The Statistician (1997)
46, *No.* 1, *pp.* 49–56

Odds ratio estimation in Bernoulli smoothing spline analysis-of-variance models

By YUEDONG WANG†

University of Michigan, Ann Arbor, USA

[Received June 1995. Revised September 1996]

SUMMARY

Wahba and co-workers introduced the smoothing spline analysis-of-variance (SS ANOVA) method for data from exponential families. In this paper, we estimate the odds ratios based on an SS ANOVA model for binary data and construct Bayesian confidence intervals. We give a calculation using a real data set from the 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 1 when comparing estimates at peaks with those in troughs. A bootstrap procedure is proposed to correct possible biases and it works very well in our simulation.

Keywords: Bias correction; Bootstrap; Odds ratio; Smoothing spline analysis of variance

1. Smoothing spline analysis-of-variance 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: $\mathbf{t} = (t_1, \ldots, t_d)$. Let

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

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

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

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

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

i.e., when other covariates are fixed, the effect of an increase in t_j from t_j^1 to t_j^2 is to increase the 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 researchers have used nonparametric methods. O'Sullivan *et al.* (1986) and Gu (1990) used the penalized likelihood method with smoothing splines and thin plate splines. Hastie and Tibshirani (1990) used additive models. Wahba *et al.* (1995) introduced the smoothing spline analysis-of-variance (SS ANOVA) models using the penalized likelihood and SS ANOVA methods. See also Wahba *et al.* (1994a, b), Wang (1994a) and Wang *et al.* (1995, 1996) for details of SS ANOVA models.

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

0039–0526/97/46049

An SS ANOVA model assumes that $t_j \in \mathscr{T}^{(j)}$, where $\mathscr{T}^{(j)}$ 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 $\mathscr{M}$ of an SS ANOVA model contains elements

$$f(\mathbf{t}) = C + \sum_{j \in J_1} f_j(t_j) + \sum_{(j_1,j_2) \in J_2} f_{j_1,j_2}(t_{j_1}, t_{j_2}) + \ldots + \sum_{(j_1,\ldots,j_d) \in J_d} f_{j_1,\ldots,j_d}(t_{j_1}, \ldots, t_{j_d}), \tag{4}$$

where J_k is a subset of the set of all k-tuples $\{(j_1, \ldots, j_k): 1 \leqslant j_1 < \ldots < j_k \leqslant d\}$ for $k = 1, \ldots, d$. Identifiability conditions are imposed such that each term in the sums is intergrated to 0 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. As with ANOVA higher order interactions are usually eliminated from the model space to reduce the complexity of the model. See Wahba *et al.* (1995) for details on model construction. When a model has been chosen, we can regroup and write the model space as

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

where $\mathscr{H}^0$ is a finite dimensional space containing functions which will not be penalized, usually lower order polynomials. An SS ANOVA estimate is the solution to the variational problem

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

The first part of expression (6) is the negative log-likelihood. It measures the goodness of fit. In the second part, P_j is the orthogonal projector in $\mathscr{M}$ onto $\mathscr{H}^j$ and $\|P_j f\|^2$ is a quadratic roughness penalty. The λ_j 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 on how to calculate an SS ANOVA estimate and how to choose smoothing parameters based on data.

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

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

where $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_M)^{\mathrm{T}} \sim N(\mathbf{0}, \xi I)$, Z_β are independent, zero-mean Gaussian stochastic processes, independent of $\boldsymbol{\tau}$, with

$$E\, Z_\beta(\mathbf{t})\, Z_\beta(\mathbf{s}) = R_\beta(\mathbf{t}, \mathbf{s}).$$

R_β is the reproducing kernel of $\mathscr{H}_\beta$. 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 problem (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 an 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 $\mathbf{t}$ and $\mathbf{s}$, the odds ratio of $\mathbf{t}$ and $\mathbf{s}$ is

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

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

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

Often, we are interested in how one covariate affects the odds when other risk factors are fixed (at their medians or means). Suppose that 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 $\mathbf{t} = (t_1^2, t_2, \ldots, t_d)$ and $\mathbf{s} = (t_1^1, t_2, \ldots, t_d)$ equals

$$\log \widehat{\mathrm{OR}}(\mathbf{t}/\mathbf{s}) = \sum_{1 \in J_1} \{f_1(t_1^2) - f_1(t_1^1)\} + \sum_{(1,j) \in J_2} \{f_{1,j}(t_1^2, t_j) - f_{1,j}(t_1^1, t_j)\}$$
$$+ \sum_{(1,j_2,\ldots,j_d) \in J_d} \{f_{1,j_2,\ldots,j_d}(t_1^2, t_{j_2}, \ldots, t_{j_d}) - f_{1,j_2,\ldots,j_d}(t_1^1, t_{j_2}, \ldots, t_{j_d})\}.$$

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

Often, the SS ANOVA estimate of the probability function has relatively large biases at peak and trough points. These biases add if we pick t_1^1 at a peak and t_1^2 at a trough. This effect is obvious from the simulations in Section 4. A correction to the possible biases is necessary for these cases. We propose a bootstrap procedure to correct these biases:

(a) generate bootstrap samples of binary data from the SS ANOVA estimate of the probability function p;
(b) calculate SS ANOVA estimates of odds ratios from these bootstrap samples and denote the median of them as $\widehat{\mathrm{OR}}^*(\mathbf{t}/\mathbf{s})$;
(c) estimate the bias of the log-odds ratio by

$$\widehat{\mathrm{bias}} = \ln \widehat{\mathrm{OR}}(\mathbf{t}/\mathbf{s}) - \ln \widehat{\mathrm{OR}}^*(\mathbf{t}/\mathbf{s});$$

(d) calculate the bias-corrected estimate by

$$\widehat{\mathrm{OR}}_{\mathrm{corrected}}(\mathbf{t}/\mathbf{s}) = \exp\{2\,\widehat{\mathrm{OR}}(\mathbf{t}/\mathbf{s}) - \widehat{\mathrm{OR}}^*(\mathbf{t}/\mathbf{s})\}.$$

Similar bias correction in a nonparametric regression setting using the 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 practice. Simulations in Section 4 indicate that this procedure works very well.

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

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

δ^2 can be calculated from the formulae in theorem 1 in Wahba *et al.* (1995). See Wang (1994b) for details on calculations of posterior covariances. Hence we can approximate the distribution of OR($\mathbf{t}$/$\mathbf{s}$) by a log-normal distribution and construct the $100(1 - \alpha)\%$ Bayesian confidence interval as

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

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 that it does.

3. 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 *et al.* (1988, 1989) and references therein for a detailed description of the data and some analyses using linear logistic regression.

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

(a) *age*, age in years at the time of the base-line examination;
(b) *duration*, the duration of diabetes at the time of the base-line examination;
(c) *glycosylated haemoglobin*, a measure of hyperglycaemia;
(d) *pressure*, systolic blood pressure in millimetres of mercury.

The following model was used in Wang (1994a) (model IV):

$$\begin{aligned}\operatorname{logit}\{P(\text{age, duration, glycosylated haemoglobin, pressure})\} \\ = \mu + f_1(\text{age}) + a_1\,\text{duration} + a_2\,\text{glycosylated haemoglobin} + a_3\,\text{pressure}. \quad (12)\end{aligned}$$

The main effect of age is plotted in Fig. 1(a). We see that patients between 20 and 30 years of age are at higher risk. To compare the risk at some particular ages, we may want to calculate the odds ratios at these ages. Suppose that we fix the variables duration, glycosylated haemoglobin and pressure at their median values and pick age = 25 years as the base and compare its risk with the other ages. The estimated odds ratios and their 90% Bayesian confidence intervals are plotted in Fig. 1(b). We see that the odds at age = 25 years are significantly higher than the odds at age $\leqslant$ 13 years.

From Fig. 1, the odds at age = 25 years are not significantly higher than the odds at age = 40 years. However, since age = 25 years is a peak point, it is likely that $\widehat{\mathrm{OR}}$ overestimates the true OR. This is supported by our simulations in Section 4, where $\widehat{\mathrm{OR}} = 0.45$ and $\widehat{\mathrm{OR}}^* = 0.73$ in Table 1. We can estimate the bias of the log-odds by

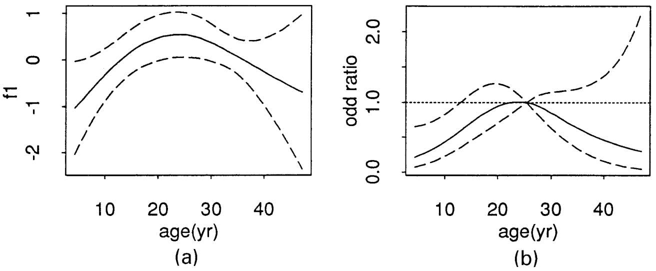

Fig. 1. (a) Estimates of the main effect f_1(age); (b) estimates of odds ratios OR{age/(age = 25 years)} (– – – –, 90% Bayesian confidence intervals)

$$\widehat{\text{bias}} = \ln \widehat{\text{OR}} - \ln \widehat{\text{OR}}^* = \ln 0.45 - \ln 0.73 = -0.48.$$

Then we can correct this bias in the original estimate:

$$\widehat{\text{OR}}_{\text{corrected}} = 0.45 \exp(-0.48) = 0.28.$$

The 95% Bayesian confidence interval for the corrected estimate of the odds ratio becomes (0.07, 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 conducted a simulation to evaluate the performance of the bias correction procedure.

In the first two simulations, cases A and B, we used the univariate logit functions

$$f(t) = \tfrac{1}{2}\beta_{10,5}(t) + \tfrac{1}{2}\beta_{7,7}(t) + \tfrac{1}{2}\beta_{5,10}(t) - 1,$$

$$f(t) = 3\{10^5 t^{11}(1-t)^6 + 10^3 t^3 (1-t)^{10}\} - 2$$

respectively, where $0 \leqslant t \leqslant 1$. $\beta_{p,q}$ 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 Fig. 2. Bernoulli responses y_i were generated on grid points $t_i = (i - 0.5)/n$, $i = 1, \ldots, n$, according to the true probability function, where n is the sample size. Two sample sizes were used: $n = 100$ and $n = 200$. In case A, we used $t = 0.2$ as the base and calculated odds ratios at points $t = 0.4$, $t = 0.6$, $t = 0.8$ and $t = 1$. In case B, we used $t = 0.5$ as the base and calculated odds ratios at points $t = 0.1$, $t = 0.2$, $t = 0.3$ and $t = 0.4$.

In the third simulation, we used the estimated probability function (12) as the true model. The design is the same as for the data. We call it case C. As in the above odds ratio calculations, we used age = 25 years as the base and calculated odds ratios at points age = 10, age = 15, age = 35 and age = 40 years.

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 labelled $\widehat{\text{OR}}$ with standard deviations in 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 labelled coverage with the 95% coverage number in parentheses. We conclude from Table 1 that these odds ratio estimates and their Bayesian confidence intervals work well. Estimates of OR are generally biased towards 1 if one of the two points of

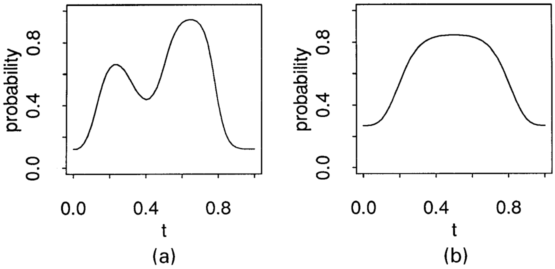

Fig. 2. Probability functions used in the simulations: (a) case A; (b) case B

TABLE 1
Odds ratios, estimates of odds ratios and coverages of Bayesian confidence intervals

	Results for case A for the following combinations of t:				*Results for case B for the following combinations of t:*				*Results for case C for the following combinations of age:*			
	0.4/0.2	*0.6/0.2*	*0.8/0.2*	*1/0.2*	*0.1/0.5*	*0.2/0.5*	*0.3/0.5*	*0.4/0.5*	*10/25*	*15/25*	*35/25*	*40/25*
OR	0.45	6.77	0.47	0.08	0.08	0.21	0.58	0.92	0.41	0.70	0.63	0.45
$n = 100$									$n = 256$			
$\widehat{OR}$	0.64 (0.73)	6.66 (34.69)	0.65 (1.08)	0.06 (0.56)	0.12 (0.10)	0.24 (0.18)	0.50 (0.82)	0.84 (0.58)	0.47 (0.20)	0.71 (0.24)	0.83 (0.61)	0.73 (0.57)
Coverage	86 (90)	89 (94)	81 (89)	92 (95)	82 (90)	89 (93)	98 (98)	99 (99)	84 (92)	93 (99)	73 (80)	75 (79)
$n = 200$												
$\widehat{OR}$	0.54 (0.38)	6.65 (10.57)	0.66 (0.44)	0.06 (0.13)	0.10 (0.06)	0.23 (0.11)	0.51 (0.34)	0.83 (0.22)				
Coverage	90 (95)	95 (97)	81 (92)	92 (97)	88 (94)	91 (95)	92 (94)	98 (100)				

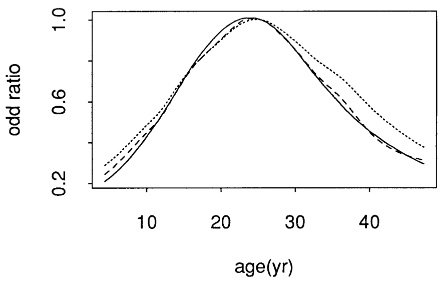

Fig. 3. True odds ratio function for case C (———), average of odds ratio estimates (·······) and average of bias-corrected odds ratio estimates (– – – –)

the OR is at the peak and/or the other is at the trough (for instance, 0.4/0.2 from case A). This is because the SS ANOVA estimate $\hat{f}$ may underestimate f at a peak and overestimate f at a trough. The coverages of Bayesian confidence intervals at high bias points are lower than the nominal value, whereas 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 the bias correction procedure, we use case C as our true model. At each of 100 replications the bootstrap bias correction procedure with 100 bootstrap samples is used to obtain bias-corrected estimates of odds ratios. Fig. 3 shows the true odds ratio function (full curve), the average of the odds ratio estimates (dotted curve) and the average of the bias-corrected odds ratio estimates (broken curve). The bootstrap correction procedure works very well.

Acknowledgements

This research was supported in part by National Institutes of Health grants R01 EY09946, P60 DK20572, P30 HD18258, U54 HD29184 and National Science Foundation grant DMS9121003. The author is indebted to Grace Wahba, Douglas Bates and Brian Yandell for their advice and encouragement. The author thanks Ronald Klein, MD, and Barbara Klein, MD, for providing the WESDR data (supported by National Institutes of Health grant EY03083).

References

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

Fan, J. and Hu, T. C. (1992) Bias correction and higher order kernel functions. *Statist. Probab. Lett.*, **14**, 235–243.

Gu, C. (1987) What happens when bootstrapping the smoothing spline? *Communs Statist. Theory Meth.*, **16**, 3275–3284.

———(1990) Adaptive spline smoothing in non-Gaussian regression models. *J. Am. Statist. Ass.*, **85**, 801–807.

Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. New York: 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. *J. Am. Med. Ass.*, **260**, 2864–2871.

———(1989) Is blood pressure a predictor of the incidence or progression of diabetic retinopathy? *Arch. Intern. Med.*, **149**, 2427–2432.

Nychka, D. (1988) Bayesian confidence intervals for smoothing splines. *J. Am. Statist. Ass.*, **83**, 1134–1143.

O'Sullivan, F., Yandell, B. and Raynor, W. (1986) Automatic smoothing of regression functions in generalized linear models. *J. Am. Statist. Ass.*, **81**, 96–103.

Wahba, G. (1983) Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. R. Statist. Soc.* 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. *Santa Fe Inst. Stud. Sci. Complex.*, **20**, 329–360.

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. *Adv. Neural Inform. Process.*, **6**, 415–422.

———(1995) Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. *Ann. Statist.*, **23**, 1865–1895.

Wang, Y. (1994a) Smoothing spline analysis of variance of data from exponential families. *PhD Thesis*. University of Wisconsin, Madison.

———(1994b) GRKPACK: fitting smoothing spline analysis of variance models to data from exponential families. *Communs Statist. Simuln Computn*, to be published.

Wang, Y. and Wahba, G. (1995) Bootstrap confidence intervals for smoothing splines and their comparison to Bayesian confidence intervals. *J. Statist. Computn Simuln*, **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. *Communs Statist. Simuln Computn*, **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. *Statist. Med.*, to be published.