SSC Annual Meeting, June 2001
Proceedings of the Survey Methods Section, pp. 143-148

SMOOTHING AND ADDITIVE MODELS FOR SURVEY DATA

David R. Bellhouse, Hugh A. Chipman and James. E Stafford[1]

ABSTRACT

Survey sampling is a statistical domain which has been slow to take advantage of flexible regression methods such as scatterplot smoothing and additive models. To make these methods more accessible, this paper introduces techniques that account for the complex survey structure of the data. We focus on smooth regression with a normal error model. This model is complicated by the survey design, which could include stratification, cluster sampling, and other complex survey designs. The presence of tied covariate values results in the smoothing of binned means. The estimation of smooths is seen to depend on the sampling design only via the sampling weights, meaning that standard software can be used for estimation. Inference for these curves is more challenging, due to correlations induced by the sampling design. We propose tests which account for the sampling design. Illustrative examples are given using the Ontario health survey.

KEY WORDS: Bootstrap; Binning; Cross validation; Mallow's Cp; Ontario health survey; Sampling; Scatterplot Smoothing.

RÉSUMÉ

L'échantillonnage est un domaine de la statistique qui tarde à prendre avantage des techniques flexibles de la régression comme le lissage de nuages de points et les modèles additifs. Afin de rendre ces méthodes plus accessibles, cet article présente des techniques tenant compte de la structure complexe des données d'enquête. Nous mettons l'emphase sur la régression lisse utilisant un terme d'erreur ayant une distribution normale. Ce modèle se complique si on est confronté à un plan stratifié, à de l'échantillonnage par grappes ou toutes formes de plans complexes. La présence de covariables aux valeurs égales crée des lissages de <<moyennes binned>> L'estimation de ces lissages est considérée comme dépendante du plan de sondage via les poids de sondage seulement. Ceci signifie que les logiciels informatiques usuels peuvent être utilisé pour l'estimation. L'inférence sur ces courbes devient un grand défi à cause de la corrélation induite par le plan de sondage. Dans cet article, nous proposons des tests qui tiennent compte du plan de sondage. Nous illustrons ces concepts avec des exemples utilisant des données de l'Enquête sur la santé de l'Ontario.

Mots clés : <<Binning>>; bootstrap; Cp de Mallow; échantillonnage; Enquête sur la santé de l'Ontario; lissage de nuages de points; validation croisée.

NOTE: This is a shortened version of Bellhouse, Chipman, and Stafford (2001) (hereafter BCS), which is currently available online.

1. INTRODUCTION

Although scatterplot smooths and additive regression models are effective tools for data exploration and flexible modelling, they have been under-utilized in the analysis of survey data. One of the main stumbling blocks is the complex survey structure, which invalidates many of the assumptions made when smoothing or fitting additive models.

These complications lead to practical and conceptual challenges. At a practical level, complex surveys yield a sample of observations which are neither independent nor identically distributed. Instead, observations have sampling weights and covariances, both of which are induced by the sampling design and the population structure. Flexible models must accommodate both sampling weights and the induced covariance structure. An additional practical challenge that occurs in this context but is not unique to survey data is that of efficient computation with large datasets. The conceptual challenge is to develop a theoretical framework which includes the sample, population and superpopulation, allowing correct

[1] David R. Bellhouse, Department of Statistical and Actuarial Sciences, University of Western Ontario, London, ON, N6A 5B7; Hugh A. Chipman; Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, N2L 3G1, hachipman@uwaterloo.ca; James E. Stafford, Department of Public Health Sciences, University of Toronto, Toronto, ON, M5S 1A8.

interpretation and inference for data originating from any complex survey design.

To motivate these challenges and give a flavour of the proposed approach, we introduce an example involving the 1990 Ontario health survey (OHS). The data consist of 33,355 individuals sampled in a stratified two-stage clustered design, with the strata corresponding to public health units. Within each stratum enumeration areas were selected by pps and then households were selected in each chosen EA. All residents within a sampled household were questioned. Here, we focus on models for body mass index (BMI), defined as weight in kilograms divided by the square of height in meters. BMI values less than 20 are associated with health problems like eating disorders and values over 27 with hypertension or heart disease.

Figure 1 displays two plots of BMI against age. The volume of raw data in the left plot obscures any trend, although the rounding of age to the nearest year is evident. In the right plot, the mean BMI for each of the 47 distinct ages is given, with plotting symbol size proportional to the number of observations in each age category. A smooth of the means with six degrees of freedom summarizes the trend. BMI appears to increase nonlinearly with age, flattening out at higher ages. This nonlinearity is statistically significant.

All the practical challenges come into play in smoothing the means. Weights were used in standard procedures for the estimation of the smoothing spline. The covariance between observations was used in a new hypothesis test of nonlinearity. Weights were also used in a modified cross-validation procedure to select an appropriate number of degrees of freedom.

In addition to weights and covariances, another feature of this data is the discreteness of the predictors, such as age in Figure 1. The estimation and inference methods developed in this paper capitalize on this commonly occurring feature, yielding fast estimation and central limit theorems for inference.

When multiple predictors $x_1, x_2, \ldots, x_p$ are available, a scatterplot smoother such as $E(Y) = g(x)$ displayed in Figure 1 can be extended to an additive model (Hastie and Tibshirani 1990, Green and Silverman 1994)

$$E(Y) = g_1(x_1) + g_2(x_2) + \ldots + g_p(x_p)$$

where $g_1, g_2, \ldots, g_p$ are estimated using scatterplot smoothing in a backfitting algorithm.

Figure 1 – Scatterplots of the Ontario Health Survey Data. Raw data (left) and mean BMI for each distinct level of age, with a 6df smooth overlaid (right).

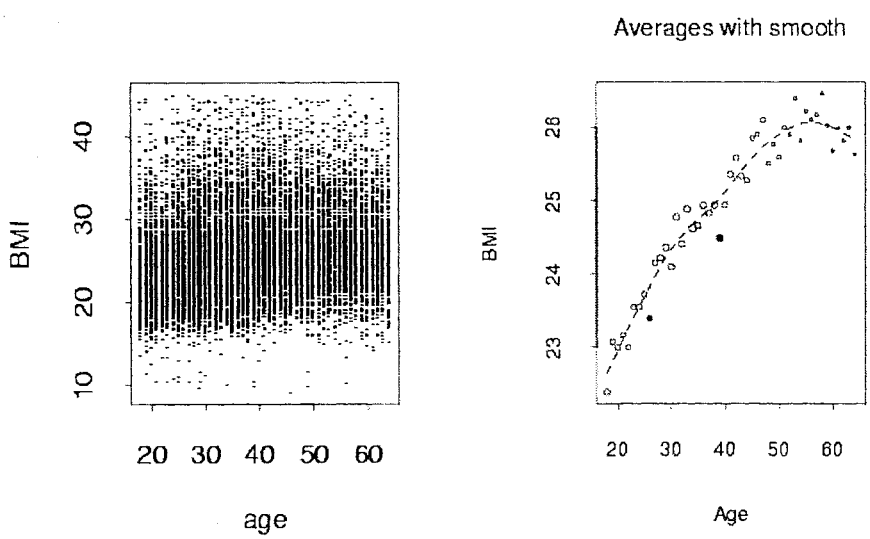
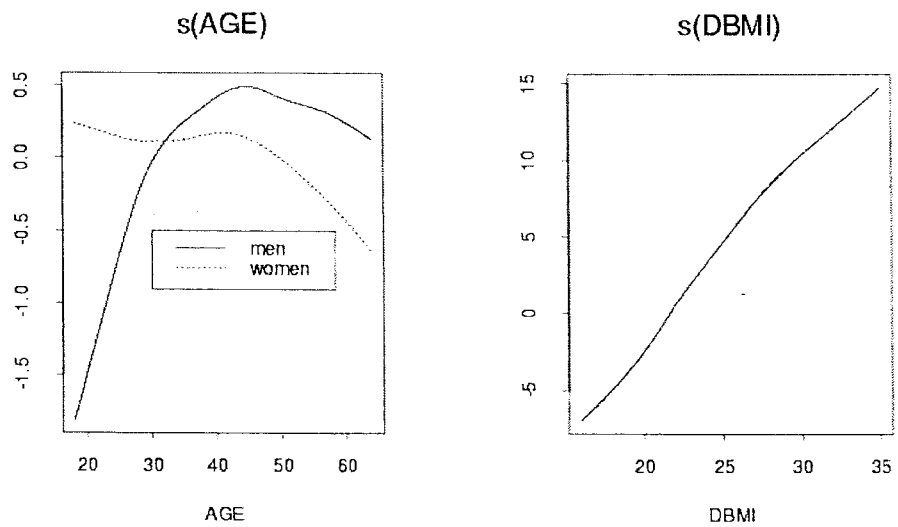


Figure 2 – Additive model for OHS data. Separate age smooths for men and women were estimated, and a common term for DBMI.



For the OHS data, we use age, gender, and DBMI (desired BMI) to predict BMI. Analysis in BCS suggests an interaction between age and gender, so the model is constructed using g_1(age,gender) and g_2(DBMI). Both terms are significant, and are displayed in Figure 2.

The remainder of the paper develops the theory and techniques necessary for the smoothing and additive modelling illustrated in this section. Section 2 briefly mentions estimation of scatterplot smooths and additive models. The survey sampling context is discussed in Section 3. Section 4 develops new testing procedures based on Wald statistics. In Section 5, we return to the OHS data and illustrate the techniques developed in the smoothing of BMI against age.

2. ESTIMATION OF SMOOTHS AND ADDITIVE MODELS IN THE SAMPLING CONTEXT

BCS discuss estimation of a scatterplot smooth and additive models for complex survey data. They show that conventional estimation methods may be used, and that in the scatterplot smoothing problem, the binned means for each unique covariate value can be treated as the raw data, provided that the associated sampling weights are used. In additive models, similar results are given, but with the important distinction that the binning is different along each covariate. This necessitates repeated binning of partial residuals during estimation.

3. THE SAMPLING CONTEXT

Generally the vector of distinct covariate values $\boldsymbol{x}$ arises from truncation that is either deliberate, such as an aggregation of responses into non-overlapping bins along the x-axis, or from reported digits of accuracy. For instance, in the OHS data age was reported in years only and anywhere from 500 to 800 people had a particular age. The assumption that $\boldsymbol{x}$ is the same in the population and sample has numerous technical advantages that ease interpretation. For instance, the vector $\boldsymbol{g}$ of expected values of Y corresponding to $\boldsymbol{x}$ has fixed length, meaning that sample and population smoothing matrices will have the same dimension. The asymptotic framework discussed in BCS assumes all k distinct values of the covariate are "observed" in the limit.

The central limit theorems developed for $\bar{\boldsymbol{y}}$ (an observed vector of bin means with expected value $\boldsymbol{g}$) are more important for survey sampling that in IID sampling. In survey sampling, a superpopulation may be composed of heterogeneous groups, making observed sample responses unlikely to follow the usual IID normal error model. Inference for $\boldsymbol{g}$ thus rests on the central limit theorems.

Different theoretical interpretations of the superpopulation are possible. For example, the finite population may be viewed as an IID sample from the superpopulation. This approach, which we shall refer to as "superpopulation 1" provides a natural motivation to scatterplot smoothing and additive models. A more accurate interpretation would be to consider a nested sequence of finite populations such that as the finite population size increases, the binned

means converge to g ("superpopulation 2"). BCS motivate their models in terms of superpopulation 1, and provide justification for the central limit theorem in terms of superpopulation 2.

4. TESTS OF HYPOTHESES

Standard hypothesis tests are inappropriate due to the structure of complex survey data. The difficulty in testing lies in the routine violation of normality assumptions and developing the appropriate tests relies on manipulating survey data so such assumptions apply.

Hypothesis tests can take advantage of the binned structure of the data to obtain normal approximations for $\bar{y}$ via central limit theorems. In addition, the variances of observations and covariances between observations must be accounted for. Therefore, approximations lead to Wald-type tests, which are less common in smoothing than likelihood ratio tests. We use Wald tests instead because they allow covariance structure to be incorporated at the level of the binned data.

4.1 Central limit theorems

Suppose we wish to test in the univariate case the finite population null hypothesis H_0:g(x) = g_0(x) against the alternative H_1:g(x) = g_1(x), where g_0 and g_1 are smooth functions with specific smoothing parameters. For example, to test linearity for the data in Figure 1, we take $g_0(x) = \beta_0 + \beta_1 x$ and $g_1(x)$ could be a six degree of freedom smooth. Finite population central limit theorems, and the assumptions of superpopulation 2, assert

$$\bar{y} \sim N(g_0(x), V) \qquad (1)$$

where $\bar{y}$ is the vector of response means corresponding to the unique values of the covariate and V is the covariance matrix of $\bar{y}$ with estimate $\hat{V}$. This can be used in tests based on sample estimates $\hat{\hat{g}}_0 = \hat{\hat{S}}_0 \bar{y}$ and $\hat{\hat{g}}_1 = \hat{\hat{S}}_1 \bar{y}$. The matrix $\hat{\hat{S}}_1$ is the smoother matrix, as discussed in BCS. Letting $C = \hat{\hat{S}}_1 - \hat{\hat{S}}_0$ we will then have

$$\hat{\hat{g}}_1 - \hat{\hat{g}}_0 = C\bar{y} \sim N(0, CVC')$$

This multivariate normal result will be used to construct Wald statistics for hypothesis testing.

Similar techniques are used for additive models, with binned partial residuals taking the place of the binned means $\bar{y}$.

4.2 A Wald test statistic

Assuming the approximation (1) is appropriate, Rao (1973, pg 188) asserts that

$$X^2 = \bar{y}'C'(CVC)^- C\bar{y} \sim \chi_r^2$$

where X^2 is commonly referred to as the Wald statistic and $r = rank(CVC') = rank(C) \le k$, where k is the number of distinct levels of x. The notation A^- denotes generalized inverse of A. In the usual additive models context an analogy with linear regression gives the trace of C as the appropriate degrees of freedom The implicit assumption is made that $C=C'$, $CC'=C$, and the eigenvalues of C are all either 0 or 1. Mimicking the usual context has appeal and use of this analogy here means approximating r by tr(C).

Use of X^2 requires computation of the estimate of CVC and $r=rank(C)$. Standard packages provide degree of freedom estimates that can be used to approximate r. Thus we need only concern ourselves with the estimation of CVC'. The obvious estimate is $C\hat{V}C'$. This can be justified for $C = \hat{\hat{S}}$ using the approach of Binder (1983), as described in BCS. The estimate $\hat{V}$ can be obtained through standard survey analysis software such as *SUDAAN* (Shah et. al. 1996).

BCS consider two additional tests. If only the diagonal elements of $\hat{V}$ are available a conservative test may be constructed. If available software packages do not provide the smoother matrix S, an alternate method can be used to calculate the test statistic X^2.

As an alternative to hypothesis tests, the smoothing penalty can be selected by cross-validation. BCS provide details, noting that regular approaches to cross-validation can be used provided that weights are accounted for.

The effectiveness of these tests and cross validation are assessed via two simulation studies in BCS. Those simulations demonstrate the critical role played by good estimates of variance. The X^2 statistic will perform poorly if a poor estimate of covariance is used, due to instabilities in the inversion of V.

4.3 Diagnostics

The covariance structure introduced for Wald tests can also be used for model diagnostics. Taking $C=I-S$ gives unstandardized residuals of $C\bar{y} = \bar{y} - \hat{\hat{g}}$. These residuals may be standardized by calculating

$$\bar{y}C(CVC')^{-1/2} = (y-g)C^{-}V^{-1/2}$$

If the null model (represented by the smoother matrix S) is appropriate, then these residuals should be independent standard normals. An illustration is given in Section 5.

5. EXAMPLE: PREDICTING BMI USING AGE

We first consider selection of the smoothing parameter via cross validation. Four different fitted lines are given in the scatterplot of BMI against age in Figure 3: a linear regression and smoothing splines with 3, 6, and 12 degrees of freedom. Comparing the plot to the binned means in Figure 1, it appears that the 3 df fit may be insufficient, while 12 df is perhaps overfitting the data. When the cross-validated residual sum of squares is used to select an appropriate amount of smoothing (or equivalently the degrees of freedom), six degrees of freedom seems optimal.

Tests for the goodness of fit, linearity and significance of the variable may also performed, using the Chi-square approximations of Section 4. The Wald statistics utilizing full covariance information are appropriate, and we consider a range of models, with 47, 20, 12, 9, 6, 3, 2, and 1 degrees of freedom. For each model, a sequence of tests against all models with smaller degrees of freedom are performed.

Figure 3 –Smooths of BMI on Age with a variety of degrees of freedom.

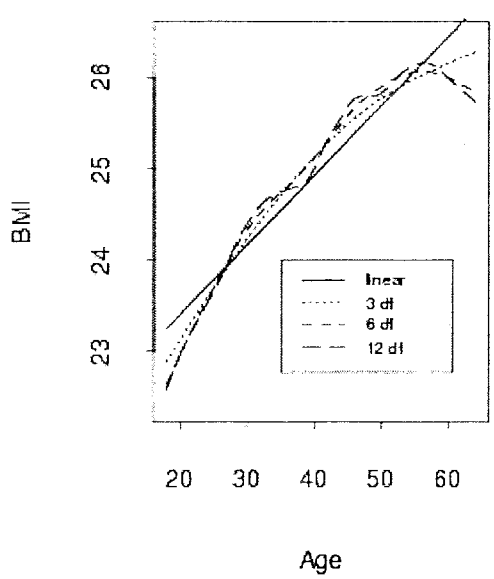


Table 1: Chi-Square test statistics X^2 for smoothing BMI on age, using full covariance information. Test marked by * and ** are significant at 0.05 and 0.01 levels respectively.

	20	12	9	6	3	2	1
47	42.0*	50.2*	52.7	61.4*	79.9**	176**	1247**
20		5.0	8.8	12.2	37.4**	122**	1181**
12			4.4	7.8	28.0**	117**	1169**
9				4.1	25.2**	115**	1166**
6					18.0**	107**	1154**
3						88**	1140**
2							1071**

The test statistics are given in Table 1. In the first row, various goodness of fit tests indicate that even some of the larger models may fail (at a 5% level) to capture all patterns in the data. The last three columns imply that that any model with 6 or more degrees of freedom cannot be simplified to a model with 3 df or less (at a 1% level). This roughly agrees with the results of the cross-validation.

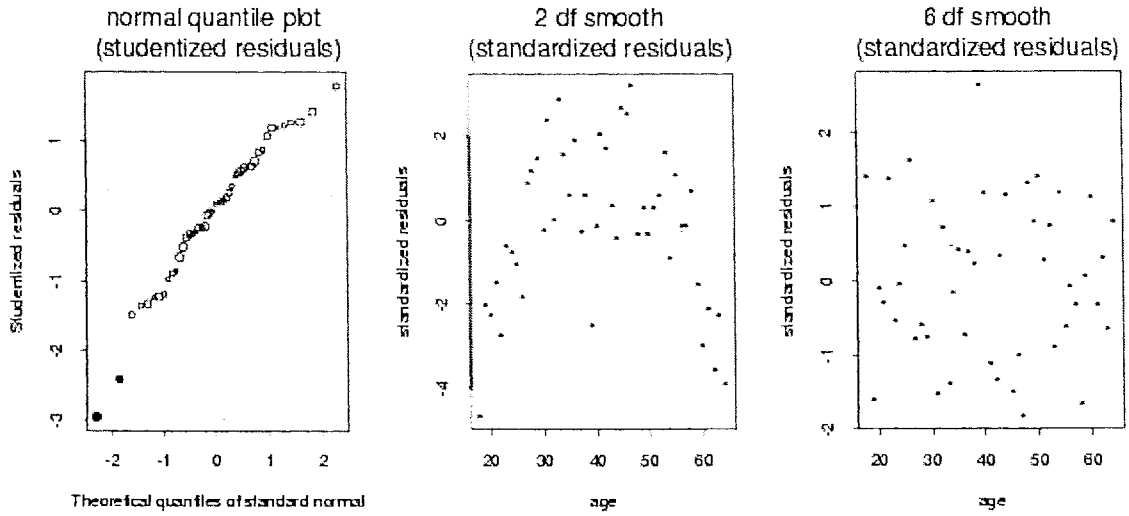
BCS consider alternative methods for hypothesis testing, including a bootstrap technique and a scenario in which only the diagonal of the covariance matrix is available. In this example, tests based on these methods were similar to the results given above.

We also consider residual diagnostics. The studentized residual, $e_i=(y_i - \hat{y}_i)/se(\hat{y}_i)$ can be calculated, and is useful to detect outliers or other anomalous patterns involving specific points. For example, the left panel of Figure 4 is a normal probability plot of residuals from a 6df model. Although the residuals are roughly normal, a few observations are outliers (indicated with filled circles in this figure and Figure 1). The largest negative outlier corresponds to 39-year-olds, who apparently go on a midlife crash diet.

Standardized residuals may also be calculated using the more general method described in Section 4.3. In Figure 4, the standardized residuals are plotted against age for a 2 and 6 degree of freedom model. Curvature in the residuals from the linear model indicates presence of a nonlinearity.

Standardized and studentized residuals have slightly different interpretations. An individual standardized residual does not correspond to a particular observation, since standardization involves taking linear combinations of the raw residuals. The studentized residuals can identify single anomalous points while the standardized residuals will identify

Figure 4 – Residual plots for smoothing BMI on age: Normal probability plot for studentized residuals, with two large residuals noted (left), and standardized residuals plotted against age, for the linear (2df) and 6df models (center and right respectively). Plotting characters of various sizes indicate weights for the associated points.



trends, nonlinearities, and violations of distributional assumptions that apply to collections of points. For example, a normal probability plot of standardized residuals would assess the assumption of normality rather than identify outliers (as in Figure 4).

Finally, we note that although BMI has a significant nonlinear relationship with age, there remains considerable unexplained variation in the data. This may be seen by comparing the range of the 6df smooth with the range of the data in Figure 1. A weighted version of the coefficient of determination,

$$R^2 = 1 - \left(\sum_{i=1}^{n} (y_i - \hat{\hat{y}}_i)^2 w_i \Big/ \sum_{i=1}^{n} (y_i - \bar{y}_i)^2 w_i \right)$$

quantifies the explained variation. Here, y_i is an individual response in the sample with corresponding weight w_i and fitted value $\hat{\hat{y}}_i$. The weighted sample mean is the scalar $\bar{y}$. For the smoothing of BMI on age, $R^2 = 5.6\%$.

Another possible analysis is to fit separate smooths of BMI on age for men and women. BCS give details on this, finding for the OHS data that there are significant differences in the smooths for men and women.

ACKNOWLEDGMENTS

We wish to thank Rob Tibshirani and David Binder for very useful suggestions. We also wish to thank MITACS and the Natural Sciences and Engineering Research Council of Canada for support.

REFERENCES

Bellhouse, D. R.. Chipman, H. A., and Stafford, J.E. (2001) Smoothing and additive models for survey data. Technical report, Department of Statistics and Actuarial Science, University of Waterloo. Available online at www.stats.uwaterloo.ca/~hachipma/

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*. 51, 279-292.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric regression and generalized linear models: A roughness penalty approach* London: Chapman and Hall.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*, London: Chapman and Hall.

Rao (1973). *Linear statistical inference and its applications*. 2nd Edition, New York: John Wiley.