Odds Ratio Estimation in Bernoulli Smoothing Spline ANOVA Models

Yuedong Wang *
Department of Biostatistics, University of Michigan

September 1, 1996

Abstract

Wahba et al (1995) introduced Smoothing Spline ANalysis of VAriance (SS ANOVA) method for data from exponential families. In this paper, we estimate the odds ratios based on a SS ANOVA model for binary data and construct Bayesian confidence intervals. We show the calculation using a real data set from Wisconsin Epidemiological Study of Diabetic Retinopathy. We conduct simulations to evaluate the performance of these estimates and their Bayesian confidence intervals. Our simulations suggest that the odds ratio estimates are quite reasonable in general but may be biased towards unity when comparing estimates at peaks with those in valleys. A bootstrap procedure is proposed to correct possible biases and it works very well in our simulation.

Key Words: bias correction; bootstrap; odds ratio; smoothing spline ANOVA.

1 SS ANOVA Models

Binary data occur very often in medical science and other areas. Suppose that for each individual, the response Y takes two possible values: $Y = 0$ or $Y = 1$. Each individual is associated with a vector of covariates: $\boldsymbol{t} = (t_1, \cdots, t_d)$. Let

$$P(Y_i = 0|\boldsymbol{t}_i) = 1 - p(\boldsymbol{t}_i), \quad P(Y_i = 1|\boldsymbol{t}_i) = p(\boldsymbol{t}_i), \qquad i = 1, \cdots, n. \tag{1}$$

Define the odds at $\boldsymbol{t}$ as $p(\boldsymbol{t})/(1 - p(\boldsymbol{t}))$. A logistic regression model

$$\log \frac{p(\boldsymbol{t})}{1 - p(\boldsymbol{t})} = f(\boldsymbol{t}) \tag{2}$$

is often used to investigate the relationship between the response probability $p(\boldsymbol{t})$ and the covariate vector $\boldsymbol{t}$. Furthermore, a linear logistic regression model assumes that

$$f(\boldsymbol{t}) = C + \sum_{j=1}^{d} \beta_j t_j \ . \tag{3}$$

* *Address for correspondence: Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA. E-mail: yuedong@umich.edu*

That is, when other covariates are fixed, the effect of an increase in t_j from t_j^1 to t_j^2 is to increase odds ratio by an amount $\exp(\beta_j(t_j^2 - t_j^1))$, which depends on the difference between t_j^1 and t_j^2 only. This model is easy to explain, but too restrictive in some applications. To build more flexible models than a linear regression surface, many authors used nonparametric methods. O'Sullivan, Yandell and Raynor (1986) and Gu (1990) used penalized likelihood method with smoothing splines and thin plate splines. Hastie and Tibshirani (1990) used additive models. Wahba, Wang, Gu, Klein and Klein (1995) introduced the SS ANOVA models using the penalized likelihood and Smoothing Spline ANalysis of VAriance methods. See also Wahba, Gu, Wang and Chappell (1994a), Wahba, Wang, Gu, Klein and Klein (1994b), Wang (1994), Wang, Wahba, Chappell and Gu (1995) and Wang, Wahba, Gu, Klein and Klein (1996) for details of the SS ANOVA models.

A SS ANOVA model assumes that $t_j \in \mathcal{T}^{(j)}$, where $\mathcal{T}^{(j)}$ is a measurable space. f belongs to a subspace of tensor products of reproducing kernel Hilbert spaces (Aronszajn (1950), Wahba et al. (1995)). More precisely, the model space $\mathcal{M}$ of a SS ANOVA model contains elements

$$\begin{aligned} f(\boldsymbol{t}) &= C + \sum_{j \in J_1} f_j(t_j) + \sum_{(j_1,j_2) \in J_2} f_{j_1,j_2}(t_{j_1}, t_{j_2}) + \cdots \\ &\quad + \sum_{(j_1,\cdots,j_d) \in J_d} f_{j_1,\cdots,j_d}(t_{j_1}, \cdots, t_{j_d}), \end{aligned} \tag{4}$$

where J_k is a subset of the set of all k-tuples $\{(j_1, \cdots, j_k) : \quad 1 \le j_1 < \cdots < j_k \le d\}$ for $k = 1, \cdots, d$. Identifiability conditions are imposed such that each term in the sums is intergrated to zero with respect to any one of its arguments. Each term in the first sum is called a *main effect*, each term in the second sum is called a *two-factor interaction*, and so on. Similar to the analysis of variance, usually higher-order interactions are eliminated from the model space to reduce the complexity of the model. See Wahba et al. (1995) for details of the model construction. When a model is chosen, we can regroup and write the model space as

$$\mathcal{M} = \mathcal{H}^0 \oplus \sum_{j=1}^{q} \mathcal{H}^j, \tag{5}$$

where $\mathcal{H}^0$ is a finite dimensional space containing functions which are not going to be penalized, usually lower order polynomials. A SS ANOVA estimate is the solution to the following variational problem:

$$\min_{f \in \mathcal{M}} \left\{ -\sum_{i=1}^{n} (y_i f(\boldsymbol{t}_i) - \log(1 + e^{f(\boldsymbol{t}_i)})) + \frac{n}{2} \sum_{j=1}^{q} \lambda_j \|P_j f\|^2 \right\}. \tag{6}$$

The first part in (6) is the negative log likelihood. It measures the goodness of fit. In the second part, P_j is the orthogonal projector in $\mathcal{M}$ onto $\mathcal{H}^j$ and $\|P_j f\|^2$ is a quadratic roughness penalty. λ_j's are a set of smoothing parameters. They control the trade-off between the goodness of fit and the roughness of the estimate. See Wahba et al. (1995) and Wang et al. (1995) for details of how to calculate a SS ANOVA estimate and how to choose smoothing parameters based on data.

The solution to (6) is approximately equal to the posterior mean of the following Bayesian model. Let the prior for $f(\boldsymbol{t})$ be

$$F_\xi(\boldsymbol{t}) = \sum_{\nu=1}^{M} \tau_\nu \phi_\nu(\boldsymbol{t}) + b^{\frac{1}{2}} \sum_{\beta=1}^{q} \sqrt{\theta_\beta} Z_\beta(\boldsymbol{t}), \tag{7}$$

where $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_M)^T \sim N(\mathbf{0}, \xi I)$, Z_β are independent, zero mean Gaussian stochastic processes, independent of $\boldsymbol{\tau}$, with $\mathrm{E}Z_\beta(\boldsymbol{t})Z_\beta(\boldsymbol{s}) = R_\beta(\boldsymbol{t}, \boldsymbol{s})$. R_β is the reproducing kernel of $\mathcal{H}_\beta$. With $\xi \to \infty$, Wahba et al. (1995) proved that the posterior means of the overall function and its components are approximately equal to the solution to (6) and its components. Posterior covariances are listed in Theorem 1 in Wahba et al. (1995). These posterior covariances can be used to construct Bayesian confidence intervals for a SS ANOVA estimate and its components.

2 Estimation of Odds Ratios and Bias Correction

The SS ANOVA estimate of the probability function for binary data can be used to calculate the odds ratios. For any two points $\boldsymbol{t}$ and $\boldsymbol{s}$, the odds ratio of $\boldsymbol{t}$ and $\boldsymbol{s}$ equals

$$OR(\boldsymbol{t}/\boldsymbol{s}) = \exp\{f(\boldsymbol{t}) - f(\boldsymbol{s})\}, \tag{8}$$

where the function f is the logit of the probability function. It depends on both $\boldsymbol{t}$ and $\boldsymbol{s}$ since we do not assume a linear relationship. A natural estimate of $OR(\boldsymbol{t}/\boldsymbol{s})$ is

$$\widehat{OR}(\boldsymbol{t}/\boldsymbol{s}) = \exp\{\hat{f}(\boldsymbol{t}) - \hat{f}(\boldsymbol{s})\}. \tag{9}$$

Often, we are interested in how one covariate affects the odds when other risk factors are fixed (at their medians or means). Suppose the covariate we are interested in is t_1. For any two possible values t_1^1 and t_1^2 of t_1, the log odds ratio of $\boldsymbol{t} = (t_1^2, t_2, \cdots, t_d)$ and $\boldsymbol{s} = (t_1^1, t_2, \cdots, t_d)$ equals

$$\begin{aligned}\log(\widehat{OR}(\boldsymbol{t}/\boldsymbol{s})) &= \sum_{1 \in J_1} (f_1(t_1^2) - f_1(t_1^1)) + \sum_{(1,j) \in J_2} (f_{1,j}(t_1^2, t_j) - f_{1,j}(t_1^1, t_j)) \\ &\quad + \sum_{(1,j_2,\cdots,j_d) \in J_d} (f_{1,j_2,\cdots,j_d}(t_1^2, t_{j_2}, \cdots, t_{j_d}) - f_{1,j_2,\cdots,j_d}(t_1^1, t_{j_2}, \cdots, t_{j_d})).\end{aligned}$$

Notice that the odds ratio depends on t_j, $j = 2, \cdots, d$, if there is an interaction between t_1 and t_j in the model space $\mathcal{M}$.

Often, the SS ANOVA estimate of the probability function has relatively large biases at peak and valley points. These biases are added up if we pick t_1^1 at a peak and t_1^2 at a valley. This effect is obvious from the simulations in section 4. Correction of the possible biases is necessary for these cases. We propose a bootstrap procedure to correct these biases: (1) generate bootstrap samples of binary data from SS ANOVA estimate of the probability function p; (2) calculate SS ANOVA estimates of odds ratios from these bootstrap samples and denote the median of them as $\widehat{OR}^*(\boldsymbol{t}/\boldsymbol{s})$; (3) estimate the bias of log odds ratio by $\widehat{bias} = \ln \widehat{OR}(\boldsymbol{t}/\boldsymbol{s}) - \ln \widehat{OR}^*(\boldsymbol{t}/\boldsymbol{s})$; (4) calculate the bias correctted estimate by $\widehat{OR}_{corrected}(\boldsymbol{t}/\boldsymbol{s}) = \exp(2\widehat{OR}(\boldsymbol{t}/\boldsymbol{s}) - \widehat{OR}^*(\boldsymbol{t}/\boldsymbol{s}))$. Similar bias correction in nonparametric regression setting using bootstrap has been studied previously by Gu (1987) and Fan and Hu (1992). These studies are primarily theoretical and it is not clear whether this technique will necessarily work in prcatice. Simulation in section 4 indicates that this procedure works very well.

Based on the same Bayes model (7), we can approximate the posterior distribution of $f(\boldsymbol{t}) - f(\boldsymbol{s})|\boldsymbol{y}$ by a Gaussian distribution with mean $\hat{f}(\boldsymbol{t}) - \hat{f}(\boldsymbol{s})$ and variance

$$\begin{aligned} \delta^2 &= \mathrm{Var}(f(\boldsymbol{t}) - f(\boldsymbol{s})|\boldsymbol{y}) \\ &= \mathrm{Var}(f(\boldsymbol{t})|\boldsymbol{y}) + \mathrm{Var}(f(\boldsymbol{s})|\boldsymbol{y}) - \mathrm{Cov}(f(\boldsymbol{t}), f(\boldsymbol{s})|\boldsymbol{y}). \end{aligned} \tag{10}$$

δ^2 can be calculated from the formulas in Theorem 1 in Wahba et al. (1995). See Wang (1997) for details for calculations of posterior covariances. Hence we can approximate the distribution of $OR(\boldsymbol{t}/\boldsymbol{s})$ by a log Normal distribution and construct the $(1-\alpha) \times 100\%$ Bayesian confidence interval as

$$(\widehat{OR}(\boldsymbol{t}/\boldsymbol{s}) \exp\{-z_{\frac{\alpha}{2}}\delta\}, \widehat{OR}(\boldsymbol{t}/\boldsymbol{s}) \exp\{z_{\frac{\alpha}{2}}\delta\}). \tag{11}$$

It is well established that the Bayesian confidence intervals for the function f have good frequentist properties (Wahba 1983, Nychka 1988, Wang and Wahba 1995). It is not clear whether the Bayesian confidence interval (11) for the odds ratio has similar frequentist properties since it involves two points. Our simulations in section 4 indicate that the answer is yes.

3 A Practical Example

In this section, we use a data set from the Wisconsin Epidemiology Study of Diabetic Retinopathy (WESDR) to demonstrate the SS ANOVA method and odds ratio estimation. See Klein, Klein, Moss, Davis and DeMets (1988), Klein, Klein, Moss, Davis and DeMets (1989) and references therein for a detailed description of the data and some analyses using the linear logistic regression.

In brief, the data set contains 256 insulin-dependent diabetic patients diagnosed as having diabetes before 30 years of age ("younger onset group"). None of them had diabetic retinopathy at the baseline. At the follow-up examination, all 256 patients were checked to see if they had diabetic retinopathy. The response $Y = 1$ if an individual had diabetic retinopathy at the follow-up. $Y = 0$ otherwise. Several covariates were recorded. We only list the variables pertinent to our analyses:

1. `age`: age in years at the time of baseline examination;
2. `duration`: duration of diabetes at the time of baseline examination;
3. `glycosylated hemoglobin`: a measure of hyperglycemia;
4. Systolic blood `pressure` in mmHg.

Following model was used in Wang (1994) (Model IV):

$$\begin{aligned} & \text{logit}(\text{P}(\texttt{age, duration, glycosylated hemoglobin, pressure})) \\ = \; & \mu + f_1(\texttt{age}) + a_1 \times \texttt{duration} + a_2 \times \texttt{glycosylated hemoglobin} + \\ & a_3 \times \texttt{pressure}. \end{aligned} \tag{12}$$

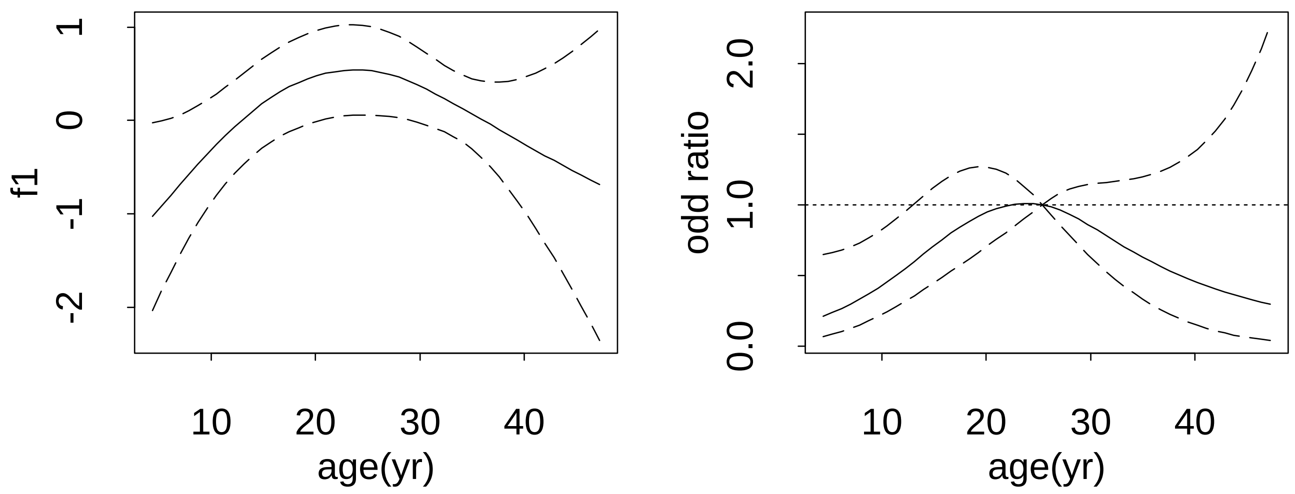

Figure 1: Left: estimates of the main effects $f_1(\texttt{age})$. Right: estimates of odds ratios $OR(\texttt{age}/\texttt{age} = 25)$. Dashed lines are 90% Bayesian confidence intervals.

The main effect of `age` is plotted in the left panel of Figure 1. We see that patients with age from 20 to 30 are at higher risk. To compare the risk at some particular ages, one may want to calculate the odds ratios at these ages. Suppose we fix `duration`, `glycosylated hemoglobin` and `pressure` at their median values, and pick `age` $= 25$ as the base and compare its risk to other ages. The estimated odds ratios and their 90% Bayesian confidence intervals are plotted in the right panel of Figure 1. We see that the odds at `age` $= 25$ is significantly higher than the odds at `age` ≤ 13.

From Figure 1, the odds at `age` $= 25$ is not significantly higher than the odds at `age` $= 40$. But since `age` $= 25$ is a peak point, it is likely that $\widehat{OR}$ overestimates the true OR. This is supported by our simulation in section 4, where $\widehat{OR} = .45$ and $\widehat{OR}^* = .73$ in Table 1. We can estimate the bias of log odds by $\widehat{bias} = \ln \widehat{OR} - \ln \widehat{OR}^* = \ln .45 - \ln .73 = -.48$. Then we can correct this bias in the original estimate: $\widehat{OR}_{corrected} = .45 \times \exp(-.48) = .28$. The 95% Bayesian confidence interval for the corrected estimate of odds ratio becomes (.07,1.00), which is just significant.

4 Simulations

We conducted three simulations to evaluate the performance of the estimates of the odds ratios and their Bayesian confidence intervals. We also conduct a simulation to evaluate the performance of the bias correction procedure.

In the first two simulations, we used two univariate logit functions:

$$\begin{array}{ll} \text{Case A} & f(t) = \frac{1}{2}\beta_{10,5}(t) + \frac{1}{2}\beta_{7,7}(t) + \frac{1}{2}\beta_{5,10}(t) - 1, \\ \text{Case B} & f(t) = 3[10^5 t^{11}(1-t)^6 + 10^3 t^3 (1-t)^{10}] - 2, \end{array}$$

where $0 \leq t \leq 1$. $\beta_{p,q}$ is the Beta function: $\beta_{p,q}(t) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} t^{p-1}(1-t)^{q-1}$. The true probability functions of these two cases are plotted in Figure 2. Bernoulli responses y_i were generated on grid points $t_i = (i - 0.5)/n$, $i = 1, \cdots, n$, according to the true probability function, where n is the sample size. Two sample sizes are used: $n = 100$ and $n = 200$. In Case A, we use $t = 0.2$ as the base and calculate odds ratios at points $t = 0.4$, $t = 0.6$, $t = 0.8$ and $t = 1$. In Case B, we use $t = 0.5$ as the base and calculate odds ratios at points $t = 0.1$, $t = 0.2$, $t = 0.3$ and $t = 0.4$.

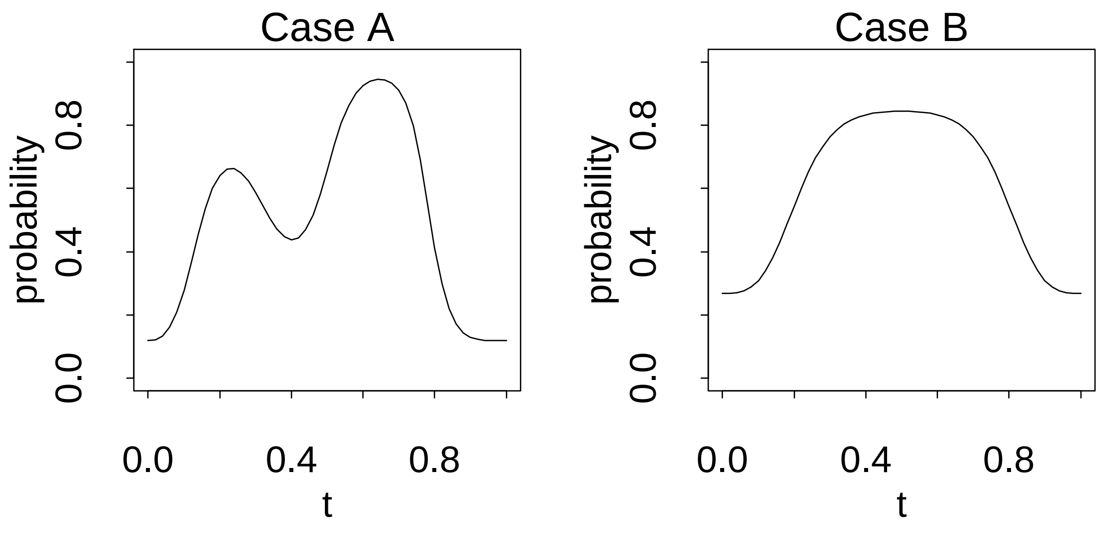

Figure 2: Plots of probability functions used in the first two simulations.

In the third simulation, we use the estimated probability function of (12) as the true model. The design is the same as the data. We call it Case C. As in the above odds ratio calculations, we use `age` $= 25$ as the base and calculate odds ratios at points `age` $= 10$, `age` $= 15$, `age` $= 35$ and `age` $= 40$.

We repeated all three simulations 100 times. In Table 1, the true odds ratios are listed in rows as OR; medians of the 100 estimates of the odds ratios and the standard deviations are listed in the rows labeled $\widehat{OR}$ with standard deviations inside parentheses; the number of times in the 100 replications that the 90% and 95% Bayesian confidence intervals covered the true odds ratios are listed in the rows labeled Coverage with the 95% coverage number inside parentheses. We conclude from this table that these odds ratio estimates and their

Table 1: Odds ratios, estimates of odds ratios and coverages of Bayesian confidence intervals.

Case A	0.4/0.2	0.6/0.2	0.8/0.2	1/0.2
OR	0.45	6.77	0.47	0.08
n=100				
$\widehat{OR}$	0.64 (0.73)	6.66 (34.69)	0.65 (1.08)	0.06 (0.56)
Coverage	86 (90)	89 (94)	81 (89)	92 (95)
n=200				
$\widehat{OR}$	0.54 (0.38)	6.65 (10.57)	0.66 (0.44)	0.06 (0.13)
Coverage	90 (95)	95 (97)	81 (92)	92 (97)
Case B	0.1/0.5	0.2/0.5	0.3/0.5	0.4/0.5
OR	.08	.21	.58	.92
n=100				
$\widehat{OR}$	.12 (.10)	.24 (.18)	.50 (.82)	.84 (.58)
Coverage	82 (90)	89 (93)	98 (98)	99 (99)
n=200				
$\widehat{OR}$	.10 (.06)	.23 (.11)	.51 (.34)	.83 (.22)
Coverage	88 (94)	91 (95)	92 (94)	98 (100)
Case C	10/25	15/25	35/25	40/25
OR	.41	.70	.63	.45
$\widehat{OR}$	.47 (.20)	.71 (.24)	.83 (.61)	.73 (.57)
Coverage	84 (92)	93 (99)	73 (80)	75 (79)

Bayesian confidence intervals work well. Estimates of OR's are generally biased toward unity if one of the two points of OR is at the peak and/or the other is at the valley (for instance, .4/.2 from Case A). This is because the SS ANOVA estimate $\hat{f}$ may underestimate f at a peak and overestimate f at a valley. The coverages of Bayesian confidence intervals at high bias points are lower than the nominal value, while the coverages are higher at other points. So these Bayesian confidence intervals behave similarly to the Bayesian confidence intervals for the estimates of probabilities on the logit scale.

To evaluate the performance of bias correction procedure, we use Case C as our true model. At each of 100 replications, bootstrap bias correction procedure with 100 bootstrap samples is used to get bias corrected estimates of odds ratios. Figure 3 shows the true odds ratio function (solid line), the average of the odds ratio estimates (dotted line) and the average of bias corrected odds ratio estimates (dashed line). The bootstrap correction procedure works very well.

5 Acknowledgements

This research was supported in part by NIH Grants R01 EY09946, P60 DK20572, P30 HD18258 and NSF Grant DMS9121003. The author is indebted to Grace Wahba, Douglas

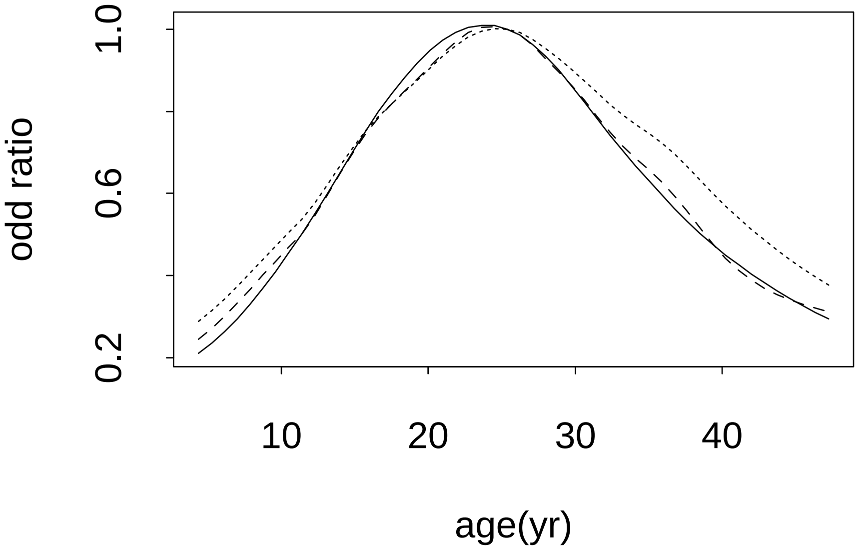

Figure 3: Solid line: true odds ratio function of Case C. Dotted line: average of odds ratio estimates. Dashed line: average of bias corrected odds ratio estimates.

Bates and Brian Yandell for their advice and encouragement. The author thanks two referees for their helpful comments and constructive suggestions. Thanks are also due to Ronald Klein, MD and Barbara Klein, MD for providing the WESDR data.

References

Aronszajn, N. (1950). Theory of reproducing kernels, *Trans. Amer. Math. Soc.* **68**: 337–404.

Fan, J. and Hu, T. C. (1992). Bias correction and higher order kernel functions, *Statistics and Probability Letters* **14**: 235–243.

Gu, C. (1987). What happens when bootstrapping the smoothing spline?, *Communications in Statistics: Theory and Methods* **16**: 3275–3284.

Gu, C. (1990). Adaptive spline smoothing in non-Gaussian regression models, *Journal of the American Statistical Association* **85**: 801–807.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, Chapman and Hall.

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

Klein, R., Klein, B. E. K., Moss, S. E., Davis, M. D. and DeMets, D. L. (1989). Is blood pressure a predictor of the incidence or progression of diabetic retinopathy, *Archives of Internal Medicine* **149**: 2427–2432.

Nychka, D. (1988). Bayesian confidence intervals for smoothing splines, *Journal of the American Statistical Association* **83**: 1134–1143.

O'Sullivan, F., Yandell, B. and Raynor, W. (1986). Automatic smoothing of regression functions in generalized linear models, *Journal of the American Statistical Association* **81**: 96–103.

Wahba, G. (1983). Bayesian confidence intervals for the cross-validated smoothing spline, *Journal of the Royal Statistical Society B* **45**: 133–150.

Wahba, G., Gu, C., Wang, Y. and Chappell, R. (1994a). Soft classification, a.k.a. risk estimation, via penalized log likelihood and smoothing spline analysis of variance, *The Mathematics of Generalization, Santa Fe Institute Studies in the Sciences of Complexity*, Vol. XX, 329-360, D. Wolpert eds,.

Wahba, G., Wang, Y., Gu, C., Klein, R. and Klein, B. (1994b). Structured machine learning for 'soft' classification with smoothing spline ANOVA and stacked tuning, testing and evaluation, *Advances in Neural Information Processing 6* pp. 415–422. J. D. Cowan, G. Tesauro and J. Alspector, eds.

Wahba, G., Wang, Y., Gu, C., Klein, R. and Klein, B. (1995). Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy, *Annals of Statistics* **23**: 1865–1895.

Wang, Y. (1994). Smoothing spline analysis of variance of data from exponential families, Ph.D. Thesis, University of Wisconsin-Madison, Dept. of Statistics. Available by ftp in `ftp.stat.wisc.edu/pub/wahba/theses/ywang.thesis.shar.gz`.

Wang, Y. (1997). GRKPACK: fitting smoothing spline analysis of variance models to data from exponential families, *Communications in Statistics: Simulation and Computation* **26**: 765–782.

Wang, Y. and Wahba, G. (1995). Bootstrap confidence intervals for smoothing splines and their comparison to Bayesian confidence intervals, *J. Statist. Comput. Simul.* **51**: 263–279.

Wang, Y., Wahba, G., Chappell, R. and Gu, C. (1995). Simulation studies of smoothing parameter estimates and Bayesian confidence intervals in Bernoulli SS ANOVA models, *Communications in Statistics: Simulation and Computation* **24**: 1037–1059.

Wang, Y., Wahba, G., Gu, C., Klein, R. and Klein, B. (1996). Using smoothing spline anova to examine the relation of risk factors to the incidence and progression of diabetic retinopathy, Technical Report 956, University of Wisconsin-Madison, Dept. of Statistics. To appear, *Statistics in Medicine*.