Certification

This is to certify that this report was written by Antonio Rungo under our supervision.

Signature: ______________________	Date: ______________________
Prof. dr. Niel HENS	**Internal Supervisor**
Signature: ______________________	Date: ______________________
Prof. dr. Philippe BEUTELS	**External Supervisor**
Signature: ______________________	Date: ______________________
Dr. Joke BILCKE	**External Supervisor**
Signature: ______________________	Date: ______________________
Antonio Rungo	**Student**

Thesis submitted in partial fulfilment of the requirements for the degree of Master of Statistics: Biostatistics.

Acknowledgements

First of all, I dedicate this work to Lord Jesus Christ and I thank Him for giving me life, health, protection, wisdom and strength to complete successfully this Master program. I also thank all my brothers in Christ that days and nights remember me in their prayers.
I wish to thank several people who have contributed to the successful realization of this project. My sincere appreciation goes to my supervisors, Prof. dr. Niel Hens, Prof. dr. Philippe Beutels and Dr. Joke Bilcke for their guidance, suggestions and advice while working on this project. I also would like to express my heartfelt appreciation for my professors at Hasselt University. It was a great honor for me to have attended this program. To all the University Staff, especially Mrs. Martine Machiels who helped us to settle, my heartfelt thanks.
I extend my gratitude to Dr. Nelson Cuamba and Dr. Francisco Mbofana for the financial support, which has enabled me to fulfill my dreams.
Special thanks to all my classmates for all the intellectual enrichment, interesting, and fruitful discussions that I had with them. I am indebted to them for making my school experience a very memorable one! To all my group members, colleagues and friends from I have learnt a lot from you.
Finally, I wish to thank all my family for their support during this 2 years of Master program. Especially, I dedicate this work to Naory.

Antonio Rungo
University of Hasselt
Belgium, December, 2013

Abstract

Background: Health-related quality of life (HRQoL) is an important outcome parameter in clinical trial and epidemiological research to support policy decision making or to monitor population health. With scarce resources for the provision of health care, choices have to be made about how those resources are allocated. The impact on the HRQoL of the population should be an important consideration when these choices are made. The aim of this study was to identify background characteristics of children, adults and elderly that are important in determining the HRQoL of these 3 age groups; to model HRQoL as a function of these covariates and to investigate if HRQoL is more alike in members from the same household.

Methodology: Statistical models were applied on two datasets: one sample of individuals belonging to one of three age groups (children, adults or elderly), another sample of households, with information of all members of each household. HRQoL was measured in two different ways, resulting in a VAS and EQ-5D score for each individual. Regression tree, random forest, lasso and elastic net were used to identify possibly important background characteristics. Thereafter, the relationships between the two HRQoL responses and these factors were modeled using beta regression, one-inflated beta regression and beta GLMM, for separated and joint responses.

Results and Conclusions: Age was significantly associated with both responses in all age groups. Girls and children who had experienced serious disease had significantly lower EQ5D scores. The effect of the number of persons in the household on the probability to be in perfect health is different for girls than for boys. If not in perfect health, adults who had experienced serious disease and adult who had primary and vocational level of education had significantly lower EQ5D scores. Having one or more domestic animal, VAS score increases more in adults. For elderly who had history of smoking (quit smoking) and for those not smoking, EQ5D score is higher than for actively smoking elderly. Elderly who had experienced serious disease, and elderly with primary and vocational level of education are estimated to have significantly lower VAS scores. It was found that individuals from the same household had EQ5D health scores more similar to each other than to any person from a random household. Significant association between the health scores of EQ5D and VAS was present.

Keywords: *Beta regression, Generalized linear mixed model, Health-related quality of life, One-inflated beta regression.*

Contents

List of Tables

List of Figures

List of Abbreviations

AIC	Akaike Information Criteria
BC	Bias Correction
BMI	Body mass index
CDC	Centers for Disease Control and Prevention
EQ-5D	Euroqol 5 Dimensions
GLMM	Generalized Linear Mixed Model
HRQoL	Health-related quality of life
Lasso	Least Absolute Shrinkage and Selector Operator
LRT	Likelihood Ratio Test
ML	Maximum Likelihood
OLS	Ordinary Least Squares
QoL	Quality of Life
RF	Random Forest
SD	Standard Deviation
VAS	Visual Analogue Scale

Chapter 1

Introduction

1.1 Background

Health-related quality of life (HRQoL) is an important outcome parameter in clinical trial and epidemiological research to support policy decision making or to monitor population health (Hunger et al., 2012). With scarce resources for the provision of health care, choices have to be made about how those resources are allocated. The impact on the HRQoL of the population should be an important consideration when these choices are made (Dolan, 1997). HRQoL measures have been widely used in health research in recent years and have been the endpoint in many clinical studies. The widespread use of HRQoL measures reflects the recognition that many treatments for chronic diseases fail in providing a cure, and therefore, the benefits of therapy may be limited. In some circumstances, the clinical therapeutic benefits may be outweighed by HRQoL considerations (Santana and Feeny, 2008).

HRQoL is a multidimensional concept referring to how people perceive aspects of their lives that relate to their health (Alsén, 2009), whereas Quality of Life (QoL) has a broader concept and is related to individuals' perceptions of their position in all areas of life. Therefore, HRQoL rests on both the concept of health and the concept of QoL (WHOQOL, 1998). There is no single and accepted definition of HRQoL, but a consensus that assessments should include perceptions of general health, physical functioning, physical symptoms, emotional functioning, cognitive functioning, role functioning, social well-being and functioning, sexual functioning and existential issues (Alsén, 2009; Guyatt, 1993; Guyatt et al. 1993).

As Dominick et al., 2002 and DeSalvo et al., 2006 said, "HRQoL questions about perceived physical and mental health and function have become an important component of health surveillance and are generally considered valid indicators of service needs and intervention outcomes. Self-assessed health status also proved to be more powerful predictor of mortality and morbidity than many objective measures of health". HRQoL measures make it possible to demonstrate scientifically the impact of health on quality of life, going well beyond the old paradigm that was limited to what can be seen under a microscope.

1.2 Types of HRQoL measures

There are a large number of measures that differ in the range of health dimensions that they cover. **Preference-based measures** give scores on scale from 0.00 (dead) to 1.00 (perfect health) and, unlike generic profiles, are able to integrate morbidity and mortality. There are two types of preference-based measures: direct and multi-attribute.

Direct preference-based measures assess the preference for a health state. Direct assessments are typically designed for specific purposes and therefore allow the researcher/individual/analyst to incorporate items that are more relevant for the particular population being studied. An advantage of using the direct preference-based approach is that the patients can be asked to provide global assessments of the net effect of treatment on their HRQoL. Therefore, HRQoL responses by the patients capture their assessments of positive treatment effects and the negative side effects. The Visual Analogue Scale (VAS) is a method used for measuring preferences for health outcomes. Death may be the worst health state (equals to zero) and placed at the bottom of the scale and, perfect health (equals to 100) may be placed at the top of the scale (Santana and Feeny, 2008).

Multi-attribute preference instruments describe the health status of a subject using a multi-attribute health status classification system and using a scoring system to value health status. The EuroQol EQ-5D (Kind, 1996; Dolan, 1997; Robin and de Charro, 2001) contains five attributes (mobility, self-care, usual activities, pain or discomfort, and anxiety or depression) with three levels per attribute. Two hundred forty-three possible health states are generated by the EQ-5D system. The instrument can be translated to a quality-adjusted life year (QALY) score, which enables comparisons between different diagnoses and with the general population. Single index values for each of these health states can be obtained using scoring functions estimated with time trade off scores. Details of the algorithm to generate the index are described in detail elsewhere (Dolan, 1997; Cleemput, 2010). Applicable to a wide range of health conditions and treatments, the QALY score provides a simple descriptive profile and a single index value for health status that can be used in the clinical and economic evaluation of health care as well as in population health surveys (Cheung et al., 2009).

1.3 Objectives

✓ To determine which socio-demographic characteristics are associated with HRQoL in the general population (measured with VAS and EQ-5D), for children, adults and elderly respectively and to develop a statistical model describing the relationship between these characteristics and their HRQoL experience;

✓ To investigate whether HRQoL is more alike within households than between households.

Chapter 2

Data Description

2.1 Study design and Sample

A survey on HRQoL was conducted in the general population in Flanders (Belgium) using the standard EuroQol questionnaire with a Visual Analogue Scale. The survey was conducted in a random sample of 3118 individuals of all ages (886 children [0-12 years], 1868 adults [13-60 years] and 363 elderly [60+ years]). The sample was divided into two subsamples: 1773 (57%) participants belonged to a unique household ('sample of individuals') and for 1345 (43%) participants, the information was collected from all members of the household ('sample of households'). Sample selection was based on random digit dialing (including mobile phones), with quota for age, gender and province. For province as such, the geographical distribution of respondents was representative for Flanders. For individuals from the same household, additional quota were set.

2.2 Description of variables

Using a diary, all participants were asked about their HRQoL (VAS and EQ-5D), general socio-demographic factors such as: age, gender, if they had experienced serious disease themselves or a member of their family, province, number of domestic animals, number of parents in the family, number of persons in the household and if they filled in the diary on a normal day. Additional questions were asked to each of the three subgroups: (1) for children: mother's education; (2) for adults: smoking behaviour, profession, education level, whether the adults worked/had worked for a health care facility and (3) for elderly: frequency of alcohol consumption, frequencies by which children and grandchildren visited them, work status, whether the person had worked for a health care facility, profession, education level, smoking behaviour and experience with serious disease by taking care of someone. The height and weight was only recorded for the 1200 respondents of all respondents grouped in households. A list of all variables (short name + explanation) can be found in the Appendix (Table A.1).

.

Chapter 3

Methodology

3.1 Data management

Removing observations and correcting misspecified values:
One respondent with a negative value of HRQoL EQ5D was removed from the dataset. Although a negative value of EQ5D is possible, it was chosen not to consider it for this analysis, as it occurs rarely, especially when measuring HRQoL in the general population. Moreover, participants from three provinces from Wallonia ('Waals-Brabant', 'Luik' and 'Luxemburg') were removed, as the study focused on the Flanders provinces and the Brussels capital area. Sixteen participants with age ranging from 13 to 16 years were wrongly classified as children; one participant aged 21 years and four participants aged 60 years were wrongly classified as elderly; seven participants with age ranging from 61 to 74 years were wrongly considered as adults. All those participants were included in the correct age category.
Body mass index (BMI) was calculated based on the reported height and weight. BMI value was considered missing if height and/or weight fell far outside the normal range for a certain age group. The average BMI value by age and gender can be found elsewhere (Wilson, 2013; CDC). Specifically, for 22 children with age ranging from 3 to 6 years the BMI value was considered missing, because the heights of those children were all higher than what is considered normal. Also, one participant aged one year reported a height of 0.20 meters, which was lower than what is considered normal, and two participants with ages 38 and 40 reported a weight of 7 kg and 2 kg respectively.

Collapsing and scaling variables:
Based on exploratory analysis, variables with many categories were regrouped into fewer meaningful categories. The variables frequencies with which children and grandchildren visited the elderly were collapsed from 8 to 4 levels; the variable frequency of alcohol consumption was collapsed from 5 to 4 levels; the variables mother's education and education were collapsed from 9 to 5 levels; and the variable profession was collapsed from 15 to 4 levels. The BMI variable was scaled, i.e. was subtracted from the average BMI value for a specific age and gender. As a result, negative values represent persons who weigh less than average, and positive values represent persons weighing more than average.

3.2 Variable Selection

In machine learning and statistics, variable selection is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a variable selection technique is that the data contain many redundant or irrelevant variables (Guyon and Elisseeff, 2003).

3.2.1 Regression Tree

Morgan and Sonquist (1963) proposed a simple method for fitting trees to predict a quantitative variable. They called the method Automatic Interaction Detection. The algorithm performs stepwise splitting. It begins with a single cluster of cases and searches a candidate set of predictor variables for a way to split this cluster into two clusters. Each predictor is tested for splitting as follows: sort all the n cases on the predictor and examine all $n-1$ ways to split the cluster in two. For each possible split, compute the within-cluster sum of squares about the mean of the cluster on the dependent variable. Choose the best of the $n-1$ splits to represent the predictor's contribution. This process is repeated for every other predictor. For the actual split, choose the predictor and its cut point, which yields the smallest overall within-cluster sum of squares (Wilkinson, 1992; Hastie et al., 2001). Categorical predictors require a different approach. Since categories are unordered, all possible splits between categories must be considered. For deciding on one split of k categories into two groups, this means that $2^k - 1$ possible split must be considered. Once a split is found, its suitability is measured on the same within-cluster sum of squares as for a quantitative predictor (Wilkinson, 1992; Ritschard, 2010).

3.2.2 Random Forest

Random forest (RF) for regression is widely used in many research fields for prediction and interpretation purposes. Their popularity is rooted in several appealing characteristics, such as their ability to deal with high dimensional data, complex interactions and correlations between variables. Another important feature is that RF provides variable importance measures that can be used to identify the most important predictor variables (Hapfelmeier, et al. 2013).

The main idea of the RF is to grow many regression trees to obtain a forest of trees. The goal is to reduce the correlation between the individual trees by using bootstrapping and a randomized variable selection method, which results in reduced variance when the trees are aggregated (Melnychuk, 2013).
RF returns several measures of variable importance. The most reliable measure of variable importance is based on the decrease of classification accuracy when values of a variable in a node of a tree are permuted randomly (Breiman, 2001; Bureau et al., 2003; Remlinger, 2004). This measure is sometimes reported as such, and sometimes it is reported after scaling it, or dividing by a quantity somewhat analogous to its standard error.

3.2.3 Lasso Regression

The lasso is a shrinkage and selection method for regression models, originally applied to OLS regression. The lasso is best described as a constraint on the sum of the absolute values of the model parameters, where this sum has a specified constant as an upper bound. Compared to OLS parameter estimates, the estimates obtained using the lasso are generally more accurate and some parameters will be shrunk towards zero, allowing for better interpretation of the model and identification of those covariates most strongly associated with the outcome (Tibshirani, 1996).

The lasso is defined by

$$\hat{\beta}^{lasso} = argmin_{\beta}\Big\{\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\Big\} \tag{3.1}$$

Here λ is a complexity parameter that controls the amount of shrinkage. It is chosen such that the mean squared prediction error is minimum. The lasso solutions have the property that tends to produce some coefficients to be exactly zero. The tuning parameter may be selected by the user or calculated via numerous methods including cross-validation. Therefore, lasso is in between subset selection and ridge regression (Tibshirani, 1996; Wu and Lange, 2008).

3.2.4 Elastic net

The elastic net method overcomes the limitations of the lasso method which uses a penalty function based on $||\beta||_1 = \sum_{j=1}^{p}|\beta_j|$ (Tibshirani, 1996). Use of this penalty function has several limitations. Consider the following three scenarios.

(a) In the $p > n$ case, the lasso selects at most n variables before it saturates. This seems to be a limiting feature for a variable selection method. Moreover, the lasso is not well defined unless the bound on the L1-norm of the coefficients is smaller than a certain value.

(b) If there is a group of variables among which the pairwise correlations are very high, then the lasso tends to select only one variable from the group and does not care which one is selected.

(c) For usual $n > p$ situations, if there are high correlations between predictors, it has been empirically observed that the prediction performance of the lasso is dominated by ridge regression (Tibshirani, 1996).

Scenarios (a) and (b) make the lasso an inappropriate variable selection method in some situations. To overcome these limitations, the elastic net adds a quadratic part to the penalty $(||\beta||^2)$, which when used alone is ridge regression. The estimates from the elastic net method are defined by

$$\hat{\beta}^{ENet} = argmin_{\beta}\Big\{||y - X\beta||^2 + \lambda_2||\beta||^2 + \lambda_1||\beta||_1\Big\} \tag{3.2}$$

where $||\beta||^2 = \sum_{j=1}^{p}\beta_j^2$, $||\beta||_1 = \sum_{j=1}^{p}|\beta_j|$.

As a result, the elastic net method includes the lasso and ridge regression: in other words, each of them is a special case where $\lambda_1 = \lambda, \lambda_2 = 0$ or $\lambda_1 = 0, \lambda_2 = \lambda$.
Similar to the lasso, the elastic net simultaneously does automatic variable selection and continuous shrinkage, and it can select groups of correlated variables. The elastic net significantly improves on the lasso in terms of prediction accuracy (Efron et al., 2004).

These models were estimated in R software (version 3.0.2) using *rpart*, *randomForest*, *LARS* and *elasticnet* packages.

3.3 Beta regression

The beta distribution is a continuous probability distribution defined over the unit interval with density function

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} \tag{3.3}$$

$0 < y < 1$, where $\Gamma(.)$ is the gamma function (Ferrari and Cribari-Neto, 2004). The parameter μ denotes the expected value of y, i.e. $E(y) = \mu$. The parameter ϕ fulfills the definition of a precision parameter because for fixed μ the greater the value of ϕ, the smaller the variance of the dependent variable. More specifically,

$$Var(y) = \frac{V(\mu)}{1+\phi}$$

where $V(\mu) = \mu(1-\mu)$ denotes the "variance function".
In classical beta regression model, as in generalized linear model framework, only the mean parameter μ of the beta distribution is expressed as a function of covariates, whereas the precision parameter ϕ is treated as nuisance.
The extended beta regression model relates both parameters to covariates through distinct linear predictor (Simas et al., 2010; Smithson and Verkuilen, 2006). This model is also referred to as "double index regression model" because it contains two regression parts: one for the mean and one for the precision. Given observations on n independent beta-distributed random variables y_i $(i = 1, \cdots, n)$, the corresponding parametersμ_i and ϕ_i are linked to linear predictors η_i and ζ_i as follows

$$g_1(\mu_i) = \eta_i = x_i^T\beta$$
$$g_2(\phi_i) = \zeta_i = z_i^T\gamma$$

where x_i and z_i are p- and q-dimensional vectors of covariates observed along with y_i $(i = 1, \cdots, n)$, and $\beta = (\beta_1, \cdots, \beta_p)^T, \gamma = (\gamma_1, \cdots, \gamma_q)^T$ are the vectors of coefficients associated with means and precision respectively. The function $g_1(.)$ and $g_2(.)$ are monotonic link functions, preferably with the property of mapping the range of μ_i $(0, 1)$ and ϕ_i $(0, \infty)$, respectively to the real line. Suitable candidates for $g_1(.)$ are the logit, probit, complementary log-log, log-log and Cauchy, and for $g_2(.)$ the log function (Cribari-Neto and Zeileis, 2010; Grun et al., 2012).

The logit link

$$g_1(\mu_i) = log\frac{\mu_i}{1-\mu_i} = x_i^T\beta, \tag{3.4}$$

has the advantage that it provides a straightforward interpretation and is commonly used as the link of choice, which restrict $0 < \mu < 1$. The log link $g_2(\phi_i) = z_i^T\gamma$ it is used to ensure that ϕ is always positive (Zimprich, 2010; Hunger et al., 2012; Smithson and Verkuilen, 2006).
Typically, the coefficients β and γ are estimated by maximum likelihood (ML) and inference is based on the usual central limit theorem with its associated asymptotic tests (Grun et al., 2012). With the precision parameter ϕ being an inverse measure of dispersion, it reflects the idea that the ϕ is of interest on its own and that in many situations covariates have an effect on the variation of the dependent variable, thus involving heteroscedasticity (Smithson and Verkuilen, 2006).

The beta distribution is defined on the open unit interval only. If ones and zeros are observed, these values need to be transformed in order to fall into the open unit interval $(0, 1)$. This can be achieved by either minimally compressing the entire range of observed values, or by only transforming the boundary points to slightly smaller or greater values, respectively (Smithson and Verkuilen, 2006). Alternatively, it has been suggested to add a small amount ϵ to the lower bound, and to subtract the same amount from the upper bound (Smithson and Verkuilen, 2006; Verkuilen and Smithson, 2012).
Both methods are likely to bias the estimates towards no effect. Verkuilen and Smithson (2012) advised the use of sensitivity analysis to ensure that the estimates and inference are not affected by the choice of ϵ. The latter technique was used in this analysis and as such, bias-correction and bootstrap techniques were implemented to investigate bias in the outcome.

3.4 One inflated beta regression

Many studies in areas involve data in the form of fractions, rates or proportions that are measured continuously in the open interval (0, 1). However, frequently the data contain observations at the extremes (either zero or one). Our focus is on the case where only one of the extreme appears in the data (i.e. many ones). Having this problem, Ospina and Ferrari (2010) proposed a class of model using a mixture of two distributions: a beta distribution and a degenerate distribution in a known value c, where c equals one. Under this approach, the probability density function of the response variable y with respect to the measure generated by the mixture is given by

$$f(y;\alpha,\mu,\phi) = \begin{cases} \alpha, & \text{if y=c} \\ (1-\alpha)f(y;\mu,\phi), & \text{if y} \in \text{(0,1)} \end{cases} \tag{3.5}$$

where $f(y; \mu, \phi)$ is the beta density described in 3.3. Note that α is the probability mass at c and represents the probability of observing one $(c = 1)$. If $c = 1$, the density is called a one-inflated beta distribution (Ospina and Ferrari, 2010).

The mean of the response and its variance can be written as

$$E(y) = \alpha c + (1-\alpha)\mu$$

$$Var(y) = (1-\alpha)\frac{\mu(1-\mu)}{\phi+1} + \alpha(1-\alpha)(c-\mu)^2$$

Note that $E(y)$ is the weighted average of the mean of the degenerate distribution at c and the mean of the beta distribution (μ, ϕ) with weights α and $1-\alpha$. Also, $E(y|y \in (0,1)) = \mu$ and $Var(y|y \in (0,1)) = \mu(1-\mu)/(1+\phi)$.

A general class of one-inflated beta regression model can be defined as follows. Let $y_1, \cdots, y_n$ be independent random variables such that each y_t, for $t = 1, \cdots, n$, has probability density function given in 3.5 with parameters $\alpha = \alpha_t$, $\mu = \nu_t$, and $\phi = \phi_t$. We assume that α_t, ν_t and ϕ_t are defined as

$$h_1(\alpha_t) = \eta_{1t} = f_1(v_t, \rho)$$
$$h_2(\mu_t) = \eta_{2t} = f_2(x_t, \beta)$$
$$h_3(\phi_t) = \eta_{3t} = f_3(z_t, \gamma)$$

where $\rho = (\rho_1, \cdots, \rho_p)^T$, $\beta = (\beta_1, \cdots, \beta_k)^T$ and $\gamma = (\gamma_1, \cdots, \gamma_m)^T$ are vectors of unknown regression parameters; $(p + k + m < n)$, $\eta_1 = (\eta_{11}, \cdots, \eta_{1n})^T$, $\eta_2 = (\eta_{21}, \cdots, \eta_{2n})^T$ and $\eta_3 = (\eta_{31}, \cdots, \eta_{3n})^T$ are predictors vectors; and $f_1(.,.)$, $f_2(.,.)$ and $f_3(.,.)$ are linear or nonlinear twice continuously differentiable functions. According to Ospina and Ferrari (2010), the link functions $h_1 : (0,1) \to R$, $h_2 : (0,1) \to R$ and $h_3 : (0,\infty) \to R$ can be assumed. For μ and α one may choose logit, probit, complementary log-log link functions, and for ϕ is $h_3(\phi) = log\phi$ (log link).

Beta regression and one-inflated beta regression were estimated in R 3.0.2 using *betareg* and *gamlss* packages.

3.5 Model selection

Linear predictors for both HRQoL outcomes were implemented using an extension to polynomials in order to allow for more functional forms of the responses. Likewise, using fractional polynomials were preferable under a certain set of the powers, $S = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$, because they provide flexible curvilinear shapes. For comparative measures of model fit under each response we based on Akaike Information Criteria (AIC) and the likelihood ratio test (Agresti, 2002) for comparing nested models for the need for interactions and as well as inclusion of covariates in the dispersion sub models.

3.6 Models for a Single Beta GLMM Response

In longitudinal analyses or in the case that subjects are clustered within sampling units or geographical entities, measurement within the same person or unit are typically correlated, violating the assumption of independent observations in regression models (Molenberghs and Verbeke, 2005; Fitzmaurice et al., 2009). One possibility to account for these dependencies is to add random cluster or subject effects into the linear predictor.

Without loss of generalizability, consider the case of longitudinal designs where $j = 1, \cdots, n_i$ observations are nested within $i = 1, \cdots, N$ subjects. Let b_i denote a vector of subject-specific random effects for individual i. Adding random effects to the beta regression model described in 3.4 yields the beta GLMM (Zimprich, 2010; Verkuilen and Smithson, 2012 and Hunger et al., 2012) given by

$$\begin{aligned} g_1(\mu_i) &= log\frac{\mu_i}{1-\mu_i} = x_{ij}^T\beta + w_{ij}^T b_i \\ g_2(\phi_i) &= z_i^T\gamma, \end{aligned} \tag{3.6}$$

with $b_i \sim N(0, G)$. In this case, w_{ij}^T is a vector of covariates, and G denotes the positive definite covariance matrix of the random effects. Note that although the assumption of normality for the random effects is common and statistically convenient, other distribution assumptions are possible in principle (Hunger et al., 2012). In a longitudinal design, b_i typically is a scalar (for random intercept only models) or a bivariate vector (for models with random intercept and random slope). In the first case, $w_{ij} = 1$, while in the second case, $w_{ij}^T = (1, t_{ij})$, where t_{ij} is the time of measurement j for subject i. Models with random slopes allow the linear effect of time to vary across subjects.

Model parameters are estimated by maximizing the marginal likelihood, which is obtained by integrating out the unobserved random effects b_i from the likelihood function (Verkuilen and Smithson, 2012). In the beta GLMM, the regression parameters only have a subject-specific interpretation and no longer describe the effect of the respective variable on the population in general (Molenberghs and Verbeke, 2005; Fitzmaurice et al., 2009).

3.7 Models for Joint Beta GLMM Responses

Difficulties in analyzing clustered or repeated measures arise because of correlations usually present between observations on the same subject or within the same cluster. In the case of multiple outcomes two types of correlations must be taken into account: correlations between measurements on different variables and correlations between measurements on the same variable within cluster or subject (Gueorguieva, 2001).

In a joint-modeling approach using mixed models, random-effects are assumed for each response process and by imposing a joint multivariate distribution on the random effects, the different processes are associated (Fieuws and Verbeke, 2004). The approach allows to joint models for responses of the same response type as well as models for responses of different types. The approach has been used in a non-longitudinal setting to validate surrogate endpoints in meta-analyses (Buyse et al., 2000) or to model multivariate clustered data (Thum, 1997). Also, joint models are popular owing to the fact that they ensure unbiased statistical inferences in a variety of settings (Iddi and Molenberghs, 2012).

In the context of jointly modeling, let us consider a bivariate response. Denote the response vector for the ith subject by $\mathbf{y}_i = (\mathbf{y}_{i1}^T, \mathbf{y}_{i2}^T)^T$, where $\mathbf{y}_{i1} = (y_{i11}, \cdots, y_{i1n_{i1}})^T$ and $\mathbf{y}_{i2} = (y_{i21}, \cdots, y_{i2n_{i2}})^T$ are the repeated measurements on the first and second variable. We assume that $\mathbf{y}_{i1j}$, $j = 1, \cdots, n_{i1}$, are conditionally independent given $\mathbf{b}_{i1}$ with density $f_1(.)$ in the exponential family. Analogously, $\mathbf{y}_{i2j}$, $j = 1, \cdots, n_{i2}$, are conditionally independent given $\mathbf{b}_{i2}$ with density $f_2(.)$ in the exponential family. Also $\mathbf{y}_{i1}$ and $\mathbf{y}_{i2}$ are conditionally independent given $\mathbf{b}_i = (\mathbf{b}_{i1}^T, \mathbf{b}_{i2}^T)^T$ and the responses on different subjects are independent. Let $g_1(.)$ and $g_2(.)$ be appropriate link functions for f_1 and f_2. Denote the conditional means of $\mathbf{y}_{i1j}$ and $\mathbf{y}_{i2j}$ by μ_{i1j} and μ_{i2j}, respectively.
Let $\mu_{i1j} = (\mu_{i11}, \cdots, \mu_{i1n_{i1}})^T$ and $\mu_{i2} = (\mu_{i21}, \cdots, \mu_{i2n_{i2}})^T$. At stage one of the linear mixed model specifications we assume

$$\begin{aligned} g_1(\mu_{i1}) &= \mathbf{X}_{i1}\beta_{\mathbf{1}} + \mathbf{W}_{i1}b_{i1} \\ g_2(\mu_{i2}) &= \mathbf{X}_{i2}\beta_{\mathbf{2}} + \mathbf{W}_{i2}b_{i2} \\ g_1(\phi_{i1}) &= z_i^T\gamma \\ g_2(\phi_{i2}) &= z_i^T\gamma \end{aligned} \tag{3.7}$$

where β_1 and β_2 are $(p_1 \times 1)$- and $(p_2 \times 1)$-dimensional unknown parameters vectors, $\mathbf{X}_{i1}$ and $\mathbf{X}_{i2}$ are $(n_{i1} \times p_1)$- and $(n_{i2} \times p_2)$-dimensional design matrices for the fixed effects, $\mathbf{W}_{i1}$ and $\mathbf{W}_{i2}$ are $n_{i1} \times q_1$ and $n_{i2} \times q_2$ design matrices for the random effects and g_1 and g_2 are applied componentwise to μ_{i1}, μ_{i2}, ϕ_{i1} and ϕ_{i2}. At stage two

$$\mathbf{b}_i = \begin{pmatrix} b_{i1} \\ b_{i2} \end{pmatrix} \sim i.i.d\, \mathbf{MVN}(0, \Sigma) = \mathbf{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{bmatrix} \right)$$

where Σ, Σ_{11} and Σ_{22} are in general unknown positive-definite matrices.
When $\Sigma_{12} = 0$ then the above model is equivalent to two separate beta GLMM's for the two response variables. Advantages of joint over separate fitting include better control over the type I error rates in multiple tests, possible gains in efficiency in the parameter estimates (Gueorguieva, 2001; Fieuws and Verbeke, 2004).

All mixed beta regression models were estimated using $SAS^{®}$ 9.3 procedure NLMIXED (SAS Institute, 2012) by maximum likelihood estimation. A particularly useful resource on how NLMIXED is used in fitting non-linear models is Molenberghs and Verbeke (2005).

Chapter 4

Results

This section presents the descriptive analyses and the application of the models discussed in section 3 for the analysis of health related quality of life in respondents from the sample of individuals and the sample of households. Explanations for each covariate can be found in Appendix (Table A.1).

4.1 Descriptive statistics

In both samples of HRQoL, 63 observations in EQ5D response and 123 observations in VAS response were deleted due to missing values in the response variable. This reduced the final sample size from 3117 to 3054 in EQ5D and 2994 in VAS responses. The average age was 32.8 years (SD = 22.5), 52.2% of the participants were female. 28.42% of the participants were children, 59.93% adults and 11.65% elderly. 22% of the participants were from East Flanders and 27% from Antwerp. Only 5% of the participants were from Brussels capital area. 63.3% had one or more domestic animals in their family, and 73% of the participants filled in the diary on a normal day. Around 36% of the participants had four members in the household and two participants had reported 9 and 11 members in the household respectively. The majority of the participants (82%), are living with the husband and wife in a family. Around 13% of the participants had experienced serious disease themselves, whereas 43.7% had experienced serious disease with a member in the family.

For the sample of individuals, three groups of categories were formed (child, adults and elderly). In the child category the mean age was 5.2 years (SD=3.5), and 47% of participants were female. Around 64% of the mothers who participated in the study had higher (not-) university/postgraduate level, and less than 2% with none or primary level of education. In the adult and elderly category the mean age was 38.4 years (SD=12.3) and 74.1 years (SD=9.9), with 57% and 53% of the participants were female respectively. More than 60% of the participants in those groups reported they have never smoked, have never worked in a health care facility and have not experienced serious disease by taking care of someone. Moreover, 50% of the participants had a white-collar job. The distribution over the different education levels was similar as for the child group. The socio-demographic characteristics of the sample of individuals are summarized in Table A.1 (see appendix A.1).

Relationships between the HRQoL outcomes and these characteristics are shown in Appendix A.2 (boxplots). Differences between boys and girls and those children having experienced serious disease in family were observed in VAS outcome. The EQ5D and VAS outcome may be different for children having experienced severe disease, and children being sick on the day the diary were filled in. Similar results were observed for the adults and the elderly (Appendix A.2).

For the sample of households, the mean age was 23.6 years (SD=16.3) with a median of 18 years, and 52.2% of the participants were female. From the 1200 respondents of whom the height and weight were recorded, the average BMI was 20.8 kg/m2 (SD= 5.6). After rescaling the BMI and taking into account the age and gender, 37.1% had a BMI below the average. Relationships between the HRQoL outcomes and the background characteristics in sample of households are shown in Appendix A.2 (boxplots). Only variable 'normal day' seems to have a (clear) impact on EQ5D and VAS.
The distributions of EQ5D and VAS are negatively skewed: most participants reported a very high HRQoL score (Figure 4.1). From Figure 4.1 is it also clear that not only the mean of the HRQoL index scores but also the shape of its distribution changes across age groups. As age increases, the distribution gets broader and the skewness reduces.

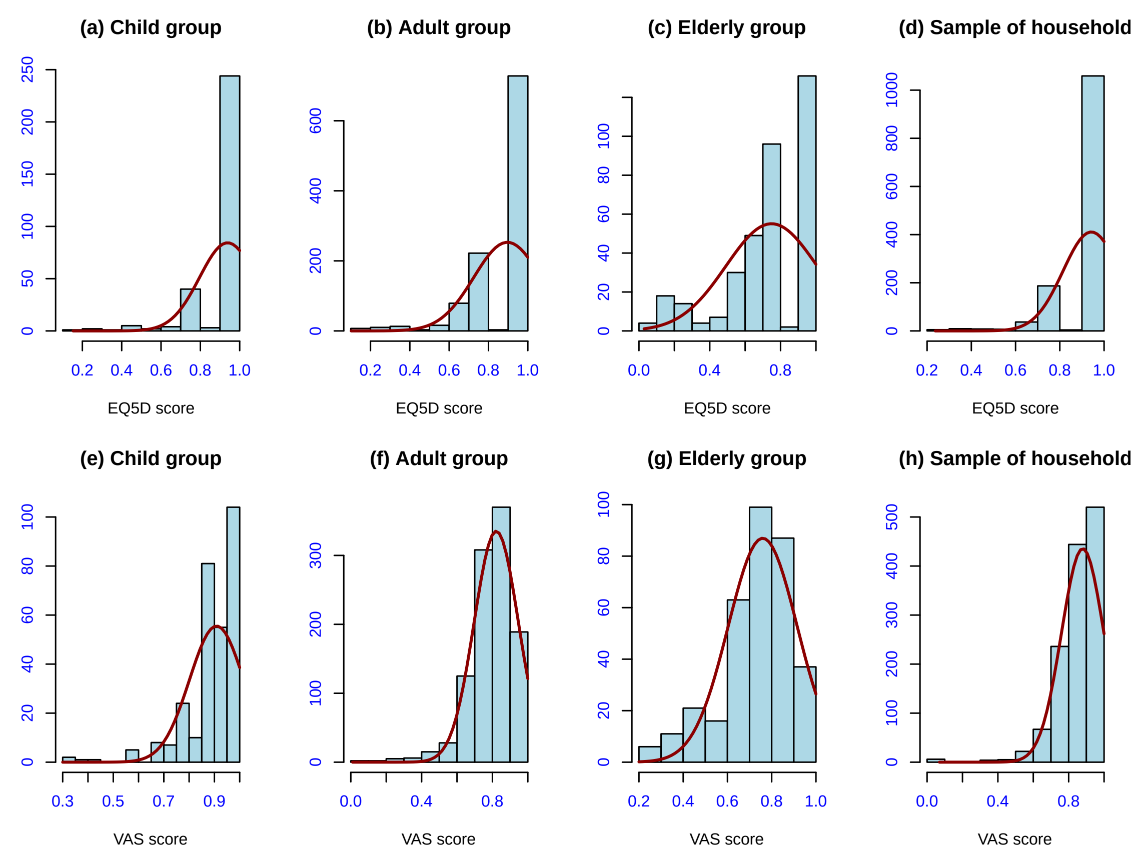

Figure 4.1: *The distributions of the EQ5D and VAS index scores for the different age categories, and for the sample of households (which includes all ages).*

4.2 Results for the sample of individuals

4.2.1 Variable selection

The data applied in this section come from the sample of individuals of HRQoL. The predictors were based on each age category group as described in section 2.2. Regression tree, RF, the lasso and the elastic net were all applied to these data and the corresponding graphs are displayed in appendix B.1. Table 4.1 below gives the general overview of the most important variables selected based on each method.

Table 4.1: *Variables selected based on regression tree, random forest, lasso regression and elastic net*

Group category	Outcome	Method	Variables selected
Child	EQ5D	Regression tree	age, illnessy, province, and mumEducation
		RF	age, illnessy, province, mumEducation and peoplehouse
		lasso	age, illnessy, province, normalday, illnessf and mumEducation
		elastic net	age, illnessy, province, normalday, illnessf and mumEducation
	VAS	Regression tree	illnessy, normalday, province and peoplehouse
		RF	age, illnessy, normalday, province and peoplehouse
		lasso	age, illnessy, normalday, peoplehouse mumEducation and parent
		elastic net	age, illnessy, normalday, mumEducation and peoplehouse
Adult	EQ5D	Regression tree	age, illnessy, profession and normalday
		RF	age, illnessy, province, education and profession
		lasso	age, illnessy, normalday, illnessf, profession, education, animal, province and peoplehouse
		elastic net	age, illnessy, normalday, illnessf, education, profession, peoplehouse, smokestatus and animal
	VAS	Regression tree	age, illnessy, profession, normalday and education
		RF	age, illnessy, profession, normalday, education, province and peoplehouse
		lasso	illnessy, normalday, illnessf and animal
		elastic net	illnessy, normalday, illnessf and animal
Elderly	EQ5D	Regression tree	age, illnessy, education, profession, freq1 and freq3
		RF	age, illnessy, education, profession, parent, province, freq1, freq2 and freq3
		lasso	age, illnessy, education, profession, smokestatus, illnessc, province, workedinHCare, freq1, freq2 and freq3
		elastic net	age, illnessy, education, profession, smokestatus, freq1 and freq3
	VAS	Regression tree	age, illnessy, education, profession, province, smokestatus, province, freq1, freq2 and freq3
		RF	age, education, profession, province, freq2 and freq3
		lasso	age, illnessy, education, profession, illnessf, workedinHCare, smokestatus, freq1 and freq3
		elastic net	age, illnessy, education, profession, illnessf, workedinHCare, freq1 and freq3

For children, all four methods show that age, illnessy, province and mother education are important for determining EQ5D. Random forest additionally selected peoplehouse. Lasso and elastic net selected additionally normalday and illnessf. Similar variables were found to be important to determine VAS, where the lasso selected additionally number of parents in a family. For adults, more or less the same set of variables as in the child group was selected for both EQ5D and VAS, with additional inclusion of profession in all methods. Lasso and elastic net also selected animal as important in this age group. For elderly, all methods show that age, illnessy, education, profession, freq1 and freq3 were important for determining both EQ5D and VAS. In all age groups, the variables age, illnessy and education were important based on the four different methods applied, and we also observed that normalday is an important variable for the children and adults group.

The variables that will be included initially as covariates when building the statistical models for EQ5D and VAS (see further) are presented in Table 4.2, and are based on the results of the initial selection methods (Table 4.1).

Table 4.2: *Variables selected based on combining the results of four variable-selection methods.*

Group category	**Outcome**	**Variables selected**
Child	EQ5D	age, illnessy, nornalday, illnessf, province, peoplehouse and mumEducation
	VAS	age, illnessy, normalday, peoplehouse and mumEducation
Adult	EQ5D	age, illnessy, normalday, profession, education, animal, illnessf and peoplehouse
	VAS	age, illnessy, normalday, illnessf, profession and animal
Elderly	EQ5D	age, illnessy, freq1, freq3, freq2, education, profession, smokestatus
	VAS	age, illnessy, freq1, freq3, education, profession, smokestatus and workedinhCare

Additionally, it was decided to include gender as a covariate. Gender was not considered as an important covariate by none of techniques applied for variable selection. Bisegger et al. (2005), studied gender and age differences in different aspects of HRQoL of children and adolescents, where they applied a newly developed HRQoL questionnaire, the "Kidscreen 52" in seven European countries. They found that children have higher HRQoL than adolescents in many aspects. With increasing age, HRQoL is frequently worse for females than for males. Thus, based on literature it was decided to use gender as a covariate in this analysis.

4.2.2 Statistical analysis

1. Child group

The analysis in health related quality of life was applied for different modeling techniques described in section 3.3 and 3.4. As has been mentioned in section 4.1, visual inspection of the distributions of EQ5D and VAS scores in child category suggest that one inflated beta distribution may be a suitable model to be applied for this age group in both responses.

Modeling One inflated beta regression in EQ5D

We considered one-inflated beta regression and fitted different possible models based on the extended polynomials and fractional polynomials. The results for their comparisons in terms of AIC and likelihood ratio tests are presented in Tables C.1 and C.2 (see appendix C.1). Tests for interactions indicated the need for interactions and/or dispersion sub-model was significant in all models. The smaller the AIC value, the better the model. Therefore, the third order polynomial model was selected and based on the likelihood ratio the inclusion of the variable dispersion model and interactions in covariates were supported. Predictions based on best models under each link function are shown in Figure C.1, with non-linear regression that could be considered using cubic splines, which resulted to a good fit to the data.

Only the clog-log link function in polynomial model was not fitting well the data, even though polynomial models were the best in terms of AIC. The third order polynomial model with logit link function was taken as a final model for ease of interpretation. The non-significant parameters were systematically eliminated from the model by backward selection. The parameter estimates with the corresponding standard errors and significance tests for the final model are summarized in Table 4.3.

Table 4.3: *Parameter estimates and standard error for the mean and dispersion sub-model parameters based on third order logit polynomial.*

Parameter	Estimates	Std. error	p-value	Estimates	Std. error	p-value
	location sub-model			dispersion sub-model		
Intercept	0.6651	0.1790	0.0002	1.5469	0.4852	0.0016
age	0.0489	0.0224	0.0299	0.2866	0.0750	0.0002
Female	-0.0948	0.0139	<0.0001	0.4584	0.4696	0.3299
No because sick	-0.9769	0.1877	<0.0001	-12.6673	1.4693	<0.0001
No because other reason	0.9514	0.2068	<0.0001	-2.0714	1.1760	0.0794
Illnessy: Yes	-0.1759	0.0533	0.0011	-0.9128	0.7979	0.2537
age*No because sick	0.1886	0.0224	<0.0001	4.3069	0.2776	<0.0001
age*No because other reason	-0.0993	0.0240	<0.0001	0.6391	0.1491	<0.0001

The location sub-model models the average EQ5D score for children not in perfect health. It is noteworthy that all the main effects in the location sub-model were significant. If not in perfect health (EQ5D score lower than 1), girls and children who had experienced serious disease had significantly lower EQ5D scores.

The presence of a significant interaction indicates that the effect of one predictor variable on the response variable is different at different values of the other predictor variables, i.e. the effect of age on health scores is different for values of 'normalday'. For children not having a normal day because of being sick, EQ5D score increases by age than for children having a normal day (Table 4.3).

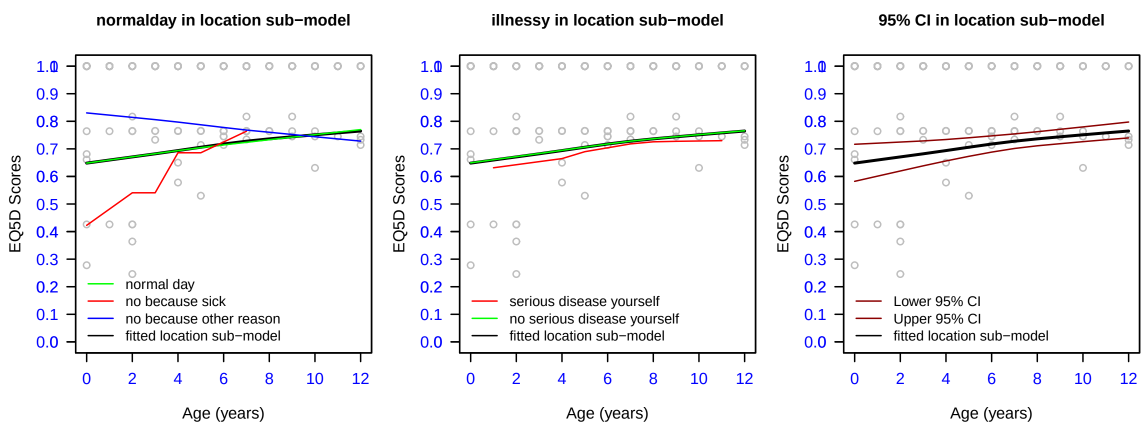

Figure 4.2: *EQ5D predictions by normalday, illnessy and 95% prediction confidence interval in location sub-model.*

Figure 4.2 above shows the fitted location sub-model. A difference in health score by age is noted for different values of 'normal day'. Children who reported not a normal day because of other reasons, have high scores at an earlier age. For children who referred not normal because of being sick, health scores increased strongly from zero years up to seven years, while for those who reported not a normal day because of other reasons there is a decrease in health as age increases. For those children who experienced serious disease their health scores remained below the average fitted, for all ages. The confidence intervals are wider for children below 3 years and narrower in older ages.

The alpha sub-model (Table 4.4) models the probability that children are in perfect health (EQ5D=1).

Table 4.4: *Parameter estimates and standard error for the alpha sub-model based on third order logit polynomial.*

Parameter	Estimates	Std. error	p-value
modelling the probability at one			
Intercept	1.2165	0.3052	0.0001
age	0.0421	0.0514	0.4126
No because of sick	-0.0226	1.5244	0.9882
No because other reason	2.4797	0.9979	0.0136
age*No because sick	-0.3197	0.3770	0.3972
age*No because other reason	-0.3066	0.1277	0.0171

Figure 4.3 below shows the fitted alpha sub-model. The age main effect was not significant, but the interactions indicate that the effect of age on the probability to be in perfect health is different for children for whom it was not a normal day because of being sick or because of another reason. In both categories, there is a decrease in the probability to be in perfect health with a steep decrease for those who reported not normal because of sick. The confidence intervals are wider at earlier age and older ages.

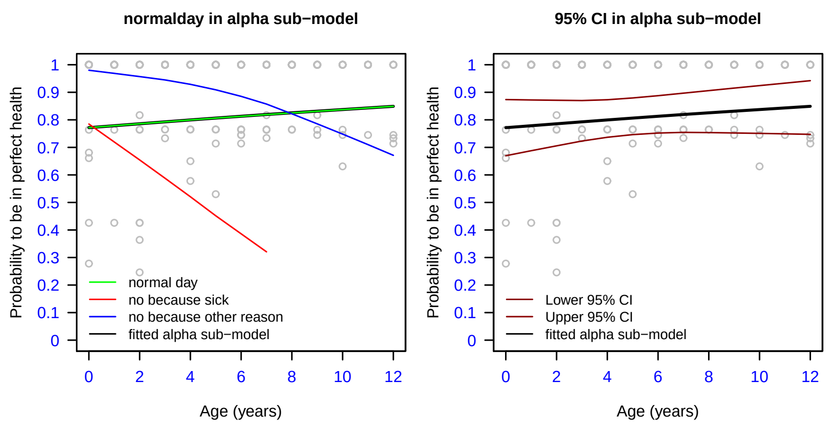

Figure 4.3: *EQ5D predictions by normalday, illnessy and 95% prediction confidence interval in alpha sub-model.*

Modeling One inflated beta regression in VAS

We considered one-inflated beta regression and fitted different possible models based on the extended polynomials and fractional polynomials. The results for their comparisons in terms of AIC and likelihood ratio test are presented in Tables C.3 and C.4 (see appendix C.2). Tests for interactions indicated the need for interactions and/or dispersion sub-model was significant in all polynomials models. Tests for fractional polynomials degree two had non-significant p-values indicating that interactions may not be useful (p-values=0.1930). Therefore, second order polynomial model was selected and based on the likelihood ratio the inclusion of variable dispersion model and interactions in covariates were supported. Predictions based on best models under each link function are shown in Figure C.2, with non-linear regression that could be considered using cubic splines, which resulted to a good fit to the data.
However, polynomial models were the best in terms of AIC. The second order polynomial model with logit link function was taken as a final model for ease of interpretation. The non-significant parameters

were systematically eliminated from the model by backward selection. The parameter estimates with the corresponding standard errors and significance tests for the final model are summarized in Table 4.5.

Table 4.5: *Parameter estimates and standard error for the mean and dispersion sub-model parameters based on third order logit polynomial.*

Parameter	Estimates	Std. error	p-value	Estimates	Std. error	p-value
	location sub-model			dispersion sub-model		
Intercept	1.6897	0.2091	<0.0001	2.1110	0.3938	<0.0001
age	0.0165	0.0151	0.2774	0.0695	0.0315	0.0284
age2	0.0010	0.0001	<0.0001	-	-	-
peoplehouse	0.0903	0.0470	0.0557	0.0778	0.0845	0.3582
No because sick	-1.5405	0.2099	<0.0001	-0.1927	0.5025	0.7017
No because other reason	0.0377	0.1151	0.7435	0.2359	0.2387	0.3241
Illnessy: Yes	-0.6574	0.3024	0.0306	-21.1749	1.5260	<0.0001
Illnessy: Yes*No because sick	0.7410	0.3680	0.0451	8.8435	0.9235	<0.0001
Illnessy: Yes*No because other reason	-0.8202	0.7290	0.2615	-2.7937	1.1129	0.0127

The location sub-model models the average VAS score for children not in perfect health. It is evident that age was not significant on the effect of health scores, but the higher order of age was highly significant with positive effect. For participants not in perfect health (VAS score lower than 1), the VAS score was estimated to increase (borderline not significantly) with the number of persons in the household. The effect of children who had experienced serious disease on health scores is different for values of 'normalday'. For children not having a normal day because of being sick, VAS score increases by age than for children not having experienced serious disease before (Table 4.5 and Figure 4.4). Narrow confidence intervals were observed in all ages.

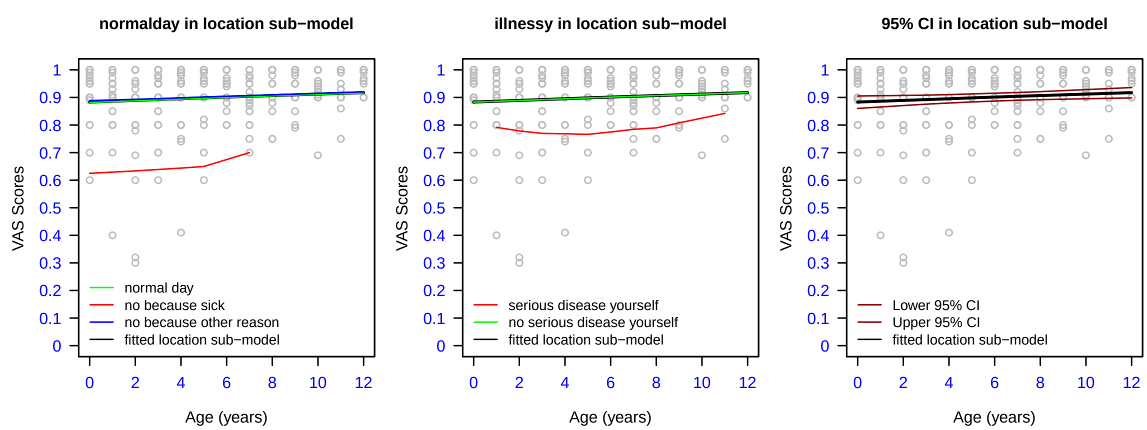

Figure 4.4: *VAS predictions by normalday, illnessy and 95% confidence interval in location sub-model.*

The alpha sub-model (Table 4.6) models the probability that children are in perfect health (VAS=1). The age was not significant, even with inclusion of higher order term. The effect of the number of persons in the household on the probability to be in perfect health is different for girls with an increase in the probability to be in perfect health.

Table 4.6: *Parameter estimates and standard error for the alpha sub-model based on logit polynomial order two.*

Parameter	**Estimates**	**Std. error**	**p-value**
modelling the probability at one			
Intercept	-0.6964	0.7861	0.3765
age	0.0744	0.0407	0.0684
peoplehouse	-0.2085	0.1889	0.2706
Female	-2.6687	1.2558	0.0345
peoplehouse*Female	0.6151	0.2911	0.0355

2. Adult group

From section 4.1, visual inspection of the distributions of EQ5D scores in adult category, the plot clearly suggest that one inflated beta distribution may be a suitable model to be applied for this age group.

Modeling One inflated beta regression in EQ5D

One-inflated beta regression was considered and fitted different possible models based on the extended polynomials and fractional polynomials. The results for their comparisons in terms of AIC and likelihood ratio test are presented in Tables C.5 and C.6 (see appendix C.3). Tests for interactions indicated the need for interactions and/or dispersion sub-model was significant in all models. Therefore, third order polynomial model was selected and based on the likelihood ratio the inclusion of the variable dispersion model and interactions in covariates were supported. Predictions based on best models under each link function are shown in Figure C.3, with non-linear regression that could be considered using cubic splines, which resulted to equally a good fit to the data.
Therefore, polynomials model were the best in terms of AIC. The third order polynomial model with logit link function was taken as a final model for ease of interpretation. The non-significant parameters were systematically eliminated from the model by backward selection. The parameter estimates with the corresponding standard errors and significance tests for the final model are summarized in Table 4.7.

Table 4.7: *Parameter estimates and standard error for the mean and dispersion sub-model based on third order logit polynomial.*

Parameter	**Estimates**	**Std. error**	**p-value**	**Estimates**	**Std. error**	**p-value**
	location sub-model			dispersion sub-model		
Intercept	1.0581	0.1088	<0.0001	3.6991	0.4206	<0.0001
age	-0.0017	0.0026	0.5071	0.0013	0.0094	0.8865
Illnessy: Yes	-0.4444	0.0931	<0.0001	-1.3921	0.2040	<0.0001
education: higher technical/secondary	-0.0727	0.0615	0.2375	-0.3905	0.2072	0.0598
education: Lower technical/secondary	-0.1575	0.0956	0.0998	-0.6802	0.2986	0.0230
education: None/Primary	-0.4810	0.2141	0.0249	-1.9440	0.3609	<0.0001
education: Vocational	-0.2772	0.1164	0.0175	-1.3545	0.2572	<0.0001

The location sub-model models the average EQ5D score for adult not in perfect health. The age was not significant on the effect of EQ5D score. If not in perfect health (EQ5D score lower than 1), adults who had experienced serious disease and adults who had primary and vocational level of education had significantly lower EQ5D scores (Table 4.7).

Figure 4.5 shows the fitted location sub-model. For adults who experienced serious disease their health scores remain below the average fitted in all ages. For adult with high education the EQ5D scores remained high in all ages, while for those with primary education level, their health scores were below the average in all ages. The confidence intervals are wider in younger age and in older ages.

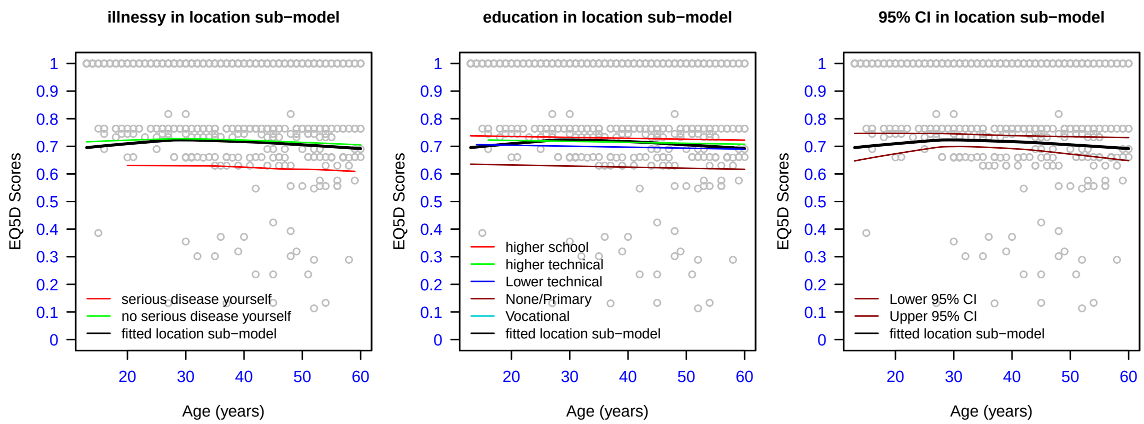

Figure 4.5: *EQ5D predictions by illnessy, education and 95% confidence interval in location sub-model.*

The alpha sub-model (Table 4.8) models the probability that adult are in perfect health (EQ5D=1). It is noteworthy that all the main effects in the alpha sub-model were significant. For an additional year in age and for adults who had experienced serious disease themselves or with a member in family, the probability to be in perfect health was significantly lower.

Table 4.8: *Parameter estimates and standard error for the alpha sub-model based on third order logit polynomial.*

Parameter	**Estimates**	**Std. error**	**p-value**
modelling the probability at one			
Intercept	2.4458	0.2755	<0.0001
age	-0.0296	0.0066	<0.0001
Illnessy: Yes	-1.1501	0.2193	<0.0001
Illnessf: Yes	-0.4965	0.1600	0.0020

Figure 4.6 below show the fitted alpha sub-model. Difference in health scores is noted between adults who experienced serious disease and those who experienced serious disease with a member in family. In both categories, there is a decrease in the probability to be in perfect health with a steep decrease for those who experienced with serious disease themselves. The confidence interval is narrow at earlier age and wider in older ages.

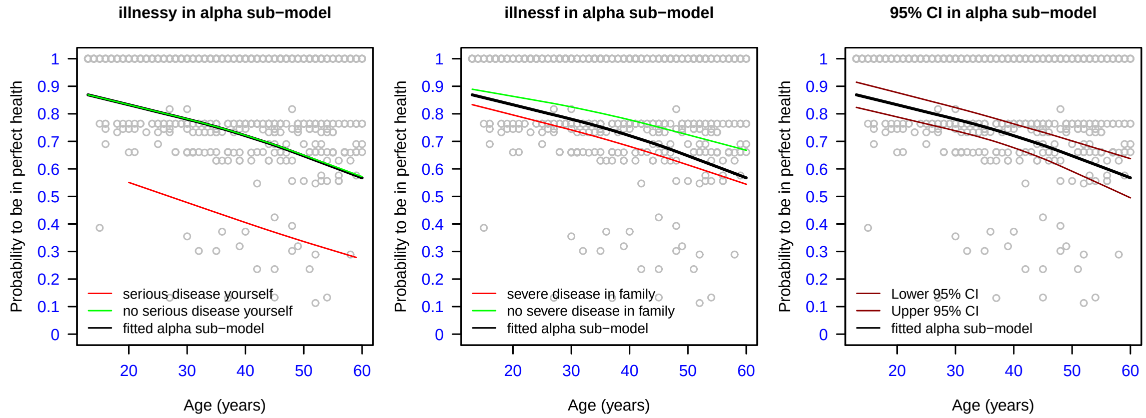

Figure 4.6: *EQ5D predictions by illnessy, illnessf and 95% confidence interval in alpha sub-model.*

Modeling beta regression in VAS

We considered beta regression and fitted different possible models based on the extended polynomials and fractional polynomials. The results for their comparisons in terms of AIC and likelihood ratio test are presented in Tables C.7 and C.8 (see appendix C.4). Tests for interactions indicated the need for interactions and/or dispersion sub-model was significant in all models. Only test for constant dispersion without any interaction in covariates was not significant in all polynomials and fractional polynomials. Therefore, first order polynomial model was selected and based on the likelihood ratio the inclusion of variable dispersion model and interactions in covariates were supported. Predictions based on best models under each link function are shown in Figure C.4, with non-linear regression that could be considered using cubic splines, which resulted to a good fit to the data.

Polynomials model was the best in terms of AIC. The first order polynomial model with logit link function was taken as a final model for ease of interpretation. The non-significant parameters were systematically eliminated from the model by backward selection. The parameter estimates with the corresponding standard errors and significance tests for the final model are summarized in Table 4.9.

Table 4.9: *Parameter estimates and standard error for the mean and dispersion sub-model based on first order logit polynomial.*

Parameter	Estimates	Std. error	p-value	Estimates	Std. error	p-value
	location sub-model			dispersion sub-model		
Intercept	1.8890	0.1002	<0.0001	1.9474	0.0664	<0.0001
age	-0.0061	0.0024	0.0109	-	-	-
Illnessy: Yes	-0.6631	0.0918	<0.0001	-0.1498	0.1371	0.2744
normalday: No because sick	-0.8998	0.1549	<0.0001	0.0164	0.2732	0.9521
normalday: No because other reason	-0.0173	0.0835	0.8359	-0.3167	0.1129	0.0050
animal: Yes	0.1488	0.0679	0.0283	0.0192	0.0979	0.8443

The location sub-model models the average VAS scores for adult not in perfect health. It is noteworthy that all the main effects in the location sub-model were significant. If not in perfect health, with additional years of age and adults who had experienced serious disease and not having a normal day because of being sick had significantly lower VAS scores. For adults having a one or more domestic animals, VAS score increases (Table 4.9).

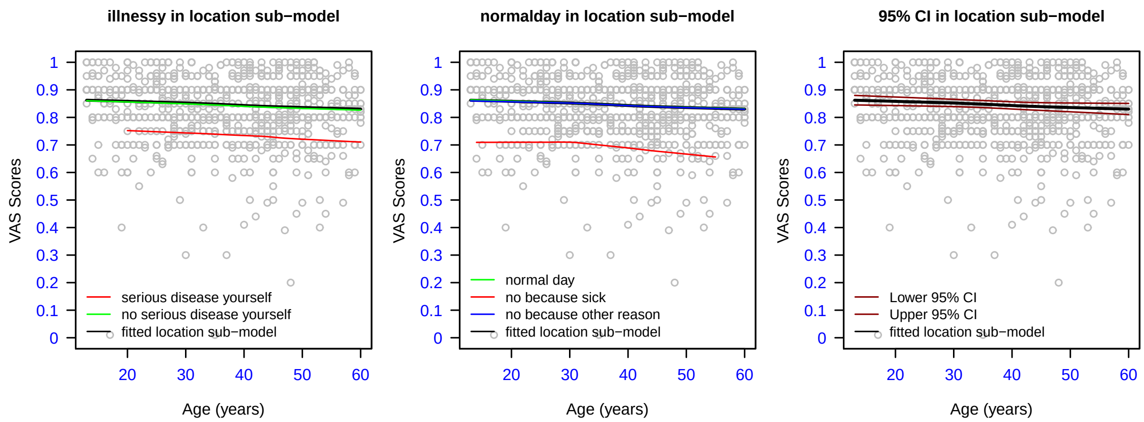

Figure 4.7: *VAS predictions by illnessy, normalday and 95% confidence interval in location sub-model.*

Figure 4.7 shows the fitted location sub-model. A difference in health scores is noted between adults who reported experience of serious disease and those who reported not normal day because of being sick with the corresponding categories for each level, where their health scores remain below the average fitted with a slight decrease. Narrow confidence intervals were observed in all ages.

To investigate whether the results may have been affected by severe bias in the ML estimator, the resulting coefficients estimates and standard errors of bias-corrected and bootstrap method based on 2000 samples were performed as shown in Table C.9 (see appendix C.4). The obtained estimates were similar to the proposed estimates in the model, meaning that, the use of small values of ϵ to move observations away from the boundary points did not appreciable affect parameter estimates.

3. Elderly group

From section 4.1, visual inspection of the distributions of EQ5D scores in elderly category, the plot clearly suggest that one inflated beta distribution may be a suitable model to be applied for this age group.

Modeling one inflated beta regression in EQ5D

We considered one-inflated beta regression and fitted different possible models based on the extended polynomials and fractional polynomials. The results for their comparisons in terms of AIC and likelihood ratio test are presented in Tables C.10 and C.11 (see appendix C.5). Tests for interactions indicated the need for interactions and/or dispersion sub-model was significant in all models. Therefore, the first order polynomial model was selected and based on the likelihood ratio the inclusion of the variable dispersion model and interactions in covariates were supported. Predictions based on best models under each link function are shown in Figure C.5, with non-linear regression that could be considered using cubic splines, which resulted to a good fit to the data.

However, polynomials model was the best in terms of AIC. The first order polynomial model with logit link function was taken as a final model for ease of interpretation. The non-significant parameters were systematically eliminated from the model by backward selection. The parameter estimates with the corresponding standard errors and significance tests for the final model are summarized in Table 4.10.

Table 4.10: *Parameter estimates and standard error for the mean and dispersion sub-model based on first order logit polynomial.*

Parameter	Estimates	Std. error	p-value	Estimates	Std. error	p-value
	location sub-model			dispersion sub-model		
Intercept	2.8949	0.4374	<0.0001	8.7438	1.0057	<0.0001
age	-0.0362	0.0060	<0.0001	-0.1074	0.0144	<0.0001
Illnessy: Yes	-0.2535	0.0905	0.0055	-0.4528	0.2219	0.0424
Ex-smoker	0.3976	0.2003	0.0483	1.8655	0.3544	<0.0001
Non-smoker	0.6185	0.1975	0.0020	2.3029	0.3712	<0.0001

The location sub-model models the average EQ5D scores for elderly not in perfect health. It is remarkable that all the main effects in the location sub-model were significant. If not in perfect health, with additional years of age and elderly who had experienced serious disease had significantly lower EQ5D scores. For elderly who had a history of smoking (they had quit smoking) and for those not smoking, EQ5D is higher than for actively smoking elderly (Table 4.10).

Figure 4.8 below shows the fitted location sub-model. For those elderly who experienced serious disease their health scores remain below the average fitted in all ages. Wider confidence intervals were observed from the age of 85 onwards.

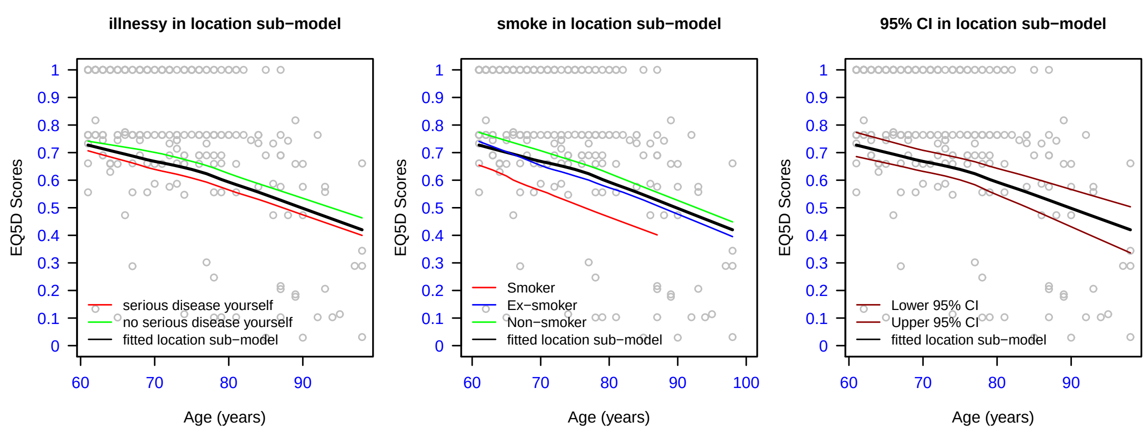

Figure 4.8: *EQ5D predictions by illnessy, smoke status and 95% prediction confidence interval in location sub-model.*

The alpha sub-model (Table 4.11) models the probability that elderly are in perfect health. Age, illnessy and education level were all significant. The probability to be in perfect health decreases significantly with age. For those who had experienced serious disease had significantly lower probability to have an EQ5D score of 1. The level of education had an impact on the probability to have an EQ5D score of 1.

Table 4.11: *Parameter estimates and standard error for the alpha sub-model based on first order logit polynomial.*

Parameter	**Estimates**	**Std. error**	**p-value**
modelling the probability at one			
Intercept	6.7817	1.4068	<0.0001
age	-0.0836	0.0195	<0.0001
Illnessy: Yes	-0.8100	0.3117	0.0099
education: higher technical/secondary	-0.9266	0.3889	0.0180
education: Lower technical/secondary	-1.2209	0.4322	0.0051
education: None/Primary	-1.4716	0.4549	0.0014
education: Vocational	-1.3578	0.5876	0.0217

Figure 4.9 below show the fitted alpha sub-model. Difference in health scores is noted between elderly who experienced serious disease. There is a decrease in the probability to be in perfect health and from age 90 onwards, no difference was observed. For those who had higher (not) university or postgraduate level of education, remain above the average fitted with a general the decrease on probability to be in perfect health. Wider confidence intervals were observed in ages below 90 with slightly narrow confidence intervals from age of 90 onwards.

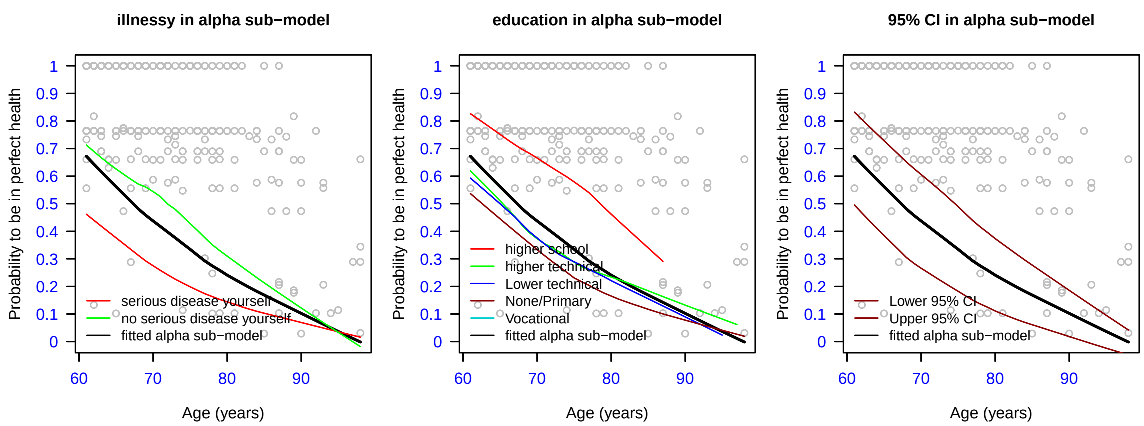

Figure 4.9: *EQ5D predictions by illnessy, education and 95% prediction confidence interval in alpha sub-model.*

Modeling beta regression in VAS

We considered beta regression and fitted different possible models based on the extended polynomials and fractional polynomials. The results for their comparisons in terms of AIC and likelihood ratio test are presented in Tables C.12 and C.13 (see appendix C.6). Tests for interactions indicated the need for interactions and/or dispersion sub-model was significant in all models. Only test for constant dispersion without any interaction in covariates was not significant in all polynomials and fractional polynomials degree one and two. Therefore, second order polynomial model was selected and based on the likelihood ratio the inclusion of variable dispersion model and interactions in covariates were supported. Predictions based on best models under each link function are shown in Figure C.6, with non-linear regression that could be considered using cubic splines. Fractional polynomial model was the best in terms of AIC, but for the ease interpretation, the second order polynomial model with logit link function was taken as a final model. The non-significant parameters were systematically eliminated from the model by backward selection. The parameter estimates with the corresponding standard errors and significance tests for the final model are summarized in Table 4.12.

Table 4.12: *Parameter estimates and standard error for the mean and dispersion sub-model based on second order logit polynomial.*

Parameter	**Estimates**	**Std. error**	**p-value**	**Estimates**	**Std. error**	**p-value**
	location sub-model			dispersion sub-model		
Intercept	3.4891	0.4667	<0.0001	2.4302	0.8181	0.0030
age	-0.0265	0.0064	<0.0001	-0.0127	0.0115	0.2683
Illnesy: Yes	-0.3950	0.1227	0.0013	0.3450	0.2357	0.1433
education: higher technical/secondary	-0.2503	0.1714	0.1443	1.0068	0.2875	0.0005
education: Lower technical/secondary	-0.2103	0.2364	0.3737	-0.1451	0.3159	0.6459
education: None/Primary	-0.6172	0.2162	0.0043	0.2893	0.3399	0.3947
education: Vocational	-0.6916	0.2207	0.0017	0.9803	0.4418	0.0265

The location sub-model (Table 4.12) models the average VAS score for elderly not in perfect health. The age was significant on the effect of VAS score. If not in perfect health (VAS score lower than 1), the effect of age had significantly lower VAS scores for every additional year. For elderly who had experienced serious disease, and elderly with primary and vocational level of education had significantly lower VAS scores.

Figure 4.10 shows the fitted location sub-model. A difference in health scores is noted between elderly who reported experience of serious disease before. The VAS scores remain below and it is decreasing in both levels. For those who reported primary education level, the VAS scores remain below the

average fitted model when compared with other levels. The confidence intervals are wider from 70 years onwards.

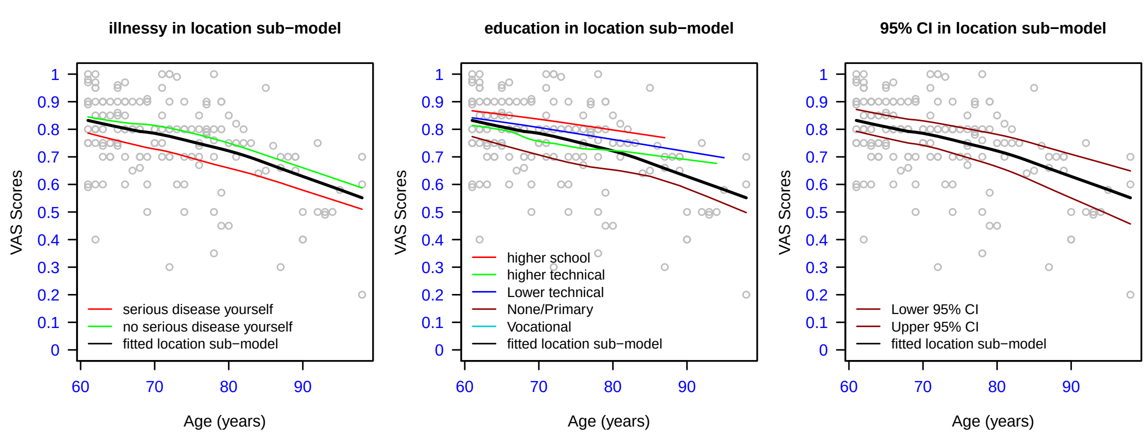

Figure 4.10: *VAS predictions by illnessy, education and 95% confidence interval in location sub-model.*

To investigate whether the results may have been affected by severe bias in the ML estimator, the resulting coefficients estimates and standard errors of bias-corrected and bootstrap method based on 2000 samples were performed as reported in Table C.14 (see appendix C.6). The obtained estimates were similar to the proposed estimates in the model, meaning that the use of small values of ϵ to move observations away from the boundary points did not appreciable affect parameter estimates.

4.3 Results for the sample of households

The data applied in this section come from the sample of households of HRQoL. To figure out the most important covariates, variable selection based on RF, lasso and elastic net were applied to select a subset of relevant covariates in model construction. Therefore, age, BMI rescaled, gender, normalday, province and profession were selected as the most important variable for both EQ5D and VAS outcomes (output not shown).

4.3.1 Beta GLMM

To allow for subject-specific inference a random effects model was considered and the results for the EQ5D and VAS scores are shown in Table 4.13 below. The age was significant on the effect of EQ5D and VAS score respectively. If not in perfect health, the effect of age had significantly lower EQ5D and VAS scores for every additional year. The random intercept is an intercept for each household. Thus, the variance of the random intercept is a measure of how much the households vary in their health scores. Therefore, the variance estimate of random intercept was approximately zero in EQ5D, meaning that no variability was observed in EQ5D response. For the VAS response, the variability of random intercept is significant.

Table 4.13: *Parameter estimates of beta GLMM in the sample of households for EQ5D and VAS outcome.*

Parameter	**Beta GLMM - EQ5D**			**Beta GLMM - VAS**		
	Estimates	**Std. error**	**p-value**	**Estimates**	**Std. error**	**p-value**
location sub-model						
Intercept	2.8486	0.0840	$<$0.0001	2.4517	0.0658	$<$0.0001
Age	-0.0090	0.0026	0.0005	-0.0137	0.0016	$<$0.0001
Female	0.0721	0.0885	0.4160	-0.0454	0.0514	0.3773
dispersion sub-model						
γ_0	1.8292	0.0919	$<$0.0001	2.2390	0.0926	$<$0.0001
γ_1	-0.0067	0.0029	0.0217	0.0089	0.0028	0.0014
γ_2	0.0786	0.1019	0.4409	-0.2018	0.0893	0.0245
σ_1^2	0.000000027	0.0000087	0.9975	0.3142	0.0424	$<$0.0001

4.3.2 Joint Beta GLMM

The results from the joint models of the two response variables using the NLMIXED procedure were estimated and summarized in Table 4.14. Significant differences were observed for the age of the participants in the households (p-value$<$0.0001) but not for the gender of the participant. The random effects for the two outcomes were also significantly positively associated. The estimate of the correlation between the random effects is far from one (0.79), with a high correlation between the health scores of both outcomes. Estimates were found to be very close to those from single analysis per outcome but the joint model yields with precision and allows for quantifying the association between outcomes.

Table 4.14: *Parameter estimates of multivariate beta GLMM in the sample of households.*

Parameter	**Description**	**Estimates**	**Std. error**	**p-value**
location sub-model in EQ5D				
β_{10}	Intercept	2.8609	0.0841	$<$0.0001
β_{11}	Age	-0.0092	0.0026	0.0004
β_{12}	Female	0.0643	0.0881	0.4656
dispersion sub-model in EQ5D				
γ_{10}	Intercept	1.8367	0.0917	$<$0.0001
γ_{11}	Age	-0.0064	0.0029	0.0277
γ_{12}	Female	0.0770	0.1016	0.4488
location sub-model in VAS				
β_{20}	Intercept	2.4507	0.0658	$<$0.0001
β_{21}	Age	-0.0136	0.0016	$<$0.0001
β_{20}	Female	-0.0488	0.0513	0.3418
dispersion sub-model in VAS				
γ_{20}	Intercept	2.2350	0.0923	$<$0.0001
γ_{21}	Age	0.0092	0.0028	0.0010
γ_{22}	Female	-0.2004	0.0891	0.0252
σ_1^2	Random intercept (EQ5D)	0.0036	0.1865	0.9847
σ_2^2	Random intercept (VAS)	0.1124	0.0315	0.0004
ρ	Correlation between random effects	0.7947	0.0412	$<$0.0001

Chapter 5

Discussion

Health related quality of life still remains a public health concern in the population and resources for the provision of health care are scarce. So, choices have to be made about how they are allocated. In this study, the interest was to determine and explain the quality of life in the general population in Flanders. Statistical models were applied on two datasets, motivated in part by the design of the study. In this analysis more than one outcome was of interest resulting into a sample of individuals with categories of all age (child, adult and elderly) groups and sample of households. Specifically EQ5D and VAS scores were considered in both datasets. Therefore, this section presents the discussion of the results divided according to the datasets used in the analysis.

For sample of individuals, the objective of this study was to analyze with different approaches to see which covariates would be considered more important with respect to either of both HRQoL outcome in different groups of categories and to model those covariates to describe the relationship with the outcome of interest.

The regression tree is conditional on the first split, and it has certain problem of being unstable. If we have to observe another sample in a population, it could have a different split. That is why the RF was used to provide the important variables. First, in the context of RFs, we fitted an unpruned tree. Recall that pruning is the important aspect of the regression tree methodology. The second notable difference is that for each node only a subset of the variables are considered as potential predictors, that is, instead of determining the best split among all potential predictors, a random sample of these variables are considered as potential splitting variables. A primary advantage of drawing a random subset of potential predictor variables at each node is that it offers a natural approach to handling collinearity in the data. The results from lasso are generally more accurate and some parameters will be shrunk towards zero, allowing for better interpretation of the model and identification of those covariates most strongly associated with the outcome. But lasso has problems with correlated data. So, the elastic net extends the lasso and uses the second penalty. If they are correlated, both covariates are going to the same point. Based on selection method one of the variables was not selected as important variable using the four methods, but based on the literature it was decided to include it in the model and was found to be important in some groups.

The distribution of health indices is commonly non-normal, exhibiting skewness to the left and a boundary at one. This study examined the applicability of beta regression and one-inflated beta regression to address the relationship between significant characteristics and both responses. Results showed that the best parametric model, according to AIC, was a polynomial model with the inclusion of interactions and dispersion. Also, by modeling dispersion in terms of covariates, beta regression provided information about the shape of the distribution, something that is not available in other methods. The logit was the selected link function, and according to Hosmer and Lemeshow (2000), is usually the parameter of interest due to its ease of interpretation.

In the child group, the covariates: age, whether a child had a normal day or not and whether a child had experienced serious disease before were related to the change in HRQoL for EQ5D and VAS. Also, in this group, age and whether a child had experienced serious disease before were related to the change in dispersion: The results suggest that age is associated with an increased variation of the HRQoL index scores. Girls' HRQoL scores are declining more than the scores for boys. A similar result was given by Michel et al. (2009), who reported that girls showed a more profound decrease in HRQoL with increasing age. And from age 12, female adolescents are in a worse position than male adolescents regarding subjective health and HRQoL.

For the adult group, the covariates: age, for those who had none or primary education and vocational education level, and whether the person had experienced serious disease before were related to the change in HRQoL for EQ5D and VAS. Having one or more domestic animal was mostly related with the change in VAS score. Levine et al. (2013) studied the pet ownership and systemic hypertension, and found the association between pet ownership and lower blood pressure, and they studied also pet ownership and physical activity, where they found that in all pets, dogs are more likely to positively influence the level of human physical activity.

In the elderly group, the covariates: age, smoke behaviour, for those who had none or primary education and vocational education level, and whether had experienced serious disease before were related to the change in HRQoL for EQ5D and VAS. Also, age, and whether the person had experienced serious disease before were related to the change in dispersion: The results suggest that age and whether the person had experienced serious disease before is associated with an increased variation of the HRQoL index scores. Lima et al. (2009) studied the health related quality of life among the elderly from the age of 60 years or more, where HRQoL was found to be worse among women, in individuals at advanced ages, those who practiced evangelical religions and those with lower levels of income and schooling.

At the sample of households, the main research question in this analysis was to investigate if the HRQoL measures are clustered in households. In this report, we examined the potential of beta regression methods in the analysis of clustered HRQoL data. Beta GLMM for the separated response and joint beta GLMM for both responses simultaneously, were fitted using adaptive Gaussian quadrature for numerical approximations in order to draw inference at the subject specific level. With a subject-specific approach, the responses were modeled as a function of covariates and parameters for the mean sub-model and precision sub-model, specific to a subject, providing interpretation of fixed-effect parameters conditional on a constant level of random-effects parameter. The use of the adaptive Gaussian quadrature points assisted in ensuring more stable results in the SAS NLMIXED procedure. This model is very simple in some sense and more things can be done (e.g. adding random-effects for the dispersion), but of course there is a computational issue on it, and interpretation will then become more difficult.

It was observed in both methods that the health scores decrease significantly with increasing age. Individuals from the same household had EQ5D health scores more similar to each other than to any other person from a random household. There was an association between the linear predictors of the EQ5D and VAS index responses.

.

Chapter 6

Conclusion

In this study, different approaches were applied to assess the health related quality of life in Flanders and possible factors influencing it. These methods showed that the covariates age, gender, experience with serious disease before, if they filled in the diary on a normal day and number of persons in the household in child group; age, experience with serious disease before, experience with serious disease in family, education level, if they filled in the diary on a normal day and if the family has one or more domestic animals in adult group; age, experience with serious disease before, smoke behaviour and education level in elderly group were considered the most predictive among those considered in study and were thus worthy of further investigation. Statistical analysis showed that age, experience with serious disease before, experience with serious disease in family, if they filled in the diary on a normal day, education level, if the family one or more domestic animals, smoke behaviour and gender were statistically significant characteristics of the participants related to their HRQoL experience. It was found that individuals from the same household had EQ5D health scores more similar to each other than to any person from a random household. This was not the case for the VAS index. Significant association between the health scores of EQ5D and VAS was present.

Limitations and recommendations

The findings of this report are constrained by some limitations concerning the definition of the variables used. It was not possible to specify the type of domestic animal during the data collection. This could help understanding if different types of domestic animals could influence HRQoL of the individuals studied.

Finally, it should be mentioned that this report did not exhausted the statistical methods for the analysis of health related quality of life in Flanders, and other methods could be also considered as well. For instance, it was observed in this dataset that there is some systematic frequency of digits in both responses. Therefore, digit preference approach could be plausible to apply to this dataset. Furthermore, methods allowing for negative EQ-5D values could be used, so that the whole range of possible EQ-5D values can be considered.

.

Bibliography

Agresti, A. (2002). *Categorical Data Analysis*. Second Edition. New York: John Wiley and Sons Inc.

Alsén, P. (2009). Illness Perception and Fatigue after Myocardial Infarction. University of Gothenburg. Gothenborg.

Bisegger, C., Cloetta, B., von Rueden, U., Abel, T. and Ravens-Sieberer, U. (2005). Health-related quality of life: gender differences in childhood and adolescence. *Soz Praventivmed*. **50**(5):281-91.

Breiman, L. (2001). Random forests. *Machine Learning*, **45**, 5-32.

Breiman, L. (2001a). Statistical modeling: the two cultures (with discussion). *Statistical Science*, **16**, 199-231.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D., Geys, H. (2000). The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*. **1**:49-67.

Centers for Disease Control and Prevention. Body Mass Index: Considerations for Practitioners. Available at: http://www.cdc.gov/obesity/downloads/bmiforpactitioners.pdf. Accessed on 8/11/2013.

Cheung, K., Oemar, M., Oppe, M. and Rabin, R. (2009). User Guide. Basic Information on how to use EQ5D. EuroQol Group.

Cribari-Neto, F., Zeileis, A. (2010). Beta Regression in R. *Journal of Statistical Software*, **34** (2), 1-24.

Cleemput, I. (2010). A social preference valuation set for EQ-5D health states in Flanders, Belgium. *Eur J Health Econ* **11**:205-213.

Dominick, K. L, Ahern, F. M., Gold, C. H. and Heller, D. A. (2002). Relationship of health-related quality of life to health care utilization and mortality among older adults. *Aging Clin Exp Res*. **14**(6):499-508.

DeSalvo, K. B., Bloser, N., Reynolds, K., He, J. and Muntner, P. J. (2006). Mortality prediction with a single general self-rated health question. A meta-analysis. *Gen Intern Med*. **21**(3):267-75.

Dolan, P. (1997). Modeling evaluations for EuroQol health states. *Medical Care*. **35**(11): 1095-1108.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*. **32**, 407-499.

Ferrari, S. L. P. and Cribari-Neto, F. (2004). Beta regression for modeling rates and proportions. *Journal of Applied Statistics*. **31**, 799-815.

Fitzmaurice, G. M., Laird, N. M., Ware, J. H. (2009). *Applied Longitudinal Analysis*. Second edition. New York: Wiley.

Fieuws, S. and Verbeke, G. (2004). Joint modelling of multivariate longitudinal profiles: pitfalls of the random-effects approach. *Statist. Med*. **23**:3093-3104.

Filzmoser, P. (2008). Linear and nonlinear methods for regression and classification and applications in R. Department of Statistics and Probability Theory, Vienna University of Technology.

Foulkes, A. S. (2009). *Applied Statistical Genetics in R For Population-based Association Studies*. Springer. New York.

Gueorguieva, R. (2001). A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family. *Statistical Modelling*. **1**:177-193.

Guyatt, G. H., Feeny, D. H. and Patrick, D. L. (1993). Measuring Health-related Quality of Life. *Annals of Internal Medicine*.**118**:622-629.

Guyon, I. and Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*. **3** 1157-1182.

Grun, B, Kosmidis, I. and Zeileis, A. (2012). Extended Beta Regression in R: Shaken, Stirred, Mixed, and Partitioned. *Journal of Statistical Software*. **48**(11), 1-25.

Guyatt, G. H. (1993). Measurement of Health-Related Quality of Life in Heart Failure. *JACC*. Vol. **22**, No. 4, 185A-191A.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Second Edition. Springer New York Inc.

Hapfelmeier, A., Hothorn, T., Ulm, K. and Strobl, C. (2012). A new variable importance measure for random forests with missing data. Available at: http://link.springer.com/article/10.1007/s11222-012-9349-1. Accessed on 22/07/2013.

Hapfelmeier, A. (2012): Analysis of missing data with random forests. Dissertation, LMU Munchen: Faculty of Mathematics, Computer Science and Statistics.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*. Second Edition. John Wiley & Sons, INC.

Hunger, M., Doring, A. and Holle, R. (2012). Longitudinal beta regression models for analyzing health-related quality of life scores over time. BMC *Medical Research Methodology*,**12**:144.

Iddi, S. and Molenberghs, G. (2012). A joint marginalized multilevel model for longitudinal outcomes. *Computational Statistics and Data Analysis*. **39**:1944-1951.

Kind, P. (1996). The Euroqol instrument: an index of health-related quality of life. In Bert Spilker, ed. Quality of Life and Pharmacoeconomics in Clinical Trials. Second Edition. Philadelphia: Lippincott-Raven Press.

Levine, G. N., Allen, K., Braun, L. T., Christian, H. E., Friedmann, E., Taubert, K. A., Thomas, S. A., Wells, D. L. and Lange, R. A. (2013). Pet Ownership and Cardiovascular Risk. *American Heart Association*.**127**:2353-2363.

Lima, M. G., Barros, M. B. A., Cesar, C. L.G., Goldbaum, M., Carandina, L. and Ciconelli, R. M. (2009). Health related quality of life among the elderly: a population-based study using SF-36 survey. *Cad. SaÃºde PÃºblica*, Rio de Janeiro, **25**(10):2159-2167.

Muhwezi, W. W., Okello, E. S. and Turiho, A. K. (2010). Gender-based profiling of Quality of Life (QOL) of primary health care (PHC) attendees in central Uganda: a cross sectional analysis. *African Health Sciences*. **10**(4): 374-385.

Michel, G., Bisegger, C., Fuhr, D. C. and Abel, T. (2009). Age and gender differences in health-related quality of life of children and adolescents in Europe: a multilevel analysis. *Qual Life Res*. **18**:1147-1157.

Morgan, J. N. and Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*. **58**, 415-434.

Melnychuk, M. (2013). Boosted regression trees and random forests. *Statistical Consulting Report*, Fisheries Centre, UBC.

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

Patrick, D. L. and Erickson, P. (1993) *Health status and health policy: Quality of life in health care evaluation and resource allocation*. New York: Oxford University Press.

Ritschard, G. (2010). CHAID and Earlier Supervised Tree Methods. Dept of Econometrics, University of Geneva, Switzerland.

Robin, R. and deCharro, F. (2001). EQ-5D: a measure of health status from the EuroQol group. *Annals of Medicine*. **33**(5):337-343.

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T. and Zeileis, A. (2008). Conditional variable importance for random forests. BMC *Bioinformatics*. **9**:307:1471-2105.

Santana, M. J. and Feeny, D. (2008). The Importance of Measuring Health-related Quality of Life. *Institute of Health Economics*. Alberta Canada.

Simas, A. B., Barreto-Souza, W. and Rocha, A. V. (2010). Improved estimators for a general class of beta regression models. *Comput Stat Data Anal*. **54**:348-66.

Smithson, M. and Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta distributed dependent variables. *Psychol Methods*. **11**:54-71.

Tibshirani, R. (1996): Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*. Series B **58**(1), pp.267-288.

Thum, Y. M. (1997). Hierarchical linear models for multivariate outcomes. *Journal of Educational and Behavioral Statistics*. **22** (1):77-108.

Verkuilen, J. and Smithson, M. (2012). Mixed and mixture regression models for continuous bounded responses using the beta distribution. *Journal of Education and Behavioral Statistics*. **37**(1):82-113.

Wu, T. T. and Lange, K. (2008). Coordinate Descent Algorithms For Lasso Penalized Regression. *The Annals of Applied Statistics*. Vol. **2**, No. 1, 224-244.

Wilkinson, L. (1992). Tree Structured Data Analysis: AID, CHAID and CART. Department of Statistics, Northwestern University, Evanston, IL 60201.

Wilson, S. (2013). Diet and Fitness. Discovery Health "BMI for Children". Available at: http://health.howstuffworks.com/wellness/diet-fitness/weight-loss/bmi2.htm. Accessed on 8/11/2013.

WHOQOL, G. (1998). The World Health Organization Quality of Life Assessment (WHOQOL). Development and psychometric properties. *Soc. Sci. Med*. Vol. **46**, No. 12, pp. 1569-1585.

Zou, H. and Hastie, T. (2003). Regression Shrinkage and Selection via the Elastic Net, with Applications to Microarrays. Department of Statistics, Stanford University, Stanford, CA 94305.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc*. Ser. B, **67**(2):301-320.

Zimprich, D. (2010). Modeling Change in Skewed Variables Using Mixed Beta Regression Models. *Research in Human Development*. **7**:1, 9-26.

Appendix A

Descriptive statistics

A.1 Socio-demographic characteristics of the sample of individuals

Table A.1: *Socio-demographic characteristics of the sample of individuals in HRQoL.*

Population level Variables (Child, Adult and Elderly shared variables) : n=3117				
Variable	**Levels**	**%**	**Type**	**Remark**
Gender	Female	52.20	categorical	Gender of respondent
	Male	47.80		
Agecat /Age	Child [0-12]	28.42	categorical	Age category of a person, of diaries also distinguishes the 3 types
	Adult [13-60]	59.93		
	Elderly [61 and older]	11.65		
Age	Observed	100.00	continuous	Age in years
	Missing	-		
BMI	Observed	0.28	continuous	Body Mass Index
	Missing	0.72		
Province	Antwerpen	27.11	categorical	Provinces
	Oost-Vlaanderen	21.69		
	West-Vlaanderen	17.65		
	Vlaams-Brabant	14.31		
	Limburg	13.73		
	Brussels Hoofdstedelijk Gewest	4.91		
	Missing	0.61		
Animal	Yes	63.23	categorical	Has the family one or more domestic animals
	No	36.12		
	Missing	0.64		
Normalday	Yes	73.02	categorical	Normal day
	No because other reason	23.97		
	No because sick	2.18		
	Missing	0.83		
Parents	2	81.55	categorical	Number of parents in a family
	1	15.72		
	Missing	2.73		
illnessy	No	81.91	categorical	Experience with serious disease with yourself
	Yes	12.93		
	Missing	5.17		
illnessf	No	48.48	categorical	Experience with serious disease with member of your family
	Yes	43.79		
	Missing	7.73		

Table A.2: *Socio-demographic characteristics of the sample of individuals in HRQoL (cont.).*

Variable	Levels	%	Type	Remark
Population level Variables (Child, Adult and Elderly shared variables): n=3117				
illnessf	No	48.48	categorical	Experience with serious disease with member of your family
	Yes	43.79		
	Missing	7.73		
Peoplehousehold	0	0.19	Categorical (ordinal)	Number of persons in the household
	1	3.14		
	2	10.84		
	3	17.71		
	4	36.06		
	5	15.17		
	6	4.20		
	7	0.67		
	8	0.10		
	9	0.03		
	11	0.03		
	Missing	11.84		
Child level additional variables: n=886				
mumEducation	higher (not-)university/postgraduate	64.45	categorical	Education level for a childs mother
	higher technical/secondary	19.64		
	Vocational	9.03		
	lower technical/secondary	4.18		
	None/Primary	1.47		
	Missing	1.24		
Adults and elderly shared variables: n=2231				
Education	higher (not-)university/postgraduate	43.12	categorical	Education level
	higher technical/secondary	25.28		
	lower technical/secondary	10.71		
	Vocational	9.95		
	None/Primary	9.14		
	Missing	1.79		
Smokestatus	Non-smoker	61.50	categorical	Smoke behaviour
	Ex-smoker	20.80		
	Smoker	16.27		
	Missing	1.43		
WorkedinHCare	No	76.47	categorical	work(ed) in health care sector?
	Yes	21.69		
	Missing	1.84		
illnessc	No	66.92	categorical	Experience with serious disease because you cared for someone
	Yes	8.47		
	Missing	24.61		
Profession	White collar job	49.89	categorical	Respondents profession
	Other	22.81		
	Blue collar job	14.97		
	Self-employed	9.32		
	Missing	3.00		

Table A.3: *Socio-demographic characteristics of the sample in HRQoL (cont.).*

Variable	Levels	%	Type	Remark
Elderly level additional variables: n=363				
Working	No	90.63	categorical	Elderly work status
	Yes	5.23		
	Missing	4.13		
Freq1	a couple of times a week	52.07	categorical	Frequency see children
	a couple of times a month	24.79		
	a couple of times a year	6.89		
	once a year or less	2.20		
	Missing	14.05		
Freq2	a couple of times a week	33.06	categorical	Frequency see grandchildren
	a couple of times a month	25.62		
	a couple of times a year	16.80		
	once a year or less	3.86		
	Missing	20.66		
Freq3	rarely or never	36.09	categorical	Frequency drinking alcohol
	weekly	22.31		
	daily	20.94		
	monthly	15.70		
	Missing	4.96		
Response variables				
VAS	Observed	0.96	continuous	Outcome measures by VAS
	Missing	0.04		
EQ5D	Observed	0.98	continuous	Outcome measures by Cleemput EQ5D
	Missing	0.02		

A.2 Boxplot at individual and household sample

1. Child category

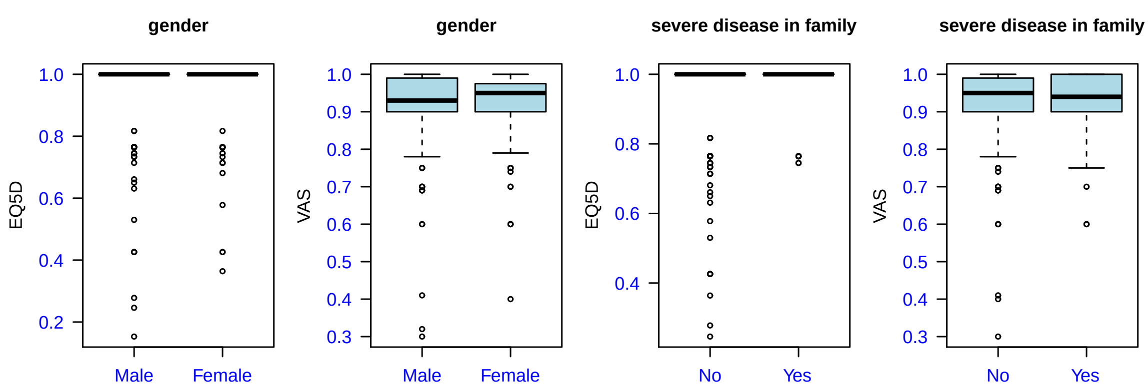

Figure A.1: *Boxplots for HRQoL by gender and illnessf.*

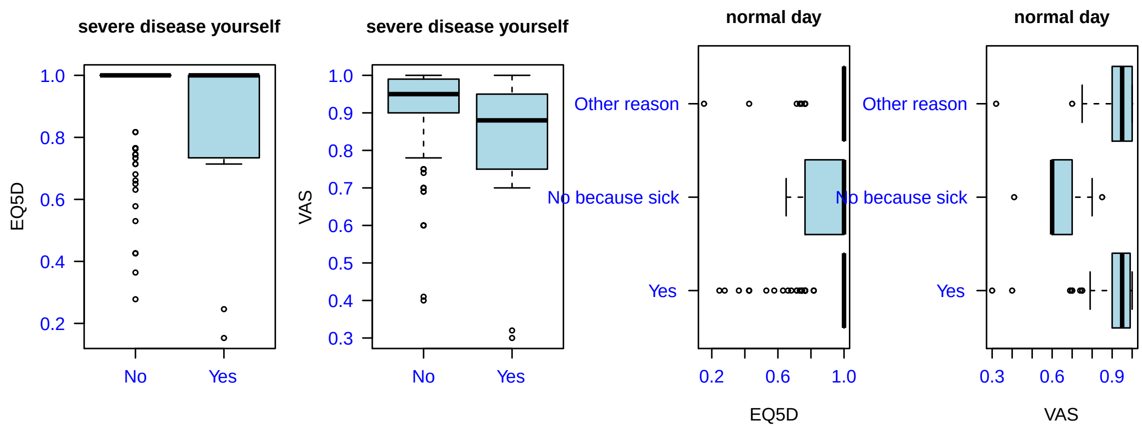

Figure A.2: *Boxplots for HRQoL by illnessy and normal day categories.*

2. Adult category

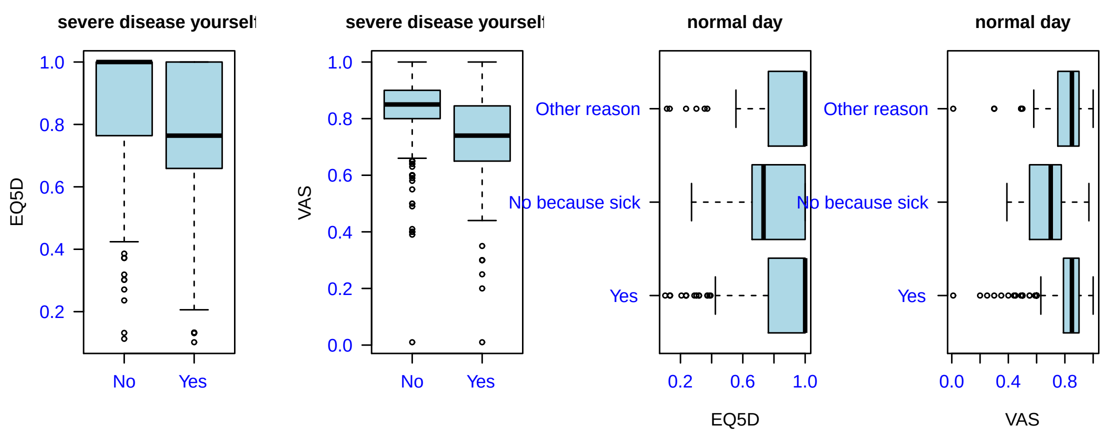

Figure A.3: *Boxplots for HRQoL by illnessy and normal day categories.*

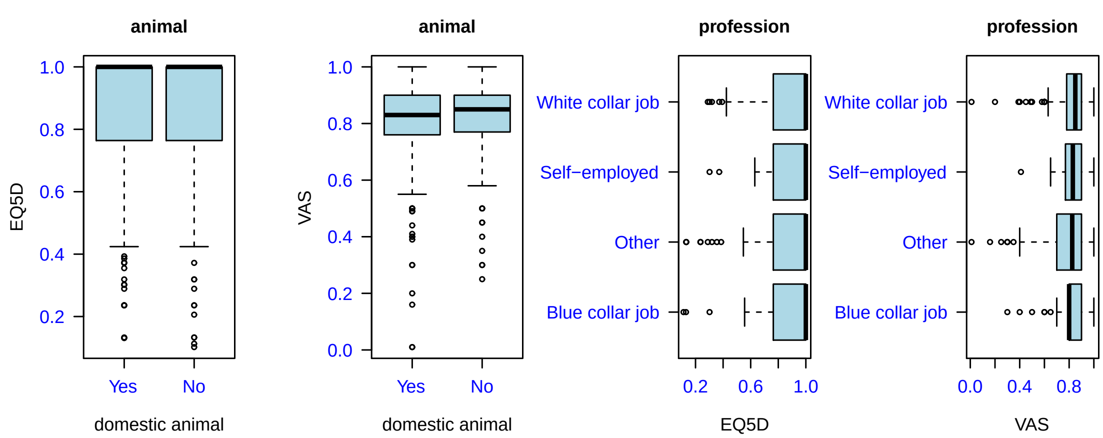

Figure A.4: *Boxplots for HRQoL by animal and profession categories.*

3. Elderly category

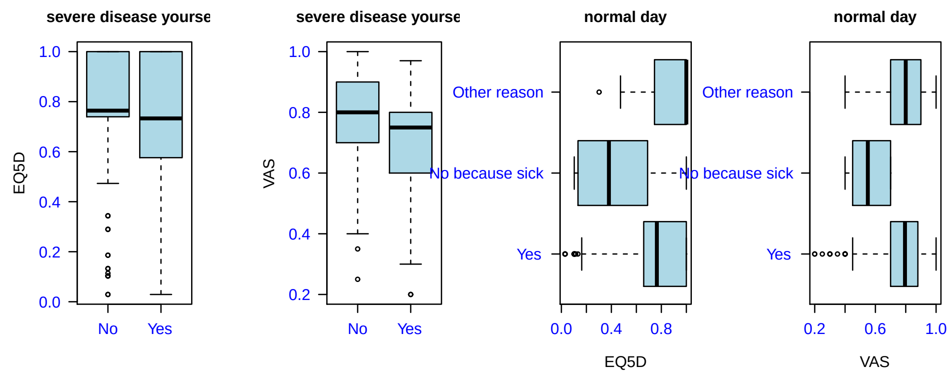

Figure A.5: *Boxplots for HRQoL by illnessy and normal day categories.*

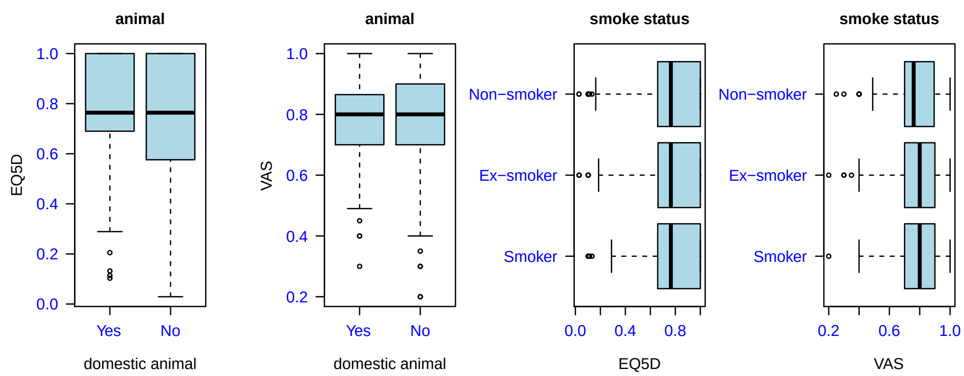

Figure A.6: *Boxplots for HRQoL by animal and smoke status categories.*

4. Household sample

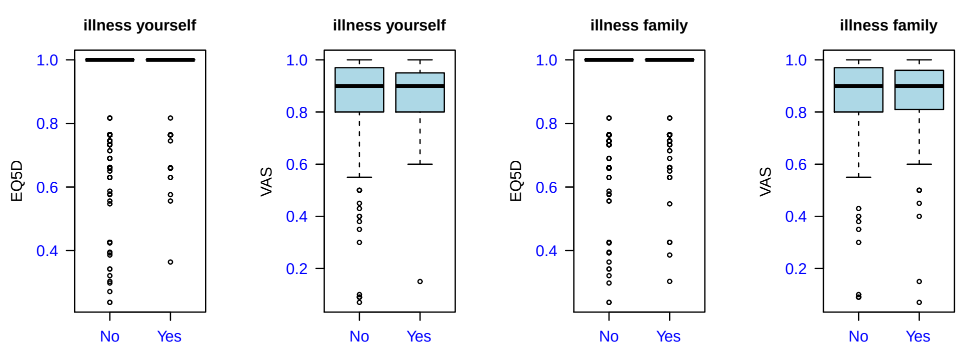

Figure A.7: *Boxplots for HRQoL by illnessy categories.*

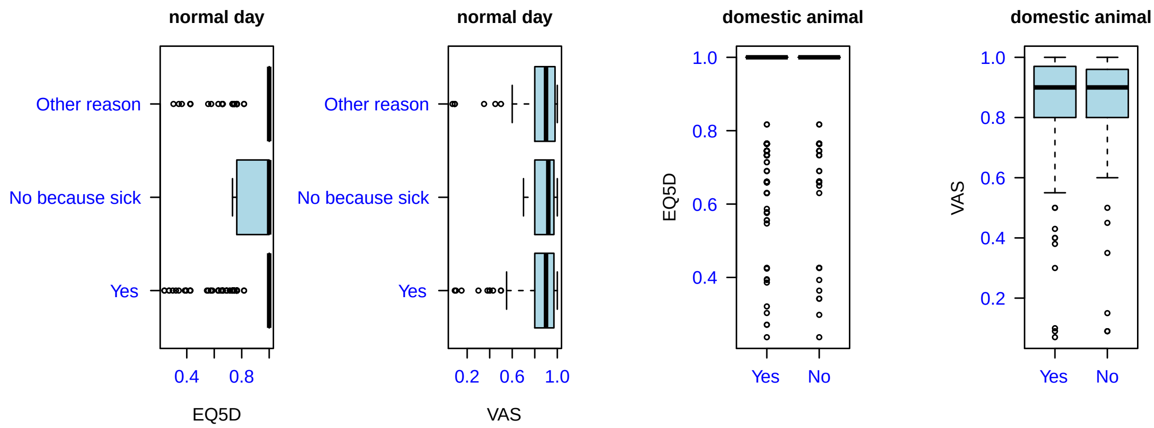

Figure A.8: *Boxplots for HRQoL by normal day and animal categories.*

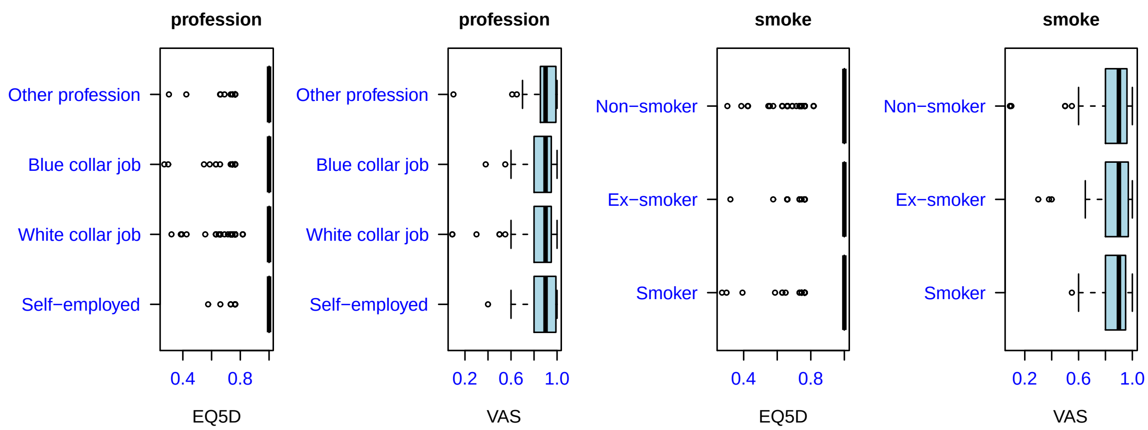

Figure A.9: *Boxplots for HRQoL by profession and smoke status categories.*

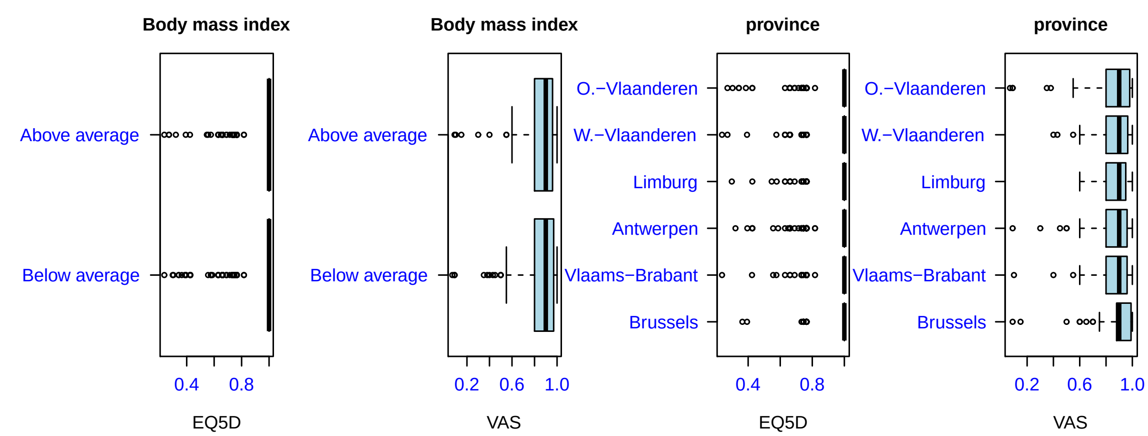

Figure A.10: *Boxplots for HRQoL by BMI and province categories.*

Appendix B

Variable selection

B.1 Variable selection plots for individual sample

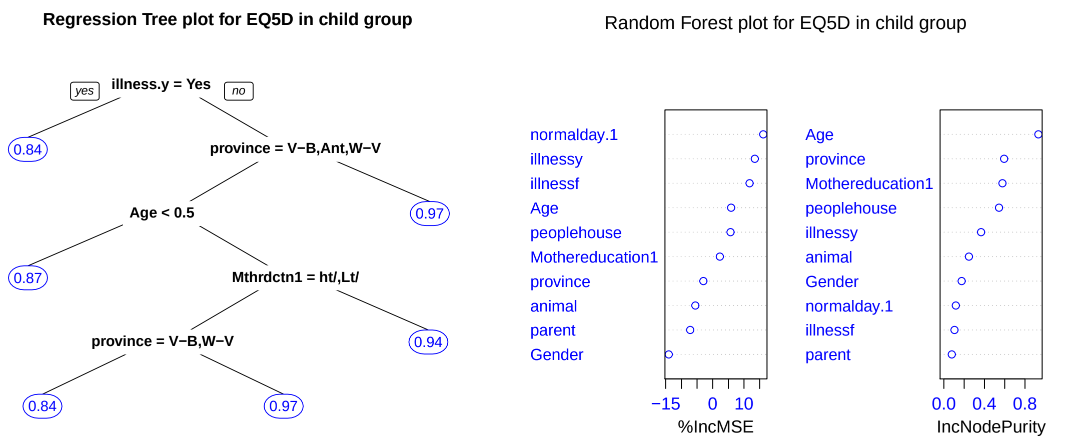

Figure B.1: *Regression tree (left) and Random forest (right) for the EQ5D in child group.*

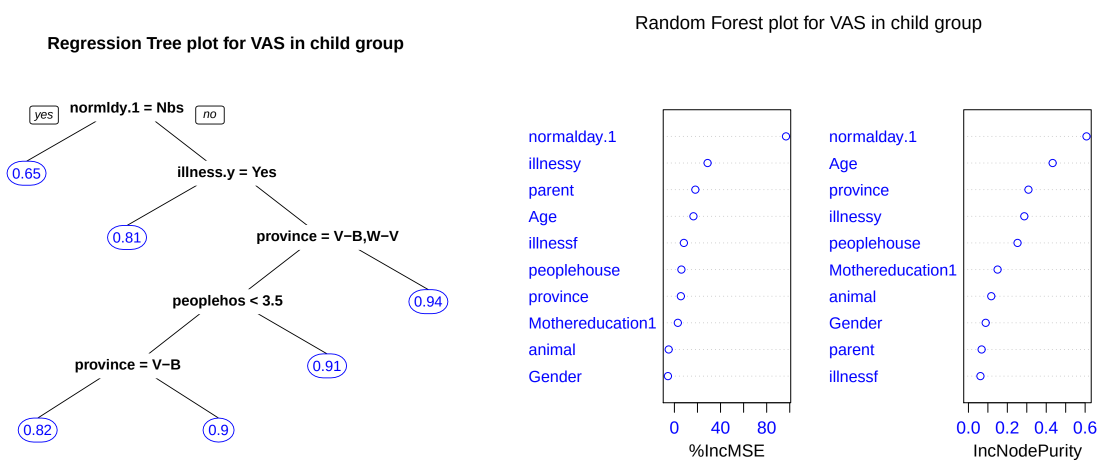

Figure B.2: *Regression tree (left) and Random forest (right) for the VAS in child group.*

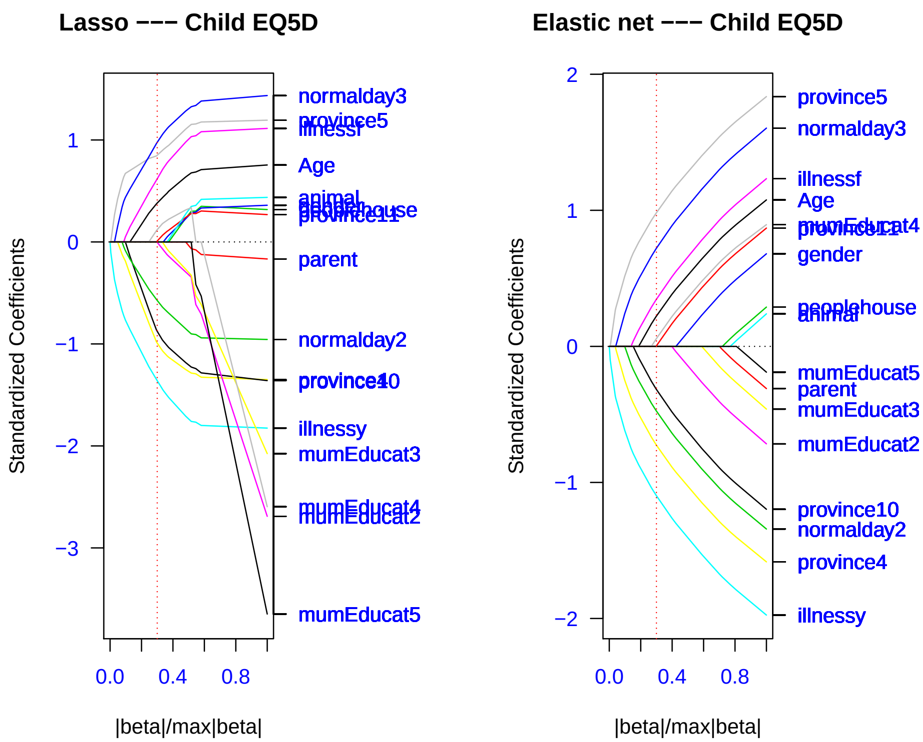

Figure B.3: *Lasso estimates (left) and elastic net estimates (right) for the EQ5D in child group.*

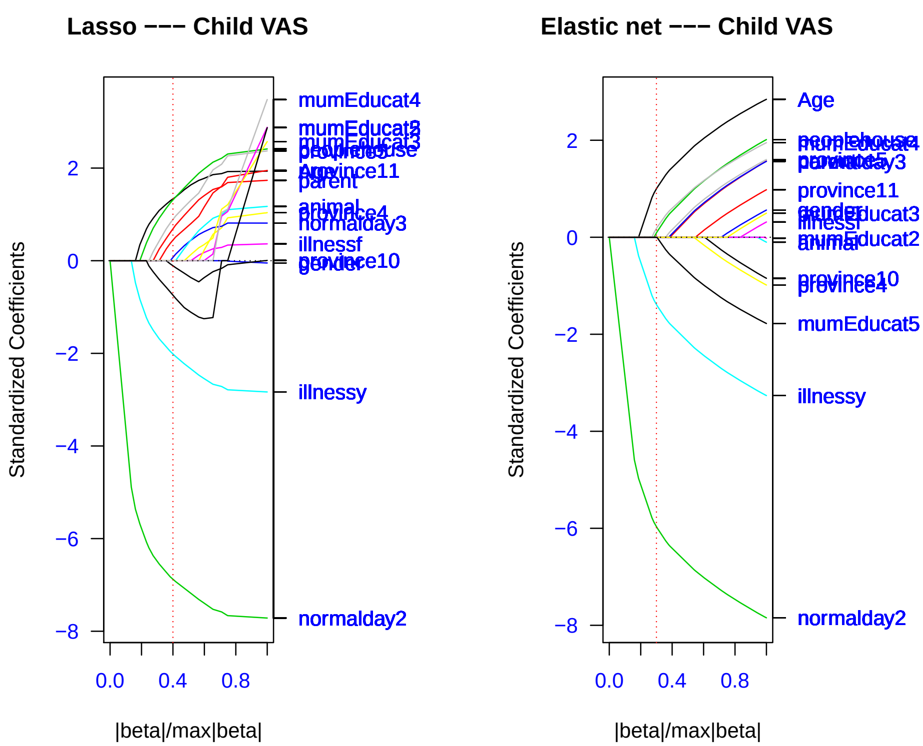

Figure B.4: *Lasso estimates (left) and elastic net estimates (right) for the VAS in child group.*

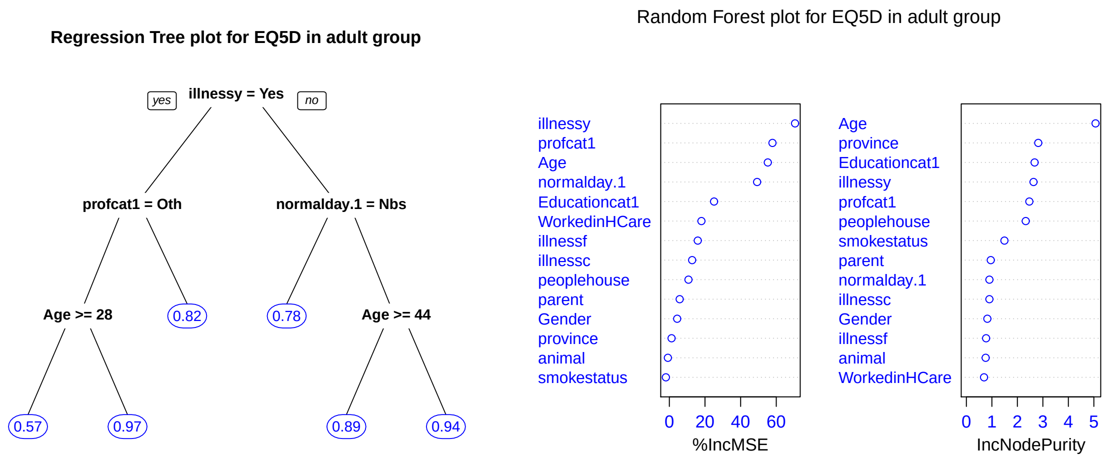

Figure B.5: *Regression tree (left) and Random forest (right) for the EQ5D in adult group.*

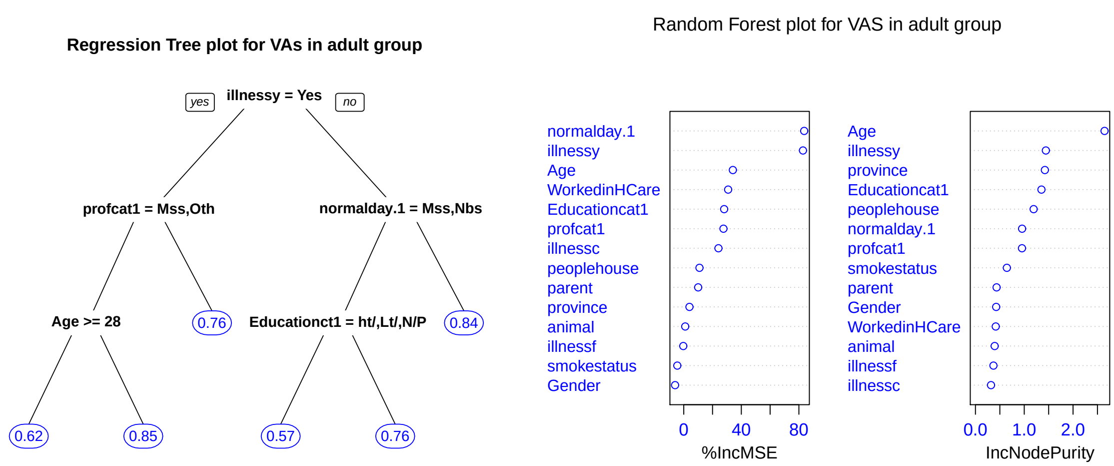

Figure B.6: *Regression tree (left) and Random forest (right) for the VAS in adult group.*

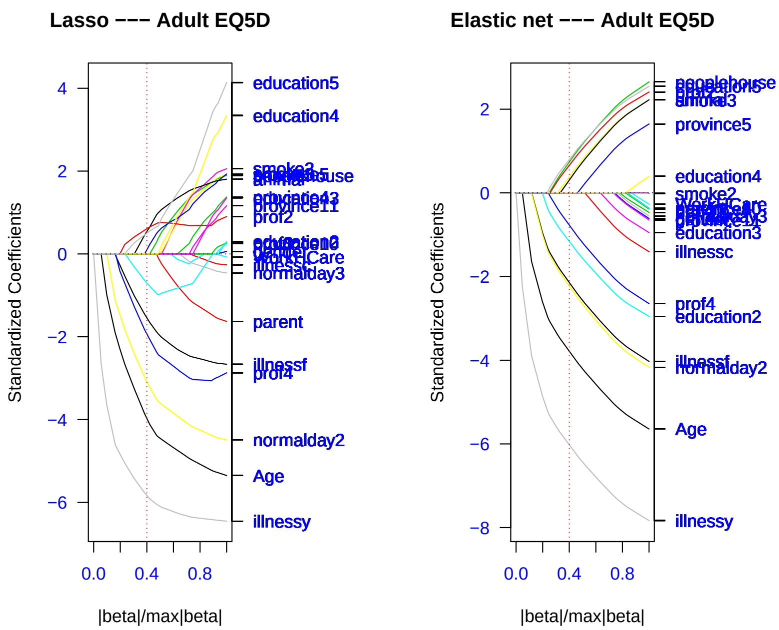

Figure B.7: *Lasso estimates (left) and elastic net estimates (right) for the EQ5D in adult group.*

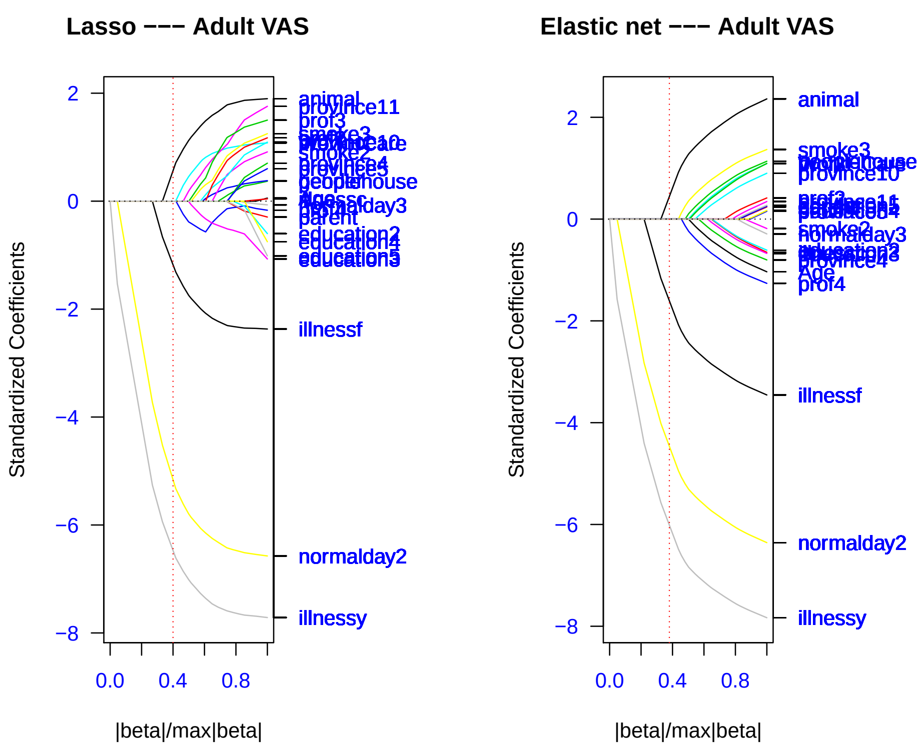

Figure B.8: *Lasso estimates (left) and elastic net estimates (right) for the VAS in adult group.*

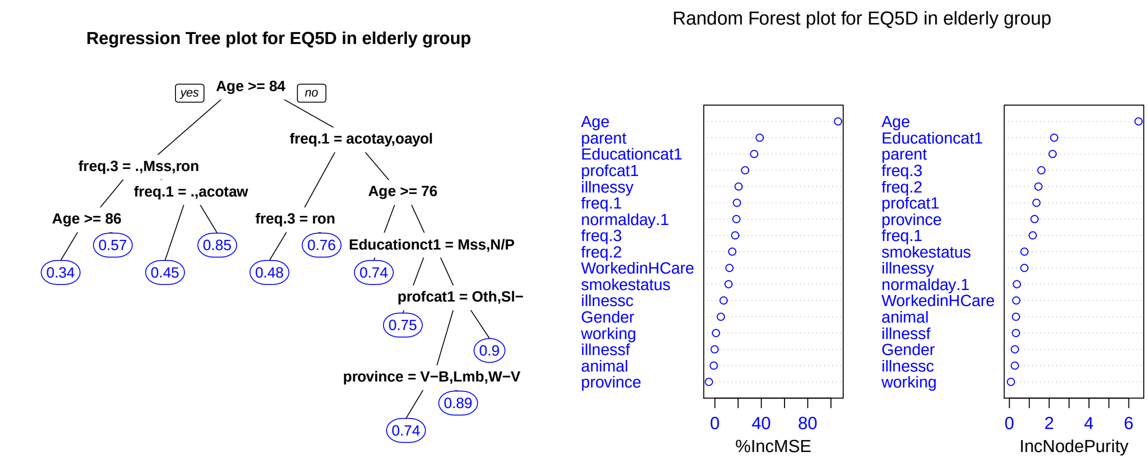

Figure B.9: *Regression tree (left) and Random forest (right) for the EQ5D in elderly group.*

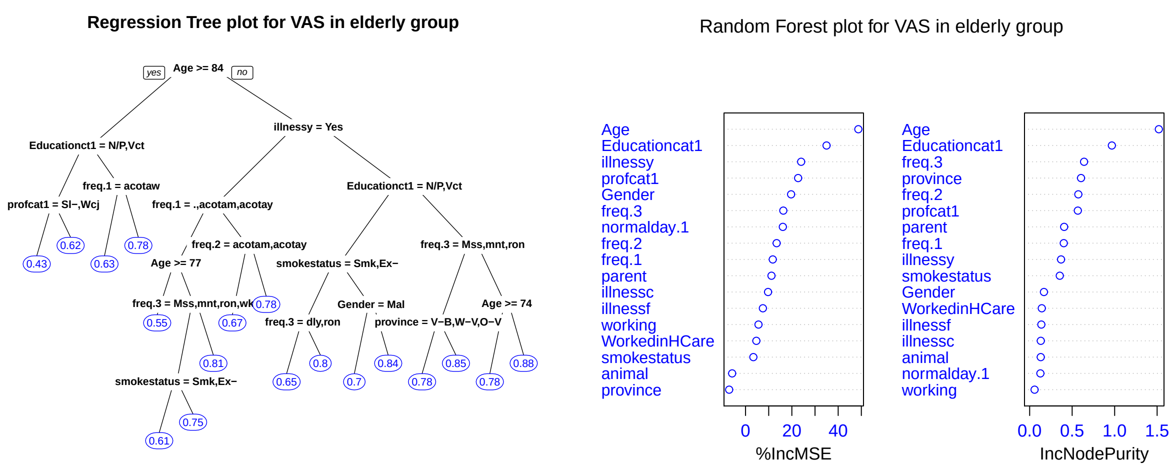

Figure B.10: *Regression tree (left) and Random forest (right) for the VAS in elderly group.*

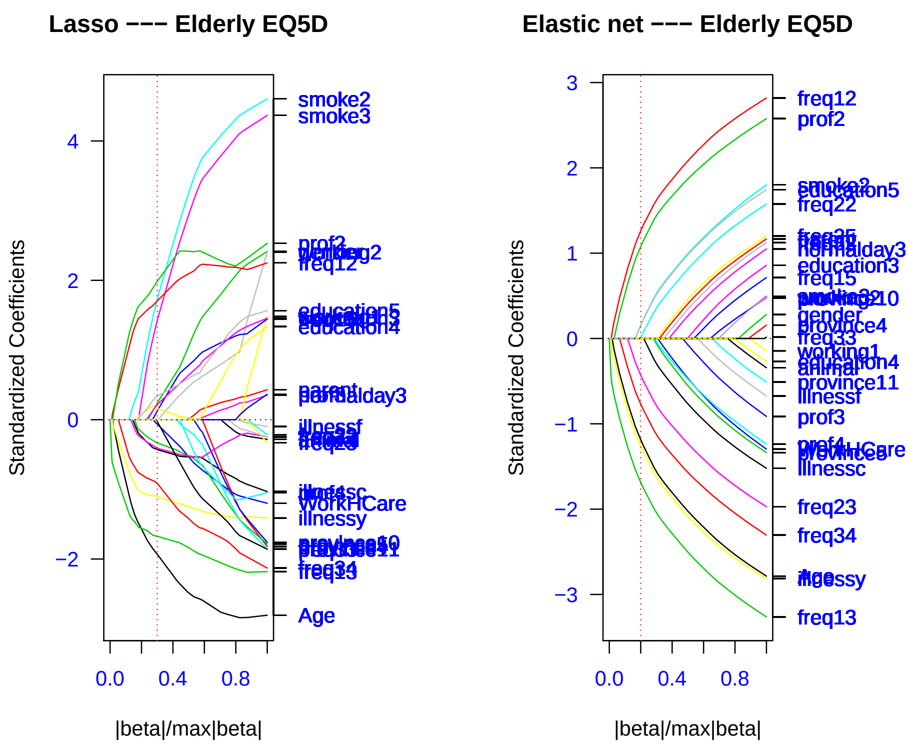

Figure B.11: *Lasso estimates (left) and elastic net estimates (right) for the EQ5D in elderly group.*

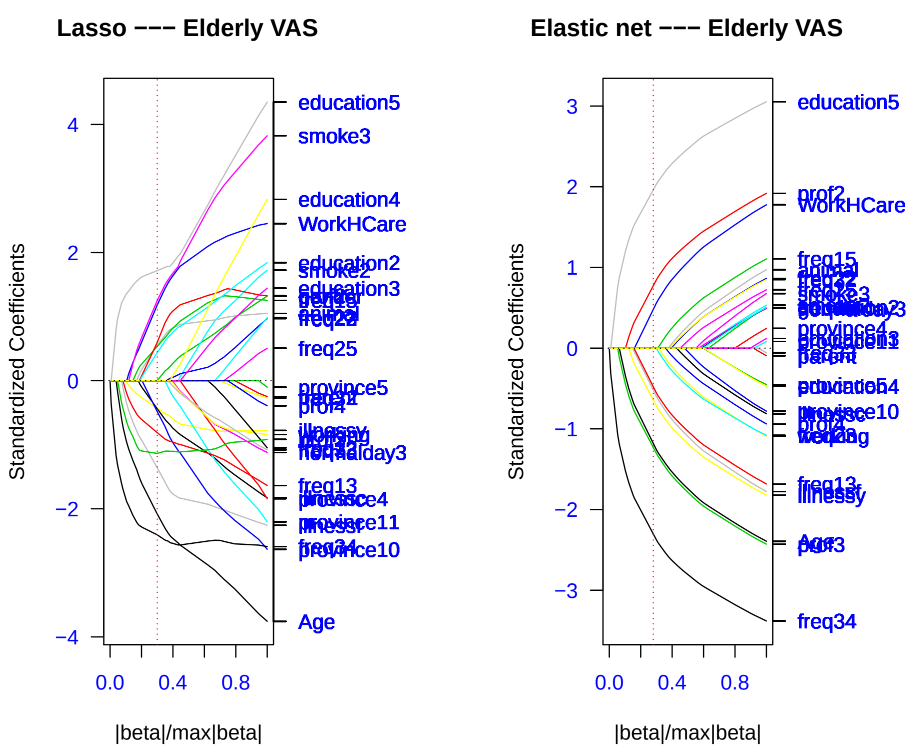

Figure B.12: *Lasso estimates (left) and elastic net estimates (right) for the VAS in elderly group.*

Appendix C

Statistical analysis

C.1 One inflated beta regression in EQ5D-child group

Table C.1: *Model comparison based on AIC and Likelihood Ratio Tests for EQ5D polynomial models.*

Polynomial Order	**Model**	**Model with interactions**	**Model with Dispersion**	**AIC**			**Comparison on logit link**	**LRT) (p-value)**
				logit	**probit**	**cloglog**		
	1	Yes	Variable	-204.353	-216.776	-209.312	1 vs. 2	$<$0.0001
1	2	Yes	Fixed	279.555	280.044	281.544	1 vs. 3	$<$0.0001
	3	No	Variable	124.532	122.968	115.484	3 vs. 4	$<$0.0001
	4	No	Fixed	224.022	224.237	225.329	2 vs. 4	0.0004
	5	Yes	Variable	-253.301	-259.903	-257.392	5 vs. 6	$<$0.0001
2	6	Yes	Fixed	241.347	241.745	242.869	5 vs. 7	$<$0.0001
	7	No	Variable	112.582	112.250	111.746	7 vs. 8	$<$0.0001
	8	No	Fixed	222.196	222.234	222.893	6 vs. 8	0.0001
	9	Yes	Variable	-305.180	-305.014	-305.935	9 vs. 10	$<$0.0001
3	10	Yes	Fixed	249.508	249.800	230.425	9 vs. 11	$<$0.0001
	11	No	Variable	96.313	95.992	95.008	11 vs. 12	$<$0.0001
	12	No	Fixed	224.507	224.633	225.670	10 vs. 12	$<$0.0001

Table C.2: *Model comparison based on AIC and Likelihood Ratio Tests for EQ5D fractional polynomial models.*

Fractional polyn. degree	**Power** (μ, ϕ, α)	**Model with interactions**	**Model with Dispersion**	**AIC**			**Comparison on logit link**	**LRT) (p-value)**
				logit	**probit**	**cloglog**		
	-0.5;	Yes	Variable	-204.184	-209.005	-206.852	1 vs. 2	$<$0.0001
1	0.5;	Yes	Fixed	256.751	257.256	258.836	1 vs. 3	$<$0.0001
	0	No	Variable	117.300	116.860	116.206	3 vs. 4	$<$0.0001
		No	Fixed	222.845	222.722	222.889	2 vs. 4	0.0001
	1,2;	Yes	Variable	-269.775	-269.240	-259.485	5 vs. 6	$<$0.0001
2	-2,-1;	Yes	Fixed	245.727	245.857	246.371	5 vs. 7	$<$0.0001
	-2,-2	No	Variable	94.323	93.759	92.818	7 vs. 8	$<$0.0001
		No	Fixed	221.317	221.233	221.706	6 vs. 8	0.0004
	-2,-2,2;	Yes	Variable	-301.092	-309.822	-281.606	9 vs. 10	$<$0.0001
3	3,3,3;	Yes	Fixed	229.404	230.019	231.345	9 vs. 11	$<$0.0001
	-2,-2,-2	No	Variable	53.251	53.607	54.475	11 vs. 12	$<$0.0001
		No	Fixed	226.334	226.158	226.196	10 vs. 12	0.0001

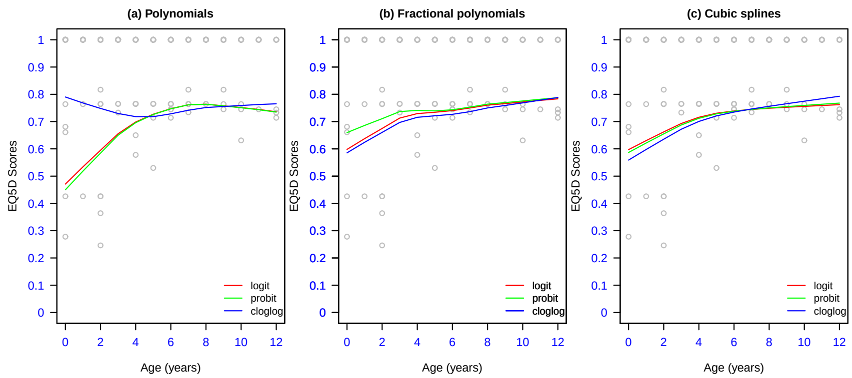

Figure C.1: *Representation of the best fits for the polynomial, fractional polynomial and cubic splines under different link functions in EQ5D child response.*

C.2 One inflated beta regression in VAS -child group

Table C.3: *Model comparison based on AIC and Likelihood Ratio Tests for VAS polynomial models.*

Polynomial Order	Model	Model with interactions	Model with Dispersion	AIC logit	AIC probit	AIC cloglog	Comparison on logit link	LRT) (p-value)
	1	Yes	Variable	-551.067	-558.935	-555.378	1 vs. 2	<0.0001
1	2	Yes	Fixed	-182.991	-182.711	-179.914	1 vs. 3	<0.0001
	3	No	Variable	-297.192	-297.606	-297.961	3 vs. 4	<0.0001
	4	No	Fixed	-226.236	-226.339	-226.271	2 vs. 4	0.0423
	5	Yes	Variable	-656.367	-653.013	-666.119	5 vs. 6	<0.0001
2	6	Yes	Fixed	-159.008	-158.531	-156.135	5 vs. 7	<0.0001
	7	No	Variable	-292.286	-292.812	-292.937	7 vs. 8	<0.0001
	8	No	Fixed	-219.400	-219.585	-219.300	6 vs. 8	0.0055
	9	Yes	Variable	-614.791	-616.701	-656.193	9 vs. 10	<0.0001
3	10	Yes	Fixed	-152.317	-151.565	-151.230	9 vs. 11	<0.0001
	11	No	Variable	-281.614	-282.925	-283.367	11 vs. 12	<0.0001
	12	No	Fixed	-211.831	-211.946	-211.739	10 vs. 12	0.0014

Table C.4: *Model comparison based on AIC and Likelihood Ratio Tests for VAS fractional polynomial models.*

Fractional polyn. degree	Power (μ, ϕ, α)	Model with interactions	Model with Dispersion	AIC logit	AIC probit	AIC cloglog	Comparison on logit link	LRT) (p-value)
1	3;	Yes	Variable	-496.414	-485.227	-484.965	1 vs. 2	<0.0001
	1;	Yes	Fixed	-181.406	-181.043	-177.712	1 vs. 3	<0.0001
	2	No	Variable	-297.253	-297.681	-298.069	3 vs. 4	<0.0001
		No	Fixed	-226.168	-226.287	-226.247	2 vs. 4	0.0534
2	-2,-2;	Yes	Variable	-472.686	-454.751	-442.101	5 vs. 6	<0.0001
	-2,-0.5;	Yes	Fixed	-183.097	-182.232	-179.242	5 vs. 7	<0.0001
	-2,-2	No	Variable	-295.894	-296.290	-296.776	7 vs. 8	<0.0001
		No	Fixed	-228.311	-228.363	-228.365	6 vs. 8	0.1930
3	-1,-1,-1;	Yes	Variable	-396.462	-426.178	-401.852	9 vs. 10	<0.0001
	-2,0.5,3;	Yes	Fixed	-202.855	-201.740	-198.878	9 vs. 11	<0.0001
	-2,-2,3	No	Variable	-290.278	-290.640	-291.108	11 vs. 12	<0.0001
		No	Fixed	-224.559	-224.580	-224.572	10 vs. 12	0.0075

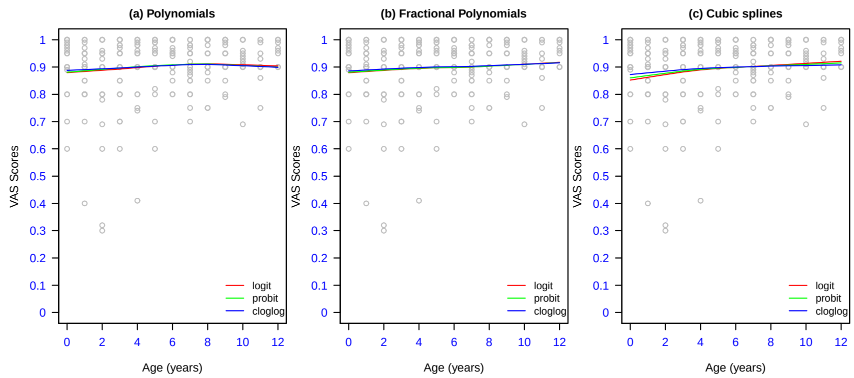

Figure C.2: *Representation of the best fits for the polynomial, fractional polynomial and cubic splines under different link functions in VAS child response.*

C.3 One inflated beta regression in EQ5D-adult group

Table C.5: *Model comparison based on AIC and Likelihood Ratio Tests for EQ5D polynomial models.*

Polynomial Order	Model	Model with interactions	Model with Dispersion	AIC logit	AIC probit	AIC cloglog	Comparison on logit link	LRT) (p-value)
	1	Yes	Variable	101.606	102.636	103.632	1 vs. 2	<0.0001
1	2	Yes	Fixed	750.398	752.429	759.981	1 vs. 3	<0.0001
	3	No	Variable	561.241	561.525	563.174	3 vs. 4	<0.0001
	4	No	Fixed	667.580	667.889	669.690	2 vs. 4	0.0206
	5	Yes	Variable	-114.436	-103.295	-96.961	5 vs. 6	<0.0001
2	6	Yes	Fixed	756.480	759.276	770.118	5 vs. 7	<0.0001
	7	No	Variable	548.966	549.431	553.110	7 vs. 8	<0.0001
	8	No	Fixed	660.829	661.309	665.088	6 vs. 8	0.0071
	9	Yes	Variable	-309.756	-293.786	-292.178	9 vs. 10	<0.0001
3	10	Yes	Fixed	765.365	767.771	779.110	9 vs. 11	<0.0001
	11	No	Variable	548.024	548.420	552.306	11 vs. 12	<0.0001
	12	No	Fixed	660.334	660.807	664.947	10 vs. 12	0.0015

Table C.6: *Model comparison based on AIC and Likelihood Ratio Tests for EQ5D fractional polynomial models.*

Fractional polyn. degree	Power (μ, ϕ, α)	Model with interactions	Model with Dispersion	AIC logit	AIC probit	AIC cloglog	Comparison on logit link	LRT) (p-value)
	-2;	Yes	Variable	93.541	94.657	96.020	1 vs. 2	<0.0001
1	-1;	Yes	Fixed	733.338	735.060	742.197	1 vs. 3	<0.0001
	-2	No	Variable	545.047	545.224	548.416	3 vs. 4	<0.0001
		No	Fixed	654.022	654.195	657.442	2 vs. 4	0.0127
	0.5,0.5;	Yes	Variable	-92.022	-89.126	-99.322	5 vs. 6	<0.0001
2	3,3;	Yes	Fixed	731.248	733.185	742.026	5 vs. 7	<0.0001
	-2,-2	No	Variable	542.267	542.489	545.834	7 vs. 8	<0.0001
		No	Fixed	656.071	656.330	659.832	6 vs. 8	0.0003
	-2,-2,-2;	Yes	Variable	143.606	144.796	150.503	9 vs. 10	<0.0001
3	-2,-2,-2;	Yes	Fixed	729.876	732.250	741.135	9 vs. 11	<0.0001
	-2,-2,-2	No	Variable	525.884	526.157	529.438	11 vs. 12	<0.0001
		No	Fixed	658.498	658.744	662.070	10 vs. 12	0.0084

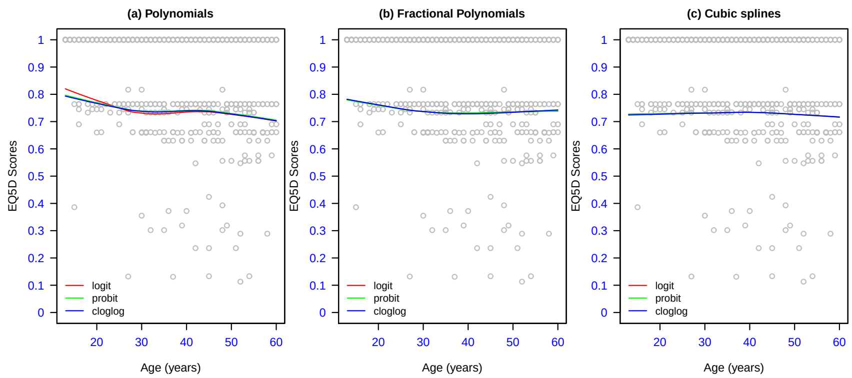

Figure C.3: *Representation of the best fits for the polynomial, fractional polynomial and cubic splines under different link functions in EQ5D adult response.*

C.4 Beta regression in VAS-adult group

Table C.7: *Model comparison based on AIC and Likelihood Ratio Tests for VAS polynomial models.*

Polynomial Order	**Model**	**Model with interactions**	**Model with Dispersion**	**AIC**			**Comparison on logit link**	**LRT) (p-value)**
				logit	**probit**	**cloglog**		
	1	Yes	Variable	-1568.875	-1568.843	-1568.817	1 vs. 2	<0.0001
1	2	Yes	Fixed	-1509.919	-1509.608	-1509.243	1 vs. 3	<0.0001
	3	No	Variable	-1557.017	-1557.044	-1557.060	3 vs. 4	<0.0001
	4	No	Fixed	-1523.665	-1524.109	-1524.683	2 vs. 4	0.1576
	5	Yes	Variable	-1562.164	-1561.602	-1560.899	5 vs. 6	<0.0001
2	6	Yes	Fixed	-1510.131	-1509.228	-1508.052	5 vs. 7	<0.0001
	7	No	Variable	-1557.184	-1557.153	-1557.093	7 vs. 8	<0.0001
	8	No	Fixed	-1526.965	-1527.388	-1527.892	6 vs. 8	0.1081
	9	Yes	Variable	-1557.240	-1556.787	-1556.261	9 vs. 10	<0.0001
3	10	Yes	Fixed	-1507.962	-1506.980	-1505.614	9 vs. 11	<0.0001
	11	No	Variable	-1561.688	-1561.654	-1561.589	11 vs. 12	<0.0001
	12	No	Fixed	-1532.428	-1532.614	-1532.785	10 vs. 12	0.1689

Table C.8: *Model comparison based on AIC and Likelihood Ratio Tests for VAS fractional polynomial models.*

Fractional polyn. degree	**Power** (μ, ϕ, α)	**Model with interactions**	**Model with Dispersion**	**AIC**			**Comparison on logit link**	**LRT) (p-value)**
				logit	**probit**	**cloglog**		
	2;	Yes	Variable	-1570.802	-1570.683	-1570.548	1 vs. 2	<0.0001
1	-1	Yes	Fixed	-1509.270	-1508.869	-1508.370	1 vs. 3	<0.0001
		No	Variable	-1563.454	-1563.432	-1563.383	3 vs. 4	<0.0001
		No	Fixed	-1522.325	-1522.720	-1523.234	2 vs. 4	0.1368
	-2,-2;	Yes	Variable	-1569.608	-1569.648	-1569.888	5 vs. 6	<0.0001
2	-2,-2	Yes	Fixed	-1511.948	-1511.373	-1510.534	5 vs. 7	<0.0001
		No	Variable	-1561.051	-1561.057	-1561.034	7 vs. 8	<0.0001
		No	Fixed	-1532.694	-1532.979	-1533.270	6 vs. 8	0.2189
	3,3,3;	Yes	Variable	-1566.794	-1566.320	-1565.739	9 vs. 10	<0.0001
3	3,3,3	Yes	Fixed	-1512.056	-1510.882	-1509.247	9 vs. 11	<0.0001
		No	Variable	-1565.789	-1565.712	-1565.592	11 vs. 12	<0.0001
		No	Fixed	-1533.234	-1533.403	-1533.548	10 vs. 12	0.0958

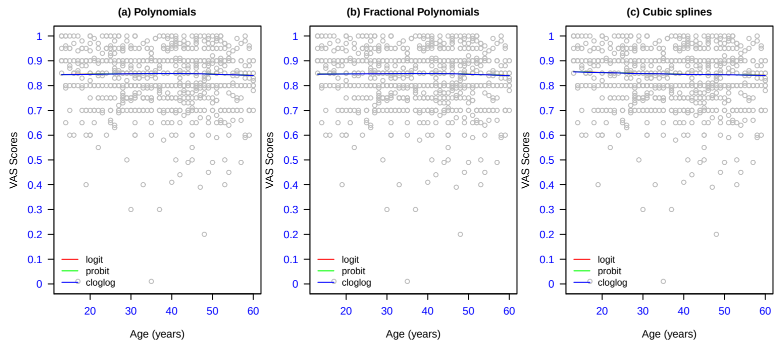

Figure C.4: *Representation of the best fits for the polynomial, fractional polynomial and cubic splines under different link functions in VAS adult response.*

Table C.9: *Comparison of parameter estimates and standard error (in parentheses) for the ML, BC and Bootstrap for the mean and dispersion sub-model.*

Parameter	**ML**	**BC**	**Bootstrap**
location sub-model			
Intercept	1.8890(0.1002)	1.8876(0.1005)	1.8898(0.1488)
age	-0.0061(0.0024)	-0.0061(0.0024)	-0.0061(0.0033)
Illnessy: Yes	-0.6631(0.0918)	-0.6640(0.0926)	-0.6589(0.1040)
normalday: No because sick	-0.8998(0.1549)	-0.9035(0.1601)	-0.8926(0.1580)
normalday: No because other reason	-0.0173(0.0835)	-0.0191(0.0839)	-0.0099(0.0871)
animal: Yes	0.1488(0.0679)	0.1481(0.0682)	0.1432(0.0667)
dispersion sub-model			
Intercept	1.9474(0.0664)	1.9414(0.0663)	1.9593(0.1497)
age	-	-	-
illnessyYes	-0.1498(0.1371)	-0.1654(0.1369)	-0.1032(0.2729)
normalday: No because sick	0.0164(0.2732)	-0.0612(0.2721)	0.1091(0.3760)
normalday: No because other reason	-0.3167(0.1129)	-0.3232(0.1129)	-0.3028(0.1901)
animal: Yes	0.0192(0.0979)	0.0164(0.0978)	0.0208(0.1722)

C.5 One inflated beta regression in EQ5D-elderly group

Table C.10: *Model comparison based on AIC and Likelihood Ratio Tests for EQ5D polynomial models.*

Polynomial Order	**Model**	**Model with interactions**	**Model with Dispersion**	**AIC**			**Comparison on logit link**	**LRT) (p-value)**
				logit	**probit**	**cloglog**		
	1	Yes	Variable	-1307.559	-1279.689	-1219.025	1 vs. 2	<0.0001
1	2	Yes	Fixed	317.147	319.019	347.389	1 vs. 3	<0.0001
	3	No	Variable	158.899	159.394	200.249	3 vs. 4	<0.0001
	4	No	Fixed	277.446	278.437	287.442	2 vs. 4	<0.0001
	5	Yes	Variable	-1130.107	-1125.714	-1088.820	5 vs. 6	<0.0001
2	6	Yes	Fixed	258.953	263.947	284.545	5 vs. 7	<0.0001
	7	No	Variable	100.983	134.755	136.715	7 vs. 8	<0.0001
	8	No	Fixed	253.676	253.839	258.272	6 vs. 8	<0.0001
	9	Yes	Variable	-1113.729	-1135.950	-1164.404	9 vs. 10	<0.0001
3	10	Yes	Fixed	248.946	254.645	280.151	9 vs. 11	<0.0001
	11	No	Variable	102.855	136.697	138.369	11 vs. 12	<0.0001
	12	No	Fixed	255.305	255.457	259.033	10 vs. 12	<0.0001

Table C.11: *Model comparison based on AIC and Likelihood Ratio Tests for EQ5D fractional polynomial models.*

Fractional polyn. degree	**Power** (μ, ϕ, α)	**Model with interactions**	**Model with Dispersion**	**AIC**			**Comparison on logit link**	**LRT) (p-value)**
				logit	**probit**	**cloglog**		
	3;	Yes	Variable	-1324.183	-1333.472	-1207.094	1 vs. 2	<0.0001
1	3;	Yes	Fixed	302.771	305.773	336.669	1 vs. 3	<0.0001
	3	No	Variable	142.025	143.043	185.201	3 vs. 4	<0.0001
		No	Fixed	263.320	263.597	271.314	2 vs. 4	<0.0001
	1,3;	Yes	Variable	-1005.826	-1053.163	-902.204	5 vs. 6	<0.0001
2	0.5,3;	Yes	Fixed	253.342	257.119	294.496	5 vs. 7	<0.0001
	3,3	No	Variable	134.136	134.619	136.522	7 vs. 8	<0.0001
		No	Fixed	253.112	253.271	257.229	6 vs. 8	<0.0001
	-2, 3, 3;	Yes	Variable	-1118.572	-1093.178	-1060.770	9 vs. 10	<0.0001
3	-2, 3, 3;	Yes	Fixed	273.855	277.898	310.089	9 vs. 11	<0.0001
	-2, 3, 3	No	Variable	107.521	138.732	140.155	11 vs. 12	<0.0001
		No	Fixed	257.208	257.380	260.901	10 vs. 12	<0.0001

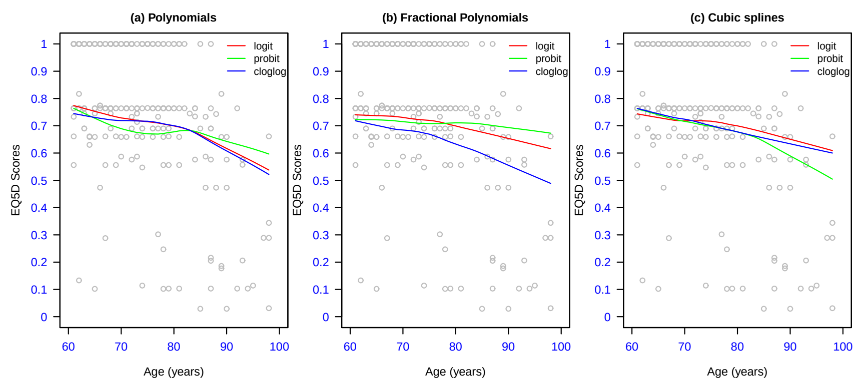

Figure C.5: *Representation of the best fits for the polynomial, fractional polynomial and cubic splines under different link functions in EQ5D response in elderly group.*

C.6 Beta regression in VAS-elderly group

Table C.12: *Model comparison based on AIC and Likelihood Ratio Tests for VAS polynomial models.*

Polynomial Order	**Model**	**Model with interactions**	**Model with Dispersion**	**AIC**			**Comparison on logit link**	**LRT) (p-value)**
				logit	**probit**	**cloglog**		
	1	Yes	Variable	-248.701	-251.196	-248.822	1 vs. 2	<0.0001
1	2	Yes	Fixed	-185.987	-188.476	-193.369	1 vs. 3	<0.0001
	3	No	Variable	-223.859	-224.120	-224.518	3 vs. 4	0.0006
	4	No	Fixed	-214.964	-217.607	-221.994	2 vs. 4	0.3685
	5	Yes	Variable	-243.945	-265.463	-273.260	5 vs. 6	<0.0001
2	6	Yes	Fixed	-193.672	-196.314	-201.342	5 vs. 7	<0.0001
	7	No	Variable	-221.668	-221.841	-222.071	7 vs. 8	0.0007
	8	No	Fixed	-213.615	-216.520	-222.030	6 vs. 8	0.0761
	9	Yes	Variable	-242.527	-264.429	-271.611	9 vs. 10	<0.0001
3	10	Yes	Fixed	-192.004	-194.678	-199.693	9 vs. 11	<0.0001
	11	No	Variable	-220.628	-220.756	-220.743	11 vs. 12	0.0007
	12	No	Fixed	-212.806	-215.381	-220.150	10 vs. 12	0.0894

Table C.13: *Model comparison based on AIC and Likelihood Ratio Tests for VAS fractional polynomial models.*

Fractional polyn. degree	**Power** (μ, ϕ, α)	**Model with interactions**	**Model with Dispersion**	**AIC**			**Comparison on logit link**	**LRT)** **(p-value)**
				logit	**probit**	**cloglog**		
	3;	Yes	Variable	-221.533	-222.818	-296.292	1 vs. 2	<0.0001
1	-2	Yes	Fixed	-202.841	-204.998	-209.412	1 vs. 3	0.0044
		No	Variable	-224.116	-224.513	-225.301	3 vs. 4	0.0006
		No	Fixed	-215.504	-218.328	-223.268	2 vs. 4	0.1780
	3,3;	Yes	Variable	-374.723	-368.902	-346.357	5 vs. 6	<0.0001
2	3,3	Yes	Fixed	-177.955	-181.836	-187.385	5 vs. 7	<0.0001
		No	Variable	-222.208	-222.260	-222.124	7 vs. 8	0.0005
		No	Fixed	-213.514	-216.393	-221.912	6 vs. 8	0.0719
	3,3,3;	Yes	Variable	-320.766	-290.399	-281.196	9 vs. 10	<0.0001
3	-2,-2,-2	Yes	Fixed	-195.626	-199.093	-204.156	9 vs. 11	<0.0001
		No	Variable	-224.561	-224.352	-223.351	11 vs. 12	0.0002
		No	Fixed	-213.331	-215.871	-220.437	10 vs. 12	0.0303

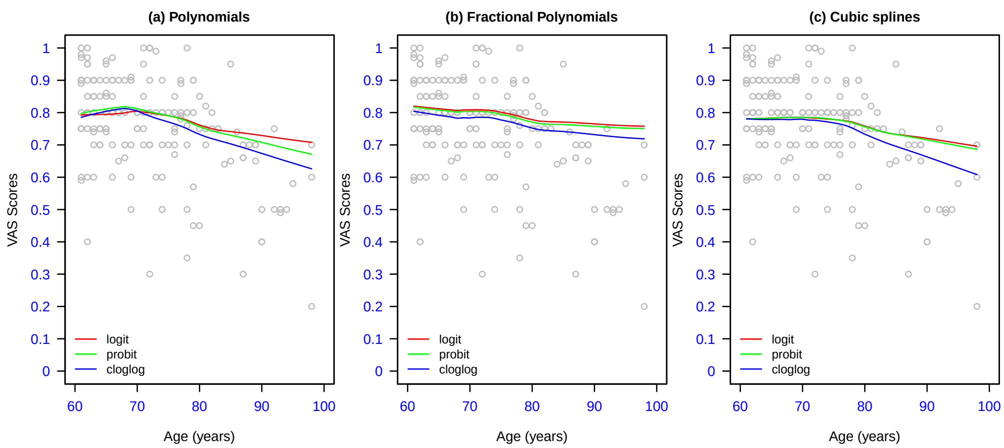

Figure C.6: *Representation of the best fits for the polynomial, fractional polynomial and cubic splines under different link functions in VAS response in elderly group.*

Table C.14: *Comparison of parameters estimates and standard error for the ML, BC and Bootstrap for the mean and dispersion sub-model.*

Parameter	**ML**		**BC**		**Bootstrap**	
location sub-model	Estimates	Std. error	Estimates	Std. error	Estimates	Std. error
Intercept	3.4891	0.4667	3.4891	0.4667	3.5019	0.8996
age	-0.0265	0.0064	-0.0265	0.0064	-0.0266	0.0789
Illnessy: Yes	-0.3950	0.1227	-0.3950	0.1227	-0.3968	0.1294
education: higher technical/secondary	-0.2503	0.1714	-0.2503	0.1714	-0.2552	0.2957
education: Lower technical/secondary	-0.2103	0.2364	-0.2103	0.2364	-0.2063	0.2367
education: None/Primary	-0.6172	0.2162	-0.6172	0.2162	-0.6212	1.4623
education: Vocational	-0.6916	0.2207	-0.6916	0.2207	-0.6952	0.2296

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Assessing the health related quality of life in Flanders and the possible factors influencing it

Richting: **Master of Statistics-Biostatistics**
Jaar: **2014**

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,

Rungo, Antonio

Datum: **12/12/2013**