0	946.94	170	74.799	0	0	595.29
-5	248	207.914	22.74	0	735.93	248	57.897	17	0	136.06
0	278	1874.93	1245.08	15.3	6277.77	278	61.233	15.51	0	218.52
5	274	129.113	21.44	0	421.56	274	54.502	0	0	209.09
10	266	134.681	17.68	0	333.95	266	27.765	0	0	109.65
15	246	97.496	0	0	302.96	246	23.887	0	0	110.53
20	235	93.557	11.6	0	374.91	235	43.707	0	0	151.41
25	207	77.696	0	0	291.06	207	32.711	0	0	95.28
30	196	49.618	0	0	242.49	196	29.969	0	0	85.03

Table 9: Parameter estimates of semiparametric mixed model

Effect	Age1 after			Effect	Age1 before		
	Estimate	Std.error	p-value		Estimate	Std.error	p-value
Intercept	-4.9137	0.4598	<.0001	Intercept	-3.4552	0.5257	<.0001
Match	-5.2662	0.6503	<.0001	Match	-5.3352	0.7434	<.0001
Time	-0.2321	0.0214	<.0001	Time	0.1208	0.0322	0.0002
Time*match	0.1364	0.0302	<.0001	Time*match	0.1066	0.0331	<.0001
Covariances							
Var of intercept	25.7302	2.0489			32.1066	2.9170	
Var[RSmooth(time)]	0.0008	0.0002			0.0008	0.0002	
Residual	144.09	1.7757			134.75	2.2522	
	Age2 after				Age2 before		
Intercept	-6.6912	0.6148	<.0001	Intercept	-7.9781	0.5877	<.0001
Match	-3.7800	0.8695	<.0001	Match	-3.1013	0.8312	0.0002
Time	-0.1320	0.0275	<.0001	Time	0.1226	0.0245	<.0001
Time*match	1.2093	0.1329	0.0014	Time*match	-0.1018	0.0346	0.0033
Covariances	Estimate						
Var of intercept	42.2063	3.4999			40.1271	3.2484	
Var[RSmooth(time)]	0.0016	0.0002			0.0015	0.0002	
Residual	0.1244	0.0389			135.37	1.7730	
	Age3 after				Age3 before		
Intercept	0.9786	0.2474	0.0001		-4.2017	0.9081	<.0001
Match	2.3010	0.3499	<.0001		-5.6524	1.2843	<.0001
Time	0.0050	0.0099	0.6143		0.1331	0.0338	<.0001
Time*match	-0.02990	0.0141	0.0336		-0.1350	0.0477	0.0047
Covariances							
Var of intercept	3.8597	0.4374			51.5516	51.5516	
Var[RSmooth(time)]	0.0001	0.00002			0.0012	0.00023	
Residual	7.3464	0.1538			127.72	2.4622	
	Age4 after				Age4 before		
Intercept	-3.7724	0.5813	<.0001		-4.5929	0.6134	<.0001
Match	5.2970	0.8221	<.0001		4.2048	0.8675	<.0001
Time	0.0440	0.0223	0.0483		0.0203	0.0194	0.2948
Time*match	-0.1357	0.0315	<.0001		0.0932	0.0274	<.0001
Covariances							
Var of intercept	40.6515	3.4612			49.8407	3.9725	
Var[RSmooth(time)]	0.0005	0.0001			0.0006	0.0001	
Residual	79.0444	1.3195			84.3704	1.2031	

Appendix B: Figures

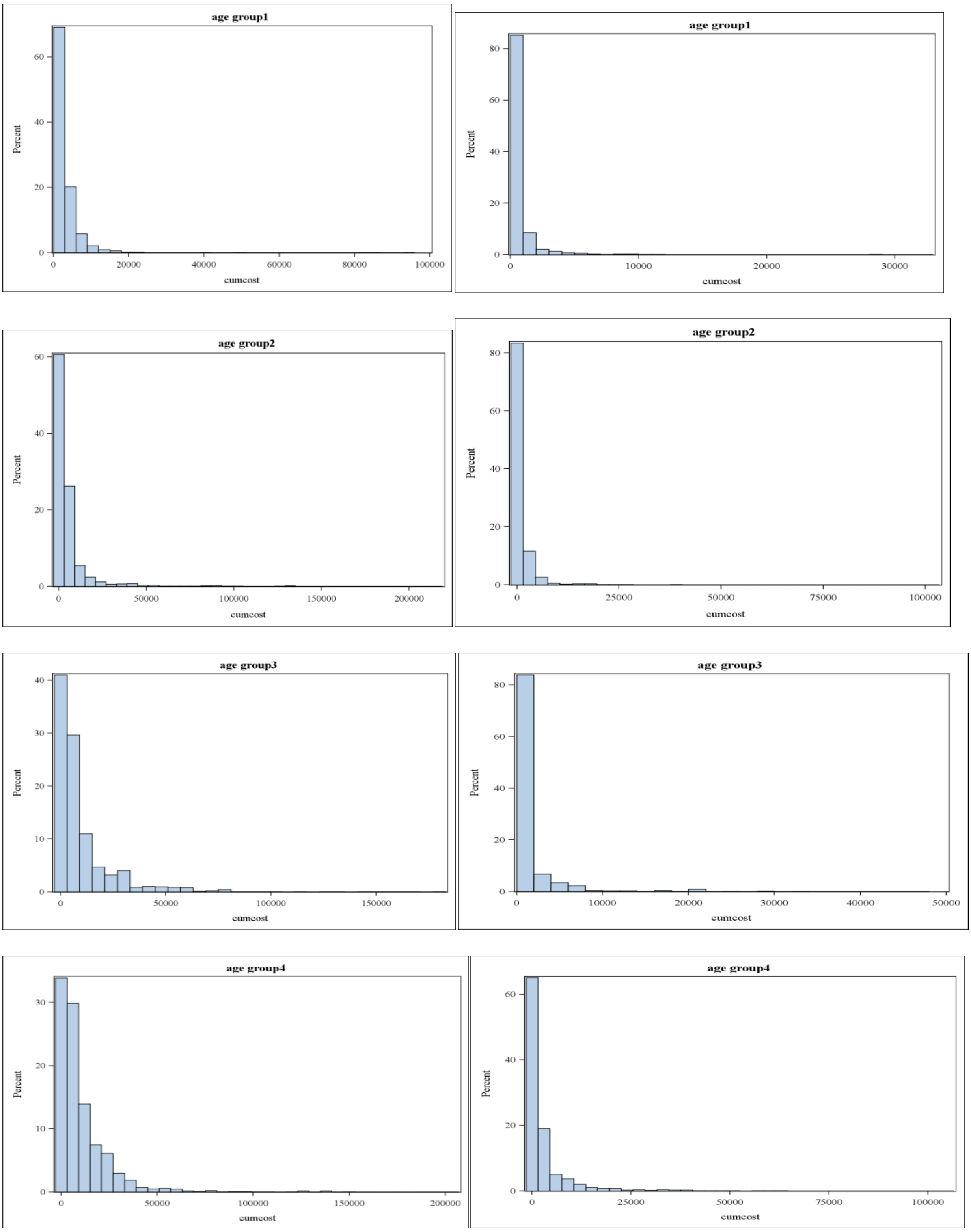


Figure 13: distribution of cumulative cost for each of the pneumococcal and control group across the age categories

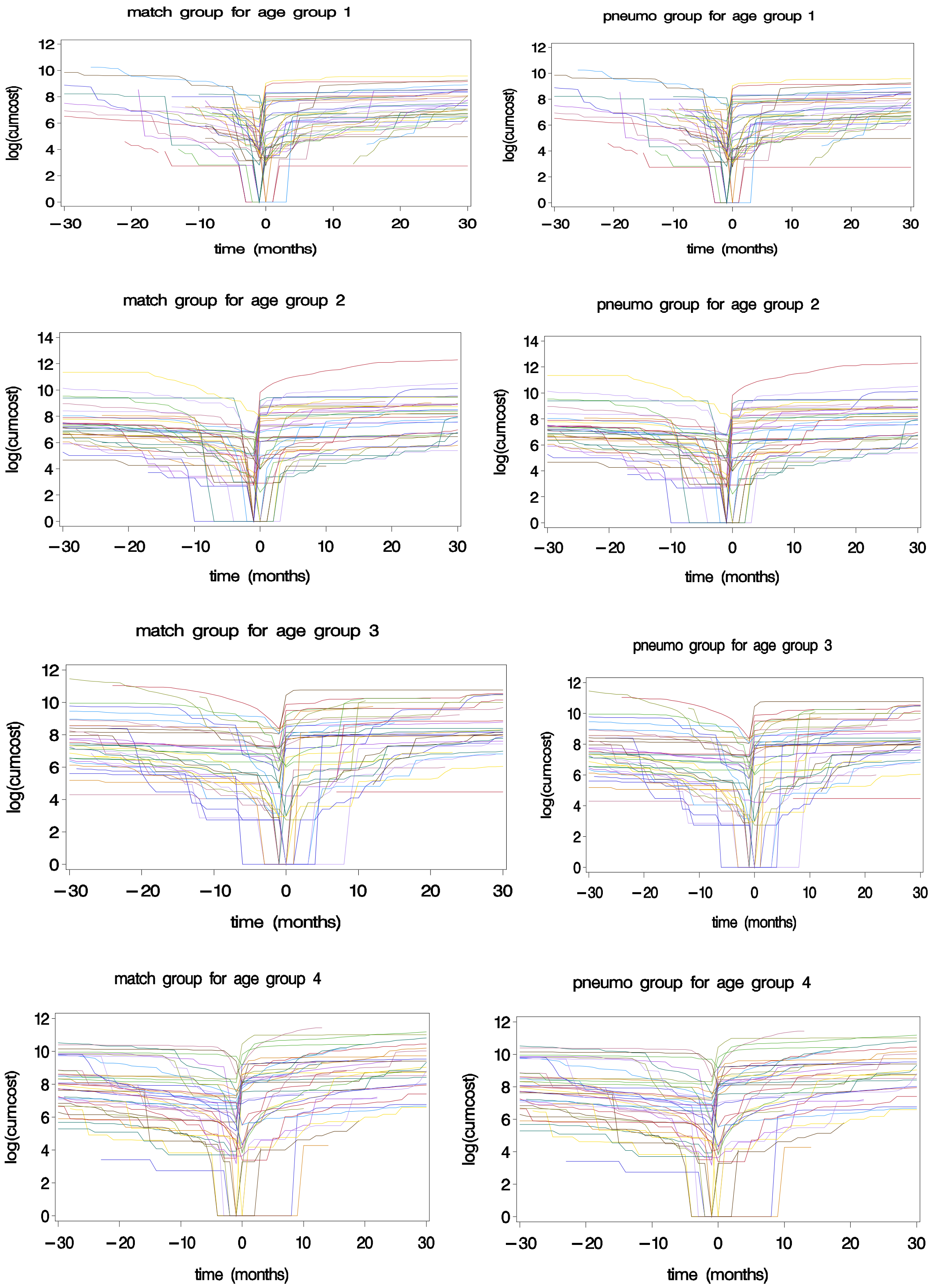


Figure 14: Individual profiles for pneumococcal and matched patients across age category based on the log cumulative costs

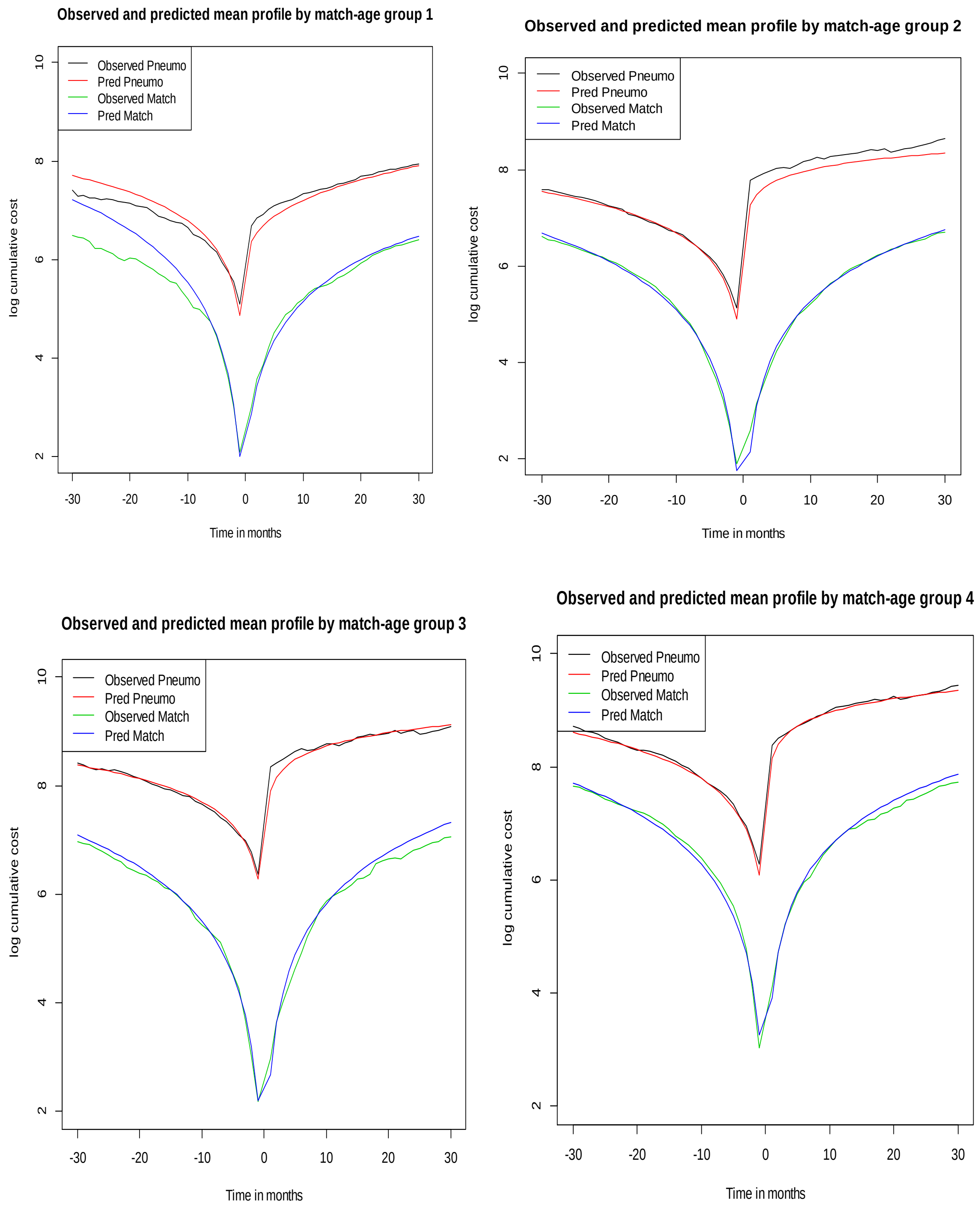
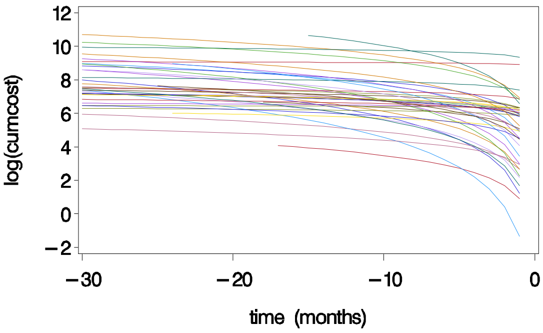
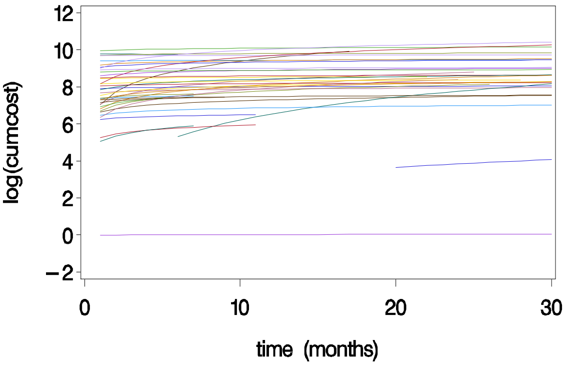


Figure 15: Observed and fitted evolutions by match with outliers

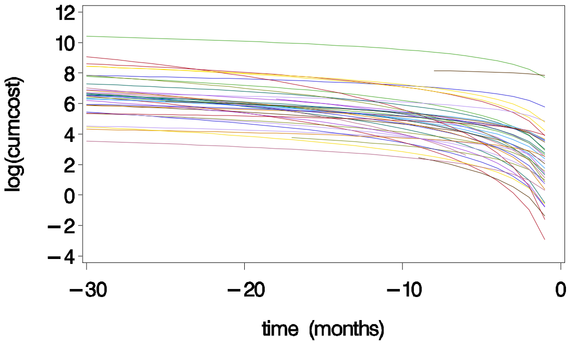
pneumo group for age group 2



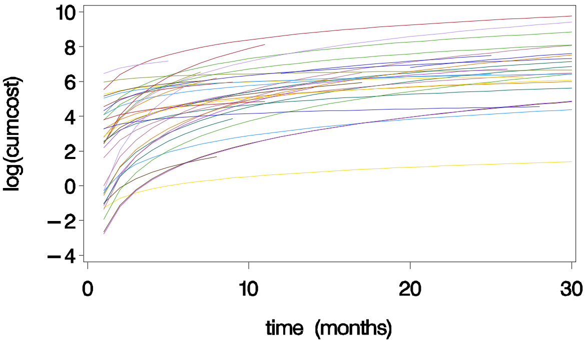
pneumo group for age group 2



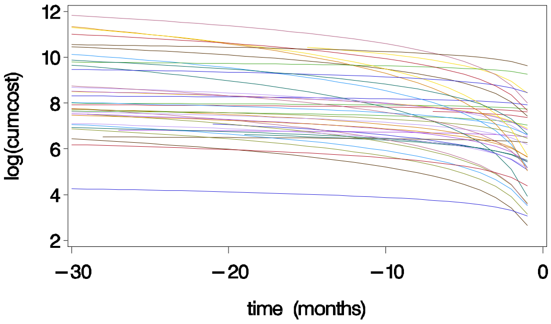
match group for age group 2



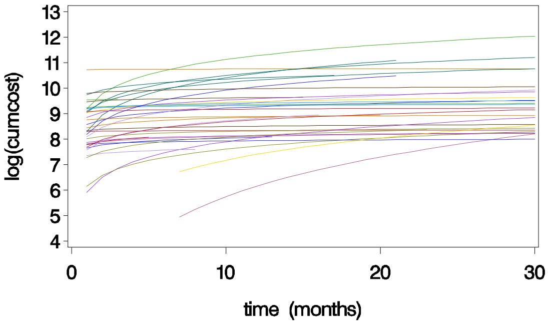
match group for age group 2



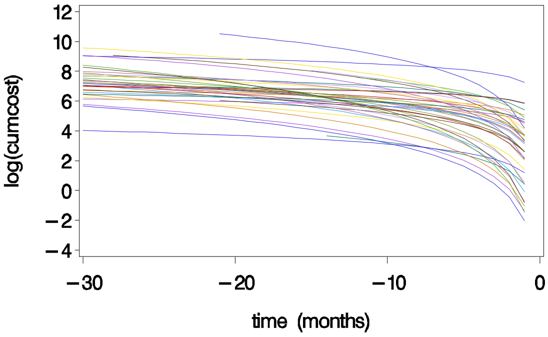
pneumo group for age group 3



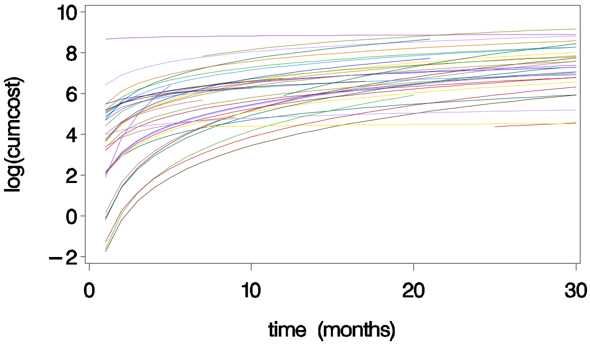
pneumo group for age group 3



match group for age group 3



match group for age group 3



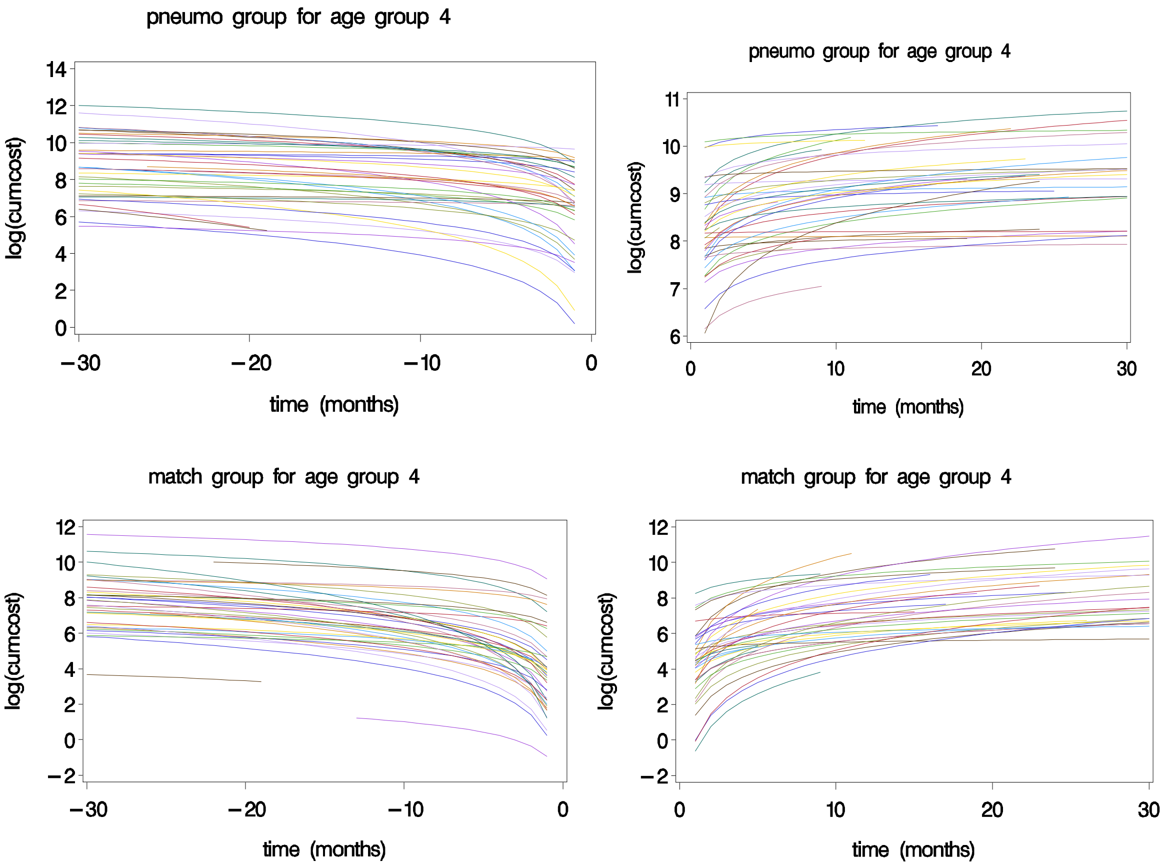


Figure 16: Prediction of individual profiles for pneumococcal and matched group across age category.

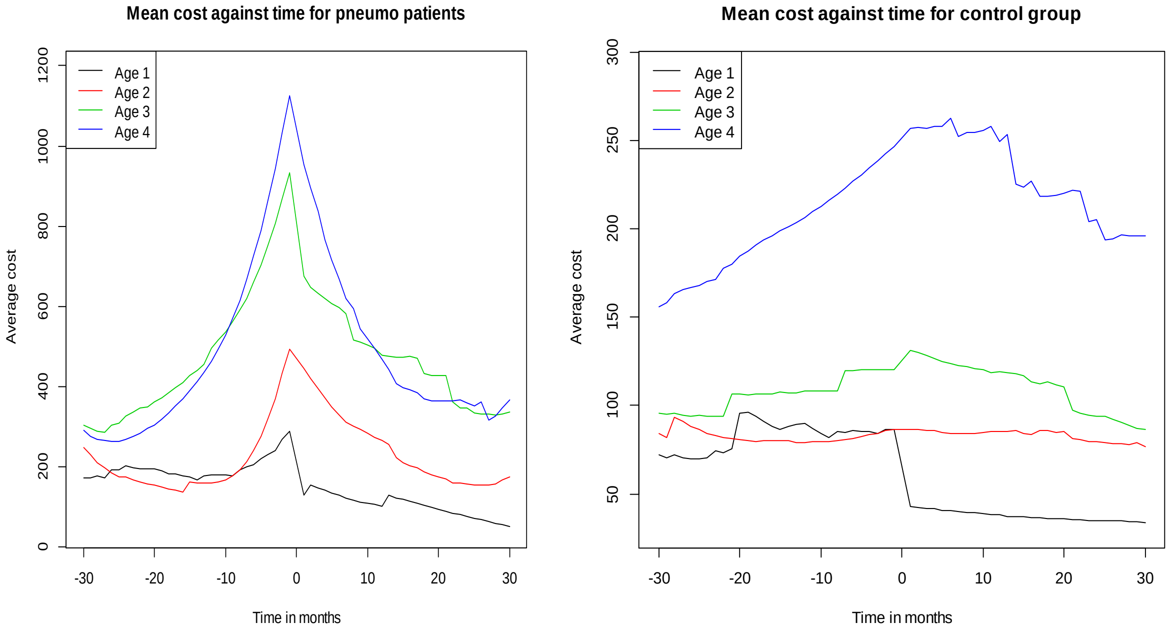
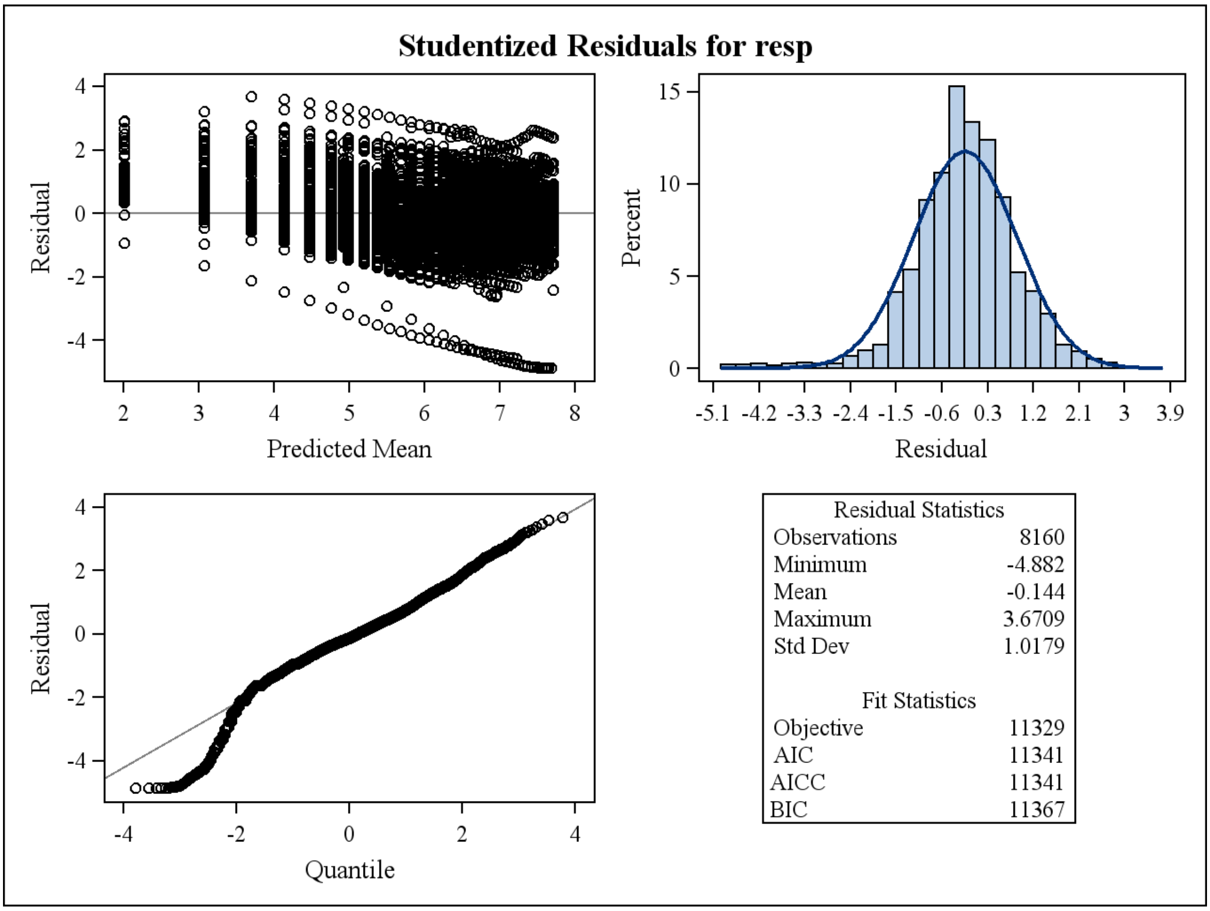


Figure 17*: Predicted cost against time across age categories for pneumococcal and matched groups based on the spline model.*



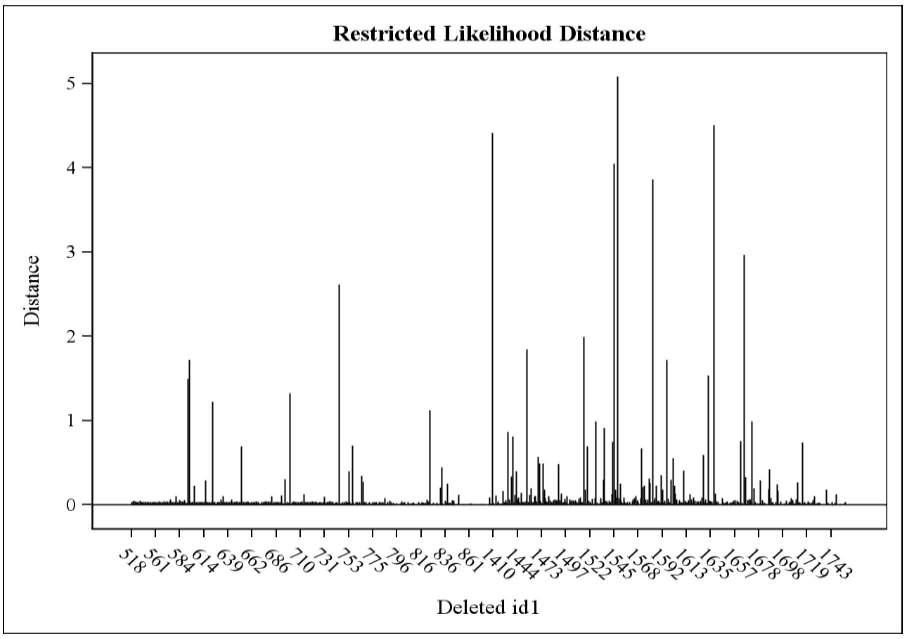
Residual Statistics	
Observations	8160
Minimum	-4.882
Mean	-0.144
Maximum	3.6709
Std Dev	1.0179

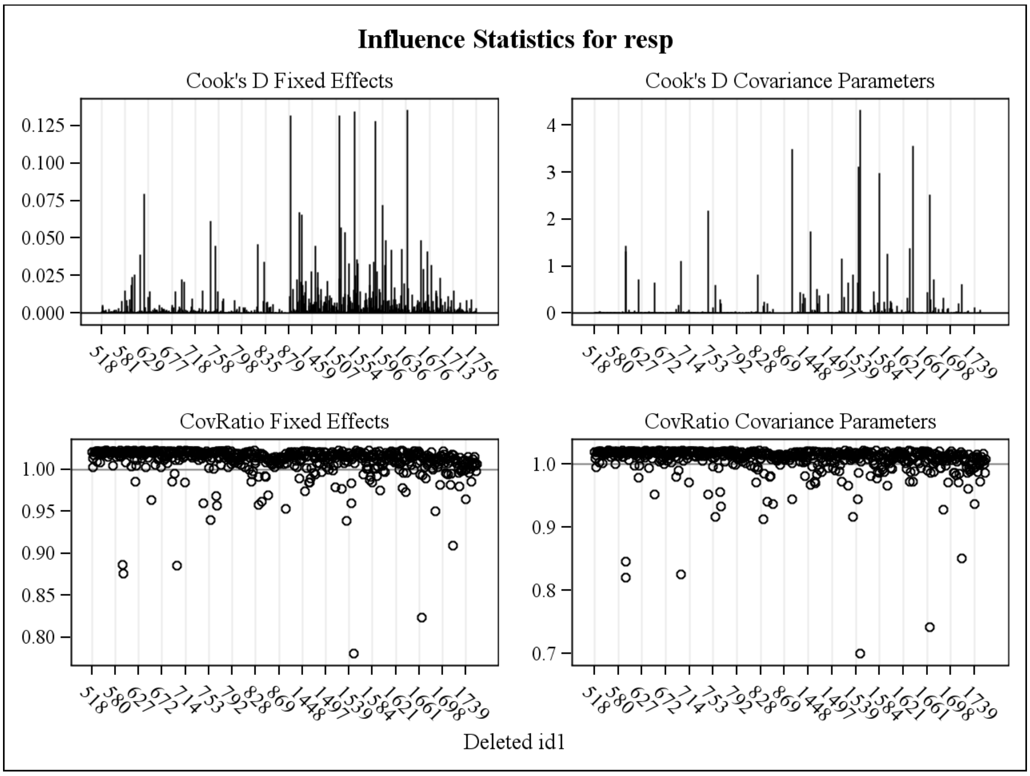
Fit Statistics	
Objective	11329
AIC	11341
AICC	11341
BIC	11367

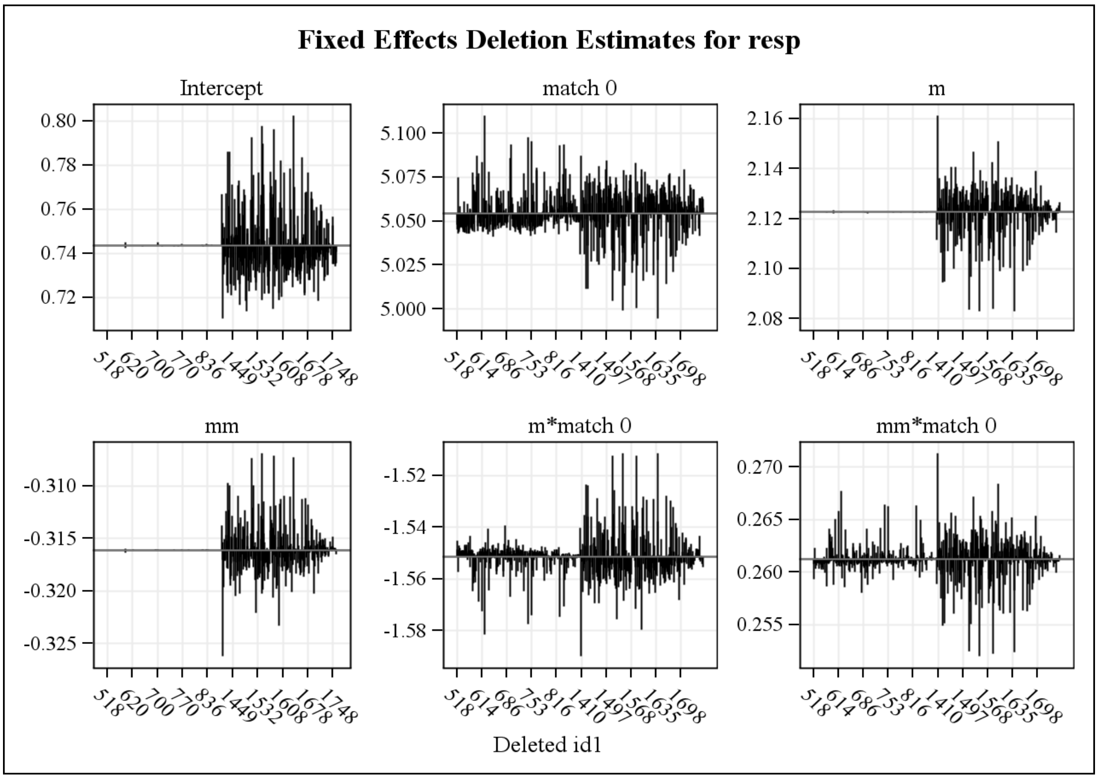
Figure 18: Studentized residual for the lowest age group for before diagnosis

Plots to detect influential observations for the lowest age group are presented here for before and after diagnosis.

Age1 after







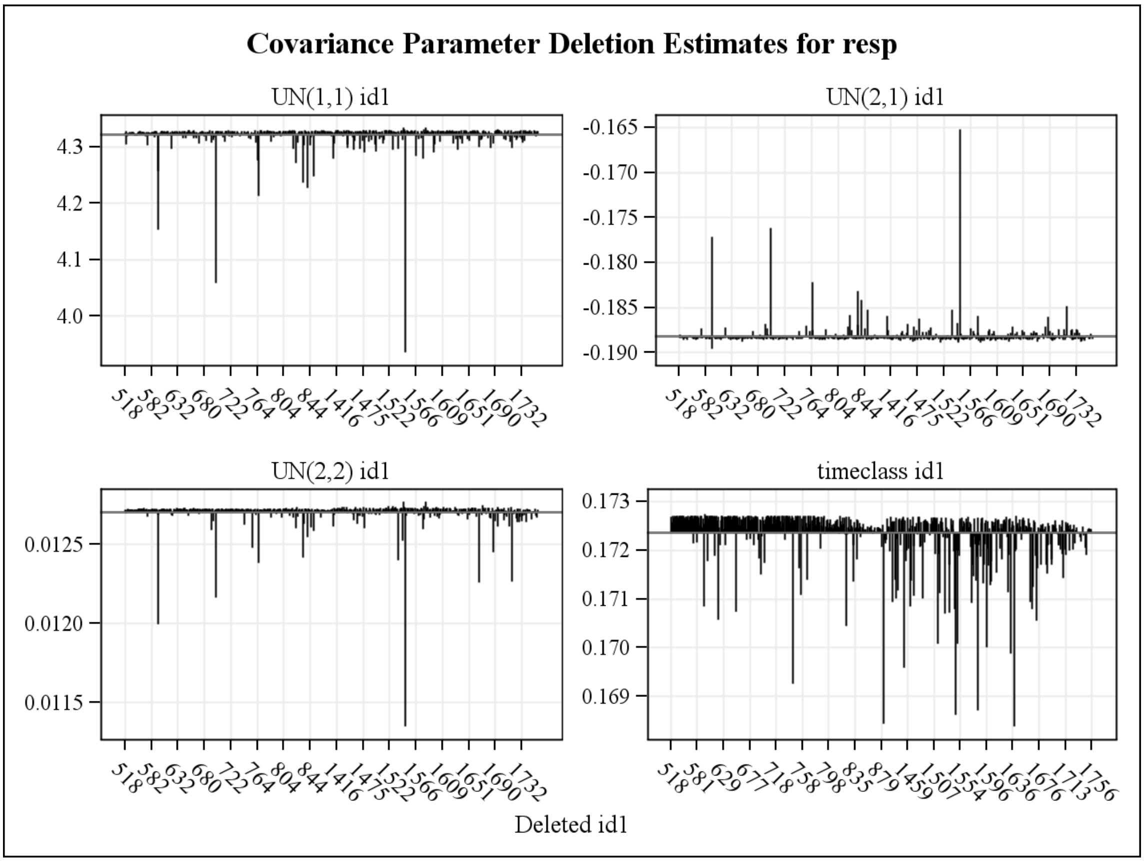
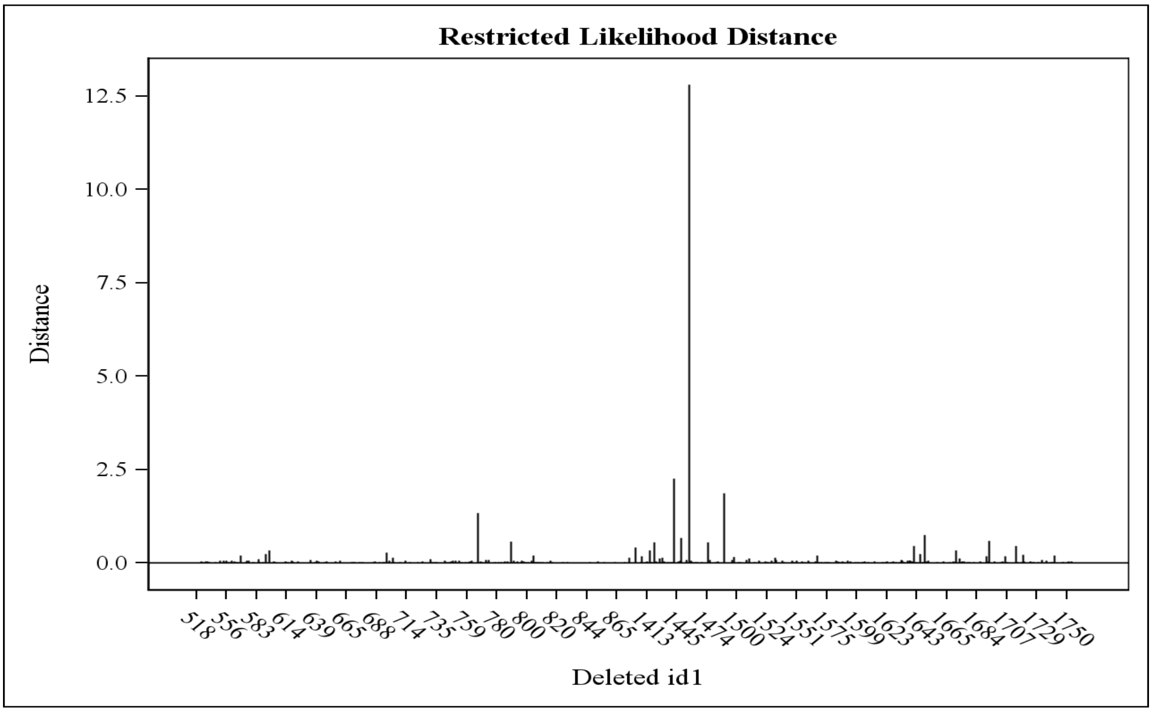
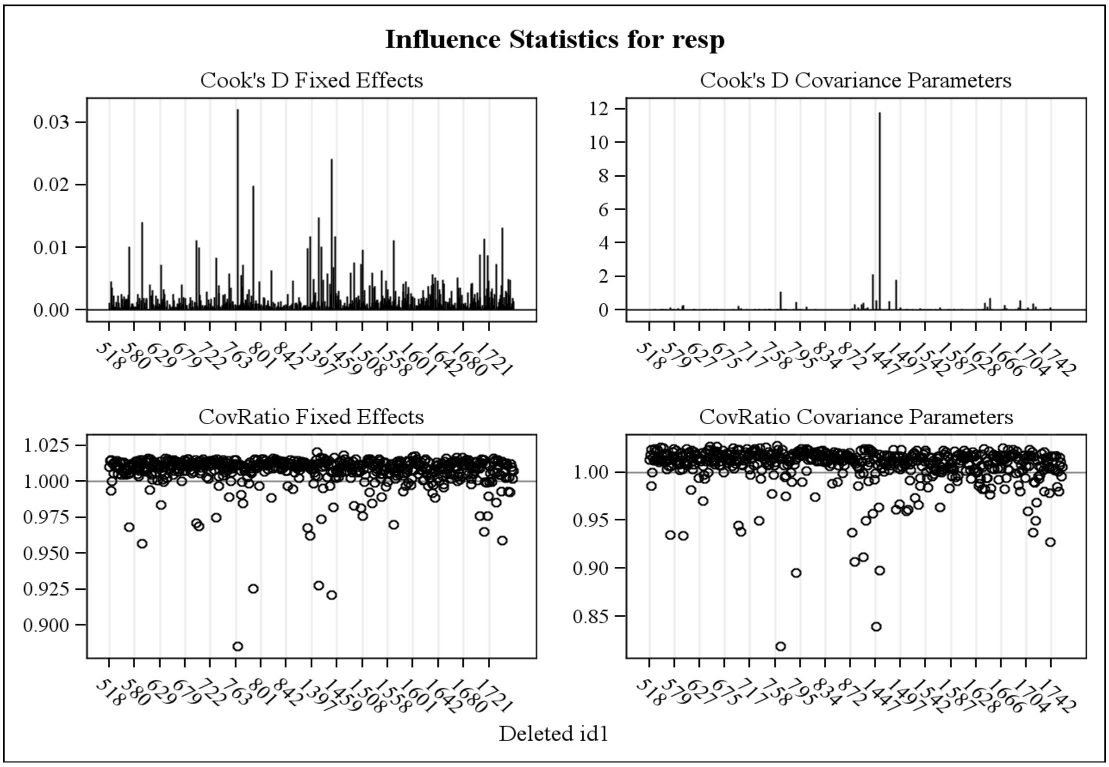


Figure 19 A: Age1 before


Restricted Likelihood Distance
Distance
12.5
10.0
7.5
5.0
2.5
0.0
518 556 583 614 639 665 688 714 735 759 780 800 820 844 865 1413 1445 1474 1500 1524 1551 1575 1599 1623 1643 1665 1684 1707 1729 1750
Deleted id1


Influence Statistics for resp
Cook's D Fixed Effects
0.03
0.02
0.01
0.00
518 580 629 679 722 763 801 842 1397 1459 1508 1558 1601 1642 1680 1721
Cook's D Covariance Parameters
12
10
8
6
4
2
0
518 579 627 675 717 758 795 834 872 1447 1497 1542 1587 1628 1666 1704 1742
CovRatio Fixed Effects
1.025
1.000
0.975
0.950
0.925
0.900
518 580 629 679 722 763 801 842 1397 1459 1508 1558 1601 1642 1680 1721
CovRatio Covariance Parameters
1.00
0.95
0.90
0.85
518 579 627 675 717 758 795 834 872 1447 1497 1542 1587 1628 1666 1704 1742
Deleted id1

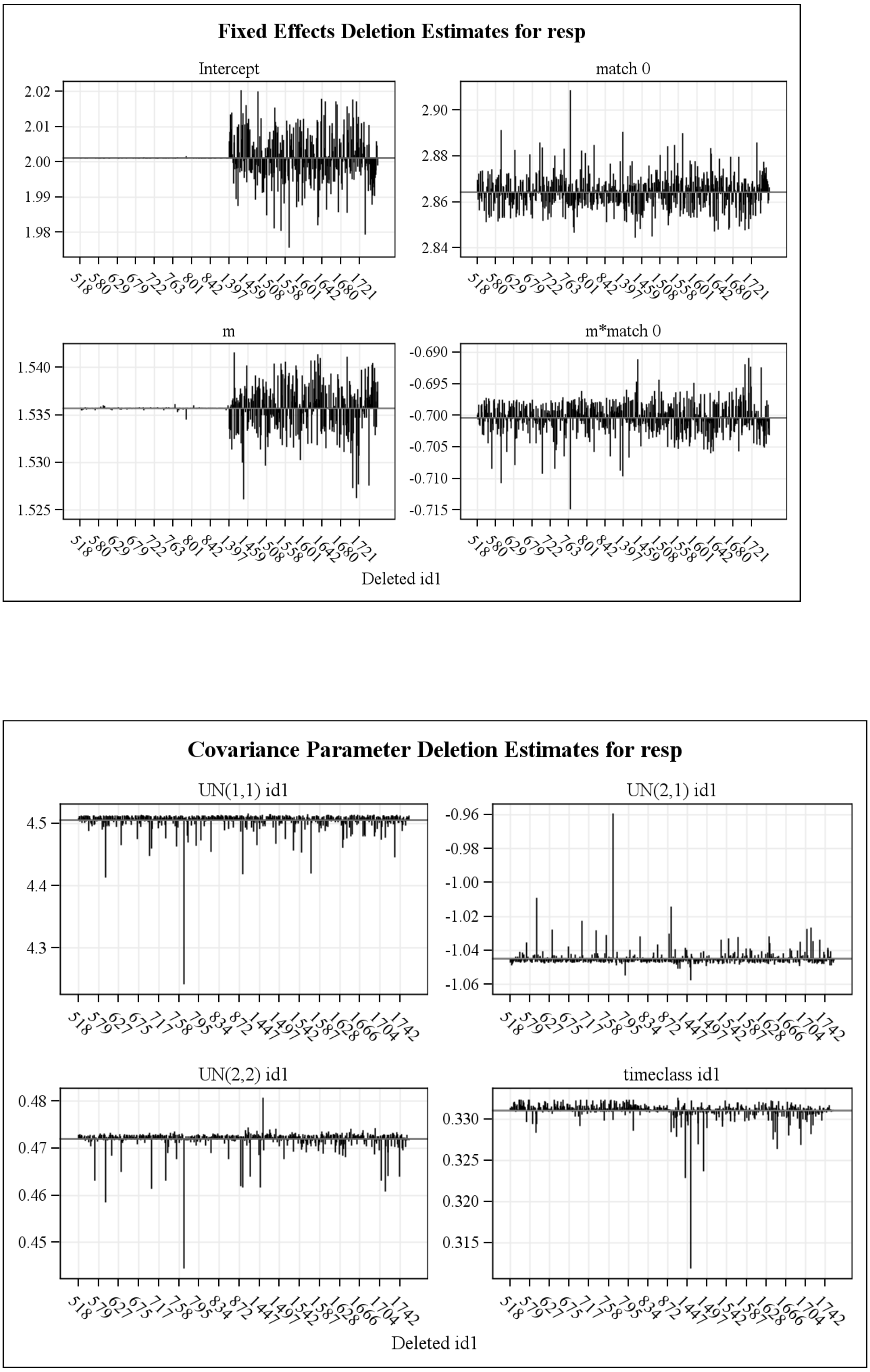
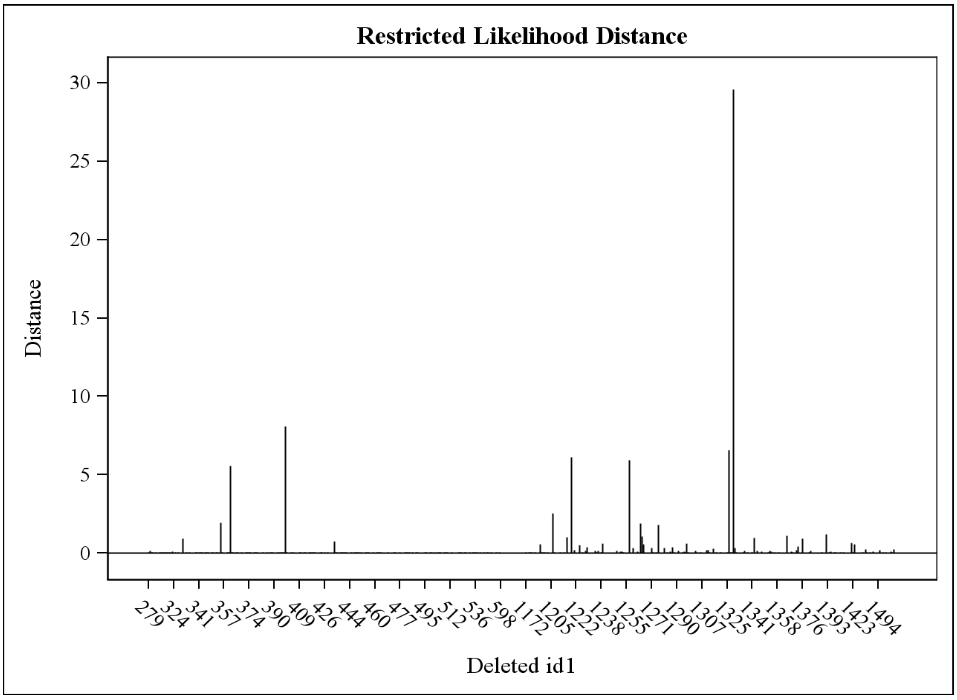
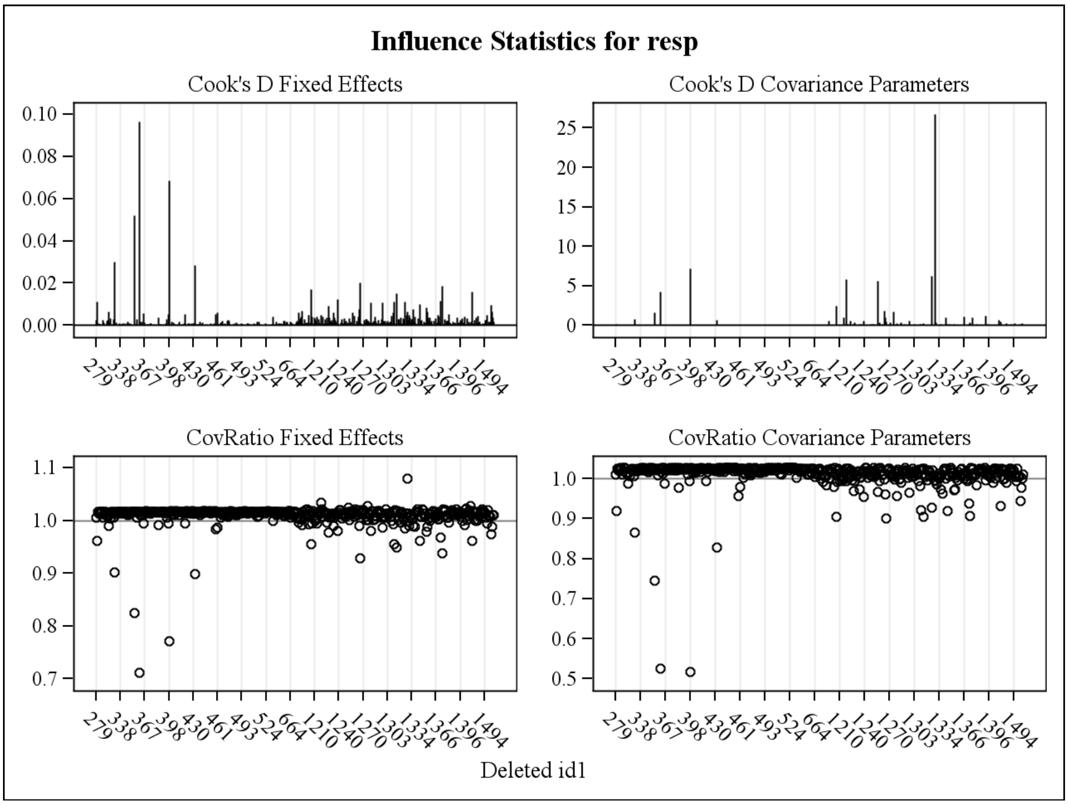
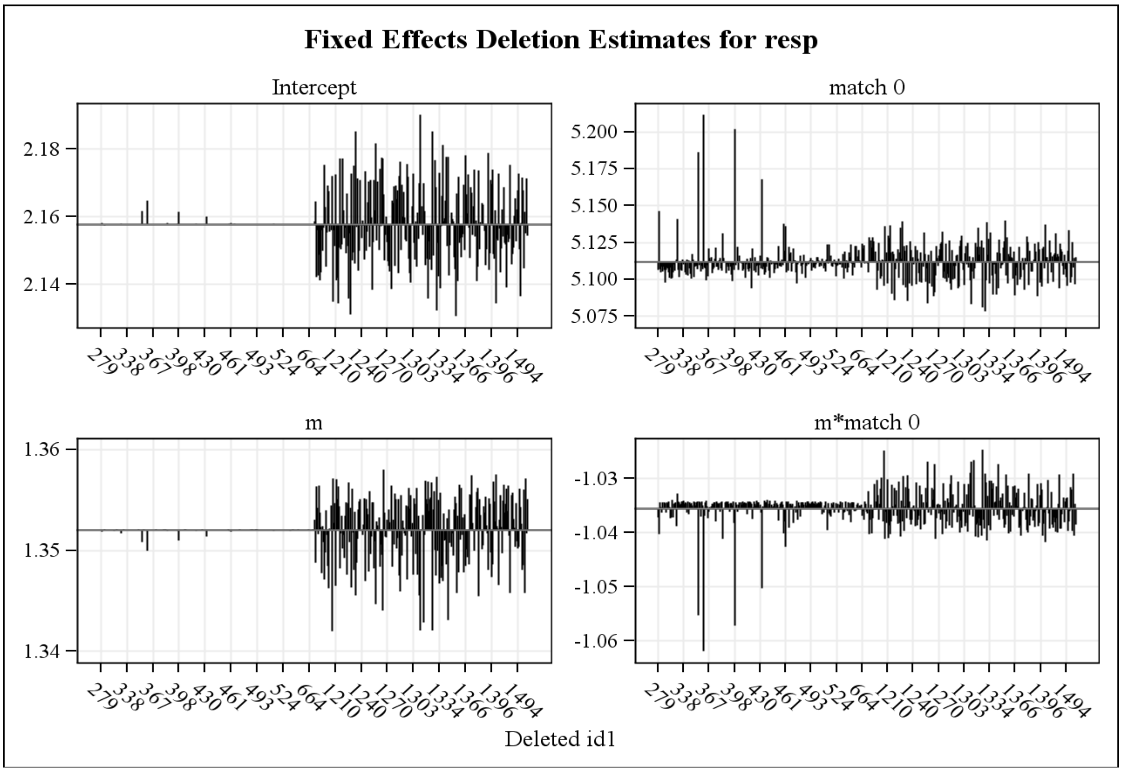


Figure 19 B: Age2 after


Restricted Likelihood Distance
Distance
30
25
20
15
10
5
0
279 324 341 357 374 390 409 426 444 460 477 495 512 536 598 1172 1205 1222 1238 1255 1271 1290 1307 1325 1341 1358 1376 1393 1423 1494
Deleted id1


Influence Statistics for resp
Cook's D Fixed Effects
0.10
0.08
0.06
0.04
0.02
0.00
Cook's D Covariance Parameters
25
20
15
10
5
0
CovRatio Fixed Effects
1.1
1.0
0.9
0.8
0.7
CovRatio Covariance Parameters
1.0
0.9
0.8
0.7
0.6
0.5
279 338 367 398 430 461 493 524 664 1210 1240 1270 1303 1334 1366 1396 1494
Deleted id1



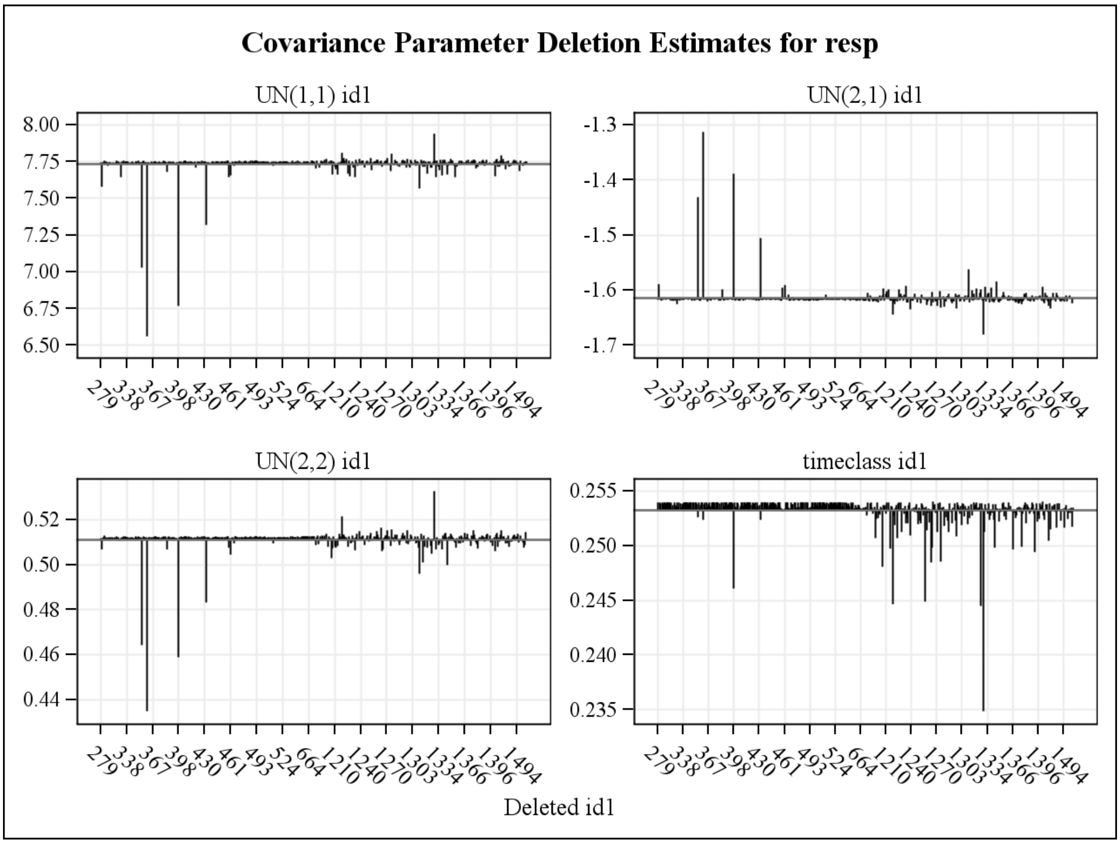
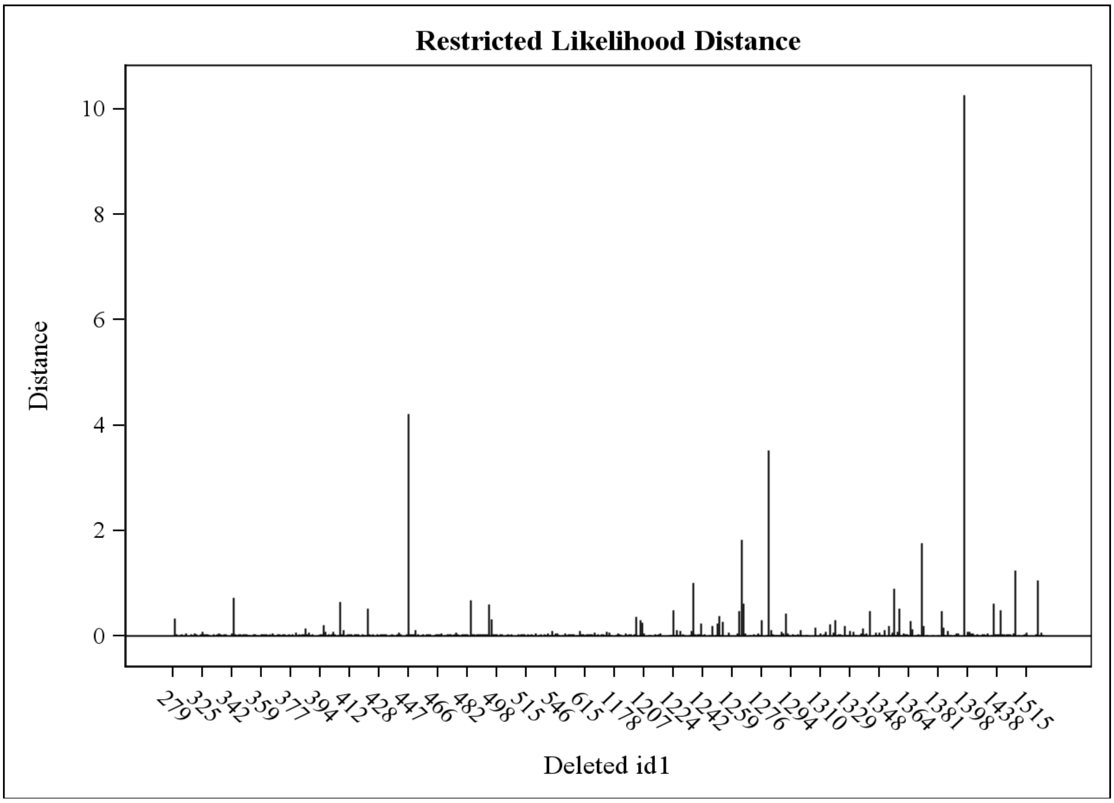
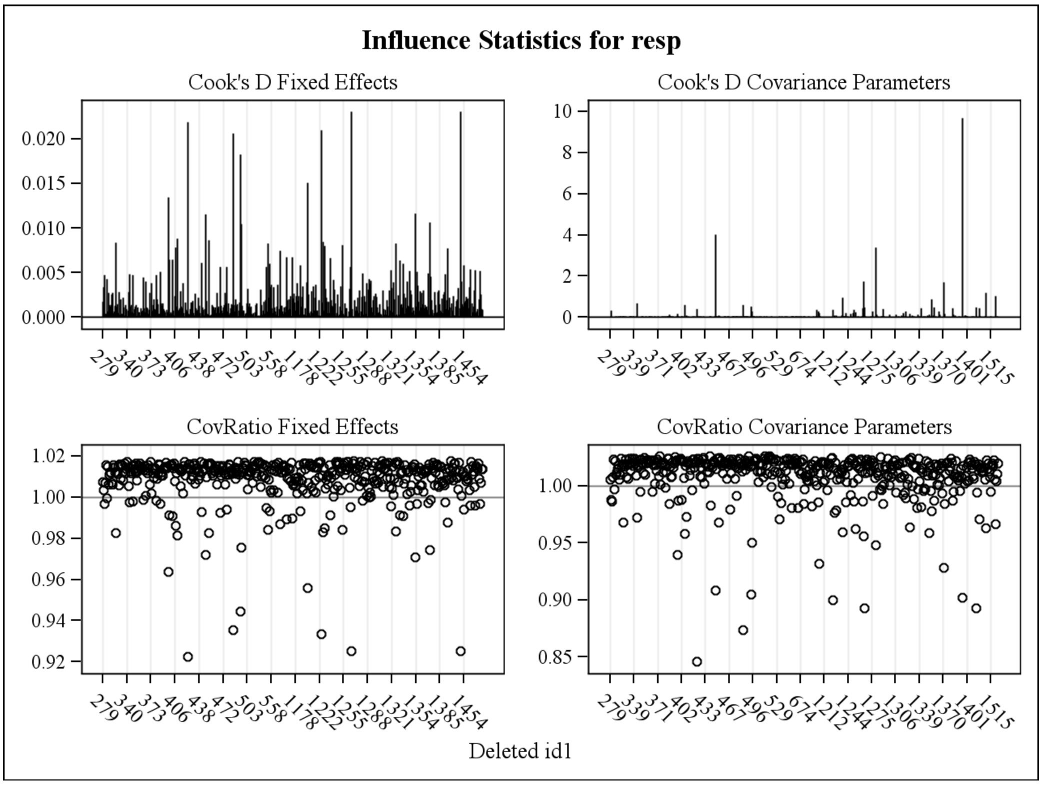
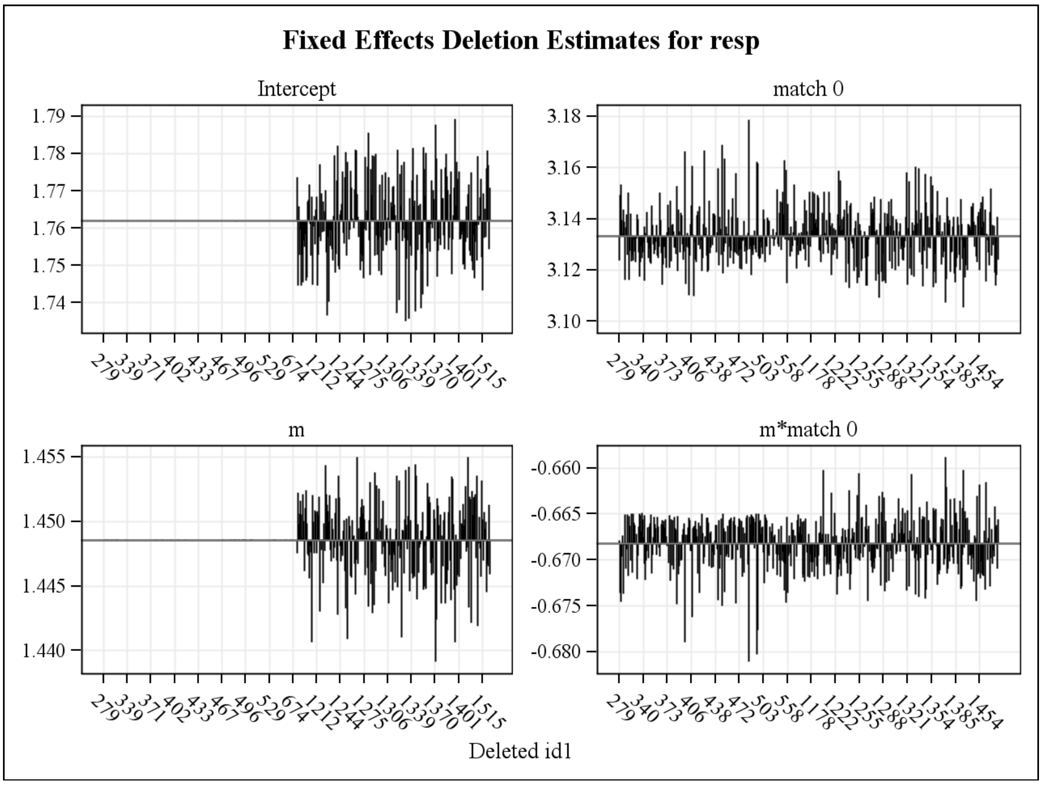


Figure 19 C:Age2 before


Restricted Likelihood Distance
Distance
0
2
4
6
8
10
279
325
342
359
377
394
412
428
447
466
482
498
515
546
615
1178
1207
1224
1242
1259
1276
1294
1310
1329
1348
1364
1381
1398
1438
1515
Deleted id1


Influence Statistics for resp
Cook's D Fixed Effects
0.000
0.005
0.010
0.015
0.020
279
340
373
406
438
472
503
558
1178
1222
1255
1288
1321
1354
1385
1454
Cook's D Covariance Parameters
0
2
4
6
8
10
279
339
371
402
433
467
496
529
674
1212
1244
1275
1306
1339
1370
1401
1515
CovRatio Fixed Effects
0.92
0.94
0.96
0.98
1.00
1.02
279
340
373
406
438
472
503
558
1178
1222
1255
1288
1321
1354
1385
1454
CovRatio Covariance Parameters
0.85
0.90
0.95
1.00
279
339
371
402
433
467
496
529
674
1212
1244
1275
1306
1339
1370
1401
1515
Deleted id1



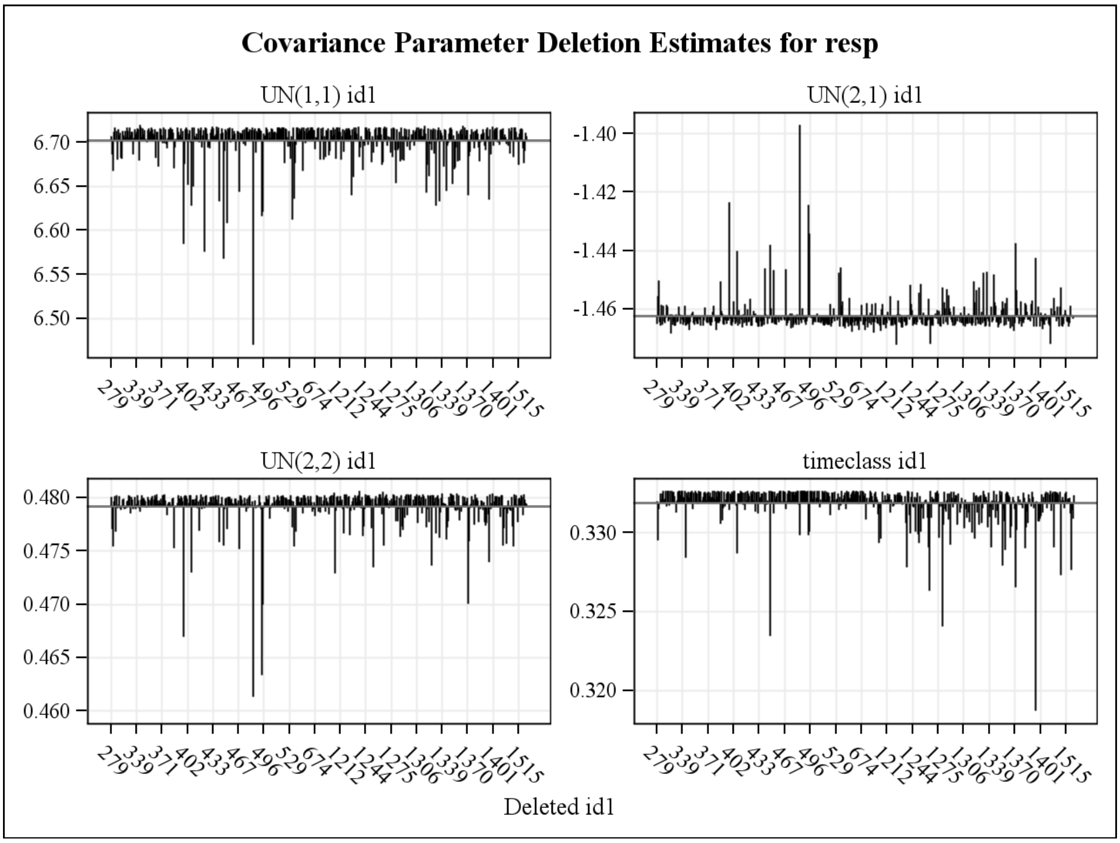
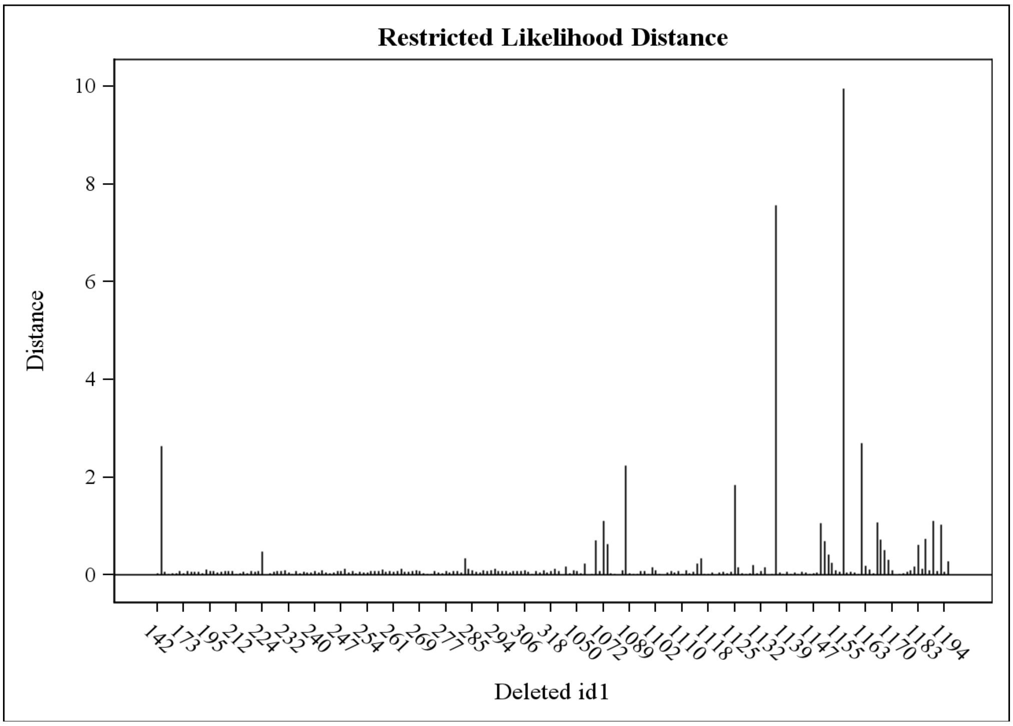
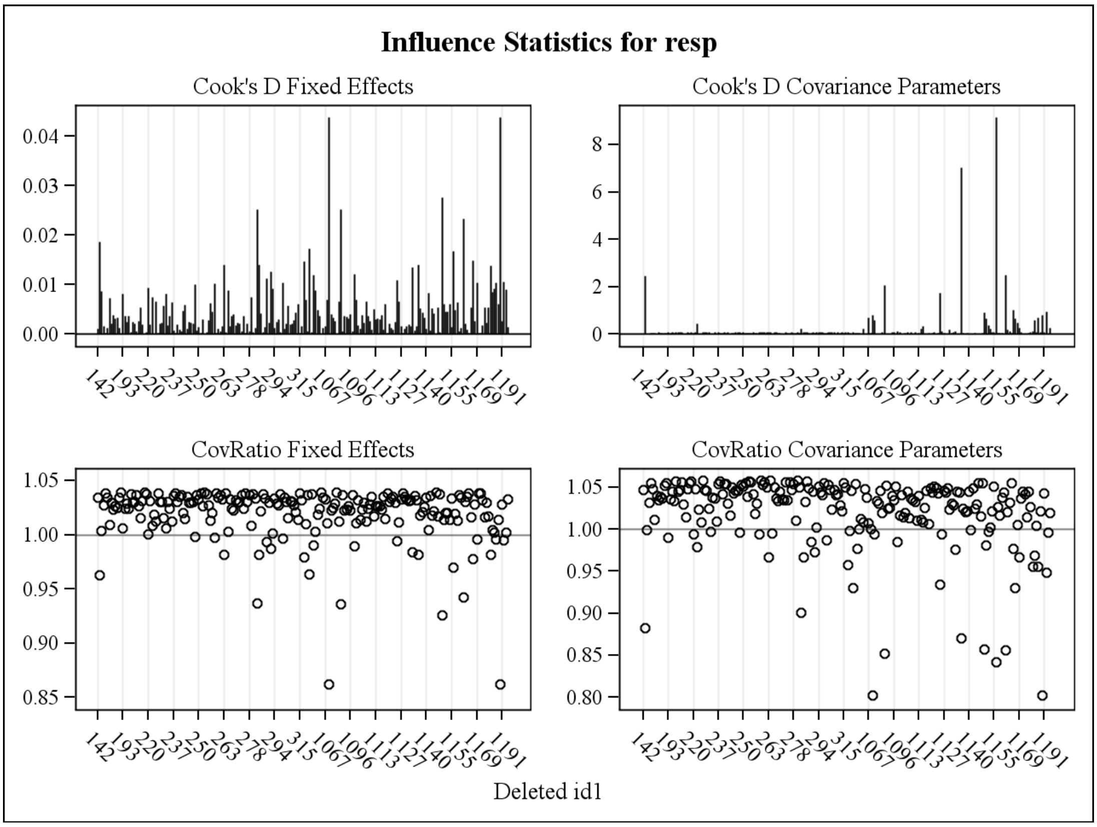
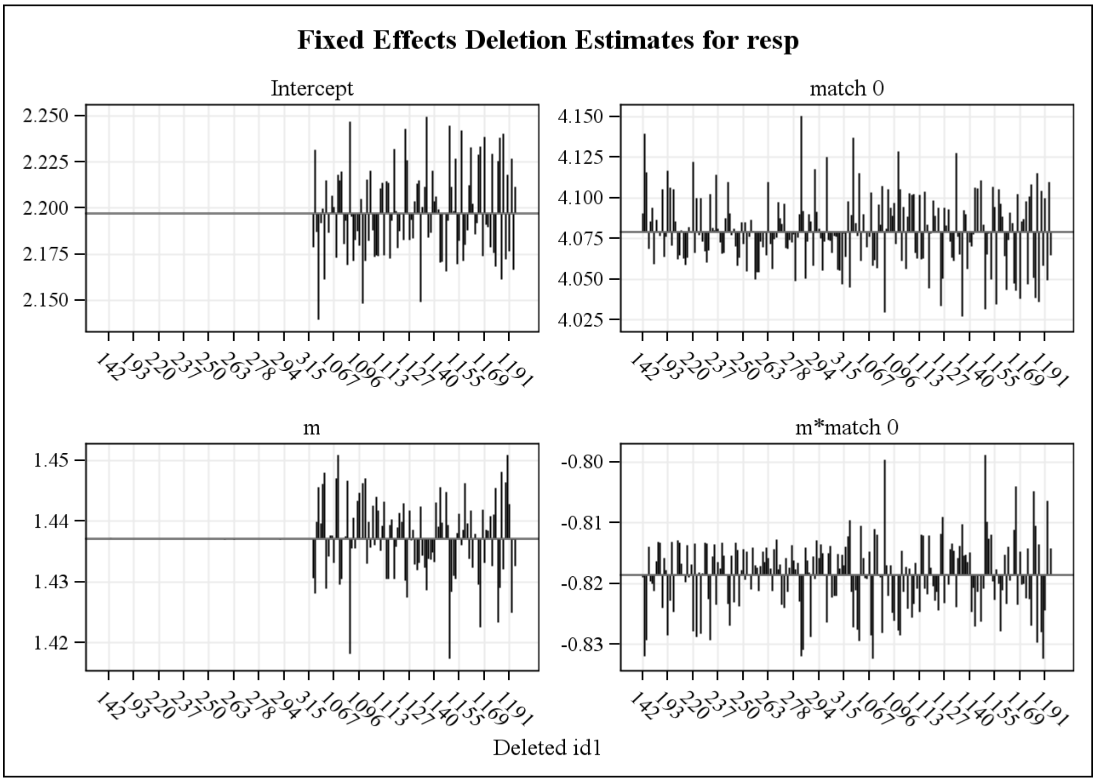


Figure 19 D: Age3 before


Restricted Likelihood Distance
Distance
0
2
4
6
8
10
142
173
195
212
224
232
240
247
254
261
269
277
285
294
306
318
1050
1072
1089
1102
1110
1118
1125
1132
1139
1147
1155
1163
1170
1183
1194
Deleted id1


Influence Statistics for resp
Cook's D Fixed Effects
0.00
0.01
0.02
0.03
0.04
Cook's D Covariance Parameters
0
2
4
6
8
CovRatio Fixed Effects
0.85
0.90
0.95
1.00
1.05
CovRatio Covariance Parameters
0.80
0.85
0.90
0.95
1.00
1.05
142
193
220
237
250
263
278
294
315
1067
1096
1113
1127
1140
1155
1169
1191
Deleted id1



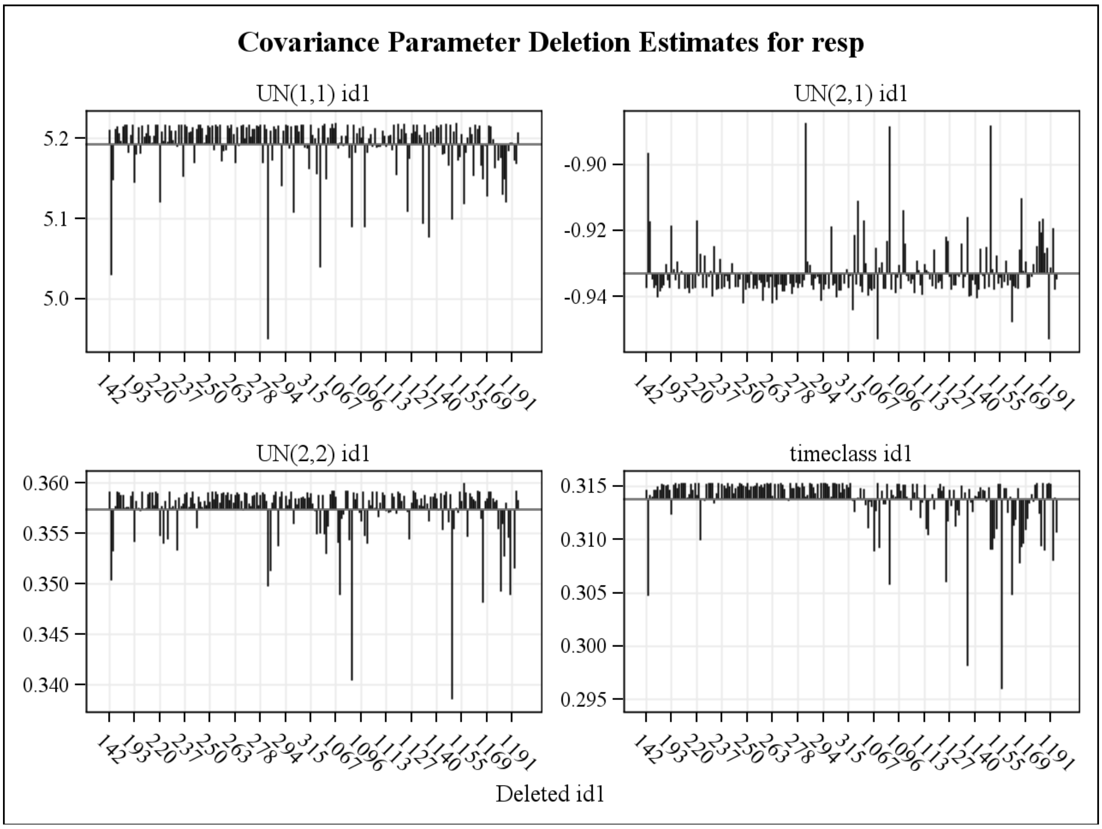
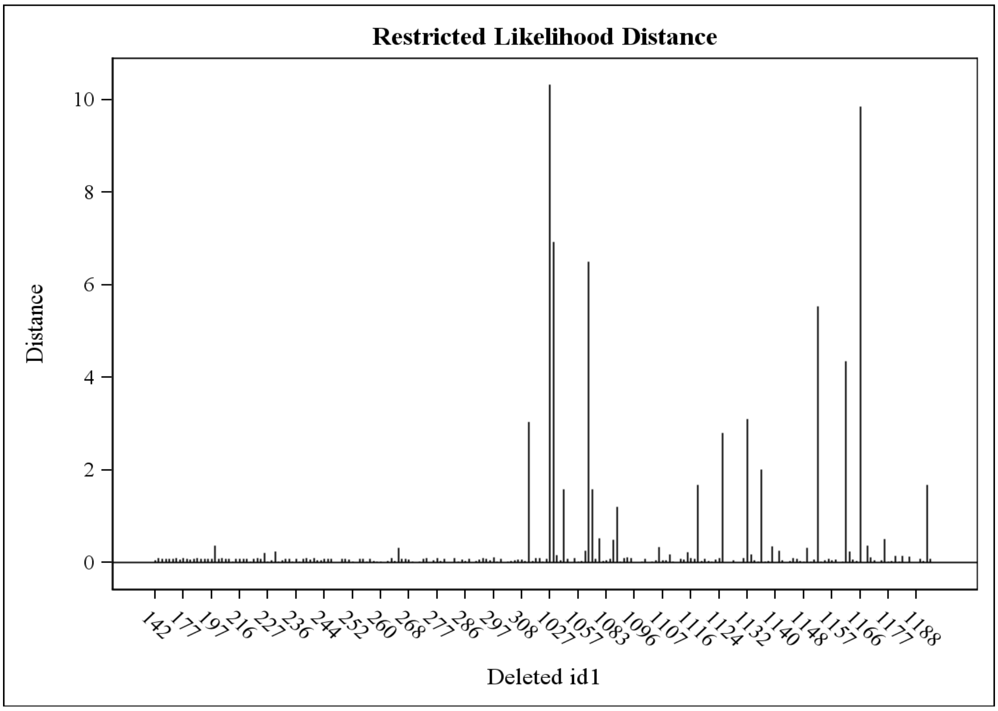
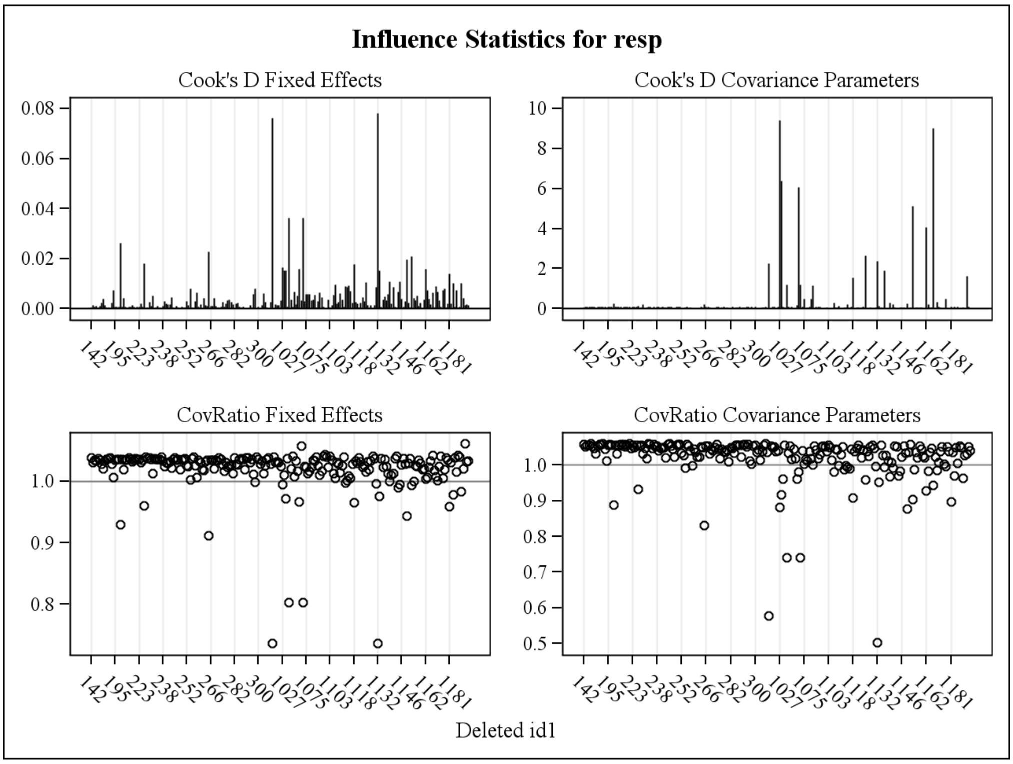
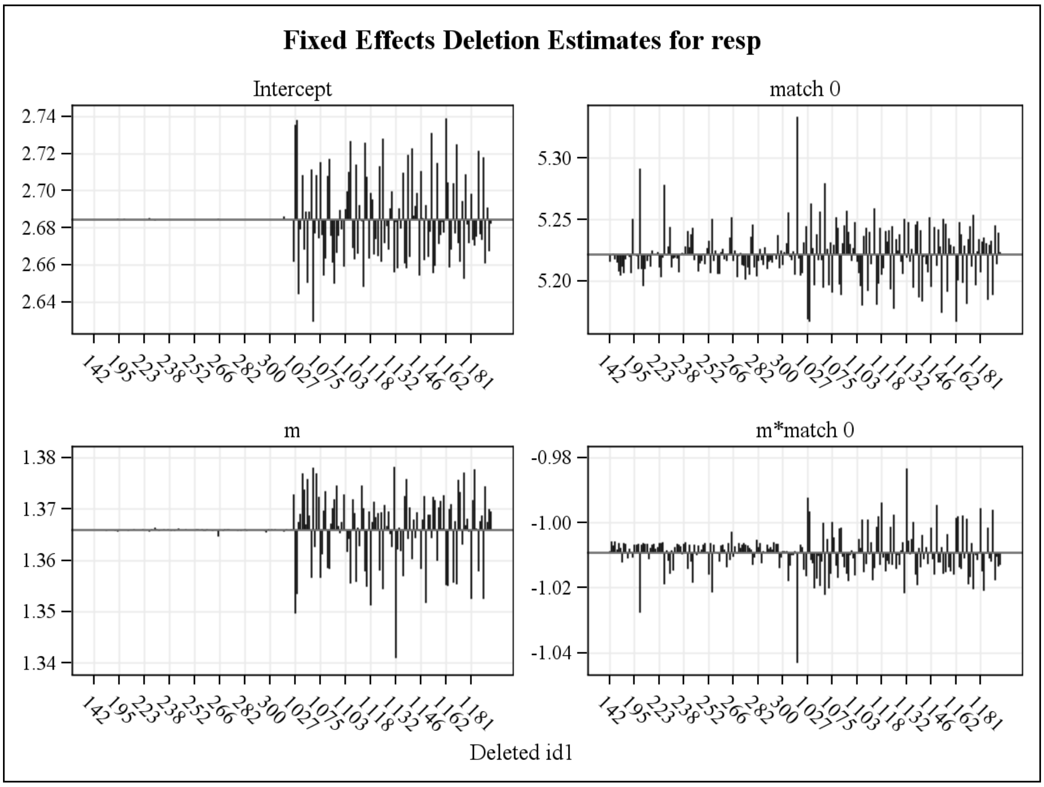


Figure 19 E: Age3 after


Restricted Likelihood Distance
Distance
0
2
4
6
8
10
142
177
197
216
227
236
244
252
260
268
277
286
297
308
1027
1057
1083
1096
1107
1116
1124
1132
1140
1148
1157
1166
1177
1188
Deleted id1


Influence Statistics for resp
Cook's D Fixed Effects
0.00
0.02
0.04
0.06
0.08
Cook's D Covariance Parameters
0
2
4
6
8
10
CovRatio Fixed Effects
0.8
0.9
1.0
CovRatio Covariance Parameters
0.5
0.6
0.7
0.8
0.9
1.0
142
195
223
238
252
266
282
300
1027
1075
1103
1118
1132
1146
1162
1181
Deleted id1



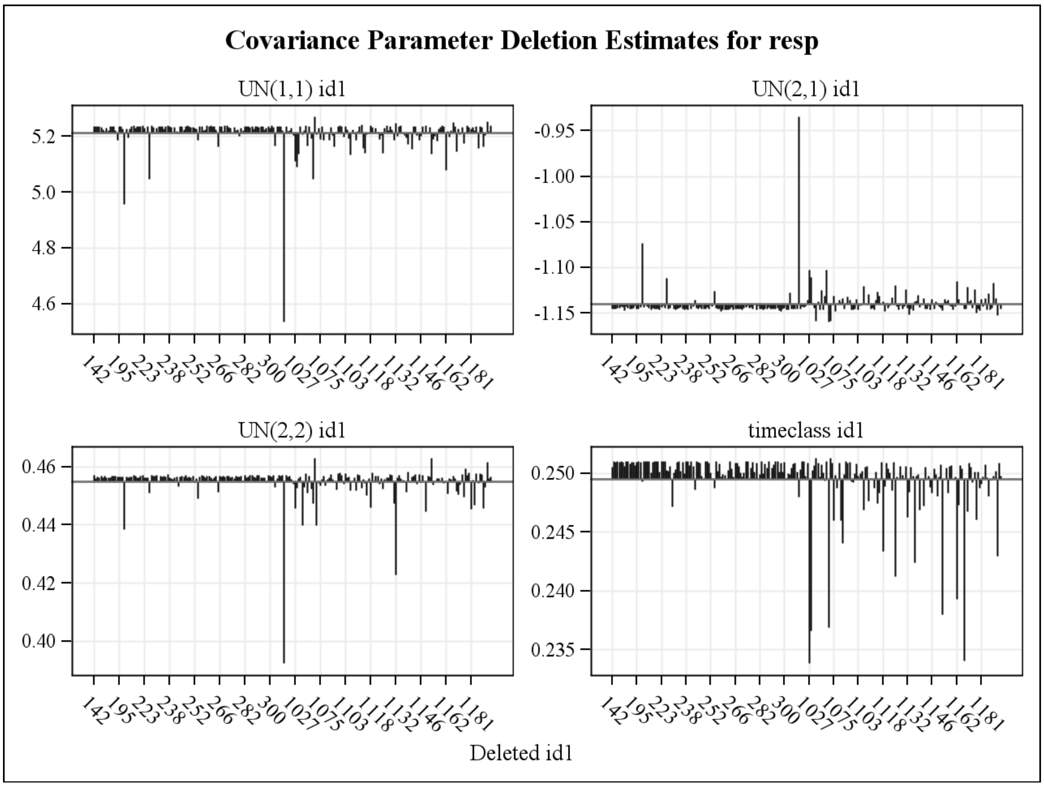
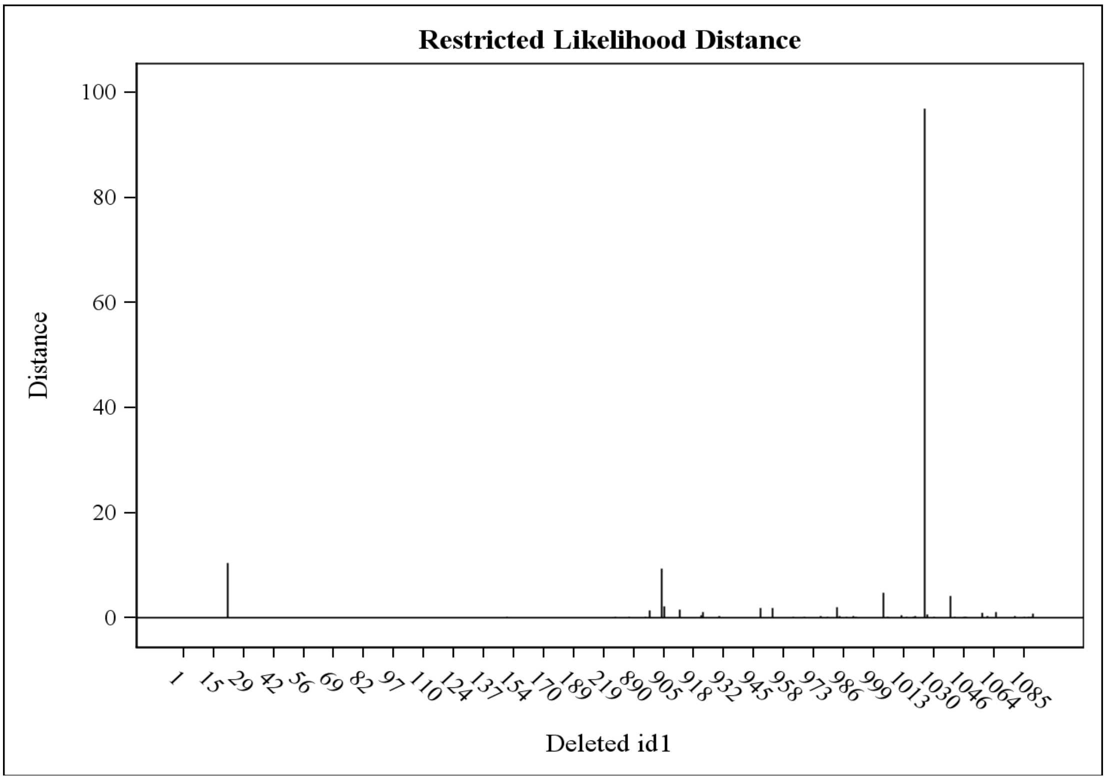
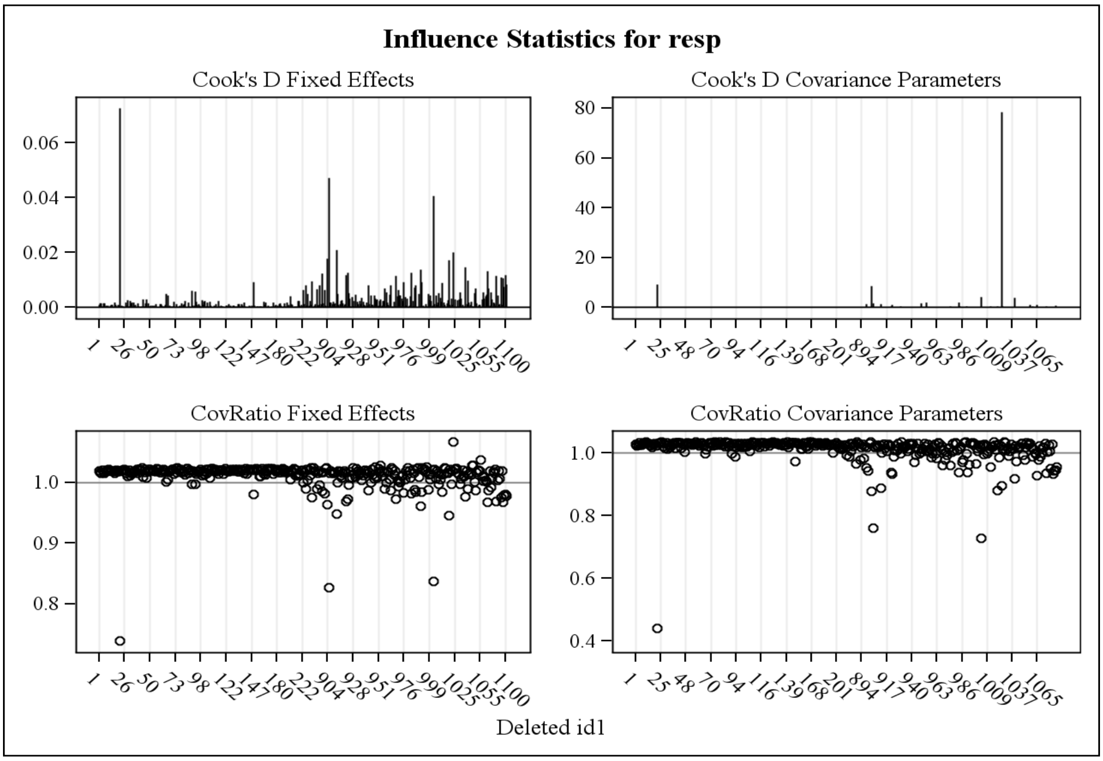


Figure 19 F: Age4 after


Restricted Likelihood Distance
Distance
100
80
60
40
20
0
1
15
29
42
56
69
82
97
110
124
137
154
170
189
219
890
905
918
932
945
958
973
986
999
1013
1030
1046
1064
1085
Deleted id1


Influence Statistics for resp
Cook's D Fixed Effects
0.06
0.04
0.02
0.00
Cook's D Covariance Parameters
80
60
40
20
0
1
26
50
73
98
122
147
180
222
904
928
951
976
999
1025
1055
1100
1
25
48
70
94
116
139
168
201
894
917
940
963
986
1009
1037
1065
CovRatio Fixed Effects
1.0
0.9
0.8
CovRatio Covariance Parameters
1.0
0.8
0.6
0.4
1
26
50
73
98
122
147
180
222
904
928
951
976
999
1025
1055
1100
1
25
48
70
94
116
139
168
201
894
917
940
963
986
1009
1037
1065
Deleted id1

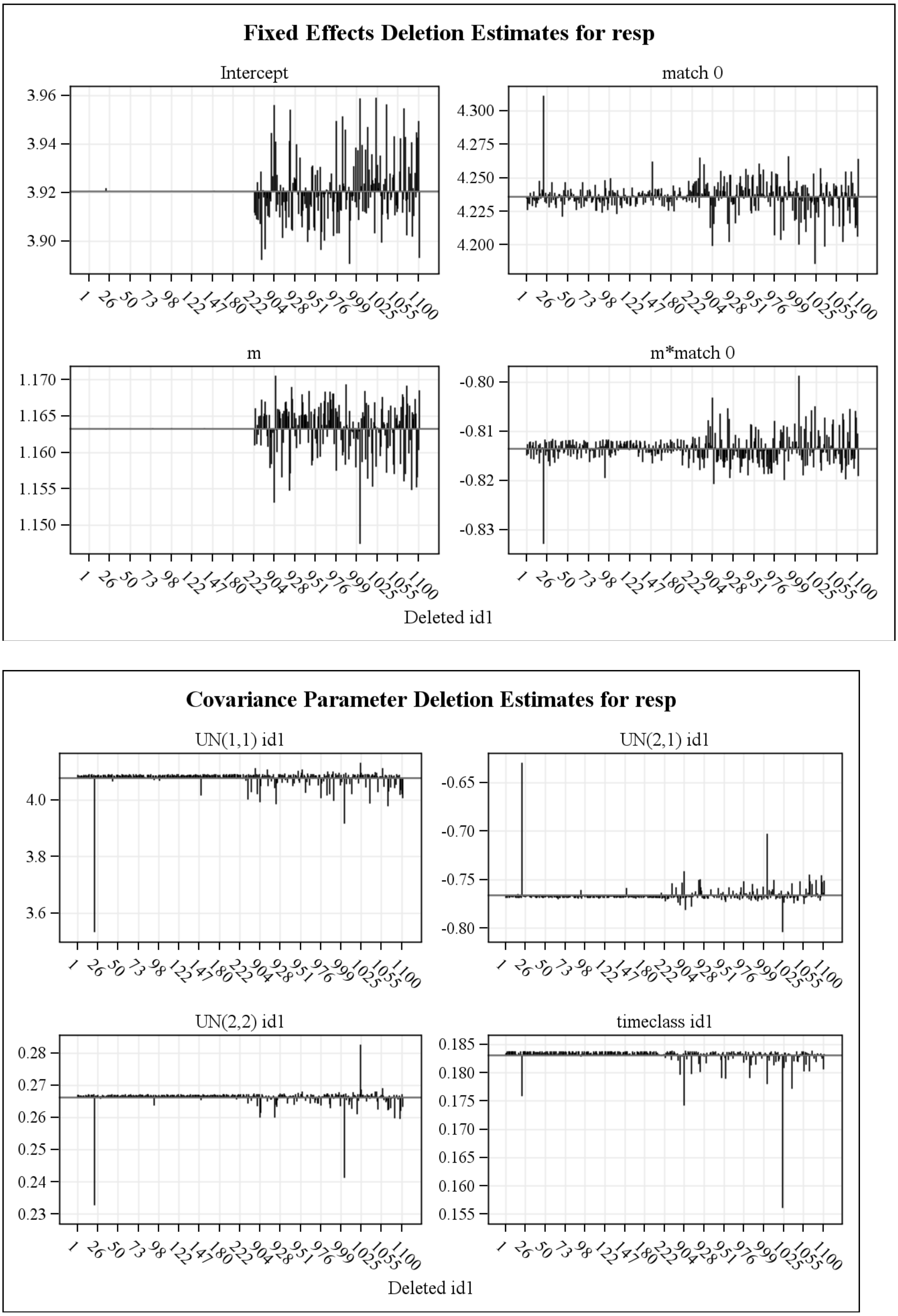
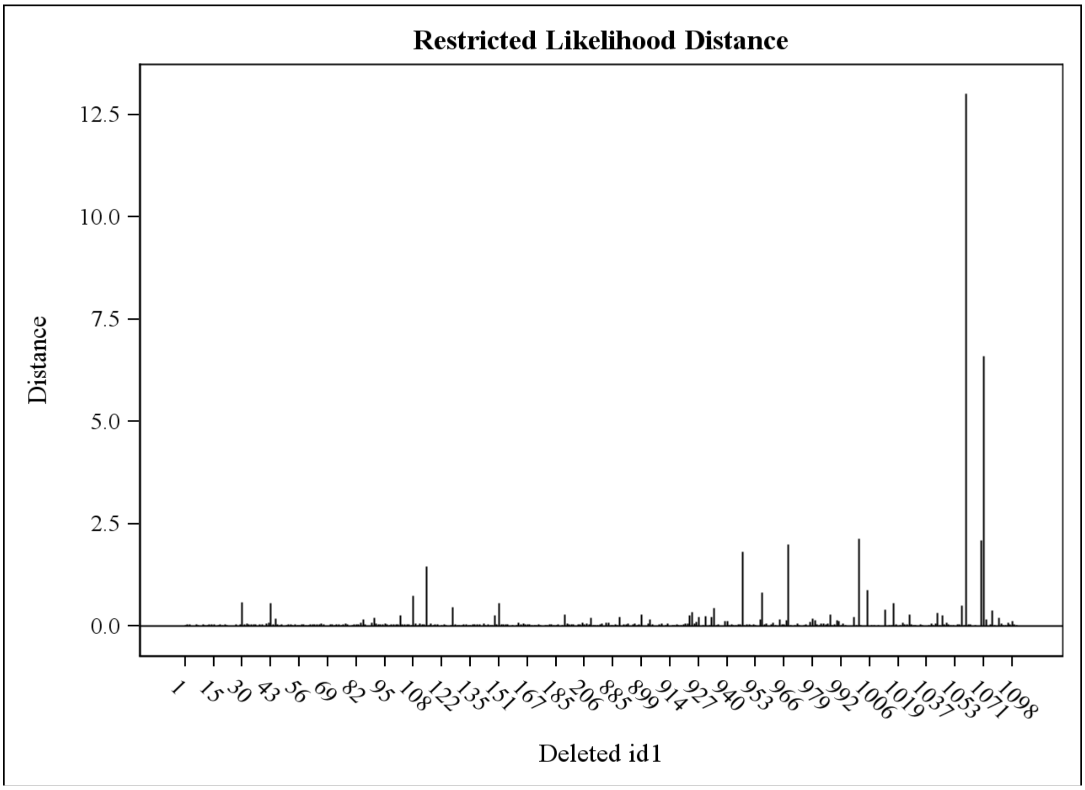
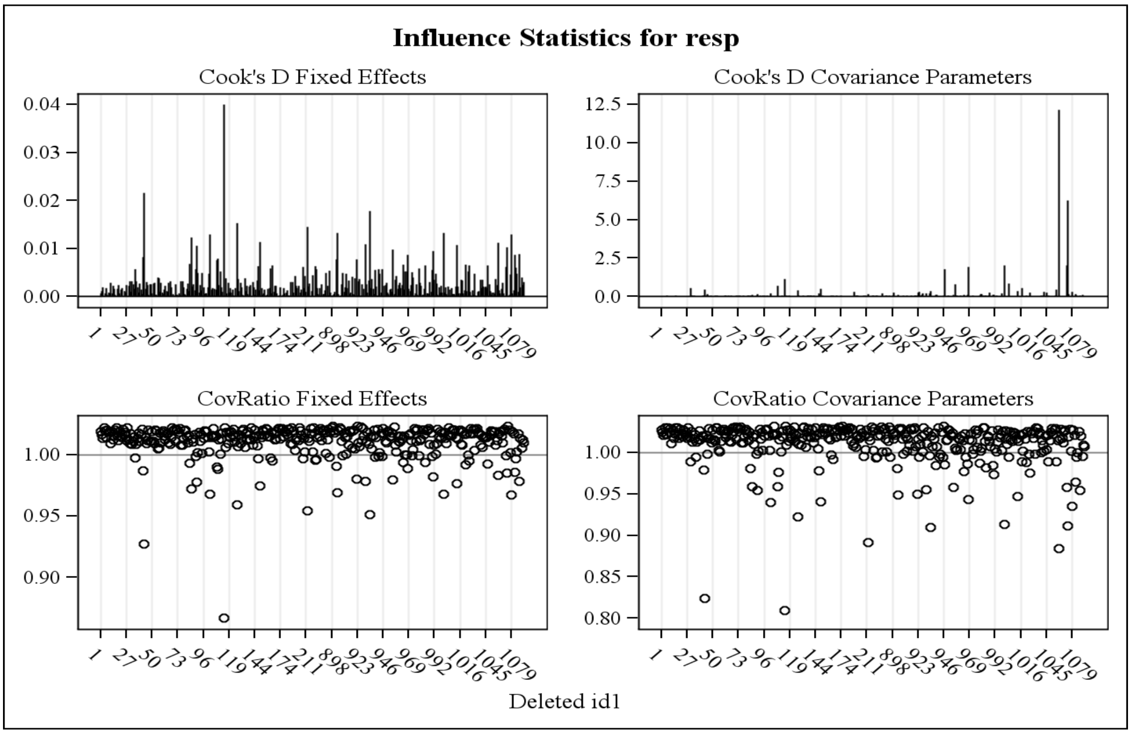
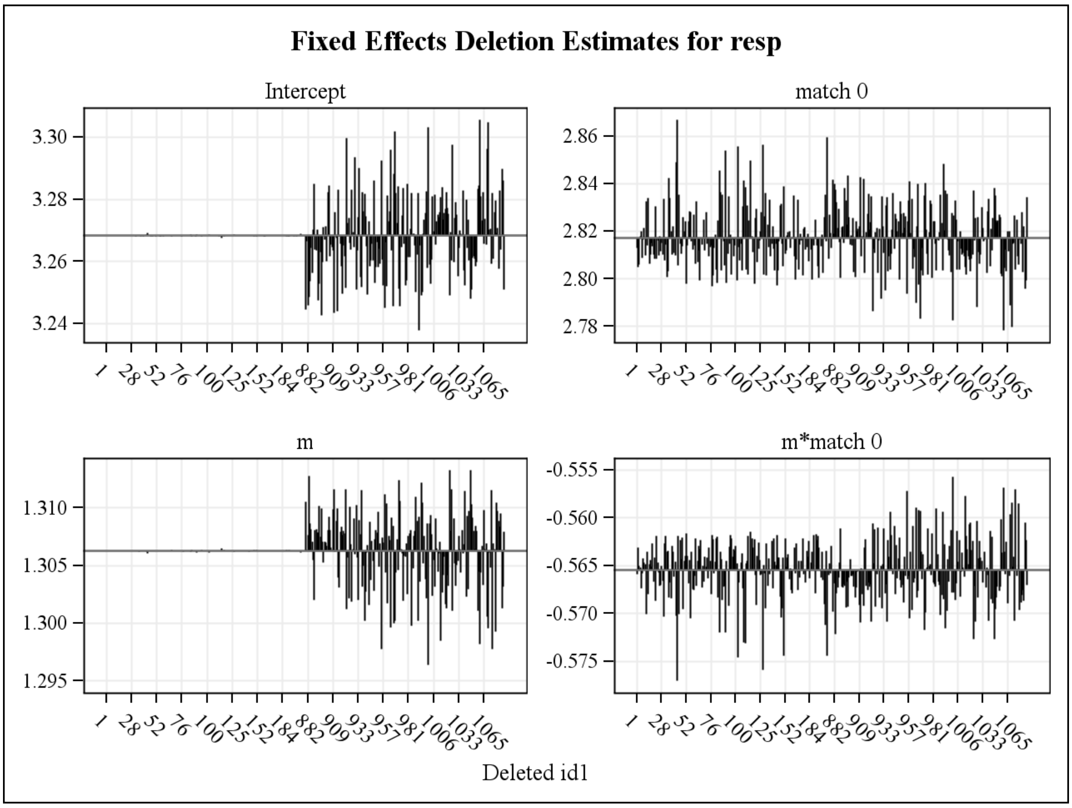
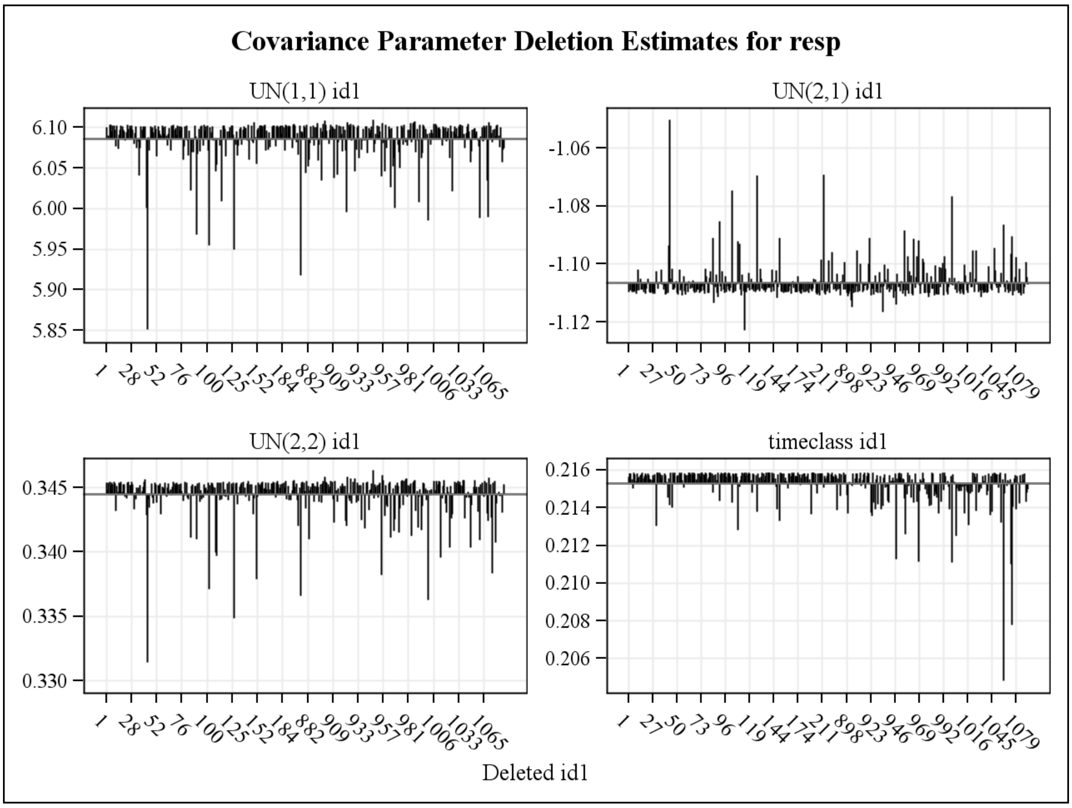


Figure 19 G: Age4 before


Restricted Likelihood Distance
Distance
12.5
10.0
7.5
5.0
2.5
0.0
1
15
30
43
56
69
82
95
108
122
135
151
167
185
206
885
899
914
927
940
953
966
979
992
1006
1019
1037
1053
1071
1098
Deleted id1


Influence Statistics for resp
Cook's D Fixed Effects
0.04
0.03
0.02
0.01
0.00
Cook's D Covariance Parameters
12.5
10.0
7.5
5.0
2.5
0.0
CovRatio Fixed Effects
1.00
0.95
0.90
CovRatio Covariance Parameters
1.00
0.95
0.90
0.85
0.80
1
27
50
73
96
119
144
174
211
898
923
946
969
992
1016
1045
1079
Deleted id1


Fixed Effects Deletion Estimates for resp
Intercept
3.30
3.28
3.26
3.24
match 0
2.86
2.84
2.82
2.80
2.78
m
1.310
1.305
1.300
1.295
m*match 0
-0.555
-0.560
-0.565
-0.570
-0.575
1 28 52 76 100 125 152 184 882 909 933 957 981 1006 1033 1065
Deleted id1


Covariance Parameter Deletion Estimates for resp
UN(1,1) id1
6.10
6.05
6.00
5.95
5.90
5.85
UN(2,1) id1
-1.06
-1.08
-1.10
-1.12
1 27 50 73 96 119 144 174 211 898 923 946 969 992 1016 1045 1079
UN(2,2) id1
0.345
0.340
0.335
0.330
timeclass id1
0.216
0.214
0.212
0.210
0.208
0.206
1 28 52 76 100 125 152 184 882 909 933 957 981 1006 1033 1065
Deleted id1

Appendix C: SAS Codes

```
/*** semiparametric mixed mode ***/

ods rtf file= "before.rtf";
proc glimmix data=prj.before method=mmpl;
class id1 match;
model cost= match month match*month/  solution;
random month /type=rsmooth
knotmethod=equal(10) subject=id1;
random int  /type=un subject=id1;
run;
ods rtf close;


        /***second degree***/


%macro fracpol;

data Fitsummary; run;

%let p1=%sysevalf(-2);

%do %while (%sysevalf(&p1<=2));

data hlp;
set pneumoc;
agem=month;
if &p1=0 then agep1=log(agem); else
agep1=agem**(&p1);
run;

ods output FitStatistics=Fit;
proc mixed data=hlp method=ml ic covtest update
noinfo absolute noclprint maxiter=200;
class id1 timeclass;
model resp= agep1 /solution ;
random int /type=un subject=id1  ;
repeated timeclass/ type=simple subject=id1  ;
run;

data Fit; set Fit (keep=Descr value); run;
proc transpose data=Fit
out=Fit2(keep=col1 col2 col3 col4);
run;

data Fit3 (keep=power1 power2 LogLik AIC AICC
BIC);
set Fit2;
LogLik=col1;
AIC=col2;
AICC=col3;
BIC=col4;
power1=&p1;
run;

data FitSummary; set FitSummary Fit3; run;
```

```
%let p1=%sysevalf(&p1+0.5);
%end;


%let p1=%sysevalf(-2);
%do %while (%sysevalf(&p1<=2));
%let p2=%sysevalf(&p1);
%do %while (%sysevalf(&p2<=2));


data hlp;
set pneumoc;
agem=month;
if &p1=0 then agep1=log(agem); else
agep1=agem**((&p1)/1);
if &p1=&p2 then do;
if &p2=0 then agep2=log(agem)**2; else
agep2=(agem**((&p2)/1))*log(agem); end;
else do;
if &p2=0 then agep2=log(agem); else
agep2=agem**((&p2)/1); end;
run;

ods output FitStatistics=Fit;
proc mixed data=hlp method=ml ic covtest update
noinfo absolute noclprint maxiter=200;
class id1  timeclass;
model resp= agep1 agep2 /solution ;
random  int /type=un subject=id1  ;
repeated timeclass/ type=simple subject=id1  ;
run;

data Fit; set Fit (keep=Descr value); run;
proc transpose data=Fit
                    out=Fit2(keep=col1 col2 col3
col4);
run;
data Fit3 (keep=power1 power2 LogLik AIC AICC
BIC);
set Fit2;
LogLik=col1;
AIC=col2;
AICC=col3;
BIC=col4;
power1=&p1;
power2=&p2;
run;

data FitSummary; set FitSummary Fit3; run;

%let p2=%sysevalf(&p2+0.5);
%end;
%let p1=%sysevalf(&p1+0.5);
%end;
```

```
proc print data=FitSummary; run;

proc sql;
create table SelectModel1 as
select power1
from FitSummary
having BIC=min(BIC);
quit;

proc print data=SelectModel1;
run;

proc sql;
create table SelectModel2 as
select power2
from FitSummary
having BIC=min(BIC);
quit;

proc print data=SelectModel2;
run;

proc sort data=FitSummary;
by AIC;
run;

%mend;
%fracpol
```

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Flexible modeling of the cost evolution of pneumococcal infections

Richting: **Master of Statistics-Biostatistics**
Jaar: **2012**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Tohye, Abera Mulugeta

Datum: **14/09/2012**