DEPARTMENT OF STATISTICS
University of Wisconsin
1210 West Dayton St.
Madison, WI 53706

TECHNICAL REPORT NO. 956

December 21, 1995

Using Smoothing Spline ANOVA to Examine the Relation of Risk Factors to the Incidence and Progression of Diabetic Retinopathy

Yuedong Wang [1]
Department of Biostatistics, University of Michigan Ann Arbor MI

Grace Wahba [2]
Department of Statistics, University of Wisconsin, Madison WI

Chong Gu [3]
Department of Statistics, Purdue University, West Lafayette IN

Ronald Klein, MD[4] Barbara Klein, MD[5]
Department of Ophthalmalogy, University of Wisconsin,Madison WI

[1] yuedong@umich.edu Research supported in part by NIH Grant EY09446, P60 DK20572 and P30 HD18258

[2] wahba@stat.wisc.edu Research supported in part by NIH Grant EY09946 and NSF Grant DMS9121003

[3] chong@stat.lsa.umich.edu Research supported by DMS9301511

[4] kleinr@epi.ophth.wisc.edu Research supported by NIH Grant EY03083

[5] Research supported by NIH Grant EY03083

AMS 1991 subject classifications. Primary 62G07, 92C60, 68T05, 65D10, 62A99, 62J07; Secondary 41A63, 41A15,62G07,62M30,65D15,92H25.

Key words and phrases. Smoothing spline ANOVA, nonparametric regression, exponential families, risk factor estimation, diabetic retinopathy

Using Smoothing Spline ANOVA to Examine the Relation of Risk Factors to the Incidence and Progression of Diabetic Retinopathy

Yuedong Wang, Grace Wahba, Chong Gu,
Ronald Klein, MD and Barbara Klein, MD

December 21, 1995

Abstract

Smoothing Spline ANOVA (ANalysis Of VAriance) methods provide a flexible alternative to the standard parametric GLIM (Generalized Linear Models) methods for analyzing the relationship of predictor variables to outcomes from data from large epidemiologic studies. These methods allow the visualization of relationships which are not readily fit by simple GLIM models, and provide for the ability to visualize interactions between the variables, while at the same time they reduce to GLIM models if the data suggest that the added flexibility is not warranted. Using this method, we investigate risk factors for incidence and progression of diabetic retinopathy in a group of patients with older onset diabetes from the Wisconsin Epidemiological Study of Diabetic Retinopathy. We carry out four analyses illustrating various properties of this class of methods. Some of the results confirm what has been previously found using standard methods, while others allow the visualization of more complex relationships not evident from the application of parametric methods.

1 INTRODUCTION

In many demographic medical studies, data of the form $\{y_i, t(i), i = 1, \cdots n\}$ are collected, where i indexes the ith study participant, y_i is 1 or 0 indicating whether some medical condition of interest at followup is present or absent, and $t(i) = (t_1(i), \cdots, t_d(i))$ is a vector of d predictor variables at baseline, which may or may not be related to the likelihood that y_i is a 1. From this data it is desired to estimate $p(t) = Prob\{y = 1|t\}$, the probability that a person with predictor vector t will present the condition of interest at followup. Such estimates are used to estimate the prevalence of the condition of interest in general populations and to study the sensitivity of p to the predictor variables, or, if possible, to combinations of them. The traditional GLIM models [1] do this by defining $f(t)$, the logit, as

$$f(t) = log[p(t)/(1 - p(t))] \tag{1.1}$$

and assuming that f is a simple parametric function of the components of t. When the t_α are continuous variables, the most commonly used model is linear in the components of t,

$$f(t) = f(t_1, \cdots, t_d) = \mu + \sum_{\alpha=1}^{d} \beta_\alpha t_\alpha, \tag{1.2}$$

but sometimes second or even third degree polynomials are used if it appears that a linear model may not be adequate. If some of the t_α are categorical variables, then indicator functions can be used. More generally, given M parametric functions $\phi_\nu, \nu = 1, \cdots, M$ subject to some identifiability conditions, f is modeled parametrically as

$$f(t) = \sum_{\nu=1}^{M} \beta_\nu \phi_\nu(t). \tag{1.3}$$

Then (μ and) the β's are obtained by minimizing the negative log likelihood $\mathcal{L}(y, f)$ given by

$$\mathcal{L}(y, f) = \sum_{i=1}^{n} [y_i f(t(i)) - log(1 + e^{f(t(i))})] \tag{1.4}$$

with (1.2) or, more generally, (1.3), substituted in for f. If the 'true but unknown' f is actually some linear combination of the specified ϕ_ν then this is, of course the proper thing to do. Unfortunately, as large data sets are examined more closely, it becomes clear that linear models, or even quadratic or cubic models, may not be adequate. What happens if, for example, the dependence on t_α is "J" shaped, or, even has two peaks, possibly representing two distinct subpopulations? If the 'true' log odds ratio f is not well approximated by some function of the specified form, then possibly large biases may be introduced in the estimates of f, furthermore, the common statistical hypothesis tests, confidence intervals, P-values and so forth are not necessarily valid if f is not of the specified form. The purpose of this paper is to describe and demonstrate the use of smoothing spline ANOVA (SS-ANOVA) methods for estimating f. These methods get around many of the above difficulties, while providing the ability to visualize some of the relationships between the variables not easily observed via the use of more traditional methods. We will demonstrate their use to examine the relation of risk factors to the incidence and progression of diabetic retinopathy, using data from the Wisconsin Epidemiological Study of Diabetic Retinopathy (WESDR).

The analysis here is based on (a special case of) the smoothing spline ANOVA method for modeling and estimating f which appears in Wahba, Wang, Gu, Klein and Klein [2](WWGKK), and we have implemented the analysis via the publicly available code GRKPACK described in Wang [3]. The code itself is available from `netlib@research.att.com` in the `gcv` directory there, and through `statlib@lib.stat.cmu.edu`.

An SS-ANOVA estimate is a form of penalized likelihood estimate. We first note two important penalized likelihood estimates that appear in the literature. Then in the remainder of this introduction we describe the main features of SS-ANOVA in general and the details of the SS-ANOVA models in particular that we will be applying to the WESDR data.

Major precursors of the work under discussion include O'Sullivan [4] and and O'Sullivan, Yandell and Raynor [5], who proposed a penalized log likelihood estimate for f based on thin plate splines. These splines are useful in many contexts, and can be incorporated in an SS-ANOVA model [6][7], but are not employed in the present work.

Hastie and Tibshirani [8] (see also other references there) discussed estimates of f of the form

$$f(t)=f(t_1,\cdots,t_d)=\mu+\sum_{\alpha=1}^{d}f_\alpha(t_\alpha) \tag{1.5}$$

where the f_α are 'smooth' functions obtained by some smoothing process, including the use of cubic smoothing splines. The S code (Chambers and Hastie [9]) provides the facility for fitting models of the form (1.5). They also note that some of their work extends to the SS-ANOVA methods that we will be considering here. Cubic smoothing splines for the f_α in (1.5) are the solution to the minimization problem: Find f_α (in an appropriate function space), and subject to some identifiability criteria such as $\int f_\alpha(t_\alpha)dt_\alpha=0$, to minimize the penalized log likelihood functional

$$\mathcal{I}_\lambda(y,f)=\mathcal{L}(y,f)+\sum_{\alpha=1}^{d}\lambda_\alpha J_\alpha(f_\alpha), \tag{1.6}$$

where the penalty functionals J_α are defined by

$$J_\alpha(f_\alpha)=\int_0^1(f''_\alpha(t_\alpha))^2dt_\alpha. \tag{1.7}$$

As the smoothing parameter λ_α tends to infinity, f_α tends to a linear function in t_α, so that the minimizer of (1.6) will tend to the linear function as in (1.2) if all the λ_α's become large. Penalized likelihood methods with more general penalty functionals and unpenalized terms are discussed in [10].

Recent research has focussed on more general models for f which allow the explicit modeling and visualization of possible interactions between variables, via functional analysis of variance decompositions (to be described) and SS-ANOVA estimation methods. Given a fairly arbitrary function $f(t_1,\cdots,t_d)$ of several variables, a (functional) ANOVA decomposition of f may be defined (generalizing ideas from parametric ANOVA) as

$$f(t_1,\cdots,t_d)=\mu+\sum_{\alpha=1}^{d}f_\alpha(t_\alpha)+\sum_{\alpha<\beta}f_{\alpha\beta}(t_\alpha,t_\beta)+\cdots+f_{1,\cdots,d}(t_1,\cdots,t_d), \tag{1.8}$$

where the f_α are the main effects, $f_{\alpha\beta}$ are the two factor interactions, and so on. The components are uniquely determined given a set of averaging operators $\mathcal{E}_\alpha$, which average *functions* over the t_α in some specified way. For example, if t_α is a continuous variable in the interval $[0,1]$, then a possible choice for $\mathcal{E}_\alpha$ is

$$(\mathcal{E}_\alpha f)(t_1,\cdots,t_{\alpha-1},t_{\alpha+1},\cdots,t_d)=\int_0^1 f(t_1,\cdots,t_d)dt_\alpha. \tag{1.9}$$

Other averaging operators include weighted sums over possible or observed values of t_α. Then the mean μ is given by $\mu=\prod_{\alpha=1}^{d}\mathcal{E}_\alpha f$, the main effects are given by $f_\alpha=(I-\mathcal{E}_\alpha)\prod_{\beta\neq\alpha}\mathcal{E}_\beta f$, the two factor interactions are given by $f_{\alpha\beta}=(I-\mathcal{E}_\alpha)(I-\mathcal{E}_\beta)\prod_{\gamma\neq\alpha,\beta}\mathcal{E}_\gamma f$, and so forth.

Since $\mathcal{E}_\alpha \mathcal{E}_\alpha = \mathcal{E}_\alpha$, it can be seen that the terms in the functional ANOVA decomposition satisfy side conditions analogous to those in ordinary parametric ANOVA, for example, $\mathcal{E}_\alpha f_\alpha = 0$. In general, for model fitting purposes higher order terms in the functional ANOVA decomposition are eliminated and some or all of the lower order terms are estimated by finding μ, and (some of) the f_α, $f_{\alpha\beta}$, etc. to minimize a penalized log likelihood functional $\mathcal{I}_\lambda(y, f)$ given by an expression which generalizes (1.6), with a separate penalty functional for each independently smoothed term in the model. Indicator functions for categorical variables may also be incorporated in the model.

It is well known [10] that solutions to variational problems like the minimizer of $\mathcal{I}_\lambda(y, f)$ and generalizations to $\mathcal{I}_\lambda(y, f)$ with f containing interaction terms as in (1.8) have a representation as a linear combination of the unpenalized terms in the model plus n (basis) functions, which can be constructed from the $t(i)$, the reproducing kernels associated with the penalty functionals, and the smoothing parameters, see [11] [12] [4] [13][10] [3]. Given the smoothing parameters, the numerical problem of computing the minimizer f_λ can be reduced to finding the coefficients of a representation for it as a linear combination of the unpenalized terms and the above mentioned basis functions. The numerical problem for the coefficients may be solved by a Newton-Raphson iteration in conjunction with matrix decompositions. The history of these numerical methods includes [14][15][16][17][4] [5][10]. Various approximate numerical methods for data sets that are too large for matrix decomposition methods are available and under development, but are beyond the scope of this article.

Objective methods for choosing the smoothing parameters are desirable. This paper and the code GRKPACK employ the iterative unbiased risk method given in Gu [14][15] for the non-Gaussian case and extended to the p smoothing parameter case $\lambda = (\lambda_1, \cdots, \lambda_p)$ in Wang [18] and WWGKK. This method is a computable proxy for the Kullback-Liebler information distance from the estimate to the 'truth', see [15][19] [2][20]. Thus, the method is attempting to choose the smoothing parameters to minimize the Kullback-Liebler information distance from the estimate f_λ and the unknown 'true' f. Other related references are [21][22][23].

As with any estimation method, it is important to have some indication of the accuracy of the method. It is particularly important in the case here of nonparametric regression with Bernoulli data obtained from medical records because this kind of data tends to be irregular, and may contain influential outliers. The rigidity of parametric models may mask the effect of outliers, as well as obscure the fact that inferences are sometimes made in fairly data sparse regions. The nonparametric estimates may be more sensitive to outliers, and it is important to be able to delineate a region in predictor variable space in which the estimate may be relied upon, as well as provide some sort of confidence statement. In this work we use the Bayesian 'confidence intervals' proposed in [24], adapted to the component-wise multiple smoothing parameter case in [7], to non-Gaussian data [25], and to the multiple smoothing parameter non-Gaussian case in [18] and WWGKK. The Bayesian 'confidence interval' for f_λ is used to delineate the region in predictor variable space for which the overall estimate is deemed to be reliable, by computing an appropriate level curve in a contour plot of the width of the confidence interval, and using this to enclose a 'reliable' region. It is also used in conjunction with cross sectional plots so that the accuracy estimates can be visualized. The component-wise confidence intervals have been used to eliminate some terms which cannot be distinguished from noise, by deleting terms whose confidence intervals contain 0 over most of their domain. We remark that these confidence intervals are based on an across-the-function property - that is, a 95% confidence interval has the property that, (approximately) the confidence intervals are expected to cover the true curve at about 95% of the n data points, see [25][3] [20].

We now proceed to the details of the particular models that we will be employing in the analysis of the WESDR data. We will be analyzing four data sets, each of which illustrates some particular feature of this class of models. In each case the (continuous) predictor variables $t_1, \cdots, t_d$ have been rescaled to $[0,1]$, and we will use the averaging operator $\mathcal{E} = \mathcal{E}_\alpha$ defined in (1.9). We will only consider main effects and two factor interactions. We will employ the penalty functional J_α of (1.7) and a two dimensional relative $J_{\alpha\beta}$ to be defined below. If all of the two factor interactions were eliminated, then the models described immediately below would reduce to the form (1.5) as studied by Hastie and Tibshirani, although our estimation method is different.

We need to make some definitions in order to define the terms in the particular ANOVA decomposition (1.8) that we will be using. Define the (linear) 'trend' function $\phi(u) = u - 1/2$ for $u \in [0,1]$, and, for continuous functions g defined on $[0,1]$ let $\mathcal{B}g$ stand for $\mathcal{B}g = g(1) - g(0)$. For future reference note that $\mathcal{E}\phi \equiv \int_0^1 \phi(u)du = 0$ and $\mathcal{B}\phi = 1$. $\mathcal{E}$ and $\mathcal{B}$ will be used to define ANOVA and identifiability side conditions. A subscript α on $\mathcal{E}$ or $\mathcal{B}$ means that it applies to what follows considered as a function of t_α. The typical main effect term $f_\alpha(t_\alpha)$ in this setup will have a decomposition of the form

$$f_\alpha(t_\alpha) = d_\alpha \phi(t_\alpha) + f_{s\alpha}(t_\alpha) \tag{1.10}$$

where $d_\alpha\phi(t_\alpha)$ is the linear and unpenalized part of f_α and $f_{s\alpha}$ is the (detrended) 'smooth' part of f_α. $f_{s\alpha}$ satisfies the side conditions $\mathcal{E}_\alpha f_{s\alpha} = \mathcal{B}_\alpha f_{s\alpha} = 0$ and will appear inside a penalty functional as $J_\alpha(f_{s\alpha})$. Due to the side conditions on f_α, $J_\alpha(f_{s\alpha}) = 0$ implies that $f_{s\alpha} = 0$.

The two factor interaction $f_{\alpha\beta}(t_\alpha, t_\beta)$ has an analogous decomposition into four interaction terms, namely

$$f_{\alpha\beta}(t_\alpha, t_\beta) = d_{\alpha\beta}\phi(t_\alpha)\phi(t_\beta) + \phi(t_\alpha) f_{s\beta}^{(\alpha)}(t_\beta) + \phi(t_\beta) f_{s\alpha}^{(\beta)}(t_\alpha) + f_{s\alpha\beta}(t_\alpha, t_\beta) \tag{1.11}$$

where the 'smooth' factors of the trend × smooth interactions satisfy the side conditions $\mathcal{E}_\alpha f_{s\alpha}^{(\beta)} = \mathcal{B}_\alpha f_{s\alpha}^{(\beta)} = \mathcal{E}_\beta f_{s\beta}^{(\alpha)} = \mathcal{B}_\beta f_{s\beta}^{(\alpha)} = 0$ and the 'smooth-smooth' interaction term satisfies $(\mathcal{E}_\alpha f_{s\alpha\beta})(t_\beta) = (\mathcal{B}_\alpha f_{s\alpha\beta})(t_\beta) = (\mathcal{E}_\beta f_{s\alpha\beta})(t_\alpha) = (\mathcal{B}_\beta f_{s\alpha\beta})(t_\alpha) = 0$, for all t_α, t_β. Letting $J_{\alpha\beta}(f_{s\alpha\beta}) = \int_0^1\int_0^1 (\frac{\partial^4}{\partial t_\alpha^2 \partial t_\beta^2} f_{s\alpha\beta}(t_\alpha, t_\beta))^2 dt_\alpha dt_\beta$, then it can be shown that the side conditions insure that $J_{\alpha\beta}(f_{s\alpha\beta}) = 0$ implies that $f_{s\alpha\beta} = 0$. Letting $\lambda = (\lambda_\alpha, \lambda_\beta, \lambda_\alpha^{(\beta)}, \lambda_\beta^{(\alpha)}, \lambda_{\alpha\beta})$ and

$$\mathcal{J}_\lambda(f) = \lambda_\alpha J_\alpha(f_{s\alpha}) + \lambda_\beta J_\beta(f_{s\beta}) + \lambda_\alpha^{(\beta)} J_\alpha(f_{s\alpha}^{(\beta)}) + \lambda_\beta^{(\alpha)} J_\beta(f_{s\beta}^{(\alpha)}) + \lambda_{\alpha\beta} J_{\alpha\beta}(f_{s\alpha\beta}), \tag{1.12}$$

f is estimated by finding $\mu, d_\alpha, d_\beta, d_{\alpha\beta}$, and $f_{s\alpha}, f_{s\beta}, f_{s\alpha}^{(\beta)}, f_{s\beta}^{(\alpha)}$,and $f_{s\alpha\beta}$ to minimize

$$\mathcal{I}_\lambda(y, f) = \mathcal{L}_\lambda(y, f) + \mathcal{J}_\lambda(f). \tag{1.13}$$

Due to the side conditions, as a component of λ becomes large, then the estimate of the 'smooth' term which it multiplies will become small. Thus, with a good method for choosing the components of λ from the data, unneeded terms in the expansion may effectively be eliminated if their companion smoothing parameters are estimated as large.

Categorical variables may be included in the model, by, for example, letting

$$f(t_\alpha, t_\beta, z) = \mu + f_\alpha(t_\alpha) + f_\beta(t_\beta) + f_{\alpha\beta}(t_{\alpha,\beta}) + \sum_{k=2}^{K} \gamma_k I_k(z) \tag{1.14}$$

where z is a variable with K possible values $z_1, \cdots, z_K$ and $I_k(z) = 1$ if $z = z_k$ and 0 otherwise. The γ's are then included in the minimization of (1.13).

The mathematics behind the representation of f, the numerical method for the minimization of $\mathcal{I}_\lambda$, the data-based choice of λ according to the iterative unbiased risk method, and the calculation of the confidence intervals is described in WWGKK. Further details are given in [3] and the documentation for GRKPACK. In the four data sets that we analyze below, there are between $d = 2$ and $d = 4$ predictor variables. In the $d = 4$ case, considering the penalty functional in (1.12), including all of the possible main effects and two factor interaction terms with their own smoothing parameters would result in 4 main effects smoothing parameters (λ_α), 12 trend × smooth smoothing parameters ($\lambda_\alpha^{(\beta)}$) and 6 smooth × smooth parameters ($\lambda_{\alpha\beta}$), more than we would want to simultaneously fit with the sample sizes here (all less than 500). The number of terms with smoothing parameters was reduced to a maximum of 7 based on previously published analyses of this data by more traditional methods, by extensive pre-screening by fitting parametric polynomial GLIM models with terms as high as cubic order to detect and delete terms which did not give any evidence of significance, and by fitting some marginal SS-ANOVA models as demonstrated in Section 6. More systematic pre-screening methods are an area of active research. Once the number of smoothing parameters was reduced to 7 or less, informal model selection methods as discussed in WWGKK were used, including the deletion of component terms which were too small to have an observable effect on cross-sectional plots of f, and the deletion of terms whose component-wise confidence intervals included the 0 function. Further details specific to each data set are given below. Leaving-out-one-third model selection procedures were discussed in [19], but have not been used here.

Section 2 discusses the Wisconsin Epidemiologic Study of Diabetic Retinopathy, and Sections 3, 4, 5 and 6 discuss the analysis of four data sets from this study. Section 7 is a summary and conclusion.

2 WISCONSIN EPIDEMIOLOGIC STUDY OF DIABETIC RETINOPATHY (WESDR)

The WESDR is an ongoing epidemiological study of a cohort of patients receiving their medical care in an 11-county area in southern Wisconsin, who were first examined in 1980-82, then again in 1984-86 and 1990-92. A third followup is currently in progress. Detailed descriptions of the data have been given in [26] and [27] and references there. In brief, a sample of 2990 diabetic patients was selected in an 11-county area in southern Wisconsin. This sample was composed of two groups. The first group consisted of 1210 patients diagnosed as having diabetes before 30 years of age and who took insulin ("younger onset group"). The second group consisted of 1780 patients who had the diagnosis of diabetes made at 30 years of age or older. Of these, 824 were taking insulin ("older onset group taking insulin") and 956 were not ("older onset grout not taking insulin "). Of the 2990 eligible patients, 2366 participated in the baseline examination from 1980 to 1982.

A large number of medical, demographic, ocular and other covariates were recorded at the baseline and later examinations along with a retinopathy score for each eye (to be described). Relations between various of the covariates and the retinopathy scores have been extensively analyzed by standard statistical methods including categorical data analysis and parametric GLIM models, and the results reported in a series of WESDR manuscripts. See [28][29][30][31][32][33][34]. SS-ANOVA methods were applied to a subset of the data from the younger onset group in WWGKK in conjunction with an extensive technical account of the mathematical theory and numerical methods behind the method. It is the purpose of this account to carry out an SS-ANOVA analysis for four data sets from the older onset WESDR group,and in the process to illustrate some of the more important features of the method, with the goal of explaining the possible results, advantages and disadvantages in a less technical manner, aimed at an clinicians and medical researchers.

The present study is limited to building predictive models for incidence and for progression (to be defined) of diabetic retinopathy at the first followup, as a function of some of the covariates available at baseline. We will do this both for the "older onset group taking insulin" (ID group) and the "older onset group not taking insulin " (NID group). We will analyze for an 'incidence event' and a 'progression event' for both the ID and the NID groups. We only list the covariates pertinent to our analysis:

1. `Age`: age at the examination (years);

2. `Duration`: duration of diabetes at the examination (years);

3. `Glycosylated hemoglobin`: a measure of hyperglycemia (%)[26];

4. Body mass index (`bmi`): weight in kg / (height in m)^2;

5. `Pulse` rate counted for 30 seconds;

6. Baseline retinopathy severity levels (`base-retinopathy-level`). See explanation below.

At the baseline and follow-up examinations, each eye was graded as one of the 6 levels: 10, 21, 31, 41, 51 and 60+, in order of increasing retinopathy severity with 10 indicating no retinopathy and 60+ indicating the most severe stage, proliferative retinopathy. The retinopathy level for a participant was derived by giving the eye with the higher level (more severe retinopathy) greater weight. For example, the level for a participant with level 31 retinopathy in each eye is specified by the notation "level 31/31", whereas that for a participant with level 31 in one eye and less severe retinopathy in the other eye is noted as "level 31/<31". This scheme provided an 11-step scale: 10/10, 21/<21, 21/21, 31/<31, 31/31, 41/<41, 41/41, 51/<51, 51/51, 60+/<60+ and 60+/60+. Participants in the analysis for *incidence* consisted of subjects with level 10/10 at the baseline and no missing data, and the model will provide an estimate of the probability of incidence, which is defined as the probability that such a participant has level 21/<21 or worse at the follow-up examination. Participants in the analysis for *progression* consisted of subjects with no or non-proliferative retinopathy at baseline and no missing data, and the model will provide an estimate of the probability of progression of such a participant, where progression is defined as an increase in baseline level by two steps or more (10/10 to 21/21 or greater, or 21/<21 to 31/<31 or greater, for instance). These analyses correspond to analyses previously carried out by traditional methods in some of the WESDR references cited. Note that this allows subjects who are included in the analysis for incidence to also be part of the analysis for progression.

3 INCIDENCE IN THE OLDER ONSET GROUP NOT TAKING INSULIN

After excluding participants with missing values, there were 297 participants in the older onset NID group with a baseline score of 10/10, thus qualifying them for inclusion in the NID 'Incidence' analysis. One observation of `glycosylated hemoglobin` was recorded as 23.6%, which is much greater than the others. We decided to delete this influential observation. Our conclusions remain the same with this observation in the analysis.

Klein *et al* [26] found, using GLIM, that `glycosylated hemoglobin` is the only significant predictor of incidence of retinopathy in older onset patients. Using GLIM as a screening tool, we found that the effect of `age` is not linear and that there is a strong interaction between `age` and `glycosylated hemoglobin`.

We first fit an SS-ANOVA model with the main effects of `age` and `glycosylated hemoglobin` and all three interaction terms. The main effect of `glycosylated hemoglobin` is linear. All interaction terms except *trend*(`glycosylated hemoglobin`) × *smooth*(`age`) are near zero. The final model is

$$\begin{aligned} & f(\texttt{age, glycosylated hemoglobin}) \\ = \; & \mu + f_1(\texttt{age}) + a_1 \times \texttt{glycosylated hemoglobin} + \\ & \mathit{trend}(\texttt{glycosylated hemoglobin}) \times \mathit{smooth}(\texttt{age}). \end{aligned} \tag{3.1}$$

We plot `age` vs `glycosylated hemoglobin` on left of Figure 3.1. Those participants who had an incidence of retinopathy are marked as solid circles and those with no incidence are marked as open circles. We superimpose the contour lines of estimated posterior standard deviations. These contours agree well with the distribution of the observations. Thus they can be used to delineate a region in which the estimate of the probability function is deemed to be reliable. We decided to use the region with estimated posterior standard deviations less then or equal to 1, in the logit scale.

The probability function estimate is plotted on right of Figure 3.1. Figure 3.2 gives cross sections of the estimated probability of incidence as a function of age with the Bayesian confidence intervals at the cross sections, at four quantiles of `glycosylated hemoglobin`. The width of the confidence intervals suggests that the small bumps are probably an artifact, however the general shape of the response is evident, and it appears that the age effect might not be particularly well modeled by a second or even third degree polynomial.

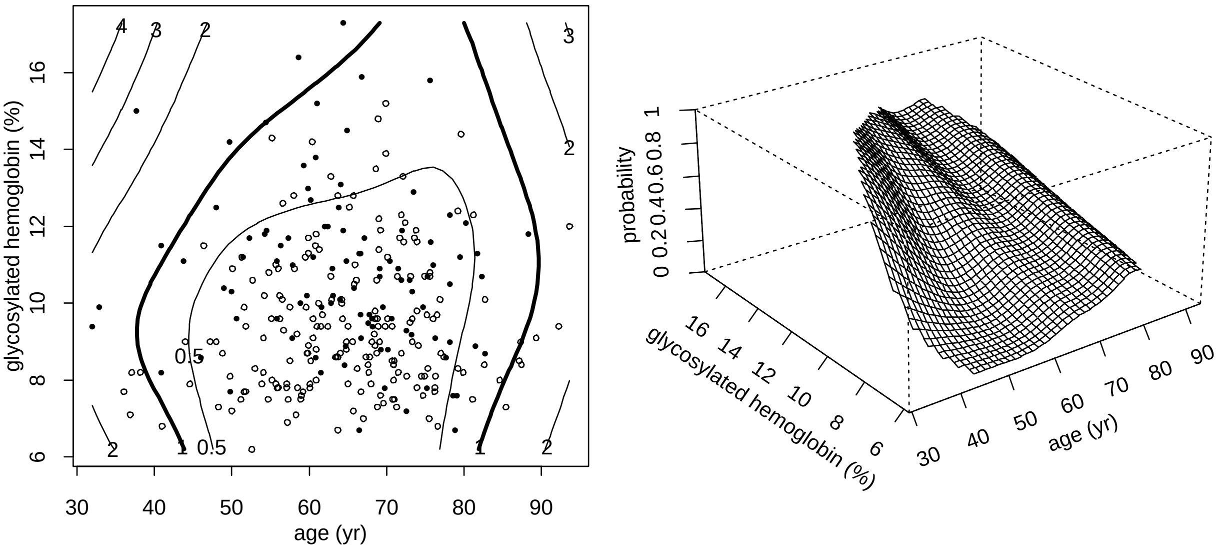


Figure 3.1: NID 'Incidence' analysis. Left: data and contours of constant posterior standard deviation. Solid circles indicate incidence and open circles indicate non-incidence. Right: estimated probability of incidence in the defined region, as a function of `age` and `glycosylated hemoglobin`.

The risk for incidence of retinopathy increases with increasing glycosylated hemoglobin. The risk increases with increasing age for lower glycosylated hemoglobin. This increase might be a result of development of other conditions such as hypertension and atherosclerotic vascular disease in older compared to younger subjects that lead to an increased risk of developing retinopathy even when the glycosylated hemoglobin level is relatively low. The risk decreases with increasing age for higher glycosylated hemoglobin. This decrease may be caused by mortality since an old participant with high glycosylated hemoglobin is more likely to die during the observation period.

4 PROGRESSION OF THE OLDER ONSET GROUP NOT TAKING INSULIN

After excluding participants with missing values, there were 432 participants in the older onset NID group qualified to be included in the analysis for 'Progression'. Klein *et al* [26] found that `age`, `duration` and `glycosylated hemoglobin` were significant in a GLIM model. Using GLIM, we found that the `duration` main effect of polynomials up to the cubic is significant and the `bmi` main effect of polynomials up to the quartic is significant. We also found that the multiplication interaction between `age` and polynomials up to cubic of `duration` is significant. These indicate that a GLIM with lower order polynomials may not

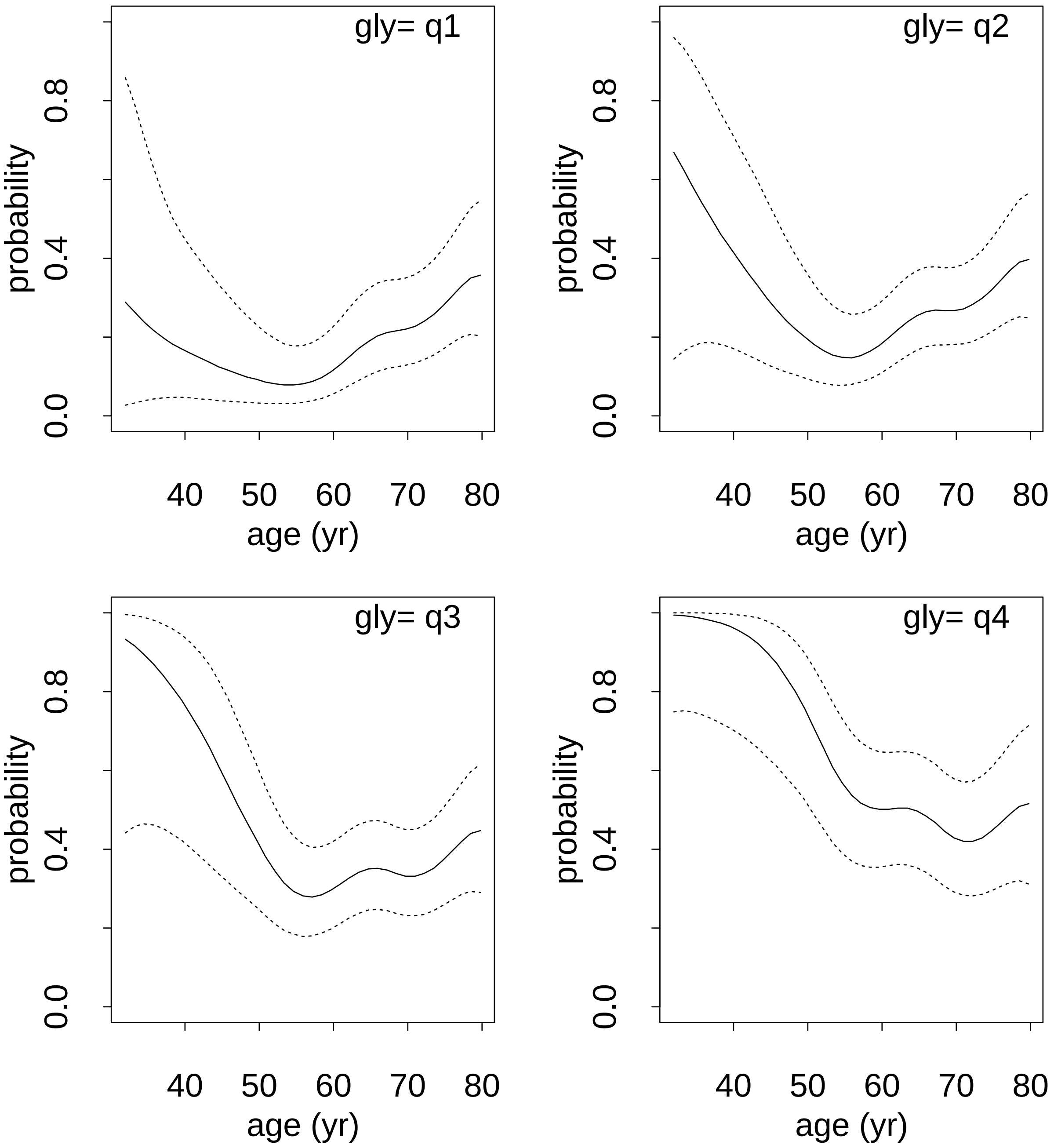


Figure 3.2: NID 'Incidence' analysis. Cross sections of estimated probability of incidence as a function of `age`, with Bayesian confidence intervals, at four quantiles of `glycosylated hemoglobin`.

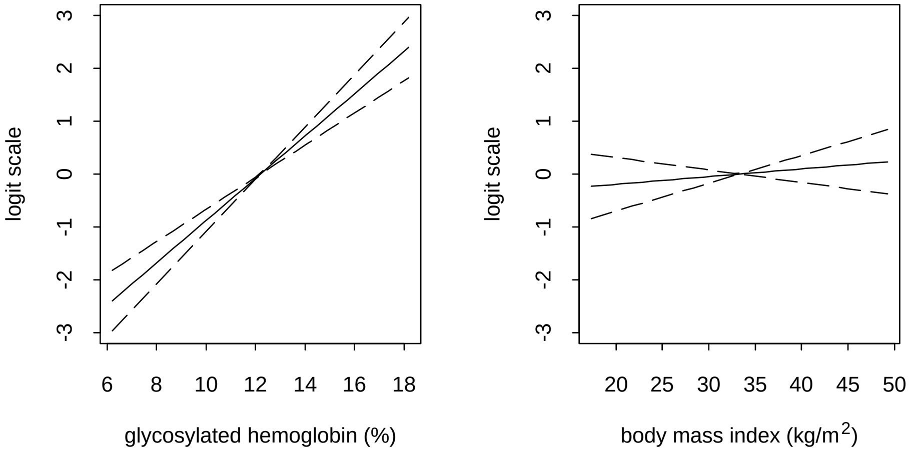


Figure 4.1: NID ‘Progression’ analysis. Main effects estimates of `glycosylated hemoglobin` and `bmi` along with their component-wise Bayesian confidence intervals, logit scale.

be good enough. We decided to fit the SS-ANOVA model:

$$\begin{aligned} & f(\texttt{age, duration, glycosylated hemoglobin, bmi}) \\ = \; & \mu + f_1(\texttt{age}) + f_2(\texttt{duration}) + f_{1,2}(\texttt{age}, \texttt{duration}) + \\ & f_3(\texttt{glycosylated hemoglobin}) + f_4(\texttt{bmi}). \end{aligned} \tag{4.1}$$

The main effects estimates and their Bayesian confidence intervals for `glycosylated hemoglobin` and `bmi` are plotted in Figure 4.1. We see that the effect of `glycosylated hemoglobin` is strong. The effect of `bmi` is small, and 0 is contained within its confidence interval. The estimates are effectively linear on a logit scale even though a ‘smooth’ term for each of them was allowed in the model. This illustrates the ability of the method to reduce to a partially parametric form if that is warranted by the data.

The effects of `glycosylated hemoglobin` and `bmi` are additive in the logit scale. In the remaining plots, we fix `glycosylated hemoglobin` and `bmi` at their median values.

We plot `age` vs `duration` on the left of Figure 4.2. Those participants who had progression of retinopathy are marked as solid circles and those with no progression are marked as open circles. We superimpose contour lines of the estimated posterior standard deviations as a function of `age` and `duration` with `glycosylated hemoglobin` and `bmi` fixed at their median values. These contours agree well with the distribution of the observations. We decided to use the region with estimated posterior standard deviations less than or equal to 0.5. The probability function estimate is plotted on right of Figure 4.2.

To see more clearly how the probability of progression depends on `age` and `duration`, we plot the cross sections of the estimate in Figure 4.3. The cross sections with their 90% Bayesian confidence intervals are plotted in Figure 4.4 and Figure 4.5. From these plots, we

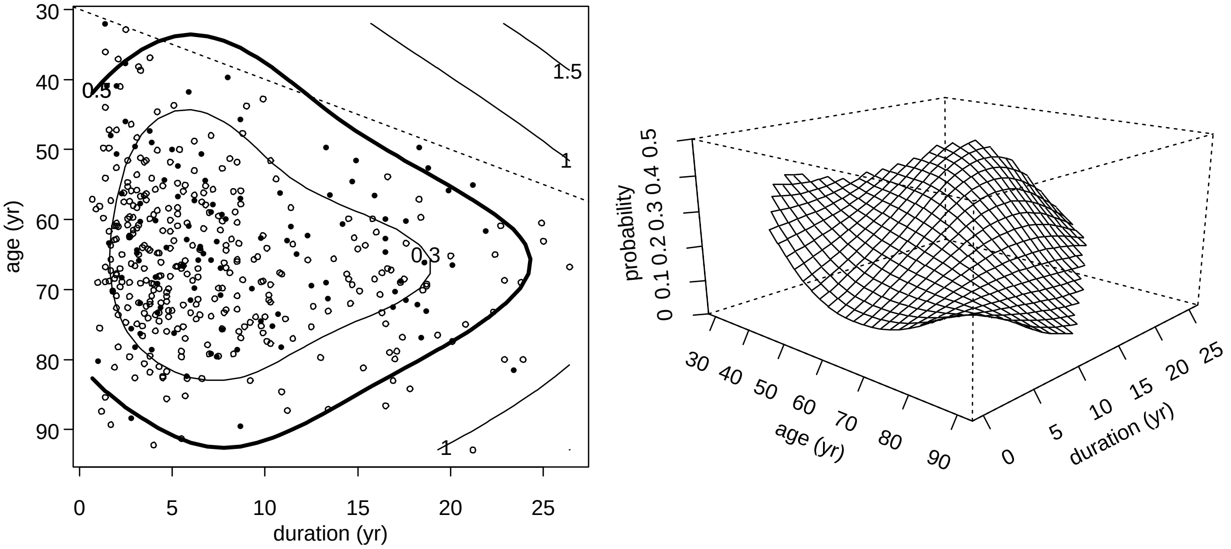


Figure 4.2: NID 'Progression' analysis. Left: data and contours of constant posterior standard deviation as a function of `age` and `duration` at the median value of `glycosylated hemoglobin` and `bmi`. The dotted line is `age - duration` = 30 years. Right: estimated probability of progression in the defined region, as a function of `age` and `duration`, at the median value of `glycosylated hemoglobin` and `bmi`.

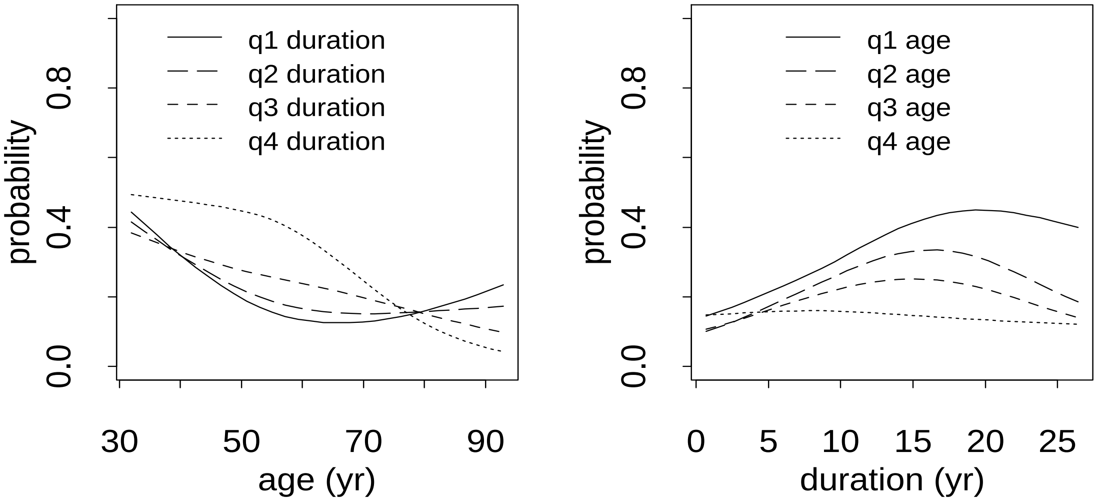


Figure 4.3: NID 'Progression' analysis. Cross sections of estimated probability of progression as a function of age and duration, at the median value of glycosylated hemoglobin and bmi. q1, q2, q3 and q4 are the quantiles at .125, .375, .625 and .875.

see that the risk of progression of retinopathy decreases with increasing age and the shapes are different between the fourth quantile and other quantiles of duration. The risk increases with increasing duration up to about 17 years and does not increase after that. Increased mortality of those with longer duration and more severe retinopathy may explain, in part, this finding. Table 4.1 gives the correspondence between the percentiles and the physical units.

Table 4.1: Percentiles used in plots.

percentile	12.5	37.5	62.5	87.5
age (yr)	53.0	63.5	69.8	78.2
duration (yr)	2.5	4.2	7.8	16.7

Plots in Figure 4.2 indicate that participants who were diagnosed to have diabetes just above 30 years of age (points just below the dotted line in the left panel) are at higher risk of progression of their retinopathy than those diagnosed later in life.

5 INCIDENCE IN THE OLDER ONSET GROUP TAKING INSULIN

After excluding participants with missing data, there were 143 participants in the older onset ID group that had a baseline score of 10/10, and hence, are included in ID 'Incidence'

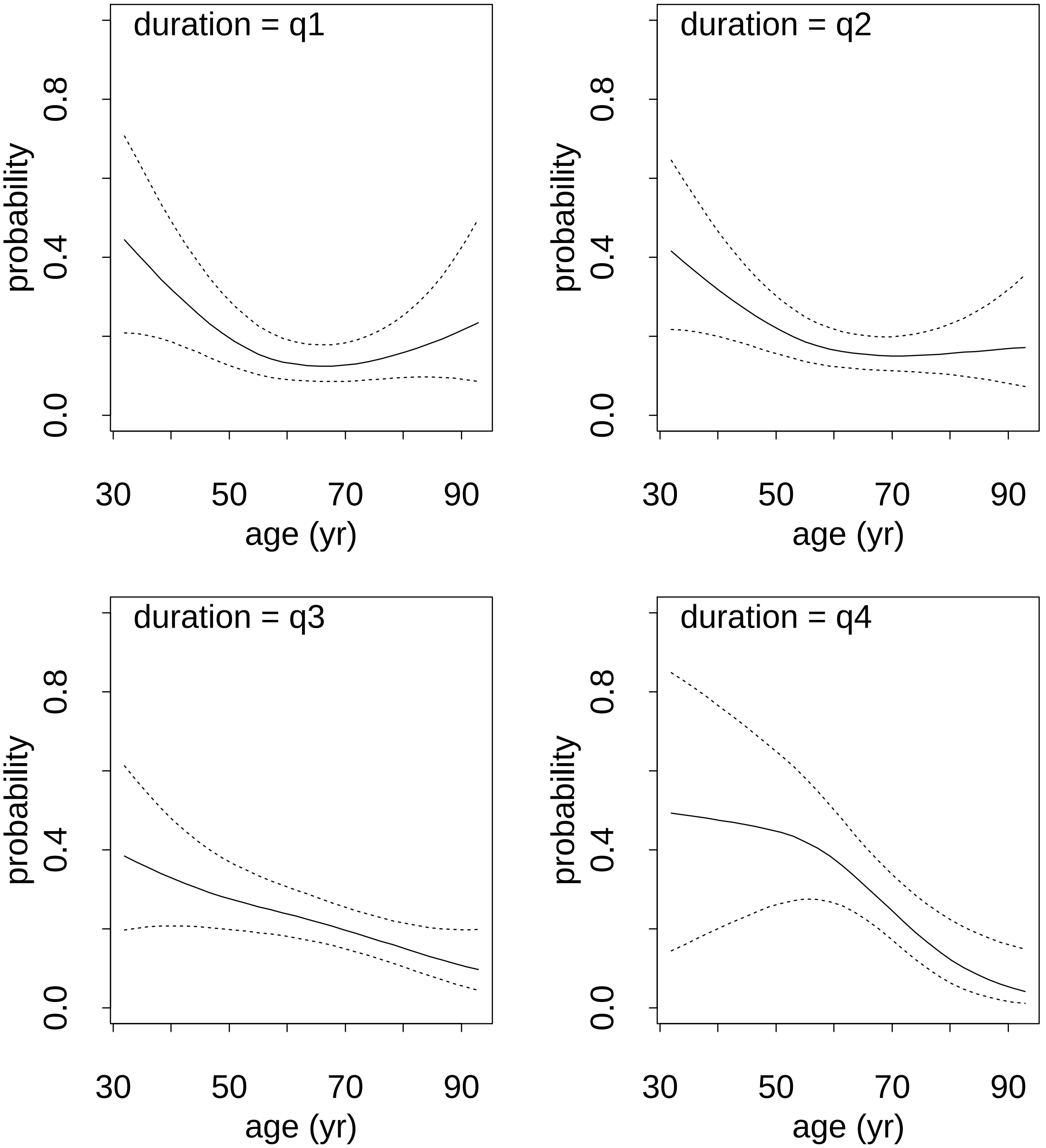
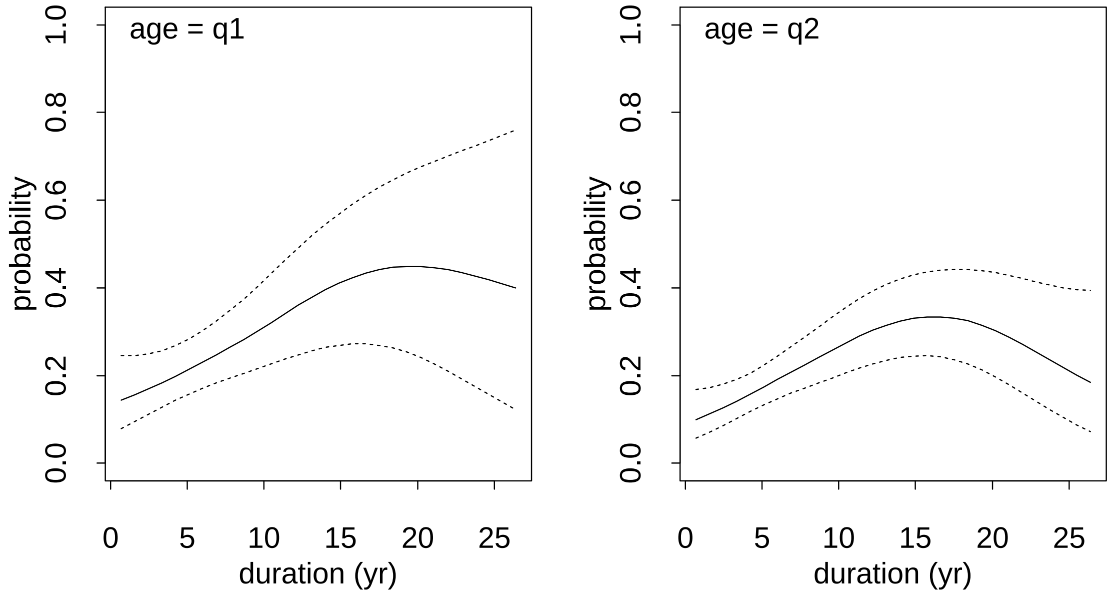


Figure 4.4: NID 'Progression' analysis. Cross sections of estimated probability of progression as a function of `age` with Bayesian confidence intervals, at four quantiles of `duration` and at the median values of `glycosylated hemoglobin` and `bmi`.



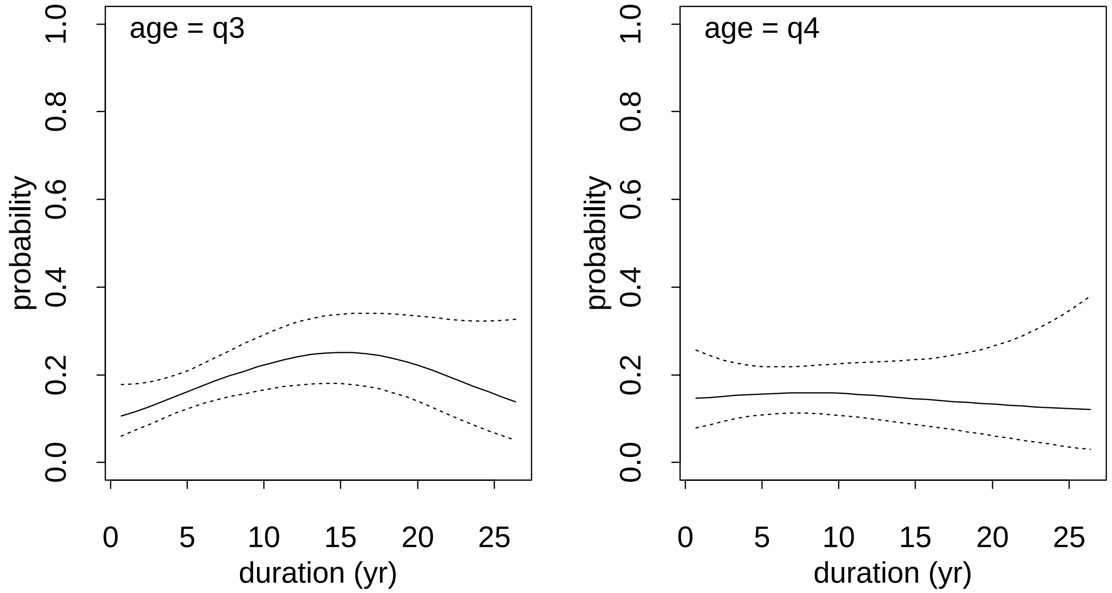


Figure 4.5: NID 'Progresion' analysis. Cross sections of estimated probability of progression as a function of `duration` with Bayesian confidence intervals, at four quantiles of `age` and at the median values of `glycosylated hemoglobin` and `bmi`.

analysis. This sample size is probably not large enough to estimate interactions well. Klein *et al*[27] found that `age` and `glycosylated hemoglobin` are significant using a GLIM. We found that quadratic terms of `duration` and `pulse` are significant using a GLIM:

$$\begin{aligned}&\text{logit}(p(\texttt{age, duration, glycosylated hemoglobin, pulse}))\\ &\quad= \mu + a_1 \times \texttt{age} + a_2 \times \texttt{duration} + a_3 \times \texttt{duration}^2 + \\ &\qquad a_4 \times \texttt{glycosylated hemoglobin} + a_5 \times \texttt{pulse} + a_6 \times \texttt{pulse}^2. \end{aligned} \tag{5.1}$$

Fitting an SS-ANOVA model with main effects of `age`, `duration`, `glycosylated hemoglobin` and `pulse`, we found that the main effects of `age` and `glycosylated hemoglobin` are linear. Then we fitted the model:

$$\begin{aligned}&f(\texttt{age, duration, glycosylated hemoglobin, pulse})\\ &\quad= \mu + a_1 \times \texttt{age} + f_1(\texttt{duration}) + a_2 \times \texttt{glycosylated hemoglobin} + \\ &\qquad f_2(\texttt{pulse}). \end{aligned} \tag{5.2}$$

The estimates of the main effects and their 90% Bayesian confidence intervals are plotted in Figure 5.1. We believe the ups and downs in the middle of the main effect of `duration` are caused by the poor choice of the smoothing parameter (too small) due to the small sample size. But the pattern of the main effect of `duration` is reliable and agrees with the previous conclusion. That is, the risk increases with the increasing duration up to about 6 years and does not increase any more thereafter. From the main effect of `pulse`, we conclude that the risk is higher for participants with higher pulse rate. Higher resting pulse rate may be secondary to diabetic neuropathy involving the autonomic nervous system, a condition which has been postulated to be involved in the development of diabetic retinopathy. The fits from model (5.1) are well inside the Bayesian confidence intervals of the SS-ANOVA estimates. It is difficult to distinguish between these two models with such a small sample size. The quadratic form of the GLIM model for pulse is necessarily symmetric about its minimum at around 35 and suggests that the risk is increasing as pulse decreases below 35 while the SS-ANOVA model suggests that the minimum is a little higher, and the curve is flattens out below its minimum.

The SS-ANOVA method needs relatively large sample sizes to get good estimates of multiple smoothing parameters. Our experience with real data and simulations suggest that about 100 observations for each smoothing parameter will give fairly reliable estimates.

6 PROGRESSION OF THE OLDER ONSET GROUP TAKING INSULIN

Due to its flexibility, the SS-ANOVA method can be used in several stages of data analyses. In the previous sections, we explained the SS-ANOVA method as a tool for model building and estimation. In practice, we can also use the SS-ANOVA method to explore the behavior of raw data. For example, we can get a marginal estimate for each covariate to investigate whether a covariate has a marginal nonlinear effect. Alternatively, we can get some marginal

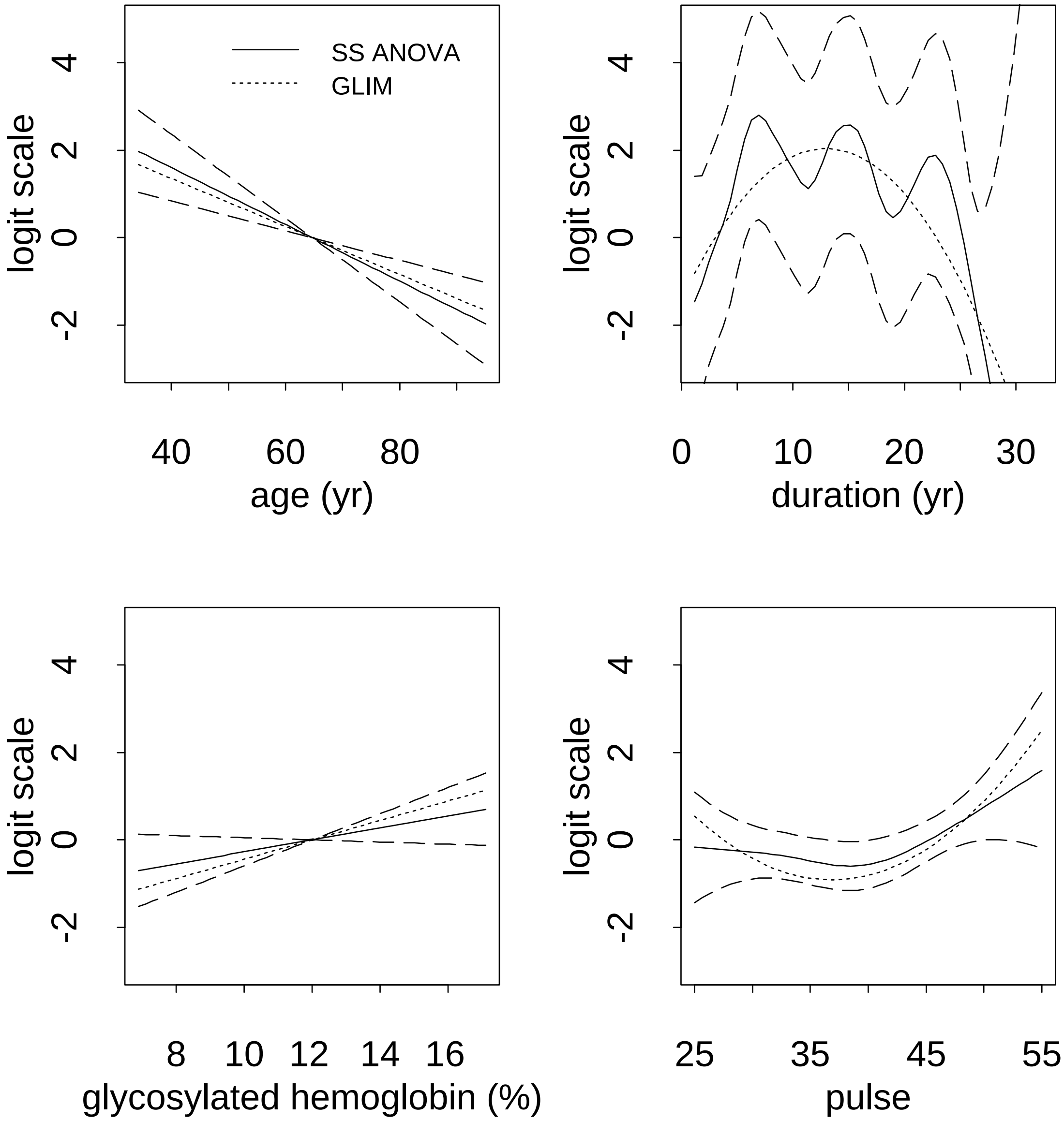


Figure 5.1: ID 'Incidence' analysis. Estimates of the main effects for incidence, logit scale. Dashed lines are component-wise 90% Bayesian confidence intervals.

estimates on pairs of 2 or more covariates to investigate their interactions. Furthermore, the SS-ANOVA method can be used as a diagnostic tool [35].

After deleting participants with missing data, there were 374 participants in the older onset ID group qualified for inclusion in the analysis for 'Progression'. Klein *et al*[26] found that `age`, `duration`, `glycosylated hemoglobin` and `base-retinopathy-level` are significant using a GLIM model. They concluded that the risk of progression is higher if `duration` is longer and if `base-retinopathy-level` is lower (less severe). To investigate whether a GLIM is appropriate, we obtained SS-ANOVA estimates for `age`, `duration`, `glycosylated hemoglobin` and `bmi` separately. These marginal estimates and proportions within each decile on the logit scale are plotted in Figure 6.1. Using GLIM, we find that the effect of `age` is nonlinear (significant up to 4th order of polynomials in a GLIM model). The effect of `bmi` is borderline significant and nonlinear (a 4th order polynomial has a p value of 0.0609).

This data set provides an opportunity to include `base-retinopathy-level` as a categorical variable in the model. We divided `base-retinopathy-level` into 5 categories: 10/10 as category 1, 21/<21 and 21/21 as category 2, 31/<31 and 31/31 as category 3, 41/<41 and 41/41 as category 4, 51/<51 and 51/51 as category 5. We first fit an SS-ANOVA model with main effects of `age`, `duration`, `glycosylated hemoglobin`, `bmi` and `base-retinopathy-level`, we found that the main effect of `age` and `glycosylated hemoglobin` are linear. Finally we fitted the model:

$$\begin{aligned} & f(\texttt{age, duration, glycosylated hemoglobin, bmi, base-retinopathy-level}) \\ = \;& \mu + a_1 \times \texttt{age} + f_1(\texttt{duration}) + a_2 \times \texttt{glycosylated hemoglobin} + \\ & f_2(\texttt{bmi}) + \sum_{k=2}^{5} \gamma_k I_k(\texttt{base-retinopathy-level}). \end{aligned} \tag{6.1}$$

The estimates of the main effects and their 90% Bayesian confidence intervals are plotted in Figure 6.2. The estimates of γ_2, γ_3, γ_4, γ_5 and their posterior standard deviations (inside parentheses) are 0.73(0.35), -0.68(0.37), -1.05(0.40) and -0.17(0.51). Our conclusions basically agree with Klein *et al*[26] with the following modifications:

1. The risk of progression of retinopathy increases with increasing duration of diabetes up to around 8 years and does not increase after that;

2. The effect of body mass index is small;

3. Baseline level has a significant effect on progression of retinopathy, but not monotonically. A subject with baseline level 10/10 has about same risk as a subject with baseline level 51/<51 or 51/51. A subject with baseline level 10/10 has higher risk than a subject with baseline level 31/<31 to 41/41. A subject with baseline level 10/10 has lower risk than a subject with baseline level 21/<21 or 21/21. This may reflect the arbitrary division of the levels which may not be evenly spaced, or, persons with more severe retinopathy may progress more rapidly but those who progress may be less likely to survive to be examined at the follow-up examination.

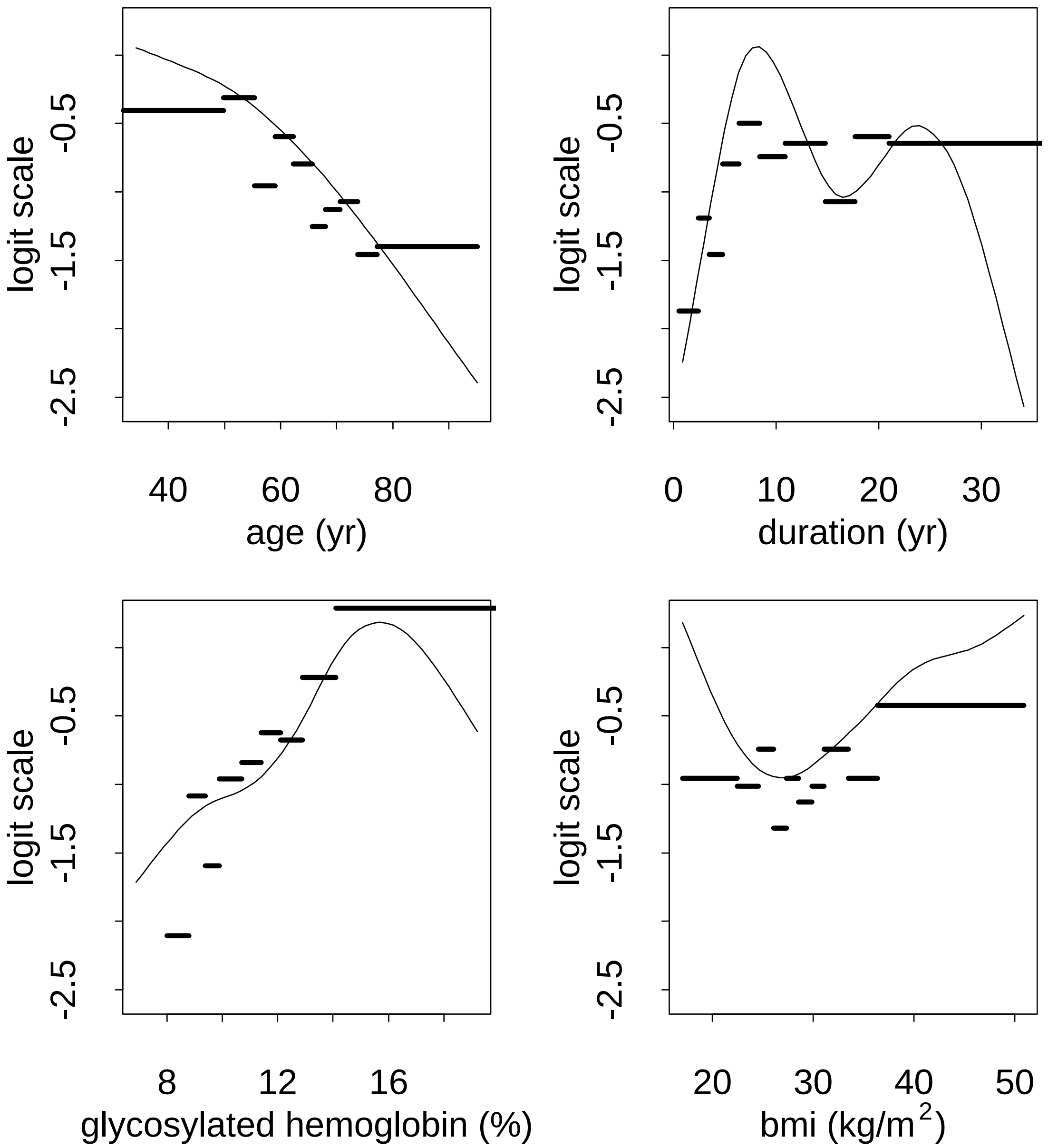


Figure 6.1: ID 'Progression' analysis. Estimates of the marginal effects. Solid lines: SS-ANOVA estimates; segments: the logit of proportions within each decile.

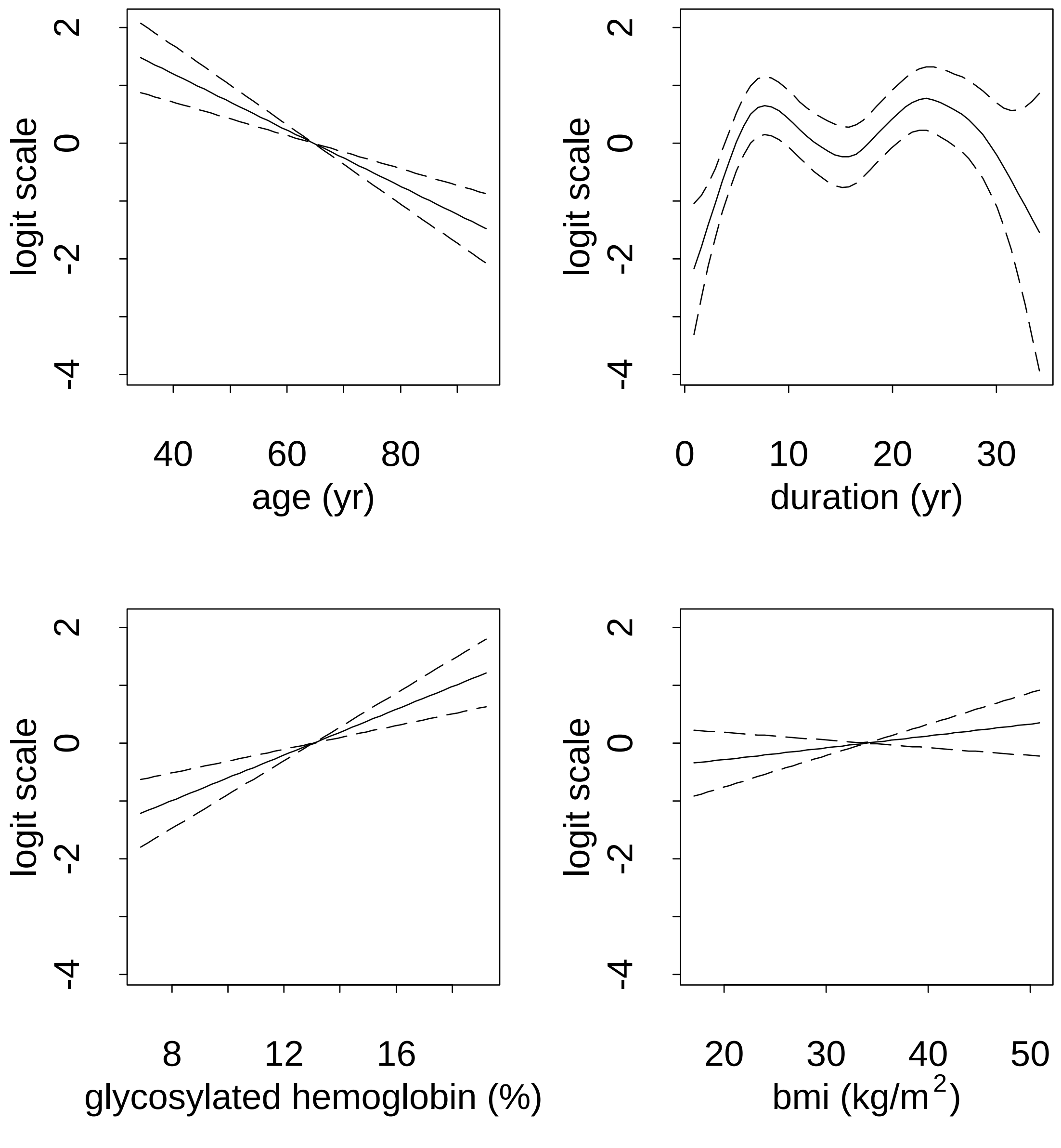


Figure 6.2: ID 'Progression' analysis. Estimates of the main effects for progression. Dashed lines are component-wise 90% Bayesian confidence intervals.

7 CONCLUSIONS

We first summarize the new findings from our analyses, and then make some concluding remarks concerning the SS-ANOVA estimation methods.

1. `Age` effects are nonlinear and different for different groups. For progression of the older onset ID group, the risk decreases with increasing `age`. Since mortality may change the population under study, especially for the older onset group, more frequent follow-up of a cohort (perhaps yearly) might provide further understanding of this relation.

2. The effect of `duration` is consistent in all groups. The risk generally increases with increasing `duration` up to a certain point and does not increase after that. This indicates that if a person with older onset diabetes has not had an event (incidence or progression of retinopathy) after some years of diabetes, the risk will not substantially increase after that, although it remains higher than newly diagnosed persons.

3. The effect of `glycosylated hemoglobin` is large, which agrees with previous studies. We find that there is a strong interaction between `age` and `glycosylated hemoglobin` for incidence in the older onset NID group.

4. The effect of body mass index (`bmi`) is small.

5. `Pulse` rate has a moderate effect on the incidence of retinopathy in the older onset ID group. In a main effects model which also included `age`, `duration`, and `glycosylated hemoglobin` , the main effect for `pulse` rate is constant for lower pulse rates and then veers upward linearly (in the logit scale) for higher pulse rates.

6. Baseline retinopathy severity level (`base-retinopathy-level`) was found to have a significant effect on progression in the older onset ID group in a main effects model which also included age, duration, glycosylated hemoglobin and body mass index, but the effect was not monotonic. Participants with baseline level 21/<21 and 21/21 had the highest risk and participants with baseline levels 31/<31 to 41/41 had the lowest risk of progression.

The SS-ANOVA models have provided us with a family of flexible penalized log likelihood estimates, which specifically include the possible fitting of interaction terms, as well as allowing for combinations of continuous and categorical variates, which, moreover reduce to standard GLIM models as the smoothing parameters tend to infinity. These models are well adapted to the irregular distribution of multiple predictor variables commonly found in demographic data. The data-based method for estimating smoothing parameters allow the data to suggest when the flexible or 'smooth' components of the model are not necessary. This family of models allows the user to visualize complex relationships between variables that might not otherwise be evident from the use of purely parametric GLIM models. For example, the analysis here leads to new questions regarding why higher glycosylated hemoglobin in the younger subjects in the NID incidence group leads to increased risk compared to higher glycosylated hemoglobin in the older subjects in this group.

As with any nonparametric function estimation method, various assumptions about the function being estimated must be made. The primary assumption in the models of this paper is that f is 'smooth' as measured by the integrals of $(\frac{\partial^2 f}{\partial t_\alpha^2})^2$, $(\frac{\partial^2 f}{\partial t_\beta^2})^2$, and $(\frac{\partial^4 f}{\partial t_\alpha^2 \partial t_\beta^2})^2$ (which give the five terms in the penalty functional of (1.12)).

There is a limit to the number of smoothing parameters that can be reliably estimated with the sample sizes here. Some of the local 'wiggles' are probably a manifestation of that, particularly in Figure 5.1. However, the Bayesian confidence intervals do suggest when bumps are not 'real'. When used for exploratory purposes, the smoothing parameters chosen by the automatic method here may be used as starting guesses for 'eyeball' or subjective smoothing parameter choices.

One drawback of this family of methods (aside from the fact that larger data sets are required than for standard purely parametric models) is the fact that the publicly available software for using them is not at the present time quite at the 'cookbook' level. If a sample program for the particular model of interest is not available, the user must write a driver which contains details of their model. Examples of drivers for three models, including the one used in WWGKK, are packaged with GRKPACK. Drivers for the four SS-ANOVA models used in this paper may be obtained from the first author. Since the GRKPACK code is based on matrix decompositions, it is relatively slow, and requires a relatively large amount of storage, placing an upper limit on the sample sizes can be analyzed. Approximate numerical methods that will reduce both the time and space required for a given sample size are an area of active research at the present time. In the meantime, we believe that the present methods represent an important new approach to the analysis of demographic data sets similar to those analyzed here.

ACKNOWLEDGMENTS

We thank Scot Moss who provided some of the data and invaluable technical support. The second autor wishes to express her special appreciation to Marian Fisher who introduced her to some important issues in demographic data analysis. Yuedong Wang's research supported in part by NIH Grant EY09946, P60 DK20572, P30 HD18258 and U54 HD29184, Grace Wahba's research supported in part by NIH Grant EY09946 and NSF Grant DMS9121003, Chong Gu's research supported by NSF Grant DMS9301511, Ronald Klein, MD and Barbara Klein, MD, research supported by NIH Grant EY03083.

References [2],[18],[3] and [22] are available through the second author's home page URL `http://www.stat.wisc.edu/~wahba`.

References

[1] P. McCullagh and J. Nelder. *Generalized Linear Models, Second Edition.* Chapman and Hall, 1989.

[2] G. Wahba, Y. Wang, C. Gu, R. Klein, and B. Klein. Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. Technical Report 940, Department of Statistics, University of Wisconsin, Madison, WI, 1994, to appear in the December, 1995 issue of *Ann. Statist.*

[3] Y. Wang. GRKPACK: Fitting smoothing spline analysis of variance models to data from exponential families. Technical Report 942, Dept. of Statistics, University of Wisconsin, Madison, WI, 1995.

[4] F. O'Sullivan. *The analysis of some penalized likelihood estimation schemes*. PhD thesis, Dept. of Statistics, University of Wisconsin, Madison, WI, 1983. Technical Report 726.

[5] F. O'Sullivan, B. Yandell, and W. Raynor. Automatic smoothing of regression functions in generalized linear models. *J. Amer. Statist. Assoc.*, 81:96–103, 1986.

[6] C. Gu and G. Wahba. Semiparametric analysis of variance with tensor product thin plate splines. *J. Royal Statistical Soc. Ser. B*, 55:353–368, 1993.

[7] C. Gu and G. Wahba. Smoothing spline ANOVA with component-wise Bayesian "confidence intervals". *J. Computational and Graphical Statistics*, 2:97–117, 1993.

[8] T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman and Hall, 1990.

[9] J. Chambers and T. Hastie. *Statistical Models in S*. Wadsworth and Brooks, 1992.

[10] G. Wahba. *Spline Models for Observational Data*. SIAM, 1990. CBMS-NSF Regional Conference Series in Applied Mathematics, v. 59.

[11] C. Gu and G. Wahba. Comments to 'Multivariate Adaptive Regression Splines', by J. Friedman. *Ann. Statist.*, 19:115–123, 1991.

[12] G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Applic.*, 33:82–95, 1971.

[13] G. Wahba. Partial and interaction splines for the semiparametric estimation of functions of several variables. In T. Boardman, editor, *Computer Science and Statistics: Proceedings of the 18th Symposium*, pages 75–80. American Statistical Association, Washington, DC, 1986.

[14] C. Gu. Adaptive spline smoothing in non-Gaussian regression models. *J. Amer. Statist. Assoc.*, 85:801–807, 1990.

[15] C. Gu. Cross-validating non-Gaussian data. *J. Comput. Graph. Stats.*, 1:169–179, 1992.

[16] C. Gu, D.M. Bates, Z. Chen, and G. Wahba. The computation of GCV functions through householder tridiagonalization with application to the fitting of interaction spline models. *SIAM J. Matrix Anal.*, 10:457–480, 1989.

[17] C. Gu and G. Wahba. Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM J. Sci. Statist. Comput.*, 12:383–398, 1991.

[18] Y. Wang. *Smoothing spline analysis of Variance of Data from Exponential Families*. PhD thesis, Technical Report 928, University of Wisconsin-Madison, Madison, WI, 1994.

[19] G. Wahba, C. Gu, Y. Wang, and R. Chappell. Soft classification, a. k. a. risk estimation, via penalized log likelihood and smoothing spline analysis of variance. In D. Wolpert, editor, *The Mathematics of Generalization, Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol. XX*, pages 329–360, Reading, MA, 1995. Addison-Wesley.

[20] Y. Wang, G. Wahba, R. Chappell, and C. Gu. Simulation studies of smoothing parameter estimates and Bayesian confidence intervals in Bernoulli SS-ANOVA models. *Commun. Stat. Comp. Sim.*, 24:1037–1059, 1995.

[21] W. Wong. Estimation of the loss of an estimate. Technical Report 356, Dept. of Statistics, University of Chicago, Chicago, Il, 1992.

[22] D. Xiang and G. Wahba. A generalized approximate cross validation for smoothing splines with non-Gaussian data. Technical Report 930, Dept. of Statistics, University of Wisconsin, Madison, WI, 1994, to appear, *Statistica Sinica*.

[23] B. Yandell. Algorithms for nonlinear generalized cross-validation. In T.J. Boardman, editor, *Computer Science and Statistics: 18th Symposium on the Interface*. American Statistical Association, Washington, DC, 1986.

[24] G. Wahba. Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Stat. Soc. Ser. B*, 45:133–150, 1983.

[25] C. Gu. Penalized likelihood regression: a Bayesian analysis. *Statistica Sinica*, 2:255–264, 1992.

[26] R. Klein, B. E. K. Klein, S. E. Moss, M. D. Davis, and D. L. DeMets. Glycosylated hemoglobin predicts the incidence and progression of diabetic retinopathy. *Journal of the American Medical Association*, 260:2864–2871, 1988.

[27] R. Klein, B. E. K. Klein, S. E. Moss, M. D. Davis, and D. L. DeMets. Is blood pressure a predictor of the incidence or progression of diabetic retinopathy. *Arch. Intern. Med.*, 149:2427–2432, 1989.

[28] B. E. K. Klein, M. D. Davis, P. Segal, J. A. Long, W. A. Harris, G. A. Haug, Y. Magli, and S. Syrjala. Diabetic retinopathy: Assessment of severity and progression. *Ophthalmology*, 91:10–17, 1984.

[29] R. Klein, B. E. K. Klein, and K. J. Moss, S. E. Cruickshanks. The relationship of hyperglycemia to long-term incidence and progression of diabetic retinopathy. *Arch. Intern. Med.*, 154:2169–2178, 1994.

[30] R. Klein, B. E. K. Klein, S. E. Moss, and K. J. Cruickshanks. The Wisconsin Epidemiologic Study of Diabetic Retinopathy. XIV. Ten year incidence and progression of diabetic retinopathy. *Arch. Ophthalmol.*, 112:1217–1228, 1994.

[31] R. Klein, B. E. K. Klein, S. E. Moss, M. D. Davis, and D. L. DeMets. The Wisconsin Epidemiologic Study of Diabetic Retinopathy. II. Prevalence and risk of diabetic retinopathy when age at diagnosis is less than 30 years. *Arch. Ophthalmol.*, 102:520–526, 1984.

[32] R. Klein, B. E. K. Klein, S. E. Moss, M. D. Davis, and D. L. DeMets. The Wisconsin Epidemiologic Study of Diabetic Retinopathy. III. Prevalence and risk of diabetic retinopathy when age at diagnosis is 30 or more years. *Arch. Ophthalmol.*, 102:527–532, 1984.

[33] R. Klein, B. E. K. Klein, S. E. Moss, M. D. Davis, and D. L. DeMets. The Wisconsin Epidemiologic Study of Diabetic Retinopathy. IX. Four year incidence and progression of diabetic retinopathy when age at diagnosis is less than 30 years. *Arch. Ophthalmol.*, 107:237–243, 1989.

[34] R. Klein, B. E. K. Klein, S. E. Moss, M. D. Davis, and D. L. DeMets. The Wisconsin Epidemiologic Study of Diabetic Retinopathy. X. Four incidence and progression of diabetic retinopathy when age at diagnosis is 30 or more years. *Arch. Ophthalmol.*, 107:244–249, 1989.

[35] Edward B. Fowlkes. Some diagnostics for binary logistic regression via smoothing. *Biometrika*, 74:503–515, 1987.