Flexible Modeling with Generalized Additive Models and Generalized Linear Mixed Models: Comprehensive Simulation and Case Studies.

Daniel Hercz, Department of

Epidemiology and Biostatistics

McGill University, Montreal

August 2012

Thesis submitted to McGill University in

partial fulfillment of the requirements of the degree of

M.Sc Biostatistics

Table of Contents

Table of Figures

PREAMBLE

Thesis Format

I have opted to write this thesis in the manuscript-based format as permitted by McGill University regulations. The following are the titles of the two manuscripts that are included in the thesis:

1. Flexible modeling using Generalized Additive Models and Linear Mixed Models

2. Modeling nonlinear trends in ICU patients with Sepsis: A comparison of GAMs and GLMMs

These manuscripts were combined and presented at 32nd Annual Conference of the International Society for Clinical Biostatistics. The first has been submitted to *Statistics in Medicine.* The manuscripts are connected in the logical progression as will be explained in the introduction and linking statements and summarized in the general conclusions. In a manuscript-based thesis some amount of redundancy is inevitable, particularly in the introductions of the individual manuscripts. To mitigate this all references were combined into a single section after the general conclusions. Acknowledgements are also in only one section near the beginning.

Contributions and Co-authors

This thesis represents the results of my own independent research. Both manuscripts have been co-authored by my thesis supervisor, Dr. Andrea Benedetti who contributed to the design, execution, analysis, and presentation of the results presented. Dr Jean Bourbeau and Dr Sandra Dial contributed the data and empirical motivation for the first and second manuscript respectively.

Acknowledgements

I would like to extend a special thanks to my advisor, Andrea. Her great patience with me as I completed this thesis has allowed me to continue on with other aspects of my career.

Declaration

Herewith I affirm that I have written this thesis on my own. I did not enlist unlawful assistance of someone else. Cited sources of literature are marked and listed at the end of this thesis. The work was not submitted previously in same or similar form to another examination committee and was not yet published.
Sydney, August 2012

Abstract

This thesis compares GAMs and GLMMs in the context of modeling nonlinear curves. The study contains a comprehensive simulation and a few real life data analyses. The simulation uses thousands of generated datasets to compare and contrast the two models' (and linear models as a benchmark) fit, extent of nonlinearity, and shape of the resulting curve. The data analyses extend the results of the simulation to GLMM/GAM curves of lung function with measures of smoking as the independent variable. An additional and larger real life data analysis with dichotomous outcomes rounds out the study and allow for more representative results

Introduction

Linear modeling is arguably the most prolific form of inferential analysis in modern statistics. However, these models carry obvious limitations in that they (in their most basic form) can only describe linear or parametric relationships between variables. In being constrained to this simplified functional format, many statistical tests have reduced power and often fail to identify trends in datasets.

This thesis seeks to compare two specific methods for modeling nonlinear associations. The Generalized Additive Model (GAM) and the Generalized Linear Mixed Model (GLMM) are two commonly used approaches to fit nonlinear curves. They are far from the only methods available (others are discussed in the literature review), but they are rigorous, effective and implemented in current statistical software such as R and SAS. They are also important in that they do not presume the form of the nonlinearity, which plagues many other techniques. However, it is unknown if applying both techniques to the same dataset would result in different curves and inferences. If so, deciding on which model is best suited for one's analysis is an important *a priori* consideration.

The first section of this thesis is a statistical simulation with datasets that vary in sample size, error variance, and degree of nonlinearity. Comparisons between GAMs and GLMMs examined included the Kullback Liebler distance, several information criteria, and graphical representations on the fitted curves. The thesis is also instructional as to the curve fitting role of GLMMs and GAMs. While many researchers may be familiar with one technique (or both in a different capacity, such as longitudinal analysis), the other may be foreign to them. In addition to a systematic review this thesis may serve as an introduction to the implementation of both GAMs and GLMMs with respect to fitting nonlinear relationships. Finally each manuscript examines a real life dataset. The first examines the relationship between smoking duration and intensity and several spirometric indicators of lung function. The second manuscript seeks to models the associations between four different vital parameters and mortality in post-surgery ICU patients.

In the literature review, there is a brief overview of the basis, mechanics and known strengths and weaknesses of GAMs and GLMMs. This is followed by background on alternative methods to modeling nonlinear data. The Kullback Liebler distance and the three information criteria used are reviewed. Finally there is a more in depth discussion on the variables involve in the real life datasets and their current and historical analytic strategies.

Development of linear statistical modeling as a research tool

Generalized Additive Models (GAMs) and Generalized Linear Mixed Models (GLMMs) are variants of Generalized Linear Models (GLMs) and Additive Models (AMs) with some of the model assumptions relaxed [1, 2]. The simplest way to approach GLMs is to understand them by their namesake, a generalization of linear models. AMs are discussed in the following section. The standard notation from the linear model is shown below.

$$y = X\beta + \varepsilon$$

X is the data matrix, whose rows contain sequential data on each subject. y is the response vector. The ε are the error terms, a vector of zero-mean normally-distributed random variables. The β vector (or vector of covariates) is chosen in such a way to minimize the magnitude of the εs. Estimation for β is most commonly achieved through least squares minimization [3].

To maintain the interpretability of the model and facilitate locating the least squares estimators $\hat{\beta}, \hat{y}$, and $\widehat{\sigma^2}$ (final parameter being the estimator for the variance of the εs). several assumptions about the data must be made [3].

I. $y = X\beta$ A linear predictor is the quantity which incorporates the information about the independent variables into the model.
II. The errors have identical normal distributions: $\varepsilon \sim N(0, \sigma^2)$
III. Independence of these error terms
IV. The fitted errors must have a mean of zero $\bar{\varepsilon} = 0$
V. The model assumes no multicollinearity in the data matrix. Mathematically this is realized by $E(x_i\ x_j)$ = a positive semi-definite matrix.

The result of these assumptions is a hypothesis where data takes on the distribution N $(X\beta, \sigma^2)$. From this model a likelihood function can be created and inferences about the effect of certain variable on the likelihood function can be made through likelihood ratio tests. Solving a multilinear regression (or Gaussian GLM) is done through Ordinary Least Squares (OLS) [4].

OLS/Maximum Likelihood Results for the Covariate Vector, Fitted Values, and Model Variance

$$\hat{\beta} = (X^t X)^{-1} X^t y \qquad \hat{Y} = X(X^t X)^{-1} X^t y \qquad \hat{\sigma} = \frac{\varepsilon^t \varepsilon}{n}$$

GLM

GLMs were first introduced as method of statistical inference and modeling in 1972, and dropped the first assumption of linear modeling. This is possible through restriction the models' class of probability distributions to those with similar enough properties to the normal distribution to allow parameter estimation and statistical inferences. This is called the single parameter exponential family of probability distributions and includes the Normal, Binomial, Poisson, Gamma, Inverse, Geometric, and Negative Binomial distributions. This greatly increased the type of response data one can model[3].

One additional quality that GLMs gained is that of a link function. The link function provides the relationship between the linear predictor and the mean of the chosen distribution function. Such a relationship aligns the response domain with independent variables' domain and allows for an interpretable model and inferences[5].

Smoothing in linear models

Smoothing methodology offers a means by which non-linear relationships can be handled without the restrictions of parametric functional forms. It has become a widely used tool for data analysis and inference. Its integration into complex models and use in applications is also becoming more and more pervasive. Two models, amongst others, which implement model smoothing are GAMs and GLMMs. Also examined in the review are Bayesian methods for smoothing.

Additive Models

The additive model (AM) is a nonparametric modeling technique suggested by [6]. The AM uses a one dimensional smoother to build a restricted class of nonparametric regression models. The general form of AMs is:

$$\mathrm{Y} = \beta_0 + \mathrm{f}_1(\mathrm{x}_1) + \mathrm{f}_2(\mathrm{x}_2) + \cdots + \mathrm{f}_\mathrm{m}(\mathrm{x}_\mathrm{m})$$

Where $E[\varepsilon] = 0$, $Var(\varepsilon) = \sigma^2$ and $E[f_j(X_j)] = 0$ on mean-centered data. The functions $f_j(x_{ij})$ are unknown smooth functions whose shape is estimated directly from the data. Fitting the AM (i.e. the functions $f_j(x_{ij})$) can be done using the back fitting algorithm proposed by Hastie and Tibshirani [7].

GAMs

$$\mathrm{g}\big(\mathrm{E}(\mathrm{Y})\big) = \beta_0 + \mathrm{f}_1(\mathrm{x}_1) + \mathrm{f}_2(\mathrm{x}_2) + \cdots + \mathrm{f}_\mathrm{m}(\mathrm{x}_\mathrm{m})$$

GAMS are a combination of the GLMs and additive models. A link function relates the response data with the independent variables, which are adapted again to estimated functions. The functions $f_i(x_i)$ may be fit using parametric or non-parametric means, thus providing the potential for better fits to data than other methods. The method hence is very general – a typical GAM might use a scatterplot smoothing function such as a locally weighted mean for $f_1(x_1)$, and then use a factor model for $f_2(x_2)$. By allowing nonparametric fits, well designed GAMs allow good fits with relaxed assumptions on the actual shape of the relationship, though perhaps at the expense of interpretability of familiar results.

One way GAMs may also be computed is by penalized estimating equations. The key is to express each function as a linear combination of basis functions common to all fs.

$$f(x) = \sum_{k}^{K} \beta_k b_k(x)$$

Where β_k are the coefficients to be estimated, and $b_k(x)$ are the basis functions chosen for convenience. The penalization term (to promote smoothness) is chosen to be the integral of the second derivative of f.

$$\int f''(x)^2\, dx = \int \beta^t\, \mathrm{b''}^{\mathrm{t}}(\mathrm{x})\, \mathrm{b''}(\mathrm{x})\, \beta\, dx$$

$$= \beta^t \int \mathrm{b''}^{\mathrm{t}}(\mathrm{x})\, \mathrm{b''}(\mathrm{x})\, dx\, \beta = \beta^t S \beta$$

where $\mathrm{b''}(\mathrm{x}) = {b_1}''(x) + \cdots + {b_K}''(x)$ by linearity and $S = \int \mathrm{b''}^{\mathrm{t}}(\mathrm{x})\, \mathrm{b''}(\mathrm{x})\, dx$ is

the matrix of coffeciants. The penalized estimating equation is:
$\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{y_i} - \mathrm{X_i}\beta)^2 + \lambda(\mu^t \mathrm{S}\mu)$.

Most statistical packages use restricted maximum likelihood (REML) or generalized cross validation (GCV) to compute a λ value or allow the user to specify one. Computation for $\hat{\beta}$ and $\hat{y}$ directly follows.
Lambda controls the trade-off between smoothness and fit.

As with all flexible modeling methods, over-fitting can be a problem with GAMs and the fitted process is potentially very sensitive to the data. The smoothing parameter can be specified, and in most epidemiologic applications this number should be reasonably small, certainly well under the degrees of freedom offered by the data. Cross-validation (in addition to the smoothing parameter) can be used to detect and/or reduce over-fitting problems with GAMs (or other statistical methods). If GAMs improve the predictive ability substantially for the application in question, such as in the nonlinear case, then their use is warranted.

Generalized Linear Mixed Models (GLMMs)

Smoothing methods that use basis functions with penalization can be formulated as fits in a mixed model framework. One of the major benefits is that software for mixed model analysis can be used for smoothing. Mixed model representations used for smoothing also allow for easy combination of smoothing with other modeling tools such as random effects for longitudinal data [8].

In this context, the LMM is of the following form:

$$\mathrm{y_i} = \beta_0 + \beta_1 \mathrm{x_1} + \cdots + \beta_m \mathrm{x_m} + \sum_{\mathrm{k=1}}^{\mathrm{K}} \underbrace{\mu_k(\mathrm{x_i} - \kappa_k)_+}_{\text{piecewise effects}} + \sum_{k'=1}^{\mathrm{K}} \underbrace{\mu_{k'}(\mathrm{x_i} - \kappa_{k'})_+}_{\text{piecewise effects}} + \varepsilon_\mathrm{i}$$

where $\mu_k(\mathrm{x_i} - \kappa_k)_+ = \begin{cases} 0 & \mathrm{x} \le \kappa_k \\ \mathrm{x} - \kappa_k & \mathrm{x} > \kappa_k \end{cases}$; *and* κ_k are the “knots”. These are pre specified points along the independent data. Their selection can be data-driven, random, uniformly space, amongst others. [8]

Most terms here are analogous to their GLM counterparts. The piecewise effect terms, $\mu_k$are modelled as identically independent. Determination of the $\mu_k s$ using ordinary least squares (as if there were m+k βs) results in a fixed effects model. This produces a smooth (up to p, the degree on polynomial used for the mixed terms) and nonlinear interpolation of the knots.

Constraining the μ_k such that $\mu_k \sim N(0, \sigma_\mu^2)$ and independent of each other changes it to a mixed effect model. With a finite variance providing additional penalty to the size of the μ_ks, slope changes around the knots are much smoother. Adding the piecewise effects is known as penalized spline regression or P-splines for short. The full model in matrix form is shown below:

$$\mathrm{X} = [1\ \mathrm{x}_i]_{1 \le i \le n} \qquad \mathrm{Z} = \begin{matrix} [\mathrm{x}_i - \kappa_{k_+}]_{\substack{1 \le i \le n \\ 1 \le k \le K}} \\ [\mathrm{x}_i - \kappa_{k'_+}]^2_{\substack{1 \le i \le n \\ 1 \le k' \le K}} \end{matrix} \qquad \mu = [\mu_1, \ldots, \mu_k]^t$$

$$y = \mathrm{X}\beta + \mathrm{Z}\mu + \varepsilon, \quad \begin{bmatrix} \mu \\ \varepsilon \end{bmatrix} \sim N(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_\mu^2 I & 0 \\ 0 & \sigma_\varepsilon^2 I \end{bmatrix})$$

The ratio of σ_{μ}^{2} to σ_{ε}^{2} is known as the smoothing parameter and determines the amount of smoothing. The interpretation of this parameter [2] is as an indicator of the relative level of smoothing around the knots. A higher ratio indicates a more nonlinear curve. Mixed model software packages typically provides the option to estimate the smoothing parameter based on restricted maximum likelihood estimation of variance components or allow the parameter to be specified by the user.

An additional challenge in using GLMMs to smooth, is specifying the number and location of the knots. Ruppert (2002) provides the following rules for knot location and density:

$$\kappa_k = \ (k+1)/(K+2)^{\ th} \text{ sample quantile of unique } \mathrm{x_i s}$$

or all k knots. The general choice for the total number of knots, K is:

$$\mathrm{K} = \ \max\left(5, \min\left(\frac{\text{number of unique } \mathrm{x_i s}}{4}, 35\right)\right)$$

Refer to [9] for further discussion on default knot specification.

The most relevant parameter for interpretation of the model are the means of the fixed (β) and random (μ) effects.

Solving for the $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$ with GLMMs:

Initialize $\hat{y} = g(y)$ where is g is the link function (identity function for the Gaussian case)

1)Minimize the penalized sums of squares to obtain a new $\hat{\mu}\ and\ \hat{\beta}$

$$\sum_{i=1}^{n}(y_i - \mathrm{X}_i\beta + \mathrm{Z}_i\mu)^2 + \ \lambda\left(\mu^t\sigma_{\mu}^{2}I^{-1}\mu\right)$$

Where λ is the smoothing parameter.

2)Set $\ \widehat{\mathrm{y}} = \mathrm{X}\hat{\beta} + \mathrm{Z}\hat{\mu}$ and repeat

Note the Linear Mixed Model case does not need to be iterated.[2] For a full break down of the minimization algorithm as well as the iterated sampling technique to for the generalized case see Wand et al. [10]

Similarities and differences between GAMs and GLMMs

The largest commonality between GLMMs and GAMs is the penalty imposed on the log-likelihood to ensure that GAMs remain economical with their use of parameters. It is similar to the constraint imposed on the predictors in GLMMs that require them to behave like a sample from the specified distribution family. With GLMMs this often causes the predictors to be less volatile and 'spread out' than would be separate parameter estimates. This could be beneficial analysis of large fragmentary data sets with many influential points sensitive to small perturbations in the models specification [11].

One important benefit of GLMMs over GAMs is the ability to accurately model correlated data in a highly interpretable manner. By assigning grouped processes to random effects terms, GLMMs may be able to capture the multi-correlation and allow the fixed term (and other possible random effects terms) an unbiased result. While not the focus of this thesis, it may be an important benefit to consider when deciding between the models [12].

GAMs are not necessarily encumbered by knots as GLMMs are. However once the number of knots is above a critical data dependent level, knot placement is not thought to significantly affect modeling [9]. Thus, once the degree of nonlinearity has been decided upon, the most prominent difference in smoothing GAMs and GLMMs are their fitting methodologies.

In GLMMs the level of smoothing may be estimated from the data [2], whereas the level of smoothing in GAMs must be specified by the user (see below on methods to do this). They also roughly obey the data structure of linear models whereas the only restriction on the predictors of GAMs are that they be additive. GAMs are more akin to numerical analysis techniques such as interpolation while LMMs rely on error reduction through a complex variance matrix.

Degrees of Freedom (df)

The number of degrees of freedom is the number of values in the final calculation of a statistic that are free to vary. In other words, it is the dimension of the vector space a statistic spans. A fitted curve within given data can be defined by its residual error. It follows that the df for the residuals is the df of the curve. For a multilinear regression (or Gaussian GLM) computed through OLS, the df of the fitted line is n-p where the response sample size is n and p is the number of parameters specified including the intercept term. The "n" points being estimated is constricted by being forced to satisfy p equations.

$$\varepsilon_1 + \cdots + \varepsilon_n = 0, \quad \mathrm{x}_{11}\varepsilon_1 + \cdots + \mathrm{x}_{1n}\varepsilon_n = 0, \quad \cdots, \quad \mathrm{x}_{(p-1)1}\varepsilon_1 + \cdots + \mathrm{x}_{(p-1)n}\varepsilon_n = 0$$

Effective Degrees of Freedom (edf)

The complexity of a GLM is proportional to the number of explanatory terms (model parameters) used. Every parameter corresponds to one degree of freedom (df) due to the assumption that each has only a linear relation with the data. Flexible regression models have nonlinear predictors that may require more than one df. The total number of equivalent degrees of freedom used by the model is known as the effective degree of freedom or edf [13]. For the purposes of this thesis df will be used interchangeably with edf.

The edf is controlled by α, the smoothing parameter. [14] The smoothing parameter controls the amount of smoothing in GLMMs and has an analog in GAMs and other non-parametric smoothing techniques. The edf is a measurement of the additive df accrued from fitting the smoothing terms to GAMs, GLMMs, or other flexible models. Calculation of the edf can be explained through the mechanics of regression [15].

GLMMs and many additive regression methods use regularized (generalized and/or penalized) least-squares, so edf defined in terms of dimensionality are generally not useful for these procedures. However, these procedures are still linear in the observations, and the fitted values of the regression can be expressed in the traditional form $\hat{y} = \mathrm{H}y$, where $\hat{y}$ is the vector of fitted values at each of the original covariate values from the fitted model and y is the original vector of responses. H is the "hat matrix" in linear regression or, more generally as in GAMs and GLMMs models, the smoother matrix. Nonlinear GLMMs generate their hat matrices via iterated fitting [2]. With respect to different existing methods of fitting GAMs, the hat matrix may be numerically constructed [16]. However, because H does not correspond to an ordinary least-squares fit (i.e. is not an orthogonal projection), the sum-of-squares no longer have (scaled, non-central) chi-squared distributions, and geometrically-defined df are not potentially as useful. Below is the explicate form of the hat matrix (available when fitting normal GLMMs). Note there is a 1 to 1 and monotonic relationship between α and $tr\big(\mathrm{H(k)}\big)$ or the edf

$$\hat{\beta} = \left(\mathrm{X^tX} - \lambda\sigma_\mu^2\mathrm{I}^{-1}\right)^{-1}\mathrm{X^tY}$$

$$\hat{y} = \mathrm{X}\left(\mathrm{X^tX} - \lambda\sigma_\mu^2\mathrm{I}^{-1}\right)^{-1}\mathrm{X^tY}$$

$$\mathrm{H(k)} = \mathrm{X}\left(\mathrm{X^tX} - \lambda\sigma_\mu^2\mathrm{I}^{-1}\right)^{-1}\mathrm{X^t}$$

The edf of the fit can be defined in various ways to implement goodness-of-fit tests, cross-validation and other inferential procedures. Here, we use the form tr(2H - H H'). [17]. Once the total edf is known then the df allocated to modeling the non-linear variable of interest can be explicitly determined.

In GAMs and GLMMs, the smoothing parameter, and the edf, can be selected a priori, graphically, or to optimize some criteria that balance the model's fit (usually a function of the likelihood function) and overspecification (dependant on sample size and the actual EDF). Here we focus on data dependent edf selection. One downside to data dependent edf selection strategies is that it precludes formal statistical inference about the nonlinearity of the curve [18]. This would be an appropriate course of action if a priori selection of the df could be well founded on previous studies and background information. However, this information is often not available, and previously unknown nonlinearities of higher degree may still be missed with a priori df selection

Alternative methods to select edf

An alternative method for selecting the df in mixed models was presented by Cantoni and Hastie in 2002. The paper suggests a test statistic as opposed to any mean squared error based criteria. However the test necessitates the declaration of null and alternative hypothesis edfs. These declarations may be subjective to the researcher's definition of an acceptable threshold of nonlinearity. Also, based on the paper's simulation, the test power (alpha = 0.05) only exceeds 80% if the difference between the hypothesis is above four edf.

Alternatives to GAMs and GLMMs for fitting nonlinear associations There are many strategies available to deal with nonlinear dose response curves. Here I outline some options.

Dichotomization

Dichotomization on independent prediction can be a useful method to simplify ones analysis and potentially more effect future decision making with the model. Interpretability and inference are straightforward and easy (ANOVA, t-test, etc). It is sometimes essential in the cases of "soft" statistics when over-quantification would introduce some bias. However there are limited situations when dichotomizing well behaved continuous data is necessary and beneficial [19]. These variables should be left to be modeled in their natural distributions (e.ge guassian). Some specific downsides to dichotomization in critical velocities are discussed below.

Transformations

Prior to the modeling of data, the response or predictor variable may be transformed with the goal of maintaining the regression assumptions (homogeneity, linearity, over/under dispersion or the correct distribution). Nonlinear dose response curves may violate one or more of these principles so a well specified transformation could possibly correct these. Transformations are generally less complex (use fewer parameters) than a polynomial regression with the similar level of accuracy. [20]

As in all types of exploratory analysis, attempting multiple transformations may lead to inflated type I errors. In small samples with linear fit featuring heterogeneity, it may be easy to infer a nonlinear relationship. Eyeballing the data and "connecting the dots" may allow post specified transformations to be a significant source of bias. Most texts recommend response and covariate transformations to be scientifically motivated as opposed to statistically motivated. Texts often recommend changing models over changing data as to avoid the issue entirely. [21]

Polynomial Regression

Often, when dealing with a smaller number of possibly nonlinear covariates it may be beneficial to add polynomial terms to a covariate. If there exists some true function relating a covariate to the response it can be approximated as:

$$\mathrm{y} = \mathrm{a}_0 + \mathrm{a}_1\mathrm{x} + \mathrm{a}_2\mathrm{x}^2 + \mathrm{a}_3\mathrm{x}^3 + \mathrm{a}_4\mathrm{x}^4 + \ \dots + \ \mathrm{a}_\mathrm{m}\mathrm{x}^\mathrm{m}$$

with the error term inversely proportional to the number of polynomial terms. This approach seems to fail with all but the simplest type of non-linear data. In lower dimensions sharp changes in slope will not be accurately modeled as there are a limited number of shapes a lower dimensional curve can hold. As the number of polynomial terms increases the correlation between each of the terms will add up. This results in multicollinearity which can severely reduce the power of a linear model. Centering the data by subtracting the means will reduce this multicollinearity, but correlation will remain.

Bayesian Methods

Given that for almost all analyses computation is no longer a restriction, it is important to mention that there are several Bayesian approaches to curve fitting. To my best knowledge there has not been any systematic comparison of Bayesian vs Frequentist flexible curve fitting. However Bayesian curve fitting does compensate for major Frequentist modeling drawbacks. The curve is no longer conditionally dependent on a

prespecified (or criterion specified) smoothing parameter. Also the specification of basis or knots make GLMM and GAM modeling work more effectively with homogeneous (variance) functions. Heterogeneous functions are cumbersome for these models [22].
The Bayesian mode is as follows:

$$\mathrm{Y}_i|\mathrm{X} \sim p\ (y|f(\mathrm{X}_i), \eta)$$

Where the goal is to estimate the unknown function f by maximizing the posterior distribution p with $(f(X_i), \eta)$ parameters. The posterior probability distribution is the unknown distribution conditional on the evidence obtained from an experiment or survey. Model terms are additive:

$$\mathrm{Y}_i = f(\mathrm{x}_i) + \varepsilon_i$$

There are many possible specifications for the structure of the Bayesian curve fitting model. Computation is dependent of this. One technique is known as piecewise polynomials. Like GLMM knots divide the data in to subintervals. Unlike GLMM each subinterval fitted with a low order polynomials (with priors and estimated again to maximize the posterior distribution) [23].

$$\mathrm{y}_i = \beta_0 + \beta_1\mathrm{x}_1 + \cdots + \beta_m\mathrm{x}_m + \sum_{j=1}^{J}\sum_{k=1}^{K} \underbrace{\mu_k(\mathrm{x}_i - \kappa_k)^j{}_+}_{piecewise\ effects} + \varepsilon_i$$

Bayesian modeling opens the possibility to highly complex models.

Once a model framework is established with formulae, distributions, and variable one must decide on the proper method of evaluating and comparing models.

Goodness of fit

Measures of goodness of fit usually summarize the discrepancy between observed values and the values expected under the model in question. Some results from this can be used in statistical hypothesis testing [24].

There are several types of fit one can measure ranging from sampling from a distribution to fitted values on linear models. Thus one can test to see if two sets of observations come from the same distribution or if one set follows a specified distribution. This is particularity useful in linear modeling since a key assumption is that the residuals are normally distributed. There are a number of commonly used, both Frequentist and Bayesian, tests for normality.

For a basic regression analysis with model assumption as specified above, an established indicator of goodness of fit is the ratios of the sum of squares. The sums of square represent various squared error totals (fitted vs observed, average vs fitted, etc). From this an F statistic can be constructed:

$$F = \frac{\mathrm{Residual\ S.S./df_r}}{\mathrm{error\ S.S./df_e}}$$

This does not give a relative definition for comparison between models.

The coefficient of determination offers a more relative (for comparison between similar models) evaluate of goodness of fit.

$$\mathrm{TSS} = \sum_{i=1}^{n}(y_i - \overline{y})^2 \quad \mathrm{RSS} = \sum_{i=1}^{n}(\hat{y} - \overline{y})^2 \quad \mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y})^2$$

Sums of square lose relevance with non-Gaussian GLMs. The analogous statistic in GLMs is called the model deviance. Two models may be tested under the null hypothesis that they are not statistically different using their model deviance. This is known as a likelihood ratio test. It forms a chi square distribution with k degrees of freedom under the said null hypothesis where k is the difference in the number of parameter of each of the models being compared.

$$\text{Deviance (Model}_1\text{)} \quad = \quad -2\left[\ \log\big(p(y|\theta_1)\big) \ - \ \log\big(p(y|\theta_{\text{full model}})\big)\ \right]$$

As shown before, when dealing with the Gaussian distribution the residual deviance can be reduced to the residual sum of squares. The same also applies to the null deviance reducing to the total sum of squares It has recently been show.[25]. Deviance/Likelihood is important for model selection. It forms the main component of data driven information criterion commonly used to select between comparable models.

Kullback Leibler distance

The Kullback Leibler distance (KL-distance) is a natural distance function [26] from a "true" probability distribution, p, to a "target" probability distribution, q. It is heavily based on information theory. The KL distance can be interpreted as summed (or integrated) log odd ratios of the two distribution weighted by a reference distribution[27-29].

Continuous

$$\mathrm{D_{KL}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx}$$

Discrete

$$\mathrm{D_{KL}(P \parallel Q) = \sum_{i} p(i) \log \frac{p(i)}{q(i)}}$$

When comparing curves generated from a single common variable, the KL-distance has an interpretation as a "distance" metric. It is essentially the logged "distance" between the curve and normalized by its location on the distribution [26].

The KL-divergence of a statistical model where the true curve is not known may be estimated, to within additional terms, by a function (like the squares summed) of the deviations observed between data and the model's predictions. Estimates of such divergence for models that share the same additional terms can in turn be used to choose between models. Many of these functions overlap with the concepts of goodness of fit and deviance mentioned about. We refer to them as data driven information criteria. [28]

Information Criteria

Burham and Anderson approach model selection as a two stage process: selection of candidate models followed by a scientific and objective comparison between such models [30]. As this thesis examines GAMs and GLMMs the first objective is not a concern. The only level of flexibility in models under direct comparison is the extent of nonlinearity of the variable of interest. Which level of nonlinearity optimizes the goodness of fit amongst the candidate model? With a basis in information theory, various information criteria provide such a scientific and objective comparison.

Akaike's Information Criterion (AIC) and AICc

In the general case, the AIC is

$$\mathrm{AIC\ (Model_1) = 2k - \log\big(p(y|\theta_1)\big)}$$

where k is the number of parameters in the statistical model, and L is the maximized value of the likelihood function for the estimated model.

From a set of potential models for the data, the preferred model is the one with the minimum AIC value. With the two parameters in the formula, AIC not only rewards goodness of fit, but also includes a penalty that is an increasing function of the number of estimated parameters. This penalty discourages overfitting since increasing the number of free parameters in the model always improves the goodness of the fit [31].

AIC is based on information theory. Suppose that the data is generated by some unknown process I. Consider two candidate models to represent the data: m_1 and m_2. Knowing the true model one could find the information lost from using m_1 to represent I by with the KL of I and m_1; similarly, the information lost from using m_2 to represent I would be found by calculating the KL of I and $m_{2.}$ One would then choose the model that minimized the information loss. [29]

Since the true model m is not known the KL distance cannot be verified. Akaike [31] showed, however, that we can estimate, via AIC, how much more (or less) information is lost by m_1 than by m_2. It is remarkable that such a simple formula for AIC is the result. The estimate, though, is only valid asymptotically; if the number of data points is small, then some correction is often necessary. The derivation examines the mean expected maximum log likelihood (MELL).

$$E_{Y|\theta}E_{X|\theta}\left[\log\ (X|\hat{\theta}^k(Y)\ )\right] = E_{X|\theta}\left[\log\ (X|\hat{\theta})\right] - \frac{k}{2}$$

$$E_{X|\theta}\left[\log\ (X|\hat{\theta})\right] = E_{X|\theta}\left[\log\ (X|\hat{\theta}^k(Y))\right] - \frac{k}{2}$$

$$E_{Y|\theta}E_{X|\theta}\left[\log\ (X|\hat{\theta}^k(Y)\ )\right] = E_{X|\theta}\left[\log\ (X|\hat{\theta}^k(Y))\right] - k$$

$$\sim 2k - 2\ln(L)$$

Where Y is the response variable and X is that data matrix. $\log\ (X|\hat{\theta}^k(Y)\ )$ is the log likelihood for a model with estimated k parameter vector $\hat{\theta}^k$.

Akaike showed that the expected maximized log likelihood of a $\hat{\theta}^k$ parameterized model is a biased estimate of MELL. The bias is asymptotically equal to k, the number of estimable parameters in the model.

As the AIC was designed to approximate the expected Kullback-Leibler distance, a measure of the difference between the true and estimated curves, a smaller AIC is more desired. However, as the model dimension increases in relation to the sample size, AIC underestimates the KL-distance, which may lead to overfitting.

The AIC is meaningless as an absolute number and is always compared with a similar model of differing size. Changes in AIC must be evaluated as a strong or weak improvement if it is expected to influence model selection. Burnham & Anderson (2002, p.446) offer the following for deciding on an absolute difference that signifies improvement.

> *As a rough rule of thumb, models having their AIC within 1–2 of the minimum have substantial support and should receive consideration in making inferences. Models having their AIC within about 4–7 of the minimum have considerably less support, while models with their AIC > 10 above the minimum have either essentially no support and might be omitted from further consideration or at least fail to explain some substantial structural variation in the data.*

AICc [30] is a proposed correction to the AIC that accounts for smaller finite sample sizes:

$$\mathrm{AIC} = \mathrm{AIC} + \frac{2k(k+1)}{n-k-1}$$

where k denotes the number of model parameters. Thus, AICc is AIC with a greater penalty for extra parameters.

Burnham & Anderson recommends using AICc, rather than AIC, if n is small or k is large. Using AIC, instead of AICc, when n is not many times larger than k^2, may increase the probability of selecting models that have too many parameters, i.e. of overfitting. The probability of AIC overfitting can be substantial, in some cases. Despite this many modern studies continues to use only AIC to help select their optimal models.

$$\lim_{n\to\infty} 2k - 2\ln(L(\hat{\theta})) = -2\ln(L(\hat{\theta}))$$

So in addition to low sample size bias, AIC is commonly optimized by a greater number of parameters even in large samples. In a sense there is just more explanatory room in large models so adding a parameter has less of a dramatic effect. Some other information criteria (BIC, HQ) are known to have an asymptotic desirable property, the so-called consistency or dimension consistent [32].

Bayesian Information Criterion (BIC)

In the general case, the BIC is

$$\mathrm{BIC}\ (\mathrm{Model}_1) = \mathrm{k} * \ln(\mathrm{n}) - \log\big(\mathrm{p}(\mathrm{y}|\theta_1)\big)$$

The philosophy behind the BIC is that the candidate model corresponding to the minimum value of the criterion is also the model corresponding to the highest Bayesian posterior probability. BIC was originally justified for the limited case of independent, identically distributed observations within a linear model under the assumption that the likelihood is from the regular exponential family [33]. This justification has become more generalized over time [34]. If we assume k to be the number of specified parameters of a candidate model (degrees of freedom), then the goal is to maximize p(k|y) which is proportional to distribution on the parameters.

$$-2\log\big(\mathrm{p}(\mathrm{k}|\mathrm{y})\big) \ = \ -2\log\left(\pi(\mathrm{k}) \int \mathrm{p}(\theta|\mathrm{y})\, \mathrm{p}_*(\theta|\mathrm{k}) d\theta / \mathrm{m}(\mathrm{y})\right)$$

$$\sim -2\log(\pi(\mathrm{k}) \int \mathrm{p}(\theta|\mathrm{y})\, \mathrm{p}_*(\theta|\mathrm{k}) d\theta)$$

$$\mathrm{p}_*(\theta|\mathrm{k}) \sim 1, \text{2nd degree Taylor expansion of } \log\big(\mathrm{p}(\theta|\mathrm{y})\big)$$

$$\sim -2\log\left(\pi(\mathrm{k})\log\left(\mathrm{p}(\hat{\theta}|\mathrm{y})\right) \int \exp\left(\frac{1}{2}(\theta-\hat{\theta})^{\mathrm{t}}\log\big(\mathrm{p}(\theta|\mathrm{y})\big)''\,(\theta-\hat{\theta})\right) d\theta\right)$$

$$\sim \ -2\log\big(\pi(\mathrm{k})\big) - \log\left(p(\hat{\theta}|y)\right) - \ (2\pi)^{\mathrm{k}/2}\big(\mathrm{nI}(\hat{\theta}|\mathrm{y})\big)^{-1/2}$$

$$= -2\log\big(\pi(\mathrm{k})\big) - \log\left(\mathrm{p}\big(\hat{\theta}|\mathrm{y}\big)\right) + \mathrm{k}\log\ \left(\frac{\mathrm{n}}{2\pi}\right) + \log\big(\mathrm{I}(\hat{\theta}|\mathrm{y}\big)$$

$$\sim \ \mathrm{k} * \ln(\mathrm{n}) - \log\big(\mathrm{p}(\mathrm{y}|\theta_1)\big)$$

This results in both AIC and BIC both being justified, but in unique and different ways. BIC makes the assumption of a flat prior distribution ($\pi(\mathrm{k}) \sim 1$) whereas AIC is affected by the decision to minimize information lost.

In this thesis, I examine three commonly used criteria, Akaike's Information Criterion (AIC) [31], the Bayesian Information Criterion (BIC) [33], and Generalized Cross Validation (GCV) [35]. BIC is generally more conservative, while AIC and GCV are more aggressive in their allocation of EDF [32]. Other criteria are less frequently used such as unbiased risk estimator, corrected GCV, but are not used here. [36].

Generalized Cross Validation (BIC)

$$\text{GCV (Model}_1) = \frac{1}{\text{n}}[\text{I}-\text{A(k)y}]^2 / \left[\frac{1}{\text{n}} Trace(\text{I}-\text{A(k)})\right]^2$$

Where n is the number of observations in the dataset, k is the number of model parameters (edf taken in to account), and A(k) = $X(X^TX\text{- }nK\mathit{I})^{-1}X^T$ [35]

GCV was originally devised as a method for selecting ridge parameters for solving general ridge regression problems [35]. Ridge regression often serves as the basis for solving GLMMs and in some software packages (e.g R's mgcv package) GAMs [2, 37]. The theoretical limitations of GCV as a popular technique for the selection of tuning parameters for smoothing and penalty for nonlinear models are acknowledged by authors [38, 39]. It is put forward as an effective criterion for a model that is low on extra df.[35] Few alternatives have been put forward that preserve the cross validation basis of GCV. Successful candidates approximate the UBRE or AIC [40].

In computationally intensive scenarios or models with many possible nonlinear covariates selecting EDF by information criteria is not ideal. Wood 2004 suggests a method based on minimizing the QR decomposition, which is both stable and efficient. It is limited to models using a GAM fitted by using penalized iteratively re-weighted least squares [41].

Real life datasets

In addition to the statistical simulation, this thesis investigates two real life data sets: The effect of smoking history on lung function and vital parameters on mortality in post-surgery ICU patients. The following is a brief review on how each is currently modeled in literature.

Review of Smoking modeling and implication on statistical inference and interpretability

The effect of tobacco smoking on lifespan and various diseases such as lung cancer has been modeled by scientists since at least the 1940s [42, 43]. Smoking is a traditionally hard to model variable. A simple and common method is the categorical (indicator) variable signaling that the subject either has or has not smoked at some point in the past. More sophisticated representations assume a linear (or nonlinear) relationship between various aspects of smoking history such as the smoking intensity and or the total duration of smoking and health outcomes. A more recent development, Pack Years combines intensity and duration. Pack Years is defined by multiplying the number of packs of cigarettes smoked per day by the number of years the person has smoked {NCI, 2009 #93}. The decision of which model to choose is based on lowering bias and/or achieving a desired efficiency (power), and interpretability [44]. It is a fine balance and too large a selection of choices introduces its own bias from multiple fittings.

A number of variables have been established to predict the extent of smoking damage (both acute and chronic). These variables are correlated with each other so including them together within any linear model would be disastrous. A non exhaustive list includes pack years, smoking duration, time since smoking cessation, smoking intensity, and cumulative smoking weighted by time [45, 46]. In addition, several investigators have shown evidence of a nonlinear relation between smoking and disease outcomes linked to sustained smoking such as lung cancer, periodontal damage, and emphysema [44]. Light and early smokers seem particularly sensitive to increasing lung damage from stepping up their habit when compared to average smokers who increase smoking. There may also be a leveling of risk at higher intensities [47]. Few studies undertake a systematic approach to assessing nonlinearity as it is usually of secondary concern. One exception is the Comprehensive Smoking Index (CSI) which seeks to develop an ideal single transformation, *a priori,* onto a linear smoking variable[44, 46]. However there is not a consensus on the exact form of the CSI.

Review of critical velocity (and other vital indicators) modeling

Several studies have used smoothing to improve inference when modeling mortality [48, 49]. Mortality is often triggered by extreme and abrupt conditions. Give this, it is logical that modeling the effect of an exposure on mortality should be flexible enough to be both nonlinear and somewhat discontinuous [48].

The benefits of flexible modeling with respect to identifying critical rates within the body have been known for some time [49]. To the author's knowledge there has been little work applying flexible regression models to vital parameters such as oxygen delivery, cardiac index, and serum lactate concentration outside of measuring kinetics in response to exercise stimulation. Much research has been focused on dichotomizing the parameter of interest.

Cardiac Output is the volume of blood being pumped by the heart, in particular by a left or right ventricle over one minute. [50]. Mixed Venous Oxygen Saturation (MVO_2) is dependent on arterial oxygen saturation, hemoglobin concentration, cardiac output, and tissue oxygen demands. O_2uptake does not desaturate blood hemoglobin more the necessary [51] and a drop in saturation implies the body is drawing an addition percentage of $O_2$from the blood. Such a drop usually implies anemia, arterial oxygen desaturation, and/or decreased cardiac output. Serum lactate is involved in the conversion of pyruvate and lactate when tissue O_2levels are low. Elevated initial through 24-hour lactate levels have been shown to be significantly correlated with mortality [52]. DO_2 is the delivered oxygen. It explains the rate at which oxygen reaches the organ tissues. The normal state DO_2 is more than sufficient to meet the demands of all tissues and organs. Even with a moderate reduction in DO_2, the MVO_2can slightly adjust to compensate. When DO_2 drops below the critical DO_2, MVO_2 becomes supply dependent [53].

A threshold may exist in the associations between mortality and these parameters. The primary method most papers adopt is to specify a specific threshold based on previous studies and then dichotomize the variable [54, 55]. This allows for direct hypothesis testing within a linear model framework. However, there is considerable loss of power and increases the prevalence confounding [19, 56]. Vital indicators within humans such as those examined in the second stage of this thesis (oxygen delivery, oxygen saturation, cardiac index, and serum lactate) are often both directly and indirectly related with each other and the outcome of mortality [57, 58]. Rather than specifying a cutpoint in advance, sometimes a cutpoint is derived from the data – however, this may lead to serious bias.

These factors lead to a significant variation in cutpoints selected in papers which seek to establish the critical values of vital indicators. Critical in this context refers to the level associated with the onset of negative symptoms. The negative outcomes associated with these thresholds are shock and increased morbidity. Critical mixed venous oxygen saturation has been measured as anywhere between 40% and 70% [54, 59, 60]. The critical Cardiac index is reported between 1.8 and 2.2 L / min / m^2 [50, 61, 62]. Critical Delivered Oxygen is reported as between 7-10 ml/kg-min [63]. Critical lactate is reported as between 2-4 mmol / L [64, 65]. However, some authors propose there is no theoretical basis to have a threshold concentration serum lactate and that the relationship tends to increase in a somewhat linear fashion [64].

Flexible modeling would address the power and bias engendered by dichotomization in linear models [1]. Thus GAMs and GLMMs have the potential to provide a better estimation of the critical value.

The interpretation of nonlinear smoothing is less straightforward than for plain GLMs. What was previously perceived as a hard and fast critical value may in fact be modeled as a rounded plateau when plotted. The sharpness (and by extent the edf) can serve as evidence for a critical value. One drawback that remains is that it is a heuristic measurement, albeit and transparent and consistent one. It also has the potential to show repeatability in other similar data sets. GAMs have been proposed for isolating thresholds, but have encountered difficulties [18].

Objectives

The objective of this thesis is an analysis, both representative and comprehensive, and comparison of flexible modeling of nonlinear curve with Generalized Additive Models and Generalized Linear Mixed Models. The goal will is to achieve repeatable results via simulation and examine the process of analysis of real life data from representative medical data

Linking Statement

This manuscript compares GAMs and LMMs in their capacity to fit non-linear curves as discussed in the literature review, and addresses both objectives of the thesis. This includes a simulation and real life data component using smoking and lung function data. The simulation examines a variety of data conditions for robustness. Different sample sizes, variances, and degrees of nonlinearity are considered. Edf is also selected by optimizing each of the information criteria discussed in the literature review in both the simulation and real life data analysis. The results are captured both in tables which describe the relevant dfs and optimized information criteria and in graph of the nonlinear (and linear for reference) curves which have been isolated from the GAMs and LMMs.

Full title: Flexible modeling using Generalized Additive Models and Linear Mixed Models

Short title: Flexible modeling using GAMs and LMMs

Daniel Hercz[1]
Andrea Benedetti[1,2]
Jean Bourbeau[1,2]

[1]Department of Epidemiology, Biostatistics and Occupational Health, McGill University
[2]Respiratory Epidemiology and Clinical Research Unit, Department of Medicine, McGill University

Corresponding Author: Andrea Benedetti

Address: Montreal Chest Institute, K-135
3650 St. Urbain
Montreal, QC
H4A 2Z6
Telephone: (514) 934-1934 ext. 32161
Fax: (514) 843-2083
Email: Andrea.benedetti@mcgill.ca

Abstract

Generalized Additive Models (GAMs) and Linear Mixed Models (LMMs) are two methods of fitting curves to data without strong *a priori* assumptions about the functional form of the association under study. The extent of the smoothing may be selected via data-dependent approaches such as the Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC) or generalized cross-validation (GCV).

A simulation study was performed to compare GAMs to LMMs with smoothing selected to optimize AIC, BIC, or GCV. Sample size, functional form and strength of the association under study were varied. LMMs outperformed GAMs in situations of high variability and low sample size. Under more ideal condition GAMs outperformed although most models were roughly in agreement in these scenarios.

The approaches were also applied to investigate the effect of smoking intensity and duration on lung function in COPD. Overall, there seemed to be limited evidence of nonlinearity in the associations between smoking and lung function.

Keywords: Linear mixed models, smoothing, generalized additive models, chronic obstructive pulmonary disease

Generalized linear models (GLMs) are the backbone of most epidemiologic statistical analyses. Usually, in these analyses, the shape of the association between the predictors and the outcome is assumed to have a linear or other *a priori* specified functional form (e.g. quadratic) [1]. While tests for significance are straightforward in this scenario, the required *a priori* specification of the functional form may introduce bias and loss of statistical power [66].

Flexible modeling can achieve higher power and better fit in the presence of nonlinear relationships and significant noise in the predictor variables. One of the challenges inherent in flexible modeling is specifying how complex a curve should be fit.

Whereas the complexity of a GLM is proportional to the number of explanatory terms used, with each parameter assumed to have linear effect on the outcome corresponding to one degree of freedom (df); nonlinear predictors may require more than one df. The total number of equivalent df used by the model is known as the effective degree of freedom (edf).

The Generalized Additive Model (GAM) generalizes the conventional regression model to:

$$y = \beta_0 + \sum_{X \in A} \beta_i X_i + \sum_{X \in B} S_j(X_j) + \varepsilon$$

where $S_j(X_j)$ are smooth functions whose shapes are estimated directly from the data for those predictors $(X \in B)$ modeled non-parametrically, and $\varepsilon \sim N(0, \sigma^2)$. A number of smoothing methodologies are available in GAMs, though here we focus on cubic smoothing splines [3, 4].

Another approach to flexible modelling is to use Linear Mixed Models (LMMs). LMMs generalize linear models by including random effects, usually in order to properly account for correlated or clustered observations (e.g. to model data arising from longitudinal studies) [5, 6]. By specifying a number of predefined points on the data, called knots [7] and including a series of knot-dependent basis functions whose regression coefficients appear in the model as random effects we can achieve a smoothed curve [8]. In LMMs the level of smoothing may be estimated directly from the data, or user-specified [7].

The LMM regression equation is:

$$y = \beta_0 + \beta_1 x + \cdots \beta_p x^p + \sum_{k=1}^{K} \mu_k (x - \kappa_k)_+^p + \varepsilon_i;$$

$$\text{where } \begin{aligned} (x-\kappa_{k=})_+^p &= 0 \text{ if } x \le \kappa_k \\ &= (x-\kappa_{k=})_+^p \text{ if } x > \kappa_k \end{aligned} ;$$

and β_i represent the fixed effects; the μ_k represent the normally distributed random effects ($\mu_k \sim N(0, \tau^2)$); the κ_k are the pre-specified knots, the ε_i are normally distributed errors ($\varepsilon_i \sim N(0, \sigma^2)$), and p is the degree of the polynomial. The variance of the random effects, τ^2, is usually estimated by Restricted Maximum Likelihood (REML), though can also be arrived at by minimizing the penalized least squares formulation: $\sum(Y-\hat{Y})^2+\lambda\sum\mu_k$ [3]. In this representation, $\sigma^2/\tau^2 = \lambda$ is the “smoothing parameter” [9].

The smoothing parameter, λ, controls the level of smoothing in LMMs, and is analogous to the smoothing term in a smoothing spline equation [2]. In both LMMs and GAMs, the smoothing parameter is directly related to the calculation of the edf [10]. The edf is a measurement of the additive df accrued from fitting the smoothed terms in GAMs, LMMs, or other nonlinear models [2].

In GAMs, and other flexible regression methods, the fitted values can be expressed in the traditional form $\hat{y} = Sy$, where $\hat{y}$ is the vector of fitted values at each of the original covariate values from the fitted model and *y* is the original vector of responses. S is the “smoother matrix” (analogous to the “hat matrix” in linear regression [3]).The edf of the fit can be defined in various ways to implement goodness-of-fit tests, cross-validation and other inferential procedures. Here, we use df=tr(2H - H H') as it is less prone to numerical error, though other definitions are possible [11].

The smoothing parameter or edf explicitly control the bias-variance trade-off [12]. Choosing the correct df is not trivial, and the estimated curve can differ qualitatively depending on the edf. In GAMs and LMMs, the smoothing parameter, and the edf, can be selected using various criteria. This paper examines three commonly used criteria, the Akaike Information Criterion (AIC) [31], the Bayesian Information Criterion (BIC) [66], and Generalized Cross Validation (GCV) [13]. BIC is generally more conservative, while AIC and GCV are more aggressive in their allocation of edf [8].

One downside to data dependent edf selection strategies such as these is that it precludes standard formal statistical inference about the nonlinearity of the curve [14][67]. While *a priori* selection of the edf may be well founded on previous studies and background information, usually there is not enough information. Even then, previously unknown nonlinearities of higher degree may be missed.

The main goal of this paper is to investigate and compare the modeling of nonlinear relationships using LMMs and GAMs via simulation study and then to illustrate their use to investigate the effect of smoking on lung function in subjects with chronic obstructive pulmonary disease.

Methods

Data generation

The independent variable (X) was generated from a uniformly spaced interval constant (0-40 and 0-200). The continuous dependent variable (Y) was generated from a linear (y = 0.03x for the large sample datasets) or nonlinear function $\left(y = 4/_{x^{0.25}} \text{ or } y = 4|\cos 0.02x| \text{ for the large sample datasets (see Figure 1)}\right)$ with normal error distribution, and similar functions for the smaller sample datasets. Two levels of variance were attached to the error term to reflect varying levels of uncertainty (standard deviations of 0.7 and 2) (see Figure 2). All the curves were scaled to fit in a response range of roughly between 0 and 6. This scaling is generalizable to any other dataset and allow for visual comparisons to be made. Sample sizes of 40 and 400 were investigated. The sample sizes were chosen to mimic common epidemiological datasets. Within each permutation of sample size, distribution of X, type of X-Y association, and variance, individual datasets were generated independently 500 times over.

Data analysis

For each data sample, LMMs and GAMs were fit to the generated data. For GAMs, smoothing splines were used and the models were estimated using the backfitting algorithm [1] using the gam package in R. For LMMs, restricted maximum likelihood was used to estimate the model, with one knot specified per every 6 observations, uniformly spaced on the domain of the covariate [15]. The "gam" package was used in R [68]. Code for the LMM was adapted from Wand [10].

Finally, for each generated data sample, a linear model was also fit.

For both GAMs and LMMs, flexibility of the curve was selected by fitting curves with 1 to 20 edfs with increments of 0.25 and choosing as the final model that which optimized AIC, BIC or GCV. The df that resulted in the lowest AIC, BIC, or GCV was chosen as the "optimal df". All models were scaled to be roughly on the same domain and range so visual comparisons in curves could be assessed.

AIC BIC, and GCV were calculated according to the following formulas:

$$\mathrm{AIC}\ (\mathrm{Model}_1) = 2\mathrm{k} - \log\big(\mathrm{p}(\mathrm{y}|\theta_1)\big)$$

$$\mathrm{BIC}\ (\mathrm{Model}_1) = \mathrm{k} * \ln(\mathrm{n}) - \log\big(\mathrm{p}(\mathrm{y}|\theta_1)\big)$$

$$\mathrm{GCV}\ (\mathrm{Model}_1) = \frac{1}{\mathrm{n}}[\mathrm{I}-\mathrm{A}(\mathrm{k})\mathrm{y}]^2 / \left[\frac{1}{\mathrm{n}} Trace(\mathrm{I}-\mathrm{A}(\mathrm{k}))\right]^2$$

Where n is the number of observations in the dataset, k is the number of model parameters (edf taken in to account), and $\mathrm{p}(\mathrm{y}|\theta_1)$ is the likelihood of all the model event occurring given the model's assumptions on its distribution.
Also $\mathrm{A}(\mathrm{k}) = \mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X}\text{- } \mathrm{nK}I)^{-1}\mathrm{X}^{\mathrm{T}}$

Performance of methods

We investigated how close the X-Y association curves estimated via GAMs or LMMs with df chosen to optimize a given criterion were to the true associations via the Kullback-Leibler distance (KL-distance) [16]. We assessed the KL-distance across the entire independent variable range as well as over the tails (defined as the lower and upper 10%), using the following formula:

$$\mathrm{D_{KL}}(\mathrm{P} \parallel \mathrm{Q}) = \int_{-\infty}^{\infty} \mathrm{p}(\mathrm{x}) \log \frac{\mathrm{p}(\mathrm{x})}{\mathrm{q}(\mathrm{x})} \mathrm{dx}$$

$$\mathrm{D_{KL}}(\mathrm{P} \parallel \mathrm{Q}) = \sum_{n \in N} \mathrm{p}(\mathrm{y}_n|\mathrm{x}_n) \log \frac{\mathrm{p}(\mathrm{y}_n|\mathrm{x}_n)}{\mathrm{q}(\mathrm{y}_n|\mathrm{x}_n)} \mathrm{dx}$$

The COLD data analysis

Chronic obstructive pulmonary disease (COPD) is one of the most common causes of morbidity and mortality in the US. COPD is characterized by airflow limitation that is not fully reversible [69]. COPD, once present, may have a lengthy and costly course, with significant impacts on quality of life [69]. Smoking is the leading cause of COPD, being a primary factor in up to 90% of all cases. Smoking is a traditionally hard to model variable [70]. Several investigators have shown evidence of a nonlinear relationship between sustained smoking and disease outcomes such as lung cancer, and COPD [71]. However, few studies have undertaken a systematic approach to assessing nonlinearity as it is usually of secondary concern.

The “Canadian Obstructive Lung Disease” Initiative (COLD study) is a national study of COPD and the first population-based lung health study including spirometry measurements in Canada. Design and data collection have been described in detail elsewhere [72]. Information on spirometric measurements, patient characteristics (such as age, sex, socio-economic status, medical history, etc.), and smoking history were collected. Here, we analyze data from the Montreal site [73]. Forced expiratory volume (FEV1), is an important in the diagnosis of obstructive and restrictive lung disease. It is the volume exhaled during the first second of a forced expiratory maneuver started from the level of total lung capacity.

Pack years was investigated as the exposure. Non-smokers were given a 0 for these variables [24]. Models adjusted for weight, age and sex. GAMs and LMMs were fit, as described in the simulation study methods section.

The primary objective for the analysis of the COLD dataset was to compare traditional regression models that assume a linear association between smoking and lung function to flexible modeling (via GAMs and LMMs).

Results

For larger datasets (n=400), GAMs and LMMs showed little difference in allocating edfs (Tables 1, 3 and 5). This was the case whether the edfs were selected by minimization on any information criteria (AIC, BIC, or GCV). As expected more edfs were allocated as the nonlinearity of the data increased (linear vs. nonlinear vs. discontinuous X-Y association).

AIC and GCV allocated roughly the same edfs at all levels of nonlinearity (Tables 1 and 5). Relative to these, mean BIC-allocated edfs were significantly lower (Table 3). When the mean edfs were not identical between GAM- and LMM-estimated curves, they were slightly higher in the former. However, GAM-derived curves had lower KL-distances. This difference was larger in the data generated with low error variance. GAMs had KL-distances ranging from 1 to 10 times smaller. KL-distance in the 10% tails at either end showed similar patterns. When the edfs were chosen to optimize AIC or GCV as opposed to BIC the KL-distance was smaller in nonlinear and discontinuous data. However, with linear data, the converse was seen and the models with edfs chosen to optimize BIC had smaller KL-distances.

For large datasets generated with a linear X-Y association with low variance the optimal edfs were near 1, no matter the criteria used. In the high variance data, this was only true when BIC was used.

The difference between GAMs and LMMs was more apparent with the smaller datasets (Tables 2, 4, and 6). When the data were generated to have both high and low error variance, any functional form between the independent and dependent variable, and edfs were chosen to optimize AIC, BIC, or GCV, fewer edfs were used by LMMs than GAMs. This trend was particularly strong when modeling data with lower error variance.

As when the sample size was large, in general more edfs were allocated with increasing levels of nonlinearity. However, when modeling the discontinuous functional form the edfs were only slightly higher compared to the smooth non linear functional form, and in some high variance cases the edfs were reduced. (See Tables 2,4, and 6.)

Again, as expected, BIC was optimized by fewer edfs in GAMs and LMMs than AIC or GCV when the sample size was small (Tables 2,4,and 6). AIC-optimal dfs were generally higher and GCV somewhere in between. GAM-derived curves had mostly lower KL-distances than LMM-derived curves when the sample size was lower, the only exception being the discontinuous functional form with high variance errors. KL-distance in the 10% tails at either end again showed similar patterns. Improvements in KL-distance were more modest in the smaller sample size data, rarely going beyond halving the distance (Tables 3 and 4).

Finally, for the smaller datasets with linear X-Y association, GAMs and LMMs consistently misallocated the edfs. The edfs were smaller when optimizing BIC, but the average edfs were closer to 2 in most cases for all criteria (Tables 3 and 4).

In smaller and larger datasets with higher error variance, LMM-estimated curves optimized the criteria with higher edfs than GAM-estimated curves. However the differences were marginal even considering the large datasets. In addition the variability in the edfs chosen also increased substantially.

COLD dataset analysis

Data from 514 COPD patients from the Montreal site of the COLD study were analyzed. These subjects were 58% female and 58% were smokers at some point in their life. The average amount of time spent as a smoker was 26 years. Both older (above 60) and female patients had lower spirometry values (FEV1, FVC, and PEF). (See Table 7).

When GAMs and LMMs were used to model the associations between pack-years and spirometry measures, GAM AIC-optimal curves used more edfs than LMM AIC-optimal curves and had correspondingly lower AIC. These curves also had lower AIC than the linear model. However, BIC-optimal curves had df near 1, suggesting little evidence for nonlinearity. (See Table 8). Moreover, the estimated AIC-optimal curves appeared overfit. (See Figures 2, 3 and 4.).

When GAMs and LMMs were used to model the associations between duration of smoking and spirometry measures, there was little evidence for nonlinearity. For PEF and FVC, the optimal df was near 1 for every criterion, and the linear model had the lowest criterion in nearly every case. For FEV1, the GAM and LMM curves were optimized by higher df, but the estimated curves were quite linear. (See Table 9 and Figures 5, 6 and 7.).

Discussion

The performance of both GAMs and LMMs with the flexibility of estimated curves selected via AIC, BIC, and GCV was compared in a variety of simulated and one real dataset.

In most instances, GAMs were optimized by a greater edf and had lower KL-distance (even when edfs were similar) than LMMs. When edfs were chosen to optimize BIC the difference in edfs were smaller, but when selected to optimize AIC or GCV there was up to 1 edf (especially in linear datasets and datasets with high variance and low sample size) at times. LMMs had lower KL-distances only when the data set was small, the X-Y association was nonlinear and the variance was high. Whereas the GAM-estimated curves were clearly over-fitting the linear X-Y associations, they may have been inferring the true extent of the non-linear curves, given that the KL-distance was often smaller. Aside from the linear case, it is difficult to estimate what should be the exact (or “true”) edfs allocated for a given curve. In fact the true curve may not even be relevent as small, non clinically significant, deviations can dramatically raise the edf.

LMMs and GAMs inherently differ in several ways. With LMMs the level of smoothing may be estimated directly from the data [25], though we did not explore that here. Knot selection may also directly impact the shape of curves estimated via LMMs. GAMs are more akin to numerical analysis techniques such as interpolation while LMMs rely on error reduction through a complex variance matrix .

Our results confirmed some well-known characteristics of information criteria. Lower edfs were allocated when BIC was used . Also, using either AIC and GCV on the simulated data resulted in models that were similar in terms of both KL-distance and edfs which has been shown to be true for large sample size linear mixed models. AIC was generally a poor performer in all measures with combined high variance and a low sample size datasets. In such scenarios, models which optimized the AIC allocated the highest edfs (including the linear data) and almost never claimed the lowest KL-distance in any category. The log likelihood component could be overcompensating for small sample size in the very nonlinear curves. Allocation based AICc instead of AIC may have lowered the edfs in these cases .

Previous simulations have shown that BIC that falsely selects spurious functions less frequently than AIC [74]. It is the more consistent of the two criteria, but often biases the result toward linearity. AIC and GCV are optimal for inferring the true extant of a highly non-linear curve. Indeed, when the X-Y association was linear, we found that BIC rarely selected for an edfs that resulted in a nonlinear curve. Conversely, when the X-Y association was nonlinear or discontinuous, the edfs chosen by BIC were greater than at least 1.5, and KL-distances were often similar to curves with df chosen via AIC for the larger datasets. Results were less clear in the smaller datasets where the BIC-curves allocated less df and had better KL-distances when the true curve was linear or smoothly nonlinear, but not when discontinuous. Thus, if a researcher was unsure of the linearity of a given covariate and wanted to be conservative, fitting a GAM or LMM with its df chosen to optimize BIC would minimize type II error and still have a strong chance of detecting nonlinearities.

An alternative method for selecting the edfs in mixed models was presented by [14]. In this work a test statistic as opposed to any mean squared error based criteria is suggested. However the test necessitates the declaration of null and alternative hypothesis edfs. These declarations may be subjective to the researcher's definition of an acceptable threshold of nonlinearity. Also, based on the paper's simulation, the test power (alpha = 0.05) only exceeds 80% if the difference between the null and alternative hypotheses is above four edfs .

In this work, we focused on proper model selection as opposed to inference on the nonlinearity of the estimated curve. While statistical significance may be an important tool in determining if a flexible regression approach is warranted, attempted inference after data-dependent selection of the smoothing parameter leads to an inflation of type I error [18]. Minimizing the number of models fit is thus desirable..

Another weakness of this work is that we generated data with only one continuous covariate. In computationally intensive scenarios or models with many possible nonlinear covariates selecting edfs by information criteria may be computationally intensive. Moreover, we focused only on GAMs and smoothing in LMMs, whereas many other flexible regression methods are available. (See the literature review p. 15) Finally, although in LMMs the amount of smoothing may be determined by the data, here we used AIC, BIC or GCV to enhance comparability with GAMs.

One strength of this work is the use of statistical simulation. Statistical simulation has its strength in its flexibility and ability to vary several features of the generated data. Any statistical simulation is the result of an artificial data generation process (DGP), driven by model choice and parameter settings, whose output is a synthetic sample. A simulation model is useful if it can be designed and calibrated so that, in terms of relevant criteria, the synthetic samples it produces approximate well the output of a real DGP.

GAM- and LMMs- estimated curves were often not similar in the COPD data, despite the large sample size. Optimal GAM and LMM estimated curves suggested that the effect of smoking duration on the three spirometric measures was linear or of non-relevant nonlinearity. With respect to pack years, the effect on FEV1 and PEF appeared to level off somewhat after 100, when the df were chosen to optimize AIC. The LMM modeled this more sharply than the GAM despite retaining the same numerical edf trends as in the simulation. A researcher may exercise caution before assuming that most nonlinear modeling methods will result in similar curves as sample size increases. However, overall we conclude that there is limited evidence of a nonlinear association between smoking and lung function in our cohort of COPD patients. The similar information criteria between the flexible models and the linear model, as well as the edfs which were mainly close to 1, especially for smoking duration certainly implied this.

In some instances, the pattern of edf allocation observed in the simulations was reversed with LMM-estimated curves optimizing the criteria with higher edfs than GAMs. Generally the difference was below one extra df but since the overall edfs were all well under 4 it made an impact on the resulting curves. It is possible that the nature of the nonlinearity in the COPD data differed from the smoother and more uniform nonlinearity generated in the simulated data. In addition the independent data was either highly skewed (Pack Years) or sparse (Smoking Duration) in the COPD dataset. The GAM's basis fitting as opposed to the LMM's knots are one of the possible differences that might have been responsible for this pattern of df-allocation .

In this work, we systematically compared LMMs and GAMs in a variety simulated, and one real-life dataset. Whereas, GAMs had the slight advantage over LMMs in large datasets, LMMs may be more reliable in small datasets with relatively large amounts of variability. In large datasets, edfs chosen by BIC with LMMs could be particularly useful for a conservative assessment of the nonlinearity of the curve. In the reverse scenario with a high sample size to variance ratio, GAMs or LMMs with AIC/GCV are most likely to capture the full extent of the nonlinear behavior.

Acknowledgements

COLD is part of the Canadian Cohort Obstructive Lung Disease (CanCOLD) funded by the Canadian Institute of Heath Research (CIHR/Rx&D Collaborative Research Program Operating Grants- 93326); industry partners Astra Zeneca Canada Ltd., Boehringer-Ingelheim Canada Ltd, GlaxoSmithKline Canada Ltd, Pfizer Canada Ltd; the Respiratory Health Network of the FRSQ and the Research Institute of the McGill University Health Centre.

Tables & Figures

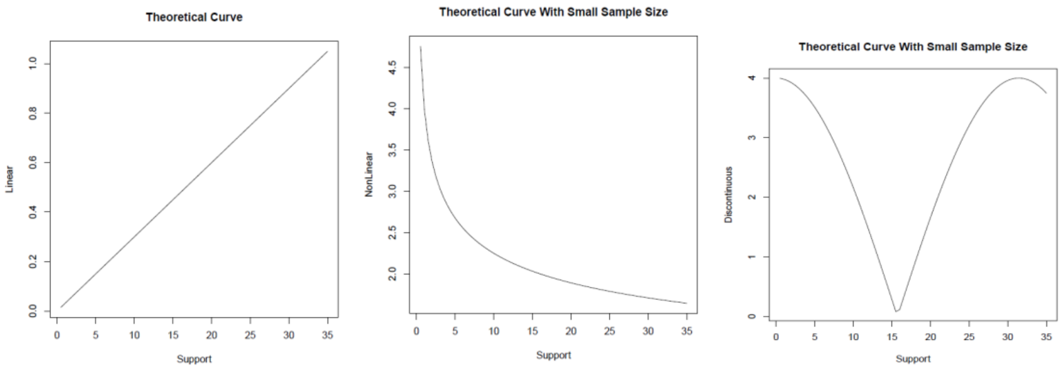


Figure 1: Shape of the X-Y association used for data generation. Three different theoretical curves (Linear, Non-Linear and Discontinuous) representing varying degrees of nonlinearity**.**

Table 1: Mean Kullback Liebler distance, and other information, for curves fit by Generalized additive models and generalized linear mixed models[1] with df chosen to minimize AIC for large datasets (n=400) with high and low variance and linear, nonlinear or discontinuous X-Y association[2].

X-Y Association		High Variance ($\sigma^2 = 2.0$)		Low Variance ($\sigma^2 = 0.7$)	
		GAM	LMM	GAM	LMM
Linear	Mean df (SD)	1.68 (1.30)	1.68 (1.30)	1.13 (0.81)	1.13 (0.81)
	Mean Criterion[3](SD)	561.58 (27.35)	561.58 (27.35)	-282.13 (28.49)	-282.12 (28.48)
	KL-total[4]	9.14E-04	1.88E-02	5.88E-03	1.00E-01
	First-10%[5]	1.39E-03	1.91E-02	9.64E-03	9.23E-02
	Last-10%[6]	1.60E-03	2.13E-02	1.09E-02	1.14E-01
Nonlinear	Mean Df (SD)	2.61 (2.10)	2.61 (2.10)	6.99 (2.42)	6.94 (2.4)
	Mean Criterion (SD)	565.15 (27.45)	565.15 (27.45)	-261.44 (28.89)	-261.45 (28.90)
	KL-total	2.64E-03	2.64E-03	5.19E-02	8.45E-02
	First-10%	4.74E-03	4.74E-03	8.51E-02	8.71E-02
	Last-10%	1.35E-02	1.35E-02	3.55E-01	1.46E-01
Discontinuous	Mean Df (SD)	5.83 (1.72)	5.83 (1.72)	9.21 (1.53)	9.22 (1.53)
	Mean Criterion (SD)	571.77 (27.79)	571.76 (27.79)	-261.46 (28.74)	-261.49 (28.74)
	KL-total	3.28E-03	1.63E-02	3.83E-02	7.51E-02
	First-10%	3.23E-03	1.67E-02	3.05E-02	7.10E-02
	Last-10%	4.05E-03	1.86E-02	3.93E-02	8.94E-02

[1] Degrees of freedom for the GAMs and LMMs were chosen from 0 to 20 by 0.25 increments to optimize the Akaike information criterion

[2] The X-Y association was for was linear y = 0.03x, nonlinear $y = {}^{4}/_{x^{0.25}}$, discontinuous $y = 4|\cos 0.02x|$ and scaled accordingly for the small samples.

[3] This is the mean minimized AIC score over all the possible degrees of freedom that could be allocated to the parameters

[4] Mean Kullback Liebler distance of the optimal df model across all simulated samples

[5] Mean Kullback Liebler distance the first 40 data points

[6] Mean Kullback Liebler distance the last 40 data points

Table 2: Mean Kullback Liebler distance, and other information, for curves fit by Generalized additive models and generalized linear mixed models with df chosen to minimize AIC for small datasets (n=40) with high and low variance and linear, nonlinear or discontinuous X-Y association.

X-Y Association		**High Variance (σ^2 = 2.0)**		**Low Variance (σ^2 = 0.7)**	
		GAM	**LMM**	**GAM**	**LMM**
Linear	Mean df (SD)	2.25 (2.10)	1.83 (0.75)	2.42 (2.13)	2.04 (0.96)
	Mean Criterion[1] (SD)	31.18 (7.30)	31.53 (6.91)	-11.06 (7.73)	-10.48 (7.39)
	KL-total[2]	9.17E-02	9.67E-02	2.71E-01	2.84E-01
	First-10%[3]	8.98E-02	9.88E-02	2.60E-01	3.01E-01
	Last-10%[4]	1.07E-01	1.20E-01	3.89E-01	4.54E-01
Nonlinear	Mean df (SD)	2.35 (1.99)	2.35 (1.99)	4.56 (2.53)	3.67 (1.85)
	Mean Criterion (SD)	33.53 (7.12)	33.53 (7.12)	-3.31 (7.83)	-1.73 (6.75)
	KL-total	1.13E-01	1.13E-01	9.76E-01	1.13E+00
	First-10%	2.02E-01	2.02E-01	3.83E+00	5.74E+00
	Last-10%	2.14E-01	2.14E-01	4.28E+00	6.00E+00
Discontinuous	Mean df (SD)	2.58 (2.14)	2.31 (1.19)	4.98 (2.31)	4.68 (1.97)
	Mean Criterion (SD)	34.50 (7.18)	34.85 (6.52)	-0.97 (7.89)	-0.73 (6.67)
	KL-total	1.54E-01	8.63E-02	8.08E-01	2.56E+00
	First-10%	1.26E-01	7.94E-02	4.96E-01	1.51E+00
	Last-10%	1.48E-01	1.04E-01	7.16E-01	2.14E+00

[1] This is the mean minimized AIC score over all the possible degrees of freedom that could be allocated to the parameters

[2] Mean Kullback Liebler distance of the optimal df model across all simulated samples

[3] Mean Kullback Liebler distance the first 4 data points

[4] Mean Kullback Liebler distance the last 4 data points

Table 3: Mean Kullback Liebler distance, and other information, for curves fit by Generalized additive models and generalized linear mixed models[1] with df chosen to minimize BIC for large datasets (n=400) with high and low variance and linear, nonlinear or discontinuous X-Y association[2].

		High Variance (σ^2 = 2.0)		**Low Variance (σ^2 = 0.7)**	
		GAM	**LMM**	**GAM**	**LMM**
Linear	Mean df (SD)	1.14 (0.36)	1.15 (0.37)	1.02 (0.05)	1.02 (0.05)
	Criterion[3](SD)	570.66 (27.54)	570.65 (27.54)	-274.03 (28.36)	-274.03 (28.36)
	KL-total[4]	7.52E-04	1.91E-02	5.52E-03	1.01E-01
	First-10%[5]	9.45E-04	1.95E-02	8.14E-03	9.27E-02
	Last-10%[6]	1.06E-03	2.16E-02	9.18E-03	1.14E-01
Nonlinear	Mean df (SD)	1.46 (0.57)	1.46 (0.57)	3.29 (1.15)	3.31 (1.14)
	Mean Criterion (SD)	576.29 (27.92)	576.29 (27.92)	-238.04 (29.35)	-238.03 (29.35)
	KL-total	2.94E-03	2.94E-03	7.88E-02	9.43E-02
	First-10%	5.43E-03	5.43E-03	1.40E-01	9.61E-02
	Last-10%	1.63E-02	1.63E-02	6.19E-01	1.52E-01
Discontinuous	Mean df (SD)	4.78 (0.64)	4.78 (0.64)	6.92 (1.27)	6.92 (1.27)
	Mean Criterion (SD)	596.37 (27.76)	596.36 (27.76)	-225.35 (28.36)	-225.41 (28.36)
	KL-total	3.82E-03	1.68E-02	5.75E-02	7.97E-02
	First-10%	3.02E-03	1.71E-02	3.23E-02	7.51E-02
	Last-10%	3.74E-03	1.91E-02	4.37E-02	9.41E-02

[1] Degrees of freedom for the GAMs and LMMs were chosen from 0 to 20 by 0.25 increments to optimize the Bayesian information criterion

[2] The X-Y association was for was linear y = 0.03x, nonlinear $y = {}^{4}/_{x^{0.25}}$, discontinuous $y = 4|\cos 0.02x|$ and scaled accordingly for the small samples.

[3] This is the mean minimized BIC score over all the possible degrees of freedom that could be allocated to the parameters

[4] Mean Kullback Liebler distance on these models

[5] Mean Kullback Liebler distance the first 40 data points

[6] Mean Kullback Liebler distance the last 40 data points

Table 4: Mean Kullback Liebler distance, and other information, for curves fit by Generalized additive models and generalized linear mixed models[1] with df chosen to minimize BIC for small datasets (n=40) with high and low variance and linear, nonlinear or discontinuous X-Y association[2].

		High Variance ($\sigma^2 = 2.0$)		**Low Variance ($\sigma^2 = 0.7$)**	
		GAM	**LMM**	**GAM**	**LMM**
Linear	Mean df (SD)	1.86 (1.58)	1.67 (0.61)	2.06 (1.86)	1.76 (0.77)
	Mean Criterion[3] (SD)	34.21 (6.92)	34.28 (6.84)	-7.85 (7.34)	-7.58 (7.38)
	KL-total[4]	6.59E-02	9.94E-02	2.41E-01	2.91E-01
	First-10%[5]	6.38E-02	1.02E-01	2.27E-01	3.09E-01
	Last-10%[6]	8.00E-02	1.23E-01	3.48E-01	4.63E-01
Nonlinear	Mean df (SD)	2.09 (1.77)	2.09 (1.77)	3.56 (2.23)	2.91 (1.22)
	Mean Criterion (SD)	36.71 (6.71)	36.71 (6.71)	1.67 (7.11)	2.45 (6.66)
	KL-total	1.07E-01	1.07E-01	1.14E+00	1.02E+00
	First-10%	2.02E-01	2.02E-01	5.09E+00	5.21E+00
	Last-10%	2.13E-01	2.13E-01	5.47E+00	5.49E+00
Discontinuous	Mean df (SD)	2.17 (1.83)	1.97 (0.88)	4.06 (1.86)	3.78 (1.14)
	Mean Criterion (SD)	37.83 (6.61)	37.96 (6.44)	4.36 (7.03)	4.42 (6.42)
	KL-total	1.57E-01	8.81E-02	9.15E-01	2.08E+00
	First-10%	1.28E-01	8.20E-02	4.71E-01	1.10E+00
	Last-10%	1.53E-01	1.06E-01	7.19E-01	1.64E+00

[1] Degrees of freedom for the GAMs and LMMs were chosen from 0 to 20 by 0.25 increments to optimize the Bayesian information criterion

[2] The X-Y association was for was linear y = 0.03x, nonlinear $y = 4/_{x^{0.25}}$, discontinuous $y = 4|\cos 0.02x|$ and scaled accordingly for the small samples.

[3] This is the mean minimized BIC score over all the possible degrees of freedom that could be allocated to the parameters

[4] Mean Kullback Liebler distance on these models

[5] Mean Kullback Liebler distance the first 40 data points

[6] Mean Kullback Liebler distance the last 40 data points

Table 5: Mean Kullback Liebler distance, and other information, for curves fit by Generalized additive models and generalized linear mixed models[1] with df chosen to minimize GCV for large datasets (n=400) with high and low variance and linear, nonlinear or discontinuous X-Y association[2].

		High Variance ($\sigma^2 = 2.0$)		Low Variance ($\sigma^2 = 0.7$)	
		GAM	**LMM**	**GAM**	**LMM**
Linear	Mean df (SD)	1.67 (1.30)	1.68 (1.30)	1.13 (0.81)	1.13 (0.81)
	Mean Criterion[3] (SD)	4.08 (0.28)	4.08 (0.28)	0.50 (0.04)	0.50 (0.04)
	KL-total[4]	9.14E-04	1.88E-02	5.88E-03	1.00E-01
	First-10%[5]	1.39E-03	1.91E-02	9.64E-03	9.23E-02
	Last-10%[6]	1.61E-03	2.13E-02	1.09E-02	1.14E-01
Nonlinear	Mean df (SD)	2.56 (1.97)	2.56 (1.97)	6.96 (2.42)	6.94 (2.41)
	Mean Criterion (SD)	4.12 (0.28)	4.12 (0.28)	0.52 (0.04)	0.52 (0.04)
	KL-total	2.61E-03	2.61E-03	5.20E-02	8.45E-02
	First-10%	4.68E-03	4.68E-03	8.52E-02	8.71E-02
	Last-10%	1.34E-02	1.34E-02	3.56E-01	1.46E-01
Discontinuous	Mean df (SD)	5.82 (1.71)	5.82 (1.71)	9.20 (1.52)	9.20 (1.52)
	Mean Criterion (SD)	4.19 (0.29)	4.19 (0.29)	0.52 (0.04)	0.52 (0.04)
	KL-total	3.28E-03	1.63E-02	3.83E-02	7.51E-02
	First-10%	3.23E-03	1.67E-02	3.05E-02	7.10E-02
	Last-10%	4.04E-03	1.86E-02	3.93E-02	8.94E-02

[1] Degrees of freedom for the GAMs and LMMs were chosen from 0 to 20 by 0.25 increments to optimize the Generalized Cross Validation

[2] The X-Y association was for was linear y = 0.03x, nonlinear $y = {}^{4}/_{x^{0.25}}$, discontinuous $y = 4|\cos 0.02x|$ and scaled accordingly for the small samples.

[3] This is the mean minimized GCV score over all the possible degrees of freedom that could be allocated to the parameters

[4] Mean Kullback Liebler distance on these models

[5] Mean Kullback Liebler distance the first 40 data points

[6] Mean Kullback Liebler distance the last 40 data points

Table 6: Mean Kullback Liebler distance, and other information, for curves fit by Generalized additive models and generalized linear mixed models[1] with df chosen to minimize GCV for small datasets (n=40) with high and low variance and linear, nonlinear or discontinuous X-Y association[2].

		High Variance ($\sigma^2 = 2.0$)		**Low Variance ($\sigma^2 = 0.7$)**	
		GAM	**LMM**	**GAM**	**LMM**
Linear	Mean df (SD)	1.81 (0.69)	1.78 90.64)	2.12 (1.62)	1.94 (0.87)
	Mean Criterion[3] (SD)	5.23 (1.71)	5.24 (1.71)	0.64 (0.22)	0.65 (0.23)
	KL-total[4]	5.01E-02	9.70E-02	2.38E-01	2.86E-01
	First-10%[5]	4.92E-02	9.91E-02	2.28E-01	3.04E-01
	Last-10%[6]	6.57E-02	1.21E-01	3.55E-01	4.57E-01
Nonlinear	Mean df (SD)	1.97 (0.99)	1.97 (0.99)	3.57 (1.89)	3.07 (1.26)
	Mean Criterion (SD)	5.91 (1.89)	5.91 (1.89)	1.00 (0.33)	1.03 (0.33)
	KL-total	8.42E-02	8.42E-02	1.06E+00	1.04E+00
	First-10%	1.79E-01	1.79E-01	4.85E+00	5.31E+00
	Last-10%	1.91E-01	1.91E-01	5.25E+00	5.59E+00
Discontinuous	Mean df (SD)	2.07 (0.92)	2.10 (0.95)	4.01 (1.56)	3.87 (1.12)
	Mean Criterion (SD)	6.21 (1.88)	6.20 (1.89)	1.13 (0.36)	1.11 (0.34)
	KL-total	1.21E-01	8.77E-02	8.73E-01	2.09E+00
	First-10%	9.34E-02	8.10E-02	4.52E-01	1.10E+00
	Last-10%	1.17E-01	1.05E-01	6.92E-01	1.63E+00

[1] Degrees of freedom for the GAMs and LMMs were chosen from 0 to 20 by 0.25 increments to optimize the Generalized Cross Validation

[2] The X-Y association was for was linear y = 0.03x, nonlinear $y = {}^{4}/_{x^{0.25}}$, discontinuous $y = 4|\cos 0.02x|$ and scaled accordingly for the small samples.

[3] This is the mean minimized GCV score over all the possible degrees of freedom that could be allocated to the parameters

[4] Mean Kullback Liebler distance on these models

[5] Mean Kullback Liebler distance the first 4 data points

[6] Mean Kullback Liebler distance the last 4 data points

Table 7: Basic characteristics of subjects (n=541) from the COLD study

Variable	Mean (SD) or Proportion
Male	41.21%
Age	54.40 (10.20)
FEV1	2.71 (0.80)
FVC	3.72 (1.03)
PEF	7.17 (2.15)
Ever smoked	58.15%
Years of smoking	26.50 (12.80)
Pack years	14.33 (20.40)
Weight	74.70 (17.30)

Table 8: Measures of fit and degrees of freedom used to model the association between pack-years[1] and lung function[2]

		GAM[3]	GLMM[4]	Linear[5]
FEV1	AIC	-655.47	-648.78	-646.47
	df	4.45	3.35	1.00
	BIC	-638.47	-614.21	-638.20
	df	1.00	2.55	1.00
	GCV	0.30	0.30	0.30
	df	4.45	3.35	1.00
FVC	AIC	-375.92	-366.87	-372.02
	df	4.50	2.95	1.00
	BIC	-363.65	-338.24	-363.44
	df	1.00	1.02	1.00
	GCV	0.50	0.51	0.50
	df	4.50	2.95	1.00
PEF	AIC	478.65	485.56	486.07
	df	4.25	3.35	1.00
	BIC	494.35	519.63	494.65
	df	1.00	2.55	1.00
	GCV	2.43	2.46	2.46
	df	4.20	3.35	1.00

[1] The effect of pack years was smoothed, with df chosen to optimize AIC, BIC or GCV.

[2] Separate models were fit using FEV1, FVC, or PEF as the response. Age, sex and weight were included as confounders in the model.

[3] Results from the generalized additive model: The optimal (lowest AIC, BIC, or GCV) df assigned to Pack Years

[4] Results from the generalized linear mixed model: The optimal (lowest AIC, BIC, or GCV) df assigned to Pack Years

[5] Results from a linear model

Table 9: Measures of fit and degrees of freedom used to model the association between smoking duration[1] and lung function[2]

		GAM[3]	GLMM[4]	Linear[5]
FEV1	**AIC**	-382.58	-374.50	-380.48
	df	2.80	2.55	1.00
	BIC	-372.98	-349.98	-372.98
	df	1.00	1.00	1.00
	GCV	0.30	0.30	0.30
	df	2.80	2.55	1.00
FVC	**AIC**	-229.65	-221.65	-229.65
	df	1.00	1.00	1.00
	BIC	-222.16	-199.15	-222.16
	df	1.00	1.00	1.00
	GCV	0.48	0.49	0.48
	df	1.00	1.00	1.00
PEF	**AIC**	280.42	288.40	280.51
	df	1.35	1.35	1.00
	BIC	288.01	311.01	288.01
	df	1.00	1.00	1.00
	GCV	2.44	2.51	2.44
	df	1.30	1.35	1.00

[1] The effect of duration was smoothed, with df chosen to optimize AIC, BIC or GCV.

[2] Separate models were fit using FEV1, FVC, or PEF as the response. Age, sex and weight were included as confounders in the model.

[3] Results from the generalized additive model: The optimal (lowest AIC, BIC, or GCV) df assigned to Smoking Duration

[4] Results from the generalized linear mixed model: The optimal (lowest AIC, BIC, or GCV) df assigned to Smoking Duration

[5] Results from a linear model

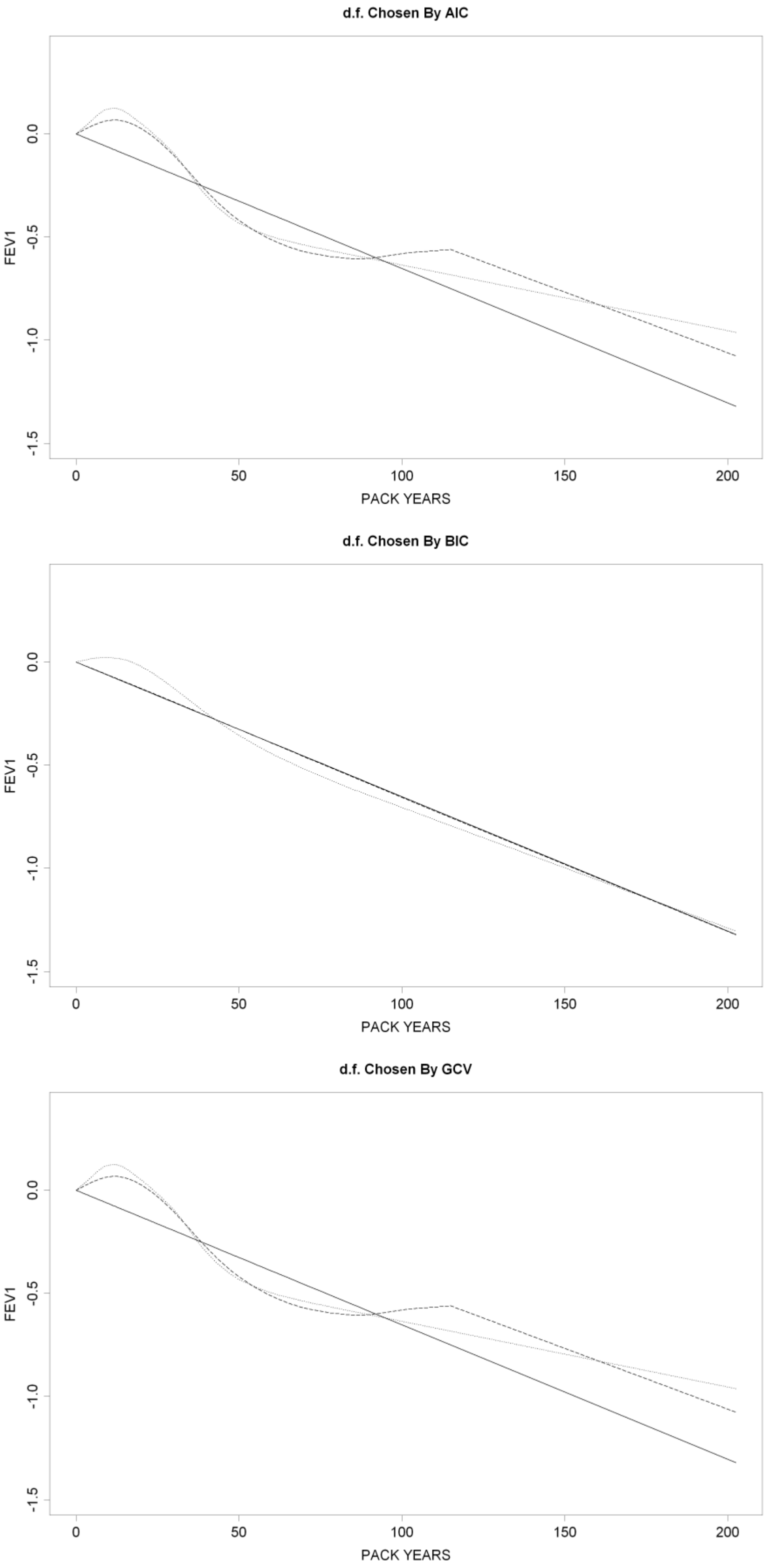


Figure 2: Association between Pack Years and FEV1 estimated via a linear model (solid line), a GAM (dotted line) and a LMM (dashed line) with df chosen to optimize AIC (upper panel), BIC (middle panel) or GCV (lower panel)

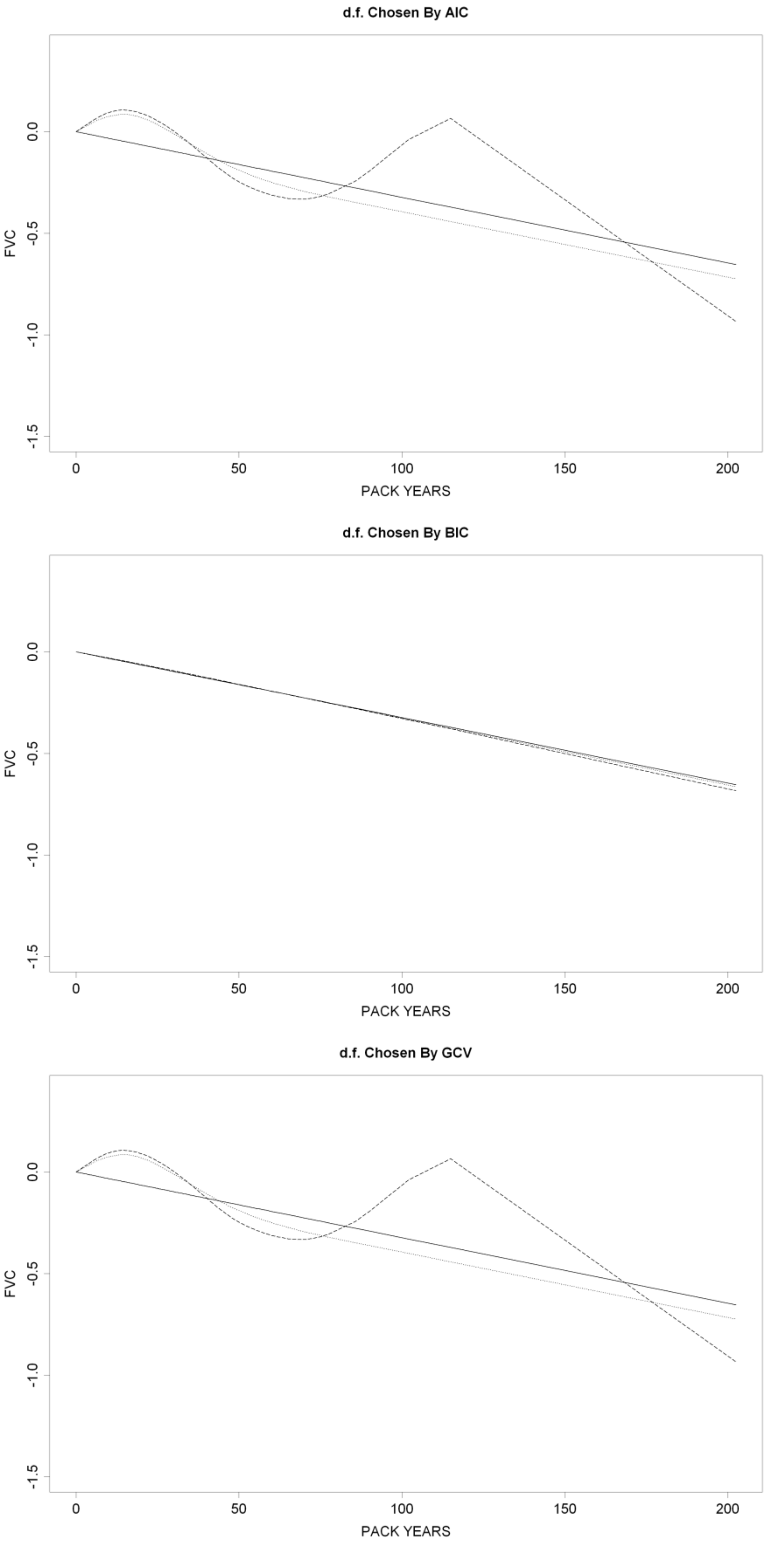


Figure 3: Association between Pack Years and FVC estimated via a linear model (solid line), a GAM (dotted line) and a LMM (dashed line) with df chosen to optimize AIC (upper panel), BIC (middle panel) or GCV (lower panel)

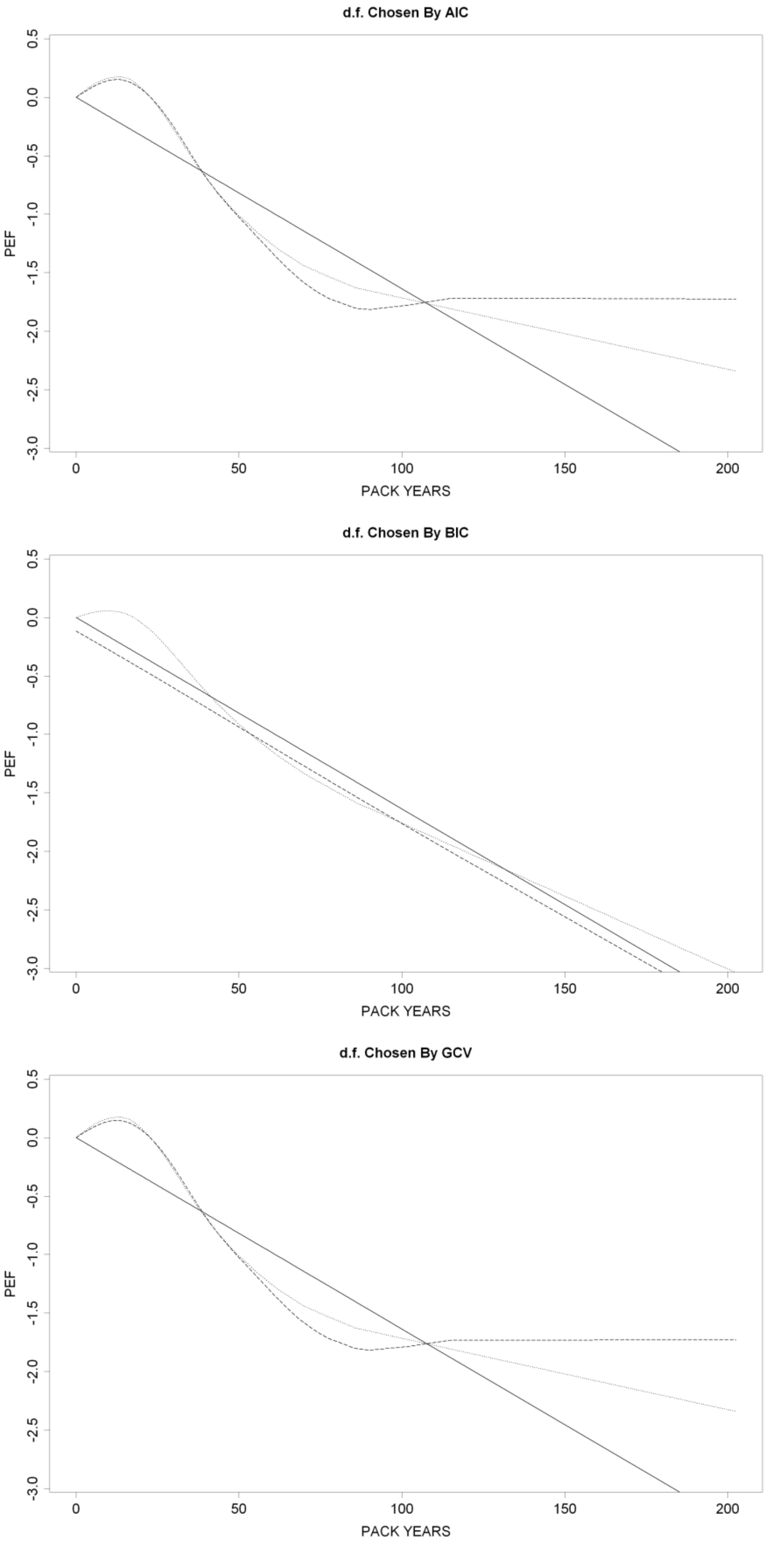


Figure 4: Association between Pack Years and PEF estimated via a linear model (solid line), a GAM (dotted line) and a LMM (dashed line) with df chosen to optimize AIC (upper panel), BIC (middle panel) or GCV (lower panel)

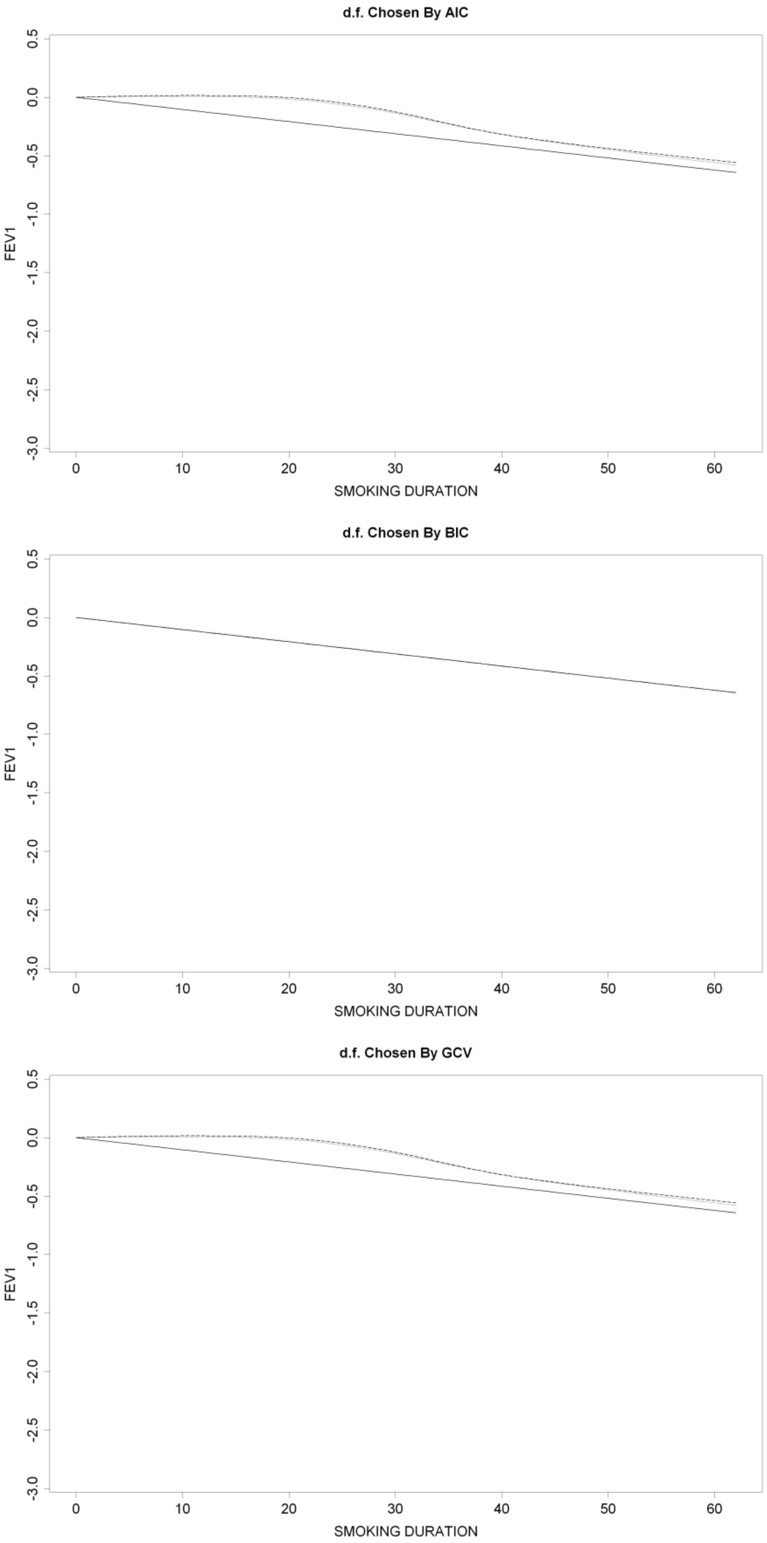
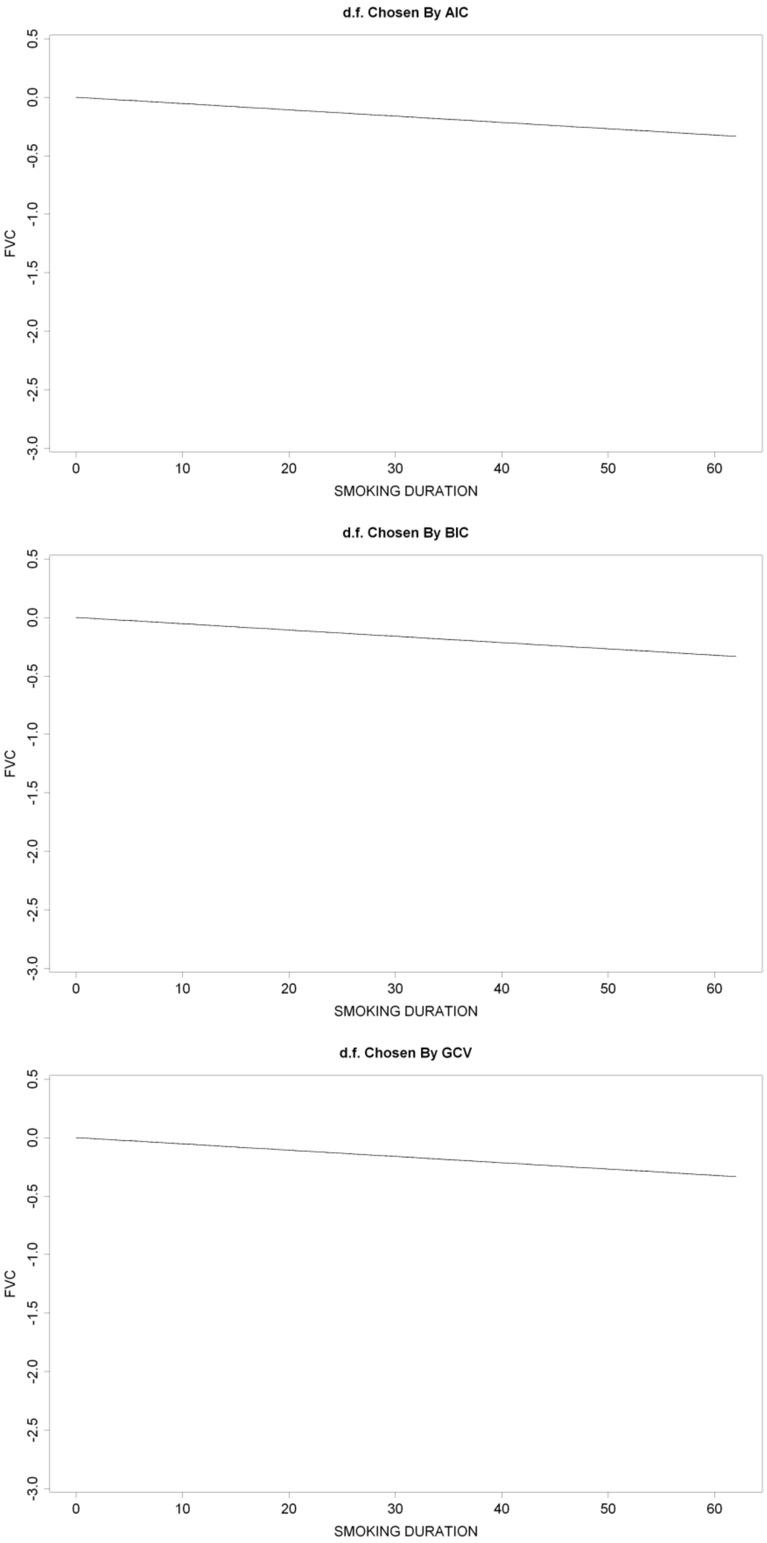


Figure 5: Association between Smoking Duration and FEV1 estimated via a linear model (solid line), a GAM (dotted line) and a LMM (dashed line) with df chosen to optimize AIC (upper panel), BIC (middle panel) or GCV (lower panel)


d.f. Chosen By AIC

d.f. Chosen By BIC

d.f. Chosen By GCV

Figure 6: Association between Smoking Duration and FVC estimated via a linear model (solid line), a GAM (dotted line) and a LMM (dashed line) with df chosen to optimize AIC (upper panel), BIC (middle panel) or GCV (lower panel)

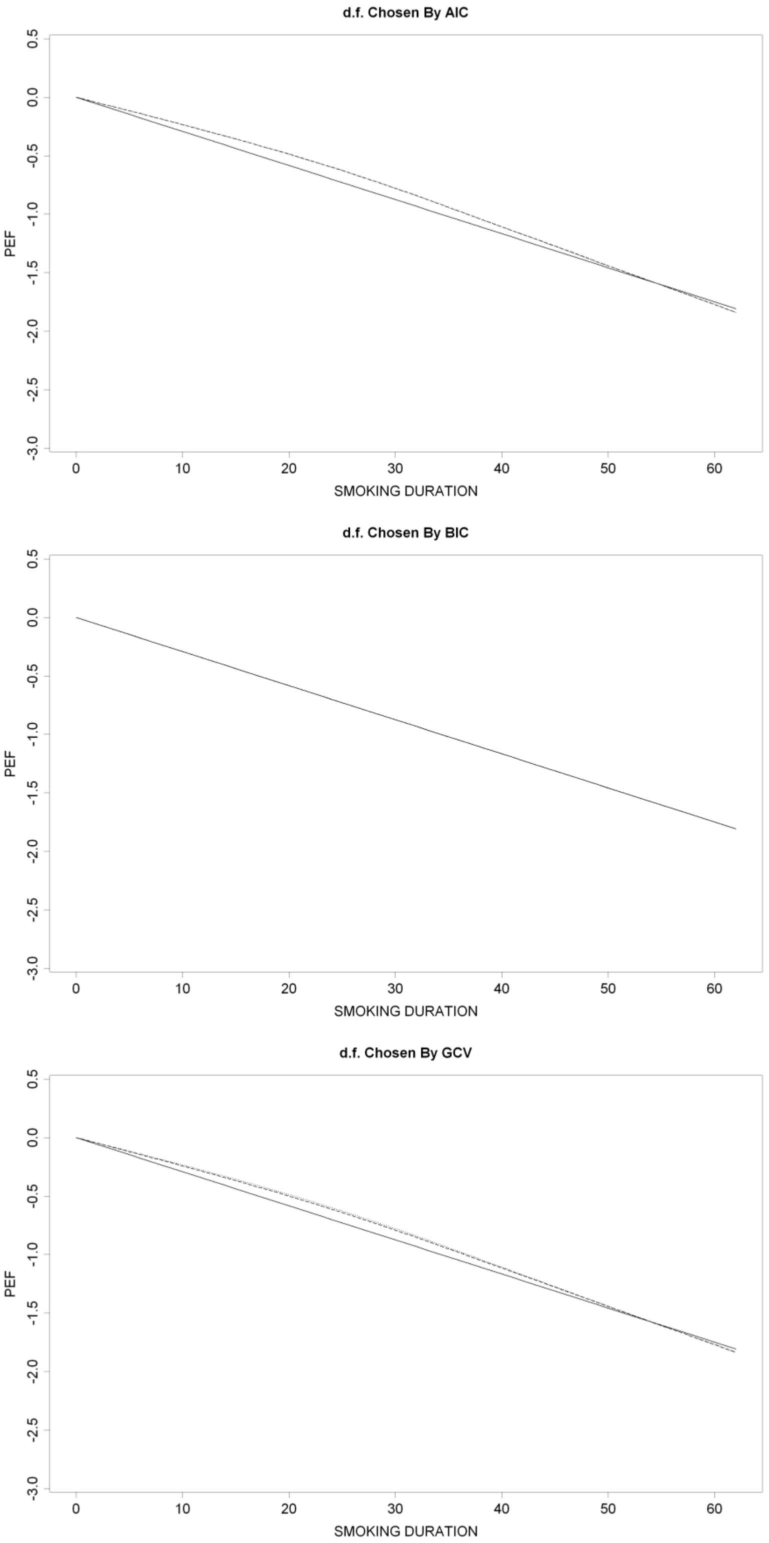


Figure 7: Association between Smoking Duration and PEF estimated via a linear model (solid line), a GAM (dotted line) and a LMM (dashed line) with df chosen to optimize AIC (upper panel), BIC (middle panel) or GCV (lower panel)

Linking Statement

The first manuscript compared GAMs and LMMs via simulation study and in one real life example for continuous outcomes. The second manuscript focuses on modeling known non-linear associations between continuous independent variables and a binary outcome variable, investigating a previously established trend. A comparison is still made between GAMs and GLMMs. However, since the study is focused on real life data it lacks a simulation. Thus the true accuracy of GAMs vs GLMMs cannot be known and so this investigation may not be as a thorough as in the first manuscript. The advantage of this manuscript lies in the fact that the nonlinear characteristics of the data are strong. These relationships are of high clinical importance. Interpretation of their nonlinear curves is clinically relevant.

Full title: Modeling Nonlinear trends in ICU patients: A Comparison of Generalized Additive Models and Generalized Linear Mixed Models

Short title: Modeling Nonlinear trends in ICU patients

Daniel Hercz[1]
Sandra Dial

Andrea Benedetti[1,2]

[1]Department of Epidemiology, Biostatistics and Occupational Health, McGill University
[2]Respiratory Epidemiology and Clinical Research Unit, Department of Medicine, McGill University

Corresponding Author: Andrea Benedetti

Address:
Montreal Chest Institute, K-135
3650 St. Urbain
Montreal, QC
H4A 2Z6

Telephone: (514) 934-1934 ext. 32161
Fax: (514) 843-2083
Email: Andrea.benedetti@mcgill.ca

Abstract

Patients with sepsis have an increased risk of mortality when several key vital parameters fall below certain threshold levels. This research focuses on whether and where this occurs with patients receiving critical care post cardiac surgery. Generalized Additive Models (GAMs) and Generalized Linear Mixed Models (GLMMs) were used to model the associations between Mortality vs Cardiac Index, Mixed Venous Oxygen Saturation, Oxygen Delivery, and Serum Lactate. Flexibility of the smooth curves was determined by data driven criteria (AIC, BIC or GCV). All variables except Serum Lactate exhibited a nonlinear relationship with Mortality. GAMs had a tendency to assign stronger nonlinearity to models than GLMMs. GLMMs were more likely to estimate curves suggestive of the existence of a threshold while at the same time not overfitting.

Background

In the intensive care unit, certain postoperative physiological variables and hemodynamic parameters, such as oxygen delivery, cardiac index, mixed venous oxygen saturation and serum lactate are monitored. These measurements are variable and may change in the hours after surgery while in the intensive care unit (ICU). It is suggested that the inability to achieve or reach certain levels, may predict an increased risk of mortality. [75]

The Cardiac Index (CI) is the volume of blood being pumped by the heart, in particular by a left or right ventricle over one minute, scaled by the patient's body size. In a healthy patient the CI is demand based; such that it can vary depending on the metabolic and oxygen requirements of the subject. The normal range of CI is 2.6 - 4.2 L/min/m^2[50].

The tissues and organs of the body require oxygen and other nutrients, in particular glucose, to meet their metabolic needs. The delivery of oxygen to the tissues (DO_2) is determined by the cardiac output, and the oxygen content of the blood. The oxygen content of the blood in turn is determined by the hemoglobin level and the oxygen saturation of the hemoglobin after it has been oxygenated in the lungs (arterial oxygen saturation). The tissues then extract oxygen from the hemoglobin for metabolism (VO_2). Under normal conditions the DO_2 is more than sufficient to meet the demands of all tissues and organs. Even with a moderate reduction in DO_2, the tissues may be able to obtain sufficient oxygen by extracting more oxygen from the hemoglobin. The O_2 uptake is denoted by VO_2 on Figure 1. The level of DO_2 at which the VO_2 is affected is known as the critical DO_2. When DO_2 drops below the critical DO_2, $\dot{V}O_2$ becomes supply dependent [53]. The tissues, because of insufficient oxygen delivery, may convert from a normal aerobic metabolism, to an anaerobic type of metabolism to try to meet their energy requirements. When this occurs lactic acid may be produced [53], leading to an increase in serum lactate.

The Mixed Venous Oxygen Saturation (MVO_2) refers to the oxygen saturation of blood returning to the heart after the tissues have extracted oxygen required for their metabolic needs. MVO_2 depends on the DO_2 and the VO_2. The VO_2 in turn is dependent on the tissue oxygen demands. As shown in Figure 1, even in the presence of increased DO_2, the oxygen uptake remains constant and does not desaturate blood hemoglobin more than necessary [51]. If there is a significant drop in saturation of the blood, this implies the body is drawing an additional percentage of O_2 from the blood. Such a drop usually implies the possibility of an inadequate DO_2, which could be because of the hemoglobin being too low (anemia), decreased arterial oxygen saturation, and/or decreased cardiac output; however, a normal or high value does not exclude such disturbances. When employed in conjunction with the other indicators of tissue oxygenation available in an intensive care unit, MVO_2 can be useful as a guide for both prognosis and urgency of therapy [76].

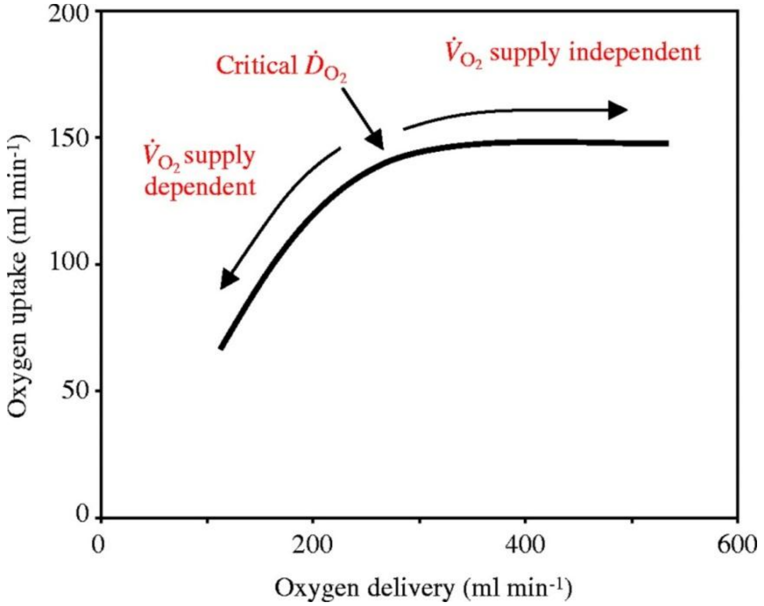


Figure 1: Oxygen Delivery per DO_2

Serum lactate is involved in the conversion of pyruvate and lactate when tissue O_2 levels are low and implies the occurrence of anaerobic metabolism, which can is considered a marker of insufficient DO_2. The normal blood lactate concentration in unstressed patients is 0.5-1 mmol/L [52].

Because the associations between these measures and mortality are often nonlinear [77, 78], methods that impose a linear functional form will produce biased effect estimates.

Categorization of the independent variable could be used, but this may result in an untrue [19] functional form and can reduce power to detect an effect. *A priori* specification of the functional form (e.g. quadratic) often requires information that is not available, and as with assuming linearity, can result in a biased estimate if the shape is wrong. A better option is to use a flexible regression method in which the shape of the association is estimated directly from the data. There are many possible approaches – here we consider two. [77, 78]

Generalized Linear Mixed Models (GLMMs) are extensions to Generalized Linear Models through the addition of "random" effect term[17]. Typically, GLMMs are used to model longitudinal data, but mixed models have a much wider generality than those used for handling correlation or clustering. By treating the terms from a regression spline as random effects a smooth curve can be estimated. Both the shape and flexibility of this curve can be estimated directly from the data.[2].

Generalized additive models (GAMS) generalize the regression equation to include smooth functions of some independent variables, conditional on the user-specified degrees of freedom (df) that control the flexibility of the curve. The smooth functions have shapes that are estimated directly from the data. Several smoothing strategies are available in GAMs – here we use smoothing splines. [79]

The Generalized Additive Model (GAM) generalizes the conventional regression model to:

$$\mathbf{Y} = \mathbf{Logit}\ (\mathrm{y})\ \mathbf{where}$$

$$y = \beta_{\mathbf{0}} + \sum_{X\in A} \beta_i X_i + \sum_{X\in B} S_j(X_j) + \varepsilon$$

where $S_j(X_j)$ are smooth functions whose shapes are estimated directly from the data for those predictors ($X \in B$) modeled non-parametrically, and $\varepsilon \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\sigma}^2)$. A number of smoothing methodologies are available in GAMs, though here we focus on cubic smoothing splines [79].

The GLMM regression equation for a binomial outcome is:

$$\mathbf{Y} = \mathbf{Logit}\ (\mathbf{y})\ \mathbf{where}$$

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \cdots \beta_p \mathbf{x}^p + \sum_{k=1}^{K} \mu_k (\mathbf{x} - \kappa_k)_+^p + \varepsilon_i \quad ;$$

$$\text{Where} \quad \begin{aligned} (x-\kappa_{k=})_+^p &= 0 \text{ if } \mathrm{x} \le \kappa_k \\ &= (x-\kappa_{k=})_+^p \text{ if } \mathrm{x} > \kappa_k \end{aligned} \quad ;$$

and β_i represent the fixed effects; the μ_k represent the normally distributed random effects; the κ_k are the pre-specified knots, the ε_i are normally distributed errors , and p is the degree of the polynomial. The variance of the random effects, is usually estimated by maximum likelihood when the outcome is continuous, though can also be arrived at by minimizing the penalized least squares [17]. The smoothing parameter, λ, is the ratio of the variance to the two normal distributions above. It controls the level of smoothing in GLMMs. [2]

When the outcome is binary, estimation involves an intractable likelihood. Two popular approaches are penalized quasi likelihood and numerical integration via adaptive Gaussian hermite quadrature [17]. Here we focus on PQL.

Several “threshold” relationships are known to exist between the vital indicators of ER patients with sepsis and mortality however, their exact values are either not well established or subject to debate [53, 64]. Moreover, it is not clear whether the associations between these parameters and mortality can be applied to other patient populations such as the one used here. The main objective of this paper is to model the associations between vital indicators and mortality using GAMs and GLMMs in a population of subjects receiving critical care post cardiac surgery.

Methods

Data was collected at two adult tertiary care university affiliated hospitals in Montreal, Canada, retrospectively between January 1, 2005 and December 31, 2005 by trained reviewers using standardized data collection sheets from patient charts. Consecutive patients who had a coronary artery bypass (CABG), valve replacement or repair, or combined CABG and valvular aortic procedures, were included in the study. Patients undergoing a heart transplant, pulmonary thromboendarterectomy, or placement of a ventricular assist device were excluded.

All patients were admitted postoperatively to the intensive care unit (ICU). A Swan-Ganz catheter was used perioperatively at both hospitals to guide patient resuscitation. Serum lactate and mixed venous oxygen saturation levels were measured in all patients at one site, and selected patients at the other site. In order to avoid bias, only the data from the patients treated at the hospital with routinely measured serum lactate and mixed venous oxygen saturation were used for the analysis of those variables.

The postoperative physiological variables (delivered oxygen, cardiac index, mixed venous oxygen saturation and serum lactate) were measured at three time points: at admission to the ICU; 6 and 24 hours post ICU admission.

The primary study outcome was hospital mortality. Data were collected on patient age, sex, Parsonnet's score [80], past medical history, procedure related variables, and six-hour postoperative physiological variables. Past medical history was abstracted from patient records. Conditions considered were any prior cardiac surgery, hypertension, diabetes, atrial fibrillation, preoperative hospitalizations for heart failure, preoperative renal dysfunction, preoperative dialysis, preoperative ejection fraction and left ventricular dysfunction.

All research was in keeping with the principles outlined in the Helsinki declaration. The research ethics committee of the McGill University Health Centre Research Institute approved the study. The ethics committee waived the Need for informed consent as the data were collected retrospectively.

Statistical analysis

A generalized linear mixed model (GLMMs), a generalized additive models (GAMs) and a linear logistic model were fit to the ER patient data at each time point and for each independent variable, with mortality as the binary outcome and adjusted for important confounders (age, sex and Parsonnet’s score) as linear effects.

Knots for the GLMMs were uniformly distributed proportionally to the density of the independent data as per Wand 2003 [2], and. estimated via penalized quasi likelihood using the algorithm provided by Wand [10]. GAMs were fitted using the “gam” library of the R statistical software package [81].

The complexity of a linear model is proportional to the number of explanatory terms; every parameter corresponds to one degree of freedom (df) due to the assumption that each has only a linear relation with the data. Flexible regression models estimate the shape of the association under study directly from the data. The smoothness of the estimated curve depends on the df – with more df resulting in a bumpier curve that follows the data more closely. The total number of equivalent degrees of freedom used by the model is known as the effective degree of freedom or edf [13]. For the purposes of this study df will be used interchangeably with edf.

Both GLMMs and GAMs can fit their nonlinear aspects as proportional to the df [1, 37]. The calculation is based on the resultant curve. In GLMMs, it is functionally connected to the ratio of variability in the random vs. the fixed effects. In GAMs (which have a broadest range of methods to fit their curves), the connection is less strict. This is convenient as all the nonlinearity of a model is expressed as a one dimensional value, the edf. Through stepwise specification of the df (from 1 to 20 with increments of 0.25 for the purposes of this paper), one can achieve a gradient of models and select the one with the best fit. Three measures of fit were used: the Akaike Information Criterion (AIC), Baysian Information Criterion (BIC), and Generalized Cross Validation (GCV) were used to select df for both GAMs and GLMMs. The minimized criterion score is reported. [31, 35, 66]

Results

Table 1 Basic characteristics of subjects (n=520) from the ICU study

Variable	**Mean (SD) or Proportion in Subjects who Survived (n=452)**	**subjects who died (n=68)**
Male	62.31%	54.41%
Age	68.4 (10.60)	74.8 (8.19)
Parsonnet’s Score	17.38 (11.10)	29.44 (14.0)

Table 2: Results for modeling the association between cardiac index at three time points and mortality via GAMs, GLMMs and a linear model [1]

Cardiac Index				
		GAM[2]	GLMM[3]	Linear[4]
At admission	AIC	-1218.27	-1196.24	-1192.25
	df	5.10	5.90	1.00
	BIC	-1199.12	-1164.24	-1183.76
	df	1.00	1.35	1.00
	GCV	9.39E-02	9.82E-02	9.88E-02
	df	5.80	5.40	1.00
At 6 hours	AIC	-1219.62	-1197.57	-1182.25
	df	4.30	3.60	1.00
	BIC	-1197.33	-1166.72	-1173.77
	df	2.20	3.60	1.00
	GCV	9.37E-02	9.82E-02	1.01E-01
	df	8.25	6.55	1.00
At 24 hours	AIC	-1208.82	-1190.27	-1177.93
	df	3.30	3.10	1.00
	BIC	-1197.71	-1163.94	-1169.46
	df	1.40	1.40	1.00
	GCV	9.35E-02	9.76E-02	9.93E-02
	df	3.30	3.10	1.00

[1] Age, sex and Parsonnet’s score as confounders were included in the model as linear effects.

[2] Results from the generalized additive model: The optimal (lowest AIC, BIC, or GCV) df assigned to Cardiac Index

[3] Results from the generalized linear mixed model: The optimal (lowest AIC, BIC, or GCV) df assigned to Cardiac Index

[4] Results from a linear model

Table 3
Results for modeling the association between MVO_2 at three time points and mortality via GAMs, GLMMs and a linear model [1]

MVO_2				
		GAM[2]	GLMM[3]	Linear[4]
At admission	AIC	-917.46	-904.35	-904.46
	df	3.10	3.95	1.00
	BIC	-904.38	-880.03	-896.54
	df	1.90	1.00	1.00
	GCV	9.29E-02	9.63E-02	9.60E-02
	df	3.05	3.00	1.00
At 6 hours	AIC	-951.12	-936.77	-937.29
	df	4.15	3.25	1.00
	BIC	-939.10	-910.94	-929.29
	df	1.00	1.00	1.00
	GCV	9.55E-02	9.94E-02	9.88E-02
	df	4.10	3.10	1.00
At 24 hours	AIC	-1006.12	-992.40	-987.97
	df	3.35	3.85	1.00
	BIC	-994.43	-966.73	-979.83
	df	1.80	1.90	1.00
	GCV	9.74E-02	1.01E-01	1.02E-01
	df	3.80	4.10	1.00

[1] Age, sex and Parsonnet's score as confounders were included in the model.
[2] Results from the generalized additive model: The optimal (lowest AIC, BIC, or GCV) df assigned to MVO_2
[3] Results from the generalized linear mixed model: The optimal (lowest AIC, BIC, or GCV) df assigned MVO_2
[4] Results from a linear model

Table 4:
Results for modeling the association between DO$_2$ at three time points and mortality via GAMs, GLMMs and a linear model [1]

DO$_2$				
		GAM[2]	GLMM[3]	Linear[4]
At admission	AIC	-1146.52	-1129.45	-1131.77
	df	2.25	1.15	1.00
	BIC	-1138.11	-1104.02	-1123.36
	df	1.00	1.00	1.00
	GCV	9.82E-02	1.02E-01	1.01E-01
	df	1.15	1.15	1.00
At 6 hours	AIC	-1072.58	-1048.79	-1032.71
	df	7.20	4.10	1.00
	BIC	-1044.92	-1019.92	-1024.46
	df	1.60	1.45	1.00
	GCV	9.57E-02	1.01E-01	1.04E-01
	df	10.00	3.85	1.00
At 24 hours	AIC	-1154.53	-1140.32	-1111.08
	df	4.9	5.15	1.00
	BIC	-1131.46	-1110.14	-1102.70
	df	3.05	1.55	1.00
	GCV	9.39E-02	9.72E-02	1.03E-01
	df	7.10	6.85	1.00

[1] Age, sex and Parsonnet's score as confounders were included in the model.

[2] Results from the generalized additive model: The optimal (lowest AIC, BIC, or GCV) df assigned to DO$_2$

[3] Results from the generalized linear mixed model: The optimal (lowest AIC, BIC, or GCV) df assigned to DO$_2$

[4] Results from a linear model

Table 5:
Results for modeling the association between Serum Lactate at three time points and mortality via GAMs, GLMMs and a linear model [1]

Lactate				
		GAM[2]	GLMM[3]	Linear[4]
At admission	AIC	-908.03	-883.39	-885.53
	df	1.00	1.00	1.00
	BIC	-882.10	-857.57	-877.79
	df	1.00	1.00	1.00
	GCV	7.75E-02	8.41E-02	8.25E-02
	df	1.10	1.35	1.00
At 6 hours	AIC	-893.43	-881.12	-876.24
	df	1.30	1.30	1.00
	BIC	-871.98	-857.48	-868.51
	df	1.20	1.20	1.00
	GCV	7.97E-02	8.28E-02	8.36E-02
	df	1.00	1.50	1.00
At 24 hours	AIC	-917.69	-897.89	-894.41
	df	1.40	2.20	1.00
	BIC	-904.75	-874.68	-886.68
	df	1.05	1.05	1.00
	GCV	7.43E-02	7.92E-02	7.94E-02
	df	5.00	4.80	1.00

[1] Age, sex and Parsonnet's score as confounders were included in the model.

[2] Results from the generalized additive model: The optimal (lowest AIC, BIC, or GCV) df assigned to Serum Lactate

[3] Results from the generalized linear mixed model: The optimal (lowest AIC, BIC, or GCV) df assigned to Serum Lactate

[4] Results from a linear model

Information was collected on 520 post-surgical ICU subjects, 68 who did not survive. Basic characteristics of the included subjects are presented in Table 1. 62.3% of the surviving group was male (54.41% in nonsurviving), with a mean age of 68.4 years (74.8 years in nonsurviving) and mean parsonnet score of 17.38 (29.44 in nonsurviving). The total incidence of mortality was 13%.

Both GAMs and GLMMs captured consistent and similar directional trends within each ER variable modeled (all figures). Mortality decreased as Cardiac Index, MVO_2, and DO_2 increased (Figures 1, 2, and 3). Mortality increased as serum lactate (Figure 4) increased. In addition, there appeared to be threshold values for both Cardiac Index and MVO_2 (Figures 1-18 and 19-36). Above 2-2.5 for Cardiac Index and slightly above 40% for MVO_2 the reduction in mortality rate leveled off (Figures 1-9 and 10-15). DO_2 showed a similar threshold at 250, but only after initial admission (Figure 22-27). There was some nonlinearity beyond this point, but for most models it appeared localized and idiosyncratic. This was usually in the form of a periodic function having little or no trend.

GAMs with higher (Tables 1 and 3: Cardiac Index and DO_2) or the same edf (Tables 2 and 4: MVO_2 and DO_2) fit better than GLMMs. Most differences were subtle, but in some variables such as serum lactate (a theoretically roughly linear relationship), GAMs yielded nonlinear curves. In some curves there were also possible signs of over fitting when using GAMs (small and idiosyncratic deviations from the curve visualized as "bumpiness") with all information criteria (Figures 4-6 and 10-12).

Selection of edf through optimization of AIC allocated more edf than through optimization of BIC. In fact, models with edf chosen to optimize BIC were mostly linear or close to it (<2 edf). This was similar for GLMMs and GAMs. Instead of estimating similarly to AIC as predicted for large datasets, in some cases GCV allocated higher edf than AIC or BIC. In some cases the extra edf seemed arbitrary in relation to those selected by the other criteria, resulting in it almost being the lone nonlinear model (e.g. The Serum Lactate models Tables 28-36).

In addition to GCV allocating the highest edf over the other information criteria, the estimated curves appear to be overly fitting noise with particularly "bumpy" results. (See Figures 47-48, 53-54, and 71-72).

Discussion

In this work, we used GAMs and GLMMs, as well as a linear logistic regression to model the associations between Cardiac Index, Mixed Venous Oxygen Saturation, Oxygen Delivery, and Serum Lactate and mortality. Edf for GAMs and GLMMs were chosen to optimize one of three criterion: AIC, BIC or GCV.

Our results suggested that selection with BIC likely failed to capture the extent of nonlinearity present in Cardiac Index, DO_2, and MVO_2. Used with both GAMs and GLMMs, BIC was repeatedly minimized with nearly linear curves (i.e. with df near 1). GCV presented the opposite problem and was often optimized with unrealistically high df. While it is not possible to know the true edf of a given model, the GCV-selected df resulted in implausibly bumpy curves. AIC picked up on the nonlinear data more than BIC and produced less extreme edfs than GCV.

Both GAM- and GLMM-estimated curves suggested the possible existence of a threshold level of vital indicators, such that values lower than these thresholds were associated with a higher rate of mortality. Visual estimates of the threshold levels, determined from the curves, were similar to those reported in septic patients. GLMM curves produced sharper curves in the regions where a threshold may exist.

Our results show some notable differences with previously published information. While levels of MVO_2 of 60% or lower are considered abnormal or dangerous [76], we found that MVO_2 had little impact on mortality until almost 40%. In one study of 17 critically ill septic and nonseptic patients, the mean critical DO_2 was approximately 300 mL/min for a 75-kg patient or 60% [59]. We found that DO_2 levels would often be much lower at time of admission and shortly after (closer to 40-50%).

On the other hand, our results were similar for other exposure variables. CI values below 1.8 L/min/m² potentially indicate that the patient is in cardiogenic shock [50]. Our results are similar, suggesting that this is roughly the case with CI's contribution to mortality starting below 2 L/min/m².

Elevated initial through 24-hour lactate levels, and lactate clearance time have been shown to be significantly correlated with mortality [52, 82], independently of clinically apparent organ dysfunction and shock in patients admitted to the ED with severe sepsis [64]. Overall, our results suggest a linear association between these variables and mortality.

The overall negative effect of low oxygen delivery when a patient is admitted (Figures 37-42) had a lower effect on mortality than at 6 or 24 hours after admission. This difference is possibly because of extra patient treatment delivered to the severe cases with low DO_2. There were 27 patients with DO_2 below 200 ml/min, but only 15 such patients at 6 hours. More serious and systemic problems can manifest with persistently low DO_2[57].

Given the established nonlinear nature of critical velocities, it follows that untransformed linear models would result in biased results. The relative drop in a given information criterion from a linear model to a GAM or GLMM can yield information on the strength of the nonlinear relationship. By this standard, Cardiac Index and MVO_2 benefitted the most from flexible modeling.

There are several weaknesses to this work. Statistical significance of nonlinearity was not assessed. Given that data driven criterion such as AIC and BIC were used to choose the edf, statistical inference is problematic [18]. Previous work on the associations between postoperative physiological variables and hemodynamic parameters and mortality focused on identification of a threshold [83]. However, previous work has indicated that estimating a threshold using more objective criteria using GAMs was difficult [18], thus we only estimated curves here.

Additionally, we used GLMMs estimated via PQL which is known to perform poorly in some situations, whereas it may be interesting to consider estimation via numerical integration[84]. While there is a large body of work comparing GLMMs estimated via PQL to those estimated using numerical integration for analyzing correlated or clustered data, we found no comparisons of the two estimation methods in this context.

Our results also demonstrated that GCV did not perform well when used in this context with a binomial outcome. GCV is not a likelihood based method for binomial GLMMs.

For normal linear mixed models the marginal distribution of Y is directly computed as a multivariate normal. For binomial mixed models the marginal distribution of Y can be approximated using penalized quasi-likelihood [85]. For small τ^2 or high n, GCV is a rough transformation of AIC on a Gaussian likelihood [86, 87]. In most binomial cases, GCV is emulating the wrong distribution. By being geared toward Gaussian distributions it is possible that GCVs are under penalizing additional df. This would require systematic testing to draw a substantial conclusion.

Previous modeling of the vital parameters has relied on a priori specification of the functional form to capture the threshold relationship (e.g. binary variables created by thresholds). GAMs and GLMMs deliver comparable results without this encumbrance. There is an apparent region for where the probability of mortality begins to increase at 250 ml/min for

DO_2. For Cardiac Index, there is a corresponding area just above 2 L/min/m^2. For MVO_2 only saturation below 40% appears to be harmful. Evidence for a threshold is less strong for Serum Lactate. GLMMs in general produced a sharper and more consistent threshold point. GAMs, however, produced lower minimized information criterion when selecting degrees of freedom with AIC, BIC, and GCV. When selecting this binomial model's effective df GCV behaved inconsistently. As well, BIC may have been too conservative, and seemed to fail to capture the extent of the nonlinearity, and repeatedly selected a linear result.

Small sample size GLMM model selection by AIC

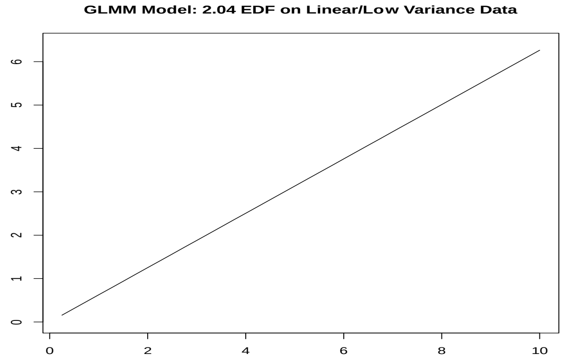


Figure 1

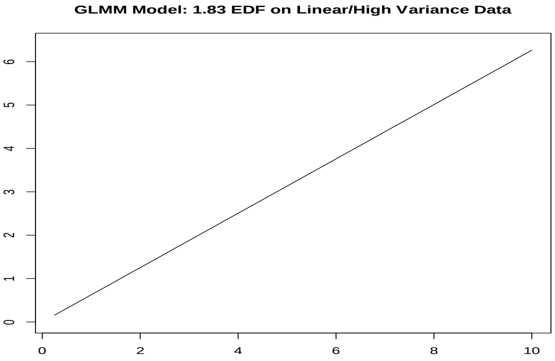


Figure 2

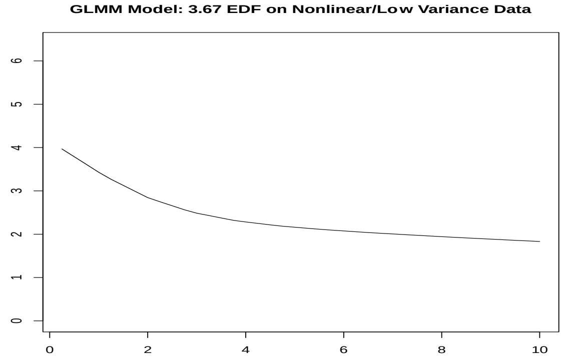


Figure 3

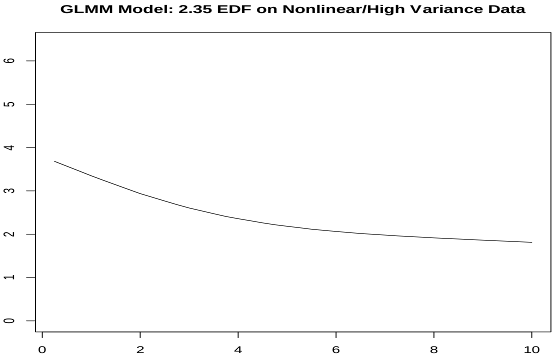


Figure 4

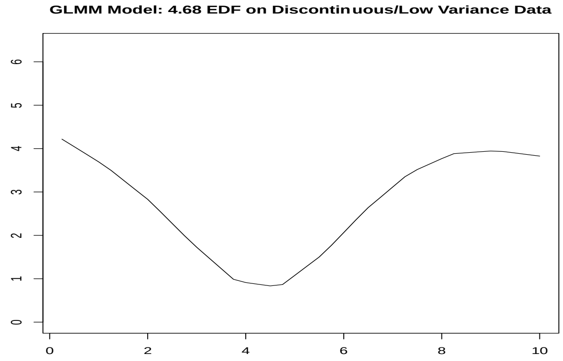


Figure 5

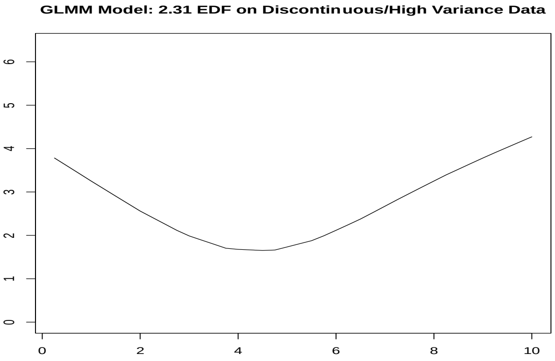


Figure 6

Small sample size GAM model selection by AIC

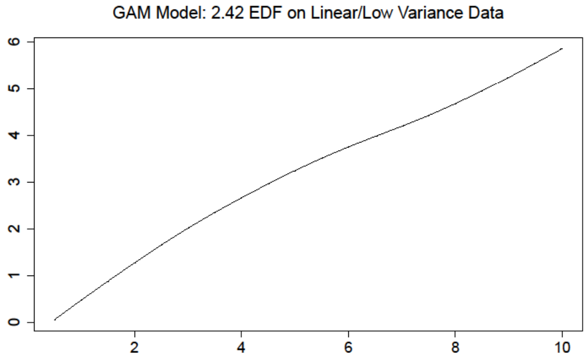

GAM Model: 2.42 EDF on Linear/Low Variance Data

Figure 7

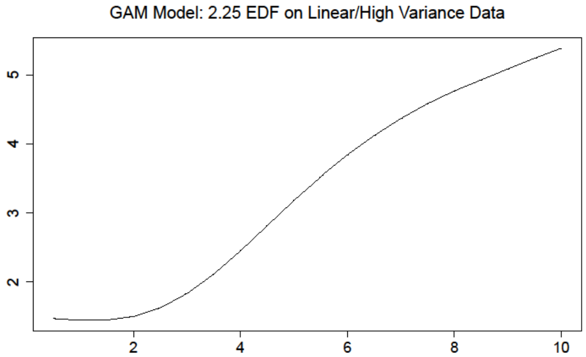

GAM Model: 2.25 EDF on Linear/High Variance Data

Figure 8

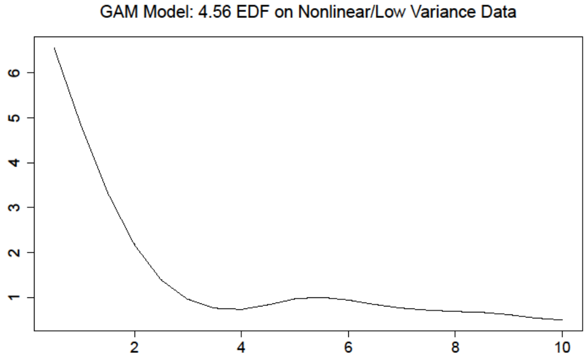

GAM Model: 4.56 EDF on Nonlinear/Low Variance Data

Figure 9

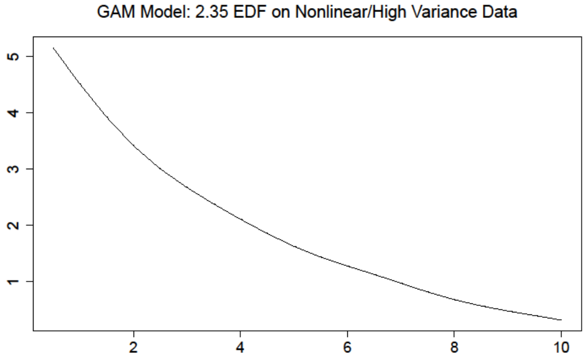

GAM Model: 2.35 EDF on Nonlinear/High Variance Data

Figure 10

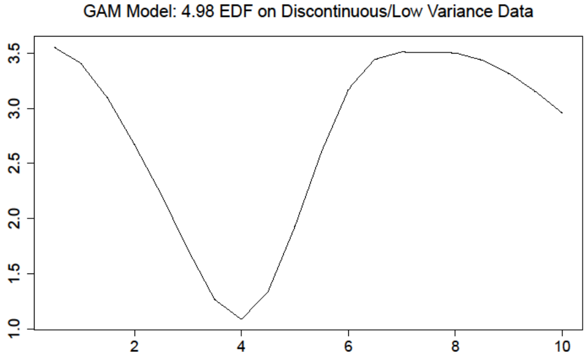

GAM Model: 4.98 EDF on Discontinuous/Low Variance Data

Figure 11

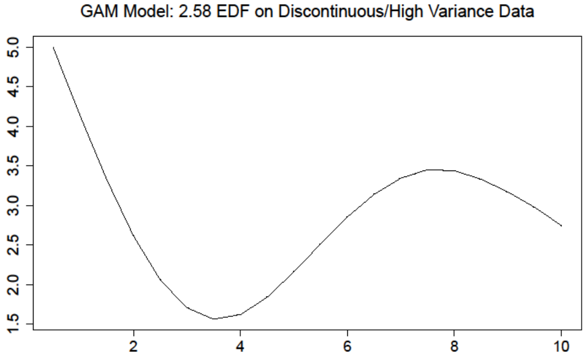

GAM Model: 2.58 EDF on Discontinuous/High Variance Data

Figure 12

Large sample size GLMM model selection by AIC

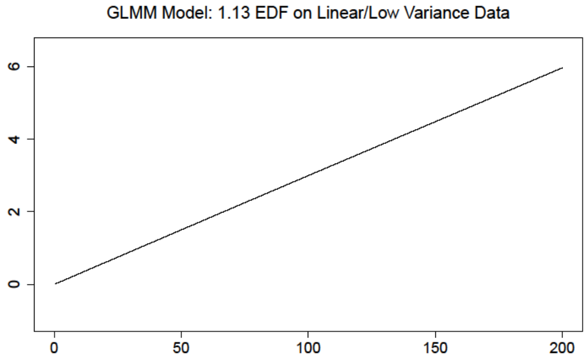

GLMM Model: 1.13 EDF on Linear/Low Variance Data

Figure 13

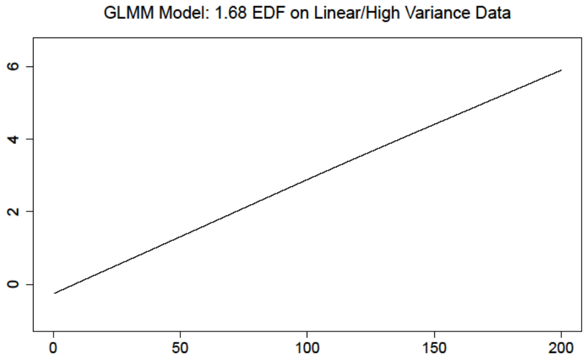

GLMM Model: 1.68 EDF on Linear/High Variance Data

Figure 14

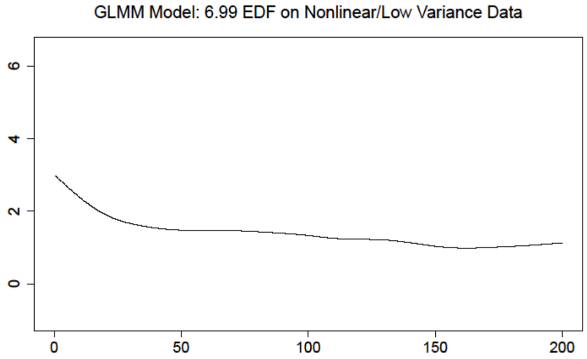

GLMM Model: 6.99 EDF on Nonlinear/Low Variance Data

Figure 15

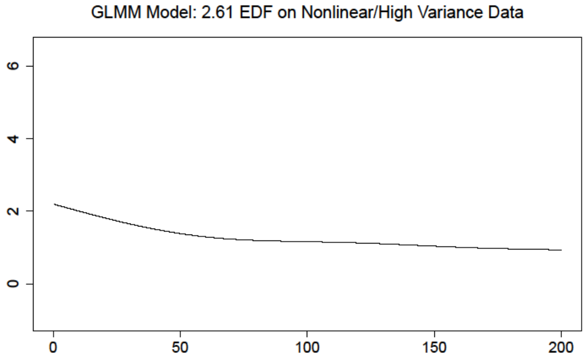

GLMM Model: 2.61 EDF on Nonlinear/High Variance Data

Figure 16

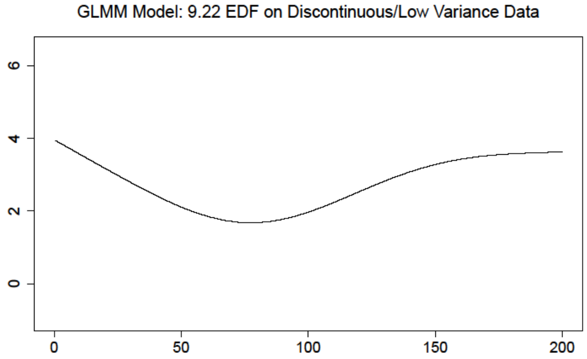

GLMM Model: 9.22 EDF on Discontinuous/Low Variance Data

Figure 17

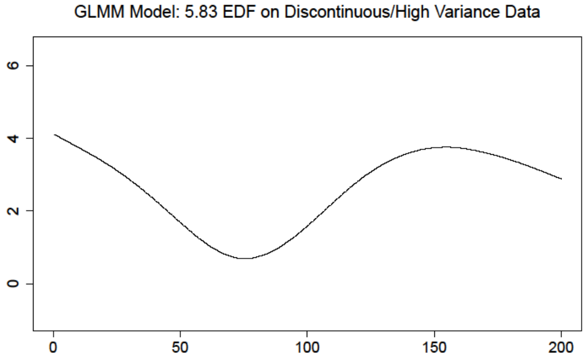

GLMM Model: 5.83 EDF on Discontinuous/High Variance Data

Figure 18

Large sample size GAM model selection by AIC

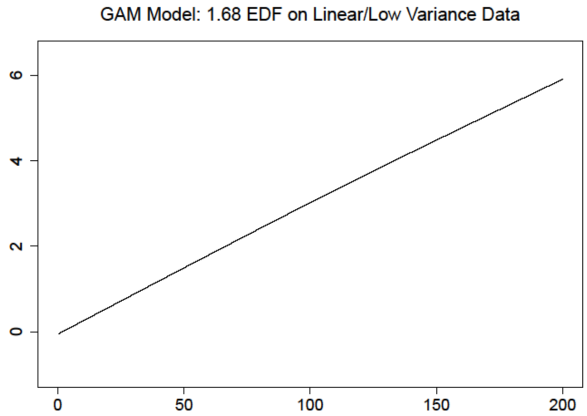

GAM Model: 1.68 EDF on Linear/Low Variance Data

Figure 19

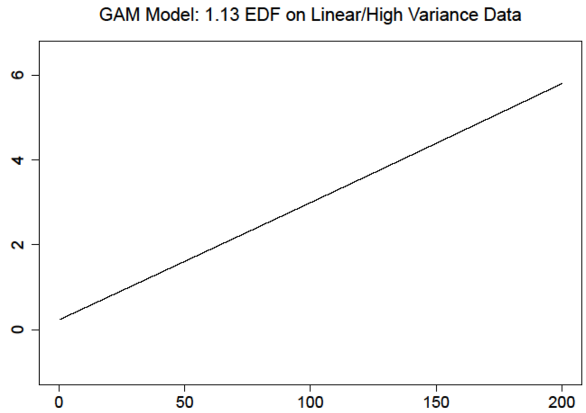

GAM Model: 1.13 EDF on Linear/High Variance Data

Figure 20

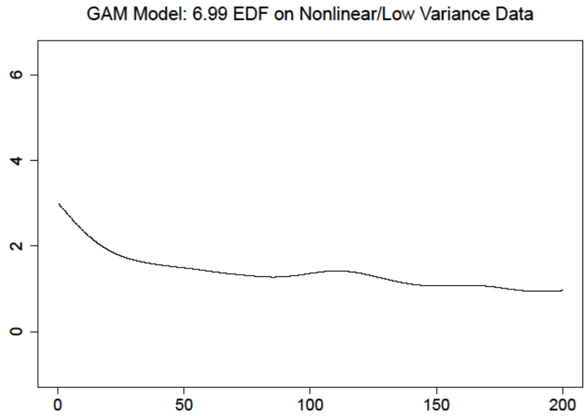

GAM Model: 6.99 EDF on Nonlinear/Low Variance Data

Figure 21

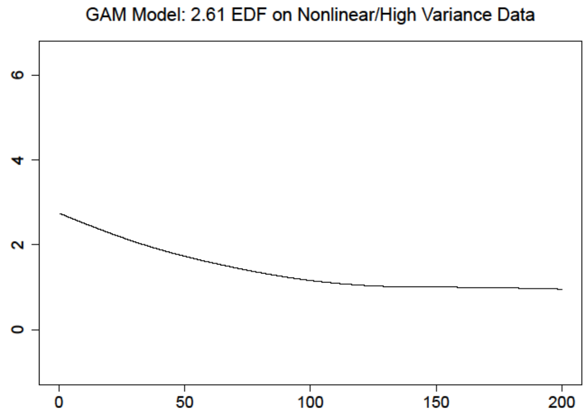

GAM Model: 2.61 EDF on Nonlinear/High Variance Data

Figure 22

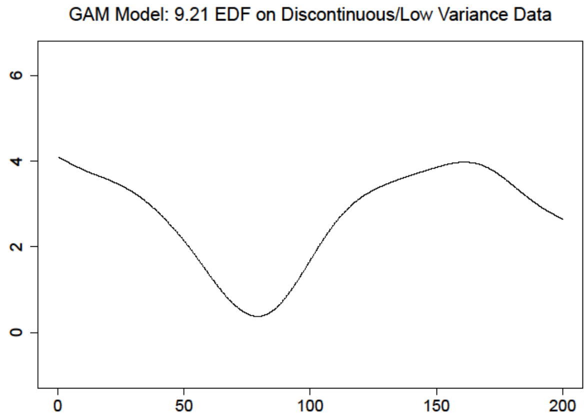

GAM Model: 9.21 EDF on Discontinuous/Low Variance Data

Figure 23

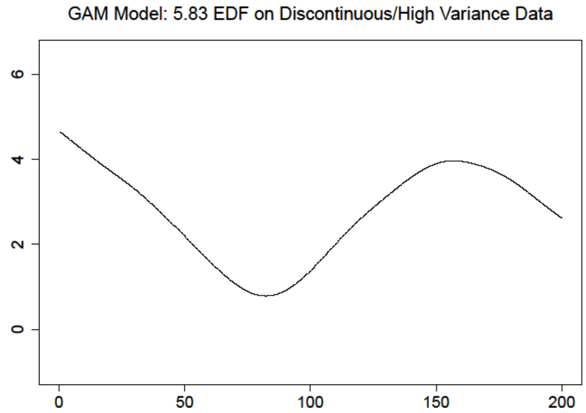

GAM Model: 5.83 EDF on Discontinuous/High Variance Data

Figure 24

Small sample size GLMM model selection by BIC

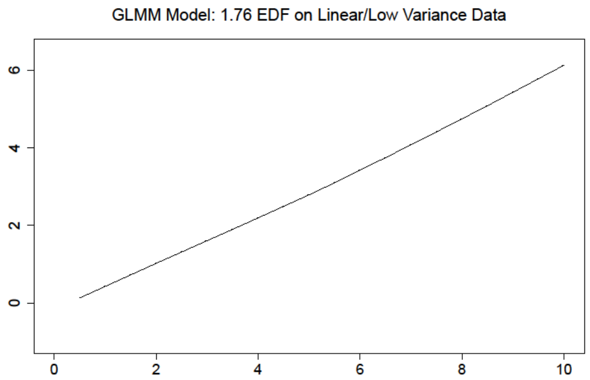

GLMM Model: 1.76 EDF on Linear/Low Variance Data

Figure 25

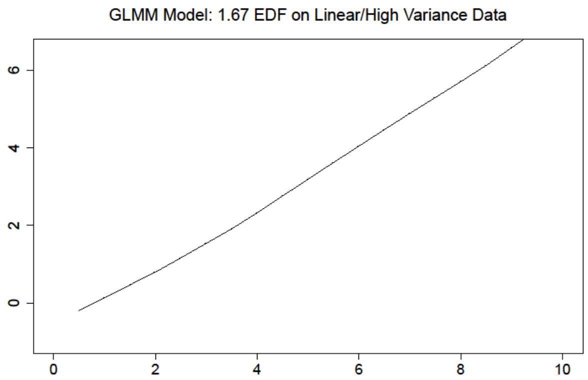

GLMM Model: 1.67 EDF on Linear/High Variance Data

Figure 26

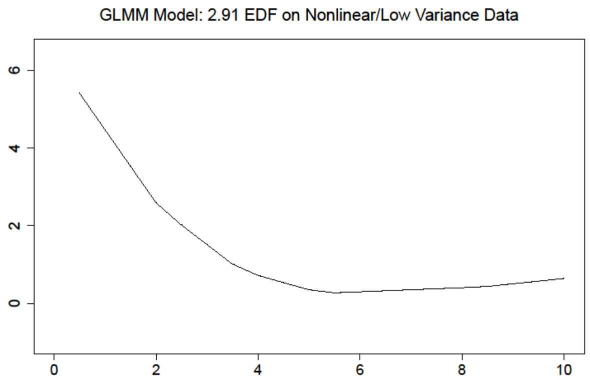

GLMM Model: 2.91 EDF on Nonlinear/Low Variance Data

Figure 27

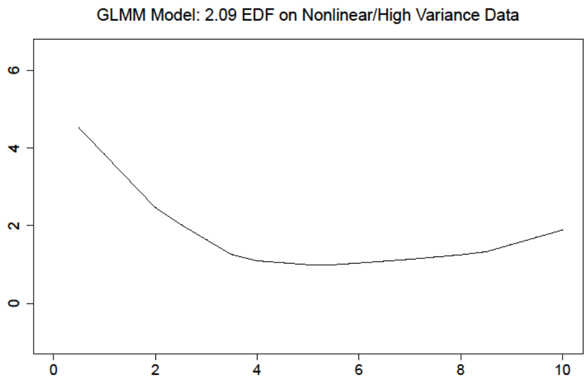

GLMM Model: 2.09 EDF on Nonlinear/High Variance Data

Figure 28

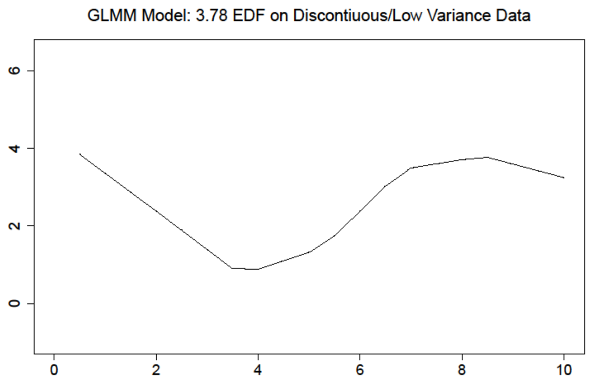

GLMM Model: 3.78 EDF on Discontiuous/Low Variance Data

Figure 29

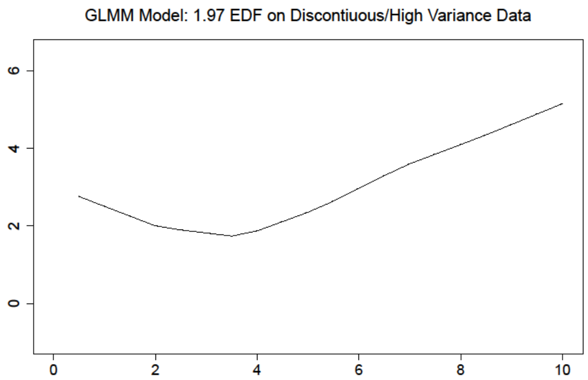

GLMM Model: 1.97 EDF on Discontiuous/High Variance Data

Figure 30

Small sample size GAM model selection by BIC

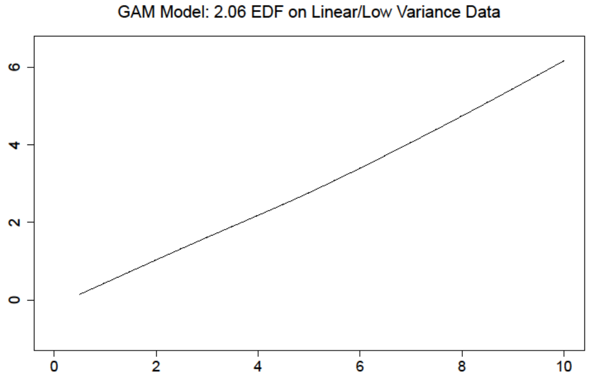


Figure 31

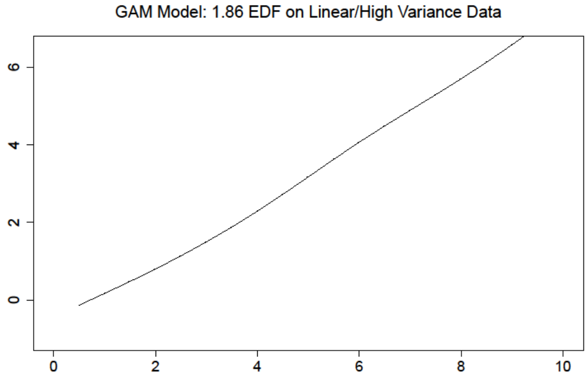


Figure 32

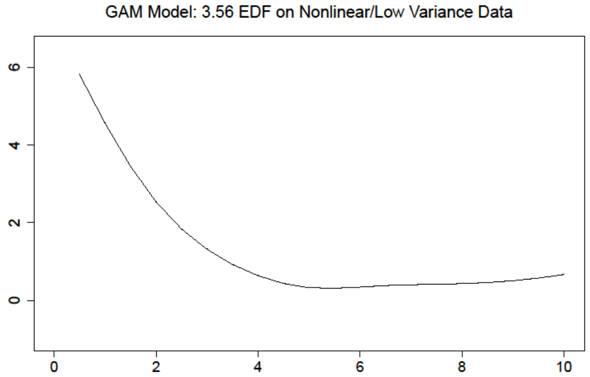


Figure 33

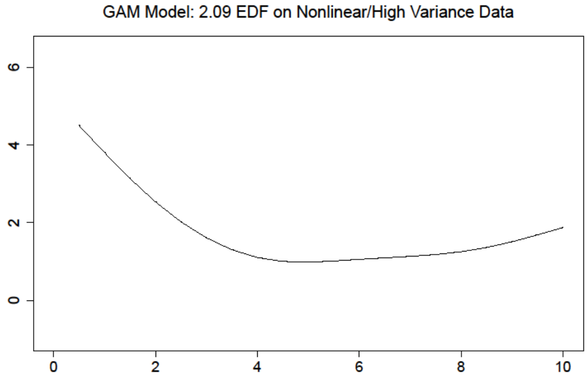


Figure 34

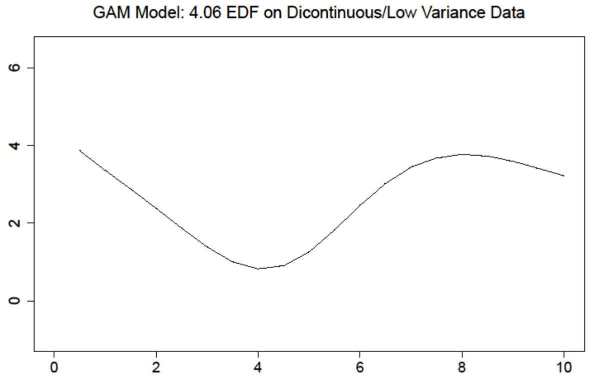


Figure 35

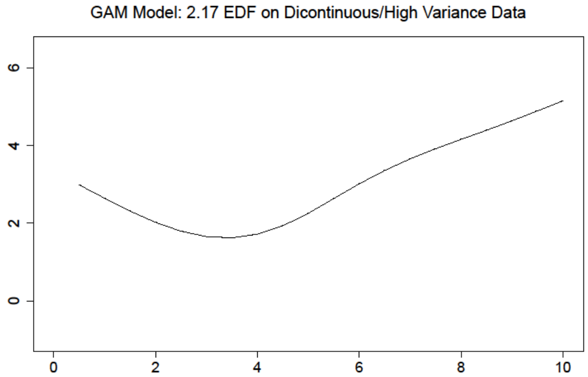


Figure 36

Large sample size GLMM model selection by BIC

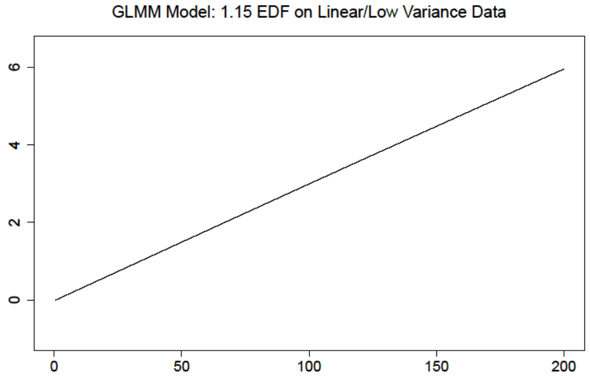

GLMM Model: 1.15 EDF on Linear/Low Variance Data

Figure 37

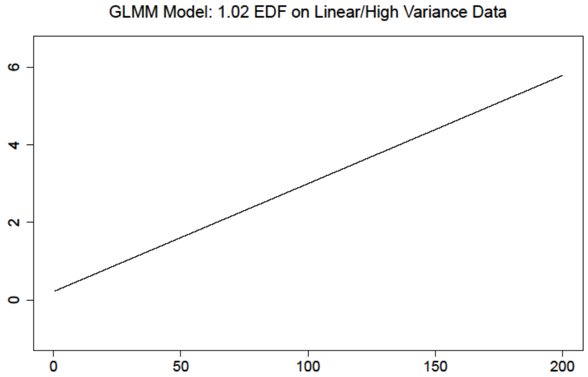

GLMM Model: 1.02 EDF on Linear/High Variance Data

Figure 38

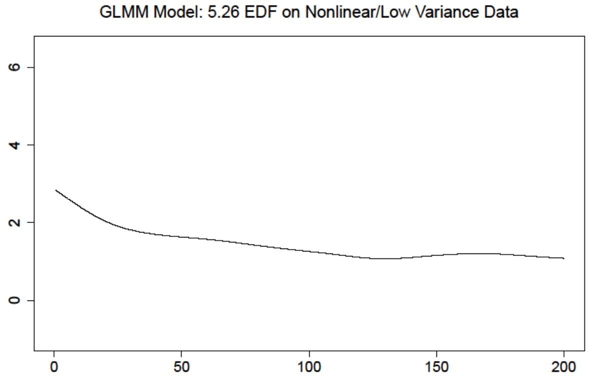

GLMM Model: 5.26 EDF on Nonlinear/Low Variance Data

Figure 39

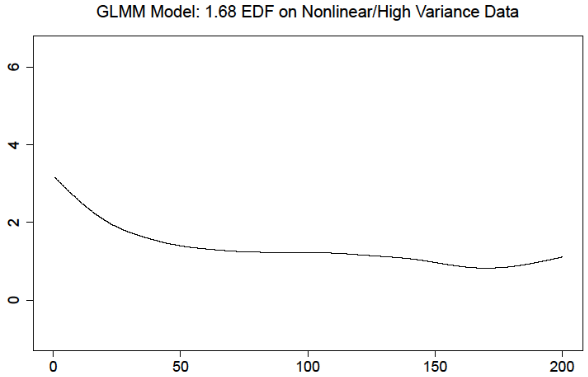

GLMM Model: 1.68 EDF on Nonlinear/High Variance Data

Figure 40

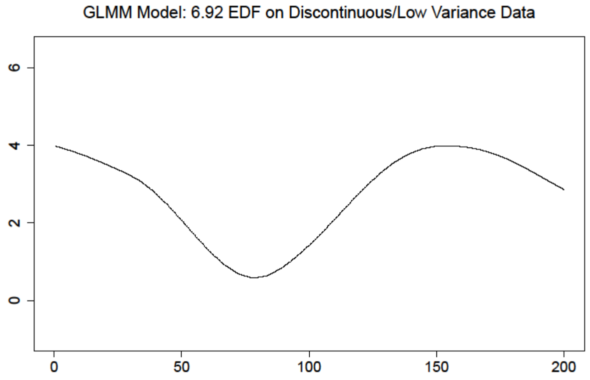

GLMM Model: 6.92 EDF on Discontinuous/Low Variance Data

Figure 41

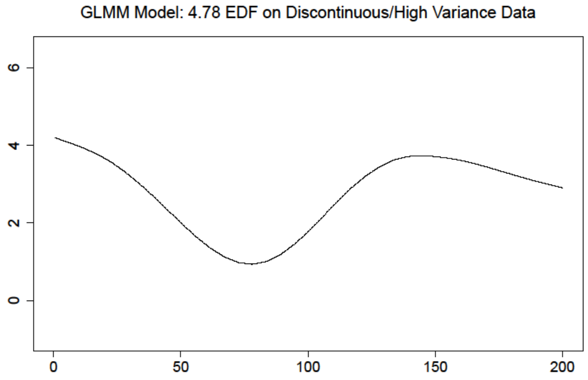

GLMM Model: 4.78 EDF on Discontinuous/High Variance Data

Figure 42

Large sample size GAM model selection by BIC

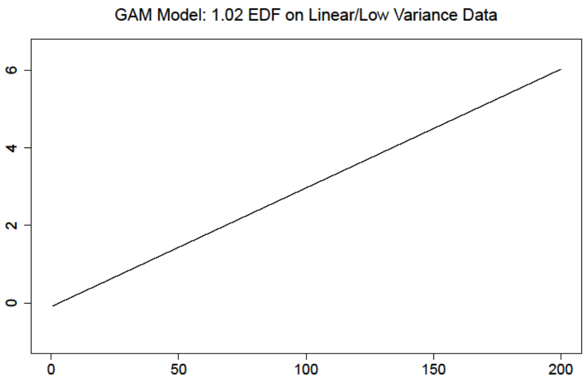

GAM Model: 1.02 EDF on Linear/Low Variance Data

Figure 43

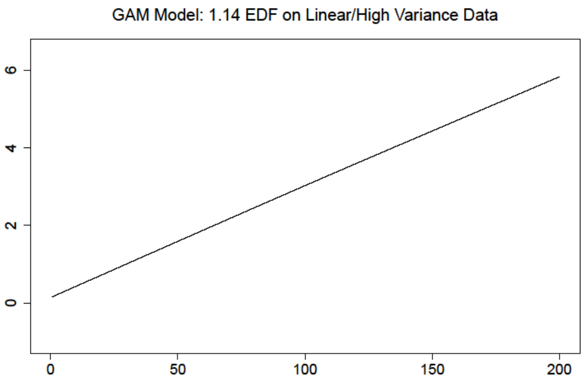

GAM Model: 1.14 EDF on Linear/High Variance Data

Figure 44

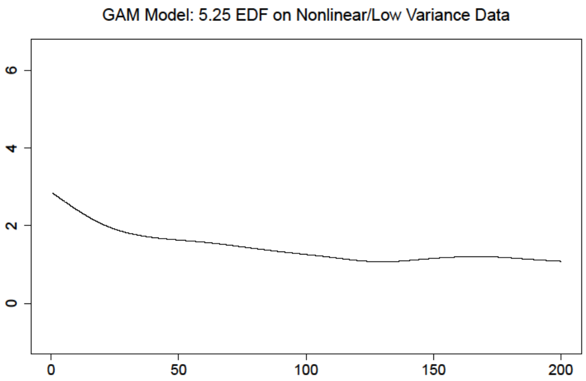

GAM Model: 5.25 EDF on Nonlinear/Low Variance Data

Figure 45

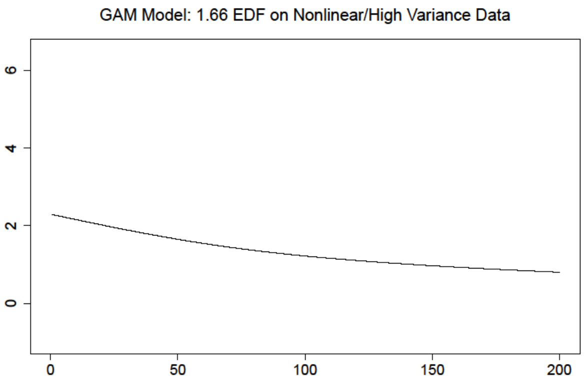

GAM Model: 1.66 EDF on Nonlinear/High Variance Data

Figure 46

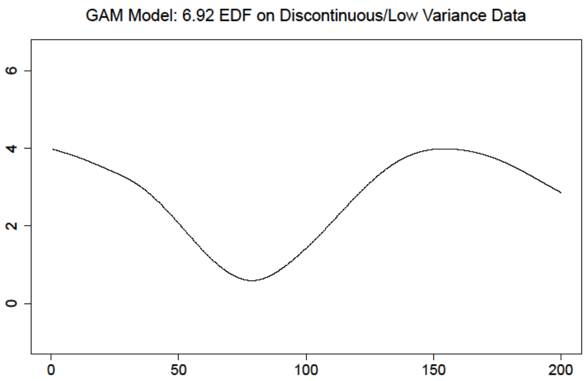

GAM Model: 6.92 EDF on Discontinuous/Low Variance Data

Figure 47

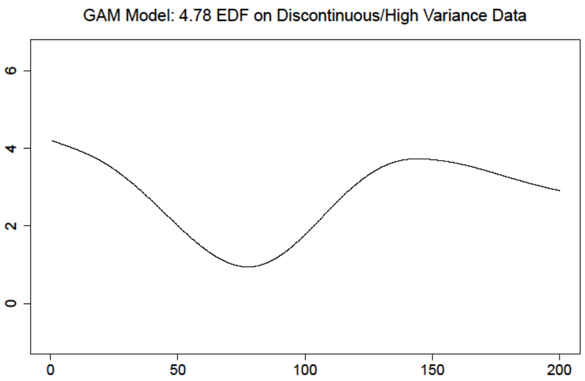

GAM Model: 4.78 EDF on Discontinuous/High Variance Data

Figure 48

Small sample size GLMM model selection by GCV

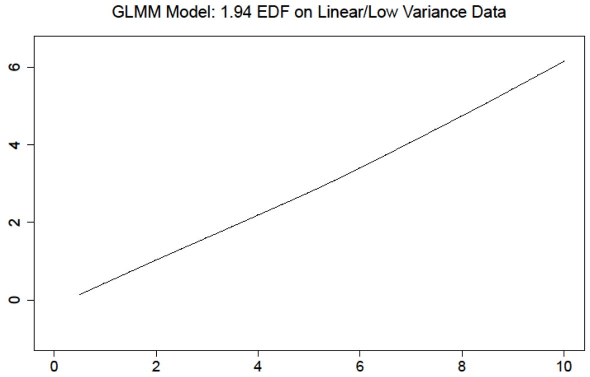

GLMM Model: 1.94 EDF on Linear/Low Variance Data

Figure 49

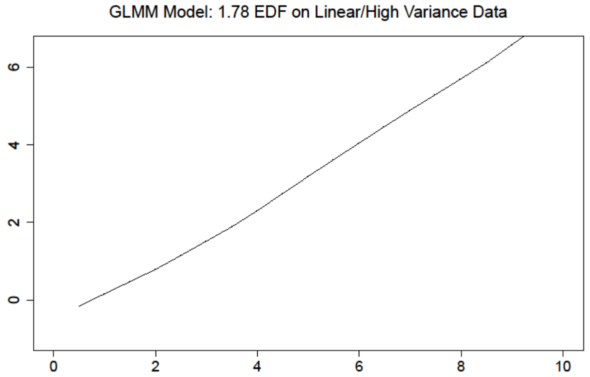

GLMM Model: 1.78 EDF on Linear/High Variance Data

Figure 50

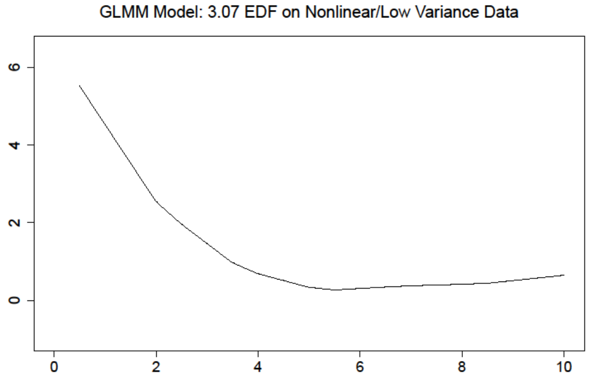

GLMM Model: 3.07 EDF on Nonlinear/Low Variance Data

Figure 51

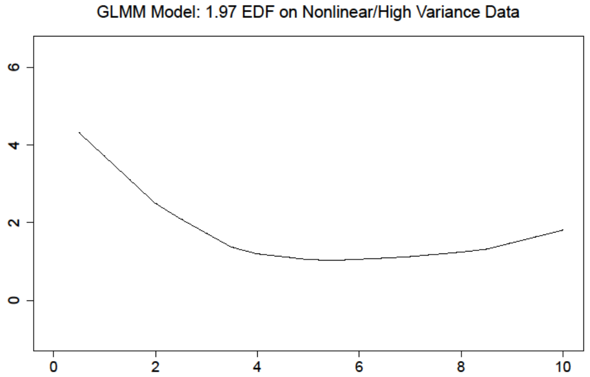

GLMM Model: 1.97 EDF on Nonlinear/High Variance Data

Figure 52

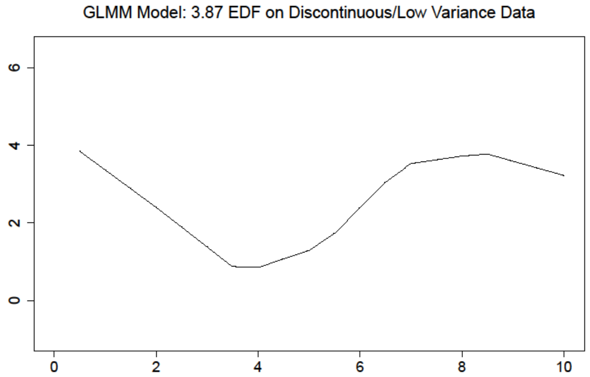

GLMM Model: 3.87 EDF on Discontinuous/Low Variance Data

Figure 53

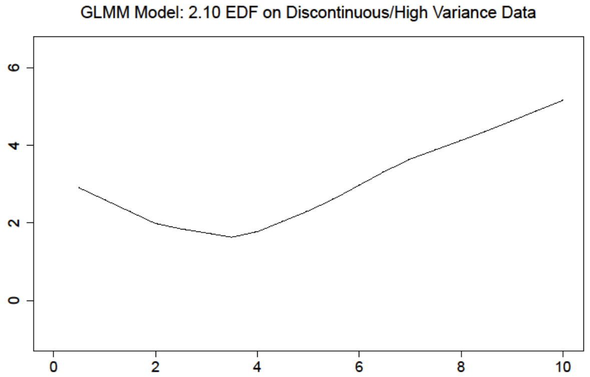

GLMM Model: 2.10 EDF on Discontinuous/High Variance Data

Figure 54

Small sample size GAM model selection by GCV

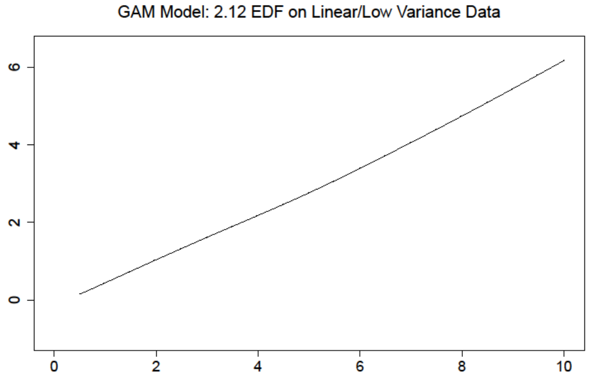

GAM Model: 2.12 EDF on Linear/Low Variance Data

Figure 55

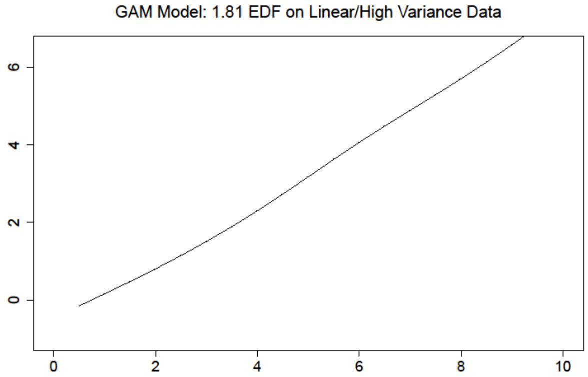

GAM Model: 1.81 EDF on Linear/High Variance Data

Figure 56

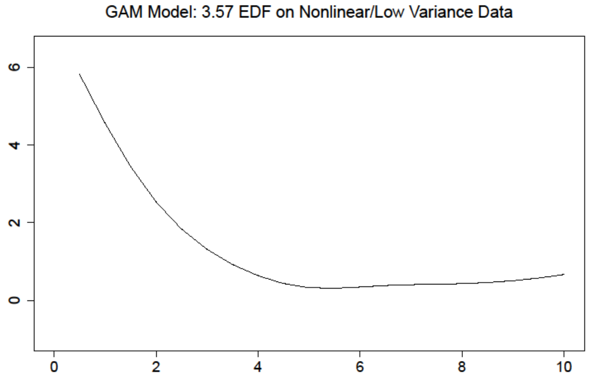

GAM Model: 3.57 EDF on Nonlinear/Low Variance Data

Figure 57

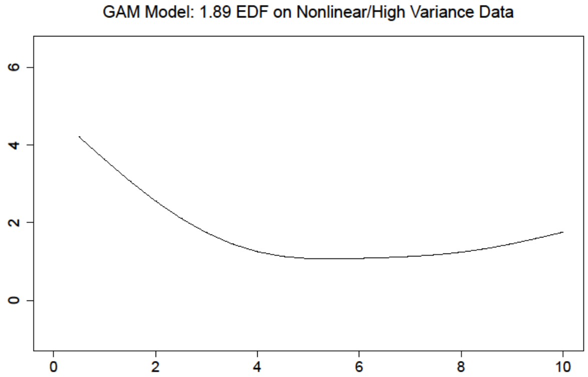

GAM Model: 1.89 EDF on Nonlinear/High Variance Data

Figure 58

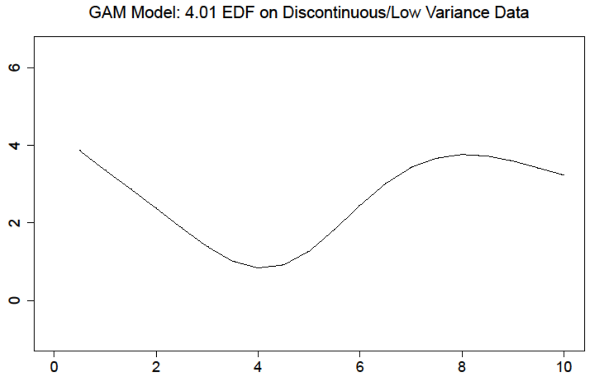

GAM Model: 4.01 EDF on Discontinuous/Low Variance Data

Figure 59

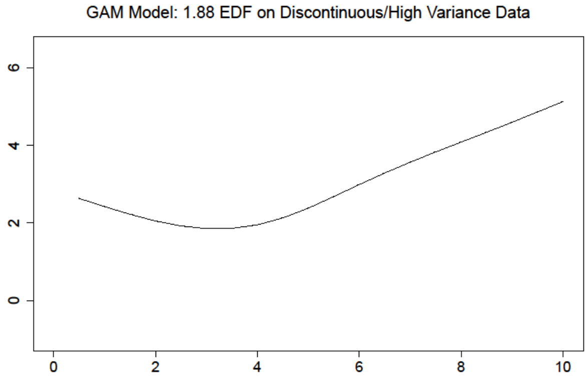

GAM Model: 1.88 EDF on Discontinuous/High Variance Data

Figure 60

Large sample size GLMM model selection by GCV

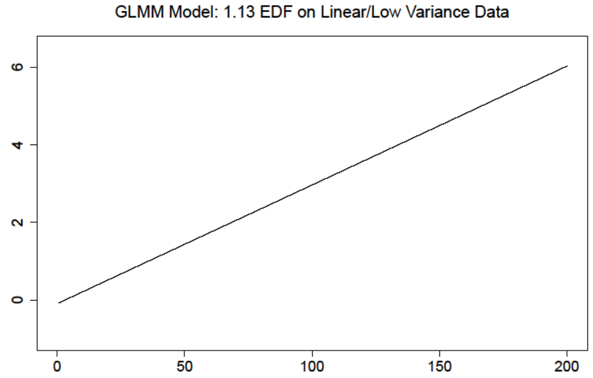

GLMM Model: 1.13 EDF on Linear/Low Variance Data

Figure 61

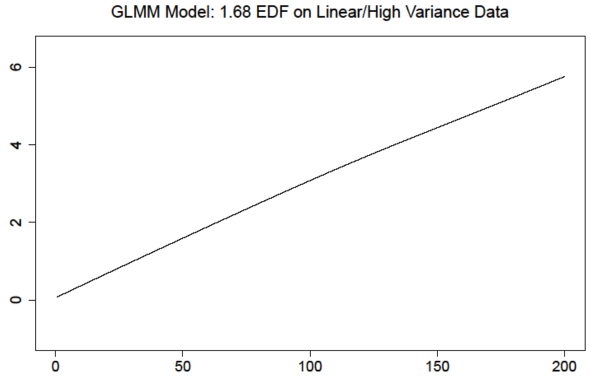

GLMM Model: 1.68 EDF on Linear/High Variance Data

Figure 62

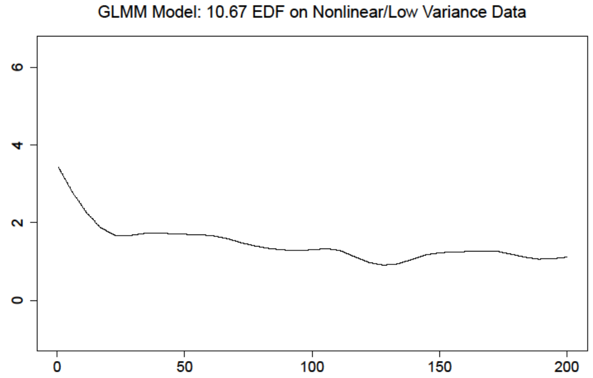

GLMM Model: 10.67 EDF on Nonlinear/Low Variance Data

Figure 63

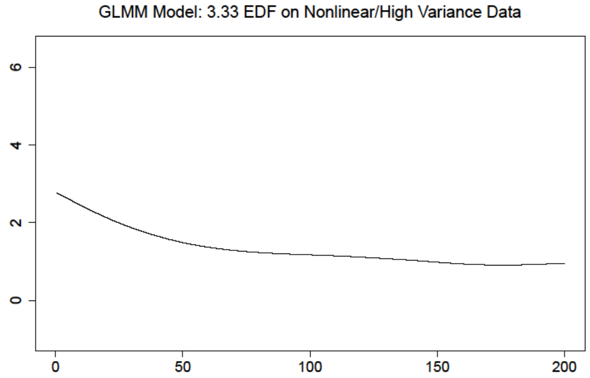

GLMM Model: 3.33 EDF on Nonlinear/High Variance Data

Figure 64

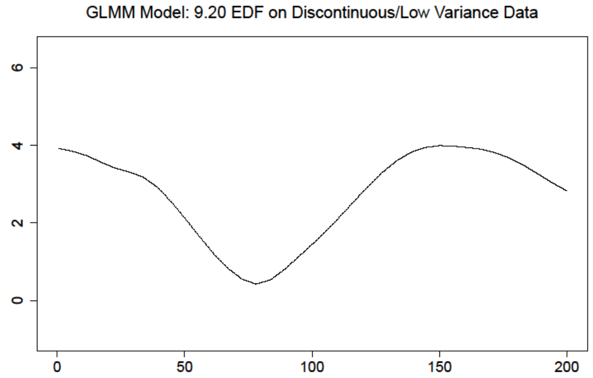

GLMM Model: 9.20 EDF on Discontinuous/Low Variance Data

Figure 65

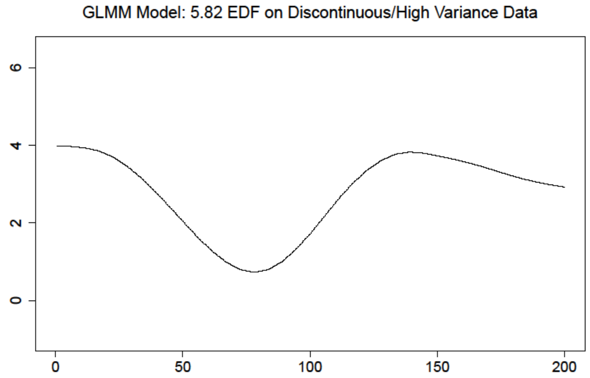

GLMM Model: 5.82 EDF on Discontinuous/High Variance Data

Figure 66

Large sample size GAM model selection by GCV

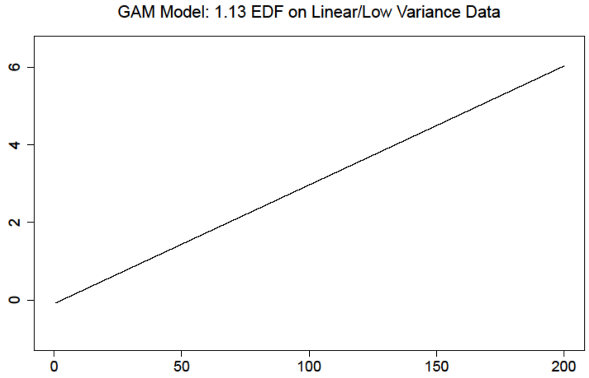

GAM Model: 1.13 EDF on Linear/Low Variance Data

Figure 67

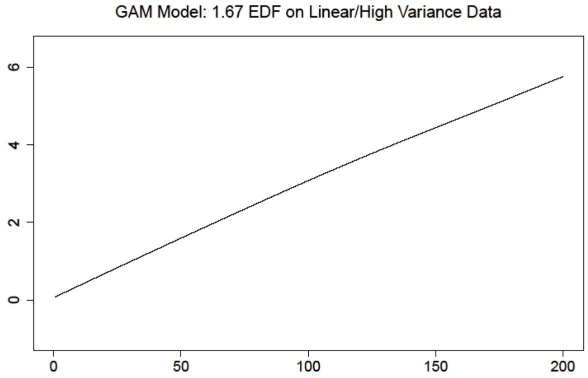

GAM Model: 1.67 EDF on Linear/High Variance Data

Figure 68

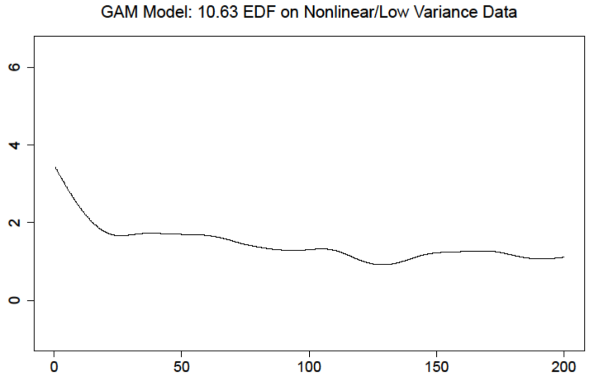

GAM Model: 10.63 EDF on Nonlinear/Low Variance Data

Figure 69

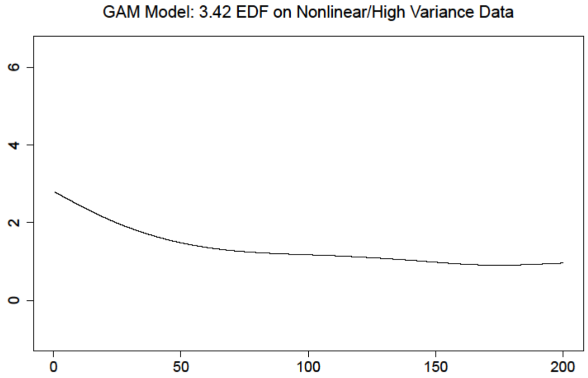

GAM Model: 3.42 EDF on Nonlinear/High Variance Data

Figure 70

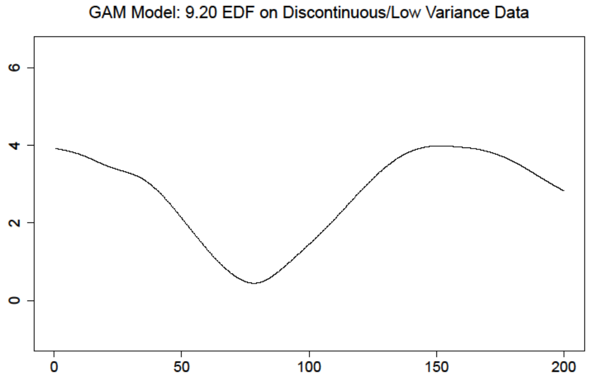

GAM Model: 9.20 EDF on Discontinuous/Low Variance Data

Figure 71

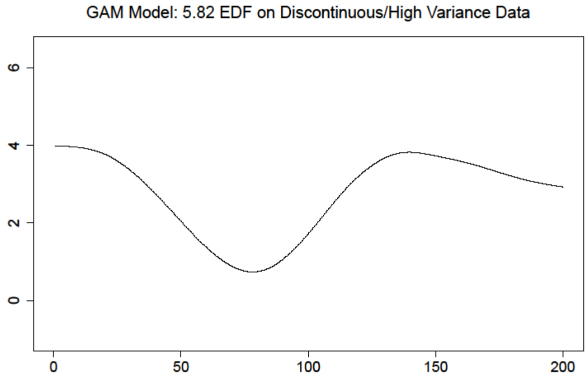
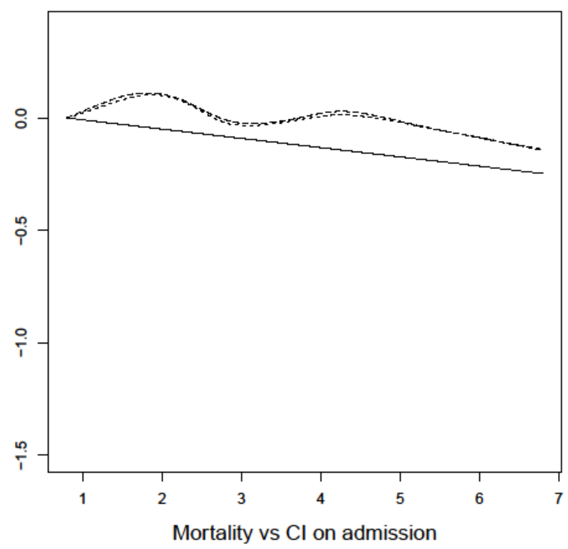
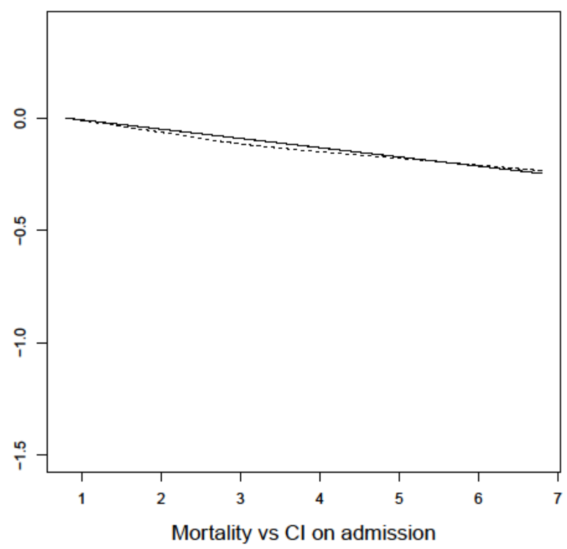

GAM Model: 5.82 EDF on Discontinuous/High Variance Data

Figure 72

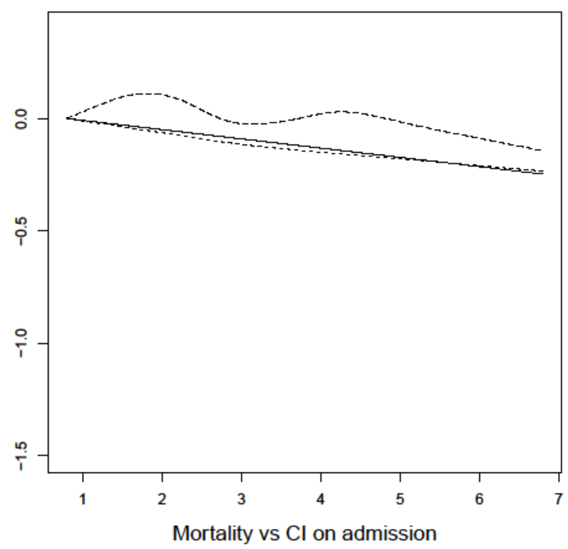
d.f. Chosen By AIC (GAM 5.10 GLMM 5.90)



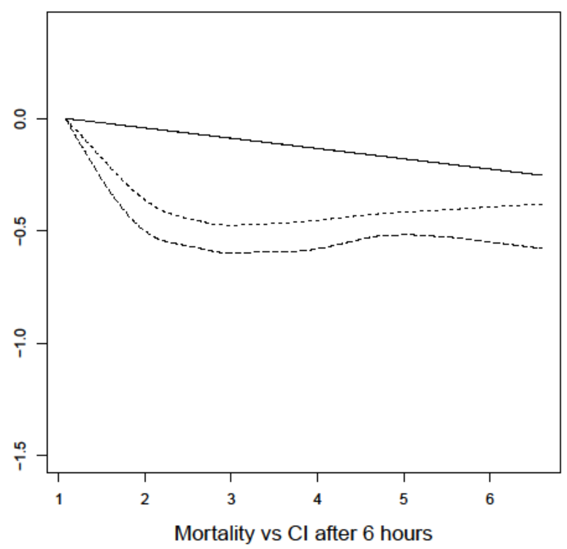
d.f. Chosen By BIC (GAM 1.00 GLMM 1.35)



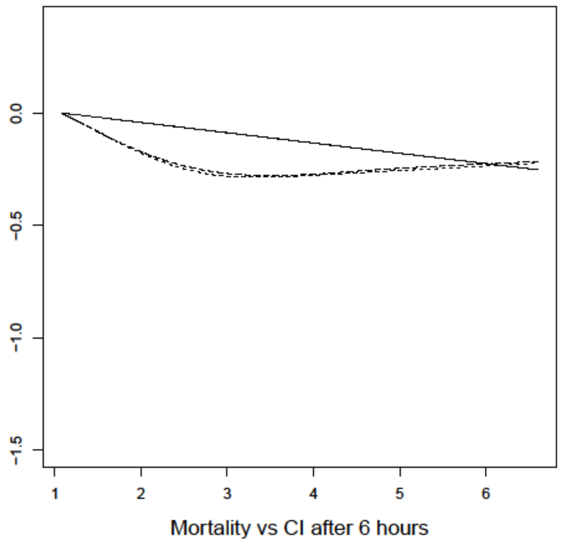
d.f. Chosen By GCV (GAM 5.80 GLMM 5.40)



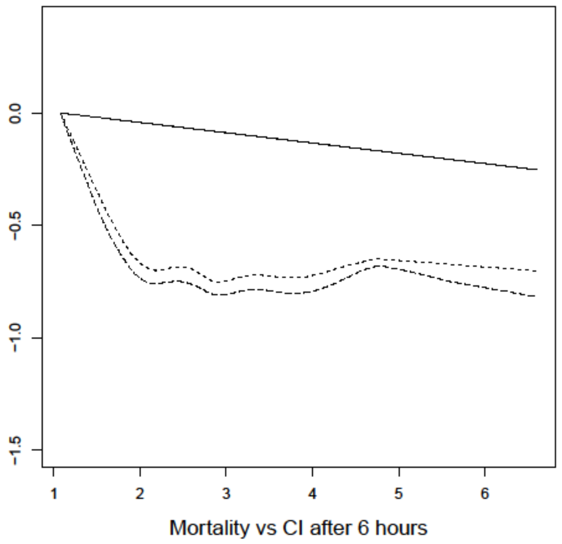
d.f. Chosen By AIC (GAM 4,30 GLMM 3.60)



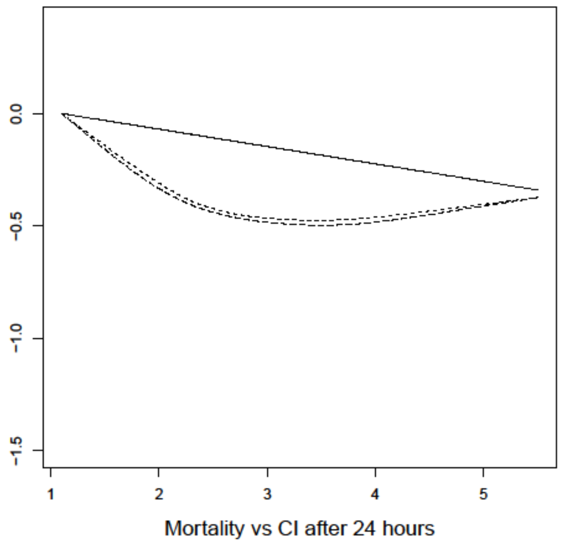
d.f. Chosen By BIC (GAM 2.20 GLMM 3.60)



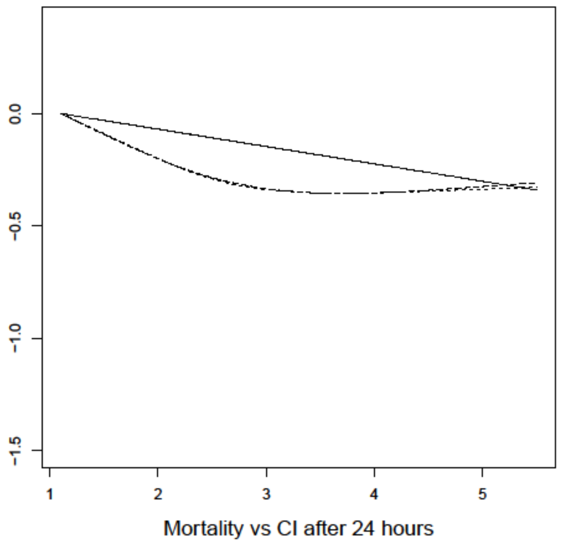
d.f. Chosen By GCV (GAM 8.25 GLMM 6.55)

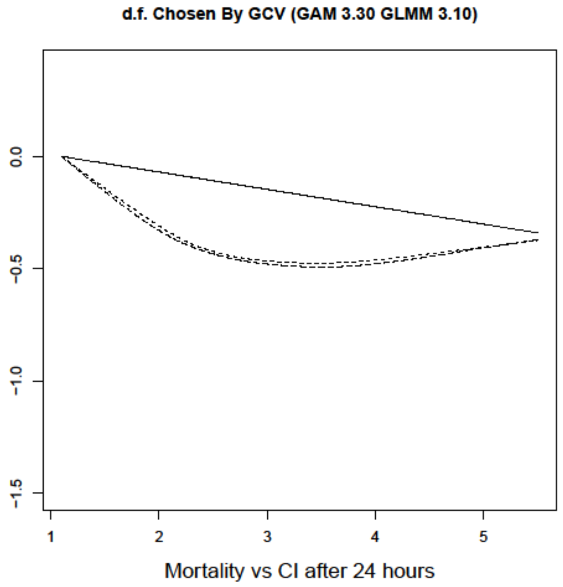


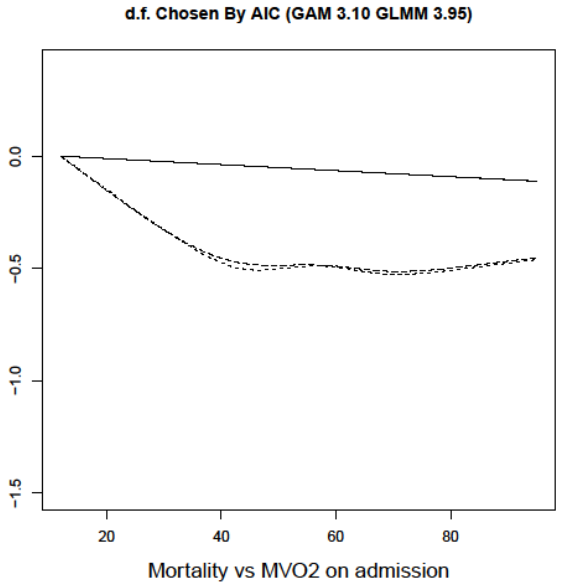
d.f. Chosen By AIC (GAM 3.30 GLMM 3.10)

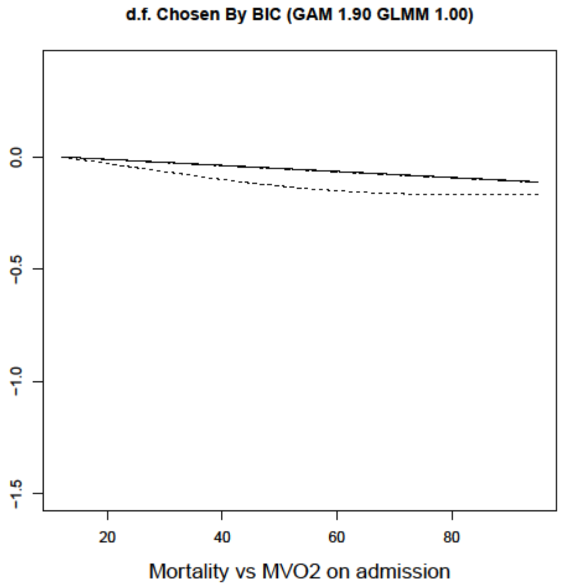


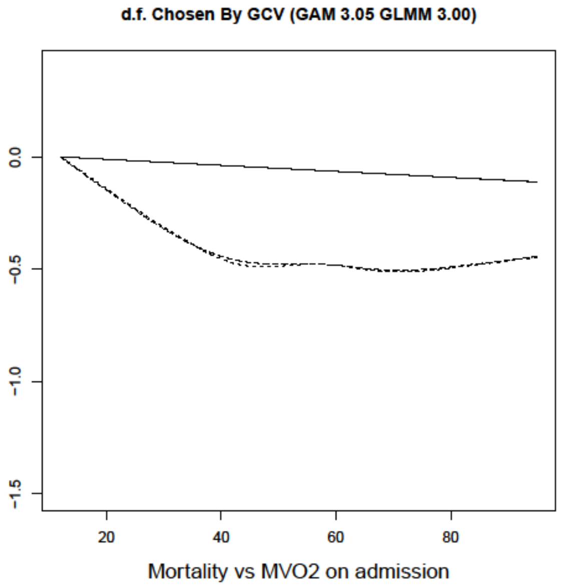
d.f. Chosen By BIC (GAM 1.40 GLMM 1.40)



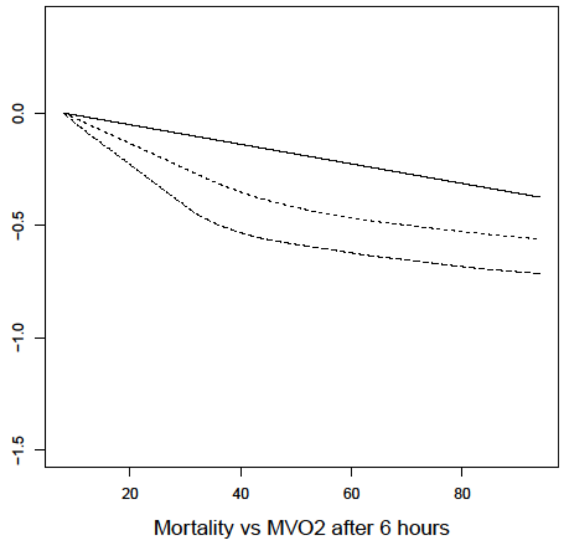

d.f. Chosen By GCV (GAM 3.30 GLMM 3.10)


d.f. Chosen By AIC (GAM 3.10 GLMM 3.95)

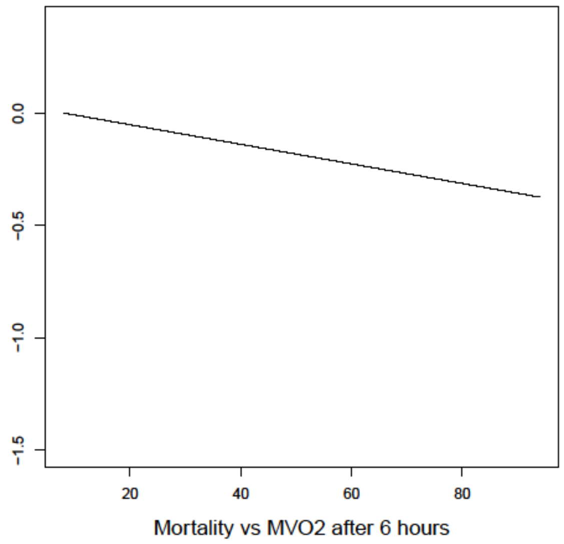

d.f. Chosen By BIC (GAM 1.90 GLMM 1.00)


d.f. Chosen By GCV (GAM 3.05 GLMM 3.00)

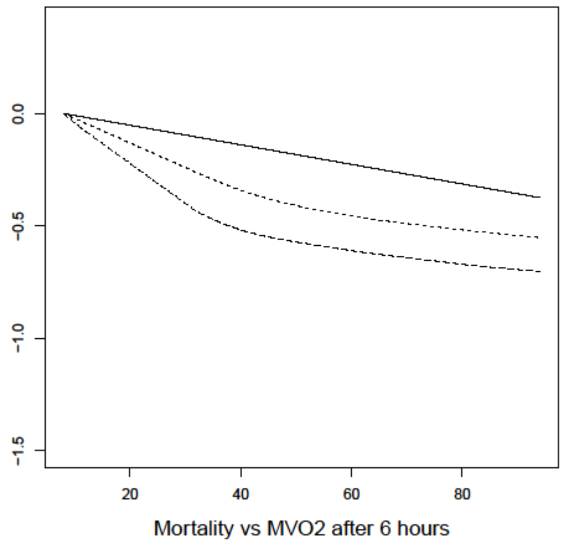
d.f. Chosen By AIC (GAM 4.15 GLMM 3.25)



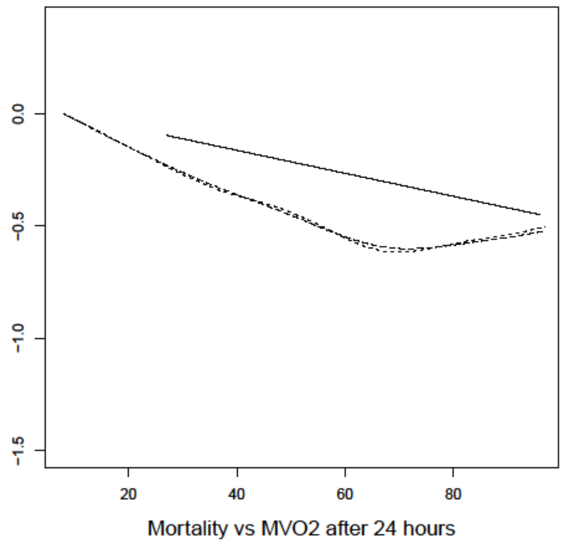
d.f. Chosen By BIC (GAM 1.000 GLMM 1.00)

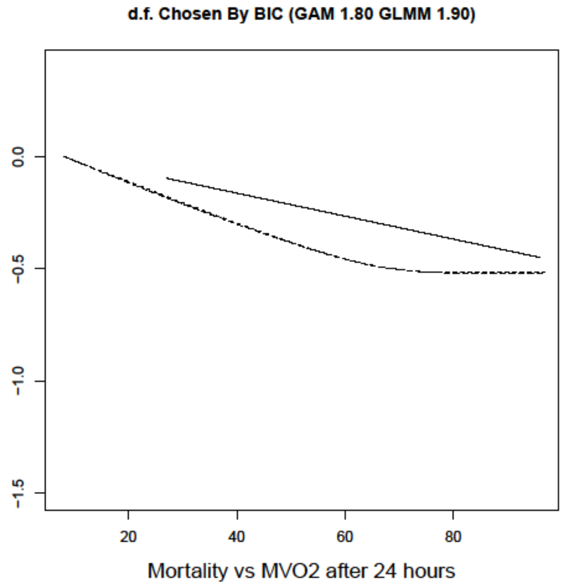


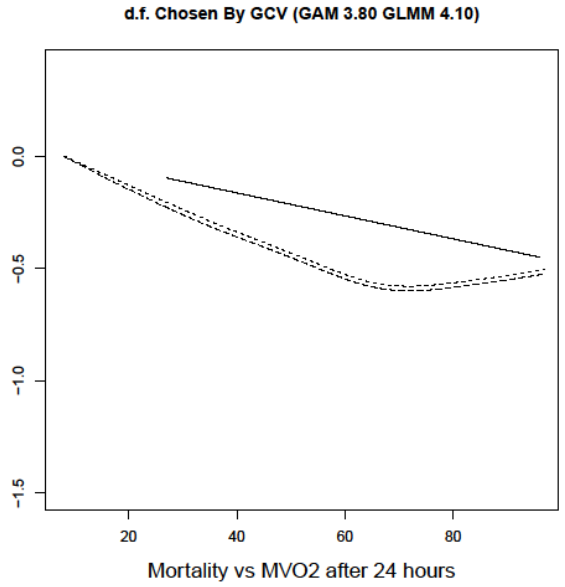
d.f. Chosen By GCV (GAM 4.10 GLMM 3.10)

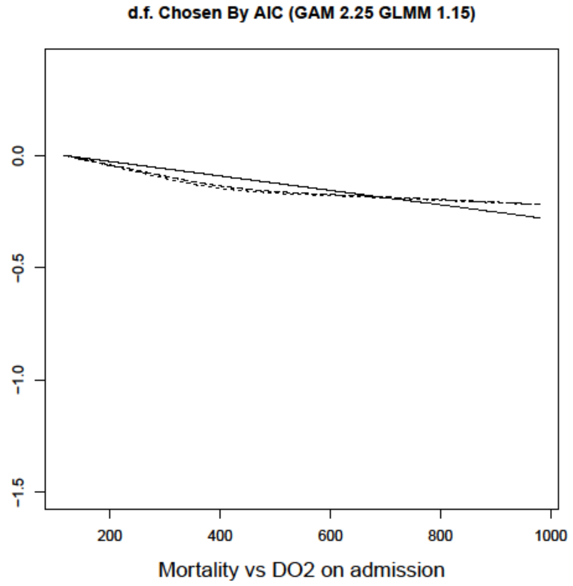


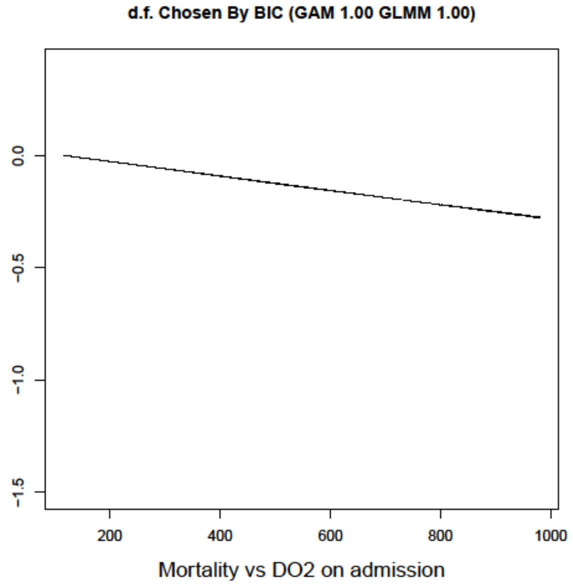
d.f. Chosen By AIC (GAM 3.35 GLMM 3.85)

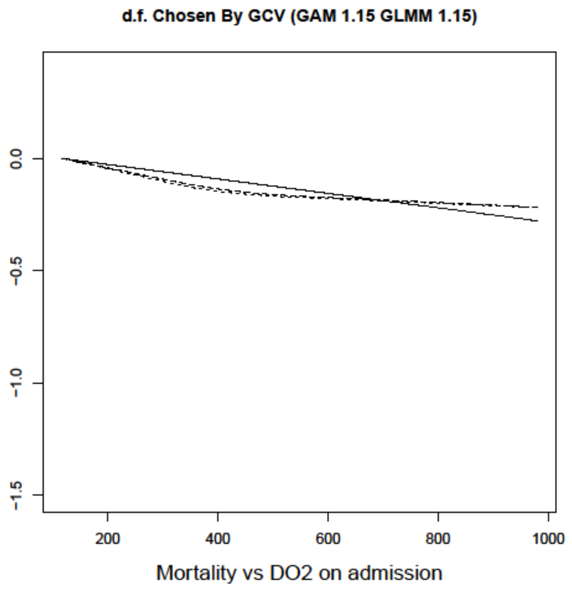


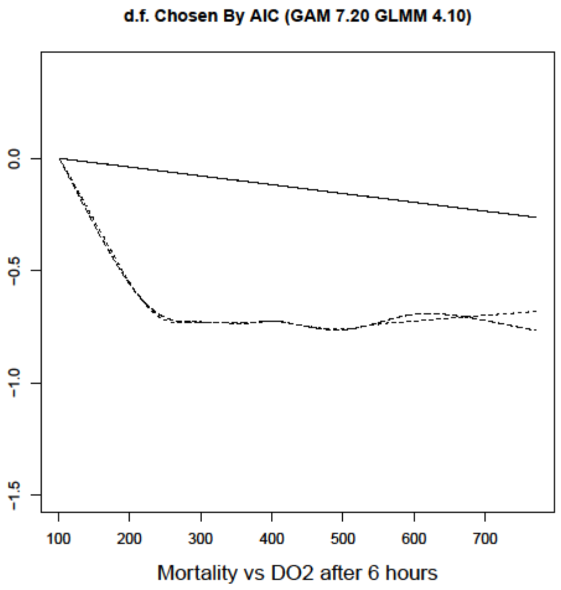

d.f. Chosen By BIC (GAM 1.80 GLMM 1.90)
0.0
−0.5
−1.0
−1.5
20
40
60
80
Mortality vs MVO2 after 24 hours

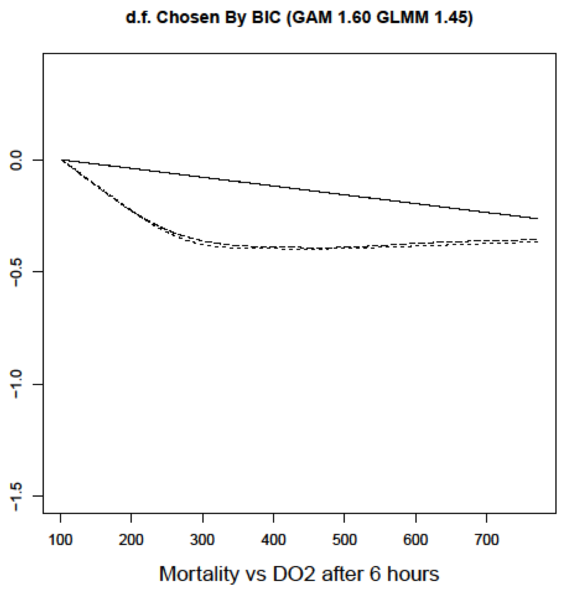

d.f. Chosen By GCV (GAM 3.80 GLMM 4.10)
0.0
−0.5
−1.0
−1.5
20
40
60
80
Mortality vs MVO2 after 24 hours

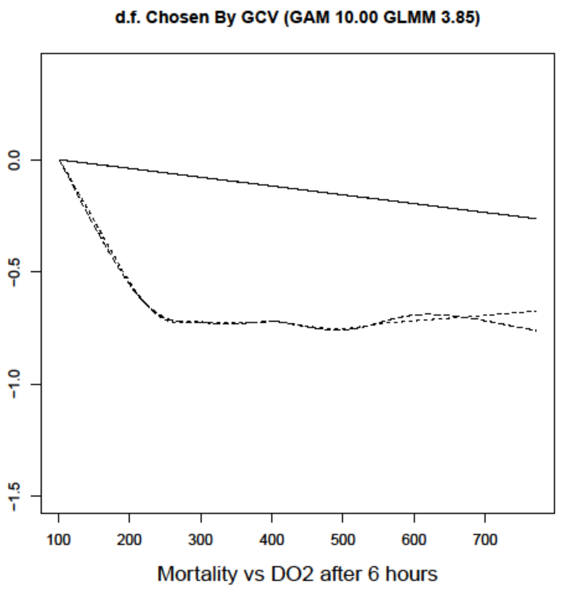

d.f. Chosen By AIC (GAM 2.25 GLMM 1.15)
0.0
−0.5
−1.0
−1.5
200
400
600
800
1000
Mortality vs DO2 on admission

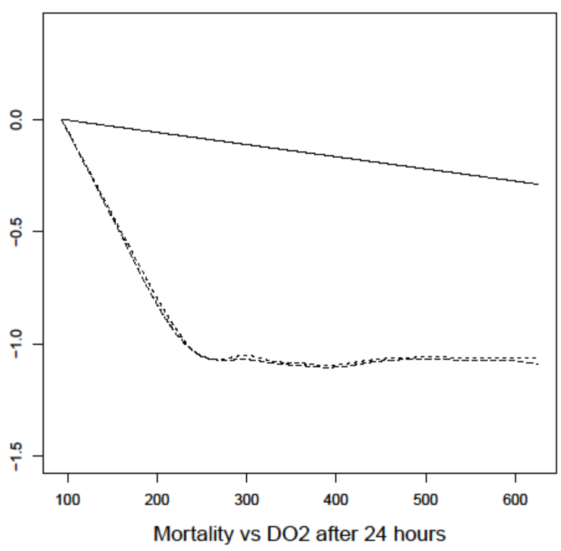

d.f. Chosen By BIC (GAM 1.00 GLMM 1.00)
0.0
−0.5
−1.0
−1.5
200
400
600
800
1000
Mortality vs DO2 on admission


d.f. Chosen By GCV (GAM 1.15 GLMM 1.15)
0.0
−0.5
−1.0
−1.5
200
400
600
800
1000
Mortality vs DO2 on admission

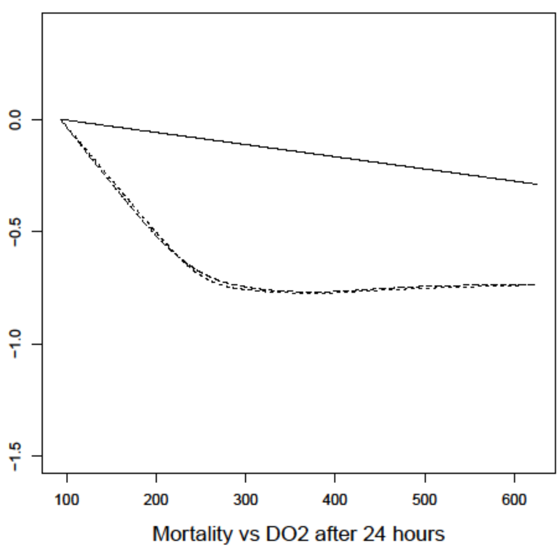

d.f. Chosen By AIC (GAM 7.20 GLMM 4.10)
0.0
−0.5
−1.0
−1.5
100
200
300
400
500
600
700
Mortality vs DO2 after 6 hours


d.f. Chosen By BIC (GAM 1.60 GLMM 1.45)
0.0
−0.5
−1.0
−1.5
100
200
300
400
500
600
700
Mortality vs DO2 after 6 hours

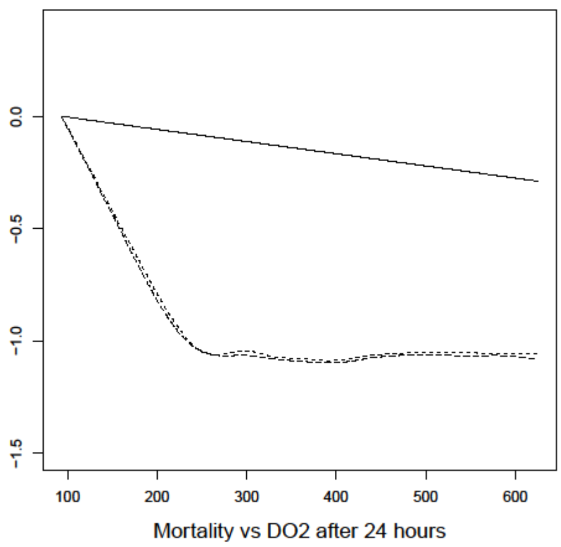

d.f. Chosen By GCV (GAM 10.00 GLMM 3.85)
0.0
−0.5
−1.0
−1.5
100
200
300
400
500
600
700
Mortality vs DO2 after 6 hours



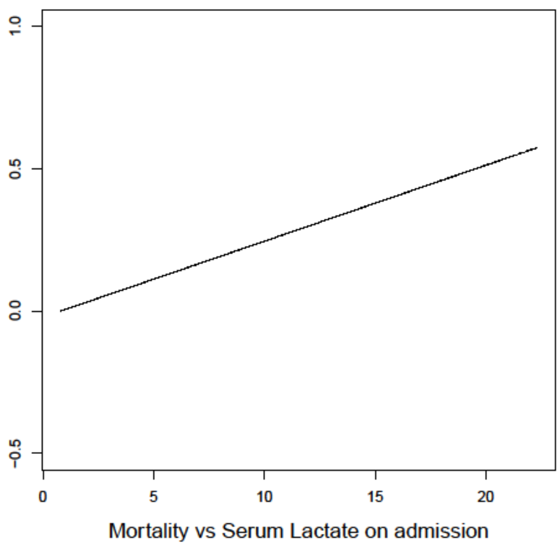
d.f. Chosen By AIC (GAM 4.90 GLMM 5.15)



d.f. Chosen By BIC (GAM 3.05 GLMM 1.55)

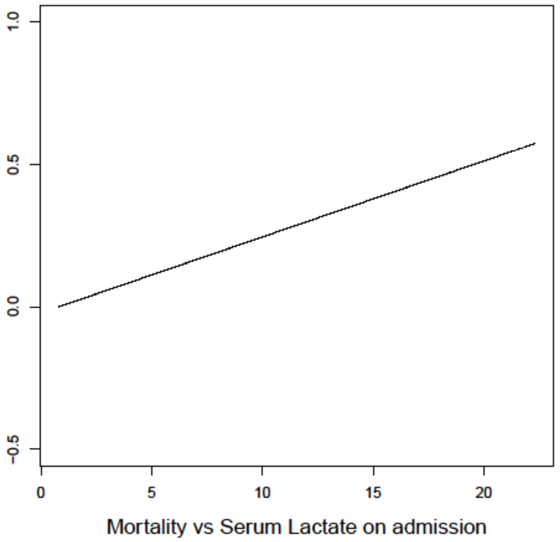


d.f. Chosen By GCV (GAM 7.10 GLMM 6.85)

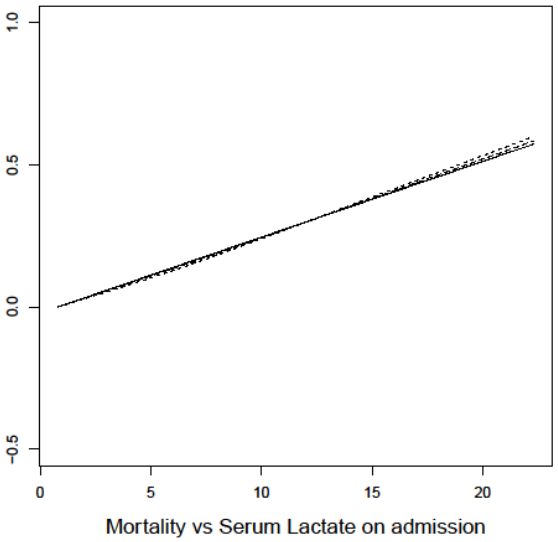


d.f. Chosen By AIC (GAM 1.00 GLMM 1.00)

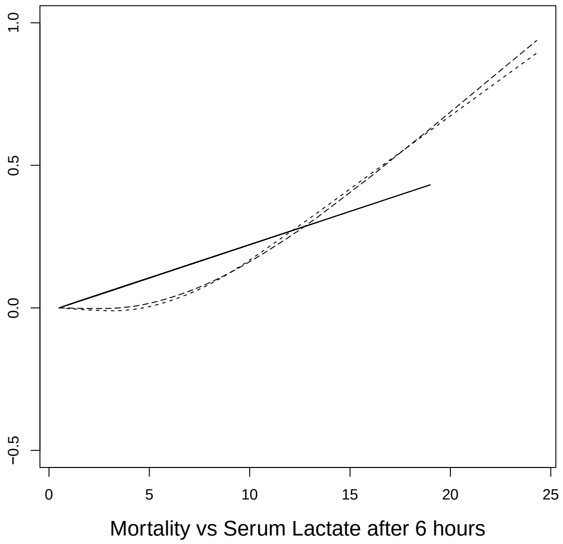
d.f. Chosen By BIC (GAM 1.00 GLMM 1.00)



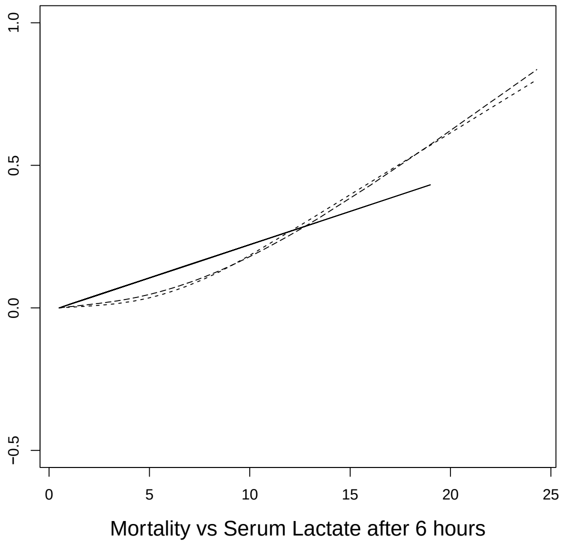
d.f. Chosen By GCV (GAM 1.10 GLMM 1.35)

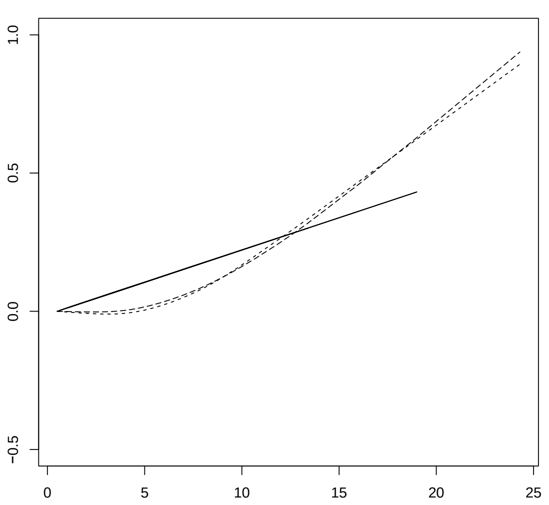


d.f. Chosen By AIC (GAM 1.25 GLMM 1.20)



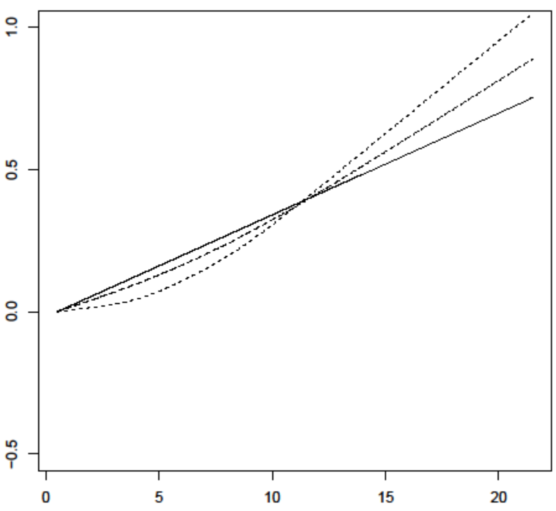
d.f. Chosen By BIC (GAM 1.20 GLMM 1.20)





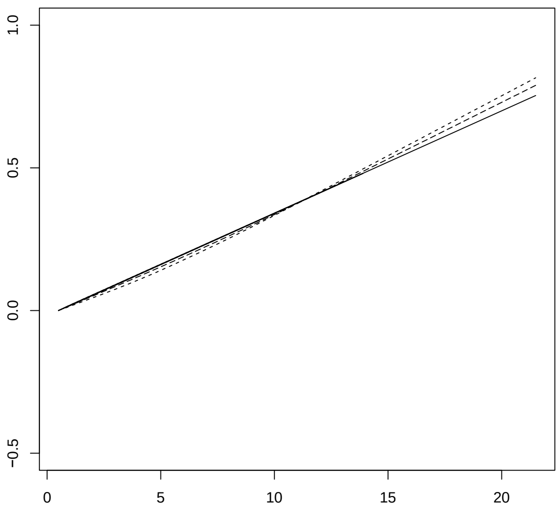
d.f. Chosen By GCV (GAM 1.45 GLMM 1.50)

Mortality vs Serum Lactate after 6 hours



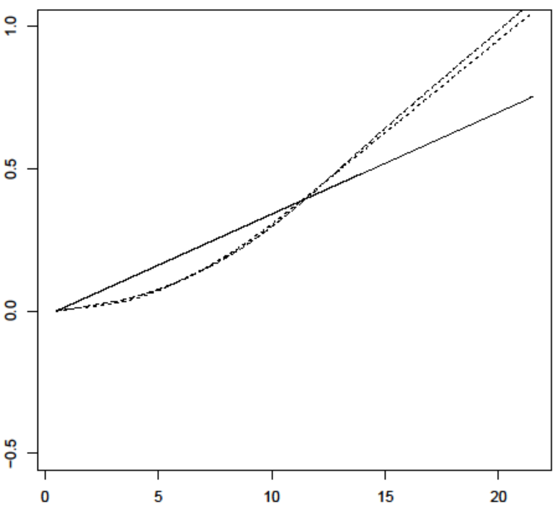
d.f. Chosen By AIC (GAM 1.4 GLMM 2.2)

Mortality vs Serum Lactate after 24 hours



d.f. Chosen By BIC (GAM 1.05 GLMM 1.05)

Mortality vs Serum Lactate after 24 hours



d.f. Chosen By GCV (GAM 5.00 GLMM 4.80)

Mortality vs Serum Lactate after 24 hours

Concluding Discussion

While the two manuscripts of this thesis had distinct objectives, in this section I review the results and findings from both. In the first manuscript I conducted a simulation study comparing GAMs and GLMMs in a series of comprehensive scenarios for continuous outcomes, which gave the real life data analyses some theoretical context. In this manuscript I also explored the relationship between smoking and lung function. In the second manuscript, I dealt with a larger and more practical real life application, in which log-linear modeling is an inadequate method for capturing curvilinear associations. Its objective was simpler: To evaluate each model's capacity to model a known nonlinear effect (the leveling off of morbidity risk in ICU patients).

The first manuscript concluded that GAMs had the slight advantage over LMMs in large datasets. LMMs were shown in some cases be more reliable in small datasets with relatively large amounts of variability. Again in the large datasets of the first manuscripts simulation study, edfs chosen by BIC with LMMs provided a consistent and relatively conservative assessment of the nonlinearity of the curve. In the reverse scenario with a high sample size to variance ratio (with less information), GAMs or LMMs with AIC/GCV used to choose the df were most likely to capture the full extent of the nonlinear behavior.

The curves that estimated via GAMs or GLMMs in the second manuscript gave roughly similar results that supported the existence of thresholds of increased ICU mortality known in the literature. However, the estimated curves for MVO_2 suggested increased mortality at significantly lower levels (closer to 40%). In general, GLMMs produced a sharper and more consistent threshold point than did GAMs. However, we did not formally estimate the threshold, and instead relied on visual identification.

Both the ICU and COLD data featured similar curves with similar character of nonlinearity. They both roughly approximated the "nonlinear" category of curves, from the simulated dataset of the first manuscript. However, the independent variable for the ICU data distribution was quite different from those used to generate data for the simulations presented in Manuscript 1. The independent variable was slightly sparser than the Smoking Duration variable and much less skewed than the Pack Year variance. There were also more observations in the suspected areas of nonlinearity. The one ICU dataset variable that lacked data in the location of nonlinearity had the weakest evidence for nonlinearity (Serum Lactate Table 4, Figures 28-36). In general for the real life data it seems that GLMMs are a better performer than GAMs. This is in contrast to the simulation. As mentioned in the methods of the first manuscript, the independent variable of the simulated data was generated from a uniform distribution while the smoking variables are heavily skewed towards zero with a small number of highly influential outliers. Such outliers increased the variance of the edf results in the simulation, but to what degree this would change edf is unclear.

This thesis has several strengths. Both Gaussian and non-Gaussian models were explored (with binomial outcomes being exceedingly common in epidemiological studies). A variety of methods commonly used in model selection were used to determine the degree of nonlinearity implemented (via the edf). Finally results were examined both graphically and analytically (via KL-distance, curve edf, and information criteria).

Overall, the simulation study described in the first manuscript gave a slight edge to GAMs in its ability to capture the extent nonlinear relations and return fitted values with a lower KL-distance from the true curves. Both sets of our models derived from the real life datasets, however, tended to favor GLMMs. Although the simulation study did not address modeling dichotomous outcomes, it seemed that GLMMs did a better job than GAMs, since the general form of the existing nonlinear relationship was well established.

There is room for further comparison between to the two classes of models; we focused on proper model selection as opposed to inference on the nonlinearity of the estimated curve. The direct inference method has seen mixed success[18]. Also, for each GAM and GLMM we only examined a single algorithm to fit the model, but several are commonly used in practice. It is possible that other algorithms may yield different results (REF to Engel B from 2nd manuscript). Finally we could examine flexible regression methods beyond GAMs and GLMMs entirely using methods discussed in the literature review.

Several overall conclusions can be drawn from both manuscripts. Despite GAMs and GLMMs both exceling in certain and district scenarios they were both appropriate for the task in almost all the examined situations. Thus if a researcher has other stronger considerations that heavily favor one type of model over another (e.g. mixed models for longitudinal data) then in general the non-optimal one may become the optimal. This is especially true with larger datasets where simulation results converged in many cases. Finally both the simulation and the real life analyses were constructed in a controlled fashion. More exotic considerations such as additional covariates (both linear and nonlinear) or selective usage of variables (for example having them determined by AIC) could have been employed. It is important to consider the results of this thesis in an “all things considered equal” setting.

The work contained in thesis can also be viewed beyond its primary purpose of a comparison and evaluation of GAMs and GLMMs in the context of nonlinear modeling. It is useful for those who wish to add these models to their research project. There is considerable variability on how both these models should be generated and for what is their exact usage and interpretation. A meta-analysis or survey on how GAMs and GLMMs are commonly implemented for nonlinear modeling could shed some light on how researchers commonly use these tools.

Future directions in the juxtaposition of GAMs and GLMMs may include a simulation and/or real life analysis of data that demands an alternative distribution (e.g. Poisson rates) or where an entirely nonparametric model must be employed. In the vein of expanding the scope of the simulation, more "types" of nonlinear and discontinuous scenarios could be examined. Similarly, the scenarios examined here could be made more realistic by adding other covariates which may or may not have a linear association with the outcome. Another possible method of evaluation is to examine the rates of type I and II errors through hypothesis testing.

This thesis has been able to provide a comparison of GAMs and GLMMs in a theoretical and comprehensive context, as well as two applied and applicable to medical research cases. The simulation has provided distinct scenarios where each GAM or GLMM may excel in accuracy and interpretability. The real life data sets have shown that these models can identify both established and exploratory nonlinear relations in clinical variables.

References

1. Hastie, T., Tibshirani, RJ., *Generalized Additive Models*, ed. C.a. Hall1990, New York.
2. Ruppert, D., M.P. Wand, and R.J. Carroll, *Semiparametric regression*2003: Cambridge University Press.
3. Nelder, J.A. and R.W.M. Wedderburn, *Generalized Linear Models.* Journal of the Royal Statistical Society. Series A (General), 1972. **135**(3): p. 370-384.
4. Nussbaum, M., *Rao, C. R., Linear statistical inference and its applications, second Edition, New York. John Wiley & Sons. 1973. XX, 625 S., £ 12.85.* ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik, 1977. **57**(8): p. 500-500.
5. Hardin, J.W., J.M. Hilbe, and J. Hilbe, *Generalized linear models and extensions*2007: Stata Press.
6. Friedman, J.H. and W. Stuetzle, *Projection Pursuit Regression.* Journal of the American Statistical Association, 1981. **76**(376): p. 817-823.
7. Buja, A., T. Hastie, and R. Tibshirani, *Linear Smoothers and Additive Models.* The Annals of Statistics, 1989. **17**(2): p. 453-510.
8. Diggle, P., *Analysis of longitudinal data*2002: Oxford University Press.
9. Ruppert, D., *Selecting the Number of Knots for Penalized Splines.* Journal of Computational and Graphical Statistics, 2002. **11**(4): p. 735-757.
10. Wand, M.P., *Smoothing and mixed models*, 2002, Department of Biostatistics, School of Public Health, Harvard University,: Boston, Massachusetts.
11. Venables, W.N. and C.M. Dichmont, *GLMs, GAMs and GLMMs: an overview of theory for applications in fisheries research.* Fisheries Research, 2004. **70**(2-3): p. 319-337.
12. Zuur, A.F., et al., *GLMM and GAMM*, in *Mixed effects models and extensions in ecology with R*2009, Springer New York. p. 323-341.
13. Ye, J., *On Measuring and Correcting the Effects of Data Mining and Model Selection.* Journal of the American Statistical Association, 1998. **93**(441): p. 120-131.
14. Cantoni, E. and T. Hastie, *Degrees-of-Freedom Tests for Smoothing Splines.* Biometrika, 2002. **89**(2): p. 251-263.
15. Hastie, T., R. Tibshirani, and J.H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*2009: Springer.
16. Wood, S.N., *Generalized Additive Models*, in *CRC*2006.
17. Diggle, P., et al. , *Analysis of Longitudinal Data 2nd*2002, Oxford, England: Oxford University Press.
18. Benedetti, A., *Using Generalized Additive Models to Detect and Estimate Threshold Associations.* The International Journal of Biostatistics, 2009. **5**.
19. Royston, P., D.G. Altman, and W. Sauerbrei, *Dichotomizing continuous predictors in multiple regression: a bad idea.* Statistics in Medicine, 2006. **25**(1): p. 127-141.
20. Myers, R.H., *Classical and modern regression with applications*1990: PWS-KENT.
21. McCullagh, P. and J.A. Nelder, *Generalized Linear Models*1989: Chapman and Hall.
22. I DiMatteo, C.R.G., R.E. KASS, *Bayesian curve-fitting with free-knot splines.* Biometrika, 2001. **88**(4): p. 1055-1071.
23. Denison, D.G.T., B.K. Mallick, and A.F.M. Smith, *Automatic Bayesian curve fitting.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 1998. **60**(2): p. 333-350.
24. Taylor, J.R., *An introduction to error analysis: the study of uncertainties in physical measurements*1997: University Science Books.

25. Aydin, D., Tuzemen, S., *A comparative study of the sum of squares and deviance in linear, additive and partial linear additive models.* J. Applied Sci, 2010. **10**: p. 919-929.
26. Devroye, L., L. Györfi, and G. Lugosi, *A probabilistic theory of pattern recognition*1996: Springer.
27. Abrahamowicz, M., Ciampi, A., *Information theoretic criteria in non-parametric density estimation: bias and variance in the infinite dimensional case.* Computational Statistics & Data Analysis 1991. **12**: p. 239-247.
28. Kullback, S., *Information theory and statistics* 1959, New York: John Wiley and Sons.
29. Shinto Eguchia, a.J.C., *Interpreting Kullback-Leibler divergence with the Neyman-Pearson lemma.* Journal of Multivariate Analysis, 2006(97): p. 2034-2040.
30. Burnham, K.P., Anderson, D. R., *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, Second Edition*. Springer Science 2002, New York.
31. Akaike, H., *A new look at the statistical model identification.* IEEE Transactions on Automatic Control 1974. **19**(6): p. 716-723.
32. Yang, Y., *Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation.* Biometrika, 2005. **92**(4): p. 937-950.
33. Schwarz, G., *Estimating the Dimension of a Model.* The Annals of Statistics, 1978. **6**(2): p. 461-464.
34. Kass, R.E. and A.E. Raftery, *Bayes Factors.* Journal of the American Statistical Association, 1995. **90**(430): p. 773-795.
35. Golub, G., M. Heath, and G. Wahba, *Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter.* Technometrics, 1979. **21**(2): p. 215-223.
36. Crawley, M.J., *The R book*2007: Wiley.
37. Fox, J., *An R and S-Plus companion to applied regression*2002: Sage Publications.
38. J. Fu, W., *Nonlinear GCV and quasi-GCV for shrinkage models.* Journal of Statistical Planning and Inference, 2005. **131**(2): p. 333-347.
39. Haber, E. and D. Oldenburg, *A GCV based method for nonlinear ill-posed problems.* Computational Geosciences, 2000. **4**(1): p. 41-63.
40. Li, K.-C., *From Stein's Unbiased Risk Estimates to the Method of Generalized Cross Validation.* The Annals of Statistics, 1985. **13**(4): p. 1352-1377.
41. Wood, S.N., *Stable and efficient multiple smoothing parameter estimation for generalized additive models.* Journal of the American Statistical Association, 2004. **99**(673-686).
42. Doll R, H.A.B., *Smoking and Carcinoma of the Lung.* British Medical Journal, 1950. **4682**: p. 739–748.
43. Doll R, H.A.B., *The mortality of doctors in relation to their smoking habits; a preliminary report.* British Medical Journal, 1954. **4877**: p. 1451-1455.
44. Hoffmann, T.D.K., *A Comprehensive Index for the Modeling of Smoking History in Periodontal Research.* Journal of Dental Research, 2004. **83**(859).
45. Neuner, B., et al., *Modeling Smoking History: A Comparison of Different Approaches in the MARS Study on Age-Related Maculopathy.* Annals of epidemiology, 2007. **17**(8): p. 615-621.
46. Leffondré, A.e.a., *Modeling Smoking History: A Comparison of Different Approaches.* Am J Epidemiol 2002, 2002. **156**: p. 813-823.
47. Cox, L.A., *A Causal Model of Chronic Obstructive Pulmonary Disease (COPD) Risk.* Risk Analysis, 2011. **31**(1): p. 38-62.
48. Kirkby, J. and I. Currie, *Smooth models of mortality with period shocks.* Statistical Modelling, 2010. **10**(2): p. 177-196.
49. Housh, T., et al., *The effect of mathematical modeling on critical velocity.* European Journal of Applied Physiology, 2001. **84**(5): p. 469-475.

50. Lim, N., et al., *Do All Nonsurvivors of Cardiogenic Shock Die With a Low Cardiac Index?*.* Chest, 2003. **124**(5): p. 1885-1891.

51. Caille, V. and P. Squara, *Oxygen uptake-to-delivery relationship: a way to assess adequate flow.* Critical Care, 2006. **10**(Suppl 3): p. S4.

52. Mizock, B.A. and J.L. Falk, *Lactic acidosis in critical illness.* Critical Care Medicine, 1992. **20**(1): p. 80-93.

53. Huang, Y.-C.T., *Monitoring Oxygen Delivery in the Critically Ill*.* Chest, 2005. **128**(5 suppl 2): p. 554S-560S.

54. Yu, M., et al., *Frequency of mortality and myocardial infarction during maximizing oxygen delivery: A prospective, randomized trial.* Critical Care Medicine, 1995. **23**(6): p. 1025-1032.

55. Nevill, A.M. and R.L. Holder, *Modelling Maximum Oxygen Uptake-A Case-Study in Non-Linear Regression Model Formulation and Comparison.* Journal of the Royal Statistical Society. Series C (Applied Statistics), 1994. **43**(4): p. 653-666.

56. Farrington, D.P. and R. Loeber, *Some benefits of dichotomization in psychiatric and criminological research.* Criminal Behaviour and Mental Health, 2000. **10**(2): p. 100-122.

57. Marino, P.L. and K.M. Sutin, *The ICU book*2007: Lippincott Williams & Wilkins.

58. Rivers, E., et al., *Early Goal-Directed Therapy in the Treatment of Severe Sepsis and Septic Shock.* New England Journal of Medicine, 2001. **345**(19): p. 1368-1377.

59. Ronco, J.J., et al., *Identification of the Critical Oxygen Delivery for Anaerobic Metabolism in Critically III Septic and Nonseptic Humans.* JAMA: The Journal of the American Medical Association, 1993. **270**(14): p. 1724-1730.

60. Scheinman, M.M., M.A. Brown, and E. RAPAPORT, *Critical Assessment of Use of Central Venous Oxygen Saturation as a Mirror of Mixed Venous Oxygen in Severely Ill Cardiac Patients.* Circulation, 1969. **40**(2): p. 165-172.

61. Konstantinov BA, S.E., Priimak VP, *Cardiac insufficiency and mortality after correction of the mitral valve defect, depending on the degree of cardiac output before the operation.* Kardiologiia, 1978. **18**(4): p. 113-9.

62. Drazner, M.H., et al., *Value of Clinician Assessment of Hemodynamics in Advanced Heart Failure / CLINICAL PERSPECTIVE.* Circulation: Heart Failure, 2008. **1**(3): p. 170-177.

63. Lieberman, J.A., et al., *Critical Oxygen Delivery in Conscious Humans Is Less Than 7.3 ml O2 · kg−1 · min−1.* Anesthesiology, 2000. **92**(2): p. 407.

64. Mikkelsen, M.E., et al., *Serum lactate is associated with mortality in severe sepsis independent of organ failure and shock *.* Critical Care Medicine, 2009. **37**(5): p. 1670-1677 10.1097/CCM.0b013e31819fcf68.

65. Shapiro, N.I., et al., *Serum Lactate as a Predictor of Mortality in Emergency Department Patients with Infection.* Annals of Emergency Medicine, 2005. **45**(5): p. 524-528.

66. Gideon, S., *Estimating the Dimension of a Model.* The Annals of Statistics, 1978. **6**(2): p. 461-464.

67. Benedetti A, A.M., Goldberg MS, *Accounting for data-dependent degree of freedom selection when testing the effect of a continuous covariate in generalized additive* Communications In Statistics, 2009. **38**(5): p. 1115-35.

68. Hastie, T., *gam: Generalized Additive Models*, T. Hastie, Editor 2010.

69. Roberto Rodriguez-Roisin, M., *Global Initiative for Chronic Obstructive Lung Disease. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease. Workshop Report*, 2001, U.S. Department of Health & Human Services, National Heart, Lung & Blood Institute.

70. Cox, L.A., *A Causal Model of Chronic Obstructive Pulmonary Disease (COPD) Risk.* Risk Analysis, 2011. **31**(1): p. 177-196.

71. Caporaso, J.H.L.a.N.E., *Cigarette Smoking and Lung Cancer: Modeling Total Exposure and Intensity.* Cancer Epidemiology, Biomarkers & Prevention, 2006. **51**(15).

72. Buist AS, V.W., Sullivan SD, Weiss KB, Lee TA, Menezes AM, Crapo RO, Jensen RL, Burney PG., *The Burden of Obstructive Lung Disease Initiative (BOLD): rationale and design.* COPD, 2005. **2**(2): p. 277-283.

73. Rabe KF, H.S., Anzueto A, Barnes PJ, Buist SA, Calverley P, Fukuchi Y, Jenkins C, Rodriguez-Roisin R, van WC, Zielinski J. , *Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary.* Am.J.Respir.Crit Care Med., 2007. **176**(6): p. 532-555.

74. Rahman, M.S. and M.L. King, *Improved model selection criterion.* Communications in Statistics - Simulation and Computation, 1999. **28**(1): p. 51-71.

75. Otero, R.M., et al., *Early Goal-Directed Therapy in Severe Sepsis and Septic Shock Revisited*.* Chest, 2006. **130**(5): p. 1579-1595.

76. Kandel, G. and A. Aberman, *Mixed Venous Oxygen Saturation: Its Role in the Assessment of the Critically Ill Patient.* Arch Intern Med, 1983. **143**(7): p. 1400-1402.

77. Williams, T.A., et al., *Effect of length of stay in intensive care unit on hospital and long-term mortality of critically ill adult patients.* British Journal of Anaesthesia, 2010. **104**(4): p. 459-464.

78. Vittinghoff, E., *Regression Methods In Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*2005: Springer.

79. Wahba, G., *Spline Models for Observatinal Data*1990, Madison, USA: SIAM. Chapter 4.

80. Gabrielle, F., et al., *Is the Parsonnet's score a good predictive score of mortality in adult cardiac surgery: assessment by a French multicentre study.* European Journal of Cardio-Thoracic Surgery, 1997. **11**(3): p. 406-414.

81. Hastie, T., *Generalized Additive Models*, 2011, CRAN.

82. Husain, F.A., et al., *Serum lactate and base deficit as predictors of mortality and morbidity.* The American Journal of Surgery, 2003. **185**(5): p. 485-491.

83. Brierley, J., et al., *2007 American College of Critical Care Medicine clinical practice parameters for hemodynamic support of pediatric and neonatal septic shock*.* Critical care medicine, 2008.

84. Engel, B., *A Simple Illustration of the Failure of PQL, IRREML and APHL as Approximate ML Methods for Mixed Models for Binary Data.* Biometrical Journal, 1998. **40**(2): p. 141-154.

85. Fan, J., X. Lin, and J.S. Liu, *New developments in biostatistics and bioinformatics*2009: Higher Education Press.

86. Eggermont, P.P.B. and V.N. LaRiccia, *Maximum Penalized Likelihood Estimation: Regression*2009: Springer.

87. Fang, Y., *Asymptotic Equivalence between Cross-Validations and Akaike Information Criteria in Mixed-Effects Models.* Journal of Data Science, 2011. **9**: p. 15-21.