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

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

Maastricht University

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

2010
2011

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

Identifying determinants of life expectancy in the EU

Promotor :
Prof. dr. Niel HENS

Promotor :
Prof. PH. BEUTELS
Prof. ADRIAAN BLOMMAERT

Francis Batomen

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

Maastricht University

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

2010
2011

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

Identifying determinants of life expectancy in the EU

Promotor :
Prof. dr. Niel HENS

Promotor :
Prof. PH. BEUTELS
Prof. ADRIAAN BLOMMAERT

Francis Batomen

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

Maastricht University

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

Dedication

To Michel

To Hermione

To my parents

To the Lord

Abstract

Life expectancy is the expected number of years of life remaining at a given age. It is a very important measure, a good proxy for health status and also it appears among the indicators used to assess the development of a country. Life expectancy is also used in public policy planning.

This report looks for the determinants of life expectancy at birth and at age 65 in the EU over the period 1999-2007. Due to the difference in health status between men and women their life expectancies were modelled separately. Potential determinants present in the data set used are numerous and therefore are more likely to be concerned by multicollinearity. There are also time-dependent. Penalized Generalized Estimating Equations (PGEE) with independence working correlation was then applied to select among the covariates those which significantly explain the life expectancy in the EU.

Among different penalty functions which could be added to the score equations of Generalized Estimating Equations (GEE), elastic-net and SCAD-L_2 have useful properties, including sparsity which is necessary to perform variable selection. Through cross-validation SCAD-L_2 proved to be the optimal penalty for both models of each gender.

It emerges from this study that in the EU some factors affect life expectancy regardless the gender and the age, namely the GDP, the population size, the practising physician density and the variability of absolute humidity. The first three factors have a positive effect and the last one a negative effect on life expectancy. Indeed, considerable differences appear between males and females. It is also noted that the determinants of female life expectancy at birth and age 65 are almost the same whereas those of male life expectancy at birth and age 65 differ slightly.

Key Words: Life expectancy, variable selection, penalized GEE, Cross-validation

Acknowledgments

This thesis would never have been started without the help of a number of persons. First, I am thankful to Prof. Dr. Herbert Thijs for providing me with this project. I am very grateful to my wonderful internal supervisor Prof. Dr. Niel Hens and external supervisors Prof. Dr. Philippe Beutels and Adriaan Blommaert for their immense availability, advices, comments, suggestions and encouragements during the writing of this thesis. Thank you for taking time from your extremely busy schedule to help me. Special thanks to Adriaan Blommaert!

It is also an opportunity for me to thank all lecturers of Master Biostatistics of the University of Hasselt. All your courses were very useful to me for the accomplishment of this work.

My sincere thanks to my family who has agreed to spend several months without my presence. Special thanks to my wife for her understanding and incessant moral support during this training.

Contents

Abbreviations

AIC:	Akaike Information Criterion
AR(1):	First-order Autoregressive
BIC:	Bayesian Information Criterion
CHERMID:	Centre for Health Economics Research and Modeling Infectious Diseases
CO2:	Carbon dioxide
ESAC:	European Surveillance of Antimicrobial Consumption
EU:	European Union
EUROSTAT	European Statistics
FCCM:	Full Covariate Conditional Mean
GDP:	Gross Domestic Product
GEE:	Generalized Estimating Equations
IDR:	Intersection Deletion Rule
LASSO:	Least Absolute Shrinkage and Selection Operator
LMM	Linear Mixed Model
LOOCV:	Leave-One-Out Cross-Validation
MAR	Missing At Random
MCAR:	Missing Completely At Random
MVDR:	Majority Vote Deletion Rule
MSE:	Mean Square Error
OECD:	Organization for Economic Cooperation and Development
OLS:	Ordinary Least Squares
PGEE:	Penalized Generalized Estimating Equations
SCAD:	Smoothly Clipped Absolute Deviation
SE:	Standard Error
RSS:	Residual Sum of Squares
UDR:	Union Deletion Rule
WHO:	World Health Organization
WGEE:	Weighted Generalized Estimating Equations

Chapter 1: Introduction

1.1 Background of study

The World Health Organization (WHO) defines the life expectancy as the average number of years persons can expect to live, if in the future they experience the current age-specific mortality rates in the population [42]. It is a common measure of population health in general, and is often used as a summary measure to compare different populations. It summarizes the mortality pattern and also gives an idea on the quality of healthcare delivery and the ageing of population. Life expectancy is used in public policy planning. Indeed, it appears among the important measures of the economic well-being of a nation [28]. It is computed for each age but for comparison between countries the life expectancies at birth, at age 60 and 65 for males and females (separately) are most of the time used.

The increase of the life expectancy in some European countries (and other developed countries around the world) during the 20th century results from the improvements in housing, sanitation, and nutrition; the control of infectious diseases and maternal mortality; and the advent of antibiotics [37]. In 1900, one in five infants died before age 10. Life expectancy for those who survived to age 10 was about 60. By 1940, over 90% of infants survived to age 10. Since 1940, mortality reductions have shifted to older ages [43]. That is partially explained by the rising living standards, the improved access to better education, including for women. Between 1940 and 1960, infectious disease mortality continued to decline with the development of (antibioticics) sulfa drugs in the 1930s and penicillin in the 1940s. Later in the century, new medical technology has been developed for treating cardiovascular disease, one of the leading killers which appeared after the decline of infectious disease. In 2010 the life expectancy in developed countries is more than 80 years [44] and is expected to increase in 2011 [45].

Even if the European continent is among the richest regions of the Earth and its economy among the largest [46], a large variation of wealth exists among its countries. The richer states tend to be in the West; some of the Eastern economies are still emerging from the collapse of the Soviet Union and Yugoslavia [46]. For instance between 1980 and 2005, the trends in life expectancy show a still growing gap between western and eastern parts of Europe; with longest life expectancy recorded in the western part [35]. Concerning the mortality, the east-west inequality is also pronounced. For instance, infant mortality is higher in the eastern part of the region between 1980 and 2005, with the highest levels in some south-eastern countries [35].

1.2 Sex differences in life expectancy

In all populations, females tend to outlive males [1]. The female life expectancy is considerably higher than that of men. That difference in life expectancy with respect to the gender is ascribed to a gender differential in health status. Men have higher rates of death (mortality) and more serious and chronic illnesses (morbidity) than women [26]. For instance, men are more likely to die from cardiovascular disease, which is acute, and often fatal at a relatively young age (40-60 years). According to Abdulraheem *et al.* [1], the differences in incidence of illness between males and females are attributed to three elements: (1) the biological factors because "*boys are more likely to have problems in utero or early infancy due to having a single X chromosome. Occasionally, reference may be made to the hormonal differences between men and women as contributory.*" (2) The behavioural/cultural factors: men's "risk-taking" behaviours and the "male role" are the real causes of poor health [15, 16]. In other words, the differences in lifestyle, behaviours and attitudes between men and women determine to a large extent their different longevities. (3) The material/structural factors: Here, social factors (such as employment patterns, income, educational opportunities, government policies, provision of health services, etc.) are considered as determinants of health outcomes. For instance it is noted that structural factors are the main cause of women's health problems [1].

1.3 Determinants of life expectancy (according to empirical studies)

As we have noticed so far, life expectancy and health are strongly related. The increase of life expectancy in developed countries is considered as the result of health care improvement. In developing countries where the infant and maternal mortality rates are high, and diseases like malaria and tuberculosis rage (Sub-Saharan Africa for instance) the life expectancy at birth is lower [10]. The life expectancy is therefore used as a proxy of health status in the estimation of the health production function in most studies. The health production function mostly defined is derived from the Grossman [17] theoretical model that treats social, economic, and environmental factors as inputs of the production system. The outcome is the health status and the covariates are: *"nutrient intake, income, consumption of public goods, education, time devoted to health related procedures, initial individual endowments like genetic makeup, and community endowments such as the environment"* [10].

For the estimation of health production function for Sub-Saharan Africa Fayissa and Gutema [10] used the GDP per capita, health expenditure per capita, food availability, illiteracy rate, population, adult alcohol consumption per capita, urbanization rate, and CO2 emissions to explain the health status proxied by life expectancy at birth.

Shaw *et al.* [32] used the income (GDP per capita), pharmaceutical consumption, non-pharmaceutical health care consumption, percentage of the population 65 years of age and older, health care consumption, alcohol consumption, smoking behavior, fruit and vegetable consumption, fat consumption and a spatial factor as determinants of the life expectancies for males and females at ages 40, 60 and 65 in OECD. In the specified model the logarithmic transformation is applied to all continuous variables. In particular, the wealth of a country (GDP) is assumed to have a strong non-linear influence on the life expectancy of its inhabitants [31]. Therefore in some life expectancy production function the logarithmic transformation is applied to the GDP [10, 32].

As summary demographic and socio-economic factors, health resources, lifestyle and environmental factors are the main determinants of life expectancy. In most studies only few variables are used to characterize each group of factors: diet, alcohol and tobacco consumption as lifestyle factors, income and education as socio-economic factors, etc. However, Poudyal *et al.* [30] included many variables in the regression model to examine the effect of natural resource amenities on human life expectancy in the United States. They extended the existing model of life expectancy production function by considering 9 demographic and socio-economic factors, 8 medical facilities and risk factors and 12 variables to characterize the natural amenities and outdoor recreation resources. The outcome is the life expectancy at birth.

1.4 Statement of the Problem

In the framework of disease prevention and control, the Centre for Health Economics Research and Modeling Infectious Diseases (CHERMID) of the University of Antwerp gathered data in EU countries over the period 1999-2007. The database established contains different groups of variables: agricultural factors, burden disease, cultural and perception of illness, education and knowledge about antibiotics, health care system, socio-economic factors and demographic factors among which the life expectancies at birth, 60 years and 65 years for males and females. The goal of this thesis is to select variables which may explain the life expectancy in the EU. The three important issues in the database are multicollinearity due to the large number of variables, time-dependent covariates and longitudinal nature of the responses.

- ***Multicolinearity and variable selection***

In multiple regression, multicollinearity occurs when two or more predictors are highly correlated and provide redundant information about the response [47]. One consequence of multicollinearity is the increase of the standard errors of the regression coefficients [21]. Multicollinearity can cause strange results when attempting to study how well individual independent variables contribute to an understanding of the dependent variable. In general, multicollinearity can cause wide confidence intervals and produce parameter estimates of implausible magnitude or incorrect sign. Parameter estimates become very sensitive to the addition or deletion of observations. In truly extreme case it may prevent the determination of the numerical solution of the model [27]. Therefore, in the analysis, it is necessary to deal with multicollinearity when it is present in the data.

In order to handle multicollinearity ridge regression has been proposed [19, 20]. That model consists in penalizing the least squares or residual sum of squares (RSS) of the linear regression model with the L$_2$-norm $\sum_{j=1}^{p} \beta_j^2$. Afterwards the L$_1$-norm $\sum_{j=1}^{p} |\beta_j|$ has also been applied as penalty to the RSS. The resulting model is termed LASSO [33] and has a property that allows variable selection contrary to the ridge penalty which deals only with multicollinearity. In the same vein, different penalties have been proposed to concomitantly handle multicollinearity and perform variable selection.

- ***Longitudinal data, time-dependent covariates and variable selection***

Data to analyze are correlated since the response variables are longitudinal. The repeated measurements over time of a life expectancy are expected to be correlated within a country. Linear regression is not suitable to model such data due to the fact that it is based on the assumption of independence of the observations. In addition, the scientific focus of this study is less on the individual’s response but, more on the population-averaged response. We are interested by the effect of covariates on the average life expectancy in the EU (population-averaged response) not on the life expectancy in each country (individual’s response). Therefore a marginal model is suitable to address our research question: “what are the determinants of life expectancy in the EU?”

Generalized Estimating Equations (GEE) proposed by Liang and Zeger in 1986 [22, 38] is the most used marginal model because many software packages offer procedures or functions to fit it. However, in 1994 Pepe and Anderson [29] drew attention on the key condition for its use by noting that the consistency of GEE parameter estimates depends on the type of covariates and the type of association among observations (Working correlation) considered. With time-independent covariates GEE parameter estimates are consistent whatever specified working correlation; that is not always the case with time-dependent covariates [29]. In CHERMID database almost all potential determinants of life expectancy vary over time. This warns us about the choice of working correlation structure for the marginal model which seems appropriate for this study.

The penalty functions aforementioned which are applied to the RSS can be generalized to the likelihood (deviance precisely) in order to proceed to variable selection [7], however in GEE approach the likelihood is not known. The estimating equations are derived without full specification of the joint distribution of the observation of a subject [5]. Thus, variable selection is done in GEE analysis by penalizing the score equations [8, 14].

1.5 Objective of the study

The main goal is to apply penalized GEE in order to select determinants of life expectancy in the EU from the database established by the CHERMID. This study focuses on the life expectancies at birth and at age 65 for males and females separately. Broadly the structure of this report is as follows. The second chapter describes the subset of the database used for the analysis. Different penalty functions and their properties are first of all described before their application to GEE. The third chapter starts by the exploration of data, the assessment for the presence missingness which characterizes longitudinal studies and ends by the final models that answer the research question. The last chapter (4) lists the limitations of the methodology used in this report, and thus sketches some tracks for the improvement of this work and also for future research.

Chapter 2: Data and methodology

2.1. Data

Data analyzed in this report is a subset of the CHERMID database, which has already been used in the ESAC3 (European Surveillance of Antimicrobial Consumption) project to investigate determinants of antibiotic use. CHERMID collected data from generic databases (such as Eurostat and OECD health data), pertaining to the period 1999-2007. The database contains different groups of variables such as agricultural factors, demographics factors, education, culture, healthcare system and socio-economic factors.

In the present study, we try to identify the determinants of life expectancy at birth and at 65 years of age for female (FemLifeExpBirth and FemLifeExp_65) and male (MaleLifeExpBirth and MaleLifeExp_65) separately. All these response variables are considered as continuous. Table A.1 in the appendix describes the potential determinants of life expectancy in the EU countries available in the database. Almost all covariates are continuous except the binary variable Guidelines that indicates if there are any guidelines for Pulmonologists or General Practitioners.

The peculiarity of the data set to analyze is the longitudinal structure of each of the 4 responses, the covariates which are time-dependent and the large number of covariates (53) that undoubtedly, will pose the multicollinearity problem.

2.2. Methodology

2.2.1. A muticollinearity remedial measure and its extensions

Consider the usual linear regression model $Y = \mathbb{1}\beta_0 + X\beta + \varepsilon$; where the response Y is an $N \times 1$ vector, $\mathbb{1} = (1, \dots, 1)^t$, β_0 the intercept, X an $N \times p$ matrix of covariates, β a $p \times 1$ vector of coefficients and $\varepsilon \sim N(0, \sigma^2 I_N)$. The estimates of the coefficients β_0 and β are determined using the Ordinary Least Squares (OLS) method. However, the presence of multicollinearity may have an undesirable effect on them [21]. Different methods have been proposed to deal with multicollinearity in regression analysis; some of them consist in combining, transforming or dropping covariates [21] others consists in modifying the regression method by considering principal components regression [18] or ridge regression [19, 20].

In principal components regression instead of regressing the predictor variables on the dependent variable directly, the principal components (index variables) that are linear combinations of the predictor variables are used [18]. Multicollinearity problem is solved because principal components are mutually uncorrelated.

As recall, in the presence of severe multicollinearity, the standard errors of the regression coefficients are inflated, making it very difficult to detect the partial effect of regressors and making interval estimates very imprecise [21]. In such situation biased estimators, but that are substantially more precise than the unbiased estimators may be preferable. Ridge regression modifies the least squares method so that biased estimates are allowed that may result in more precision (smaller variability for the estimators). The ridge coefficients minimize a penalized Residual Sum of Squares (RSS),

$$\tilde{\beta}^{ridge} = \underbrace{\arg min}_{\beta}\left\{\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2\right\}$$

equivalently $\tilde{\beta}^{ridge} = \underbrace{\arg min}_{\beta}\left\{\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2\right\}$ subject to $\sum_{j=1}^{p} \beta_j^2 \leq t$

where λ or t controls the model complexity (regularization parameter). $P(\beta) = \lambda \sum_{j=1}^{p} \beta_j^2$ is a L_2 penalty. In data mining linear regression model is recognized as often exhibiting high variance and so doesn't reduce the prediction error [18]. Ridge regression appears as a shrinkage method that proposes to trade off unbiasedness in favor of greater stability. Another shrinkage method similar to ridge proposed by Tibshirani [33] is the least absolute shrinkage and selection operator (LASSO). It minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant $\left(\sum_{j=1}^{p}|\beta_j| \leq t\right)$. The LASSO is a member of the penalized least squares family with L_1 penalty: $P(\beta) = \lambda \sum_{j=1}^{p}|\beta_j|$. The bridge regression has also been proposed with L_q penalty $P(\beta) = \lambda \sum_{j=1}^{p}|\beta_j|^q$ (where q is some number between 0 and 2) to deal with collinearity in regressions [12, 13]. Ridge and LASSO are special cases of bridge penalty; the former corresponds to $q = 2$ and the latter to $q = 1$. Zou and Hastie [40] combined linearly these two penalties to form the elastic-net penalty function $P(\beta) = \lambda \sum_{j=1}^{p}\left(\alpha|\beta_j| + (1-\alpha)\beta_j^2\right)$, with α between 0 and 1. Fan and Li [9] defined the Smoothly Clipped Absolute Deviation (SCAD) penalty function as $P(\beta) = \lambda \sum_{j=1}^{p} p_j(\beta_j)$ such that

$$p_j(\beta_j) = \begin{cases} \lambda|\beta_j| & if\ \ 0 \leq |\beta_j| < \lambda \\ \frac{(a^2-1)\lambda^2-(|\beta_j|-a\lambda)^2}{2(a-1)} & if\ \ \lambda \leq |\beta_j| < a\lambda \\ \frac{(a+1)\lambda^2}{2} & if\ \ |\beta_j| \geq a\lambda \end{cases} \qquad \text{With } a > 2.$$

Xie and Zeng [36].proposed the SCAD-L$_2$ penalty which is the linear combination of SCAD penalty and ridge penalty: $P(\beta) = \lambda \sum_{j=1}^{p} (\alpha \times p_j(\beta_j) + (1-\alpha)\beta_j^2)$. The penalty functions are sketched in figure 1 [7, 9, 36]. We can notice the convexity of L$_2$-norm. Elastic-net penalty function will fall in between ridge and LASSO if the graphs are superimposed. SCAD-L$_2$ penalty function is almost dominated by the L$_2$-norm and looks like elastic-net. As in ridge regression presented above, all these penalties are used to perform penalized least squares analysis: $\tilde{\beta}^{penalty} = \underbrace{\arg min}_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + P(\beta) \right\}$

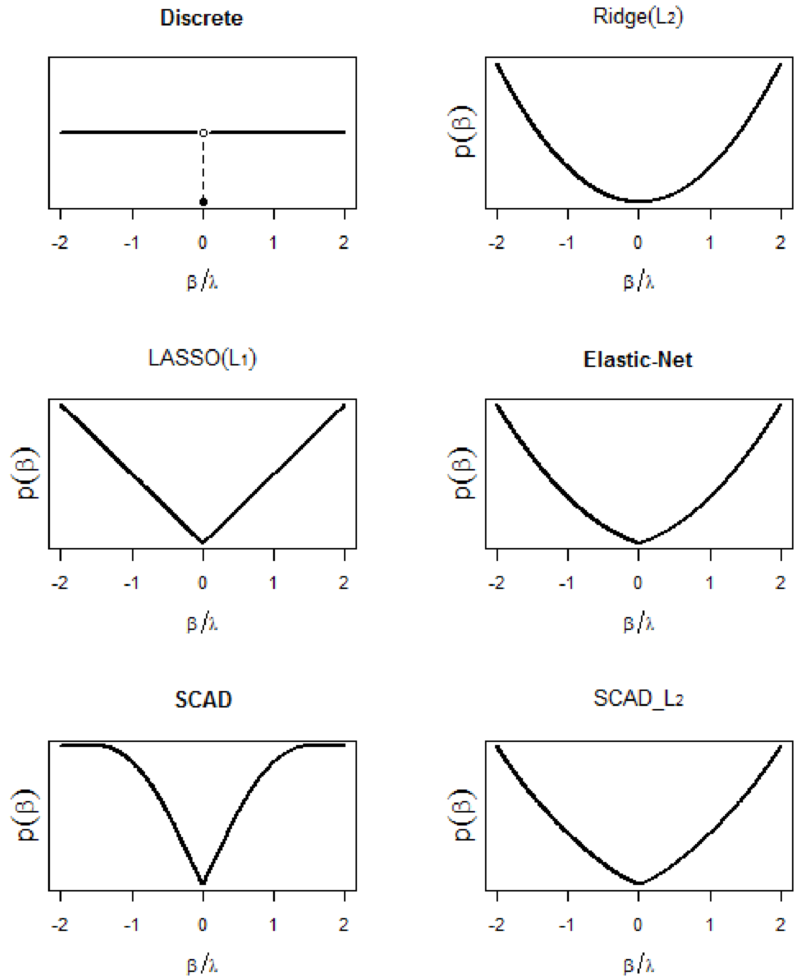


Figure 1: Penalty functions

2.2.2. Variable selection and penalization

In regression analysis parsimony is an important issue. When the number of predictors is large we would like to determine a smaller subset that put more light on the relationship between the response and covariates making the model interpretation simpler. Usually stepwise procedure (forward selection or backward elimination) and best subset selection are used for that purpose [18, 23]. These variable selection methods are based on the optimization of a criterion of model goodness like, residual sum of squares, adjusted R^2, Mallow's Cp, the Akaike information criterion (AIC), or the Bayesian information criterion (BIC). They become computationally expensive (when the number of variables is too large) and ignore stochastic errors present in the variable selection process [4, 9]. In addition to the fact that subset selection exhibits high variance and so does not reduce the prediction error of the full model [18], it is also criticized for its inherent discreteness (variables are either retained or discarded) that may provides extremely variable models [33]. Subset selection method corresponds to L_0 penalty function $\sum_{j=1}^{p}\left|\beta_j\right|^{0+}$, where we define f^{0+} as $I\{f \neq 0\}$ and $I(.)$ is the indicator function [7, 13, 33].

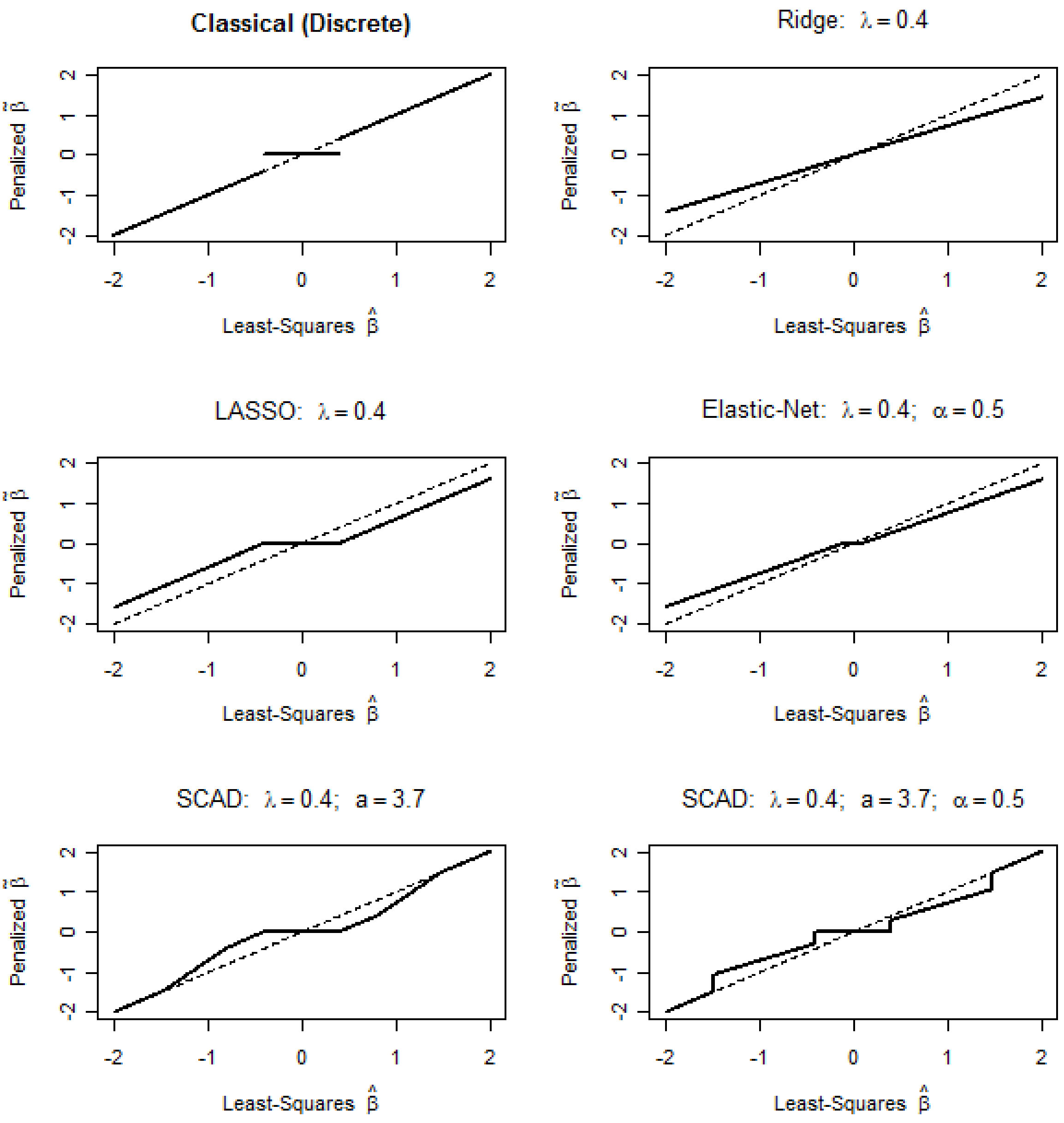
Penalized least squares methods that are continuous and do not suffer as much from high variability have been proposed for variable selection. LASSO is the simpler one which shrinks some coefficients and set others to zero (sparse representation). However the shrinkage of some coefficients may be excessive, introducing too much bias. Ridge penalty shrinks the coefficients but does not set any of them to zero [33]. Ridge regression keeps all regressors in the model and therefore cannot produce a parsimonious model event if it improves its prediction performance through a bias-variance trade-off. We can therefore notice that penalize the RSS is not enough to ensure variable selection, moreover shrinkage may be unnecessary. So, a good penalty function should satisfy three main properties according to Fan and Li [9]:

1. ***Unbiasedness****: The resulting estimator is nearly unbiased when the true unknown parameter is large to avoid unnecessary modeling bias.*
2. ***Sparsity****: Small estimated coefficients (in absolute value) are set to zero to reduce model complexity (singularity at the origin).*
3. ***Continuity****: The resulting estimator is continuous with respect to the unbiased estimate to avoid instability in model prediction.*

The SCAD penalty possesses all these properties. Another advantage of SCAD is that it possesses the “Oracle properties” defined by Fan and Li [9]; meaning in simpler words according to Xie and Zeng [36] that “*the probability of selecting the right set of variables (with* *nonzero coefficients) converges to 1 and that the estimators of the nonzero coefficients are asymptotically normal with the same means and covariances as if the zero coefficients were known in advance,*" or in other words that "*the asymptotic bias, variance, and distribution of the resulting estimate are the same as if the correct subset were known in advance,*" according to Dziak [7]. It is necessary to mention that continuity is obtained with SCAD penalty when a is larger than 2. Otherwise the SCAD penalty would converge pointwise to the L_0 penalty and lose its stability [7]. Ridge shares only the continuity property with SCAD.

Even if LASSO (L_1 penalty) accomplishes automatic variable selection and continuous shrinkage, it is criticized for the fact that when the number of observations N is lower than the number of covariates p it will select at most N variables and also for the fact it tends to select only one variable from a group of variables highly correlated [40]. The L_2 penalty term in the Elastic-net penalty confers to the latter the "grouping effect" property (which lacks to SCAD penalty), meaning the ability to drop or select groups of correlated variables. The L_1 penalty term gives to Elastic-net the LASSO characteristics (simultaneous automatic variable selection and continuous shrinkage). As regards the L_q penalty their properties intermediate between L_0 and L_2 penalties.

The behaviour of the penalty functions for an orthogonal design (orthogonal columns of X) are summarized in table A.3 in the appendix where the penalized estimates $\tilde{\beta}_j$ are simple functions of the least squares estimates $\hat{\beta}_j$ [7, 9, 18, 36, 40]. The properties of these penalties are shown in figure 2. We note the undesirable properties of some of them namely the discontinuity of L_0-norm; ridge which does not set coefficients to zero and then does not give sparse model. LASSO and elastic-net introduce considerable bias. Fortunately, with SCAD and SCAD-L_2 large coefficients are unbiased.



Classical (Discrete)

Ridge: $\lambda = 0.4$

LASSO: $\lambda = 0.4$

Elastic-Net: $\lambda = 0.4$; $\alpha = 0.5$

SCAD: $\lambda = 0.4$; $a = 3.7$

SCAD: $\lambda = 0.4$; $a = 3.7$; $\alpha = 0.5$

Figure 2: Penalized Estimators as Functions of the Least-Squares Estimator in the Orthogonal Case

2.2.3. Generalized Estimating Equations (GEE)

This study looks for the factors which may affect the life expectancy in the EU. It is less interested by the determinants of life expectancy in a specific country of the EU compared to others (heterogeneity among countries). A population-averaged approach is then suitable to model the life expectancy. Moreover data are correlated because of the longitudinal nature of the response variable [34]. We expect the life expectancy values (of different measurement times) within each country to be more correlated than across countries.

Different likelihood-based marginal models have been proposed mostly for discrete longitudinal data [5, 25], however there are computationally expensive even in situations with a small number of repeated measurements for each subject [5] since they require the specification of many parameters. GEE was proposed as marginal model by extending the quasi-likelihood approach to longitudinal and clustered data [5]. The response variable follows a distribution that belongs to the exponential family. Therefore GEE is used to model not only repeated binary data, but also count, ordinal and continuous data [11].

Let consider y_{ij} the response for the subject i at time j $(i = 1, \dots n,$ and $j = 1, \dots t_i)$, $x_{ij} = (x_{ij1}, \dots, x_{ijp})'$ the corresponding $p \times 1$ vector of covariates and $\beta = (\beta_1, \dots, \beta_p)$ a $p \times 1$ vector of unknown parameters characterizing how the response distribution depends on the explanatory variables. As marginal model GEE has three-part specification [5]. In the first step the marginal response $\mu_{ij} = E(y_{ij}|x_{ij})$ is related to a linear combination of the covariates, $g(\mu_{ij}) = x'_{ij}\beta$, where $g(.)$ is the link function. Afterwards the variance of y_{ij} is expressed as function of the mean, $Var(y_{ij}|x_{ij}) = V(\mu_{ij})\phi$, where $V(.)$ is the variance function and ϕ an unknown scale parameter. The last specification is the within-subject association among the vector of repeated responses assumed to be function of an additional set of association parameters α $(R_i(\alpha) = corr(y_i) = R(\alpha))$. $R(\alpha)$ is called working correlation matrix and approximates the average dependence among repeated observations over subjects. Different working correlation structures have been suggested, independence, exchangeability, first-order autoregressive (AR(1)) and unstructured [5, 25]. Fortunately, GEE yields consistent estimates of the regression coefficients and their variances even under misspecification of the covariance matrix structure (With time-independent covariates, indeed).

The final step of GEE is the estimation of the parameter vector β and its covariance matrix. For the i-th subject, let A_i be the $t_i \times t_i$ diagonal matrix with $V(\mu_{ij})$ as the j-th diagonal element. The working covariance matrix for $y_i = (y_{i1}, \dots, y_{it_i})'$ is $V_i(\alpha) = \phi A_i^{1/2} R_i(\alpha) A_i^{1/2}$. The parameter vector β is estimated by solving the following score equations:

$$S(\beta) = \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta}\right)' [V_i(\hat{\alpha})]^{-1} (y_i - \mu_i) = 0 \qquad (1)$$

The key property for the consistency of $\hat{\beta}$, solution to (1) is $E(S(\beta)) = 0$ [5, 11, 29]. That is true if the marginal expectation μ_{ij} is equal to the partially-conditional expectation $E(y_{ij}|x_{i1}, \dots, x_{it_i})$:

$$E(y_{ij}|x_{ij}) = E(y_{ij}|x_{i1}, \dots, x_{it_i}) \qquad (2)$$

Diggle et *al.* [6] refer to the condition (2) as the Full Covariate Conditional Mean (FCCM) assumption. Pepe and Anderson [29] showed that this condition is not always satisfied with time-dependent covariates, but with time-independent covariates. They concluded that fitting a marginal model to longitudinal data requires that either the FCCM assumption is verified or that working independence GEE is used [29].

2.2.4. Penalized GEE (PGEE)

In sections 2.2.1 and 2.2.2 we defined different penalty functions and some of their properties. In linear regression, they improve the OLS estimator in terms of mean squared error and prediction error and some of them contribute to variable selection. They can be generalized to likelihood-based models by replacing the objective function RSS by the model deviance (precisely -2 ×log-likelihood) [14]. However, the extension to GEE is not straightforward due to the unavailability of the joint likelihood function. To deal with collinearity in longitudinal data the penalized score equations approach is adopted:

$$\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta}\right)' [V_i(\hat{\alpha})]^{-1}(y_i - \mu_i) - n\dot{P}(\beta) = 0$$

where the first term corresponds to (1) and $\dot{P}(\beta) = \partial P(\beta)/\partial \beta$ is the vector derivative of the penalty function $P(\beta)$. Fu [14] applied the bridge penalty to GEE with $q > 1$. Subsequently Dziak and Li [8] applied the SCAD penalty and mentioned that the oracle property is maintained for penalized GEE with this penalty function.

In this study we considered GEE with independence working correlation assumption due to the fact that all covariates vary randomly with the time. In order to try out the penalties defined in section 2.2.1, we considered the following form for the penalty function $P(\beta)$ applied to GEE: $P(\beta) = \lambda \sum_{j=1}^{p} \left(\alpha f_{L1}(\beta_j) + (1-\alpha) f_{L2}(\beta_j)\right)$

where $f_{L1}(.)$ is the part of the penalty function that provides sparsity of the resulting estimator, $f_{L2}(.)$ is the part of the penalty function that provides the grouping-effect (has to be a convex function) and $\alpha \in [0,1]$. For instance by taking $f_{L1}(\beta_j) = |\beta_j|$ and $f_{L2}(\beta_j) = \beta_j^2$ the $P(\beta) = LASSO$ for $\alpha = 1$ and $P(\beta) = Ridge$ for $\alpha = 0$. Elastic-net penalty is obtained by taking an α value different from 0 and 1; in fact the SCAD penalty is applied for $\alpha = 0$. This general form of the penalty function makes easier the application of the $SCAD\text{-}L_2$ penalty function proposed by Xie and Zeng [36]. The $SCAD\text{-}L_2$ penalty function is a linear combination of the L_2 norm and the SCAD function. It is obtained by taking $f_{L1}(\beta_j) = p_j(\beta_j)$, $f_{L2}(\beta_j) = \beta_j^2$ and an α value different from 0 and 1. In addition to the three properties satisfied by the SCAD penalty, $SCAD\text{-}L_2$ penalty function also satisfies the grouping-effect property [36], which is important for variable selection in case of high multicollinearity.

2.2.5. Selection of tuning parameters

The elastic-net and SCAD-L_2 penalties retained for this study have the two tuning parameters λ and α. SCAD-L_2 has an additional one: a, like SCAD penalty function. However, it was shown that $a = 3.7$ is approximately optimal [9] for SCAD penalty. Moreover, Fan and Li [9] recommended to use cross-validation or generalized cross-validation method to determine tuning parameters for SCAD regularization. Therefore, for this study we considered $a = 3.7$ and cross-validation method to determine the best pair $(\alpha; \lambda)$ over the two-dimensional grid (for model selection).

Cross-validation is a technique for assessing how the results of a statistical analysis will generalize to an independent data set. In k-fold cross-validation, the original data is first randomly partitioned into k equally (or nearly equally) sized segments or folds. Subsequently k iterations of training and validation are performed such that, within each iteration a different fold of the data is held-out for validation while the remaining $k - 1$ folds are used for the learning process. Figure 3 illustrates a three-fold $(k = 3)$ cross-validation [49]. The darker sections of the data are used for the training process while the lighter sections are used for validation. In data mining 10-fold cross-validation $(k = 10)$ is commonly used. But when $k = 1$ the procedure is called leave-one-out cross-validation (LOOCV). In this study we performed a LOOCV at the country level, due the correlated nature of the data; meaning that for each iteration a country i (with its n_i repeated measurements) is used for validation and the remaining countries for training. Model fit is measured using the mean squared error (MSE) as follows:

$$\overline{MSE} = \frac{1}{k}\sum_{i=1}^{k} MSE_i; \quad \text{where } MSE_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\left(y_{ij} - \hat{y}_{ij}\right)^2$$

The goal of cross-validation is to estimate the expected level of fit of a model to a data set that is independent of the data that were used to train the model. Cross-validation over a two-dimensional grid consists in fixing a set of values for each tuning parameter λ and α, and afterwards computing the MSE (by applying cross-validation algorithm) for each pair. The pair corresponding to the minimum MSE is the best.

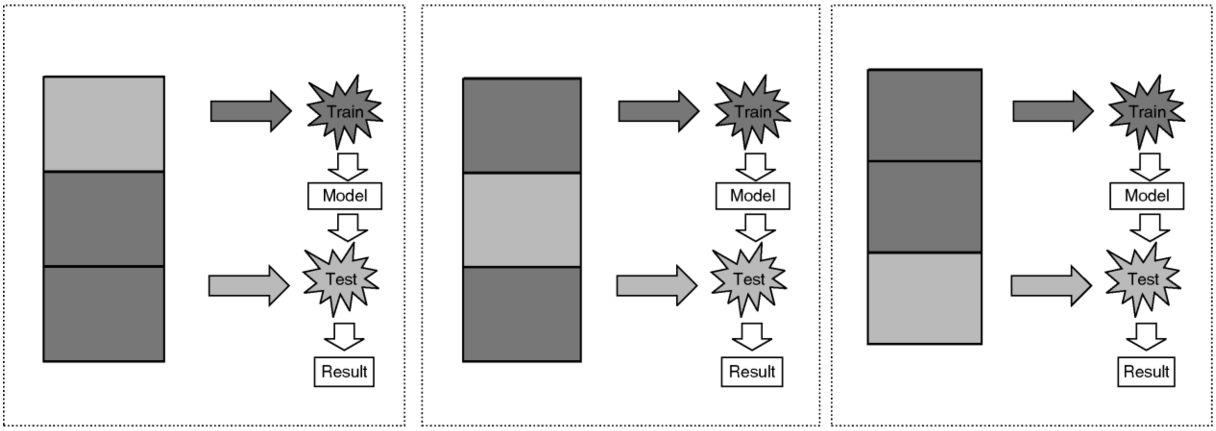


Figure 3: Procedure of three-fold cross-validation [49]

Chapter 3. Results

3.1. Exploratory Data Analysis

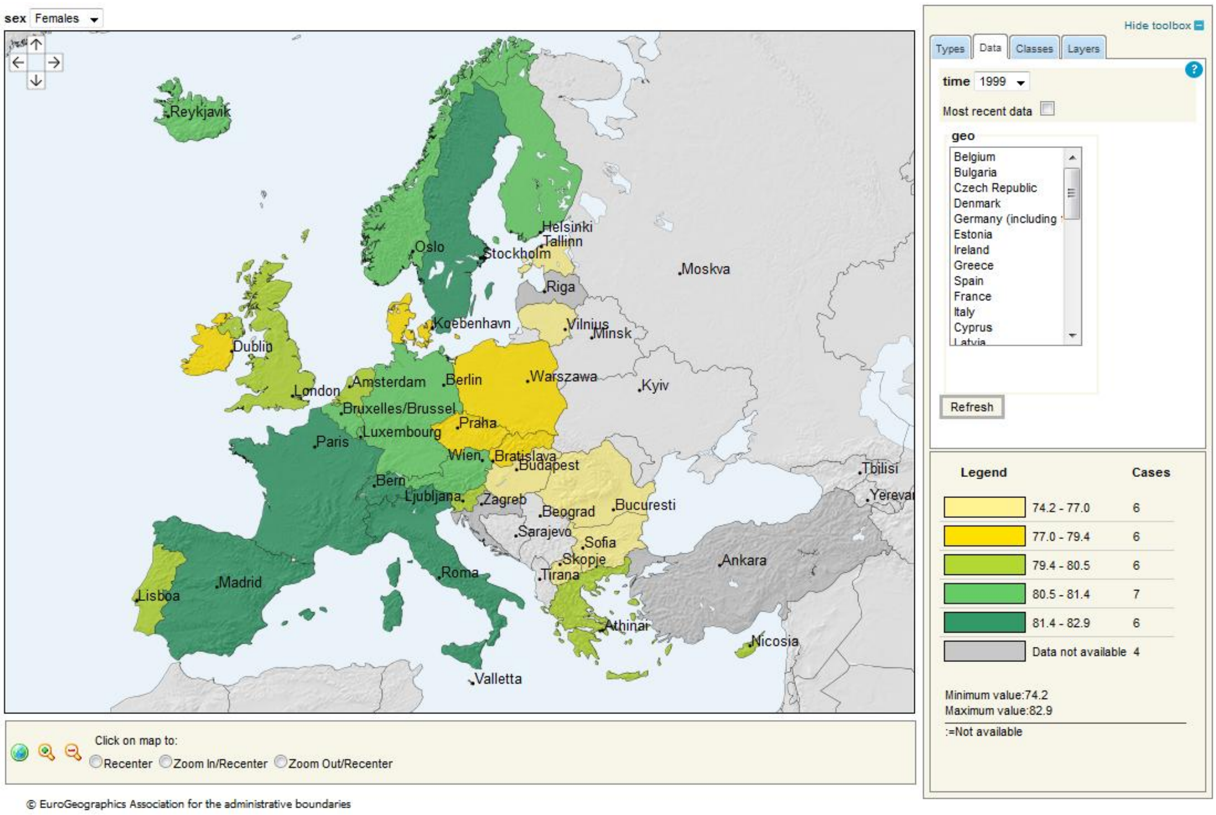
3.1.1. Context

Not all countries of the EU were included in the analysis. Table 1 shows the list of the 24 countries concerned. No selection of countries was done, the single exclusion criteria is the unavailability of data. We can notice from table 1 that data are not available for the 9 measurement occasions (1999-2007) in all countries. In Greece and Latvia for instance, data are available for 5 years only, for 6 years in Croatia, for 7 years in the Czech Republic, Norway and Poland, for 8 years in Finland, France, Italy, Slovakia and United Kingdom. We have all data for the remaining 12 countries (50%).

Table 1: List of countries and number of years that observations are available

Country	Number of years	Country	Number of years	Country	Number of years
Austria	9	Germany	9	Norway	7
Belgium	9	Greece	5	Poland	7
Croatia	6	Hungary	9	Portugal	9
Czech Republic	7	Ireland	9	Slovakia	8
Denmark	9	Italy	8	Slovenia	9
Estonia	6	Latvia	5	Spain	9
Finland	8	Luxembourg	9	Sweden	9
France	8	Netherlands	9	United Kingdom	8

The map of figure 4 illustrates the female life expectancy at birth in the EU in 1999 (beginning of the study period). We can see that the females have a higher life expectancy at birth in France (Paris), Italy (Rome), Luxembourg, Spain (Madrid), Sweden (Stockholm) and Switzerland (Bern). The map also points out the lack of data for Croatia (Zagreb) and Latvia (Riga).



sex Females

Hide toolbox

Types | Data | Classes | Layers

time 1999

Most recent data □

geo

Belgium
Bulgaria
Czech Republic
Denmark
Germany (including
Estonia
Ireland
Greece
Spain
France
Italy
Cyprus
Latvia

Refresh

Legend		Cases
▭	74.2 - 77.0	6
▭	77.0 - 79.4	6
▭	79.4 - 80.5	6
▭	80.5 - 81.4	7
▭	81.4 - 82.9	6
▭	Data not available	4

Minimum value:74.2
Maximum value:82.9

:=Not available

Click on map to:
○Recenter ○Zoom In/Recenter ○Zoom Out/Recenter

© EuroGeographics Association for the administrative boundaries

Figure 4: Life expectancy of females at birth in 1998 in the EU (Source: EUROSTAT website)

3.1.2. Response variables

The response variables of this study are life expectancy at birth and life expectancy at age 65. Figure 5 portrays their evolutions over time in each country per gender. One notes a considerable between- country variability and small within-country variability showing that data are correlated. Regardless the gender, the evolution of each response variables is almost linear in most of the countries.

In 1999 the average male life expectancy at birth is 73.5 in the 24 countries, it reaches 74.81 in 2007. The average female life expectancy is 80.09 in 1999 and 81.36 in 2007. The average life expectancy at age 65 of males starts at 15.02 in 1999 to reach 16.08 in 2007. That of females starts at 18.8 and reaches 19.83. The large disparity between males and females justifies somehow why we will look for the determinants of life expectancy separately for both genders.

One also notices a certain parallelism between the profiles. In addition to the low between-country variability, that suggests that a marginal model would reflect the trend in the evolution of life expectancy in most of the countries.

Figure 5 also shows up the missingness issue. As mentioned earlier countries with complete data represent 50%. As table A.4 in the appendix shows, monotone missingness represents 20.83% and intermittent (non-monotone) missingness 29.17%.

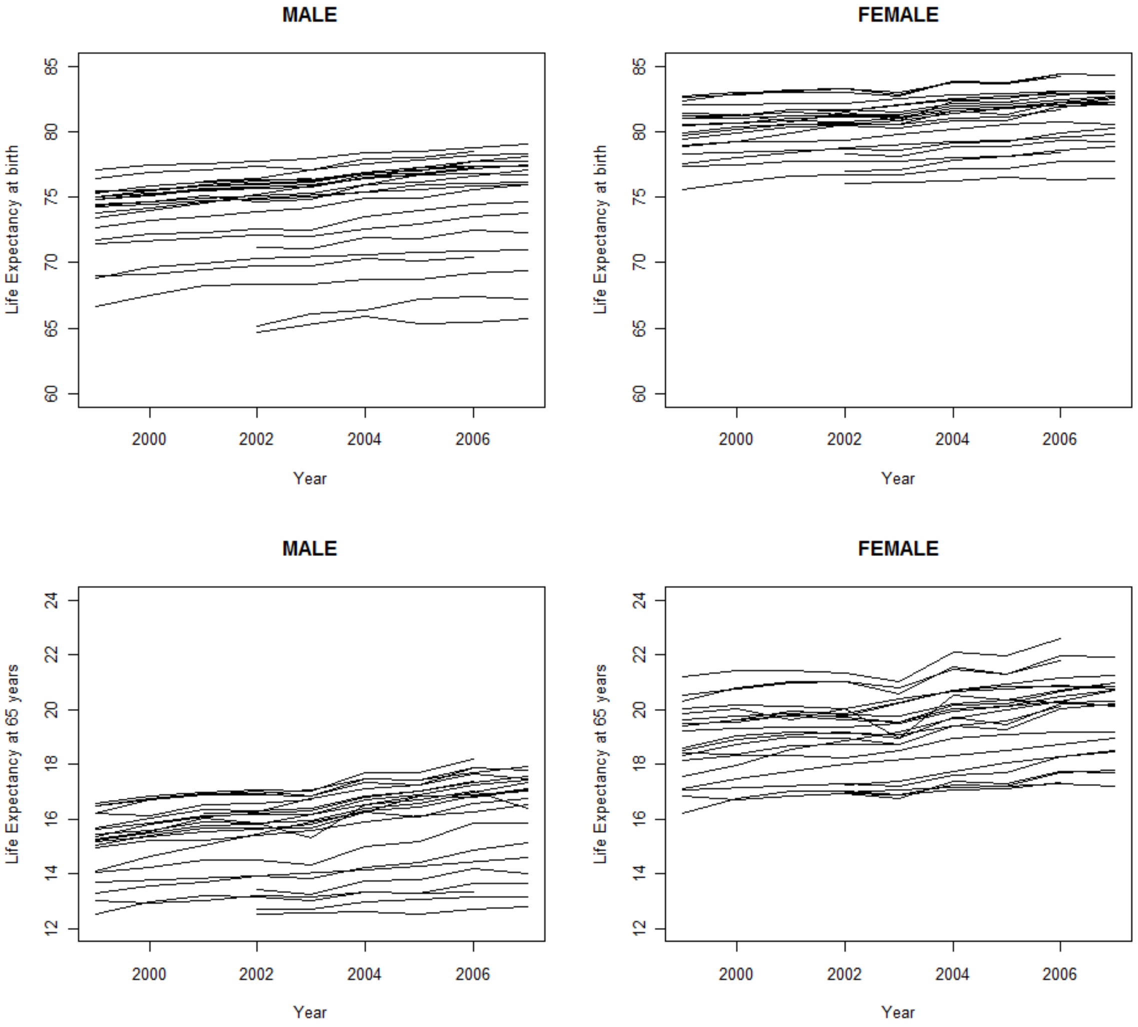


Figure 5: Profiles of the life expectancy at birth and at age 65 (Males and Females)

3.1.3. Covariates

Covariates are also affected by missingness. Multiple imputation was applied to deal with that. A total of 5 data sets were generated in order to account for the uncertainty related to the multiple imputation in the analysis [25]. Almost all covariates are time-varying. Figure 6 depicts the evolution of the single binary covariate (Guidelines) and one continuous covariate randomly selected (Gini). We can notice that within almost all country Gini varies over time. Even Guidelines changes over time within a country.

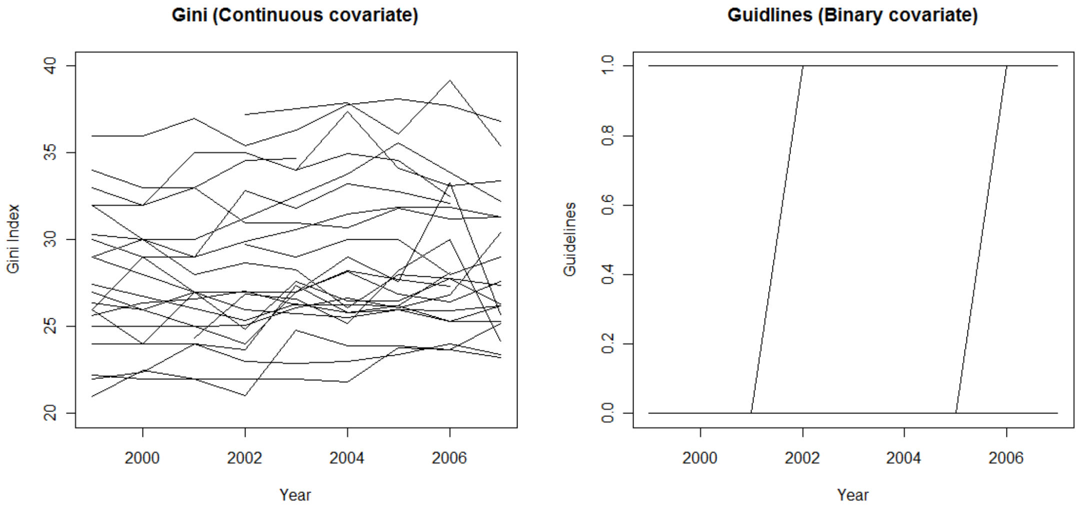


Figure 6: Profiles of two covariates (Gini and Guidelines)

3.2. Penalized GEE

In this report we look for the determinants of life expectancy for males and females separately. For each gender we also search the determinants of life expectancy at birth and at age 65 apart. Four types of models are thus considered. However, the missingness of covariates has been handled with multiple imputation and five data set were generated to that end. Each of the four types of model is then fitted five times in order to take into account the variability related to multiple imputation.

3.2.1. Comparison of Elastic-Net and SCAD-L_2: MSE

Elastic-net and SCAD-L_2 penalties depend on a tuning parameter α supported on the interval [0,1] and a tuning parameter $\lambda > 0$. As recall SCAD-L_2 depends on an additional tuning parameter $a > 2$. But in this analysis we used the optimal value $a = 3.7$ proposed by Fan and Li [9]. In order to select optimal values for the parameters α and λ we performed cross-validation over a two-dimensional grids for each of the 20 models to consider. For both penalties 10 values equally spaced were considered in the interval $[0.1; 1]$ for the parameter α. For the elastic-net penalty we generated in the interval $[0.0002;\ 0.04]$ 200 values equally spaced for the parameter λ; for SCAD-L_2 rather 400 values equally spaced were generated between 0.001 and 1 for the latter tuning parameter.

Figure 7 shows the MSE of the each of the four types of models in the grid of pair $(\alpha;\lambda)$ for the first data set. It compares elastic-net with SCAD-L_2. Figures A.1, to A.4 in the appendix present the results for the four remaining data sets. We note a high sensitivity of all models with respect to the choice of the pair $(\alpha;\lambda)$. However, the effect of the tuning parameter α is more pronounced only for larger values of the tuning parameter λ, not for its smaller values. Whatever α value considered, we note a high sensitivity of each model with respect to the λ values. These figures also point out a noticeable difference between the elastic-net penalty and the SCAD penalty for each model. They briefly indicate a substantial variability among the five data sets. They also suggest different optimal models for each gender and each age.

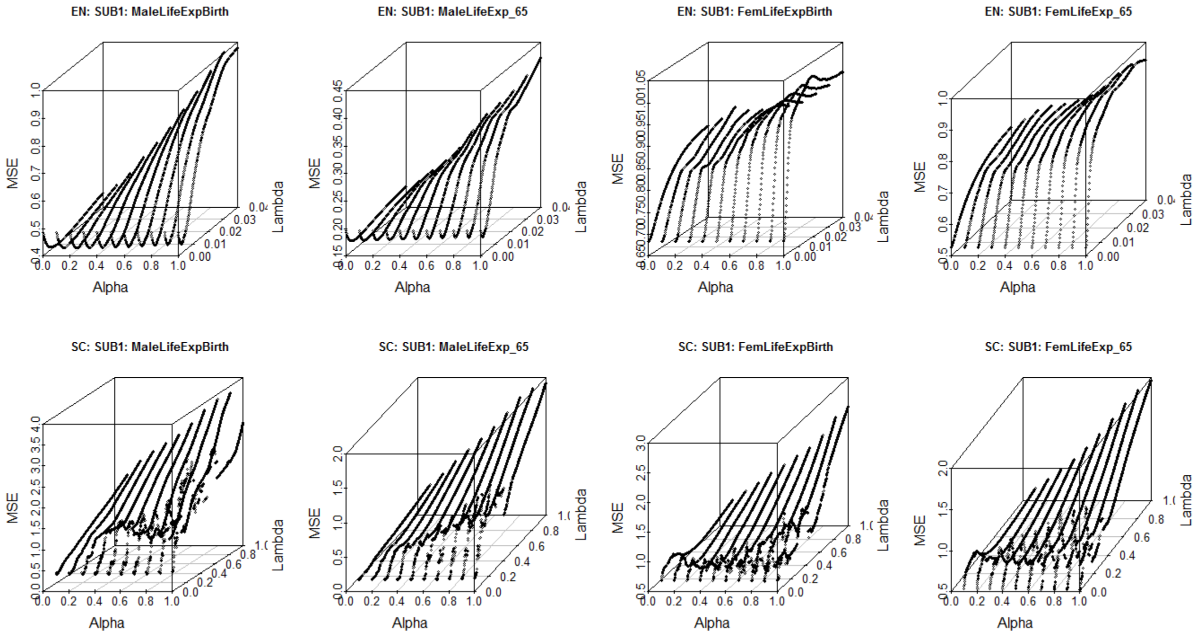


Figure 7: Two-dimensional grid cross-validation (on PGEE) for the pair (α,λ): Data set 1

As a reminder both elastic-net and SCAD-L_2 satisfy the main properties required to apply penalization for variable selection purpose. But they seem to perform differently on our data as we observed on figures 7, A.1 to A.4. Table 2 compares the two penalties first of all by the means of the minimum MSEs in the 20 grids described above. We can notice that except the model for male life expectancy on the third data set and that of male life expectancy at age 65 on the fourth data set, for each model the minimum MSE of SCAD-L_2 penalty is lower than that of elastic-net. After taking into account variability caused by multiple imputation by averaging the MSE of the five data sets for each type of model, the so-obtained MSEs indicate that SCAD-L_2 is the optimal penalty i.e. the one that minimizes the MSE on average.

Table 2: Comparison of elastic-net and SCAD-L$_2$: Minimum MSE in the two-dimensional grid

Gender	Age	Data set	Elastic-Net				SCAD-L$_2$			
			MSE	SE	Alpha	Lambda	MSE	SE	Alpha	Lambda
Male	Birth	1	0.4051	0.0155	0.1	0.0154	0.3866	0.0144	0.6	0.0210
Male	Birth	2	0.4039	0.0126	0.3	0.0064	0.3669	0.0121	0.8	0.0260
Male	Birth	3	0.3415	0.0124	0.1	0.0084	0.3421	0.0123	0.2	0.0110
Male	Birth	4	0.3566	0.0116	0.4	0.0038	0.3245	0.0115	1.0	0.0335
Male	Birth	5	0.3476	0.0116	0.1	0.0072	0.3278	0.0114	0.9	0.0210
	Mean		**0.3709**				**0.3496**			
Male	65	1	0.1702	0.0068	0.1	0.0132	0.1541	0.0069	0.5	0.0260
Male	65	2	0.1698	0.0066	0.1	0.0164	0.1548	0.0058	0.2	0.0310
Male	65	3	0.1471	0.0061	0.1	0.0104	0.1407	0.0064	0.3	0.0510
Male	65	4	0.1344	0.0056	1.0	0.0022	0.1356	0.0053	0.8	0.0185
Male	65	5	0.1295	0.0048	0.1	0.0036	0.1231	0.0045	0.6	0.0110
	Mean		**0.1502**				**0.1417**			
Female	Birth	1	0.6827	0.0287	0.9	0.0012	0.6210	0.0245	0.3	0.3385
Female	Birth	2	0.6801	0.0242	0.1	0.0042	0.5662	0.0224	0.3	0.3235
Female	Birth	3	0.4561	0.0193	0.1	0.0006	0.3253	0.0123	1.0	0.0285
Female	Birth	4	0.7152	0.0267	1.0	0.0014	0.6028	0.0229	1.0	0.0235
Female	Birth	5	0.5213	0.0215	0.1	0.0010	0.4966	0.0214	0.9	0.0185
	Mean		**0.6111**				**0.5224**			
Female	65	1	0.5284	0.0216	0.1	0.0014	0.5254	0.0217	1.0	0.0110
Female	65	2	0.5656	0.0203	0.1	0.0032	0.5083	0.0237	0.3	0.2785
Female	65	3	0.3672	0.0155	0.1	0.0006	0.3587	0.0153	1.0	0.0085
Female	65	4	0.6212	0.0227	1.0	0.0008	0.5221	0.0239	1.0	0.0235
Female	65	5	0.4772	0.0197	0.1	0.0008	0.4733	0.0196	1.0	0.0185
	Mean		**0.5119**				**0.4775**			

3.2.2. Profiles of elastic-net and SCAD-L$_2$ coefficients (Paths)

The difference between elastic-net and SCAD-L$_2$ appears in the way they bias the coefficients; that has an effect on the MSE as we noticed in table 2. Their theoretical effects (orthogonal design case) on coefficient estimates are shown in table A.3 in the appendix and figure 2. Figures 8 and 9 compare the behaviours of both penalties on the first data set for the models of male and female life expectancy at birth. For each figure, the left panels represent the profiles of all 53 coefficients (of the 53 covariates) and the right panels those of 20% of coefficients randomly selected. It is easy to notice that both penalties set coefficients to zero. SCAD-L$_2$ varies more than elastic-net, mostly for smaller values of the tuning parameter λ. We can notice for both penalty functions that for some λ values the coefficients are rather biased (in absolute value) upward not downward; that is surely the consequence of multicollinearity that our data suffer from. These graphs are built with the optimal pairs (α, λ). The vertical dashed lines represent the optimal λ (Optimal model).

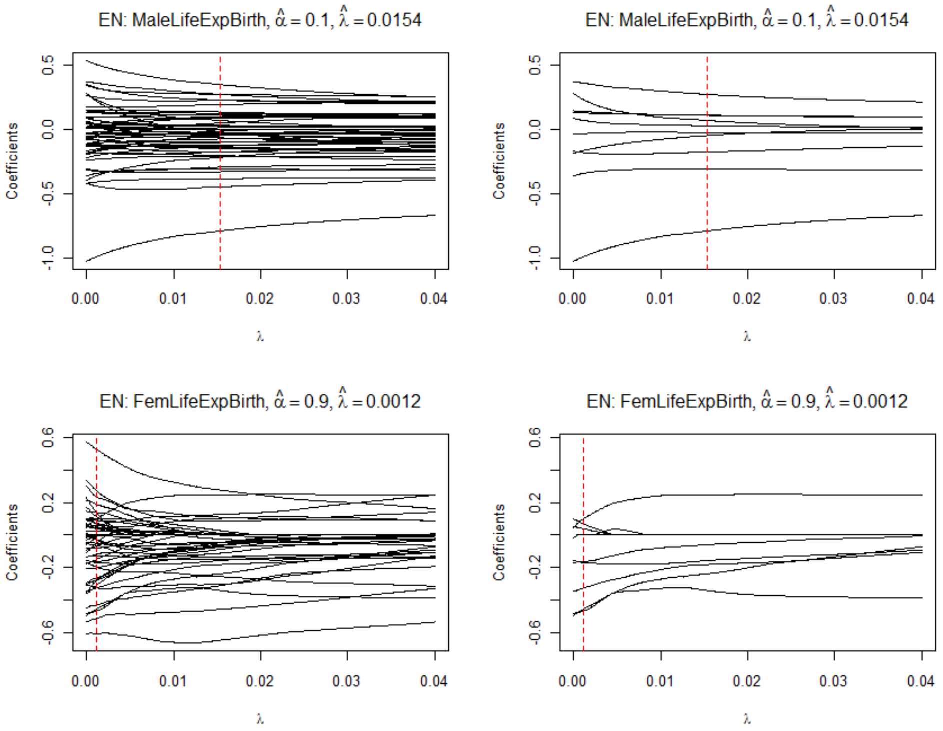


Figure 8: Profiles of elastic-net coefficients on data set 1. All 53 variables on left and 20% randomly selected on right

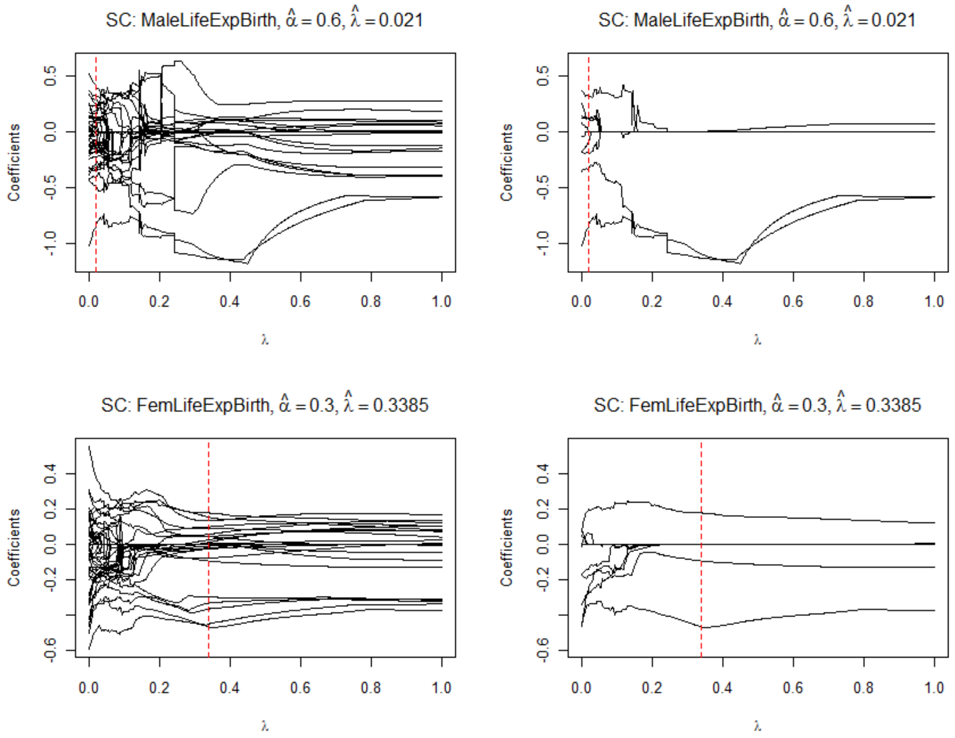


Figure 9: Profiles of SCAD-L$_2$ coefficients on data set 1. All 53 variables on left and 20% randomly selected on right

3.2.3. Selection of optimal pair $(\boldsymbol{\alpha}, \boldsymbol{\lambda})$

On one hand it emerges from section 3.2.1. that SCAD-L_2 penalty will be used for variable selection. On the other hand, the grid is made up of different pairs (α, λ). Therefore, for each of the 20 models it is necessary to select the pair that minimizes the MSE. The first idea is to select the pair with the smallest MSE. But, each MSE has a certain variability attached to it, for instance in the grid of the model for male life expectancy at birth on data set 1, the minimum MSE is 0.3866 with a standard error of 0.0144 (Table 2). This means that in a grid the model with minimum MSE may not be the single optimal model. Therefore we selected the optimal pair different, precisely in two steps. The optimal α is first of all determined. It corresponds to the estimate $\hat{\alpha}$ of the model with minimum MSE. Afterwards the optimal λ is determined with the "one-standard error" rule. The inclusion of a "one-standard error" rule provides a collection of good models in which we select the one with the largest $\hat{\lambda}$ as the optimal model. Figure 10 illustrates the application of "one-standard error" rule. Two situations are presented. The left panel is the model for male life expectancy at birth on the third data set, with minimum MSE 0.3421 and the related standard error 0.0123 (Table 2). The corresponding optimal $\hat{\alpha} = 0.2$ and the corresponding $\lambda = 0.006$. By application of "one-standard error" rule, all models with MSE between 0.3421 and 0.3421 + 0.0123 = 0.3544 are good; that are the four models under the horizontal dashed line (at "one-standard error" of the minimum MSE).

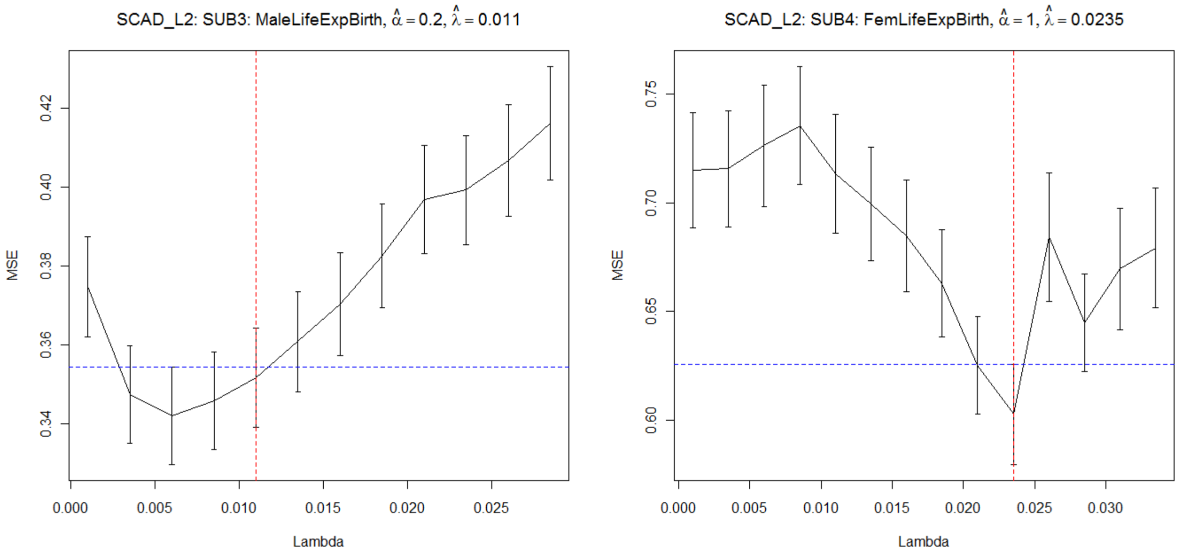


Figure 10: Examples of "one-standard error" rule application (Focus in the optimal region)

Among them the optimal model is the one with largest λ value (model covered by the vertical dashed line), meaning the optimal parameter $\hat{\lambda} = 0.0110$. In the right panel however (model of female life expectancy at birth on the fourth data set), the optimal λ corresponds to that of the model with minimum MSE, $\hat{\lambda} = 0.0235$. Indeed, the latter is the only model under the "one-standard error" line. Figures A.5 to A.8 in the appendix show the application of "one-standard error" rule to all 20 considered models.

3.2.4. Comparison of deletion rules (SCAD-L$_2$ penalty)

The considerable difference between our five data sets started to emerge from the previous analyses (difference between optimal pairs $(\alpha; \lambda)$ of the five data sets for the same type of model); in this section that is clearly visible. As we can notice for the model of life expectancy at birth (Table A. 5 in the appendix), penalized coefficients vary too much between the data sets. Coefficients set to zero are not the same across the data set. We obtain similar results with the models of male life expectancy at age 65, female life expectancy at birth and at age 65. Thus three rules were defined in order to account for the variability related to multiple imputation in the selection of variables. The first one is the ***"Majority Vote Deletion Rule" (MVDR):*** It consists in dropping variables set to zero by the majority of the five data sets, meaning by at least 3 data sets. The second is the "***Union Deletion Rule" (UDR):*** it consists in dropping all variables set to zero by at least one data set. The last one is the ***"Intersection Deletion Rule" (IDR)***: It consists in deleting a variable if its penalized coefficient is equal to zero for each of the data sets.

By definition UDR selects less variables and IDR selects more variables. MVDR appears in between the two rules. That is satisfied in our case as table 3 shows up. For example, for the model of male life expectancy at birth IDR keeps all 53 variables, UDR selects 25 and MVDR 35. However, it is necessary to mention that the three rules would have led to the same result if PGEE had set to zero the coefficients of the same variables simultaneously for the five data sets. Since it is not the case the performance of the three rules has to be assessed. Thus, with the selected variables cross-validation was performed on GEE. Table 3 contains the MSE and their standard errors which have resulted. IDR with almost the highest selection rate has the poorest performance. MVDR has the best performance for two models and UDR for two models also. However MVDR is retained to build the final models because the number of variables selected by UDR for the model of female life expectancy at birth seems too small (selection rate of 20.75%, only).

Table 3: Comparison of dropping rules through cross-validation on GEE

Gender	Age	# Covariates selected	Selection Rate	MSE per data set (SE)					Average MSE
				SUB1	SUB2	SUB3	SUB4	SUB5	
Majority Vote Deletion Rule									
Male	Birth	35	66.04%	0.2650	0.2676	0.2052	0.2173	0.2524	0.2415
				(0.0148)	(0.0124)	(0.0101)	(0.0086)	(0.0128)	
Male	65	38	71.70%	0.1548	0.1492	0.1802	0.1533	0.1181	0.1511
				(0.0040)	(0.0049)	(0.0079)	(0.0060)	(0.0034)	
Female	Birth	29	54.72%	1.1465	1.3682	0.9367	0.9783	0.9351	1.0730
				(0.0769)	(0.1095)	(0.0625)	(0.0661)	(0.0617)	
Female	65	32	60.38%	1.1573	1.1272	1.0540	0.9855	0.9426	1.0533
				(0.0881)	(0.0933)	(0.0819)	(0.0789)	(0.0703)	
Union Deletion Rule									
Male	Birth	25	47.17%	0.1786	0.1827	0.1629	0.1376	0.2299	0.1783
				(0.0059)	(0.0073)	(0.0066)	(0.0048)	(0.0094)	
Male	65	26	49.06%	0.3792	0.3611	0.3727	0.3578	0.3437	0.3629
				(0.0131)	(0.0129)	(0.0135)	(0.0139)	(0.0120)	
Female	Birth	11	20.75%	0.8738	0.8871	0.9297	0.9472	0.9508	0.9177
				(0.0378)	(0.0380)	(0.0369)	(0.0405)	(0.0430)	
Female	65	10	18.87%	1.2523	1.3272	1.3635	1.3738	1.2679	1.3169
				(0.2024)	(0.2189)	(0.2313)	(0.2306)	(0.2000)	
Intersection Deletion Rule									
Male	Birth	53	100.00%	0.6346	0.6134	0.5303	0.5146	0.4308	0.5447
				(0.0323)	(0.0271)	(0.0251)	(0.0241)	(0.0242)	
Male	65	49	92.45%	0.3629	0.3454	0.3600	0.3187	0.2289	0.3232
				(0.0154)	(0.0153)	(0.0178)	(0.0143)	(0.0078)	
Female	Birth	42	79.25%	1.7974	2.0264	1.7231	1.7899	1.3720	1.7417
				(0.1179)	(0.1436)	(0.1141)	(0.1167)	(0.0858)	
Female	65	51	96.23%	2.0643	2.0775	1.9695	2.3475	2.1182	2.1154
				(0.1461)	(0.1592)	(0.1358)	(0.1623)	(0.1425)	

3.3. Final models: GEE

3.3.1. Data preparation

In the PGEE algorithm used above the response is centered and the covariates are standardized (mean 0 and variance 1). Then covariates are used on the same scale. However, some variables exhibit a very large standard deviance namely CHICPROD, GDP, HopBeds, Pop, PopDens and TOTALHE. (Table A.2 in the appendix). They may mislead inference on parameter estimates if they are included in the model without any transformation. We can also notice from figure 11 that even for our data the relationship between GDP and life expectancy is not linear. Logarithmic transformation was then applied as attempt to capture that non-linear relationship between each outcome and GDP [10, 32]. This monotonous transformation has an additional advantage of reducing the variability of our covariate; therefore it was generalized to others covariates with large variance (Table A.2). In the final models which are GEE, covariates selected with PGEE are not centered, therefore transformation of those with larger variance is useful.

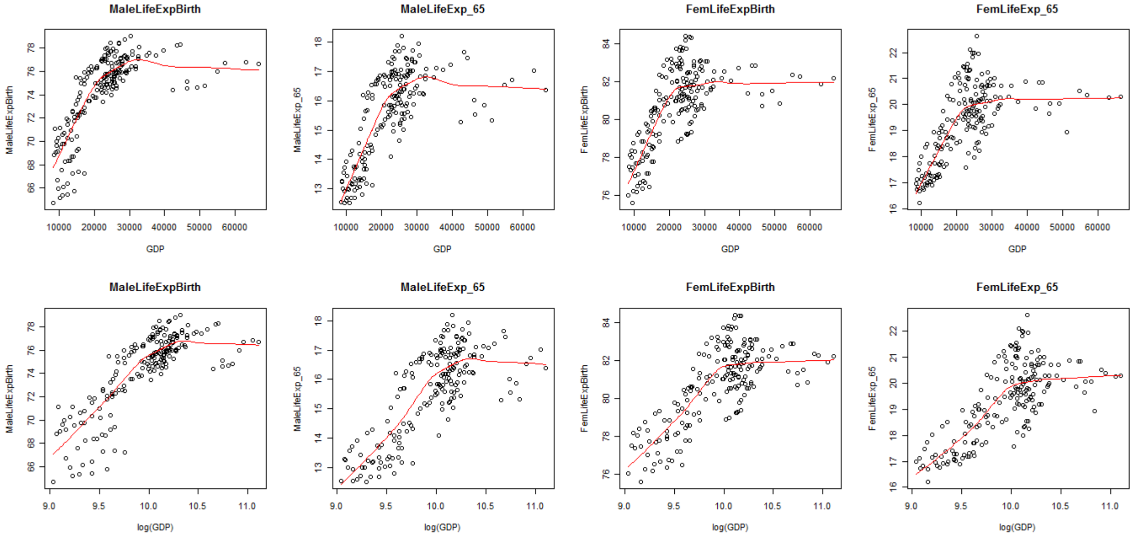


Figure 11: Scatter plot of each outcome variable against GDP and log(GDP) (First data set)

3.3.2. Estimation of final models

In previous sections we based variable selection (and model comparison) on biased coefficients estimates and prediction performance (use of cross-validation). The so-selected covariates are used to fit GEE. For each of the four types of model, GEE is fitted to each of the five data sets. But contrary to the PGEE where we defined rules in order to pool the results on the five data sets, the five analyses in this case are combined into a single one using the algorithm proposed by Rubin [25]. Therefore model refinement is now based on the pooled inference on the parameters estimates.

Finally, 25 variables out of 53 explain the male life expectancy at birth (Table A.6 in the appendix) and 21 explain the male life expectancy at age 65 (Table A.7 in the appendix). Regarding female, 19 variables out of 53 affect the life expectancy at birth (Table A.8 in the appendix) and 21 affect the life expectancy at age 65 (Table A.9 in the appendix). For each model some factors are positively related to the life expectance whereas others are negatively related to it.

3.3.3. Model interpretation

PGEE was fitted to our data using R code written by Blommaert [3]. Model selection was mainly based on cross-validation. After taking into account the variability related to the multiple imputation, method used to handle the missingness of covariates, the four types of models so-obtained can be interpreted.

3.3.3.1. Determinants of male life expectancy at birth (Appendix: Table A.6)

In the EU many factors have a positive effect on male life expectancy at birth, namely among other things the GDP, the population size, the practising physician density and the proportion of person aged 65. The increase of the percentage of public sector expenditure on health in the total government expenditure, and the increase of birth rate also affect positively the male life expectancy at birth. In the countries where the number of hours worked per week in full time employment is higher, at birth males are expected to live longer. Unexpectedly, the increase of suicide rate and also the increase of corruption index score are related with the increase of male life expectancy at birth.

Not all death rates explain the male life expectancy at birth. Deaths due to pneumonia, liver, diabetes, chronic diseases, alcohol abuse, Aids, diseases of respiratory system and accidents contribute significantly to the reduction of that life expectancy. Others factors which are negatively related to the male life expectancy at birth are the infant deaths per 1000 live births, the variability of absolute humidity, the percentage of urban population, the percentage of regular daily smokers and the ratio of females to males (the number of women per 100 men). Countries with higher male life expectancy at birth seem to be less atheist (give more attention to religion) than those with smaller male life expectancy at birth. Surprisingly this response variable seems to decrease with the hospital bed density and the percentage of total expenditure on private health in the total expenditure on health.

3.3.3.2. Determinants of male life expectancy at age 65 (Appendix: Table A.7)

Male life expectancy at age 65 increases with the total population size, the practising physician density, the GDP, the number of hours worked per week for full time employment, the total health expenditure (as percentage of GDP) and surprisingly the fact that less people in the country can be trusted. Like the male life expectancy at birth, the corruption index score is positively related to male life expectancy at age 65. Death rates due to pneumonia, ischaemic heart disease, cancer, Aids, influenza, diseases of the respiratory system and accidents significantly reduce the male life expectancy at age 65. Others factors negatively related to this response variable are the variability of absolute humidity, the average population density per square kilometer, the percentage of regular daily smokers, the hospital bed density and the total expenditure on private health as a percentage of total expenditure on health. The more a country is religious, the higher its male life expectancy at age 65 seems to be. In addition, male life expectancy at age 65 is lower in countries where most of the people think that the greater respect for authority is a bad thing.

3.3.3.3. Determinants of female life expectancy at birth (Appendix: Table A.8)

In the EU female life expectancy at birth increase with the GDP, the population size, the practising physician density, the ratio of females to males and against all the odds the total Greenhouse gas emissions. It is negatively related to the variability of absolute humidity, the average population density per square kilometer, the hospital bed density, the percentage of population aged 0-14, the infant deaths per 1000 live births and the percentage of public sector expenditure on health in the total government expenditure. Moreover, female life expectancy at birth is larger in countries where more people think that it is a very good thing of having experts who make decisions about the country and also in countries where more people think that the greater respect for authority is a good thing. It is lower in countries with larger private households' out-of-pocket payment on health (as % of total health expenditure). The out-of-pocket expenditure on health is defined as the direct outlays of households, including gratuities and in-kind payments made to health practitioners and to suppliers of pharmaceuticals, therapeutic appliances and other goods and services [48]. In addition to the infant deaths per 1000 live births, deaths that significantly contribute to the reduction of female life expectancy at birth are those due to pneumonia, ischaemic heart disease, chronic diseases, cancer and Aids.

3.3.3.4. Determinants of female life expectancy at age 65 (Appendix: Table A.9)

The determinants of female life at age 65 are not too different from those of female life expectancy at birth. The former outcome is positively related to the GDP, the population size, the practising physician density, the ratio of females to males, the total health expenditure (as percentage of GDP) and against all the odds the suicide rate and the fact that less people in the country can be trusted. It is negatively related to the variability of absolute humidity, the percentage of the population aged 65 and above, the average population density per square kilometer, the hospital bed density, the percentage of the population aged 0-14, the private households' out-of-pocket payment on health (as percentage of total health expenditure), the Public sector expenditure on health (as percentage of total government expenditure) and the fact that less people in a country think that greater respect for authority is a very bad thing. Death rates due to pneumonia, ischaemic heart disease, cancer, Aids and others acute respiratory infections importantly reduce the female life expectancy at age 65.

Chapter 4: Discussion and conclusion

4.1. Discussion

4.1.1. Missingness

The four outcomes are affected by missingness. By basing variable selection on GEE we indirectly assume that the missingness mechanism is MCAR (Missing Completely At Random), meaning that the missingness is independent of both observed and unobserved data. This assumption seems plausible since data analyzed come from administrative sources. The reason of missingness in administrative data can mostly be structural (organizational, political, economical) and may differ from one country to another. In fact many elements come into play in the production of such data; that is why we deemed MCAR assumption less doubtful. However this assumption is more often considered as too strong to hold and the weaker Missing At Random (MAR) assumption is most of the time preferred. Missing data are said to be MAR if the probability of missingness depends only on observed data.

Under MAR, valid inferences can be obtained through Weighted GEE (WGEE) by modelling the monotone missingness (dropout) process [25] and a likelihood-based analysis without the need for modelling the monotone missingness process [24]. The latter case means fitting a Linear Mixed Model (LMM) to our data. However, We could not apply this model to our data because of lack of tools on LMM with time-dependent covariates. In fact, Diggle *et al.* [6] stated that the FCCM assumption defined by Pepe and Anderson [29] is not only required for the application GEE, it is an important issue for all longitudinal data analysis methods including likelihood-based methods such as linear and generalized linear mixed models. Due to lack of theory on penalized WGEE it was not also applied; theory is available only on PGEE that we used.

4.1.2. Penalization

In regression analysis, if an important predictor is left out coefficient estimates will be biased and predictive performance may be poor [7]. However, we learnt that by controlling the biasedness of the coefficient estimates the predictive performance of the model can be improve (through bias-variance trade-off) and also we can gain in terms of interpretability of the model.

Different penalties have been defined in this report. Ridge cannot be used for variable selection because of its lack of sparsity. Among those which can be used for that purpose LASSO is sometimes blamed because its shrinkage produces biased estimates for the large coefficients and that is not always desired. Therefore adaptive LASSO has been proposed, in which adaptive (data-dependent) weights are used for penalizing different coefficients in the L_1 penalty: $\sum_{j=1}^{p} w_j \left|\beta_j\right|$. It was shown that the adaptive LASSO enjoys the oracle properties like SCAD [39].

The elastic-net penalty defined in section 2.2.1. is known as "naïve elastic-net". Elastic net (corrected) parameter estimates are obtained by multiplying the naïve ones by a factor [40]. Naïve elastic-net is defined as a linear combination of $L_1$and L_2 norms; by replacing L_1 with adaptive LASSO penalty we obtain the adaptive elastic net penalty which also enjoys oracle properties [41].

Even though SCAD-L_2 proved to be the optimal penalty in our analysis, it is more time-consuming for model selection compared to its counterpart elastic-net. As we noticed in figures 7, A.1-4, model performance varies too much with SCAD-L_2 penalty, mostly for smaller values of the tuning parameter λ. That makes selection of optimal model difficult as depicted in figure A.8 in the appendix (different minima) where a wrong specification of the λ range of values would lead to the selection of a wrong optimal model. We overcame this situation by considering a wider range of values.

4.1.3. Expectations

In our finals models the relationships between some covariates and the life expectancies are contrary to the expectations, namely the total greenhouse gas emissions (CO2) which is rather positively related to the female life expectancy at birth and the suicide rate (SUICIDE) also positively associated to the male life expectancy at birth and female life expectancy at age 65. Corruption is a symptom of deep institutional weaknesses and leads to inefficient economic, social, and political outcomes. It reduces among other things expenditures for education and health. Akçay [2] showed that corruption in all its aspects retards human development, but our models reveal contradictory results: the relationship between corruption index and male life expectancy (at bird or age 65) rather is positive for our data. These results that are in contradiction with our expectations call for further research, maybe by considering a larger period of time (more than 9 years). The positive relationship between life expectancy and the fact the "less people in the country can be trusted" can be attributed to the subjective nature of the related covariate. For the four models, none of the variables on education has been selected although the latter is known as a determinant of life expectancy.

4.2. Conclusion

It emerges from this analysis that between 1999 and 2007 in the EU, the wealth of a country, the population size, and the practising physician density positively affect the life expectancy whatever age and gender considered. The life expectancy is negatively related to the variability of absolute humidity and the hospital bed density regardless age and sex. Deaths due to Aids and pneumonia reduce the average number of years a person can expect to live without regard to the age and gender. As the density of hospital beds, a high variation of absolute humidity also seems to have an adverse effect life expectancy of every person. With the exception of those determinants that are common to everyone, we can notice that the determinants of female life expectancy at age 65 are almost the same as those of the female life expectancy at birth. However, that is not the case for males.

4.3. Prospects

At the end of this work, we can note that several improvements can be made thereto theoretically and also practically. Practically, one of the first improvements that can be made is to write R code that computes standard errors of the penalized parameter estimates. Thus model refinement could be based on PGEE, not GEE. Secondly, a new multiple imputation method can be applied to handle the missingness of covariates. The method used in this study leads to a high variability among the data sets and even to almost the loss of sparsity when the PGEE results on the five data sets are combined.

On the theoretical side the task seems heavy, and even complex. First of all we propose the extension of penalization to WGEE so that variable selection can be done not only under the strong MCAR assumption, but also under the weaker and acceptable MAR assumption. We also propose to perform nonlinear modelling to capture the true relationship between life expectancy and GDP. In that case, we can further think of penalized nonlinear model in order to proceed at the same time to variable selection.

References

[1] Abdulraheem, I.S., Jimoh, A.A.G. and Oladipo, A.R. (2011). Gender Differential in Life Expectancy: Trends, Determinants and Empirical Findings. *Journal of Peace, Gender and Development Studies*, **1,** 015-027.

[2] Akçay, S. (2006). Corruption and Human Development . *Cato Journal*, **26**, 29-48.

[3] Blommaert A. (2011). *R Code for Penalized Generalized Estimating Equations*. University of Antwerp, Unpublished.

[4] Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, **24**, 2350–2383.

[5] Davis, C.S. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. New York: Springer.

[6] Diggle, P.J., Heagerty, P.J., Liang, K.Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data*. 2nd ed. Oxford: Oxford University Press.

[7] Dziak, J.J. (2006). *Penalized Quadratic Inference Functions for Variable Selection in Longitudinal Research.* The Pennsylvania State University, Ph.D dissertation.

[8] Dziak, J.J. and Li, R. (2007). An overview on variable selection for longitudinal data. In: *Quantitative Medical Data Analysis Using Mathematical Tools and Statistical Techniques*, chapter Submitted, Hong, D. and Shyr, Y. (ed.). Singapore: World Scientific, pp. 3-24.

[9] Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, **96**, 1348-1360.

[10] Fayissa, B. and Gutema, P. (2008). *A Health Production Function for Sub-Saharan Africa (SSA).* Murfreesboro: Middle Tennessee State University.

[11] Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. (2009). *Longitudinal data analysis*. United States of America: Chapman and Hall/CRC.

[12] Frank, I. and Friedman, J. (1993). A Statistical View of Some Chemometrics Regression Tools (with discussion). *Technometrics*, **35**, 109–135.

[13] Fu, W.J. (1998). Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, **7**, 397-416.

[14] Fu, W.J. (2003). Penalized Estimating Equations. *Biometrics*, **59**, 126-132.

[15] Germov, J. (1998). *Second Opinion: An Introduction to Health Sociology*. Melbourne: Oxford University Press.

[16] Grbich, C. (1996). *Health in Australia: Sociological Concepts and Issues*. Sydney: Prentice Hall.

[17] Grossman, M. (1972). *The Demand for Health: A theoretical and Empirical Investigation.* New York: NBER

[18] Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer.

[19] Hoerl, A.E. and Kennard, R.W. (1970a). Ridge Regression: Biased estimation for Nonorthogonal Problems. *Technometrics*, **12**, 55-67.

[20] Hoerl, A.E. and Kennard, R.W. (1970b). Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics*, **12**, 69-82.

[21] Kutner, M.H., Nachtsheim, C.J., Neter, J. and Li, W. (2005). *Applied Linear Statistical Models*. 5th ed. New York: Mc Graw Hill.

[22] Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

[23] Miller, A. (2002). *Subset Selection in Regression*. 2nd ed. United States of America: Chapman & Hall/CRC.

[24] Molenberghs G. and Kenward M.G. (2007). *Missing Data in Clinical Studies*. England: John Wiley and Sons Ltd

[25] Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer.

[26] NSW Department of Health (1999). *NSW Chief Health Officers Report*. Sydney.

[27] O'Brien, R.M. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality and Quantity*, **41**(5), 673-690.

[28] O'Sullivan, A., Sheffrin, S.M. (2007). *Economics: Principles in Action*. Boston, Massachusetts, Upper Saddle River, New Jersey: Pearson Prentice Hall.

[29] Pepe, M.S. and Anderson, G.L. (1994). A Cautionary Note on Inference for Marginal Regression Models with Longitudinal Data and General Correlated Response Data. *Communications in Statistics – Simulation and Computation*, **23(4)**, 939–951.

[30] Poudyal, N.C., Hodges, D.G., Bowker, J.M. and Cordell, H.K. (2009). Evaluating Natural Resource Amenities in a Human Life Expectancy Production Function. *Forest Policy and Economics*, **11(4)**, 253–259.

[31] Schnabel, S.K. and Eilers, P.H.C. (2009). An Analysis of Life Expectancy and Economic Production Using Expectile Frontier Zones. *Demographic Research*, **21(5),** 109-134.

[32] Shaw J.W., Horrace, W.C. and Vogel, R.J. (2005) The Determinants of Life Expectancy: An Analysis of the OECD Health Data. *Southern Economic Journal*, **71**, 768-783.

[33] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, Series B, **58**, 267–288.

[34] Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

[35] World Health Organization regional Office for Europe (2008). *Atlas of Health in Europe*. 2nd ed: http://www.euro.who.int/__data/assets/pdf_file/0011/97598/E91713.pdf

[36] Xie, J. and Zeng, L. (2010). Group Variable Selection Methods and Their Applications in Analysis of Genomic Data. In: *Frontiers in Computational and Systems Biology,* Feng, J., Fu, W. and Sun, F. (ed.). New York: Springer, pp. 231-248.

[37] Yi, Z., Crimmins, E.M., Carrière, Y. and Robine, J.M. (2006). *Longer Life and Healthy Aging*. Netherlands: Springer.

[38] Zeger, S.L. and Liang, K.Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**, 121–130.

[39] Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association,* **101**, 1418-1429.

[40] Zou, H. and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society,* Series B, **67**(2), 301–320.

[41] Zou, H. and Zhang, H.H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics,* **37**, 1733–1751.

[42] http://www.who.int/topics/life_expectancy/en/ (Accessed: 1/07/2011)

[43] http://www.ipa.udel.edu/education/courses/econ490/Production_of_Health.pdf (Accessed: 16/08/2011)

[44] http://en.wikipedia.org/wiki/List_of_countries_by_life_expectancy (Accessed: 17/08/2011)

[45] https://www.cia.gov/library/publications/the-world-factbook/rankorder/2102rank.html (Accessed: 17/08/2011)

[46] http://en.wikipedia.org/wiki/Europe#cite_note-135 (Accessed: 5/07/2011)

[47] http://en.wikipedia.org/wiki/Multicollinearity (Accessed: 10/07/2011)

[48] http://www.globalhealthfacts.org/data/topic/map.aspx?ind=73 (Accessed, 27/08/2011)

[49] http://www.public.asu.edu/~ltang9/papers/ency-cross-validation.pdf (Accessed, 1/09/2011)

Appendix

Table A. 1: Description of variables

N°	Variable Name	Variable Description
Response Variables		
1	FemLifeExpBirth	Female life expectancy at birth
2	FemLifeExp_65	Female life expectancy at 65
3	MaleLifeExpBirth	Male life expectancy at birth
4	MaleLifeExp_65	Male life expectancy at 65
Covariates		
1	ALCOHOL	Pure alcohol consumption, liters per capita
2	BirthRate	Birth rate
3	CHICPROD	Production of chicken
4	CO2	Total Greenhouse Gas Emissions (in CO2 equivalent) indexed to 1990
5	CPI	Corruption Index Score
6	DEATHACC	Death rate due to accidents
7	DRAids	Death rate of Aids
8	DRAlcohol	Death rate due to alcohol abuse
9	DRCancer	Death rate due to cancer
10	DRChronic	Death rate due to chronic diseases
11	DRDiabetes	Death rate due to diabetes Mellitis
12	DRIschaemic	Death rate due to ischaemic heart disease
13	DRLiver	Death rate due to chronic liver disease
14	DRNervous	Death rate due to nervous system
15	DRPneumonia	Death rate due to pneumonia
16	ExpertCountry	Having experts make decisions about the country
17	ExpPrivPercTot	Total expenditure on private health as a percentage of total expenditure on health
18	GDP	GDP / capita at Purchasing power standard
19	Gini	Gini Measure of (income) inequality or concentration
20	Guidlines	Are there treatment guidelines available to GPs or Pulmologists for treating respiratory track infections
21	HopBeds	Hospital beds per 100000 inhabitants
22	HourWeek	Hours worked per week of full time employment
23	InfMort	Infant deaths per 1000 live births
24	IVHTB	Percentage of infants vaccinated against invasive disease due to Haemophilius influenzae type b
25	MeanHumidity	Average absolute humidity
26	OOPPH	Private households' out-of-pocket payment on health as % of total health expenditure
27	PDBEF5	Probability of dying before age 5
28	PEHPERCGSPEN	Public sector expenditure on health as % of total government expenditure, WHO estimates
29	PercSmoker	% of regular daily smokers in the population, age 15+
30	PERTUSIS	Percentage of infants vaccinated against Pertusis
31	Physicians	Practising physicians per 100,000
32	Pop	Population (on 1 Jan)

N°	Variable Name	Variable Description
33	POP0_14	% Population aged 0-14
34	PopDens	Average Population Density per km^2
35	PovertyRate	Poverty rate
36	PropPop65	% Population aged 65 and above
37	Religious	Religious person
38	RespAuthor	Greater respect for authority
39	RUBELLA	Percentage of infants vaccinated against Rubella
40	SchoolExp	Educational level School expectancy Years
41	SDHumidity	Standard deviation of absolute humidity
42	SDRBAE	Death rate due to bronchitis asthma & emphysema
43	SDRCRONLOWRES	Death rate due to chronic lower respiratory diseases
44	SDRINFLU	Death rate due to influenza
45	SDROARESINF	Death rate due to other acute respiratory infections
46	SDRRESSYS	Death rate due to diseases of the respiratory system
47	SUICIDE	Death rate due to suicide
48	TotalExpPercGDP	Total health expenditure as % of GDP
49	TOTALHE	Total health expenditure, PPP$ per capita, WHO estimates
50	TrustMost	Most people can be trusted
51	UpperSecond	Educational level Attainment upper secondary
52	URBANPOP	% of Urban Population
53	WomenMen	Women per men

Table A. 2: Effect of logarithmic transformation of variables with large variances

Descriptive Statistique	x					
	CHICPROD	GDP	HopBeds	Pop	PopDens	TOTALHE
			Data set 1			
Mean(x)	51846.46	22916.40	586.18	19827764.30	135.44	2178.40
SD(x)	56250.33	9838.32	178.10	23634764.12	109.35	1029.25
SD(log(x))	1.43	0.41	0.32	1.30	0.87	0.55
			Data set 2			
Mean(x)	51560.58	22960.29	586.72	19827764.30	135.51	2199.26
SD(x)	56230.34	9787.51	177.72	23634764.12	109.32	999.70
SD(log(x))	1.48	0.41	0.32	1.30	0.87	0.53
			Data set 3			
Mean(x)	50699.87	22984.62	586.75	19827764.30	135.50	2201.75
SD(x)	56753.10	9767.15	177.20	23634764.12	109.32	1016.12
SD(log(x))	1.56	0.41	0.32	1.30	0.87	0.54
			Data set 4			
Mean(x)	51573.50	22953.02	586.85	19827764.30	135.51	2197.29
SD(x)	56318.93	9798.49	177.14	23634764.12	109.32	1020.65
SD(log(x))	1.47	0.41	0.32	1.30	0.87	0.55
			Data set 5			
Mean(x)	50104.89	22911.22	586.95	19827764.30	135.48	2194.88
SD(x)	56284.51	9852.95	177.38	23634764.12	109.32	1047.53
SD(log(x))	1.60	0.42	0.32	1.30	0.87	0.57

Table A. 3: Orthogonal-Case Behavior of Selection and Regularization Methods

Penalty	**Estimate in Orthogonal-Predictors Linear Model Case**
Classical (Subset selection)	$\tilde{\beta}_j = \begin{cases} 0 & if\ \lvert\hat{\beta}_j\rvert < \lambda \\ \hat{\beta}_j & if\ \lvert\hat{\beta}_j\rvert > \lambda \end{cases}$
Ridge	$\tilde{\beta}_j = (1+\lambda)^{-1}\hat{\beta}_j$
LASSO	$\tilde{\beta}_j = \begin{cases} \hat{\beta}_j + \lambda & if\ \hat{\beta}_j < -\lambda \\ 0 & if\ \lvert\hat{\beta}_j\rvert < \lambda \\ \hat{\beta}_j - \lambda & if\ \hat{\beta}_j > \lambda \end{cases}$
Elastic-net	$\tilde{\beta}_j(naive\ elastic - net) = \begin{cases} 0 & if\ \lvert\hat{\beta}_j\rvert \le \alpha\lambda/2 \\ \frac{\lvert\widehat{\beta}_j\rvert + \alpha\lambda/2}{1+(1-\alpha)\lambda} sign(\hat{\beta}_j) & if\ \lvert\hat{\beta}_j\rvert > \alpha\lambda/2 \end{cases}$
SCAD	$\tilde{\beta}_j = \begin{cases} 0 & if\ \lvert\hat{\beta}_j\rvert \le \lambda \\ \hat{\beta}_j - \lambda\, sign(\hat{\beta}_j) & if\ \lambda < \lvert\hat{\beta}_j\rvert < 2\lambda \\ \left\{\frac{a-1}{a-2} - \frac{a\, sign(\widehat{\beta}_j)}{a-2}\lambda\right\} & if\ 2\lambda < \hat{\beta}_j < a\lambda \\ \hat{\beta}_j & if\ \widehat{\beta}_j > a\lambda \end{cases}$
SCAD-L$_2$	$\tilde{\beta}_j = \begin{cases} 0 & if\ \lvert\hat{\beta}_j\rvert \le \lambda \\ \hat{\beta}_j/(1+2(1-\alpha)\lambda) & if\ \lambda < \lvert\hat{\beta}_j\rvert < a\lambda \\ \hat{\beta}_j & if\ \widehat{\beta}_j > a\lambda \end{cases}$

Table A. 4: Missingness pattern and their occurrence frequencies for the life expectancy at birth and the life expectancy at age 65 for both males and females (O: observed, M: Missing)

1999	2000	2001	2002	2003	2004	2005	2006	2007	Count	%
					Complete					
O	O	O	O	O	O	O	O	O	**12**	**50.00**
					Monotone Missingness					
O	O	O	O	O	M	M	M	M	1	4.17
O	O	O	O	O	O	O	O	M	4	16.67
Subtotal									**5**	**20.83**
					Non-Monotone Missingness					
O	O	O	O	M	O	O	M	O	1	4.17
M	M	M	O	M	O	O	O	O	1	4.17
O	O	O	O	M	O	O	O	O	1	4.17
M	M	M	O	O	O	O	O	O	2	8.33
O	M	M	O	O	O	O	O	O	1	4.17
M	M	O	O	O	O	O	O	O	1	4.17
Subtotal									**7**	**29.17**

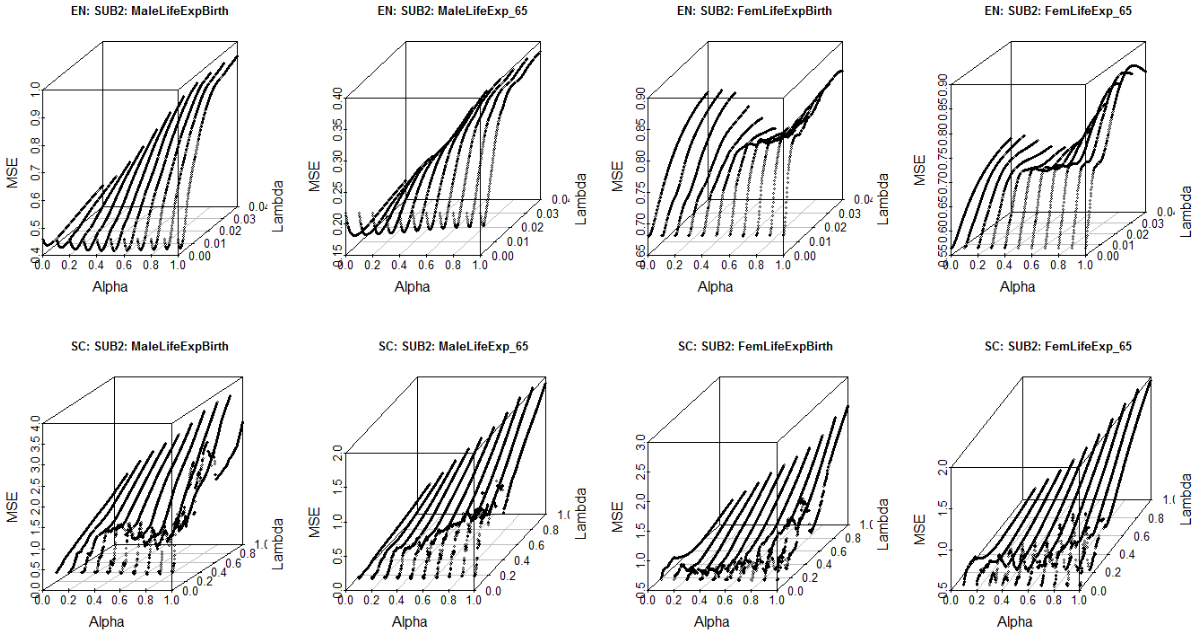


Figure A. 1: Two-dimensional grid cross-validation (on PGEE) for the pair (α,λ): Data set 2

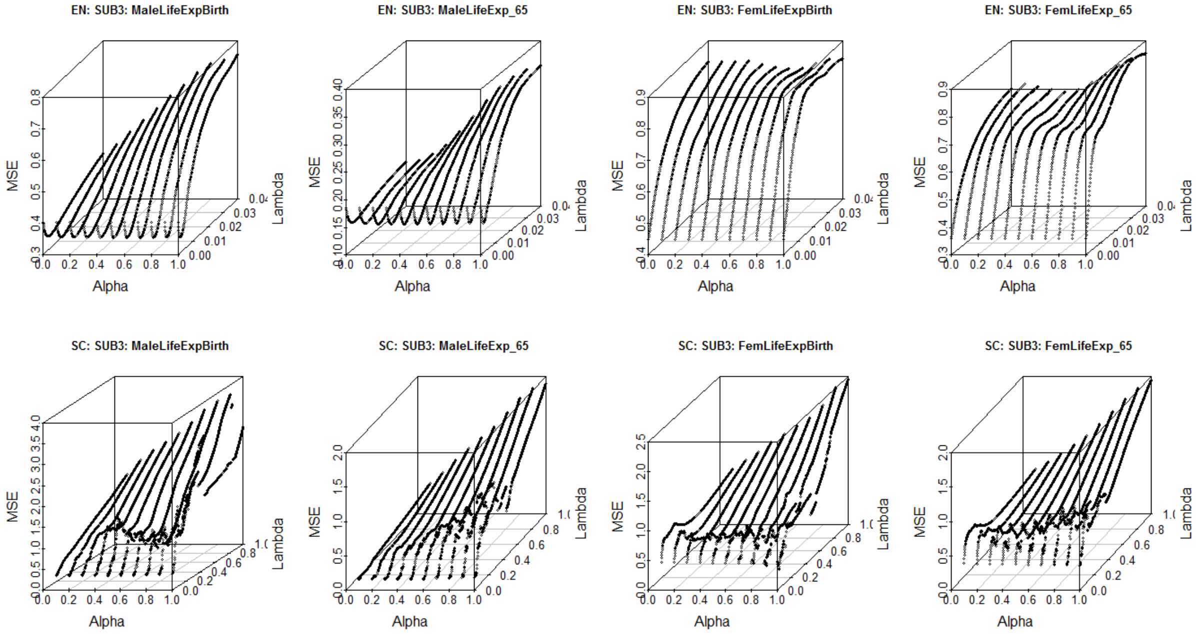


Figure A. 2: Two-dimensional grid cross-validation (on PGEE) for the pair (α,λ): Data set 3

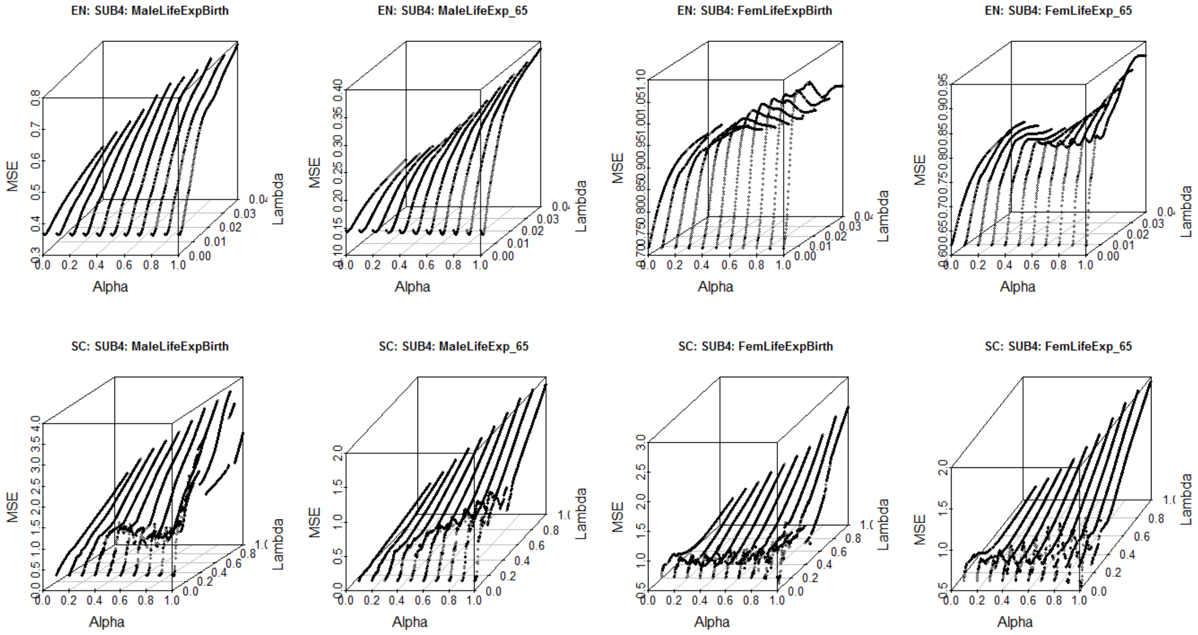


Figure A. 3: Two-dimensional grid cross-validation (on PGEE) for the pair (α,λ): Data set 4

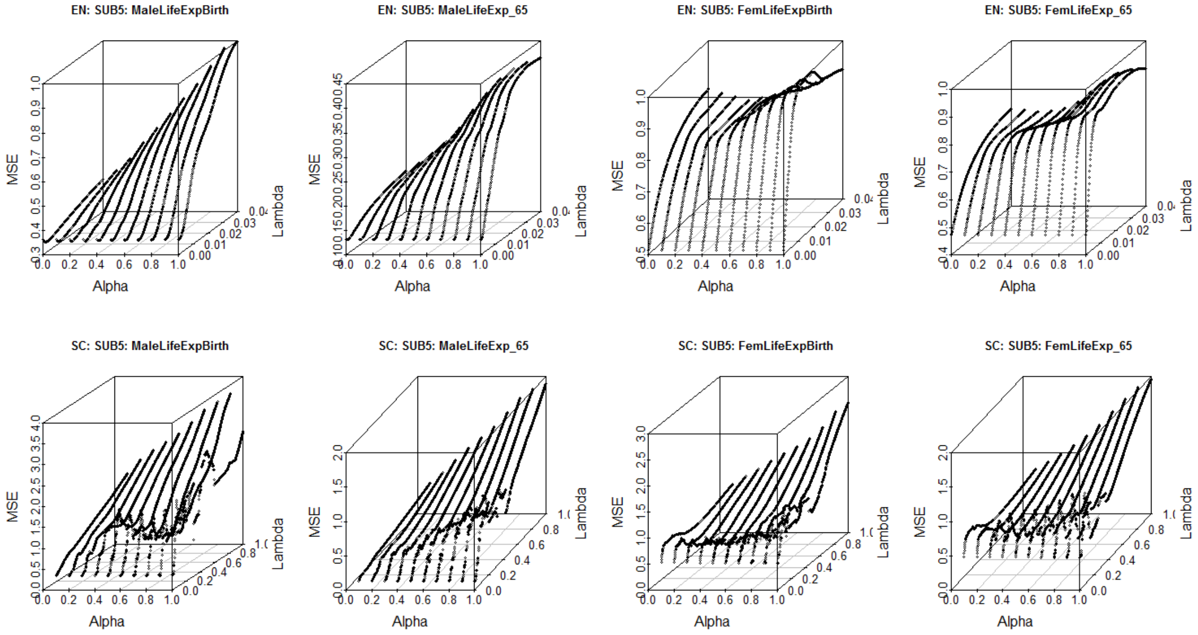


Figure A. 4: Two-dimensional grid cross-validation (on PGEE) for the pair (α,λ): Data set 5

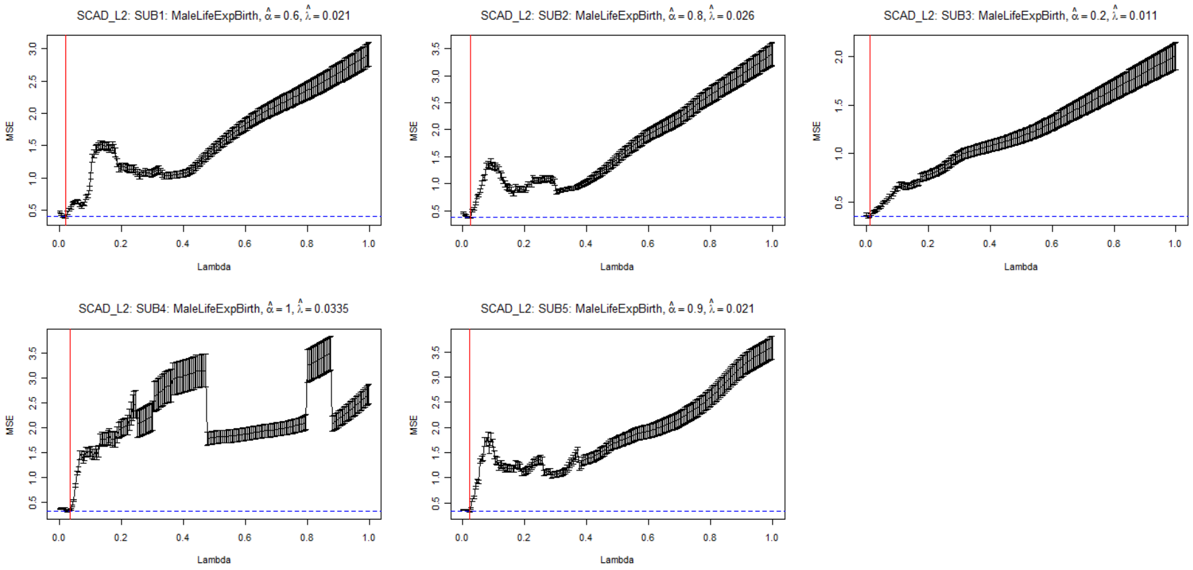


Figure A. 5: "One-standard error" rule application (MaleLifeExpBirth, Data set 1 to 5)

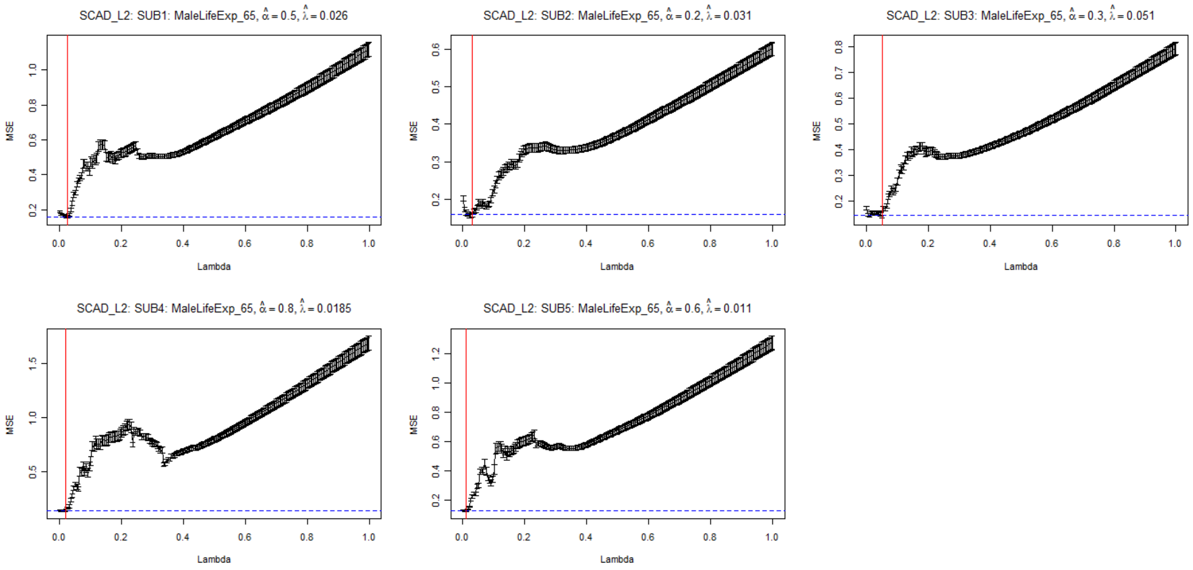


Figure A. 6: "One-standard error" rule application (MaleLifeExp_65, Data set 1 to 5)

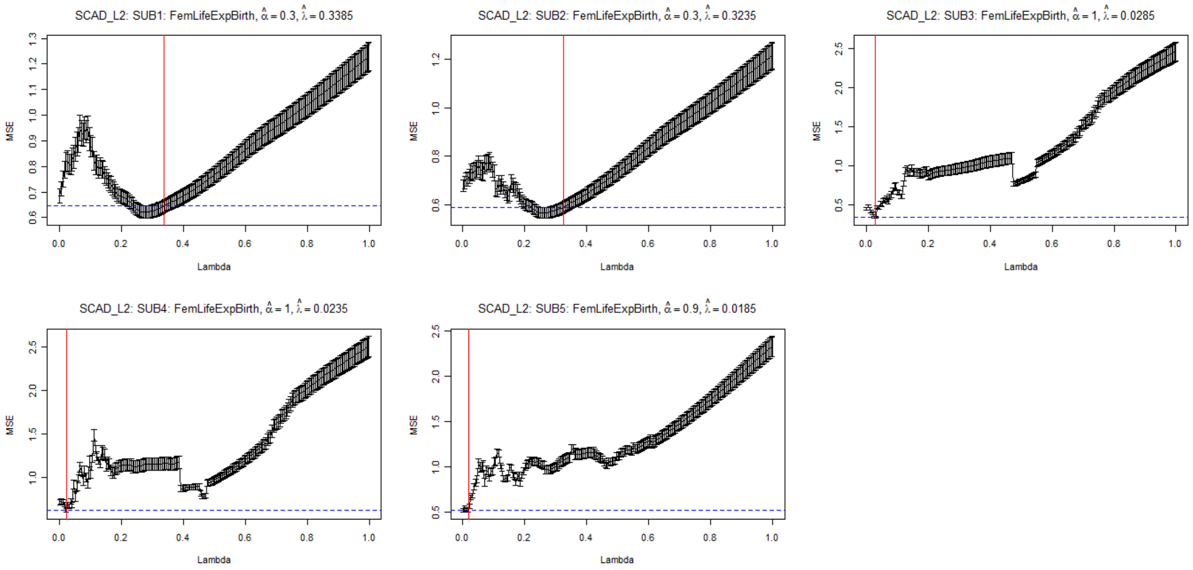


Figure A. 7: "One-standard error" rule application (FemLifeExpBirth, Data set 1 to 5)

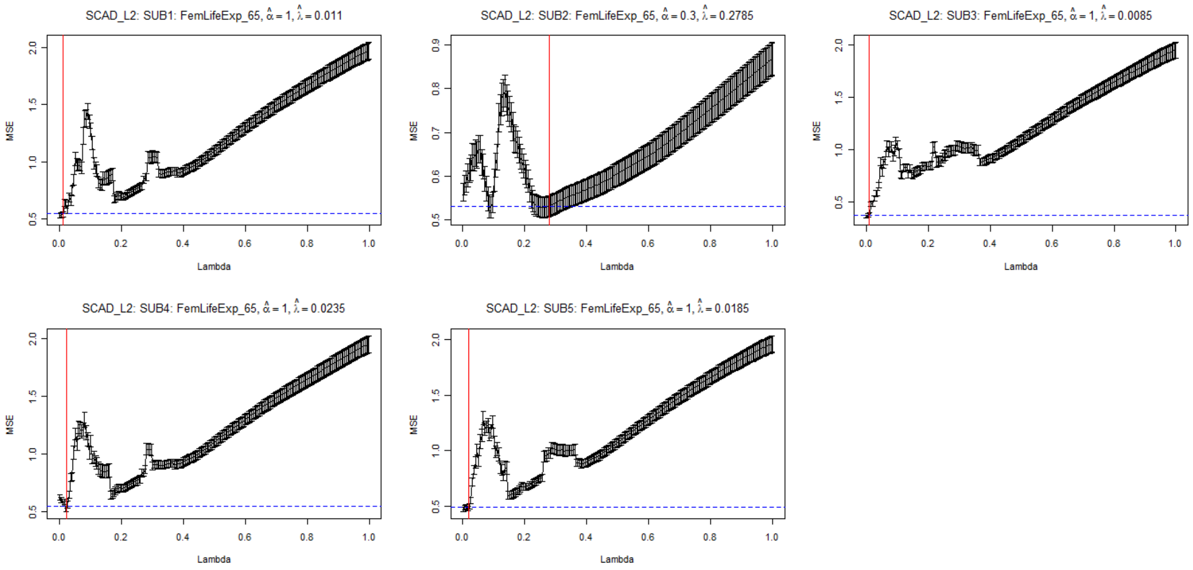


Figure A. 8: "One-standard error" rule application (FemLifeExp_65, Data set 1 to 5)

Table A. 5: Penalized coefficient estimates (Male life expectancy at birth)

NAMEVAR	Data set 1	Data set 2	Data set 3	Data set 4	Data set 5
MeanHumidity	0.1280	0.0778	0.1002	0.0000	0.0000
SDHumidity	-0.1224	-0.0786	-0.1142	0.0000	-0.1529
TrustMost	-0.0904	0.0000	-0.1037	0.0000	0.0000
ExpertCountry	0.0000	0.0000	0.0480	0.0000	0.0000
Religious	-0.1218	-0.1558	-0.1072	-0.2607	-0.1850
WomenMen	-0.3998	-0.5577	-0.4750	-0.4307	-0.5345
UpperSecond	0.1328	0.0000	0.0446	0.0000	0.0000
SchoolExp	0.0000	0.0000	0.0322	0.0000	0.0000
PropPop65	0.2504	0.3389	0.3262	0.5335	0.3642
PovertyRate	0.0000	0.0000	0.0414	0.0000	0.0000
PopDens	-0.0822	0.0000	-0.0523	0.0000	0.0000
Pop	0.2437	0.2454	0.1581	0.3596	0.2252
Physicians	0.4104	0.4747	0.3590	0.4385	0.4476
InfMort	-0.1306	-0.1482	-0.1719	-0.2994	-0.2056
HourWeek	0.1413	0.1629	0.1329	0.2481	0.2068
HopBeds	-0.2866	-0.3259	-0.2397	-0.3690	-0.4284
GDP	0.2953	0.1941	0.2688	0.2561	0.2555
DRPneumonia	-0.1635	-0.1536	-0.2701	-0.2741	-0.2742
DRNervous	0.0000	0.0000	0.0451	0.0000	0.1449
DRLiver	-0.1314	-0.2012	-0.1456	-0.2427	-0.1590
DRIschaemic	-0.3190	-0.2222	-0.1769	-0.2323	-0.2494
DRDiabetes	-0.0946	-0.1242	-0.0714	0.0000	0.0000
DRChronic	-0.3158	-0.2660	-0.4730	-0.5159	-0.4092
DRCancer	-0.3399	-0.2205	-0.1316	0.0000	-0.1453
DRAlcohol	-0.2174	-0.1697	-0.1996	-0.4161	-0.1840
DRAids	-0.4785	-0.4319	-0.4423	-0.3882	-0.4506
BirthRate	0.1712	0.2822	0.1578	0.5023	0.2275
SDRRESSYS	-0.1504	-0.2541	-0.0786	-0.1325	-0.1705
SDRINFLU	-0.0172	-0.0056	-0.0149	0.0000	0.0000
SDROARESINF	-0.1829	-0.1693	-0.1528	-0.0004	-0.1808
SDRBAE	0.0000	0.0000	-0.0665	0.0000	-0.2125
SDRCRONLOWRES	0.0000	0.0000	0.0213	0.0000	0.0912
CPI	0.3232	0.4963	0.3634	0.6724	0.4380
POP0_14	0.0000	0.0000	-0.0380	0.0000	0.0000
URBANPOP	-0.0650	-0.1351	-0.1321	-0.3070	-0.1861
ALCOHOL	0.0000	0.0000	-0.0645	0.0000	-0.0906
IVHTB	0.0945	0.0826	0.0767	0.0000	0.0898
OOPPH	0.0000	0.0000	-0.0291	0.0000	0.0000
PERTUSIS	-0.1374	-0.1726	-0.1096	-0.0011	-0.1363
RUBELLA	0.0000	0.0000	-0.0145	0.0000	0.0000
TOTALHE	0.0000	0.0000	0.0936	0.0000	0.0000
PEHPERCGSPEN	0.1194	0.1926	0.1269	0.2788	0.2127
CHICPROD	0.1139	0.1610	0.1776	0.0000	0.1462
PDBEF5	-0.1852	-0.1663	-0.1120	0.0000	-0.0975
PercSmoker	-0.1393	-0.1278	-0.1657	0.0000	0.0000
ExpPrivPercTot	-0.1112	-0.1118	-0.1135	-0.1960	-0.1475
RespAuthor	-0.0658	0.0000	-0.0606	0.0000	0.0000
TotalExpPercGDP	0.1010	0.0000	0.0764	0.0000	0.0000
Gini	0.0000	0.0000	0.0104	0.0000	0.0000
SUICIDE	0.0626	0.1575	0.1133	0.3370	0.2246
DEATHACC	-0.8331	-0.8913	-0.8683	-0.9360	-0.9601
CO2	0.1984	0.1824	0.1483	0.0000	0.0000
Guidlines	0.0664	0.0000	-0.0179	0.0000	0.0000

Table A. 6: Final model for male life expectancy at birth (MaleLifeExpBirth)

Parameter	Estimate	Std Error	95% Confidence Limits		P-value
intercept	62.7443	3.1562	56.2362	69.2525	<.0001
			Positive effect		
BirthRate	0.1421	0.0315	0.0799	0.2044	<.0001
CPI	0.1754	0.0285	0.1168	0.2341	<.0001
HourWeek	0.1623	0.0267	0.1090	0.2156	<.0001
Log(GDP)	2.2732	0.2792	1.6573	2.8891	<.0001
Log(Pop)	0.4232	0.0453	0.3269	0.5196	<.0001
PEHPERCGSPEN	0.0887	0.0173	0.0544	0.1231	<.0001
Physicians	0.0062	0.0007	0.0047	0.0077	<.0001
PropPop65	0.2307	0.0267	0.1774	0.2839	<.0001
SUICIDE	0.0755	0.0092	0.0574	0.0935	<.0001
			Negative effect		
DEATHACC	-0.0475	0.0050	-0.0574	-0.0375	<.0001
DRAids	-0.1211	0.0304	-0.1853	-0.0568	0.0010
DRAlcohol	-0.0701	0.0156	-0.1033	-0.0370	0.0004
DRChronic	-0.0139	0.0020	-0.0181	-0.0098	<.0001
DRDiabetes	-0.0162	0.0062	-0.0285	-0.0040	0.0097
DRLiver	-0.0198	0.0061	-0.0319	-0.0078	0.0015
DRPneumonia	-0.0171	0.0070	-0.0322	-0.0020	0.0298
ExpPrivPercTot	-0.0231	0.0046	-0.0323	-0.0139	<.0001
InfMort	-0.0741	0.0332	-0.1398	-0.0085	0.0271
Log(HopBeds)	-1.2851	0.1217	-1.5245	-1.0458	<.0001
PercSmoker	-0.0141	0.0077	-0.0296	0.0014	0.0729
Religious	-1.2148	0.1949	-1.6042	-0.8255	<.0001
SDHumidity	-0.1174	0.0366	-0.1910	-0.0439	0.0024
SDRRESSYS	-0.0096	0.0038	-0.0176	-0.0015	0.0228
URBANPOP	-0.0245	0.0036	-0.0316	-0.0173	<.0001
WomenMen	-0.1541	0.0250	-0.2073	-0.1008	<.0001

Table A. 7: Final model for male life expectancy at age 65 (MaleLifeExp_65)

Parameter	Estimate	Std Error	95% Confidence Limits		P-value
intercept	3.3422	2.3148	-1.3073	7.9916	0.1550
			Positive effect		
CPI	0.1194	0.0278	0.0648	0.1741	<.0001
HourWeek	0.0651	0.0224	0.0209	0.1093	0.0041
Log(GDP)	1.3512	0.1452	1.0554	1.6470	<.0001
Log(Pop)	0.3043	0.0257	0.2538	0.3547	<.0001
Physicians	0.0037	0.0005	0.0027	0.0046	<.0001
TotalExpPercGDP	0.1096	0.0248	0.0610	0.1583	<.0001
TrustMost	1.1850	0.3319	0.5321	1.8380	0.0004
			Negative effect		
DEATHACC	-0.0070	0.0025	-0.0120	-0.0020	0.0062
DRAids	-0.1821	0.0170	-0.2156	-0.1486	<.0001
DRCancer	-0.0122	0.0014	-0.0150	-0.0095	<.0001
DRIschaemic	-0.0050	0.0006	-0.0062	-0.0039	<.0001
DRPneumonia	-0.0141	0.0037	-0.0214	-0.0067	0.0003
ExpPrivPercTot	-0.0112	0.0034	-0.0181	-0.0043	0.0019
Log(HopBeds)	-0.7178	0.1042	-0.9251	-0.5104	<.0001
Log(PopDens)	-0.3857	0.0347	-0.4538	-0.3176	<.0001
PercSmoker	-0.0218	0.0055	-0.0331	-0.0106	0.0005
Religious	-0.5577	0.1167	-0.7878	-0.3276	<.0001
RespAuthor	-0.6261	0.1620	-0.9458	-0.3064	0.0002
SDHumidity	-0.1122	0.0278	-0.1684	-0.0560	0.0002
SDRINFLU	-0.1281	0.0406	-0.2102	-0.0461	0.0030
SDRRESSYS	-0.0052	0.0023	-0.0098	-0.0006	0.0289

Table A. 8: Final model for female life expectancy at birth (FemLifeExpBirth)

Parameter	Estimate	Std Error	95% Confidence Limits		P-value
intercept	76.0788	3.0715	70.0140	82.1437	<.0001
Positive effect					
CO2	0.0098	0.0024	0.0050	0.0145	<.0001
Log(GDP)	1.3131	0.1740	0.9666	1.6597	<.0001
Log(Pop)	0.2825	0.0270	0.2294	0.3356	<.0001
Physicians	0.0038	0.0007	0.0023	0.0052	<.0001
WomenMen	0.0626	0.0198	0.0227	0.1026	0.0029
Negative effect					
DRAids	-0.2567	0.0275	-0.3113	-0.2022	<.0001
DRCancer	-0.0181	0.0032	-0.0246	-0.0115	<.0001
DRChronic	-0.0117	0.0027	-0.0173	-0.0061	0.0003
DRIschaemic	-0.0080	0.0012	-0.0104	-0.0055	<.0001
DRPneumonia	-0.0243	0.0049	-0.0344	-0.0142	<.0001
ExpertCountry	-1.0738	0.1322	-1.3346	-0.8130	<.0001
InfMort	-0.1200	0.0436	-0.2080	-0.0320	0.0087
Log(HopBeds)	-0.5304	0.1118	-0.7508	-0.3100	<.0001
Log(PopDens)	-0.5844	0.0603	-0.7032	-0.4657	<.0001
OOPPH	-0.0348	0.0078	-0.0508	-0.0189	<.0001
PEHPERCGSPEN	-0.0325	0.0172	-0.0664	0.0014	0.0600
POP0_14	-0.0858	0.0263	-0.1381	-0.0336	0.0016
RespAuthor	-1.2151	0.1793	-1.5726	-0.8576	<.0001
SDHumidity	-0.1739	0.0400	-0.2529	-0.0948	<.0001

Table A. 9: Final model for female life expectancy at age 65 (FemLifeExp_65)

Parameter	Estimate	Std Error	95% Confidence Limits		P-value
intercept	9.7330	3.1590	3.4924	15.9736	0.0024
Positive effect					
Log(GDP)	1.4721	0.1747	1.1252	1.8191	<.0001
Log(Pop)	0.2985	0.0399	0.2203	0.3767	<.0001
Physicians	0.0030	0.0006	0.0019	0.0041	<.0001
SUICIDE	0.0402	0.0121	0.0151	0.0654	0.0033
TotalExpPercGDP	0.1481	0.0420	0.0616	0.2347	0.0017
TrustMost	1.2815	0.3476	0.5911	1.9719	0.0004
WomenMen	0.0644	0.0170	0.0305	0.0982	0.0003
Negative effect					
DRAids	-0.2225	0.0314	-0.2846	-0.1604	<.0001
DRCancer	-0.0174	0.0027	-0.0227	-0.0121	<.0001
DRChronic	-0.0048	0.0026	-0.0100	0.0004	0.0709
DRIschaemic	-0.0071	0.0013	-0.0098	-0.0044	<.0001
DRPneumonia	-0.0253	0.0041	-0.0336	-0.0170	<.0001
Log(HopBeds)	-0.9163	0.1286	-1.1684	-0.6642	<.0001
Log(PopDens)	-0.5463	0.0528	-0.6514	-0.4412	<.0001
OOPPH	-0.0236	0.0053	-0.0342	-0.0131	<.0001
PEHPERCGSPEN	-0.0805	0.0203	-0.1219	-0.0392	0.0004
POP0_14	-0.0978	0.0219	-0.1410	-0.0547	<.0001
PropPop65	-0.0975	0.0268	-0.1502	-0.0447	0.0003
RespAuthor	-1.3379	0.1941	-1.7188	-0.9569	<.0001
SDHumidity	-0.0991	0.0341	-0.1672	-0.0309	0.0051
SDROARESINF	-0.1091	0.0174	-0.1440	-0.0742	<.0001

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Identifying determinants of life expectancy in the EU

Richting: **Master of Statistics-Biostatistics**
Jaar: **2011**

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,

Batomen, Francis

Datum: **12/09/2011**