Smoothing and additive models for survey data

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

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. In an attempt to make these methods more accessible, this paper introduces techniques that account for the complex survey structure of the data. The use of penalized least squares in the sampling context is studied as a tool for the analysis of a general trend in a finite population. We focus on smooth regression with a normal error model. This model is complicated by the design of the survey, which could include stratification, clustering and more than one stage of sampling. Ties in covariates abound for large scale surveys resulting in the application of scatterplot smoothers to means. The estimation of smooths (for example smoothing splines) 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 and illustrate tests which account for the sampling design. Illustrative examples are given using the Ontario health survey, including scatterplot smoothing, additive models, model diagnostics, and various significance tests. Simulation studies are presented to assess the accuracy of the proposed methods.

Keywords: backfitting; bootstrap; binning; cross-validation; diagnostics; Ontario health survey; penalized least squares; sampling; scatterplot smoothing; variance estimation.

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, each observation has an associated sampling weight, and the sampling design induces a covariance structure on the sampled observations. 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 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, Statistics Canada census areas, called enumeration areas, were selected with probability proportional to the size of the area and then households were selected in each chosen enumeration area. 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.

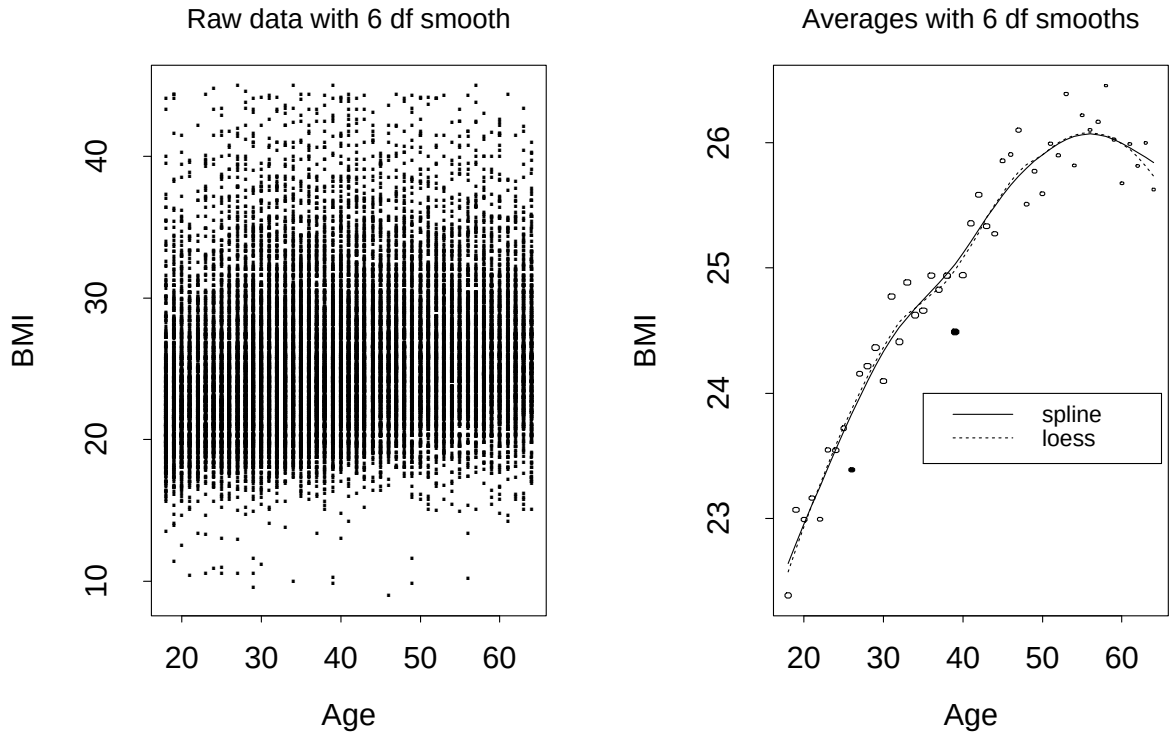

Figure 1: Scatterplots of the Ontario Health Survey data. Raw data with a 6 df smoothing spline overlaid (left), and mean BMI for each distinct level of age (right), with 6 degree of freedom smoothing spline and loess curve overlaid.

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. Two smooths of the means with approximately six degrees of freedom are overlaid, summarizing the trend. The smooths are a locally quadratic regression (Cleveland and Devlin, 1988) and a cubic smoothing spline. BMI appears to increase nonlinearly with age, flattening out at higher ages. This nonlinearity is statistically significant. Normal QQ plots reveal that the two solid scatterplot points have

sizable residuals.

All the practical challenges come into play in smoothing the means. Weights were used in standard procedures for the estimation of the smoothing spline and local regression. The covariance between observations was used in a new hypothesis test of nonlinearity. Weights were also used in a 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 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 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. For the OHS data, we use `age`, `gender`, and `DBMI` (desired BMI) to predict `BMI`. Analysis in §6.2 suggests an interaction between age and gender, so the model is constructed using $g_1(\texttt{age}, \texttt{gender})$ and $g_2(\texttt{DBMI})$. Both terms are significant, and are displayed in Figure 2.

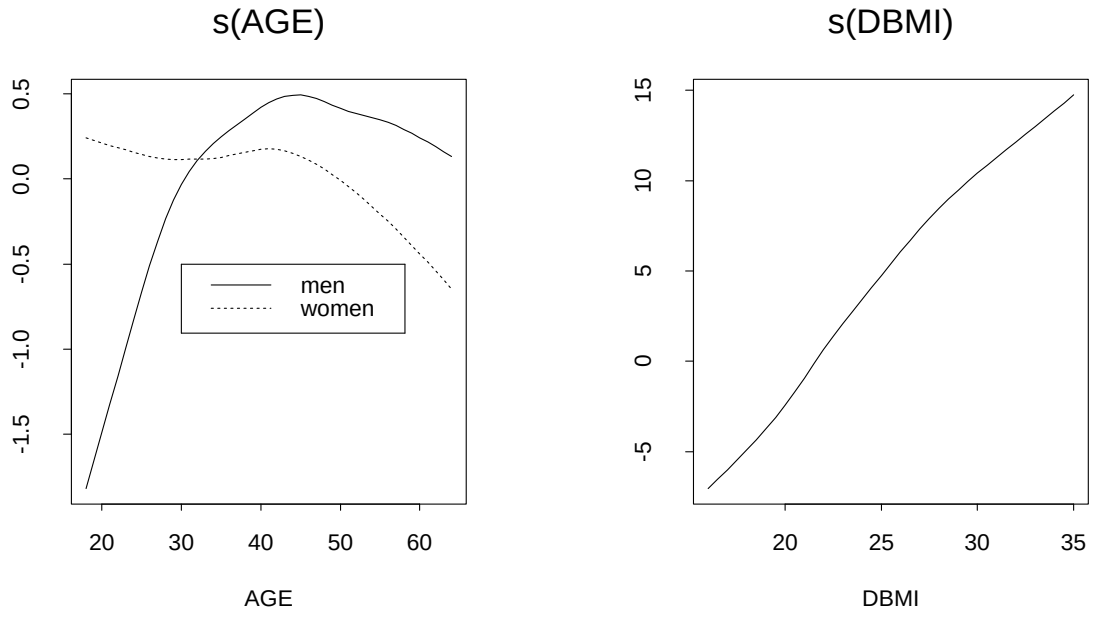

Figure 2: Additive model for OHS data. Separate `age` smooths for men and women were estimated, and a common term for `DBMI`.

Many different methods have been developed for scatterplot smoothing and additive models. In this paper we focus on linear smoothers, a class of methods given as the solution to a penalized least squares criterion. We adopt the viewpoint of Green and Silverman (1994) and Hastie and Tibshirani (1994) that penalized least squares offers a unifying framework that promotes clarity. The remainder of the paper develops the theory and techniques necessary

for the smoothing and additive modelling illustrated in this section. §2 reviews scatterplot smoothing via penalized least squares, and the application of scatterplot smoothers in additive models. In §3, we introduce the sampling context and discuss the conceptual challenge of flexible modelling of survey data. New testing procedures based on Wald statistics for scatterplot smoothers are developed in §4 along with the related issue of variance estimation. Modified versions of methods for selecting the amount of smoothing (or equivalently the degrees of freedom) such as cross-validation are also presented as alternatives in the presence of no covariance information. In the simulation studies of §5, Wald tests are shown to be accurate where variance estimation is critical in obtaining valid inferences. The usual F-tests, which ignore variance, are inflated. In addition, cross-validation is shown to behave in a reasonable fashion. In §6, we return to the OHS data and look at two full applications: the smoothing of BMI against age for men and women, and an additive model for BMI.

2 Scatterplot smoothing and additive models

The figures in the introduction display fits obtained using smoothing routines applied to means of complex survey data. This section discusses estimation via penalized least squares of g for scatterplot smoothing (§2.1) and $g_1, \ldots, g_p$ for additive modelling (§2.2) where linear regression and cubic smoothing splines are simply special cases.

One main point is that estimation, with weights and binned survey data, can be used without further modification to existing software. Another is that inferences rely on the binning of responses due to tied covariate values. Hence, when compared to Green and Silverman (1994) or Hastie and Tibshirani (1990), the developments of this section differ only in the increased emphasis on such ties.

2.1 Penalized least squares for scatterplot smoothing

We begin with the scatterplot smoothing problem, with only one covariate. For independent data, the model would typically be $Y = g(x) + \epsilon$, with $\epsilon \sim F$, for some error distribution F. For this model let $\mathbf{y}$ denote the vector of n responses, $\mathbf{x}$ the vector of k distinct values of the covariate and $\mathbf{g} = g(\mathbf{x})$ the k-vector of corresponding expected values of the response. The presence of tied covariate values can be made explicit through the use of an incidence matrix. Here the value taken by the covariate for each observation is indicated by an $n \times k$ matrix $\mathtt{I}$ with entries

$$\mathtt{I}_{ij} = \begin{cases} 1, & \text{if observation } i \text{ has the } j\text{th covariate value} \\ 0, & \text{otherwise} \end{cases},$$

$i = 1, \ldots, n,\ j = 1, \ldots, k$. Exactly one element of each row of $\mathtt{I}$ will be equal to 1 and the model for the sample can be expressed as

$$E[\mathbf{y}|\mathtt{I}] = \mathtt{I}\mathbf{g}$$

as in Green and Silverman (1994). The k-vector of sample means $\bar{\mathbf{y}}$ for the distinct covariate values, such as that displayed in Figure 1, may be computed as $\bar{\mathbf{y}} = (\mathtt{I}'\mathtt{W}\mathtt{I})^{-1}\mathtt{I}'\mathtt{W}\mathbf{y}$, where $\mathtt{W}$

is the $n \times n$ diagonal matrix of sample weights and $\mathtt{IWI}$ is another diagonal matrix whose non-zero entries are the total weight for each mean. The sample means, $\bar{\mathbf{y}}$, are trivially the solution of the least squares criterion

$$(\mathbf{y} - \mathtt{I}\mathbf{g})'\mathtt{W}(\mathbf{y} - \mathtt{I}\mathbf{g})$$

for which the addition of a penalty

$$(\mathbf{y} - \mathtt{I}\mathbf{g})'\mathtt{W}(\mathbf{y} - \mathtt{I}\mathbf{g}) + \mathbf{g}'(S^{-} - \mathtt{I}'\mathtt{WI})\mathbf{g} \tag{1}$$

results in a scatterplot smooth as the solution $\hat{\mathbf{g}} = \widehat{\mathtt{S}}\bar{\mathbf{y}}$, where $\widehat{\mathtt{S}} = S\mathtt{I}'\mathtt{WI}$. Throughout the paper we denote sample estimates as $\hat{g}$ and finite population estimates with $\tilde{g}$. Hastie and Tibshirani (1990) use the above criterion to demonstrate that all linear smoothers optimize this penalized least squares criterion. For example, if $S = (\mathtt{I}'\mathtt{WI} + \alpha K)^{-1}$ for an appropriately chosen $k \times k$ matrix K, then the solution $\hat{\mathbf{g}}$ is a cubic smoothing spline.

Inference using smooths, $\hat{\mathbf{g}}$, relies on the above model and, in particular, the error distribution F. Unfortunately neither hold for complex survey data. In the sampling context $\mathbf{y}$ is a subvector of the analogous N-dimensional population vector $\mathbf{Y}$ and not a sample from the above model. However, the presence of tied covariate values permits the decomposition:

$$(\mathbf{y} - \mathtt{I}\bar{\mathbf{y}})'\mathtt{W}(\mathbf{y} - \mathtt{I}\bar{\mathbf{y}}) + (\bar{\mathbf{y}} - \mathbf{g})'\mathtt{I}'\mathtt{WI}(\bar{\mathbf{y}} - \mathbf{g}) + \mathbf{g}'(S^{-} - \mathtt{I}'\mathtt{WI})\mathbf{g}, \tag{2}$$

where only the second and third terms depend on $\mathbf{g}$. Thus smoothing the original data is equivalent to smoothing the means of the response for each unique covariate value (i.e., each bin). The advantage is that some convenient error distribution may now apply to the means $\bar{\mathbf{y}}$ where none existed for $\mathbf{y}$ itself. The use of central limit theorems in this respect is discussed later in §3. Note weights are now given by the diagonal of $\mathtt{I}'\mathtt{WI}$ where $\mathtt{W}$ are sampling weights, reflecting the number of population units represented by each sampled point. Mathematically they are treated in the same way as variance weights. Practically their effect is different: instead of variance stabilization, they can improve bias and consistency of estimates.

2.2 Penalized least squares for additive models

This framework can be extended to include additive models of the form $Y = g_1(x_1) + g_2(x_2) + \ldots g_p(x_p) + \epsilon$ for p covariates. Assuming an additive form has special advantages in the sampling context: binning can be done separately along each covariate, keeping the number of bins small and permitting central limit theorems to be applied. For a more general model, $Y = g(x_1, \ldots, x_p) + \epsilon$, binning of responses would only occur for observations where all covariates are tied. The curse of dimensionality suggests such bins would be sparse and central limit theorems would not apply.

We assume that covariate j takes k_j distinct values given by the k_j-vector $\mathbf{x}_j$. Each additive term g_j in the model has a corresponding vector $\mathbf{g}_j = g(\mathbf{x}_j)$, and a $n \times k_j$ incidence matrix $\mathtt{I}_j$. Then the vector of sample observations $\mathbf{y}$ have expected value given by

$$\mathrm{E}(\mathbf{y}|\mathtt{I}_1, \ldots, \mathtt{I}_p) \;=\; \mathtt{I}_1\mathbf{g_1} + \ldots + \mathtt{I}_p\mathbf{g_p}.$$

Following Hastie and Tibshirani (1990) we can extend the penalized least squares criterion (1) to the multiple predictor case, yielding

$$(\mathbf{y} - \sum_{j=1}^{p} \mathtt{I}_j \mathbf{g_j})' \mathtt{W} (\mathbf{y} - \sum_{j=1}^{p} \mathtt{I}_j \mathbf{g_j}) + \sum_{j=1}^{p} \mathbf{g_j}'(S_j^- - \mathtt{I}_j' \mathtt{W} \mathtt{I}_j) \mathbf{g_j}. \tag{3}$$

The standard backfitting algorithm (see Hastie and Tibshirani 1990) exploits additivity by optimizing this criterion one covariate at a time. When optimizing the criterion for the j^{th} covariate it may be re-expressed in terms of the partial residuals $\mathbf{e}_j = (\mathbf{y} - \sum_{l \neq j} \mathtt{I}_l \hat{\mathbf{g}}_\mathbf{l})$

$$(\mathbf{e}_j - \mathtt{I}_j \mathbf{g_j})' \mathtt{W} (\mathbf{e}_j - \mathtt{I}_j \mathbf{g_j}) + \mathbf{g_j}'(S_j^- - \mathtt{I}_j' \mathtt{W} \mathtt{I}_j) \mathbf{g_j} + \Upsilon$$

where $\Upsilon = \sum_{l \neq j} \mathbf{g_l}'(S_l^- - \mathtt{I}_l' \mathtt{W} \mathtt{I}_l) \mathbf{g_l}$ does not depend on $\mathbf{g_j}$. The situation is then completely analogous to §2.1 so that the ultimate solution $\hat{\mathbf{g}}_\mathbf{j} = \widehat{\mathtt{S}}_j \bar{\mathbf{e}}_j$, where $\widehat{\mathtt{S}}_j = S_j(\mathtt{I}_j \mathtt{W} \mathtt{I}_j)$, involves smoothing the binned means $\bar{\mathbf{e}}_j = (\mathtt{I}_j \mathtt{W} \mathtt{I}_j)^{-1} \mathtt{I}_j \mathtt{W} \mathbf{e}_j$ for which central limit theorems apply. This is again of particular importance because the survey data is not a sample from the above model. Note the means $\bar{\mathbf{e}}_j$ are simply

$$\bar{\mathbf{e}}_j = \bar{\mathbf{y}}_\mathbf{j} - (\mathtt{I}_j' \mathtt{W} \mathtt{I}_j)^{-1} \sum_{l \neq j} \mathtt{I}_j' \mathtt{W} \mathtt{I}_l \hat{\mathbf{g}}_l$$

and, if necessary, smoothing $\mathbf{e}_j$ can be reduced from being an $O(N)$ computation to $O(k_j)$ since $\bar{\mathbf{y}}_\mathbf{j}$, $\mathtt{I}_j' \mathtt{W} \mathtt{I}_j'$ and $\mathtt{I}_j' \mathtt{W} \mathtt{I}_l$ can all be computed before any cycling of the backfitting algorithm. The means $\bar{\mathbf{y}}_j = (\mathtt{I}_j \mathtt{W} \mathtt{I}_j)^{-1} \mathtt{I}_j \mathtt{W} \mathbf{y}$ are simply those computed for the distinct values of $\mathbf{x}_j$.

Both of the unconventional aspects of smoothing in the sample case (weights and binning) are implemented in many smoothing packages, such as `smooth.spline` and `loess` in S (Chambers and Hastie 1992) and R (Ihaka and Gentleman 1996). This means that existing software can be used without modification for estimation. We will show later in §4 that non-standard calculations are necessary for inference, due to correlations among sampled observations. Such correlations do not enter into estimation, since the penalized least squares criterion makes no assertions about independence of the data.

3 The sampling context

In inference the motivation for use of criteria like (1) and (3) lies in model assumptions that do not apply to survey data. For example, in the univariate case neither $E[\mathbf{y}|\mathtt{I}] = \mathtt{I}\mathbf{g}$ nor $E[\bar{\mathbf{y}}|\mathtt{I}] = \mathbf{g}$ hold. This then begs the question: what is $\hat{\mathbf{g}}$ or $\hat{\mathbf{g}}_j$ estimating? In this section we address this question for the univariate case only. The generalization to the additive case is immediate.

In the sampling context (1) has an entirely different motivation as a sample estimate of an analogous finite population quantity

$$(\mathbf{Y} - \mathcal{I}\mathbf{g})'(\mathbf{Y} - \mathcal{I}\mathbf{g}) + \mathbf{g}'(S^- - \mathcal{I}'\mathcal{I})\mathbf{g}. \tag{4}$$

The usual least squares criterion for regression in sampling is motivated this way (Kish and Frankel 1974). The $N \times k$ matrix $\mathcal{I}$ is an incidence matrix for the population vector $\mathbf{Y}$ and

the covariate $\mathbf{x}$ where we make the key assumption that, while there is a population covariate vector that differs from its sample counterpart, the distinct values of the covariate are the same for both the sample and population. Generally $\mathbf{x}$ 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 first example the covariate "age" was reported in years only and anywhere from 500 to 800 people had a particular age.

A decomposition similar to (2) results in the solution $\tilde{\mathbf{g}} = \widetilde{\mathbb{S}}\bar{\mathbf{Y}}$ where $\bar{\mathbf{Y}} = (\mathcal{I}'\mathcal{I})^{-1}\mathcal{I}'\mathbf{Y}$ and $\widetilde{\mathbb{S}} = S\mathcal{I}'\mathcal{I}$ so that in the special case of a smoothing spline $S = (\mathcal{I}'\mathcal{I} + \alpha K)$ with K as before. The assumption that $\mathbf{x}$ is the same in the population and sample is now seen to have some technical advantages. Developments now have some familiar sampling characteristics where, for example, $\bar{\mathbf{y}}$ is a sample estimate of the population parameter $\bar{\mathbf{Y}}$. Similarly, $\hat{\mathbf{g}}$ may be regarded as an estimate of the finite population quantity $\tilde{\mathbf{g}}$ since the parameter $\mathbf{g}$ has fixed length for the sample and population. Finally, the smoothing smoothing matrices, $\widetilde{\mathbb{S}}$ & $\widehat{\mathbb{S}}$, have the same dimension.

It follows that use of $\hat{\mathbf{g}}$ to conduct inference for $\mathbf{g}$ will depend on the relationship between $\tilde{\mathbf{g}}$ and $\mathbf{g}$. One obvious approach to formalizing this relationship is to assume that the model which motivates the criteria, but does not hold for the sample, does in fact hold for the finite population as a superpopulation model:

> **Superpopulation 1:** The units of a finite population are realizations of the model $Y = g(x) + \epsilon$, where $\epsilon \sim IID\ F$, such that the group means have $E[\bar{\mathbf{Y}}] = \mathbf{g}$.

A practical consequence of the superpopulation assumption is that the (abstract) process of smoothing in the finite population is operationally identical to the classical setting (smoothing an IID sample). This coupled with Binder's (1983) estimating equations approach (see §4.3) permits the assertion

$$\hat{\mathbf{g}} \longrightarrow \tilde{\mathbf{g}} \longrightarrow \mathbf{g}$$

where here the notion of convergence is not made rigorous. Thus, in addition to its familiar interpretation, the advantage of this model is the ease with which one can create analogies with the usual smoothing context permitting technology to be transferred to the sampling context. This is a standard strategy is survey data analysis. A relatively simple model on the population is assumed whose parameter estimates under a census become the finite population parameters of interest. These finite population parameters are estimated from the survey data. The approach is discussed in Pfefferman (1993) and originates in the work of Kish and Frankel (1974) with respect to regression analysis. In their context, $g(x) = \beta_0 + \beta_1 x$ and the classical least squares estimates, denoted by $\tilde{\beta}_0,\ \tilde{\beta}_1$, become the finite population parameters. Their survey estimates, denoted by $\hat{\beta}_0,\ \hat{\beta}_1$, are obtained from weighted least squares through the sampling weights.

Unfortunately superpopulation 1 may not be an accurate representation of the finite population. For example, the IID assumption may not be reasonable. In desiring to retain the advantages of this type of superpopulation model several authors have considered models that reflect the finer micro structure within a population. This has typically been in the context of linear regression. In doing so they retain helpful technical tools like expectation $\mathbf{E}$. However, micro models usually means micro estimates and $\hat{\mathbf{g}}$ becomes some complicated

weighted average of estimates so that simplicity is lost. See, for example, Scott and Holt (1982), Konijn (1962) and Pfeffermann and LaVange (1989). In addition, when a micro model is assumed limit theorems become much more difficult to obtain (Valliant et al. 2000, pp. 34 - 35).

We are interested in the general structure of the population not the ingredients, or micro models, from which the structure was obtained. We would like to explore this general structure simply without the need for appealing to the micro model. Further, we would like to appeal to a finite population central limit theorem as a basis of inference via tests of significance. Although these objectives are met by the above approach, in light of its inadequacy we adopt a broader superpopulation model where the approach is still valid:

> **Superpopulation 2:** There exists a nested sequence of finite populations indexed by N such that as $N \to \infty$ we have $\bar{\mathbf{Y}} \to \mathbf{g}$. The underlying function $g(x)$ which gives rise to $\mathbf{g}$ is assumed to be continuous.

The model asserts $\bar{\mathbf{Y}} \to \mathbf{g}$ regardless of the finer micro structure that may exist and furthermore that the underlying function $g(x)$ is still a smooth function so that simplicity is retained.

In the context of this model we appeal to the asymptotic framework in a nested sequence of finite populations put forward in Shao (1996, pp. 210 - 211) to obtain central limit theorems. There a number of conditions are given in addition to the nested sequence of finite populations. These include a Liapunov-type condition related to a mean as well as a condition that the variance of that mean is asymptotically greater than 0. These conditions imply in our context that each bin proportion is asymptotically positive. In addition there are conditions that no survey weight is disproportionately large and that the total number of first-stage sampled clusters is large. Under these conditions $\bar{\mathbf{y}} \sim N(\bar{\mathbf{Y}}, V)$ in the limit and since $\bar{\mathbf{Y}} \to \mathbf{g}$ we assert

$$\bar{\mathbf{y}} \dot{\sim} N(\mathbf{g}, V).$$

Consequently, to estimate $\mathbf{g}$ we smooth $\bar{\mathbf{y}}$ and inference for $\mathbf{g}$ rests on these central limit theorems where, as we shall see in §4, 5 stable estimation of V is crucial.

In general our strategy involves the use of both superpopulations. We use Superpopulation 1 to define the finite population parameters of interest and then find survey estimates of these parameters. The distribution of the estimates is obtained under Superpopulation 2. In other words we use the first approach to motivate techniques and the second to justify them. Bellhouse and Stafford (1999, 2001) consider such issues in the context of kernel density estimation and local polynomial regression. Here integration is a key technical tool, which under the second superpopulation assumption depends crucially on truncation and binning as limits of Riemann sums. Such asymptotic calculations are not considered in the context of this paper.

4 Tests of hypotheses

In addition to estimating non-linear structure in survey data, inference about this structure is of interest. Questions concerning the goodness of fit, linearity and significance of a covariate

can be summarized by a hypothesis. Standard tests are inappropriate due to the structure of complex survey data. This problem does not occur in estimation, where we noted that standard techniques that account for weights and binning can be used without modification. The difficulty in testing lies in the absence of a model for the sample. However, a model can be induced by taking advantage of the binned structure of the data to obtain normal approximations for $\bar{\mathbf{y}}$. In addition, the variances and covariances of $\bar{\mathbf{y}}$, that must be accounted for, lead to Wald-type tests which allow covariance structure to be incorporated at the level of the binned data. The developments in this section include immediate extensions to the additive models context.

4.1 More on central limit theorems

Suppose we wish to test in the univariate case the 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. 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

$$\bar{\mathbf{y}} \dot{\sim} N(\mathbf{g_0}, V), \tag{5}$$

where V is the covariance matrix of $\bar{\mathbf{y}}$ with estimate $\widehat{V}$. This can be used in tests based on sample estimates $\hat{\mathbf{g}}_0 = \widehat{\mathbf{S}}_0 \bar{\mathbf{y}}$ and $\hat{\mathbf{g}}_1 = \widehat{\mathbf{S}}_1 \bar{\mathbf{y}}$. Letting $C = \widehat{\mathbf{S}}_1 - \widehat{\mathbf{S}}_0$ we will then have

$$\hat{\mathbf{g}}_1 - \hat{\mathbf{g}}_0 = C\bar{\mathbf{y}} \dot{\sim} N(\mathbf{0}, CVC') \tag{6}$$

and this multivariate normal result can be used to construct Wald statistics for hypothesis testing. QQ plots and other residual diagnostics also use this result in the construction of standardized residuals. Such plots are used and discussed in simulations (§5.1) and in the scatterplot smoothing example (§6.1).

The analogy in the additive models context involves testing a hypothesis of the form $H_0 : g_j(x) = g_{0j}(x)$ against $H_1 : g_j(x) = g_{1j}(x)$. Normal approximations for partial residuals

$$\bar{\mathbf{e}}_j \dot{\sim} N(\mathbf{g}_j, V_j), \tag{7}$$

lead to

$$C_j \bar{\mathbf{e}}_j \dot{\sim} N(\mathbf{0}, C_j V_j C_j') \tag{8}$$

as the basis for constructing Wald test statistics. Here, as in the univariate case, V_j is a covariance matrix with some estimate $\widehat{V}_j$ and C_j is appropriately chosen depending on the form of the alternative hypothesis. A test for the goodness of fit combines all such tests for each dimension. Letting $C_j = I - \widehat{\mathbf{S}}_j$ the goodness of fit test is defined as $G = min\{\bar{\mathbf{e}}_j' C_j' (C_j \hat{V}_j C_j')^{-1} C_j \bar{\mathbf{e}}_j\}$.

4.2 A Wald test statistic

Assuming the approximation (5) is appropriate, Rao (1973, pg 188) asserts that

$$X^2 = (C\bar{\mathbf{y}})' (\widehat{CVC'})^{-} C\bar{\mathbf{y}} \dot{\sim} \chi^2_r$$

where X^2 is commonly referred to as the Wald statistic and $r = \text{rank}\{CVC'\} = \text{rank}(C) \leq k$. 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 systematically approximating r by trC.

However, the actual structure of a smoother matrix has the first $df_1 - df_0$ eigenvalues as large, and the remaining small, where the alternative and null models have degrees of freedom df_1 and df_0. This is illustrated in the left panel of Figure 3, which shows the eigenvalues of C corresponding to tests of 47 against 6 degrees of freedom. The vertical line is drawn just to the right of 41, which is $\text{trace}(C) = df_1 - df_0 = 47 - 6 = 41$ in the smoothing of BMI on age. In the Figure we can see that the eigenvalues of C fall from a maximum of 1 to a minimum of 0 quite quickly, whereas the eigenvalues of CVC' decay more smoothly. The third plot results from an orthogonal decomposition of the test statistic using principal components that is discussed below. It is evident from these plots that since C is not a projection matrix the true rank r exceeds the trace of C. Use of trC for the degrees of freedom will likely lead to inaccurate coverage as is evident in the sharp sharp increase in the value of the test statistic.

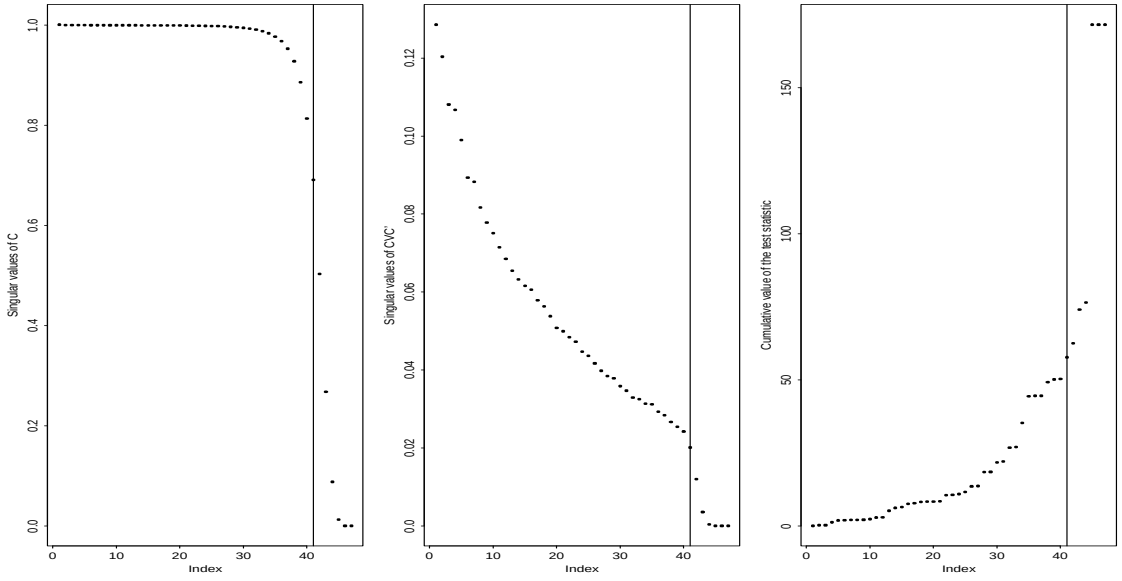

Figure 3: Eigenvalues of $C = S_1 - S_0$, the difference of two smoother matrices (left panel), and of $C\widehat{V}C'$ (centre panel). The right panel gives the cumulative value of the corresponding test statistic as more terms are included. S_1 has 47 degrees of freedom, while S_0 has 6.

The approximation can be improved by truncating the eigenvalues at trC by setting to zero those that exceed this threshold. Assuming $C\widehat{VC}'$ is available, following Rao (1973, pg 590) it can be decomposed as

$$C\widehat{VC}' = \lambda_1 \mathbf{p}_1 \mathbf{p}_1' + \cdots + \lambda_r \mathbf{p}_r \mathbf{p}_r'$$

where $\lambda_1, \ldots, \lambda_r$ and $\mathbf{p}_1, \ldots, \mathbf{p}_r$ are the eigenvalues and eigenvectors of $C\widehat{VC}'$ respectively. However $C\widehat{VC}'$ is calculated, it must be inverted to obtain the generalized inverse $(C\widehat{VC}')^-$ as

$$(C\widehat{VC}')^- = \mathbf{p}_1\mathbf{p}_1'/\lambda_1 + \cdots + \mathbf{p}_r\mathbf{p}_r'/\lambda_r$$

This permits an orthogonal decomposition of the test statistic X^2 in terms of the principal components, $u_1, \ldots, u_r$, of $C\bar{\mathbf{y}}$

$$(C\bar{\mathbf{y}})'(C\widehat{VC}')^-C\bar{\mathbf{y}} = u_1^2/\lambda_1 + \cdots + u_j^2/\lambda_j + \cdots + u_1^2/\lambda_r.$$

where $u_j = (C\bar{\mathbf{y}})'\mathbf{p}_j\mathbf{p}_j'(C\bar{\mathbf{y}})$. Truncating this sum at $j = trC$ so that

$$X_c^2 = u_1^2/\lambda_1 + \cdots + u_{trC}^2/\lambda_{trC}$$

yields a test statistic that is now Chi-squared with trC degrees of freedom (assuming the approximation (5) is appropriate). In the additive models context the Wald statistic $(C_j\bar{\mathbf{e}}_j)'(C_j\hat{V}_jC_j')^{-1}C_j\bar{\mathbf{e}}_j$ for the j^{th} covariate can be truncated in a similar way.

Computing the truncated test statistic is not onerous since the generalized inverse $(C\widehat{VC}')^-$ is required anyway. Truncation can also be justified from a computational perspective. Since C is not of full rank, numerical instabilities that inflate test statistics can arise without truncation when some λ_i are zero but are estimated by near-zero values. This happens in a region of the orthogonal decomposition where the residual has near-zero variance and there is little information. Such a problem can be seen in the right panel of Figure 3. Some of the last few eigenvalues are large enough that they were included by standard software in the calculation of the generalized inverse. They are small enough that their inversion causes the test statistic to explode. Note the scatterplot points of the right panel in Figure 3 have coordinates $(l, \sum_{j=1}^{l} u_j^2/\lambda_j)$.

An important diagnostic

Note truncation has the potential to result in a conservative test statistic. The ratio

$$\sum_{j=1}^{trC} \lambda_j \Big/ \sum_{j=1}^{r} \lambda_i$$

gives the proportion of variance retained by the truncated test statistic X_c^2. Small values of this ratio would indicate that X_c^2 is conservative. For the test in Figure 3 the ratio is .99 and X_c^2 is just significant at the 5 % level (p=0.043) indicating some evidence against the null hypothesis. This diagnostic is reported for all tests in §6.1.

4.3 Variance estimation

Use of X_c^2 requires computation of $C\widehat{VC}'$, the estimate of the covariance matrix CVC'. The obvious estimate, $C\widehat{VC}' = C\widehat{V}C'$, can be justified by Binder's (1983) estimating equations approach as follows: Use of the penalized least squares criterion (4) yields $\tilde{\mathbf{g}}$ as the solution of the population estimating equation

$$U(\mathbf{g}) = -\mathcal{I}'\mathcal{I}\bar{\mathbf{Y}} + \mathcal{I}'\mathcal{I}\mathbf{g} + (S^- - \mathcal{I}'\mathcal{I})\mathbf{g} = 0.$$

The sample version of this estimating equation

$$\hat{U}(\mathbf{g}) = -\mathtt{I}'\mathtt{W}\mathtt{I}\bar{\mathbf{y}} + \mathtt{I}'\mathtt{W}\mathtt{I}\mathbf{g} + (S^- - \mathtt{I}'\mathtt{W}\mathtt{I})\mathbf{g} = 0$$

has solution $\hat{\mathbf{g}}$ which optimizes the criterion (1). It immediately follows from the Taylor series approximations in Binder (1983) that $\hat{\mathbf{g}}$ is consistent for $\tilde{\mathbf{g}}$ and $\mathrm{Var}(\hat{\mathbf{g}})$ is approximately

$$\begin{aligned}
\mathrm{Var}(\hat{\mathbf{g}}) &= \left[\frac{\partial \hat{U}(\mathbf{g})}{\partial g}\right]^{-1} \widehat{\mathrm{Var}}\left[\hat{U}(\mathbf{g})\right] \left[\frac{\partial \hat{U}(\mathbf{g})'}{\partial g}\right]^{-1} \\
&= S^- \mathtt{I}'\mathtt{W}\mathtt{I}\ \widehat{\mathrm{Var}}[\bar{\mathbf{y}}]\ \mathtt{I}'\mathtt{W}\mathtt{I}(S^-)' \\
&= \widehat{\mathtt{S}}\widehat{V}\widehat{\mathtt{S}}'
\end{aligned}$$

where $\widehat{V}$ is an estimate of the variance-covariance matrix of $\bar{\mathbf{y}}$. In the above $C = \widehat{\mathtt{S}}$, but the extension to $C = \widehat{\mathtt{S}}_1 - \widehat{\mathtt{S}}_0$ is immediate. In the additive models context the variance estimate $C_j\widehat{V_j}C_j' = C_j\widehat{V}_jC_j'$ can be motivated in a similar way. However note that testing is somewhat more complicated in the additive case, since multiple covariance matrices for different binned residuals must be calculated, as opposed to the single covariance matrix of binned means in the scatterplot smoothing case.

The use of the variance estimate $C\widehat{V}C'$ requires computation of $\widehat{V}$ and the smoother matrices $\widehat{\mathtt{S}}_0$, $\widehat{\mathtt{S}}_1$. Standard software such as *SUDAAN* (Shah et. al. 1996) can be used to calculate $\widehat{V}$, however, smoothing routines in statistical software compute smooths efficiently without the explicit use of the full smoother matrix. Hence the smoother matrices are not routinely available and they must be computed by the user (see, for example, Green and Silverman, 1994, p. 12–13). An alternative is to resample $\bar{\mathbf{y}}_j^\star,\ j = 1, \ldots, B$ from $N(\bar{\mathbf{y}}, \widehat{V})$, use smoothing routines to compute replicates $\mathbf{e}_j^\star = C\bar{\mathbf{y}}_j^\star$ and estimate CVC' with the bootstrap covariance matrix $\widehat{C\widehat{V}C'}^\star = \frac{1}{B}\sum_j \{\mathbf{e}_j^\star - \bar{\mathbf{e}}^\star\}\{\mathbf{e}_j^\star - \bar{\mathbf{e}}^\star\}'$ where $\bar{\mathbf{e}}^\star = \frac{1}{B}\sum_j \mathbf{e}_j^\star$. Keeping the smoothing matrix C fixed throughout means

$$\widehat{C\widehat{V}C'}^\star = C\left[\frac{1}{B}\sum_j \{\bar{\mathbf{y}}_j^\star - \bar{\bar{\mathbf{y}}}^\star\}\{\bar{\mathbf{y}}_j^\star - \bar{\bar{\mathbf{y}}}^\star\}'\right]C' = C\widehat{V}^\star C'$$

and the full smoother matrix, C, enters into the calculation implicitly at the expense of replacing $\widehat{V}$ with the bootstrap analog $\widehat{V}^\star$. This bootstrap technique was not investigated in the simulation study of §5 however, results for the OHS scatterplot smoothing example are give in §6.

4.4 Incomplete covariance information

The availability of $\widehat{V}$, or rather, full information about the covariance structure of $\bar{\mathbf{y}}$, is fundamental to the use of X_c^2. If only variance estimates (i.e. the diagonal elements of $\widehat{V}$) are available, as may be the case if data are extracted from a publication, then diagnostics (§5) may support approximating $\widehat{V}$ by a diagonal matrix in the computation of X_c^2, yielding a modified test statistic

$$X_d^2 = (C\bar{\mathbf{y}})'(C\widehat{diag(V)C'})^-C\bar{\mathbf{y}} \dot{\sim} \chi_r^2$$

Simulations in §5.1 indicate that when off-diagonal elements of $\widehat{V}$ are small, X^2_d can actually outperform X^2_C in some cases.

A common alternative to X^2_d in sampling is to construct a conservative test based on matching first moments. Consider the test statistic $Y^2 = \bar{\mathbf{y}}'C'(CC')^{-}C\bar{\mathbf{y}}/\tilde{v}$, where $\tilde{v} = tr(\widehat{V})/tr(C) = \sum_{i=1}^{k} v_{ii}/tr(C)$ with v_{ii} being the i^{th} diagonal element of $\widehat{V}$. Note $\sum_{i=1}^{k} v_{ii} = \sum_{i=1}^{k} \delta_i$ where the δ's are the eigenvalues of $\widehat{V}$. From Mathai and Provost (1992, ch. 4) we have

$$Y^2 \;\sim\; \sum_{i=1}^{r} \lambda_i Z_i^2$$

where $Z_i \;\sim\; IID\; N(0,1)$ and $\lambda_1, \ldots, \lambda_r$ are the non-zero eigenvalues of $(CC')^{-}CVC'$. On applying the Theorem 2 of Scott and Styan (1985) we have $\lambda_i \leq \delta_i,\ \forall i$ so that

$$E[Y^2] = \sum_{i=1}^{r} \lambda_i E[Z_i^2]/\tilde{v} = tr(C)\sum_{i=1}^{r} \lambda_i / \sum_{i=1}^{k} \delta_i \leq tr(C)\sum_{i=1}^{r} \delta_i / \sum_{i=1}^{k} \delta_i \leq tr(C).$$

As a result the approximation $Y^2 \sim \chi^2_{tr(C)}$ yields a conservative test of the hypothesis. Note that truncation is not required here to ensure the conservative nature of the test statistic and that, in practice, Y^2 will be evaluated by replacing $tr(V)$ with $tr(\widehat{V})$.

4.5 No covariance information: cross-validation

In the absence of any information concerning the covariance of $\bar{\mathbf{y}}$, hypothesis tests cannot be used to determine an appropriate degree of smoothing. An alternative is to select a degree of smoothing to minimize expected squared residual error. Cross-validation estimates this error by removing observations from the sample one at a time, constructing a model with the remaining observations. In the context of superpopulation model 1, cross-validation for the finite population, $CV(\alpha)$, can be motivated by a model in the same way as the penalized least squares criterion (4). Here,

$$CV(\alpha) = (\mathbf{Y} - \mathcal{I}\tilde{\mathbf{g}}^{(-)})'(\mathbf{Y} - \mathcal{I}\tilde{\mathbf{g}}^{(-)}) \tag{9}$$

where $\tilde{\mathbf{g}}^{(-)}$ are the "leave-one-out" fitted values with smoothing parameter α, that is, $\tilde{\mathbf{g}}^{(-)}$ has components $\tilde{g}_i^{(-)}$ being the fitted value for the i^{th} individual in the population in the absence of the i^{th} response. The value of α that minimizes $CV(\alpha)$ is denoted $\tilde{\alpha}$.

Since the finite population is unobserved, a sample estimate of $CV(\alpha)$ that incorporates weights

$$cv(\alpha) = (\mathbf{y} - \mathtt{I}\hat{\mathbf{g}}^{(-)})'\mathtt{W}(\mathbf{y} - \mathtt{I}\hat{\mathbf{g}}^{(-)}) \tag{10}$$

can be motivated in a manner similar to (1). The value of α that minimizes $cv(\alpha)$ is denoted $\hat{\alpha}$ and comparisons of $(\tilde{\alpha}, \hat{\alpha})$ are discussed in §5.2 where simulations assess the effectiveness of cross-validation in estimating α.

5 Simulation studies

In this section, we report two simulation studies. The first concerns the accuracy of distributional assumptions required for the Wald test, and the second examines the ability of cross

validation to select an appropriate smoothing parameter. Scatterplot smoothing rather than additive modeling is considered in all simulations, but the results are suggestive of results for additive models.

5.1 Test statistic results

The Wald test statistics of §4 are based upon a chi-square approximation, which in turn arises from a central limit theorem applied to binned data. We investigate the accuracy of these approximations in a number of scenarios. Agreement of actual and theoretical distributions is assessed through the use of quantile-quantile (QQ) plots.

Recall that even in the case of conventional data the reference distributions used for these test statistics are not exact (Hastie and Tibshirani 1990, Cleveland and Devlin 1988). This is exemplified by the ambiguity with respect to degrees of freedom for a smooth. For a smoothing matrix S separate analogies with the linear regression context give any of $tr(SS') \leq tr(S) \leq tr(2S - SS')$ as the appropriate degrees of freedom. Thresholds based on each of these are given in all QQ plots to remind the reader that whatever inaccuracies are evident are not solely due to issues relating to the survey sampling context. These thresholds are the critical values for tests with a 5% level of significance. Hence the purpose of this simulation study is not to provide accurate estimates of true coverage probabilities, but to reassure the reader that the χ^2 approximations are appropriate.

Central to all scenarios considered is the need for accurate and stable variance estimates. Even in the simplest cases use of an unstable variance estimate can severely inflate a test statistic. In practice a simple QQ plot of the standardized residuals, $(\bar{\mathbf{y}} - \hat{\mathbf{g}})\widehat{V}^{-1/2}$, serves as a useful diagnostic for the adequacy of central limit theorems and variance estimates. Similarly, a QQ plot of studentized residuals, $(\bar{\mathbf{y}} - \hat{\mathbf{g}})diag(\widehat{V})^{-1/2}$, reveals whether X_c^2 can be replaced with a diagonal version X_d^2, which serves to stabilize variance estimates in cases of near-zero covariances.

The superpopulation model and sampling design

For each scenario we require that the assumption $\tilde{g}(x) \to g(x)$ of superpopulation model 2 holds for all finite populations. This is assured by generating finite populations from a model of the form

$$Y(x) = g(x) + \epsilon$$

where the interplay of the sampling design and the finite population structure determine the form of the error term. The assumptions of superpopulation model 2 assert this design should not affect central limit theorems or χ^2 approximations and this is indeed reflected in the simulations conducted.

For reasons stated in §5.2, $g(x)$ is set to be a linear function for all scenarios considered. The scenarios themselves are determined by differences in the error term; differences that, in the end, have no effect on the χ^2 approximations. The error follows the random intercept model appropriate to two stage sampling (see Rao et al. 1993 and Wu et al. 1988)

$$\epsilon_{jl} = z_{jl} + u_l$$

scenario	N	L	n	l	f	g	σ	τ	Test statistic
1	200000	1	20000	1	ϕ	-	4	0	X_c^2
2	200000	1	20000	1	t_3	-	4	0	X_c^2
3	200000	1	20000	1	χ_1^2	-	4	0	X_c^2
4	2000	1000	200	1000	ϕ	ϕ	7	0	X_c^2
5	2000	1000	200	1000	ϕ	ϕ	7	0	X_d^2
6	2000	1000	200	1000	ϕ	ϕ	7	$\sqrt{.9\sigma^2}$	X_c^2

Table 1: Parameter values for the various scenarios where ϕ denotes the standard normal density.

where

$$z_{jl} \sim \frac{1}{\sigma} f\left(\frac{\cdot}{\sigma}\right), \; u_l \sim \frac{1}{\tau} g\left(\frac{\cdot}{\tau}\right)$$

and f, g have mean 0. In addition $j = 1, \ldots, N_l, l = 1, \ldots, L$ where L is the number of primary sampling units in the population and $N_l = N$ is the number of secondary sampling units within each primary. Samples were generated by first sampling l primary units and then sampling n secondary units within each primary. Simple random sampling was used at both stages of the design. The within-primary covariance is controlled by the random intercept u_l, with $\tau = 0$ yielding independent observations. The single stage case, or simple random sampling, follows by setting $L = 1$.

Table 1 describes the scenarios considered in terms of f, g, and so on, where the parameter σ was chosen so the signal to noise ratio resembles Figure 1. The choices of N, L, n, l ensured stable variance estimates which is crucial to the behaviour of the test statistics. In the two stage scenarios (4–6), the parameter τ was tuned so the resulting variance estimate $\widehat{V}$ was either almost diagonal ($\tau = 0$, yielding off-diagonal elements of 2-3 orders of magnitude smaller than diagonal) or had off-diagonal components that are of the same order of magnitude as diagonal. In a fifth scenario the goal was to assess the effect of ignoring covariance information on the performance of the test statistic. Samples were generated as in scenario 4 but the diagonalized statistic X_d^2 was used instead of X_c^2. Finally, for each scenario 100 values of the test statistic were simulated and the covariate used was age with 47 distinct values ranging from 18 to 64 as in the Ontario health survey. The distribution of age in the finite population was determined using national census values.

Results

In each scenario, test statistics for all hypotheses in Table 2 were calculated. A representative selection of these are reported using QQ plots to assess the accuracy of the test statistic distributions. All test statistics should follow the null distribution, because the true signal has fewer degrees of freedom (2 df) than any null hypothesis. Figure 4 summarizes the behaviour of the usual F-tests, which ignore the need for variance stabilization, for scenario 1, the simplest scenario, and scenario 6. What is evident from this plot is that ignoring the variance, $\widehat{V}$, inflates the test statistic even in the simplest case and this worsens in scenario 6 where the covariance structure is more complex. The larger-than-expected number of

simulated test statistics above the threshold in scenario 6 indicates that ignoring covariance information has a serious negative effect on conventional F statistics.

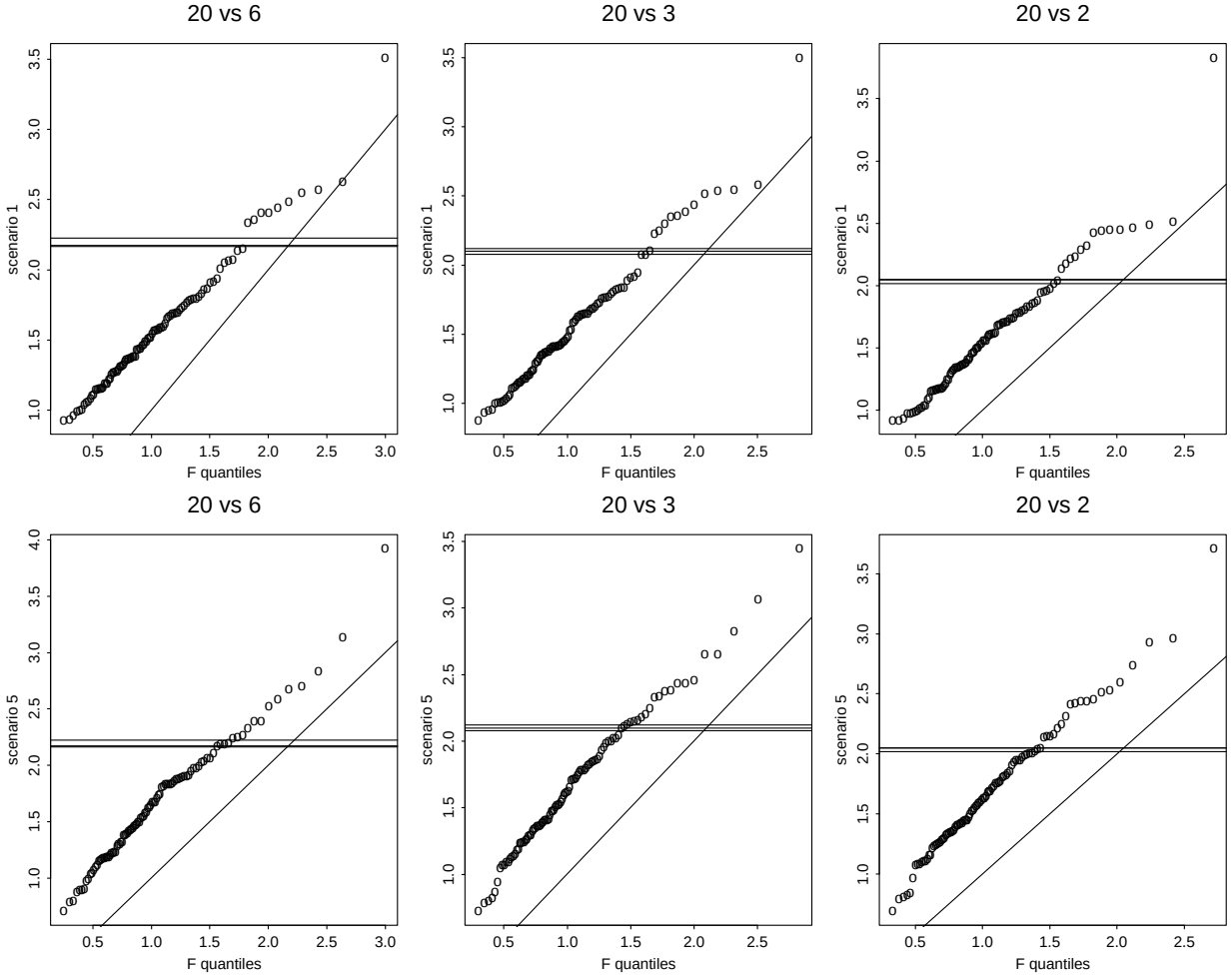

Figure 4: QQ plots for the usual F test under scenario 1 and 5. Note the horizontal thresholds, given by various analogies with linear regression, all tend to be similar

The chi-square QQ plots in Figure 5 are given for tests of 20 vs. 12, 12 vs. 9, 9 vs. 6 and 6 vs. 3 degrees of freedom, which fall on the diagonal of Table 2. In scenarios 1 and 5 the test statistics behave very well, and corresponding plots are consequently omitted. Generally speaking central limit theorems are robust to error distributions (rows 1,2) and the Wald test statistics perform well in each scenario provided variance estimation is adequate. The $k \times k$ matrix V has $k(k-1)/2$ distinct parameters and it is crucial to estimate these accurately. Instability in the estimate $\widehat{V}$ can inflate test statistics which severely affects coverage. Typically the diagonal components of $\widehat{V}$ are more stable than those off the diagonal and if covariances are small they can be ignored. This was assessed in two-stage sampling by scenarios 4 and 5, which generate zero ($\tau = 0$) within-primary correlations for the finite populations, but use different tests statistics (X_c^2 and X_d^2). The Wald tests are comparable whether covariance information is used or not, and only results for scenario 4 (X_c^2) are given, because the results for scenario 5 (X_d^2) are even better. The fourth row gives results

for scenario 6 where the full covariance matrix is used - use of X_d^2 for this scenario results in disastrous coverage and these plots are omitted.

The first two plots in the fifth row show that, in some cases, the diagonalized X_d^2 is preferable to the full covariance statistic X_c^2! The truncation of eigenvalues in the computation of the generalized inverse $(C\widehat{V}C')^-$ interacts with the effect of variance estimation on the behaviour of the Wald test statistics. In effect the fewer eigenvalues that are set to zero the greater the impact $\widehat{V}$ has on $(C\widehat{V}C')^-$ and subsequently, the test statistic. The smallest number of zero eigenvalues occur in the goodness of fit tests where instability in variance estimates can inflate test statistics dramatically. In the fifth row of the plot we give 3 QQ plots for test statistics based on 47 and 12 df. The first involves scenario 4, that is standardizing by the full covariance matrix $\widehat{V}$, and the resulting test statistic is significantly inflated. The second plot shows that using the diagonalized X_d^2 serves to offset this effectively. The normal QQ plot given for a single simulation at the end of the row, illustrates the extent to which the standardized residuals $(\bar{\mathbf{y}} - \hat{\mathbf{g}})diag(\widehat{V})^{-1/2}$ are normal and such a replacement is valid. Finally, the third QQ plot depicts the performance of the conservative test statistic of §4.4 and the improvement is surprisingly dramatic. In general, the conservative test statistic performs well but only as a goodness of fit test when covariances are near zero. Otherwise, it tends to under cover or has disastrous behaviour when covariances can not be set to zero. Although the coverage is about right, the statistic is likely to be very conservative in situations where the null is false.

Lessons learned

These simulations indicate the critical role of variance estimation in obtaining good tests. If there is no covariance between binned data, then ignoring the off-diagonal elements of $\widehat{V}$ improves performance. If on the other hand, there is covariance, then ignoring it can hurt performance severely. Conservative test statistics should be used as goodness of fit tests only, and only when the off-diagonal elements of $\widehat{V}$ can be set to zero, otherwise it use should be avoided.

In practice, we suggest examination of normal QQ plots of studentized, and standardized, residuals and based on these a decision made about whether to include covariance information in the test - that is whether or not the full covariance matrix $\widehat{V}$, or its diagonal, should be used.

5.2 Cross-validation results

Here we only assess the use of cross-validation for the sample as a surrogate for the unobserved finite population. Its use for the finite population is assumed appropriate on the basis that $CV(\alpha)$ is motivated by superpopulation model 1, a model for which cross-validation is traditionally used. Moment calculations similar to those of Binder (1983) show that the sample cross-validation sum of squares, $cv(\alpha)$, is asymptotically unbiased (in the sense of Särndal et al. 1992) for $CV(\alpha)$, that is,

$$E[cv(\alpha)] = CV(\alpha) + O(n^{-1})$$

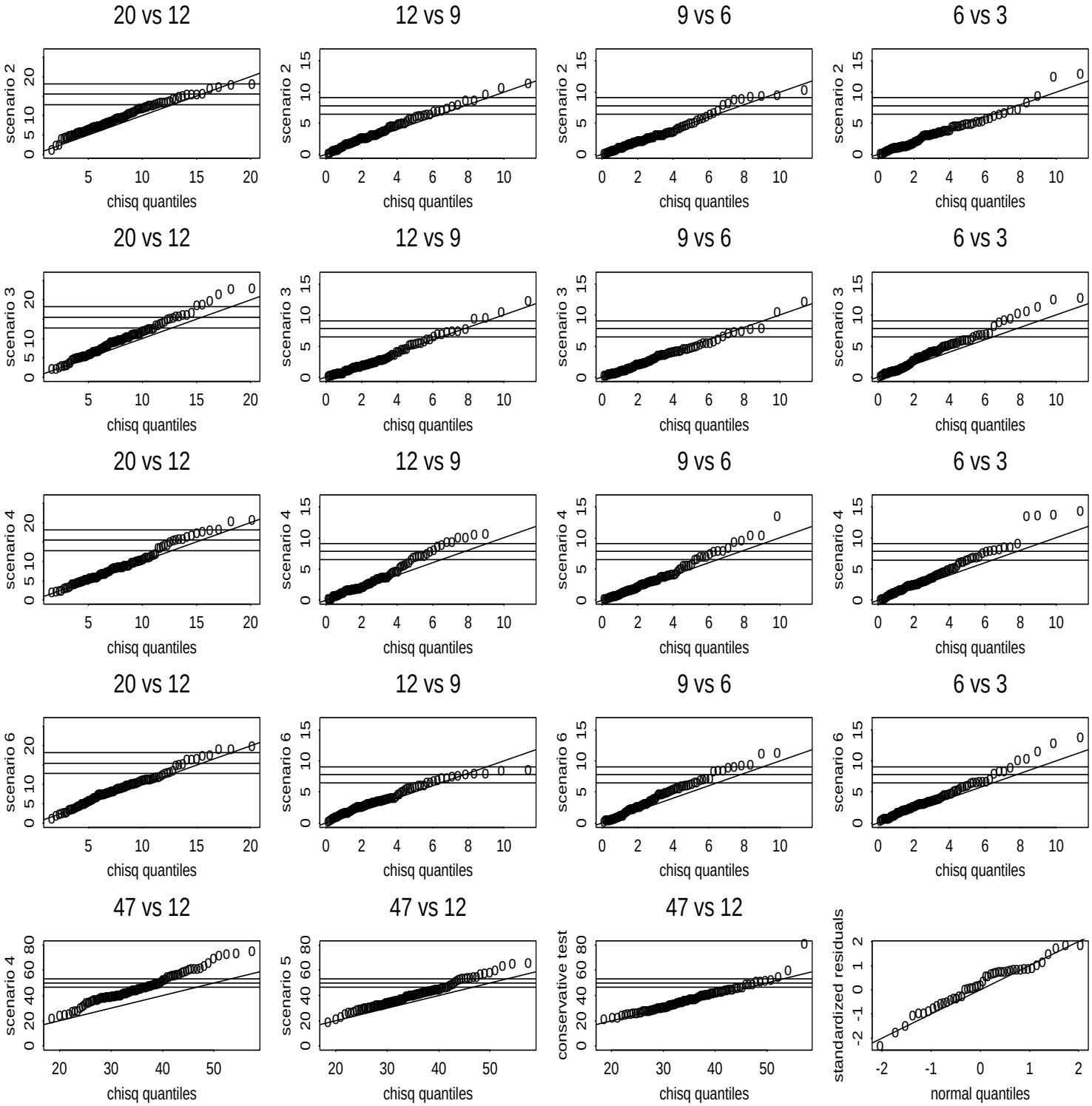

Figure 5: Chi-square QQ plots for scenarios 2, 3, 5 and 6 all reflect acceptable coverage with the exception of the goodness of fit test using full covariance information. QQ plots for scenarios 1 and 4 reveal a similar pattern but were omitted. QQ plots for the conservative test revealed it can behave well as a goodness of fit test in some circumstances but otherwise it should be avoided.

where the $O(n^{-1})$ term contributes to bias in the sample. This bias is evident in the degrees of freedom selected when the criteria were used in a simulation study. Here 1000 pairs of $(\tilde{\alpha}, \hat{\alpha})$ were generated using a superpopulation model and sampling design described in §5.1.

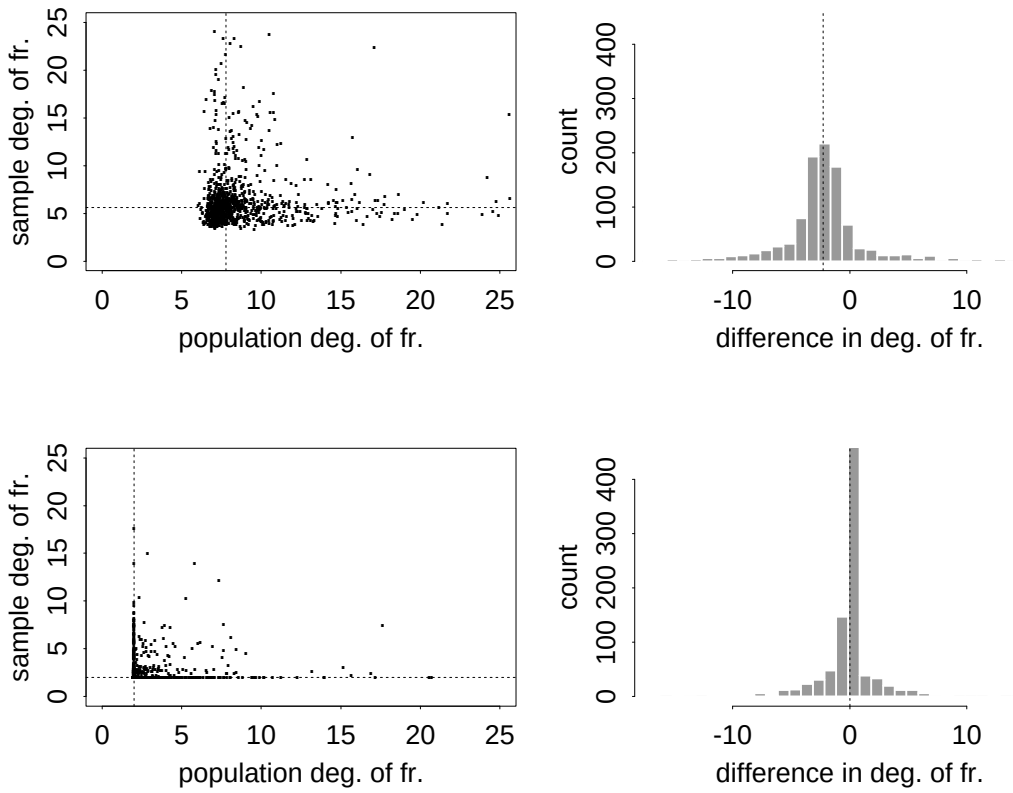

Figure 6: Distribution of smoothing parameters $\hat{\alpha}$ and $\tilde{\alpha}$ estimated from the finite population and survey sample under two simulation scenarios. In the first (top row), the true signal has 6df, while a linear (2df) signal is used in the second (bottom row). We report the effective number of degrees of freedom, rather than $\hat{\alpha}$. The bottom and top 2.5% of differences have been trimmed out of the histograms to improve presentation.

Figure 6 displays the results for two different scenarios, one where the true underlying signal was chosen to be the curve fitted with 6 degrees of freedom in Figure 1 and another where it was simply linear. A scatterplot of $\tilde{\alpha}$ versus $\hat{\alpha}$ and a histogram of the difference $\hat{\alpha} - \tilde{\alpha}$ are given for each scenario. Median values are indicated by dotted lines. The plots for the 6df case indicate that $\hat{\alpha}$ has median df slightly below 6, while in the finite population the median df is above 6. This can be thought of as a downward bias in the sample estimate. This bias results from fewer observations in the sample than the population (in this case 90% smaller), which decreases the signal to noise ratio, leading to oversmoothing. The occasional extreme value for degrees of freedom is troubling. These extrema occur both in the finite population and the sample, suggesting that this is a property of the cv criterion, rather than the sample survey design.

In the second scenario, where the signal is linear, the bias vanishes as most values of $\tilde{\alpha}$ and $\hat{\alpha}$ are extremely close to 2. The stability of both criteria in the second scenario implies a linear signal is well suited for assessing null behaviour of the Wald test statistics. Here "stability" is meant as choosing the null case when it is true. Finally, although biased, the sample cross-validation sum of squares seems to work quite well in determining when a signal is linear or not.

6 Examples

We consider three different models for BMI based on the OHS data: smoothing against age (continued from earlier), separate smooths of BMI against age for men and women, and an additive model using gender, age and desired BMI. The first example illustrates hypothesis tests and cross-validation methods developed in §4 for scatterplot smoothing, and also considers diagnostics. The other two examples illustrate more complicated models. In all examples we have access to the entire data set. Thus full covariance information is available and the Wald statistic may be used. Smoothing splines are used in all examples.

6.1 Predicting BMI using age

To illustrate cross validation, we consider smoothing BMI on age, as depicted in Figures 1 and 7. Four different fitted lines are given in the right panel of Figure 7: 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 linear or 3 df fits may be insufficient, while 12 df is perhaps overfitting the data. Cross-validation can be performed in S using `smooth.spline` applied to the original data. The cross-validated average sum of squares is given in Figure 7 (left plot) against the degrees for freedom from 2 (linear) to 12. Six degrees of freedom seems optimal, while a linear model (2 df) is clearly inappropriate.

Tests for the goodness of fit, linearity and significance of the variable may also performed, using the Chi-square approximations of §4.2-4.4. The Wald statistics utilizing full covariance information are appropriate, but for comparison we also calculate test statistics with incomplete covariance information, and complete covariance statistics using the bootstrap. 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.

To perform the test with full covariance information, we need the smoother matrices corresponding to the null and alternative hypotheses. These can be calculated quite easily; for example with cubic smoothing splines, $S = (\mathtt{I'WI} + \alpha K)^{-1}$. Details and an algorithm for construction of K for a given α or degrees of freedom are given in Sections 2.1-2.3 of Green and Silverman (1994). The test statistic X_c^2 in §4.2 requires C, which in the context of hypothesis tests will be the difference between smoother matrices S_1 and S_0, corresponding to the alternative and null hypotheses.

The results of the tests with full covariance information are given in Table 2. The row labels give the degrees of freedom for the alternative hypothesis, and all nulls of smaller size (given by column labels) are tested. In the first row, goodness of fit tests indicate that even models with 1-6 degrees of freedom may fail (at a 5% level) to capture all patterns

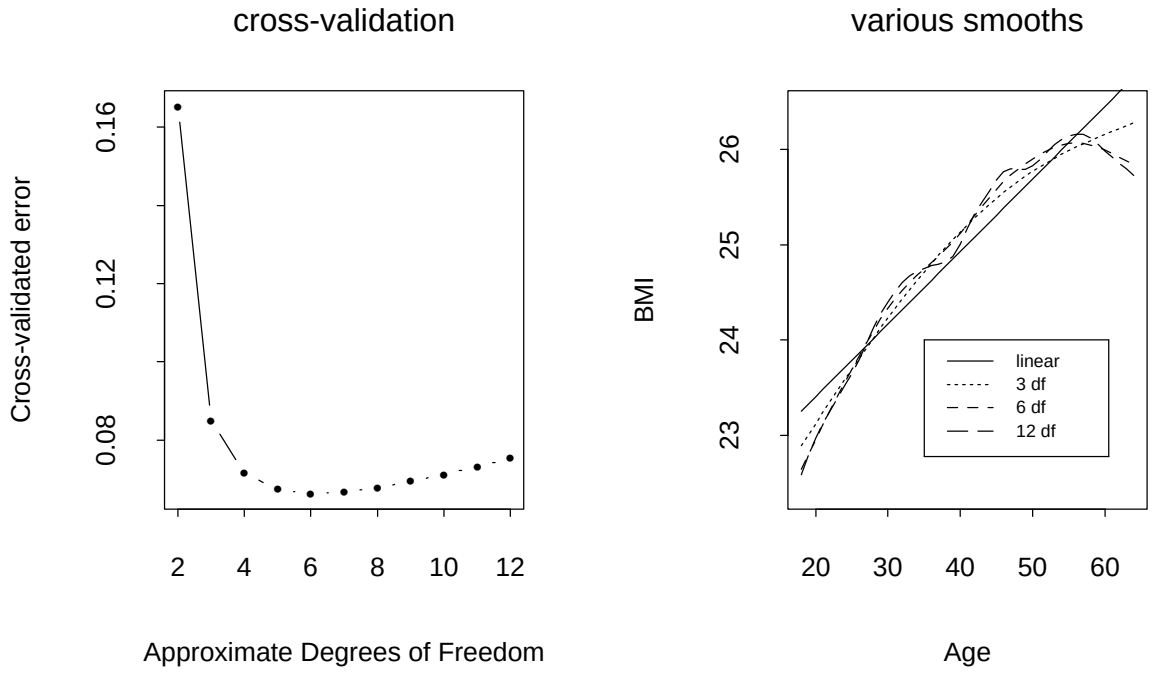

Figure 7: Plots for selection of smoothing parameter in a smooth of BMI on age. The cross-validated error is given on the left, and smooths with a variety of degrees of freedom are given on the right.

	20	12	9	6	3	2	1
47	33.8	47.0	49.0	57.7*	79.6**	172.7**	1192.3**
20	–	3.8	7.1	8.5	38.3**	107.8**	1000.1**
12	–	–	4.8	6.4	25.1**	98.9**	988.4**
9	–	–	–	3.9	23.0**	95.5**	981.7**
6	–	–	–	–	16.1**	87.8**	967.8**
3	–	–	–	–	–	65.2**	964.8**
2	–	–	–	–	–	–	851.2**

Table 2: Chi-square statistics for tests involving a variety of smooth models of BMI as a function of age. Full covariance information is used in calculating the test statistic. Values significant at 0.05 and 0.01 levels are marked by * and ** respectively.

	20	12	9	6	3	2	1
47	0.96	0.98	0.99	0.99	1.00	1.00	1.00
20	–	0.69	0.83	0.90	0.94	0.96	0.96
12	–	–	0.48	0.79	0.92	0.95	0.96
9	–	–	–	0.61	0.90	0.95	0.96
6	–	–	–	–	0.83	0.94	0.96
3	–	–	–	–	–	0.88	0.96
2	–	–	–	–	–	–	1.00

Table 3: Proportion of singular values captured in each test statistic given in Table 2.

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 3df or less (at 1% level). This roughly agrees with the results of the cross-validation. Table 3 gives the proportion of variance retained by the truncated test statistic X_c^2. In most cases this proportion is close to 1. For tests involving small changes in degrees of freedom (on the diagonal of the table), this proportion can be small. In these tests the statistic is not very close to the level of significance, so the truncation is likely to have minimal effect.

The bootstrap method of §4.3 was also used to calculate X_c^2. This makes use of $\widehat{V}$ but does not require the smoother matrices. The number of bootstrap draws (B) must be large enough to produce a reasonable estimate of the covariance matrix; we chose $B = 2000$. $m = 50$ bootstrap simulations were performed, and the test statistics for the 47 df alternative and various nulls are given in the first row of Table 4. The test statistics are given in the form of an average value of X_c^2 and a standard error (second row). The values are close to those in Table 2, although the mean values are not always within two standard errors of the full covariance test statistics. This indicates the possible presence of a mild bias. This is to be expected because of the large number of parameters to be estimated in the covariance matrix of the bootstrapped smooths.

For comparison, test statistics using the incomplete covariance information are given in the third row of Table 4. These use the variances of data within each age category, but not

Test	20	12	9	6	3	2	1
Bootstrap mean	34.17	47.55	49.77	58.90	81.05	176.63	1218.09
Bootstrap std. error	(.35)	(.23)	(.22)	(.28)	(.39)	(.91)	(5.93)
Incomplete covariance (Y^2)	15.83	24.45	28.37	34.37	50.93	124.3**	887.8**

Table 4: Chi-square statistics for a variety of goodness-of fit tests involving smooth models of BMI as a function of age. The first row gives average Bootstrap statistics calculated with full covariance information, while the second gives corresponding standard errors. The third row, the incomplete covariance information test statistic Y^2 is calculated.

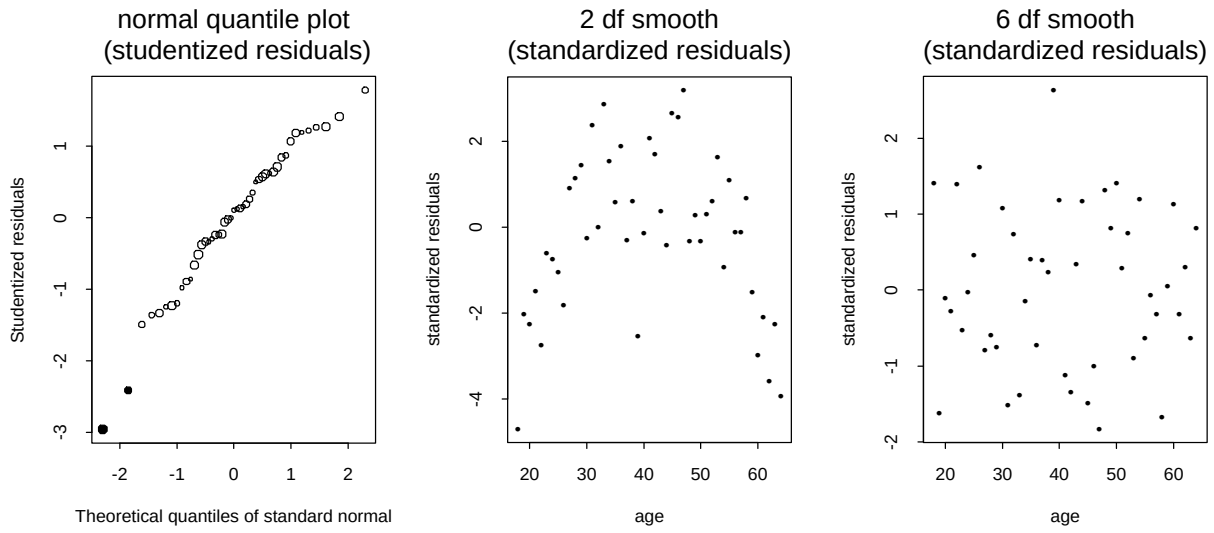

Figure 8: Residual plots for smoothing BMI on age: Normal probability plot for studentized residuals, with two large residuals noted (left), and standardized residuals plotted against age, for the linear (2df) and 6df models (center and right respectively). Plotting characters of various sizes indicate weights for the associated points.

the covariances between ages. There is strong evidence that at least a 3df smooth is required. Other tests using this statistic (not shown here) indicate that intermediate models such as 12 or 9 df may be simplified to 3 but not 2df. As one might expect with more conservative tests, the results are less significant.

We also consider residual diagnostics. The studentized residual, $e_i = (y_i - \hat{y}_i)/diag(\widehat{V})^{-1/2}$ can be calculated, and is useful to detect outliers or other anomalous patterns involving specific points. For example, the left panel of Figure 8 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.

Using (6) we calculate standardized residuals $\bar{y}C(C\widehat{V}C')^{-1/2} = (\bar{Y} - \hat{g})C^{-}\widehat{V}^{-1/2}$ where $C = I - S$. Truncation methods similar to those used in calculation of X_c^2 are used. These standardized residuals may then be used as in a usual regression analysis. For example, in

Figure 8, 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 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 8).

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,

$$R^2 = 1 - \left(\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 w_i \Big/ \sum_{i=1}^{n} (y_i - \bar{y})^2 w_i \right) \tag{11}$$

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

6.2 Separate age effects on body mass index for each gender

Although the smoothing of BMI on age provides a useful illustration, there are other important factors relating to BMI. Figure 9 gives separate smooths of BMI on age for men and women, suggesting gender differences in the shape of the relationship and mean level. The R^2 value (11) value for the separate smooths is 8.5%, an appreciable increase over the single smooth.

The testing methodology used for the age effect on BMI can be applied within genders in a manner that allows a single overall test. The modifications are as follows: Now $\bar{\mathbf{y}} = [\bar{\mathbf{y}}_\mathbf{m}, \bar{\mathbf{y}}_\mathbf{f}]$ corresponds to 94 BMI means, 47 for each gender. Two separate smooths ($\bar{\mathbf{y}}_\mathbf{m}$ on $\mathbf{x}$, and $\bar{\mathbf{y}}_\mathbf{f}$ on $\mathbf{x}$) would have 47×47 smoother matrices $\widehat{\mathbb{S}}_m$ and $\widehat{\mathbb{S}}_f$. This can be combined into a single smooth of $\bar{\mathbf{y}} = [\bar{\mathbf{y}}_\mathbf{m}, \bar{\mathbf{y}}_\mathbf{f}]$ using a block diagonal matrix $\widehat{\mathbb{S}}$:

$$\hat{g} = \widehat{\mathbb{S}}\bar{\mathbf{y}} = \begin{bmatrix} \widehat{\mathbb{S}}_m & 0 \\ 0 & \widehat{\mathbb{S}}_f \end{bmatrix} \begin{bmatrix} \bar{\mathbf{y}}_m \\ \bar{\mathbf{y}}_f \end{bmatrix}.$$

Other than requiring two separate smoothing parameters for the two models, modeling and testing proceeds as before. Cross-validation of separate models for men and women suggests that 6 degrees of freedom for men and 4 for women may suffice. We also consider a larger model with 12 df for each of the two smooths. Table 5 gives tests involving these models and various null and alternatives.

The goodness of fit test indicates a lack of fit. The same test was less significant in §6.1 when gender was ignored. A possible explanation is that once variation due to gender is modeled, the noise level in the data has been reduced, and finer patterns in the relation between age and BMI can be identified. Models with (separate) linear or constant relationships

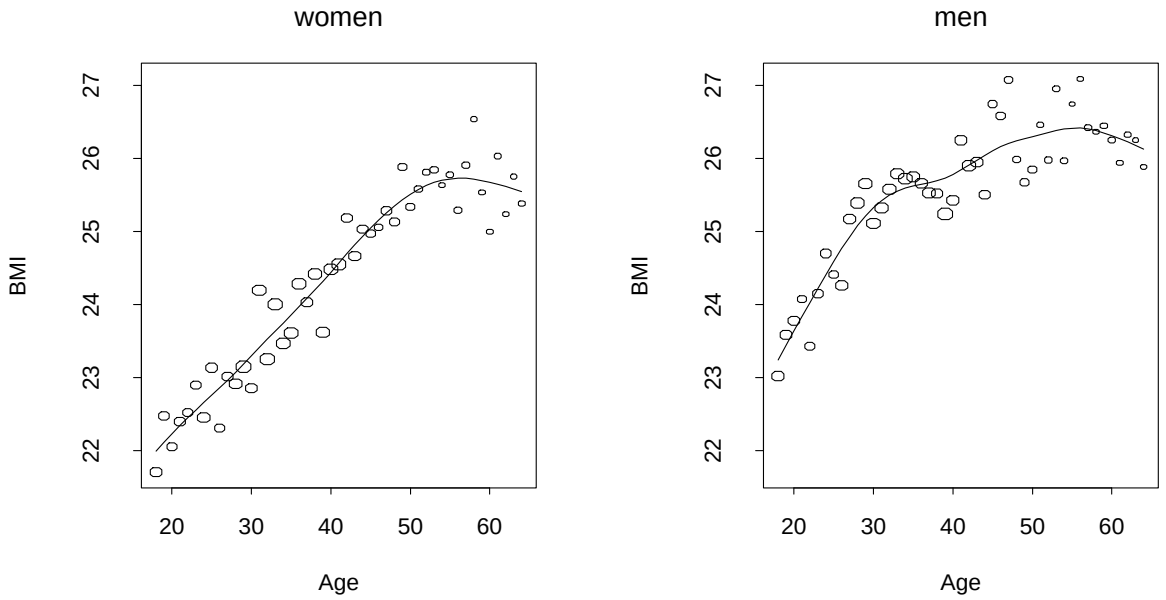

Figure 9: Separate smooths of BMI on age for women (4 df) and men (6 df)

Test	H_0 df (m,f)	H_a df (m,f)	X_c^2	Δdf	p-value
goodness of fit	(12,12)	(47,47)	121.9	70	.00012
linearity	(2,2)	(12,12)	173.6	20	0
constant relationship	(1,1)	(12,12)	1434.5	22	0
cv choice	(6,4)	(12,12)	23.7	14	.050

Table 5: Various tests for separate smooths of BMI on age for men and women.

between BMI and age for each gender are strongly rejected. Simplification from the (12,12) degrees of freedom model to the (6,4) model suggested by cross-validation is just possible at the 5 percent level. The cross-validated choice of model has similar lack of fit as the (12,12) model.

In this particular problem, the test statistic X_c^2 is quite close to the sum of two test statistics, calculated separately for men and women. In separate calculations, correlations between $\bar{\mathbf{y}}_m$ and $\bar{\mathbf{y}}_f$ would be ignored, although correlations within either vector would be used. These correlations are mostly small, making the test statistic similar to the sum of the two separate test statistics. In situations with larger correlations, results could vary, and the combined test would be preferred.

6.3 An additive model for BMI

There appears to be strong evidence for gender differences in the age effect on BMI. In this section we extend this idea, adding a smooth effect for desired body mass index (DBMI):

$$\hat{\mathbf{y}} = \hat{g}_1(\mathtt{age}, \mathtt{sex}) + \hat{g}_2(\mathtt{DBMI}) \tag{12}$$

Before discussing the model and subsequent analysis, we comment briefly on the issue of missing values. Missing values of DBMI for 1817 individuals in the survey were identified, and those cases removed. Some exploration of the data indicates that missingness may not be entirely random; for example individuals with high BMI are more likely to have a missing DBMI. Since only 5% of individuals have missing values, we treat missingness as random for the illustrative purpose of this example.

In model (12), the age $\times$ gender interaction $\hat{g}_1$ corresponds to separate smooths on age for men and women. Interactions are not usually considered in additive models, however when one of the terms is categorical, they may be accommodated by a modified backfitting algorithm (Hastie and Tibshirani 1990, p. 265-6). The working residuals are smoothed against DBMI as usual, but the two smooth functions of age are estimated using subsets of data corresponding to men and women. The three smooths account for 18 degrees of freedom in total (six each for men and women, and six for DBMI). The modified backfitting algorithm converges within 10 iterations.

The resultant smooths are displayed in Figure 2. Predictions from this combined model are very similar to predictions made from separate additive models using age and DBMI for men and women. The R^2 statistic (11) for this model is 60.6%, mostly due to the large effect of DBMI. The R^2 from two separate additive models for men and women is only slightly higher, 61.7%.

To perform tests we use the results of §4.1 and §4.2. For DBMI and the age $\times$ gender interaction we consider two tests: goodness of fit, and significance of the smooth function estimated in the backfitting algorithm. Both tests rely on the application of the central limit theorem to binned partial residuals (8).

Taking $C_j = I - \hat{\mathtt{S}}_j$, where $\hat{\mathtt{S}}_j$ is the smoother matrix for variable j, we have

$$(I - \hat{\mathtt{S}}_j)\bar{\mathtt{e}}_\mathtt{j} = \bar{\mathtt{e}}_\mathtt{j} - \hat{\mathbf{g}}_\mathtt{j} \Leftrightarrow (\mathbf{Y} - \sum_{l \neq j} \mathtt{I}_l \hat{\mathbf{g}}_\mathtt{l}) - \mathtt{I}_\mathtt{j} \hat{\mathbf{g}}_\mathtt{j} = \mathbf{Y} - \hat{\mathbf{Y}}$$

where $\Leftrightarrow$ indicates that the right hand side would need to be binned according to the levels of $\mathbf{x_j}$ to get equality with the left side. So for any j this choice of C_j corresponds to a goodness-of-fit test, with different binning schemes.

To test whether a smooth term can be dropped from the model, we need only use the binned partial residuals $\bar{\mathbf{e}}_\mathbf{j}$, calculating $X_c^2 = \bar{\mathbf{e}}_\mathbf{j}' \hat{V}_j^{-1} \bar{\mathbf{e}}_\mathbf{j}$. This is in fact a goodness of fit test for a model without $\hat{\mathbf{g}}_\mathbf{j}$, since the partial residuals are the difference between Y and the fitted model with $\hat{\mathbf{g}}_\mathbf{j}$ excluded. A test comparing the model with and without $\hat{\mathbf{g}}_\mathbf{j}$ would also be possible but is not considered here.

In the case of the interaction term, $\hat{\mathbf{S}}_j$ would be formed with a block diagonal of two smoother matrices for men and women, and applied to a stacked vector of average partial residuals for each age group and gender. This approach is similar to the tests involving two separate models for age by gender, in §6.2.

For the DBMI term, the goodness of fit test gives a Chi-square statistic of 717.5 on 34 df, and the test for dropping the term gives 34354.6 on 39 df. Both are significant, indicating lack of fit, and a very significant statistic for dropping the DBMI term. The fact that DBMI cannot be dropped seems consistent with the large increase in R^2 over the models of §6.1 and §6.2.

For the age by sex interaction, we have a goodness of fit statistic of 119.1 on 82 df and a test for dropping the term gives 4548.0 on 94 df. Again, there appears to be moderate lack of fit, and very significant evidence against dropping the smooth term. The comparatively small value relative to the DBMI test is due to the smaller range of the AGE terms in the model.

The nonlinearity of the DBMI term can also be tested. We decompose $g_2(\texttt{DBMI})$ into a linear term, $\beta \times \texttt{DBMI}$, and a smooth term, $g_2(\texttt{DBMI})$. The partial residuals (excluding g_2) are calculated, and the variance-covariance matrix of these residuals generated. Since we are comparing a model with g_1 and the linear DBMI term to a model with g_1, a linear DBMI term and a smooth DBMI term, $C = I - S_2$. This yields a test statistic of $X^2 = 76.01$ on 4 df. Although this is highly significant, it is much smaller than test statistics for removal of either the DBMI or age terms. While there is some evidence of nonlinearity of the DBMI term, the nonlinearity may be of little practical use.

In both cases, the significance of the smooth terms is very high. This continues the pattern observed in §6.2, with the separate smooths for age by gender. As more variation in BMI is explained, we are able to identify finer detail in the relationship between BMI and each predictor, and significance levels of the models increase.

Acknowledgments

We wish to thank Rob Tibshirani, David Binder, an associate editor and two anonymous referees 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. and Stafford, J. E. (1999). Density estimation from complex surveys. *Statistica Sinica.* **9**, 407-424.

Bellhouse, D. R. and Stafford, J. E. (2001). Local polynomial regression in complex surveys. *Survey Methodology*, to appear.

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

Chambers, J. M. and Hastie, T. J. (1992). *Statistical Models in S.* Wadsworth & Brooks/Cole: Pacific Grove, Ca.

Cleveland, W. S. , and Devlin, S. J. (1988), Locally weighted regression: An approach to regression analysis by local fitting, *Journal of the American Statistical Association*, **83**, 596-610 .

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

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

Ihaka, R. and Gentleman, R. (1996) R: A Language for Data Analysis and Graphics, *Journal of Computational and Graphical Statistics*, **5**, 299–314.

Kish, L. and Frankel, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society (B)* 36: 1 - 37.

Lohr, S. L. (1999). *Sampling: design and analysis.* Duxbury Press: Toronto.

Konijn, H.S. (1962). Regression analysis in sample surveys. *Journal of the American Statistical Association.* **57**, 590-606.

Mathai, A. M. and Provost, S. B. (1992). *Quadratic forms in random variables: theory and applications.* Marcel Dekker: New York.

Ontario Ministry of Health (1996). *Ontario Health Survey: User's Guide, Volumes I and II.* Queen's Printer for Ontario.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review.* **61**, 317-337.

Pfeffermann, D. and LaVange, L. (1989). Regression models for stratified multi-stage cluster samples. *Analysis of Complex Surveys*, C.J. Skinner, D. Holt and T.M.F. Smith, eds. New York: Wiley, pp. 237-260.

Rao (1973). *Linear statistical inference and its applications.* 2^{nd} Ed. Wiley: New York. Duxbury Press: Toronto.

Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization.* Wiley: New York.

Särndal, C. E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling.* Springer-Verlag: New York.

Scott, A.J. and Holt, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association.* **77**, 848-854.

Scott, A. J. and Styan, G. P. H. (1985). On a separation theorem for generalized eigenvalues and a problem in the analysis of sample surveys. *Linear algebra and its applications.* **70**, 209-24. Duxbury Press: Toronto.

Shah, B. V., Barnwell, B.G. and Bieler, G.S. (1996) *SUDAAN Users's Manual: Release 7.0*, Research Triangle Institute, Research Triangle Park, NC.

Shao, J. (1996). Resampling methods in sample surveys. *Statistics* 27: 203 - 254.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach.* Wiley: New York.