Ludwig-Maximilians-University
Munich

Department of Statistics

Bachelorthesis

Subjective Well-Being over the Life Span: Modeling Age-, Period-, and Cohort-Effects in the Additive Mixed Model Framework

Author:
Felix Günther

Supervisor:
Prof. Dr. Helmuth Küchenhoff

August 12, 2014

The analysis of panel data is often concerned with effects of time-related changes. Dealing with data on individuals, the variation of some dependent variable one is interested in can be associated to differences in the specific ages of the individuals, the different birth-cohorts they belong to, and changes in the period of observation. Because of the the Age-Period-Cohort Identification problem it is, however, in general not possible to estimate distinct effects for each of the three covariates.

This thesis provides an approach for analyzing such data in the additive mixed model framework. Estimating two smooth effects, the age-effect and the cohort- or period-effect, and additionally an interaction surface of both it can be argued that the model takes all three time-related changes into account.

This modeling approach is illustrated by an application on german socio-economic data regarding subjective well-being over the life span and various graphics for an evaluation of such models are presented in this context.

Contents

1 Introduction

1.1 Well-Being over the Life Span

The level of subjective well-being of individuals as well as its average in societies is a matter of great public, political, and scientific interest. This relevance arises from its direct connection to, respectively dependency on the individual's degree of happiness and satisfaction.

There exist many theories about influencing factors on subjective well-being. From a psychologists perspective there are generally three groups of influencing factors: the extent of fulfilled needs and suppressed discomfort, the satisfaction with the activities someone is engaged in, and personal and genetic predispositions (Diener et al. (2009)). Most empirical studies (often also done by sociologists or economists) monitor effects that can in some way be assigned to one of those three areas.

A question that arises when being concerned with subjective well-being is how it changes over the life span. Aging can be seen as a process of physiological and cognitive deterioration what suggests that the subjective well-being of individuals decreases during life. Otherwise it is possible that personal aims and needs change over time or adapt to the personal circumstances and are therefore easier fulfilled than in young ages, what would lead to a higher amount of well-being during later stages in one's life.

Empirical evidence on the effect of aging on well-being leads to the insight that there exists in general no linear development over the life-span. But the findings of previous studies diverge, some assume a U-shaped trend while others argue for the opposite, an inverse U-shape. It seems therefore reasonable to fit the age-effect on well-being in a non-parametric way and avoid to make a-priori assumptions regarding the structure of connection in this way (Wunder et al. (2013)).

1.2 Age-Period-Cohort Identification Problem

The relation of age, period and cohort to dependent factors is a matter of interest in many scientific disciplines such as epidemiology, psychology, sociology, economics and others. Especially when dealing with temporarily ordered (panel) data - measurements on individuals or groups of individuals repeated along a time dimension - the question arises how changes in those three time-related areas cohere with the factor of interest. In an Age-Period-Cohort (APC) analysis it is therefore necessary to consider three possible effects: *Age-effects* are variations of the dependent factor associated with different age-groups. Theoretical reasons can be found in biological and social processes over the lifespan, also referred to as *aging*. *Period-effects* are changes due to time-specific events or developments that have effects on all age groups simultaneously. And *Cohort-effects* are variations across groups born in the same time. Those groups experience historical and social events at same ages and share a collective environment.

Addressing the well-being over the life span, period effects could arise from economical or political conditions and specific events accompanying the survey year while cohort effects can describe e.g. the employment situation at the cohorts transition from education into to the labour market. Controlling for such period and cohort effects is necessary when being interested in the relation of age and well-being since neglecting them would result in biased estimates of the age effect.

The statistical modeling of such structures though features an identification problem. It results from the fact that the factors age, period and cohort are linearly connected. The relationship is $(age) + (cohort) = (period)$. In a regression model with age, period and cohort as covariates this leads to a collinearity problem and it is not possible to identify distinct effects of them. Clayton and Schifflers (1987) describe this identification problem as a general scientific one, that can not be solved through methodological achievements because the available information just does not allow to distinguish between the three effects. However it is still possible to get insights into the structure of the available data by analyzing it in a reasonable and systematic way.

There exist several approaches to deal with such datasets, it is for example possible to impose a linear constraint on at least one of the three influencing factors, a common practice is to define *a-priori* that two different values of age, period or cohort are estimated with the numerically same effect (Mason et al. (1973), Kupper et al. (1985)). The choice of such a constraint can be made either based on theoretical considerations or by looking at the data but is always somewhat arbitrary and has to be made very carefully. Rodgers (1982) shows by using artificial data that a wrong constraint (in the sense of determining two parameters being the same while they are not the same in the generated data) - even if it is not unreasonable (the two parameters are not that different) - leads to major effects on the estimates of the other parameters. It is not necessarily possible to detect such a more or less poor selection of the APC constraint and consequently invalid parameter estimates through measures of the model fit.

Another way of dealing with the APC problem is trying to identify specific variables lying behind the period or cohort effects and include them explicitly as covariates into the model. Instead of modeling a period effect it is for example possible to include economic indicators like the unemployment rate and the gross national product and additionally dummy variables for specific events that are assumed to have an influence on the target variable. The cohort effect could be replaced by a persons life expectancy at birth, alternatively it can be argued that the cohort effects are captured by other covariates (e.g. a cohort specific exhibition to harmful life circumstances by health variables, and cohort effects on the labor market by unemployment rates).

Of course it is disputable which factors have to be included and if the combination of covariates is adequately describing the desired effect. If not, the estimates of age and the other factor are biased as well.

Both of those approaches can lead to insights into the structure of APC effects on some dependent factors but the made assumptions are quite strong and have major effects on the results of the analysis, while one can never be sure if they are really valid. Especially when considering also the possibility of interactions between the age, period and cohort effects - it is for example possible that a specific event (period effect) can have dissimilar effects on different cohorts - it gets really difficult to judge the quality and informative value of such models.

In Biometrics (especially in the area of disease incidence and mortality data), there exists another popular approach of dealing with the APC problem that is referred to as *Holford's-parametrization* (Holford (1983)). The procedure is to partition the age, period and cohort effects each into two orthogonal components, one for their overall linear trend and the second representing deviations from this linear trend (the *curvature effects*). It is the then possible to estimate each curvature effect and certain linear combinations of the linear effects (Jiang and Carriere (2014) show that the curvature effects can also be estimated with penalized cubic regression splines). Holford's parametrization enables insights into the APC effect structure on some dependent variables but does not allow to estimate a distinct effect of each of the three predictor variables and considers no interactions between them.

An inherent feature of the APC structure is that the marginal effect of one of the three variables is automatically part of the interaction space of the two others while their marginal effects are also expressed in this interaction. Using tensor-product splines over a B-spline base Heuer (1997) estimates the APC effects in the framework of Holford's parametrization by an interaction of age and period effects (*API*-model) and compares the performance of this model with the classic Holford's APC-model including all three predictor variables on some simulated

data including age-period interactions. The API-model represents the data structure considerably better than the classic APC-model, which is heavily biased if there exist interactions between age and period in the data. Therefore this approach of estimating the APC-structure through an interaction surface of two covariates is very promising, nevertheless it did not gain a big popularity yet.
This thesis will again present such an approach, embed it into the framework of additive mixed models and apply it on socio-economic data from Germany regarding subjective well-being over the life-span.

2 Theoretical Considerations and Basics

2.1 Flexible Smoothing with Splines and Penalties

There are many tasks in which a linear dependence of the response (dependent) on the predictor (influencing) variables is not reasonable. It is possible to include covariates as higher degree polynomials into a Linear Model and estimate non-linear relations in that way, but this procedure needs a-priori informations about the structure of the connection between the response and predictor variable because the specified degree of the polynomial determines the shape of the relation. Modeling non-linear effects without such a-priori knowledge is the aim of *non-parametric regression.* It allows an automatic and data driven estimation of a flexible relationship. The general form of such a smoothing model is

$$y = \sum_{z=1}^{s} f_z(x_z) + \epsilon$$

with $f_z(\cdot)$ being a smooth function of the (continuos) covariate x_z. This covariate can also be a multivariate variable, in the bivariate case $f_z(\cdot)$ is then a surface instead of a curve. Basically the estimation of such models needs to solve two tasks: the smooth functions need to be represented in some way and the extent of smoothness has to be determined (Wood (2006a)). In the following the approach of smoothing with penalized splines (p-splines, Eilers and Marx (1996)) is described, at first in the case of a single univariate and afterwards for a bivariate variable. Section 2.2 presents a way to embed such single splines or surfaces into a general regression framework.

2.1.1 Univariate

The idea of (univariate) smoothing with a B-spline base is to construct a curve/spline from piecewise polynomials, which are linked smoothly at some predefined m knots. Those knots are often placed equidistant in the range $[a, b]$ of the covariate z. The degree l of the polynomials has to be chosen a priori.

At every point $z \in [a, b]$ the B-spline base consists of $(l + 1)$ (positive) polynomial pieces and the resulting function (defined as the weighted sum of the basis functions B_j, see Equation 2.1) is, in particular, at the knots - where two basis functions join - $(l-1)$-times continuously differentiable. Therefore it is reasonable to use cubic splines ($l = 3$) because the resulting spline $f(z)$ is then overall two-times continuously differentiable, what is an often claimed smoothness criteria. Furthermore it can be seen that every basis function is positive on $l + 2$ adjacent knots, everywhere else it is 0. It is necessary to add $2l$ knots outside the range of z (l at each side) to have the described properties even at the boarders of the predictor variables range fulfilled. Therefore it can be seen that the dimension (the amount of basis functions) of the B-spline base is $d = m + l - 1$. A last feature of the B-spline base is that the sum of all basis functions at some point $z \in [a, b]$ is $\sum_{j=1}^{d} B_j(z) = 1$. To illustrate these features Figure 2.1 shows a cubic B-spline base over 9 knots.

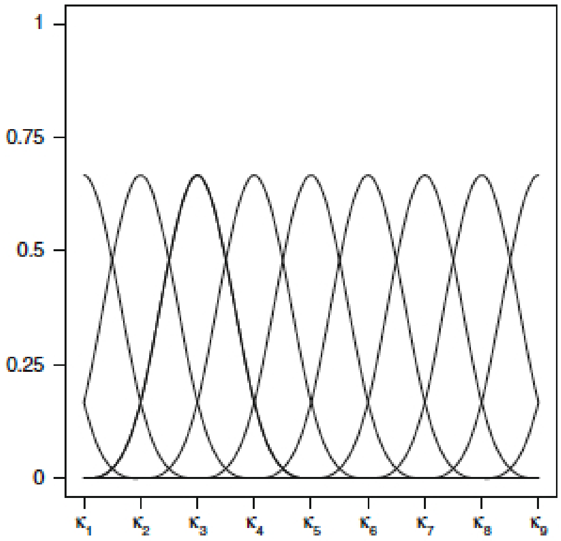


Figure 2.1: Cubic B-spline base, $m = 9$ (Fahrmeier et al. (2013), p. 428)

With a B-spline base specified like this, the smooth curve $f(z)$ is then constructed as

$$f(z) = \sum_{j=1}^{d} B_j(z)\gamma_j \tag{2.1}$$

and so the aim of the regression is to estimate the vector γ. The estimated $\hat{\gamma}_j$ values are the amplitude for the scaling of the j-th basis function.
Looking at Equation 2.1 it is easy to see that the estimation of the polynomial

spline is in fact a linear model estimation with the design matrix

$$Z = \begin{pmatrix} B_1(z_1) & \cdots & B_d(z_1) \\ \vdots & & \vdots \\ B_1(z_n) & \cdots & B_d(z_n) \end{pmatrix}.$$

With this B-Spline base there is a mathematically reasonable representation of a smooth function established, the second task is now to determine the amount of smoothness of this spline in a data driven way.
This amount of smoothness or roughness is mainly controlled by the number of knots, or basis functions, provided to the estimation. Too many knots will result in a too close interpolation of the data points and therefore *overfitting*, while too few knots do not allow the fitted curve to represent accurately the *real* underlying structure of relation. Eilers and Marx (1996) introduce penalized splines (p-splines) to deal with the tradeoff between fit and "wiggliness" of the estimated function, which are now a well established and widespread tool for nonparametric regression via spline smoothing.
The general idea is allowing the function to be fitted with a sufficiently large number of knots m what ensures the estimation to cover even very complex functions. At the same time a penalty term is introduced to prevent overfitting. It is straightforward to use derivatives as measure for the "wiggliness" of a function. Using the second derivative is particularly interesting in the context of p-splines on a cubic spline-base, since it exists over the whole range of the function and measures the curvature of the spline function. The penalty term is then constructed as

$$\lambda \int (f''(z))^2 dz.$$

Following Fahrmeier et al. (2013)(p. 430/434) the first derivative for a single spline of the B-spline base of degree l is

$$\frac{\partial}{\partial z} B_j^l(z) = l \cdot \left(\frac{1}{k_j - k_{j-1}} B_{j-1}^{l-1}(z) - \frac{1}{k_{j-1} - k_{j+1-l}} B_j^{l-1}(z) \right).$$

The derivative of the whole polynomial spline is then

$$\frac{\partial}{\partial z} \sum_j \gamma_j B_j^l(z) = l \cdot \sum_j \frac{\gamma_j - \gamma_{j-1}}{k_j - k_{j-l}} B_{j-1}^{l-1}(z), \tag{2.2}$$

and can therefore be expressed through the differences of adjacent basis coefficients and B-spline functions of one lower degree. Consequently the estimation of the coefficient vector γ leads automatically to the derivative of the spline.

Eilers and Marx (1996) propose to use just the squared sum of the (k-th order, $k \geq 2$) differences of adjacent B-spline coefficients as penalty term, since the mathematical quest is less complex than using the explicit derivations and the resulting difference penalty approximates the integrated square of the k-th derivative well.

The combination of this penalty term and the constructed B-spline base of one covariate leads to the following penalized sum of squares minimization criteria for fixed $\lambda > 0$:

$$PLS(\gamma) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{d} \gamma_j B_j(z_i) \right)^2 + \lambda \sum_{j=k+1}^{d} (\Delta^k \gamma_j)^2, \tag{2.3}$$

with $\sum_{j=k+1}^{d} (\Delta^k \gamma_j)^2$ equals $\gamma' K_k \gamma = \gamma' D_k' D_k \gamma$ and D_k being the k-th order difference matrix defined as $D_k = D_1 D_{k-1}$ with

$$D_1 = \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{pmatrix}.$$

To minimize the penalized sum of squares and consequently achieve the estimation of $\hat{\gamma}$, the derivative with respect to γ needs to be calculated and set to zero:

$$\begin{aligned} PLS(\gamma) &= (y - Z\gamma)'(y - Z\gamma) + \lambda \gamma' K \gamma \\ &= y'y - 2\gamma' Z' y + \gamma'(Z'Z + \lambda K)\gamma \qquad (2.4) \\ \frac{\partial}{\partial \gamma} PLS(\gamma) &= -2Z'y + 2(Z'Z + \lambda K)\gamma \overset{!}{=} 0 \\ \Rightarrow \qquad \hat{\gamma} &= (Z'Z + \lambda K)^{-1} Z' y. \qquad (2.5) \end{aligned}$$

Looking at Equation 2.3 helps to see how the smoothing parameter λ determines the estimation of the spline function. For $\lambda \to 0$ we get an unpenalized spline estimation over a B-spline base and therefore just an interpolation of the data points through a spline of order m (with $\lambda = 0$ Equation 2.4 is exactly the least squares estimator of an unpenalized linear model).

For $\lambda \to \infty$ the fitted function gets a polynomial of degree $k - 1$ if the degree of the basis functions $l \geq k$ (in the case of second order differences and a cubic spline base the result is a straight line). For such large λ's γ is estimated by minimizing the $\Delta^k \gamma_j$'s. In the case of first order differences this equals minimizing $\Delta^1 \gamma_j = \gamma_j - \gamma_{j-1}$, over all j. Therefore all parameters are estimated as equal and the first derivation of the whole polynomial spline (Equation 2.2) is zero. For higher order differences the corresponding higher order derivatives of the polynomial spline get zero (Fahrmeier et al. (2013), p. 435).

As a consequence, the problem of a data-driven selection of smoothness corresponds to the estimation of λ from the data. Methods for this estimation are divided into two groups, the first searches for the best λ, which minimizes the prediction error of the model, by looking at the changes of criteria like the Akaike's information criterion (AIC) or (generalized) cross-validation (G-/CV) for different λ (Eilers and Marx, 1996).

The other group reparametrizes the p-spline model as a mixed model, estimates it via (Restricted) Maximum Likelihood, and the optimal smoothing parameter $\hat{\lambda}$ results then from the mixed model's variance parameters (see Appendix I of this thesis, or alternatively Fahrmeier et al. (2004)).

2.1.2 Bivariate

The non-parametric modeling of a bivariate covariate in the spline framework, that corresponds to modeling the interaction of two univariate covariates, can be achieved analogously to the univariate case through the use of a so called *tensor product base*. This bivariate basis consists in fact just of all pairwise products of the two bases from the univariate smooths. The result is the explanation of the response variable y through a two-dimensional surface $f(z_1, z_2)$. Like above the univariate bases of z_1 and z_2 would consist of the basis functions $B_j^{(1)}(z_1)$, $j = 1, ..., d_1$ and $B_r^{(2)}(z_2)$, $r = 1, ..., d_2$ and the bivariate TP-basis is then constructed from the univariate basis functions as $B_{jr}(z_1, z_2) = B_j^{(1)}(z_1) \cdot B_r^{(2)}(z_2)$. Again the smoothing spline surface is then modeled by

$$f(z_1, z_2) = \sum_{j=1}^{d_1} \sum_{r=1}^{d_2} \gamma_{jr} B_{jr}(z_1, z_2). \tag{2.6}$$

Note that the number of basis spline functions and therefore to be estimated parameters increased from in the univariate case d_1 respectively d_2 to $d_1 d_2$, hence a big amount of data spread over the whole monitored area is needed to get a meaningful estimate. Consequently it is practically not possible to model multivariate covariates with dimensions bigger than 2 through a single spline function, the *curse of dimensionality* takes effect. Further, even in the bivariate case, it is necessary to confine the analysis on the subarea of $[min(z_1), ..., max(z_1)] \times [min(z_2), ..., max(z_2)]$ where realizations in data in fact occur.

Again the model can be seen as a conventional linear model with the rows of the design matrix Z being:

$$z_i = (B_{11}(z_{i1}, z_{i2}), ..., B_{d_11}(z_{i1}, z_{i2}), ..., B_{1d_2}(z_{i1}, z_{i,2}), ..., B_{d_1d_2}(z_{i1}, z_{i2}));\ i = 1, ..., n$$

and the corresponding vector of regression coefficients

$$\gamma = (\gamma_{11}, ..., \gamma_{d_11}, ..., \gamma_{1d_2}, ..., \gamma_{d_1d_2})'.$$

The penalty required for choosing aan appropriate number of knots while preventing overfitting can again be constructed by means of (spatial) adjacent regression coefficients. It is for example possible to use the squared differences between γ_{jk} and the regression coefficients of the four nearest (in direction of the coordinate axis) neighbors through

$$\gamma' K \gamma = \gamma' \left[I_{d_2} \otimes K_1^{z_1} + K_1^{z_2} \otimes I_{d_1}\right] \gamma,$$

with $\otimes$ denoting the Kronecker product of two matrices, I_d being the d-dimensional identity matrix and $K_1^{z_1}$, $K_1^{z_2}$ being univariate penalty matrices consisting of the squared first-order difference matrices: $K_1^{z_1} = D_1^{z_1\prime} D_1^{z_1}$, respectively $K_1^{z_2} = D_1^{z_2\prime} D_1^{z_2}$ (see Fahrmeier et al. (2013), p. 508).
Again it is possible to use higher order differences as penalty, $K_{k_1}^{z_1}$ and $K_{k_2}^{z_2}$ include then the squared k_1-/k_2-th order difference matrices $D_{k_1}^{z_1}/D_{k_2}^{z_2}$ and $(k_1 + k_2) \cdot 2$ adjacent regression coefficients are considered for smoothing.
A two-dimensional penalized regression spline surface can therefore be expressed in a similar form than a one-dimensional and hence be estimated in equal ways.

2.2 Mixed Model Regression

A linear mixed model (LMM) (see e.g. Fahrmeier et al. (2013), Chapter 7) is the extension of the classical linear model (where the effect parameters β are assumed to be unknown but be fixed)

$$y = X\beta + \epsilon, \quad \epsilon \sim N(0, I\sigma^2), \tag{2.7}$$

with (group or individual specific) random effects (whose parameters b are assumed to be realizations from a probability distribution) to

$$y = X\beta + Ub + \epsilon, \tag{2.8}$$

with

$$\begin{pmatrix} b \\ \epsilon \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} G & 0 \\ 0 & R \end{pmatrix}\right). \tag{2.9}$$

X and U are the design matrices of the fixed and the random effects, the (unknown) covariance matrices G and R for the fixed/random effect vectors are positive definite and β and b are independent.
The distribution 2.9 is the conventional assumption of Gaussian distributed random effects and error terms which is not necessarily needed, however it enables likelihood estimation of the unknown parameters in G and R. In a simple case it is possible to assume the error terms ϵ being i.i.d. $N(0, \sigma^2)$ distributed, what would lead to the covariance matrix $R = \sigma^2 I$. Alternatively it is possible to specify further correlation (e.g. the assumption of autoregressive errors) through a corresponding covariance matrix for the errors.
Such mixed models are useful for modeling of grouped data arising e.g. from longitudinal, repeated measures, or clustered data because they enable to take the correlation structure of the response variable resulting from those group structures into account. Applying models without random effects on such datasets leads to wrong standard errors and therefore also wrong confidence intervals and tests: Through combining the residual vector and the random effects into a single non-independent variable-variance residual vector $e = Ub + \epsilon$ (e.g. Wood (2006a), p. 287) it is possible to rewrite the mixed model Equation 2.8 into

$$y = X\beta + e \tag{2.10}$$

with

$$e \sim N(0, V); \quad \text{with } V = UGU' + R. \tag{2.11}$$

Fitting a classical linear model (Equation 2.7) to grouped data makes therefore a wrong assumption about the covariance of the error terms, however the resulting fixed effects parameter estimates are unbiased (because the expected values of e and ϵ are 0, respectively).
Estimation of mixed models can be achieved via (restricted) maximum likelihood estimation (Fahrmeier et al. (2013), p. 371-374): For known covariance matrices R, G and $V = UGU' + R$, β can be estimated by generalized least squares as

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y$$

and under normality assumption the conditional mean of b given the data y is the estimator of b:

$$\hat{b} = GU'V^{-1}(y - X\hat{\beta}).$$

In general the parameters υ in R, G, and V are unknown and need to be estimated before the matrices can plugged into the estimators of β and b. This can be done either by Maximum-Likelihood or restricted Maximum-Likelihood (REML) estimation. In the first case the maximization of the *profile*-log-likelihood $l_P(\upsilon)$

$$l_P(\upsilon) = -\frac{1}{2}\{log|V(\upsilon)| + (y - X\hat{\beta}(\upsilon))'V(\upsilon)^{-1}(y - X\hat{\beta}(\upsilon))\},$$

$$\hat{\beta} = (X'V(\upsilon)^{-1}X)^{-1}X'V(\upsilon)^{-1}y,$$

with respect to υ leads to the estimation of $\hat{\upsilon}_{ML}$.
For REML estimation the *marginal*-log-likelihood $l_R(\upsilon) = log(\int L(\beta, \upsilon)d\beta)$

$$l_R(\upsilon) = l_P(\upsilon) - \frac{1}{2}log|X'V(\upsilon)^{-1}X|$$

has to be maximized and leads to $\hat{\upsilon}_{REML}$.
The mixed model approach can be generalized for non-normal regression settings through connecting the conditional mean of y with an appropriate response function $h(\cdot)$ to the linear predictor:

$$E\,(y\,|\,b) = h(X\beta + Ub).$$

As already stated in Section 2.1 p-splines and mixed models are closely related. It is possible reformulate p-splines as a mixed model by dividing them into a fixed and a random effects part (see Appendix I). With this in mind it is possible to construct the *semiparametric* or *additive mixed model* (see e.g. Wood (2006b) or Ruppert et al. (2003)) of the form:

$$y = X\beta + f_z(x_z) + ... + Ub + \epsilon,$$

with β and b being coefficients of parametric fixed and random effects, X and U their design matrices and the $f_z(x_z)$ being non-parametric (uni- or multivariate) smooth functions of covariates. Reformulating those splines as mixed models and adding their fixed and random effects to the matrices X and U, respectively, one obtains a large mixed model whose (commonly REML) estimation also provides the smoothing parameters λ in case of modeling the smooth functions $f_z(\cdot)$ by p-splines.

3 Analysis of the SOEP-Data

3.1 SOEP-Data and some Descriptive Statistics

The German Socio-Economic Panel Study (SOEP) is a longitudinal study that started in 1984 and is surveying more than 20000 adult (age: 16+) people out of about 11000 households annually. As long as they stay in the panel (participation is voluntarly), the surveyed people are every year the same. Besides socio-economic data like income, labor participation, family status and members of the household the study rises also data on health and subjective well-being/life satisfaction. A detailed description of the SOEP can be found in Wagner et al. (2007).

For the analysis of the data, some editing is necessary: the first two years of every interviewed person are excluded because of panel-/learning effects, also the survey years 1990 and 1993 have to be excluded because of some missing health indicators. Because of too few observations all person-years with an age > 90 are not considered for the analysis. The whole dataset consists then out of 252406 observed person-years resulting out of the monitoring of altogether 33251 persons in the period between 1986 and 2007. Observed cohorts vary between 1900 and 1987, and the ages of the interviewed persons are between 18 and 90.

On average there is data out of 7.59 interviews per person available, with median being 6.

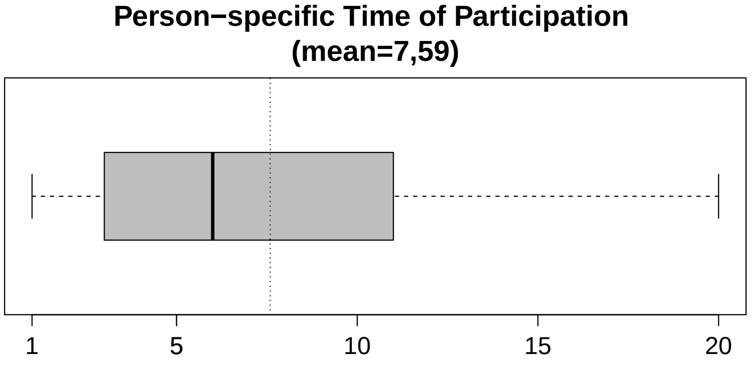


Figure 3.1: Boxplot of the Person-specific Time of Participation in SOEP-data

The subjective well-being is measured through the question "How satisfied are you with your life, all things considered?", with possible answers on a eleven point scale ranging from zero (completely unsatisfied) to ten (completely satisfied). The mean of this well-being score is 6.9, median 7 and modus 8:

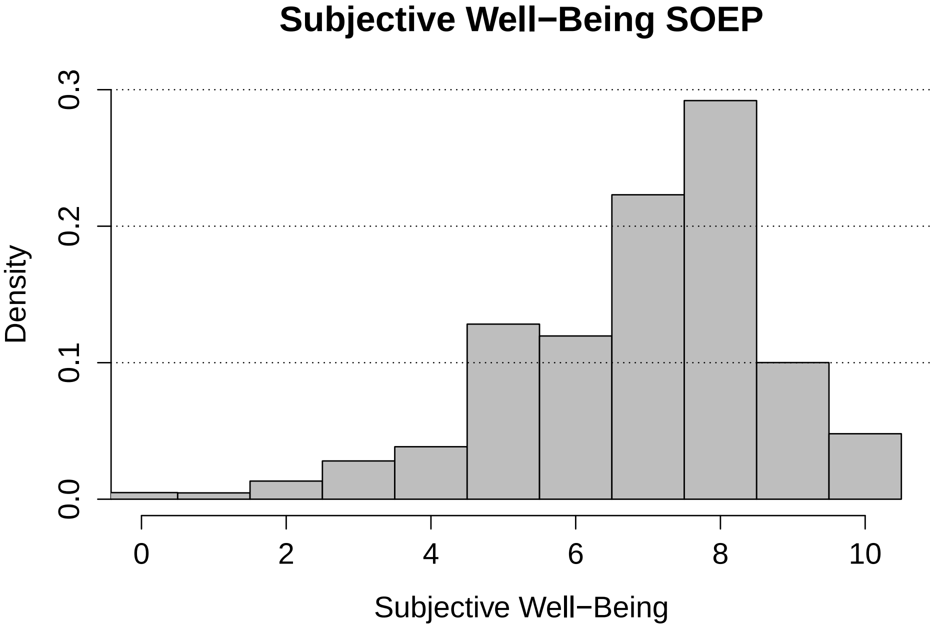


Figure 3.2: Histogram of the Subjective Well-Being in SOEP

Addressing the age and cohort structure within the 22 years' observation period it is obvious that it is only possible to look at specific age segments for each cohort that are determined by:

$$\text{Age} \mid \text{Cohort} = [\max\{18, (1986 - \text{Cohort})\}; \min\{90, (2007 - \text{Cohort})\}], \quad (3.1)$$

with Cohort $\in [1900, 1987]$.

3.2 Modeling Approach of this Thesis

Like already stated in Section 1.2 the approach of this thesis is to model the relation of age and subjective well-being in the SOEP-data in a non-parametric way, regarding the special APC-structure by estimating a spline (interaction) surface of the age and the cohort or alternatively the period effect. The use of age and cohort leads to a surface which is more intuitive to interpret since the sum of age and cohort equals the period and therefore possible effect structures that are orthogonal to the main diagonal (with age and cohort being analogously scaled) can be seen as period effects (all data points on such a line are obtained in the same year). However the results of an age-cohort and age-period interaction model should be the same (except for negligible differences due to the numerical calculation/estimation of the effects), since the interaction space includes in both times the third variable completely.

Instead of just estimating the interaction surface of the two covariates it is also possible to look at their marginal effects and the interaction surface, since the marginal effects are nested within the interaction (Wood (2006b), or Wood (2006a), p. 202-204). Therefore this model explains the same amount of variance in data and provides the same well-being predictions for all combinations of the covariates but offers additionally the marginal effect splines, which can then be compared with the results of other studies just estimating the marginal effect of age. The interaction surface can be interpreted as the deviation from the marginal effects for given values of the two covariates.

Such a model of the form

$$y_i = f_x(x_i) + f_z(z_i) + f_{zx}(x_i, z_i) + \epsilon_i$$

is referred to as *ANOVA-decomposition* model and for its estimation some identifiability conditions have to be imposed. They have to exclude the linear dependence of the basis/coefficients of the interaction surface ($f_{zx}(x, z)$)from the bases of the single terms ($f_x(x)$ and $f_z(z)$) (see Wood (2006a), p. 202-204).

Besides estimating the APC-effects it is of course necessary to control for several (socio-economic) covariates that are supposed to have an effect on the subjective well-being of individuals, whereby this work follows the previous of Wunder et al. (2013) and therefore includes linear effects for the time of education, family status, two indicators of the current health situation, weighted netto income, nationality, living in eastern or western Germany, gender, and a possible drop out of the panel study in the next or in two respectively three years.

The model is then a generalized additive mixed model or a semi-parametric model (since it includes parametric and non parametric predictor variables) and the model equation can be written as

$$y_{it} = x_{it}'\beta + f(a_{it}) + f(c_i) + f(a_{it}, c_i) + \epsilon_{it}; \quad i = 1, ..., n; \ t \in T_i \subset \{1, ..., T\}, \quad (3.2)$$

with y_{it} being the well-being of individual i at time t, x_{it} its socio-economic covariates and the $f(\cdot)$, the smooths of the age (of individual i at time t), the cohort (of individual i) and the interaction surface of both. Additionally there is an independent random error ϵ_{it} included. Non-parametric modeling of the age, respectively cohort effect is done with p-splines on a cubic B-spline base with 15 equidistant knots each, the interaction on a 15 × 15 tensor product base. The parameter estimates are achieved by restricted Maximum-Likelihood (REML) estimation.

It would also be possible to include further random effects into the mixed model, e.g. a person specific random intercept to adjust the estimates of the variances for the repeated measurements on the n individuals. Yet, it requires a lot of computational power to estimate those random effect (with big n). As described in Section 2.2 ignoring such random effects leads to unbiased estimates of the regression coefficients $\hat{\beta}$ but a biased estimation of the variance structure. When being mainly interested in the conditional means of the response variable and not in confidence intervals or testing, it is therefore possible to ignore the random effects (the authors' analysis on a subsample of the SOEP-data consisting of 5000 people provides evidence for that, since the resulting estimated coefficients and smooth functions of the models with and without a random intercept are quite similar).

After estimating this interaction model it is - besides looking at the interaction surface and the marginal effects - possible to gain (visual) insights into the effects of age, period and cohort through plotting predictions of combinations of covariates. When e.g. being interested in potential period effects it is possible to choose different fixed birth cohorts and plot their specific well-being trend over the panel's observation period, through predicting the well-being score for fixed cohorts and the specific ages of the cohort during the SOEP's period of record.

Computation in this thesis is done with R, version 3.0.1 (R Core Team (2013)), for estimating the model the packages *mgcv* (Wood (2014)) and for the overall heatplots containing the marginal age and cohort effects and their interaction *ggplot2* (Wickham (2009)) are used.

3.3 Model Evaluation

The estimation of the model described in Equation 3.2 provides directly four outcomes for evaluation: the marginal splines for age and cohort, their interaction surface and a table of the fixed effects for the socio-economic covariates. The latter can be found in Appendix II and shows in summary effects one would expect: A higher income, a longer education, employment and living in western Germany are generally affecting the subjective well-being in a positive way, while being of bad health, unmarried or unemployed results in a decrease of life satisfaction.

However the main focus of this study lies on meaningful evidence regarding APC-effects on well-being. Therefore the estimates of the smooth age and cohort (interaction) effects are of main interest. The resulting marginal age- and cohort splines can be found in Figure 3.3, respectively 3.4 and the interaction surface in Figure 3.5. In the one dimensional plots the continuous lines correspond to the estimated effects, while the dotted lines represent bounds of two (most likely biased) standard errors each, above and beyond the estimated smooth. The interaction effect is represented by a heat map including contours to specify the effect size. Looking at the second part of the table in Appendix II, it can be seen that the age spline is estimated with about 8 degrees of freedom (df) and the cohort spline with about three df. That shows that a penalization takes place during the estimation of the effects, since the upper limits of possible df are 14 for the age- (number of knots $k = 15 - 1$ df for the centering constraint) and 15 for the cohort spline. The interaction surface is estimated with about 50 df, with a possible maximum of $15 \cdot 15 - 1 = 224$. This is exactly the aim of penalized spline regression, instead of just interpolating the data through a function with a predefined amount of degrees of freedom (and consequently a somewhat predefined form of the polynomial), a smooth function is estimated with respect to the structure of the observed data.

Describing the marginal age effect (Figure 3.3) specifically, the spline shows an analogue shape as in the analysis of Wunder et al. (2013). During the life-span there exist three different episodes of well-being development: the first one holds about 35 years and ranges from the age of 18 to 53 with a steady decline of altogether about 0.9 points on the life-satisfaction scale. Afterwards there is a approx. 12 years lasting increase of the average well-being score apparent (overall ~ 0.4 points). With an age of about sixty-five a decline in life-satisfaction starts again, which lasts until the end of peoples life.

Wunder et al. (2013) present some possible theoretical reasons for this development of life-satisfaction: the first episode is predominately affected by not entirely fulfilled aspirations and the impression of faster-passing time with advancing age. In the second there is an adaption of the life situation and risen a satisfaction with the financial situation, material needs and social contacts conceivable. Further the anticipation of retirement can be a reason for a tendentially increasing subjective well-being. Altogether those reasons seem to outweigh the general deterioration of living conditions theoretically taking place during the same time. The decline in life-satisfaction in the third period of life could be explained through a deterioration of health that is not fully captured by the rather fragmentary health covariates and several other processes and situations not controlled for in the regression model (e.g. losses in the social environment, etc.).

The marginal cohort spline (Figure 3.4) shows a general decline in the average subjective well-being across the cohorts until 1960 and afterwards a rather constant, maybe slight upward trend. Looking at the shape, not the exact width of the standard errors, they increase towards the the years 1900, respectively 1987.

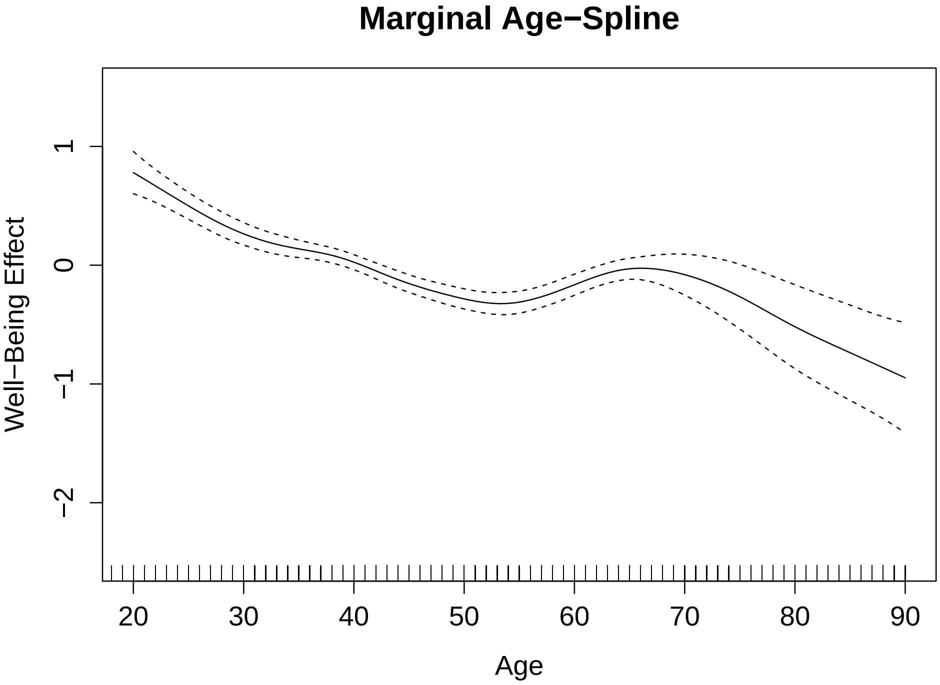


Figure 3.3: Marginal Age Spline

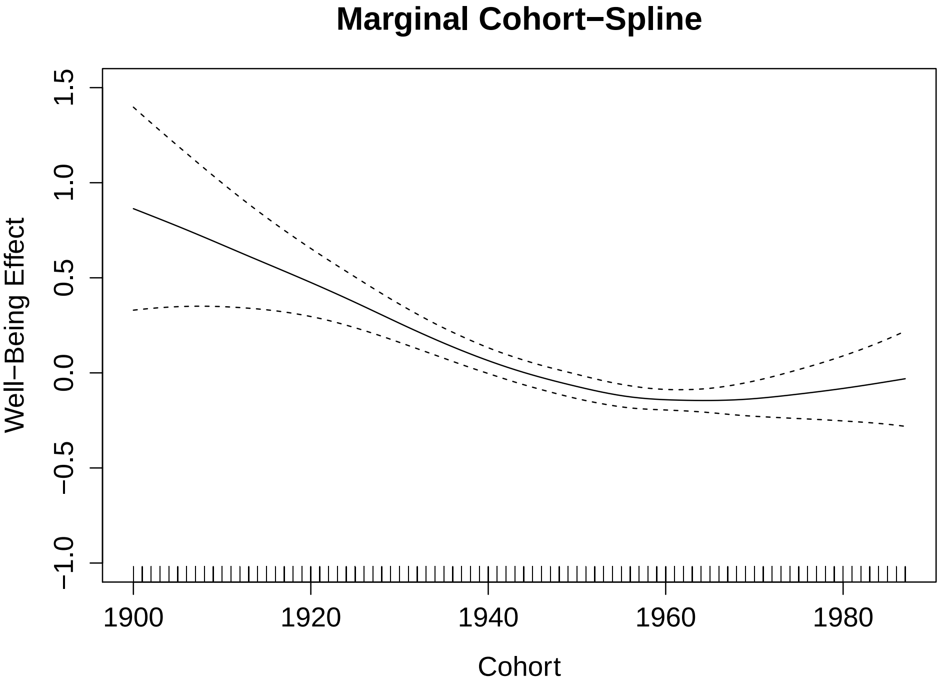


Figure 3.4: Marginal Cohort Spline

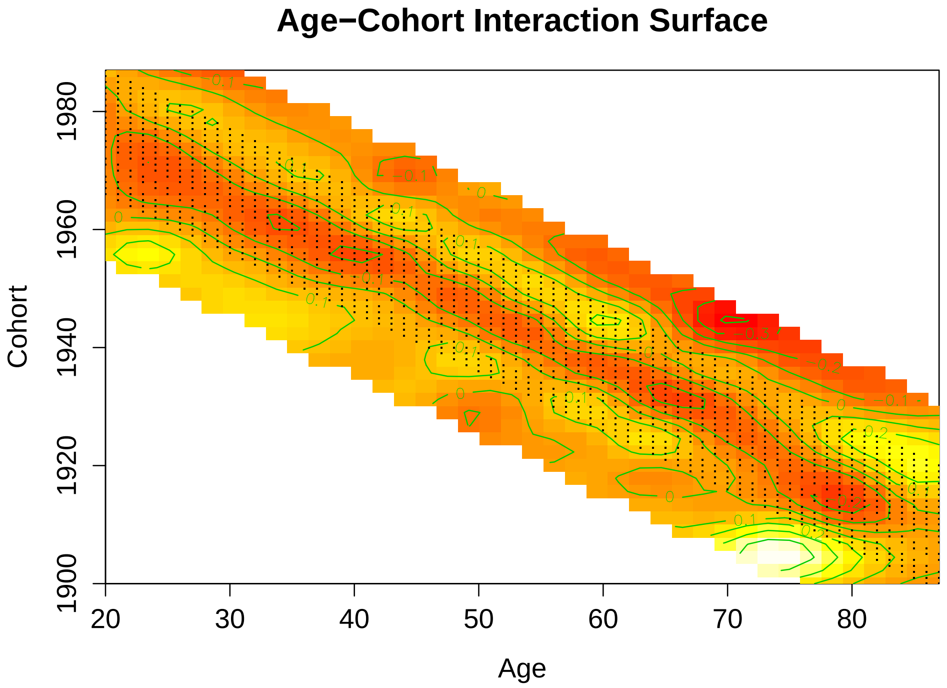


Figure 3.5: Age-Cohort Interaction Surface

This is a result from the decreasing sample size in those areas. Equation 3.1 shows that this SOEP dataset can e.g. for the cohort 1905 only contain data of people with an age ascending from 81 to 90 during the observation period between 1986 and 1995 $((period) = (age) + (cohort))$. That leads to less person-years being available for this cohort than for cohorts whose members could be observed during the whole period of the panel. This spline estimate and all further plots containing predictions/estimations for the outer cohort values should therefore be interpreted with care in the corresponding areas.

Figure 3.5 presents the estimated interaction surface of age and cohort. Note that the black dots mark the region where data occur and a reasonable interpretation of the findings should be confined to this area. With a range of $[-0.27, 0.29]$ the interaction effect is not that big. Its structure appears to be pretty constant for combinations of the age and cohort variables belonging to the same period, there are a just a few changes in the effect size for single or a few combinations of specific adjacent age and cohort values. Furthermore it seems like the marginal effects of age and cohort are properly expressed through the univariate splines, the interaction surface features neither persistent vertical nor horizontal forms. Therefore it is plausible to assume that the interaction surface mainly expresses period effects on well-being. The mean of profile curves computed over the whole area of the surface in direction of simultaneously increasing age and cohort values can cautiously be interpreted as a general period effect on well-being (Figure 3.6).

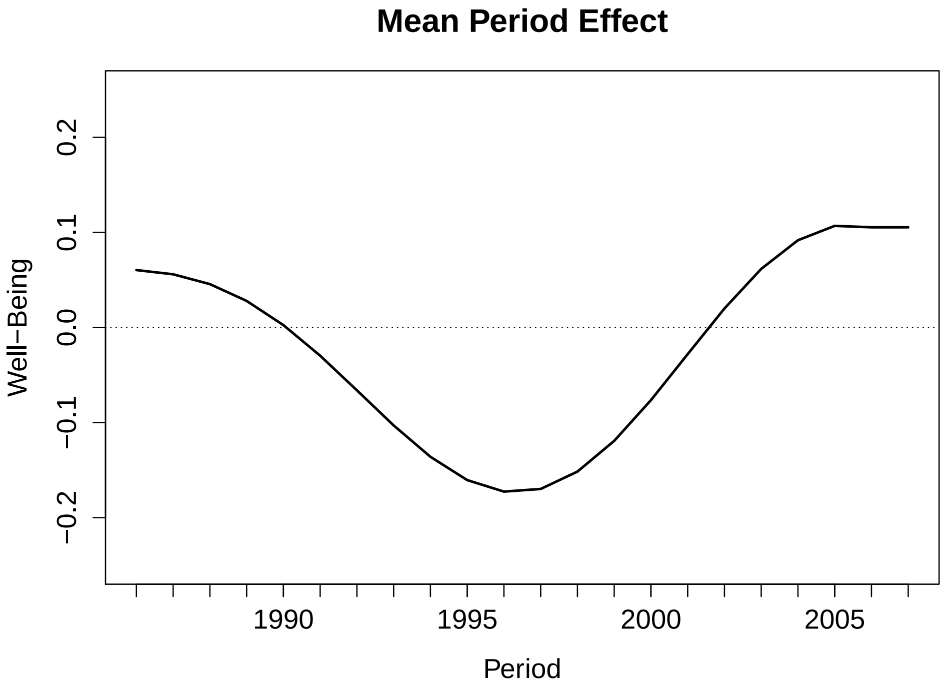


Figure 3.6: Period-effect on Well-Being

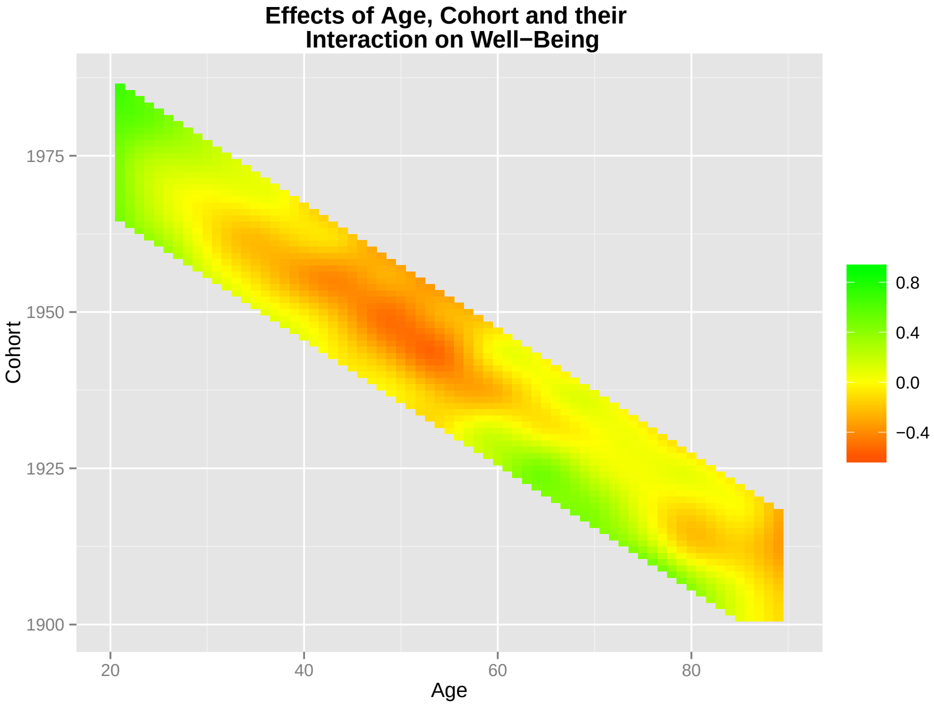


Figure 3.7: Effects of Age, Cohort and Interaction on Well-Being

With the estimated effects of age, cohort and their interaction it is possible to compute their combined expected effects for different combinations of values. Figure 3.7 shows that prediction for the whole observation area of the SOEP dataset. Depending on Age, Period and Cohort the lowest values (an effect of about -0.5) of life satisfaction were observed for people between about 45 and 55 of the approximate birth cohorts 1950 to 1940. In this area the first minimal turning point of the age spline accumulates with the relatively low well-being scores of the later cohorts and the generally low satisfaction in the observation periods around 1995. In the bottom right area of the plot the relatively high satisfaction levels of early cohorts compensate the negative connection between live satisfaction and old age to some extent. The shape of the interaction surface can be spotted over the whole observation area, with generally higher well-being values in the early and late observation periods and in the upper left area of the plot the high satisfaction values in early stages of life exceed the relatively low satisfaction levels of the late cohorts.

Another way of analyzing the APC effect structure is to predict the effects over the observation time for fixed cohorts. Thereby it is possible to plot the well-being either against period (Figure 3.8) or age (Figure 3.9). Note that the resulting curves are in fact the same, only their location on the x-axis is changing and they are in some way compressed/stretched by the scale of the axis.

The cohort specific effect is part of those curves as an additive linear factor, while the appendant age and age-cohort interaction effects determine the form of the curve. They can be seen as cross sections of the surface shown in Figure 3.7 parallel to the x-axis. Therefore such curves show the estimates of the well-being effects in the panels' observation time for chosen combinations of age and cohort values.

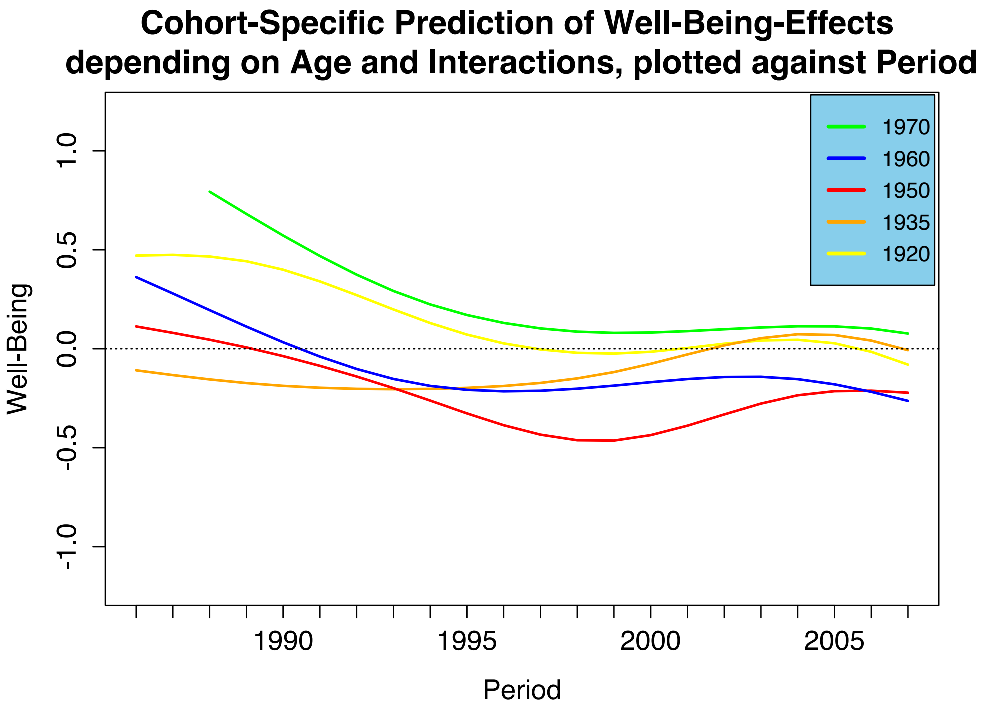


Figure 3.8: Cohort-specific Pred. of Age and Cohort-Interaction Effects I

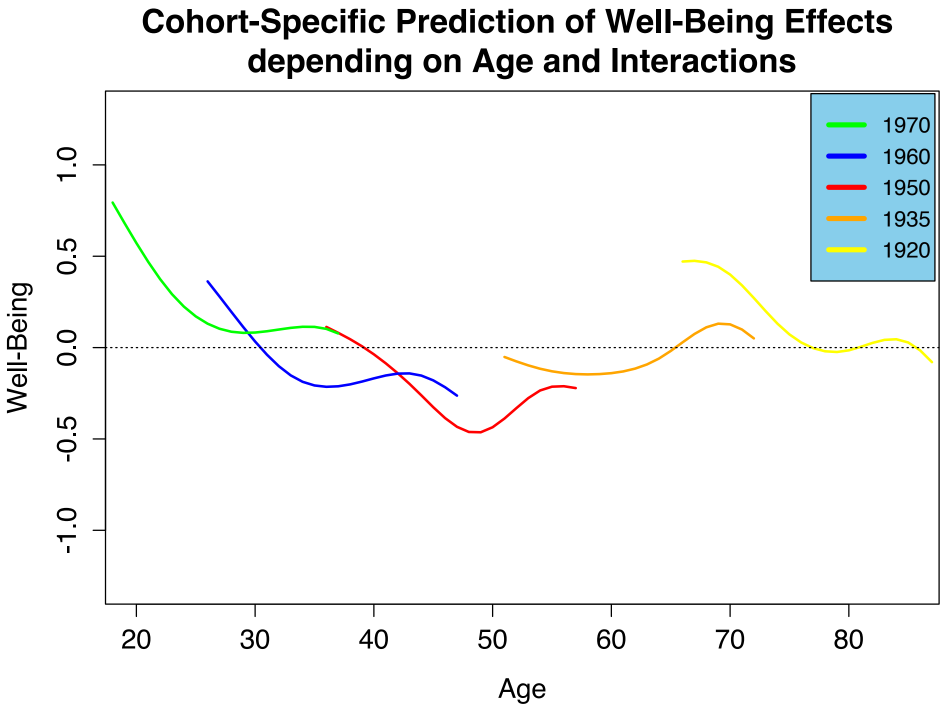


Figure 3.9: Cohort-specific Pred. of Age and Cohort-Interaction Effects II

4 Conclusion

This thesis introduced an additive mixed model approach for the analysis of panel data on individuals, regarding the estimation of effects for the time-related changes in age, period and cohort. It was applied on data of the German socio-economic panel with focus on the development of subjective well-being over the lifespan. Additionally to fixed effects for several socio-economic covariates, non-parametric effects of age and cohort are modeled to obtain estimates for their relations to well-being. Furthermore an interaction surface for age and cohort is included into the model, whereby in particular period effects are captured.

It can be assumed that the obtained age and cohort curves are appropriately describing their relations to life-satisfaction, while the interaction surface can be inspected for period related changes. Additionally to the obtained marginal splines and the interaction surface it is possible to create - based on the estimates - somehow descriptive plots for combinations of different covariate values, that allow a further explorative analysis of the effects.

To sum up, such a model features two main advantages over other approaches of analyzing Age-Period-Cohort data: Firstly, its flexible estimation via penalized splines guarantees a data-based investigation of the effects, instead of defining their form a priori. And secondly, the incorporation of the interaction surface ensures the capture of all three time-related changes in data. Of course there results no distinct estimate of the period effect (its only possible to interpret the interaction surface in some way as a period effect), but this can also be seen as an advantage: The model has the flexibility to not just estimate a general period effect for all age-groups, but also considers changes in the dependent variable for specific combinations of age and cohort values during the observation time.

Identification of the marginal and interaction effects is possible, however it does also require to include some technical constraints. The exact (non-technical) assumptions accompanying them could be a matter of further research.

Computation can be easily achieved with the *mgcv*-package in R but also done with any other software for mixed-model estimation.

Altogether the presented approach offers a reasonable and practicable way to look at the impact of age, period and cohort changes in panel data, what makes many other applications in empirical research possible.

Appendix I: Penalized Splines in Mixed Model Formulation

Formulating p-splines as a mixed models enables firstly to determine the smoothing parameter λ by estimation of the variance parameters of the mixed model, and secondly embedding them into the additive mixed model framework. The following procedure for this reformulation is also described by e.g. Wood (2004), Ruppert et al. (2003), or Fahrmeier et al. (2013):
A smooth term with parameter vector β, associated model matrix X and penalty matrix S is estimated by minimization of the penalized sum of squares

$$PLS = \parallel y - X\beta \parallel^2 + \lambda\beta' S\beta$$

with respect to β.
To ensure identifiability in models with more than one smooth covariate it is necessary to impose a constraint (e.g. the smooth sums over the covariates values up to zero). Such a constraint can be expressed through some constraint matrix C with $C\beta = 0$.
This constraint can be included into the PLS criterion by forming the QR decomposition $QR = C^T$, defining Z to be Q without its first n_c columns, with n_c being the number of rows of C and writing $\beta = Z\beta_z$:

$$PLS = \parallel y - XZ\beta_z \parallel^2 + \lambda\beta_z^T Z^T SZ\beta_z.$$

With the spectral decomposition of $Z^T SZ = UDU^T$, with D a diagonal matrix containing the eigenvalues in a decreasing order the PLS can be rewritten as:

$$PLS = \parallel y - XZU\beta_u \parallel^2 + \lambda\beta_u^T D\beta_u,$$

with $\beta_u = U^T\beta_z$.
As a result of its construction, the penalty matrix S is generally rank deficient and therefore the last few elements on the leading diagonal of D are zero. Let D^+ being the smallest possible sub matrix of D containing all positive eigenvalues. Partition β_u into $\beta_u^T = [b_u^T, b_F^T]$ so that $\beta_u^T D\beta_u = b_u^T D^+ b_u$ and XZU into $[X_u, X_F]$

in a similar manner. With $b = \sqrt{D^{+}} b_u$ and $X_R = X_u(\sqrt{D^{+}})^{-1}$ the PLS get:

$$PLS = \parallel y - X_F\beta_F - X_R b \parallel^2 + \lambda b^T b$$

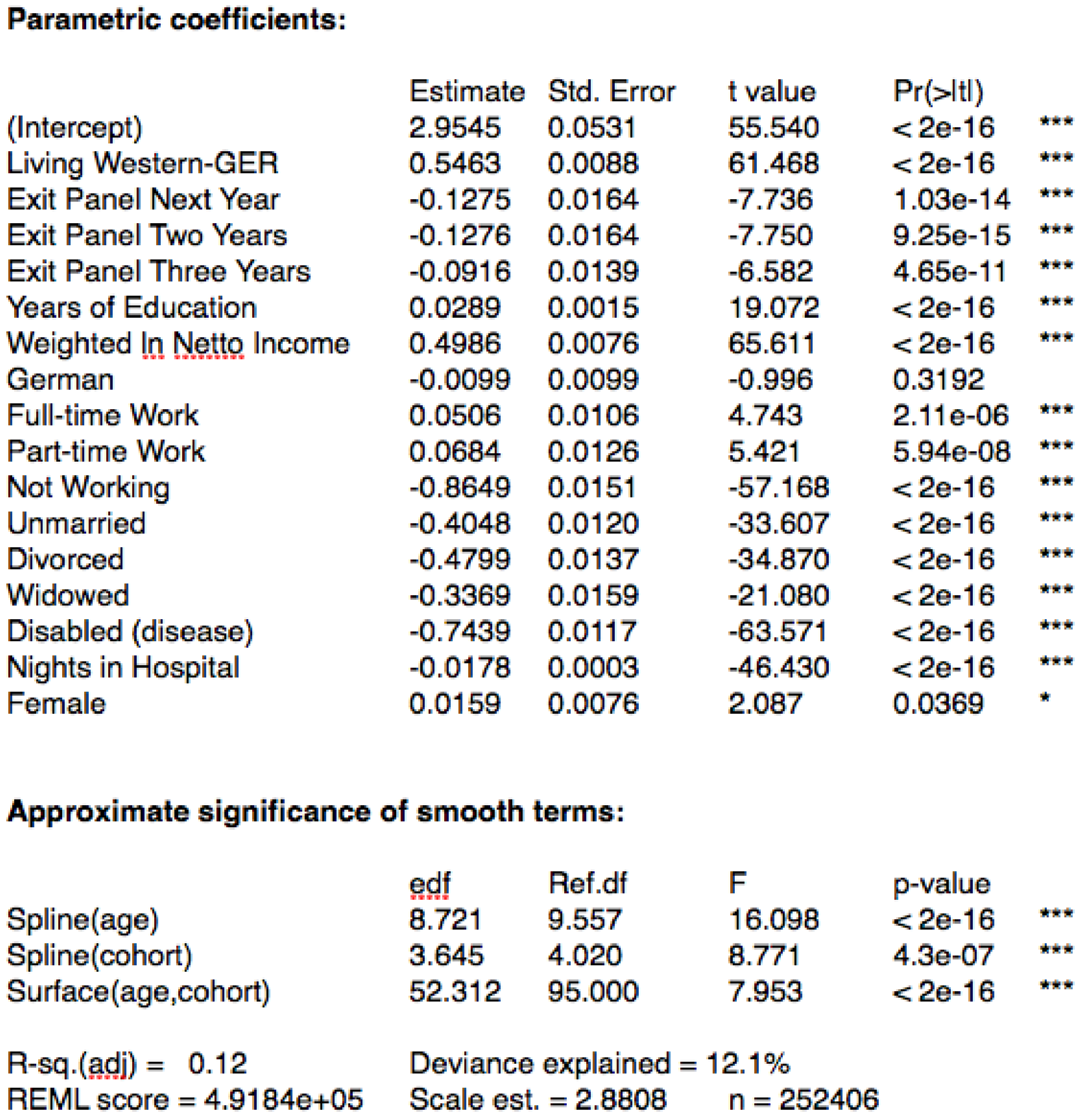
Now minimizing the PLS equals estimating a mixed model of the form:

$$y = X_F\beta_F + X_R b + \epsilon, \quad \epsilon \sim N(0, I\sigma^2), b \sim N(0, I\tau^2),$$

with $\lambda = \sigma^2/\tau^2$ being connected to the variance parameters of the random effects and the error terms. It is then straightforward to get $\hat{\lambda} = \hat{\sigma}^2/\hat{\tau}^2$ from REML estimation and therefore the reformulation of the penalized splines enables a simultaneous estimation of the regression coefficients and the smoothness parameter.

Appendix II: Coefficients

Resulting coefficients of the Age-Cohort Interaction Model (Equation 3.2). Estimated in R with the *bam*-function of the *mgcv* package:



Parametric coefficients:

	Estimate	Std. Error	t value	Pr(>ltl)	
(Intercept)	2.9545	0.0531	55.540	< 2e-16	***
Living Western-GER	0.5463	0.0088	61.468	< 2e-16	***
Exit Panel Next Year	-0.1275	0.0164	-7.736	1.03e-14	***
Exit Panel Two Years	-0.1276	0.0164	-7.750	9.25e-15	***
Exit Panel Three Years	-0.0916	0.0139	-6.582	4.65e-11	***
Years of Education	0.0289	0.0015	19.072	< 2e-16	***
Weighted ln Netto Income	0.4986	0.0076	65.611	< 2e-16	***
German	-0.0099	0.0099	-0.996	0.3192	
Full-time Work	0.0506	0.0106	4.743	2.11e-06	***
Part-time Work	0.0684	0.0126	5.421	5.94e-08	***
Not Working	-0.8649	0.0151	-57.168	< 2e-16	***
Unmarried	-0.4048	0.0120	-33.607	< 2e-16	***
Divorced	-0.4799	0.0137	-34.870	< 2e-16	***
Widowed	-0.3369	0.0159	-21.080	< 2e-16	***
Disabled (disease)	-0.7439	0.0117	-63.571	< 2e-16	***
Nights in Hospital	-0.0178	0.0003	-46.430	< 2e-16	***
Female	0.0159	0.0076	2.087	0.0369	*

Approximate significance of smooth terms:

	edf	Ref.df	F	p-value	
Spline(age)	8.721	9.557	16.098	< 2e-16	***
Spline(cohort)	3.645	4.020	8.771	4.3e-07	***
Surface(age,cohort)	52.312	95.000	7.953	< 2e-16	***

R-sq.(adj) = 0.12 Deviance explained = 12.1%
REML score = 4.9184e+05 Scale est. = 2.8808 n = 252406

Bibliography

Clayton, D. and E. Schifflers (1987). Models for temporal variation in cancer rates ii: Age-period-cohort models. *Statistics in Medicine 6*, 469–481.

Diener, E., S. Oishi, and R. E. Lucas (2009). Subjective well-being: The science of happiness and life satisfaction. *The Oxford Handbook of Positive Psychology 2*, 63–73.

Eilers, P. and B. Marx (1996). Flexible smoothing with b-splines and penalties. *Statistical Science 11*(2), 89–121.

Fahrmeier, L., T. Kneib, and S. Lang (2004). Penalized structured additive regression for space-time data: a bayesian perspective. *Statistica Sinica 14*, 731–761.

Fahrmeier, L., T. Kneib, S. Lang, and B. Marx (2013). *Regression - Models, Methods, Applications*. Springer.

Heuer, C. (1997). Modeling of time trends and interactions in vital rates using restricted regression splines. *Biometrics 53*(1), 161–177.

Holford, T. R. (1983). The estimation of age, period and cohort effects for vital rates. *Biometrics 39*(2), 311–324.

Jiang, B. and K. C. Carriere (2014). Age-period-cohort models using smoothing splines: A generalized additive model approach. *Statistics in Medicine 33*(4), 595–606.

Kupper, L. L., J. M. Janis, A. Karmous, and B. G. Greenberg (1985). Statistical age-period-cohort analysis: A review and critique. *Journal of Chronical Diseases 38*(10), 811–830.

Mason, K. O., W. M. Mason, H. H. Winsborough, and W. Kenneth (1973). Some methodological issues in cohort analysis of archival data. *American Sociological Review 38*(242-258).

R Core Team (2013). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Rodgers, W. L. (1982). Estimable functions of age, period and cohort effects. *American Sociological Review 47*(6).

Ruppert, D., M. P. Wand, and R. Carrol (2003). *Semiparametric Regression.* Cambridge.

Wagner, G. G., J. R. Frick, and J. Schupp (2007). The german socio-economic panel study (soep) – scope, evolution and enhancements. *Schmollers Jahrbuch 1*, 139–169.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis.* Springer New York.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association 99*, 673–686.

Wood, S. N. (2006a). *Generalized Additive Models: An introduction with R.* Chapman Hall.

Wood, S. N. (2006b). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics 62*, 1025–1036.

Wood, S. N. (2014). *mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation.*

Wunder, C., A. Wiencierz, J. Schwarze, and H. Küchenhoff (2013). Well-being over the life span: Semiparametric evidence from british and german longitudinal data. *The Review of Economics and Statistics 95*(1), 154–167.

Affidavit

I, Felix Günther, hereby declare that I wrote this bachelor-thesis on my own and without the use of any other than the cited sources and tools. All explanations that I copied directly or in their sense are marked as such.

City, Date *Felix Günther*