Survey Methodology, December 2001
Vol. 27, No. 2, pp. 197-203
Statistics Canada, Catalogue No. 12-001

Local Polynomial Regression in Complex Surveys

D.R. Bellhouse and J.E. Stafford [1]

Abstract

Local polynomial regression methods are put forward to aid in exploratory data analysis for large-scale surveys. The proposed method relies on binning the data on the x-variable and calculating the appropriate survey estimates for the mean of the y-values at each bin. When binning on x has been carried out to the precision of the recorded data, the method is the same as applying the survey weights to the standard criterion for obtaining local polynomial regression estimates. The alternative of using classical polynomial regression is also considered and a criterion is proposed to decide whether the nonparametric approach to modeling should be preferred over the classical approach. Illustrative examples are given from the 1990 Ontario Health Survey.

Key Words: Covariates; Exploratory data analysis; Kernel smoothing; Regression.

1. Introduction

Following Fuller (1975), multiple linear regression techniques have been studied and used extensively in sample surveys. At least three chapters of Skinner, Holt and Smith (1989) are devoted to this subject. Here we restrict attention to the case in which there is one covariate x for the variate of interest y so that we could consider polynomial regression as well as simple linear regression. In this context we could also consider the nonparametric approach of local polynomial regression, which, for the case of independent and identically distributed random variables, is described in Hardle (1990), Wand and Jones (1995), Fan and Gijbels (1996), Simonoff (1996) and Eubank (1999). Using the survey weights, Korn and Graubard (1998) introduced the use of local polynomial regression for graphical display of complex survey data. However, they did not provide any statistical properties for their procedures. Smith and Njenga (1992) used regression kernel smoothing techniques to obtain robust estimates of the mean and regression parameters for an assumed superpopulation model. Here we use local polynomial regression as an exploratory tool to discover relationships between y and its covariate x.

We assume that the covariate x is measured on a continuous scale. Due to the precision at which the data are recorded for the survey file and the size of the sample, there will be multiple observations at many of the distinct values. This feature of large-scale survey data has been exploited by Hartley and Rao (1968, 1969) in their scale-load approach to the estimation of finite population parameters. Here we exploit this same feature of the data to examine the relationship between y and its covariate x. In recognizing that the data may be naturally binned to the precision of the data, we can consider taking a further step by constructing larger bin sizes. Under this approach we examine the effect of the sampling design on estimates and second order moments.

Suppose that in the finite population of size N, x has k distinct values so that natural binning has taken place, or that x has been categorized into k bins that are wider than the precision of the data. Let x_i be the value of x representing the i^{th} bin, and assume that the values of x_i are equally spaced. The spacing or bin size $b = x_i - x_{i-1}$. The finite population mean for the $y-$values at x_i is $\bar{y}_i$. We assume that a sample of size n taken from this population has the same structure as the population in that there are k bins. From the sample data we calculate the survey estimate of $\hat{\bar{y}}_i$ of $\bar{y}_i$. The finite population proportion of the observations with value x_i is denoted by p_i. This proportion is estimated by the survey estimate $\hat{p}_i$. We assume that $\hat{\bar{y}}_i$ and $\hat{p}_i$ are asymptotically unbiased, in the sense of Särndal, Swensson and Wretman (1992, pages 166-167), for $\bar{y}_i$ and p_i respectively. The survey estimates $\hat{\bar{y}}_i$ for $i = 1, ..., k$ have variance-covariance matrix $\mathbf{V}$. On considering the distinct values x_i as domains, the estimated variance-covariance matrix $\hat{\mathbf{V}}$ may be obtained easily through survey packages such as SUDAAN and STATA.

There are several advantages to binning the data on the covariate x for exploratory data analysis:

- For large surveys, a plot of $\hat{\bar{y}}_i$ against x_i may be more informative and less cluttered than a plot of the raw data.
- By appealing to a finite population central limit theorem on $\hat{\bar{y}}_i$ and imposing a superpopulation assumption on $\bar{y}_i$, a relatively simple model for $\hat{\bar{y}}_i$ may be assumed so that the analyst may easily focus on the central issue considered here, determination of the trend function in x.

[1] D.R. Bellhouse Department of Statistical and Actuarial Sciences, Western Science Centre, University of Western Ontario, London, Ontario N6A 5B7. E-mail: bellhouse@stats.uwo.ca; J.E. Stafford, Department of Public Health Sciences, Faculty of Medecine, McMurrich Building, University of Toronto, Toronto, Ontario, M5S 1A8. E-mail: stafford@utstat.toronto.edu.

- Once $\hat{\mathbf{V}}$ hat has been obtained, then a wide variety of powerful exploratory data analyses can be easily carried out in languages such as S – Plus. Working with the raw data requires continued appeals to SUDAAN or STATA for the appropriate variance estimates.

- By binning the data, an approach to regression analysis is obtained that provides a parallel to other nonparametric approaches to survey data analysis. For example, in categorical data analysis obtained initially by Rao and Scott (1981), in the logistic regression approach of Roberts, Rao and Kumar (1987) or in the generalized linear model approach of Bellhouse and Rao (2000), the tests and associated distributions are obtained through survey estimates of domain means or proportions.

For the superpopulation, we assume that we have a model such that $E_m(\bar{y}_i) = m(x_i)$, where E_m is the superpopulation expectation. We assume further that as we move to a continuum of values on x, then $m(x)$ is a smooth function. The function $m(x)$ is the ultimate function of interest for estimation. In section 2 we provide local polynomial regression methods to estimate $m(x)$. These methods are applied to data from the 1990 Ontario Health Survey in section 3. In section 4, the question is asked: would the classical polynomial regression techniques have served equally as well in modeling $m(x)$? Some future directions for this work are given in section 5. Generally, we adopt the notation of Wand and Jones (1995) in discussing local polynomial regression here.

2. Basic Methodology

For local polynomial regression, the nestimate of $m(x)$ at any value of x is obtained upon minimizing

$$\sum_{i=1}^{k} \hat{p}_i \{\hat{\bar{y}}_i - \beta_0 - \beta_1(x_i - x) - \ldots - \beta_q (x_i - x)^q\}^2 K((x_i - x)/h)/h \quad (1)$$

with respect to $\beta_0, \beta_1, \ldots, \beta_q$. The values that minimize (1) are denoted by $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_q$. Further, for the given value of x, $\hat{m}(x) = \hat{\beta}_0$. In (1), the kernel $K(t)$ is a symmetric function with $\int K(t)dt = 1$, $\int tK(t)dt = 0$, $0 < \int t^2 K(t)dt < \infty$ and

$$R(K) = \int [K(t)]^2 dt < \infty. \quad (2)$$

Also in (1), h is the window width of the kernel. In minimizing (1) to obtain local polynomial regression estimates, there are two possibilities for binning on x. The first is to bin to the precision of the recorded data so that $\hat{\bar{y}}_i$ is calculated at each distinct outcome of x. In other situations it may be practical to pursue a binning on x that is rougher than the accuracy of the data.

In moving from the sample to the population we maintain the same window width h. This is in contrast to Breidt and Opsomer (2000) and Buskirk (1999) who assume a smoothing parameter h_N for smoothing in the full finite population. In the context here, this would yield a function $m_N(x)$, the finite population smoothed version of the $\bar{y}_i$ with smoothing parameter h_N, as a finite population parameter of interest followed by $m(x)$ the hypothetical smooth function under the asymptotic assumptions. We have kept h constant in view of the way in which binning that has been done; the bin structure is the same in the sample as in the population. The choice of the smoothing parameter h depends on the spacing of the x's and the variation in the data (Green and Silverman 1994, pages 43 – 44). The spacing of the covariate is usually dominant in the determination of h. Since the spacing has been kept constant from sample to finite population with the spacing changing only when the asymptotic assumptions are applied, we keep $h_N = h$.

Korn and Graubard (1998) provide a slightly different objective function to (1). They replace the sum over the bins in (1) by the sum over all sampled units and $\hat{p}_i$ in (1) by the sample weights. Korn and Graubard's objective function reduces to (1) plus a term that involves the weighted sum of squares of deviations of sample observations from the binned means where the weights are the sample weights scaled to sum to one. Consequently, the estimate of $m(x)$ is the same in both cases.

The estimate $\hat{m}(x)$ and its first two moments can be expressed in matrix notation. The forms are exactly the same as those that appear, for example, in Wand and Jones (1995, chapter 5.3) whose notation we have adopted. Let the vector of finite population means at the distinct values of x be $\bar{\mathbf{y}} = (\bar{y}_1, \ldots, \bar{y}_k)^{\mathrm{T}}$ and let $\hat{\bar{\mathbf{y}}}$ be its vector of survey estimates. Further, let

$$\mathbf{X}_x = \begin{bmatrix} 1 & x_1 - x & \cdots & (x_1 - x)^q \\ 1 & x_2 - x & \cdots & (x_2 - x)^q \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_k - x & \cdots & (x_k - x)^q \end{bmatrix}$$

and

$$\mathbf{W}_x = \frac{1}{h}\mathrm{diag}\begin{pmatrix} p_1 K((x_1 - x)/h), \\ p_2 K((x_2 - x)/h), \ \ldots \ p_k K((x_k - x)/h) \end{pmatrix}.$$

The matrix $\hat{\mathbf{W}}_x$ is $\mathbf{W}_x$ with p replaced by $\hat{p}$. Then

$$\hat{m}(x) = \mathbf{e}^{\mathrm{T}} (\mathbf{X}_x^{\mathrm{T}} \hat{\mathbf{W}}_x \mathbf{X}_x)^{-1} \mathbf{X}_x^{\mathrm{T}} \hat{\mathbf{W}}_x \hat{\bar{\mathbf{y}}}, \quad (3)$$

where $\mathbf{e}$ is the $k \times 1$ vector $(1, 0, 0, \ldots, 0)^{\mathrm{T}}$. The approximate design-based expectation of $\hat{m}(x)$ is

$$E_p(\hat{m}(x)) = \mathbf{e}^{\mathrm{T}} (\mathbf{X}_x^{\mathrm{T}} \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{X}_x^{\mathrm{T}} \mathbf{W}_x \bar{\mathbf{y}}, \quad (4)$$

where E_p denotes expectation with respect to the sampling design. We can also consider (4) as a smoothed estimate of $m(x)$ so that $\hat{m}(x)$ is also an estimate of $m(x)$. In the derivation of (4) we note that $E_p(\hat{\mathbf{y}}) = \bar{\mathbf{y}}$ and $E_p(\hat{\mathbf{W}}_x) = \mathbf{W}_x$ for large sample size n. Further, in (3) we can write $\hat{\mathbf{W}}_x = \mathbf{W}_x + \hat{\mathbf{A}}$, where $\hat{\mathbf{A}} = \hat{\mathbf{W}}_x - \mathbf{W}_x$. We use the first two terms in the expansion $(\mathbf{I} + \mathbf{B})^{-1} = \mathbf{I} - \mathbf{B} + \mathbf{B}^2 - \mathbf{B}^3 + \ldots$ as an approximation to complete the derivation. Using the same techniques, the approximate design-based variance is given by

$$V_p(\hat{m}(x)) = \mathbf{e}^{\mathrm{T}}(\mathbf{X}_x^{\mathrm{T}}\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{X}_x^{\mathrm{T}}\mathbf{W}_x\mathbf{V}\mathbf{W}_x\mathbf{X}_x(\mathbf{X}_x^{\mathrm{T}}\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{e}. \quad (5)$$

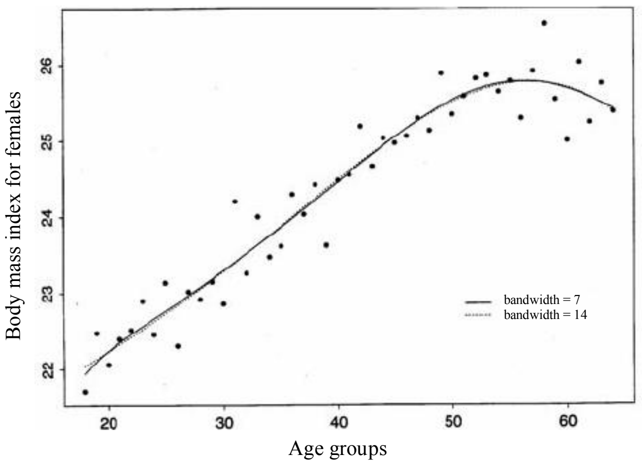
The results in (4) and (5) were obtained ignoring higher order terms in $1/n$. An estimate of the variance $\hat{V}_p(\hat{m}(x))$ is obtained on substituting the survey estimate $\hat{\mathbf{V}}$ for $\mathbf{V}$ and $\hat{\mathbf{W}}_x$ for $\mathbf{W}_x$ in (5).

3. Examples from the Ontario Health Survey

We illustrate local polynomial regression techniques with data from the Ontario Health Survey (Ontario Ministry of Health 1992). This survey was carried out in 1990 using a stratified two-stage cluster sample. The purpose was to measure the health status of the people of Ontario and to collect data relating to the risk factors of major causes of morbidity and mortality in Ontario. The survey was designed to be compatible with the Canada Health Survey carried out in 1978 – 79. A total sample size of 61,239 people was obtained from 43 public health units across Ontario. The public health unit was the basic stratum with an additional division of the health unit into rural and urban strata so that there were a total of 86 strata. The first stage units within a stratum were enumeration areas taken from the 1986 Census of Canada. An average of 46 enumeration areas was chosen within each stratum. Within an enumeration area, dwellings were selected, approximately 15 from an urban enumeration area and 20 from a rural enumeration area. Information was collected on members of the household within the dwelling.

Several health characteristics were measured. We focus on one continuous variable from the survey, Body Mass Index (BMI). The BMI is a measure of weight status and is calculated from the weight in kilograms divided by the square of the height in meters. The index is not applicable to adolescents, adults over 65 years of age and pregnant or breastfeeding women. The measure varies between 7.0 and 45.0. A value of the BMI less than 20.0 is often associated with health problems such as eating disorders. An index value above 27.0 is associated with health problems such as hypertension and coronary heart disease. Associated with the BMI is another measure, the Desired Body Mass Index (DBMI). The DBMI is the same measure as BMI with actual weight replaced by desired weight. A total of 44,457 responses were obtained for the BMI and 41,939 for the DBMI.

When there are only a few distinct outcomes of x, binning on x is done in a natural way. For example, in investigating the relationship between the body mass index (BMI) and age, the age of the respondent was reported only at integral values. The solid dots in Figure 1 are the survey domain estimates of the average BMI $(\hat{\bar{y}}_i)$ for women at each of the ages 18 through 65 (x_i). The solid and dotted lines show the plot of $\hat{m}(x)$ against x using bandwidths $h = 7$ and $h = 14$ respectively. It may be seen from Figure 1 that BMI increases approximately linearly with age until around age 50. The increase slows in the early 50s, peaks at age 55 or so, and then begins to decrease. On plotting the trend lines only for BMI and the desired body mass index (DBMI) for females as shown in Figure 2, it may be seen that, on average, women desire to reduce their BMI at every age by approximately two units.



Figuire 1. Age trend in BMI for females.

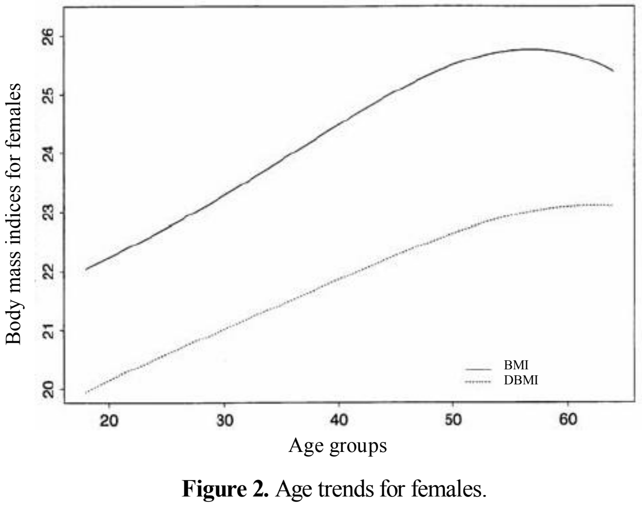


Figure 2. Age trends for females.

In other situations it is practical to construct bins on x wider than the precision of the data. To investigate the relationship between what women desire for their weight $(\text{DBMI} = \hat{\bar{y}}_i)$ and what women actually weigh $(\text{BMI} = x_i)$ the $x-$values were grouped. Since the data were very sparse for values of BMI below 15 and above 42, these data were removed from consideration. The remaining groups were 15.0 to 15.2, 15.3 to 15.4 and so on, with the value of x_i chosen as the middle value in each group. The binning was done in this way for the purposes of illustration to obtain a wide range of equally spaced nonempty bins. For each group the survey estimate $\hat{\bar{y}}_i$ was calculated. The solid dots in Figure 3 show the survey estimates of women's DBMI for each grouped value of their respective BMI. The scatter at either end of the line reflects the sampling variability due to low sample sizes. The plot shows a slight desire to gain weight when the BMI is at 15. This desire is reversed by the time the BMI reaches 20 and the gap between the desire (DBMI) and reality (BMI) widens as BMI increases.

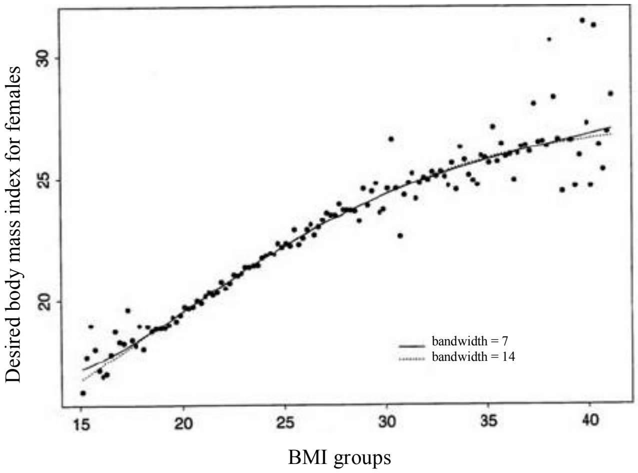


Figure 3. BMI trend in DBMI for females.

4. Parametric Versus Nonparametric Regression

Local polynomial regression allows us to obtain nonparametrically a functional relation between y and x. However, a parametric model may also be reasonable. For example, on examining Figure 1 showing the Body Mass Index against age, we might consider the parametric model that y has a quadratic relationship to x. We may also want to test in Figure 2 if the two lines are parallel, or equivalently that the difference between the Body Mass Index and the Desired Body Mass Index for females is constant over all ages. This would involve modeling the trend lines as second degree polynomials and testing for equality in the trend lines of the parameters associated with the quadratic term as well as the parameters associated with the linear term. In all cases, the question arises as to whether or not the data can be adequately modeled by a polynomial relationship between y and x. One method that we propose as an answer to this question is to calculate the confidence bands based on local polynomial regression. These bands can be thought of as providing a region of acceptable model representations. If an appropriate parametric regression line falls within the bands, then it provides a reasonable model description of the data. The $100(1-\alpha)\%$ local polynomial regression bands are obtained by ploting

$$\hat{m}(x) \pm z_{\alpha/2}\sqrt{\hat{V}_p(\hat{m}(x))} \qquad (6)$$

over a range of values of x, where $z_{\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the standard normal distribution, where $\hat{m}(x)$ is determined from (3) and where $\hat{V}_p(\hat{m}(x))$ is (5) with $\mathbf{V}$ replaced by its sample estimate $\hat{\mathbf{V}}$.

The parametric regression line to be tested may be obtained in one of two ways depending upon what sample information is available. If the complete sample file with sampling weights is available, then the standard regression approach in, for example, SUDAAN may be used. If only the binned data are available, in particular the survey estimates $\hat{\bar{y}}_i$ with estimated variance-covariance matrix $\hat{\mathbf{V}}$, then another approach is needed.

For this second approach assume that $m(x_i) = \mathbf{x}_i^T\boldsymbol{\beta}$, where $\mathbf{x}_i^T = (1, x_i, x_i^2, ..., x_i^q)$ and where $\boldsymbol{\beta}^T = (\beta_0, \beta_1, ..., \beta_q)$ is the vector of regression coefficients. For the finite population we assume that $\bar{y}_i = \mathbf{x}_i^T\boldsymbol{\beta} + \varepsilon_i$, where the errors are deviations of the actual finite from the model. For simplicity, we assume that these errors have mean 0 and variance-covariance matrix $\sigma^2\mathbf{I}$. Since the data are given by the survey estimates $\hat{\bar{y}}_i$ with variance-covariance matrix $\mathbf{V}$, the operative model is

$$\hat{\bar{y}}_i = \mathbf{x}_i^T\boldsymbol{\beta} + \delta_i, \qquad (7)$$

where the δ_i have mean 0 and variance-covariance matrix $\Sigma = \sigma^2\mathbf{I} + \mathbf{V}$. The usual weighted least squares estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\sum\nolimits^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\sum\nolimits^{-1}\hat{\bar{\mathbf{y}}}, \qquad (8)$$

where the i^{th} row of $\mathbf{X}$ is $\mathbf{x}_i^T$, $i = 1, ..., k$. In terms of data analysis it is necessary to replace Σ in (8) by its estimate $\hat{\Sigma}$. Now the survey estimate of $\mathbf{V}$ is $\hat{\mathbf{V}}$ so that it remains to find an estimate of σ^2. This may be obtained through $\text{rss} = (\hat{\bar{\mathbf{y}}} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\hat{\bar{\mathbf{y}}} - \mathbf{X}\hat{\boldsymbol{\beta}})$, the residual sum of squares, by one of two ways.

The first method is to approximate the expected residual sum of squares under model (7) and solve directly for σ^2. Upon using the expansion $(\mathbf{I}+\mathbf{B})^{-1} = \mathbf{I} - \mathbf{B} + \mathbf{B}^2 - \mathbf{B}^3 + ...$ we find

$$E(\text{rss}) \cong (n-q-1)\,\sigma^2 + \text{tr}(\mathbf{V}) - \text{tr}(\mathbf{X}^T\mathbf{V}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}). \quad (9)$$

The estimate of σ^2 is obtained on setting rss equal to the right hand side of (8) with $\mathbf{V}$ replaced by $\hat{\mathbf{V}}$ and then solving for σ^2. This leads to an iterative approach to model fitting. An initial estimate of $\boldsymbol{\beta}$ is obtained from (8) with $\mathbf{V}$ replaced by the survey estimate $\hat{\mathbf{V}}$. Then σ^2 is estimated through (9) and a new estimate of $\boldsymbol{\beta}$ using $\hat{\Sigma} = \hat{\sigma}^2\mathbf{I} + \hat{\mathbf{V}}$ is obtained. The process is repeated until convergence is obtained in the estimate of σ^2. If the estimate of σ^2 is negative, it is set to 0. The second method for estimating σ^2 is obtaining by first treating the errors in (7) as multivariate normal variables. Then a profile likelihood for σ^2 can be obtained on replacing $\boldsymbol{\beta}$ and $\mathbf{V}$ by their estimates. The most influential term in this profile likelihood is

$$\mathbf{r}^T(\sigma^2\mathbf{I}+\hat{\mathbf{V}})^{-1}\mathbf{r}, \qquad (10)$$

where $\mathbf{r}=\hat{\mathbf{y}}-\mathbf{X}(\mathbf{X}^T(\sigma^2\mathbf{I}+\hat{\mathbf{V}})^{-1}\mathbf{X})^{-1}\mathbf{X}^T(\sigma^2\mathbf{I}+\hat{\mathbf{V}})^{-1}\hat{\mathbf{y}}$ is the vector of residuals. An approximation to the profile likelihood estimate $\hat{\sigma}^2$ is that value of σ^2 which minimizes (10).

To provide examples of the question of the adequacy of parametric regression, we examined two different variables in the Ontario Health Survey and their relationship to the body mass index (BMI). These were age and fat consumption as a percentage of total energy consumption. For age the binning was natural and at the precision of the recorded data. Age was restricted to the range of 18 to 65 years since the index is not applicable outside this range and age was recorded in years. The scatterplot of BMI against age with the accompanying local polynomial regression line is shown in Figure 1. The survey data on fat consumption in percentages were recorded to three decimal places. Due to the sparseness of the data at the extremes we looked at fat consumption in the range of 14 to 56% of total energy consumption. Further, we binned the data on the covariate (fat consumption) using bins 14.0 up to 14.2, 14.2 up to 14.4 and so on; the midpoints of the bins (14.1, 14.3 and so on) were used as the x_i. At each bin the survey estimate $\hat{y}_i$ for BMI was calculated. It is the binned data that appear as a scatterplot of BMI against fat consumption in Figure 5. The solid line in Figure 5 is the local polynomial regression line with $q=1$ for BMI on fat content. As in Figure 3, the larger variability at the extremes reflects greater sampling variability due to smaller sample sizes at the extremes. From Figure 5 it appears that BMI increases slightly as fat consumption increases. Since the complete data file for the survey was available, regression lines for all variables were obtained through SUDAAN.

In Figure 4 the solid lines are the 95% confidence bands based on (6) and the dashed line is the parametric second degree polynomial regression line. Since the dashed line falls near the border for women in their thirties and outside the bands for women in their early sixties, a second degree polynomial barely adequately describes the relation between BMI and age. Another model might be preferable. Figure 6 shows the same 95% confidence bands but for the consumption of fat as a percentage of total energy consumption. In this case the dotted line is the simple linear regression line of BMI on fat consumption. For fat consumption the line falls completely within the confidence bands so that simple linear regression appears to be an adequate description of the model relationship.

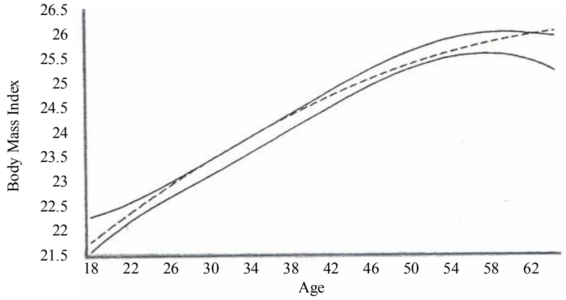


Figure 4. Confidence Bands for the Age Trend in BMI for Females.

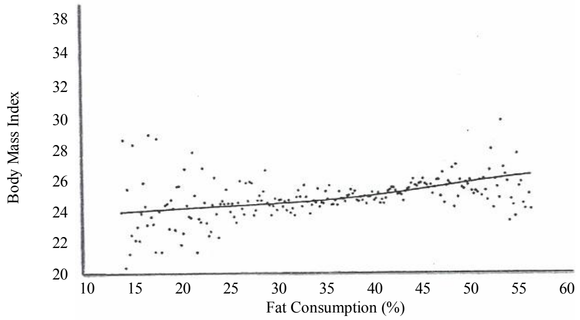


Figure 5. BMI Trend in Fat Consumption.

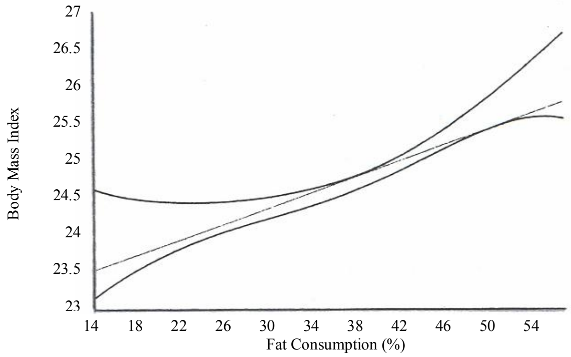


Figure 6. Confidence Bands for Fat Consumption Trend in BMI.

If the data have been binned to the precision of the data as in the case of age above, and if the exploratory analysis is complete, we can stop. The estimates and variance estimates obtained are equal to the estimates and variance estimates obtained from the raw data. This may be seen on examining (3). The term on the right hand side of (3) can be expressed as a sum over the sample of the sample weights times a new measurement obtained from the raw y – measurement times an appropriate value taken from $\mathbf{e}^{\mathrm{T}}(\mathbf{X}_x^{\mathrm{T}}\hat{\mathbf{W}}_x\mathbf{X}_x)^{-1}\ \mathbf{X}_x^{\mathrm{T}}\mathbf{W}_x^*$ times the total of the sample weights, where $\mathbf{W}_x^*$ is $\mathbf{W}_x$ with the p_i's removed. These adjusted $y-$ measurements may be fed into SUDAAN or STATA to obtain the required approximate variance estimate. It may be that the binning has been rougher than the precision of the data or that some bins have been dropped in the tails of the distribution of x due to sparseness of the data in those bins. Both of these situations occurred in analyzing the relationship of BMI to fat consumption. Once the exploratory analysis has been completed we can return with a final model and smoothing parameter, if a nonparametric approach is used in the final analysis, and apply to model to the raw data obtaining variance estimates through SUDAAN or STATA as necessary. Depending on the amount of roughness in the binning and the number of bins dropped due to sparseness in the data, the variance estimates obtained from the raw will be approximately the same as those from the binned data.

5. Future Directions

Like Bellhouse and Stafford (1999), this paper adapts a modern method of smoothing for the analysis of complex survey data. It represents an example of a host of regression techniques that could be used. To describe these we embed the current context in a general framework hinting at future work. In doing so we mimic the developments of Hastie and Tibshirani (1990).

Here a smoother is said to be linear if fitted values are obtained by applying a matrix $\mathbf{S}$ to a response vector $\mathbf{y}$. As in the case of simple linear regression for independent and identically distributed data, we let $\mathbf{H} = (\mathbf{X}^T \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}^T \Sigma^{-1}$ and further denote $(\mathbf{X}_x^{\mathrm{T}} \hat{\mathbf{W}}_x \mathbf{X}_x)^{-1} \mathbf{X}_x^{\mathrm{T}} \hat{\mathbf{W}}_x$ as $\mathbf{S}_p$. Both are examples of $\mathbf{S}$. In addition, the response vector of binned means is a type of smooth $\hat{\mathbf{y}} = \mathbf{S}_b \mathbf{y}$, where $\mathbf{y}$ is the vector of all sample responses and where $\mathbf{S}_b$ involves the sample weights. Also the usual regression context involves applying a matrix similar to $\mathbf{H}$ to the full response vector $\hat{\mathbf{y}} = \mathbf{H}_f \mathbf{y}$. So moving from usual regression to regressing means to local polynomial smoothing reduces to applying different smoothing matrices to $\mathbf{y}$:

$$\mathbf{H}_f \, \mathbf{y} \rightarrow \mathbf{H} \, \mathbf{S}_b \, \mathbf{y} \rightarrow \mathbf{S}_p \, \mathbf{S}_b \, \mathbf{y}.$$

In general $\mathbf{S}_p$ can be replaced by any smoother $\mathbf{S}$ and the methods extended to multiple covariates.

There are many advantages to binning the response from both a theoretical and practical standpoint. Standard smoothing tools, like those found in *Splus*, can be applied without modification of the smoother due to sampling issues. In addition, in the case of the additive model, finite population central limit theorems can be invoked and issues like degrees of freedom, choice of smoothing parameter, optimizing a criterion, can be handled in the usual manner. In the case of multiple covariates $x_1, \ldots, x_q$ the curse of dimensionality will result in sparse bins not allowing the use of the central limit theorem. This may be countered in the usual way by binning partial residuals one dimension at a time. Here smoothers $\mathbf{S}_j \mathbf{S}_{b_j}$, $j = 1, \ldots, q$ would be used in a backfitting algorithm.

It is our intention to study additive and generalized additive models in the above manner and to introduce these techniques to the analysis of complex survey data.

Acknowledgements

The authors would like to thank Rob Tibshirani for his helpful comments on this paper and the referees for their comments that helped to improve the presentation of the paper as well as to clarify some technical issues. This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada.

References

Bellhouse, D.R., and Rao, J.N.K. (2000). Analysis of domain means in complex surveys. *Journal of Statistical Planning and Inference*, to appear.

Bellhouse, D.R., and Stafford, J.E. (1999). Density estimation from complex surveys. *Statistica Sinica*, 9, 407-424.

Breidt, F.J., and Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. Submitted for publication.

Buskirk, T. (1999). *Using Nonparametric Methods for Density Estimation with Complex Survey Data*. Ph.D. dissertation, Arizona State University.

Eubank, R.L. (1999). *Nonparametric Regression and Spline Smoothing*. New York: Marcel Dekker.

Fan, J., and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. London: Chapman and Hall.

Fuller, W.A. (1975). Regression analysis for sample survey. *Sankhyā*, C, 37, 117-132.

Green, P.J., and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models*. London: Chapman and Hall.

Hardle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press: Cambridge.

Hartley, H.O., and Rao, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 547-557.

Hartley, H.O., and Rao, J.N.K. (1969). A new estimation theory for sample surveys, II. In *New Developments in Survey Sampling*, (Eds. N.L. Johnson and H. Smith). New York: John Wiley & Sons, Inc. Inter-Science, 147-169.

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

Korn, E.L., and Graubard, B.I. (1998). Scatterplots with survey data. *American Statistician*, 52, 58-69.

Ontario Ministry of Health (1992). *Ontario Health Survey: User's Guide, Volumes I and II.* Queen's Printer for Ontario.

Rao, J.N.K., and Scott, A.J. (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.

Roberts, G., Rao, J.N.K. and Kumar, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.

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

Simonoff, J.S. (1996). *Smoothing Methods in Statistics*. New York: Springer-Verlag.

Skinner, C.J., Holt, D. and Smith, T.M.F. (1989). *Analysis of Complex Surveys*. New York: John Wiley & Sons, Inc.

Smith, T.M.F., and Njenga, E. (1992). Robust model-based methods for analytical surveys. *Survey Methodology*, 18, 187-208.

Wand, M.P., and Jones, M.C. (1995). *Kernel Smoothing*. London: Chapman and Hall.