2009

Scatterplots with Survey Data

Barry I- Graubard

Edward L. Korn

Sample Surveys: Inference and Analysis, Vol. 29B
ISSN: 0169-7161
© 2009 Published by Elsevier B.V.
DOI: 10.1016/S0169-7161(09)00237-5

Scatterplots with Survey Data

Barry I. Graubard and Edward L. Korn

1. Introduction

The scatterplot is one of the most useful graphical displays of bivariate data. It allows one to see general trends and atypical points simultaneously, as well as other aspects of the data. Data collected in a survey, however, have some additional features that can make a simple scatterplot less useful. One such feature is that individuals in the sample represent different numbers of individuals in the population. The sample weights of the sampled individuals effectively estimate these numbers. A second feature of survey data is that some of it may be imputed to account for item nonresponse. A third feature is that the sample sizes can be large. A fourth feature is that the observations may have intraclass correlation due to cluster sampling. As will be shown below, standard scatterplots that are used for simple random samples that ignore these features can be misleading or hard to interpret. We know of no "super plot" that will be as successful in the survey setting as the simple scatterplot is in the nonsurvey setting. Instead, we present in this chapter different modifications of the scatterplot, demonstrated by examples, that can improve the presentation of survey data. By and large, these modified plots are not new, but their application to survey data may not be well known. There has been little new literature on scatterplots with complex survey data since Korn and Graubard (1998) and Korn and Graubard (1999). Most of this chapter is taken from those two sources with some minor updates.

2. Modifications of scatterplots for survey data

In this section, we present some techniques that can be used to modify a scatterplot to incorporate various aspects of survey data. First, we describe the use of bubble plots in which the sizes of the plotted circles are proportional to the sample weights of the points. Examples are given showing that such bubble plots can perform better than a simple scatterplot in (a) describing the population distribution and (b) identifying influential points in a weighted analysis (which is typically used when analyzing survey data). However, for moderate-to-large sample sizes, a bubble plot can be hard to interpret because of the overlapping bubbles. For this situation, we consider in Section 2.2 using a "sampled scatterplot," in which the sampled data is resampled proportionally to the sample weights, yielding a data set that can be plotted without circles but still represents the population distribution.

Plots of large data sets can be problematic because of overlapping plotted points. This can especially be a problem when the raw data has been implicitly or explicitly rounded. An example is given in Section 2.3, along with the possible solution of "jittering" the data, that is, adding a small amount of random noise to the data before plotting. In Section 2.4, we discuss scatterplots in which some of the plotted points represent imputed data values to account for item nonresponse. In Section 2.5, we consider using conditional mean and percentile curves constructed using kernel smoothing for nonparametrically displaying the relationship between Y and X when the sample sizes are large. Finally, in Section 2.6, a modeling alternative approach is considered that uses regression splines to investigate relationships between Y and X.

2.1. Accounting for the sample weights: bubble plots

Survey designs typically specify that individuals are to be sampled with unequal probabilities of selection. The sample weight associated with an individual is the inverse of that individual's probability of being included in the sample, adjusted, if necessary, for nonresponse. There is often an additional poststratification to ensure that the sum of the sample weights equals known population values for various subgroups (e.g., age/race/sex subgroups). The sample weights effectively represent the number of individuals in the population that the sampled individual represents.

Figure 1 is a scatterplot of daughter's birthweight versus mother's birthweight for mothers aged 30–39 years at the time of birth; the data are from the 1988 National Maternal and Infant Health Survey which sampled vital records corresponding to live births, late fetal deaths, and infant deaths in the United States (Sanderson et al., 1991). For the live birth component of the survey, mothers corresponding to sampled birth certificates were mailed a questionnaire. The birthweight of the child was taken from the birth certificate (reported in grams) and the birthweight of the mother was taken from the mother's questionnaire (reported in ounces, converted to grams for the plot). Relationships between the birthweights of mothers and their children have been studied previously using data from this survey (Wang et al., 1995). We restrict attention to first births that were daughters, and mother-daughter pairs with nonmissing birthweights ($n = 225$). Figure 1 is a misleading representation of the population because it ignores the sample weights; this survey oversampled low birthweight babies and black babies (Table 1) (Nonresponse and poststratification adjustments to the sample weights were relatively small.). One possibility to more accurately reflect the population is displayed by the bubble plot in Fig. 2; the areas of the circles are proportional to the sample weights.

Another use for using the size of bubbles to designate sample weights is to help identify influential points in an analysis. We now give an example using an analysis of the association of developing cancer with baseline transferrin saturation values based on women participating in the epidemiologic follow-up of the first National Health and Nutrition Examination Survey (National Center for Health Statistics et al., 1987). This

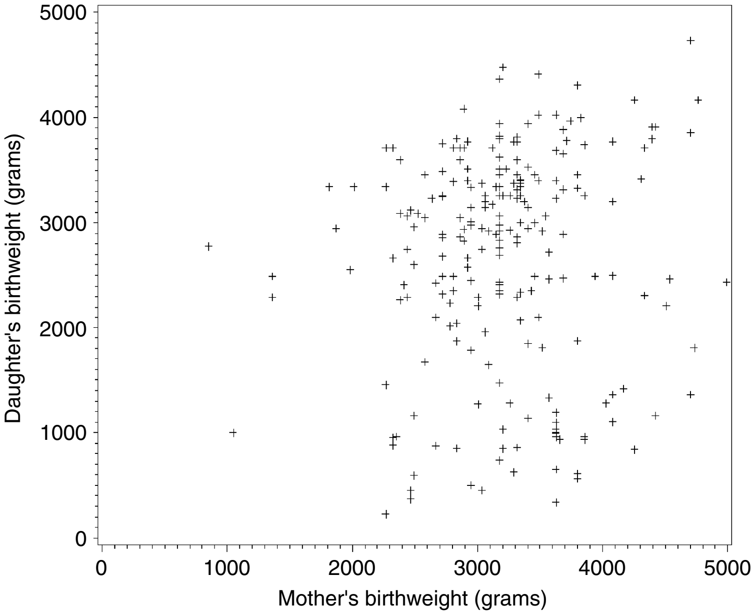

Fig. 1. Simple scatterplot based on data from mothers aged 30–39 surveyed in the 1988 National Maternal and Infant Health Survey.

Table 1
Sampling strata and sampling rates of 1988 National Maternal and Infant Health Survey

Strata		Sampling Rate
Race	Birth Weight (grams)	
Black	<1500	1/14
	1500–2499	1/55
	≥2500	1/113
Nonblack	<1500	1/29
	1500–2499	1/160
	≥2500	1/720

association has been previously studied by us Korn and Graubard (1995) and others (Stevens et al., 1988). We follow the previous analyses and remove women from the analysis who had cancer at the baseline or who developed it within four years of the baseline survey; this leaves 197 women who developed cancer and 5073 who did not. The sample weights ranged from 611 to 186,062 (coefficient of variation = 97%), with the distribution being similar for the women who developed cancer and for those who did not. We consider a logistic regression of the probability of developing cancer on transferrin saturation and other covariates described in footnote 1 of Table 2. A classical survey analysis uses weighted estimators; the weighted logistic regression coefficient for transferrin saturation is given in the first line of Table 2.

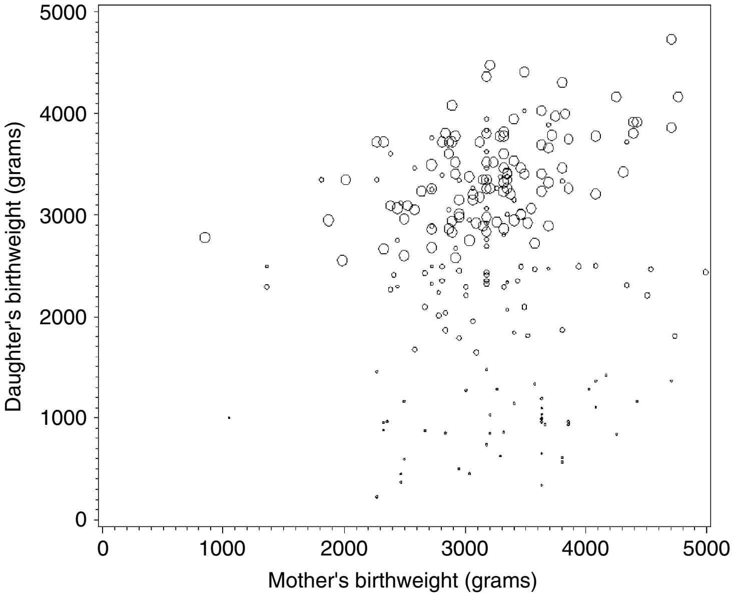

Fig. 2. Bubble plot of data plotted in Fig. 1; areas of circles are proportional to the sample weights.

Table 2
Weighted logistic regression coefficient (±standard error) for transferrin saturation from a multiple logistic regression of the probability of developing cancer on transferrin saturation and other covariates[a], dropping certain data points

Point[b] Dropped from the Analysis	Sample Size	$\beta \pm \mathrm{SE}$[c]
None	5270	.025 ± .014
Point A	5269	.009 ± .009
Point B	5269	.024 ± .014
Point C	5269	.028 ± .014

[a] Covariates included in the model are age at the baseline examination; smoking (never smoked, former smoker, current smoker, and unknown); race (white and nonwhite); senior status (age $\geq$65 and age <65 years); living in poverty census Enumeration District (yes, no); and family income (<\$3000, \$3000–6999, \$7000–9999, \$10,000–14,999, and $\geq$\$15,000).

[b] Points are designated in Fig. 3.

[c] To account for the complex sampling design, the computer program SUDAAN (Shah et al., 1995) was used to calculate the standard errors.

An added variable plot, also known as a partial regression leverage plot, is useful for identifying influential points in a multiple linear regression of Y on X and Z (Cook and Weisberg, 1994, Chapter 12.1; Atkinson, 1985, Chapter 5.2-3). It is a plot of the residuals from the regression of the dependent variable Y on the covariate vector Z (which includes the intercept) versus the residuals from the regression of the independent variable currently under study (X) on Z. The slope of the least-squares line based on this plot is the same as the regression coefficient for X in the multiple linear regression. For a multiple *logistic* regression of a binary Y on X and Z, O'Hara Hines and Carter (1993) suggest calculating the residuals from the linear regression of $\sqrt{p(1-p)}\left[\log\frac{p}{1-p}+\frac{Y-p}{p(1-p)}\right]$ on

$\sqrt{p(1-p)}X$ and $\sqrt{p(1-p)}Z$ and plotting these residuals against the residuals from the linear regression of $\sqrt{p(1-p)}X$ on $\sqrt{p(1-p)}Z$, where p is the predicted probability that $Y = 1$ based on the multiple logistic regression. The slope of the least-squares line through this plot will equal the logistic regression coefficient of X from the multiple regression.

In our application, a *weighted* multiple logistic regression is used since the observations have sample weights. To account for this in the added variable plot, the linear regressions used to obtain the residuals above need to be weighted linear regressions, and the predicted values p need to be obtained from the weighted logistic regression. With these modifications, the slope from a weighted least-squares regression through the added variable plot will equal the regression coefficient of X from the weighted logistic regression of Y on X and Z.

Figure 3 is the added variable plot for transferrin saturation; the areas of the circles are proportional to the sample weights. The dashed line in Fig. 3 is the weighted least-squares line; its slope is .025, the same at the logistic regression coefficient for transferrin saturation (Table 2). The mass of plotted points on the bottom left of the plot is not aesthetically pleasing, but for the purpose of identifying influential points is not troublesome. The point labeled A would appear to be highly influential. This is confirmed by noting that when this point is dropped from the analysis, the logistic regression coefficient for transferrin saturation changes from .025 to .009 (Table 2). This point is also highly influential for estimating the standard error of the coefficient; it changes from .014 to .009 with removal of the point.

A simple scatterplot without the circles would not be as successful as Fig. 3 in identifying influential points. For example, without the circles, the point labeled B

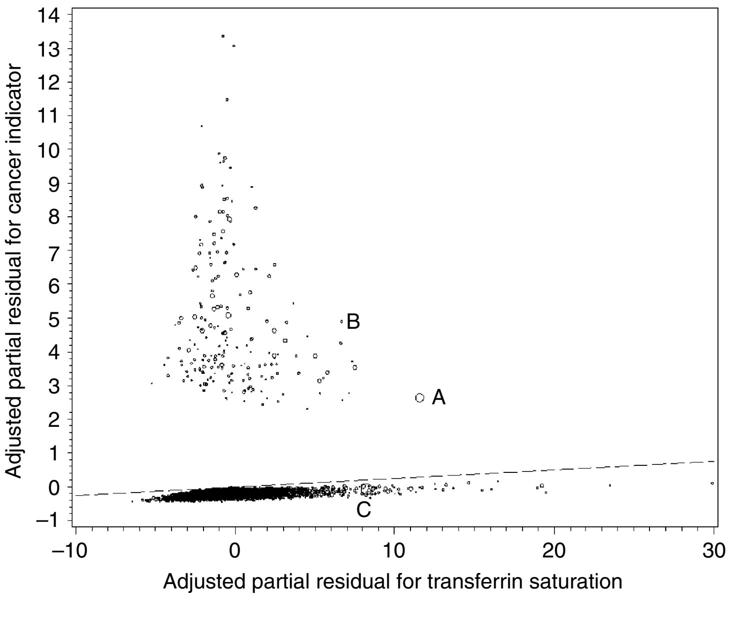

Fig. 3. Added variable plot for transferrin saturation based on weighted multiple logistic regression described in Table 2. Dashed line is weighted least-squares line; labeled points are described in the text.

might appear about as influential as point A. However, because of its small sample weight, it has very little influence on the coefficient (Table 2). On the other hand, it is not sufficient to ignore the plot and assume that observations with large sample weights will be influential. For example, the observation above the label *C* in Fig. 3 has a larger sample weight than point *A*. From its plotted position, however, we would not expect it to be influential and it is not (Table 2).

2.2. Accounting for the sample weights: sampled scatterplots

An alternative strategy to using a bubble plot is to use a "sampled scatterplot." The idea is to sample the data with probabilities proportional to the sample weights; the resulting sampled data is then approximately representative of the population and can be plotted ignoring the sample weights. Figure 4 ($n = 100$) is a sampled scatterplot of the data displayed in Fig. 2. The *i*th observation from the original data set was included in Fig. 2 if a uniform (pseudo-)random number was less than $w_i/w_{\max}$, where w_i is the sample weight of the *i*th observation and $w_{\max}(= 1008.515)$ is the largest sample weight of the 225 observations in Fig. 2. In general, one samples the *i*th data point to be plotted an expected number of times equal to $w_i/(cw_{max})$, where c is chosen to control the expected sample size of the plot. In some cases, the same observations may be sampled multiple times resulting in overlapping points on a plot. In these cases, one might consider jittering the data in the plot to separate the overlapping points as described in Section 2.3. The idea of resampling survey data to eliminate the effects of the sample weights in further analysis has been used by Murthy and Sethi (1965) and Hinkins et al. (1994) to use conventional nonsurvey methods of analysis for survey data.

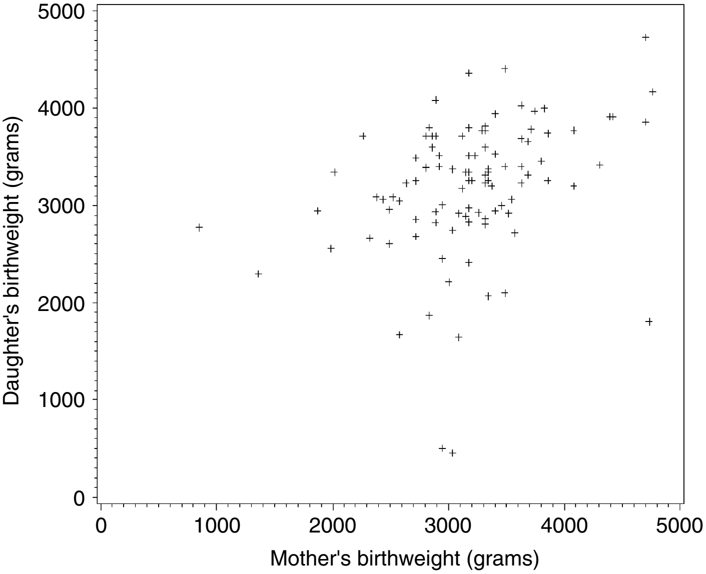

Fig. 4. Sampled scatterplot of data plotted in Fig. 2. Points were chosen for plotting with probability proportional to their sample weights.

There is no question that there is a loss of information in going from Fig. 2 to Fig. 4. Therefore, Fig. 2 would be the preferred plot for data cleaning. Additionally, weighted estimation using the full data set should be used for estimating population parameters. However, as a visual display of the population, we prefer Fig. 4 to Fig. 2, and this preference would become stronger if the sample size were larger, see the height/age example given below.

For some applications, it may be useful to sample points for a sampled scatterplot not just proportionally to the sample weights. For example, suppose we are interested in the relationship of mother's and daughter's birthweights for black and nonblack daughters. Only four of the data points in Fig. 4 correspond to black daughters, and this is reflective of the population. Since black babies were oversampled in the survey, there is much more information available. Figure 5 is a sampled scatterplot in which data points corresponding to black daughters were sampled with probability $w_i/166.642$ (166.642 is the largest sample weight corresponding to a black baby in the original data), whereas data points corresponding to nonblack daughters were sample with probability $w_i/1008.515$. Therefore, although Fig. 5 is not representative of the population, it is representative of the black and nonblack populations separately. It appears from Fig. 5 that there is a stronger positive correlation among the nonblack mother-daughter pairs than among the black mother-daughter pairs. This can also be demonstrated numerically by comparing the weighted correlations using all the sampled data for the nonblack and black pairs: $0.32(n = 170)$ versus $0.07(n = 55)$, respectively.

Figure 5 also displays an additional characteristic of the data that may not have been apparent before—there are many observations with mother's birthweight equal to 3175.133 grams, converted from 7 pounds, 0 ounces. A better representation of the

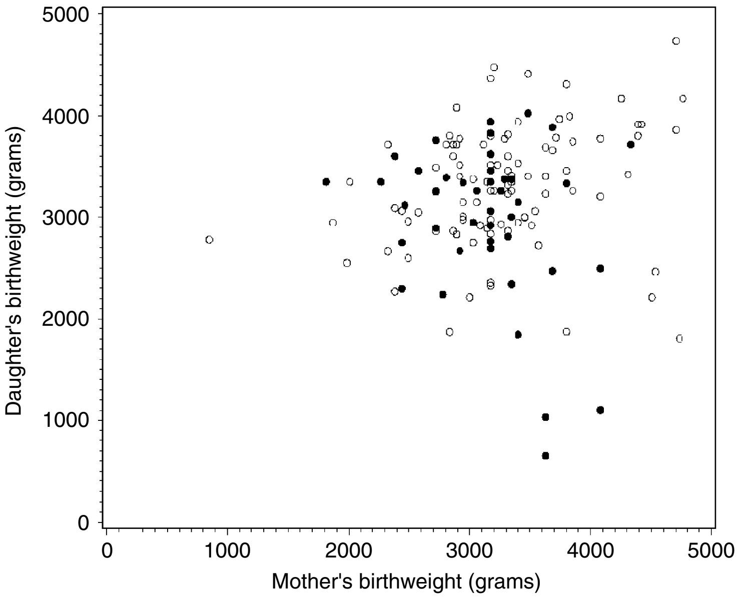

Fig. 5. Sampled scatterplot of data plotted in Fig. 2. Black daughters (filled-in circles) were sampled for plotting at approximately six times the rate as nonblack daughters (open circles).

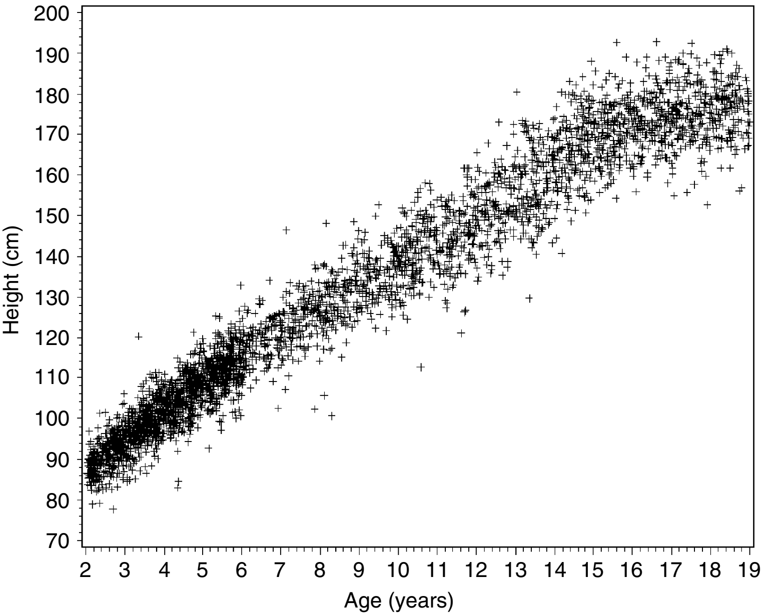

Fig. 6. Simple scatterplot of height versus age for boy aged less than 19 years sampled in the second National Health and Nutrition Examination Survey.

population might be obtained by randomly jittering the data to account for the rounding in the reporting (see Section 2.3 below).

Another application of the sampled scatterplot is when the sample size is large. Figure 6 is a simple scatterplot of height versus age for the 3667 boys aged 2–19 years sampled in the second National Health and Nutrition Examination Survey. The sample weights for these boys ranged from 1359 to 47,385, with a coefficient of variation of 71%, see McDowell et al. (1981) for full details of this survey. Besides being an unappealing plot because of the mass of points being plotted, the plot is also not representative of the population because of the differing sample weights. In particular, boys aged five years or younger were sampled in this survey at three times the rate of boys six years or older. This is reflected in Fig. 6 in the increased density of plotted points for age less than six. Because of the large number of plotted points, a bubble plot version of Fig. 6 would not be useful. We can solve the two problems of excessive density and representativeness at once by using a sampled scatterplot, see Fig. 7 in which $n = 699$ points are plotted.

2.3. *Accounting for overlap and rounding: jittering*

In plotting a small number of observations, occasionally multiple observations will have values so close (or identical) as to make their plotted points indistinguishable. The easy solution to this problem is to displace by a small amount such points. With larger data sets, the problem can become more acute. For example, Fig. 8 is a bubble plot of systolic blood pressure versus the logarithm of blood lead values for 595 white

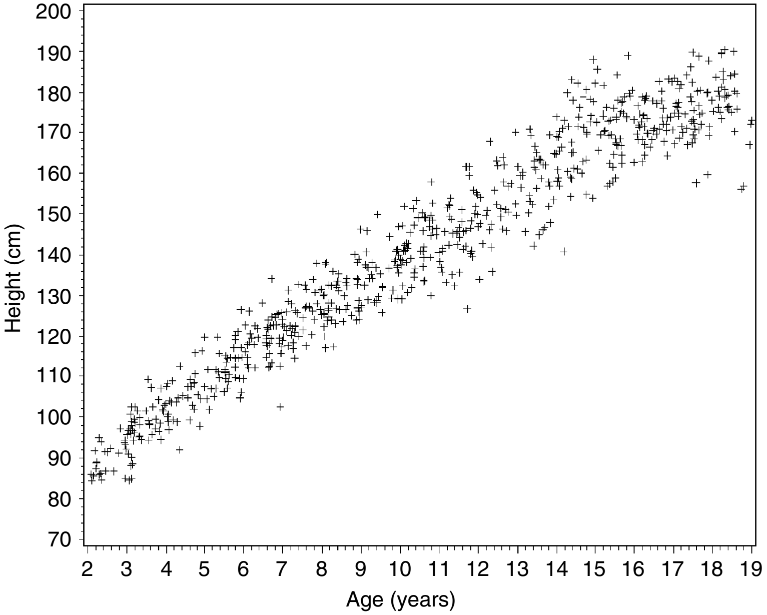

Fig. 7. Sampled scatterplot of data plotted in Fig. 6. Points were chosen for plotting with probability proportional to their sample weights.

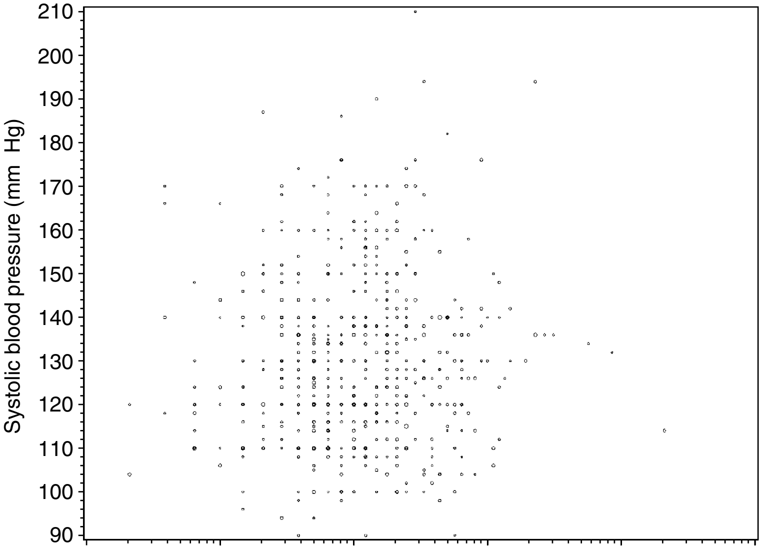

Fig. 8. Bubble plot based on data from white males aged 40–59 years sampled in the second National Health and Nutrition Examination Survey; areas of circles are proportional to the sample weights. There are many overlapping circles in this plot.

males aged 40–59 years. The data are from the second National Health and Nutrition Examination Survey, with the areas of the bubble being proportional to the sample weights (range = 11601–79176, coefficient of variation = 41%). The relationship of blood pressure and lead levels has been previously studied using these data by Pirkle et al. (1985). The lattice pattern of Fig. 8 is because blood pressure was recorded to the nearest mm Hg and blood lead values were recorded to the nearest microgram/deciliter. The overlap of the circles gives a misleading impression of the distribution of values. With this type of "rounding" of the data, a natural solution to the problem of overlapping points is to jitter the data (Chambers et al., 1983, pp. 106–107). In this case, random uniform $(-1/2, +1/2)$ variates are added to the blood pressure and lead values before plotting because it is reasonable to treat the observed values as if they had been rounded to the nearest integer from true values. The jittered plot displayed in Fig. 9 not only avoids the overlap of plotted points but also gives a better representation of the prerounded blood lead levels.

An alternative solution to the overlap problem is to sum the sample weights for points that are plotted at the same location. Figure 10 is the bubble plot using these summed sample weights. This approach has been suggested in the nonsurvey setting, in which "sunflowers" (with the number of lines in the sunflowers equal to the number of data points at the location) are used instead of bubbles (Cleveland and McGill, 1984). Additionally, continuous data can be artificially rounded to apply this approach (Cleveland and McGill, 1984). In the survey setting, this approach is less attractive than jittering because one cannot distinguish in the plot single individuals with large sample weights versus many individuals with small sample weights plotted at the same location.

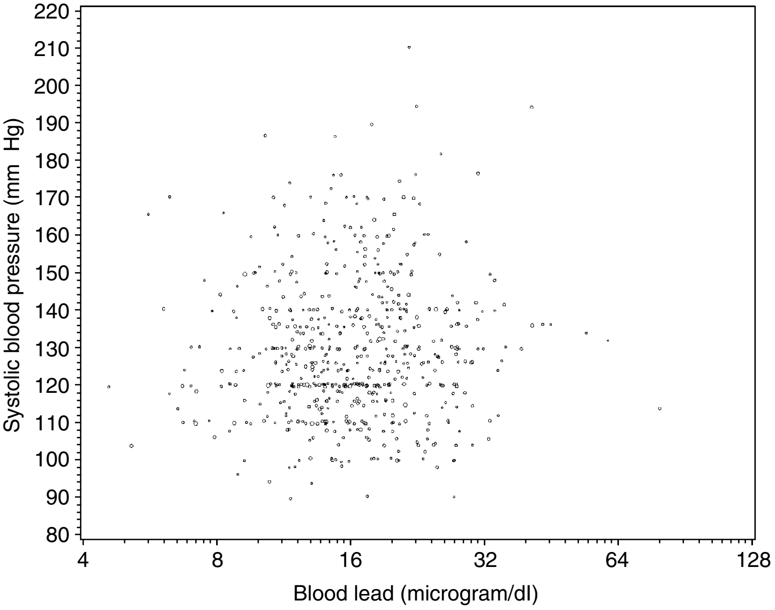

Fig. 9. Jittered bubble plot of data plotted in Fig. 8.

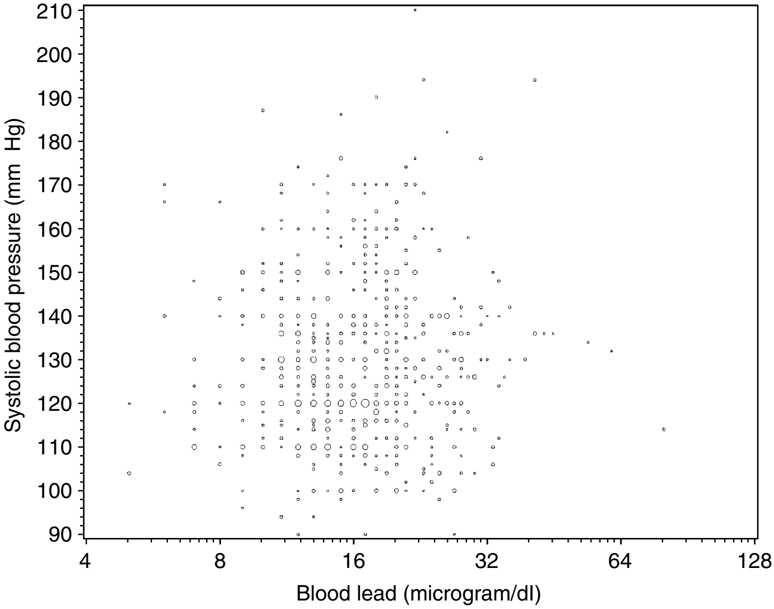

Fig. 10. Summed bubble plot of data plotted in Fig. 8. Areas of circles are proportional to the sum of the sample weights of the individuals with the same data values to be plotted.

2.4. *Accounting for missing data: imputation*

Although missing data can be a problem in any data analysis, survey data are especially susceptible because of the possibility of nonresponse. Data can be missing completely from a sampled individual (unit nonresponse) or partially missing because some questions remain unanswered (item nonresponse). A nonresponse adjustment to the sample weights is frequently used for unit nonresponse; the sample weights are adjusted upwards for respondents with values of other variables similar to those of nonrespondents. The sample weights can be accounted for in a scatterplot as described in Sections 2.1 and 2.2. Item nonresponse is sometimes handled by imputing values for the missing values. There are many ways to do this (Little and Rubin, 2002, Chapters 4 and 5), one of which is described below.

As a preliminary, it can be useful to plot the data without any imputations. Returning to the mother-daughter birthweight data (Fig. 2), the full sample size is 286 of which 225 observations have both mother's and daughter's birthweight nonmissing. Sixty observations are solely missing mother's birthweight and one observation is solely missing daughter's birthweight. Figure 11 displays the sampled scatterplot of Fig. 4, but now also contains (modified) box plots for the estimated distributions of daughter's birthweight for observations not missing, and missing, mother's birthweight (For plotting, the single observation missing daughter's birthweight is ignored.). For these box plots, the edges of the boxes represent the 25th and 75th percentiles, the line in the box represents the median, and the lines extending from the box represent the 10th and 90th percentiles. These percentiles are estimated from using weighted percentiles of the complete samples and not just the (re)sampled observations displayed on the left-hand side of Fig. 11. The

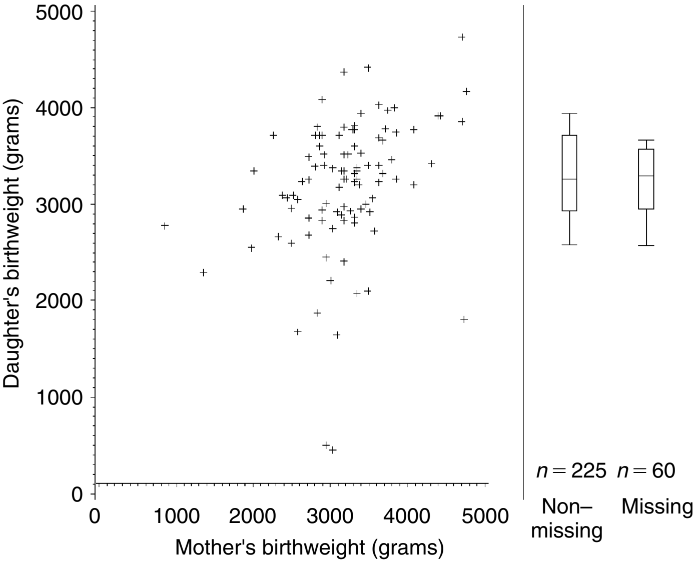

Fig. 11. Sampled scatterplot of nonmissing data with weighted box plots of nonmissing and missing data. Data are from mothers aged 30–39 surveyed in the 1988 National Maternal and Infant Health Survey.

box plots suggest that missingness of mother's birthweight may be less prevalent for high birthweight daughters, but the two-sided p-value for comparing the means is 0.18. An alternative to using the box plots in Fig. 11 would be to display weighted histograms of the distributions. As mentioned above, there are many ways for imputing values for missing data. For graphical displays, it is important that the variability of the imputed values should be consistent with the population variability. We will demonstrate the point with the mother-daughter birthweight data (no imputed values were supplied by the National Center for Health Statistics for mother's birthweight). We use the regression model

$$\begin{aligned}\text{mother's birthweight} = \alpha &+ \beta_{\text{M-HT}} X_{\text{M-HT}} + \beta_{\text{M-RACE}} X_{\text{M-RACE}} \\ &+ \beta_{\text{D-BW}} X_{\text{D-BW}} + \text{ error,}\end{aligned} \qquad (1)$$

where $X_{\text{M-HT}}$ and $X_{\text{M-RACE}}$ denote mother's height and race (1 = nonblack, 2 = black) and $X_{\text{D-BW}}$ denotes the daughter's birthweight. The regression coefficients in model (1) are estimated using (sample-)weighted least-squares for those observations with nonmissing mother's birthweight (the one observation of missing daughter's birthweight was assigned the mean daughter's birthweight). The fitted regression was

$$\begin{aligned}\text{predicted mother's birthweight} = &- 202 + 37.3 X_{\text{M-HT}} + 123 X_{\text{M-RACE}} \\ &+ 0.270 X_{\text{D-BW}}.\end{aligned} \qquad (2)$$

To impute a mother's missing birthweight, we substitute the mother's height and race and her daughter's birthweight into (2) to obtain the predicted mother's birthweight, and then add on an error term obtained as follows. The error terms for the imputed values were obtained by sampling the residuals from the fitted model (2) using probability

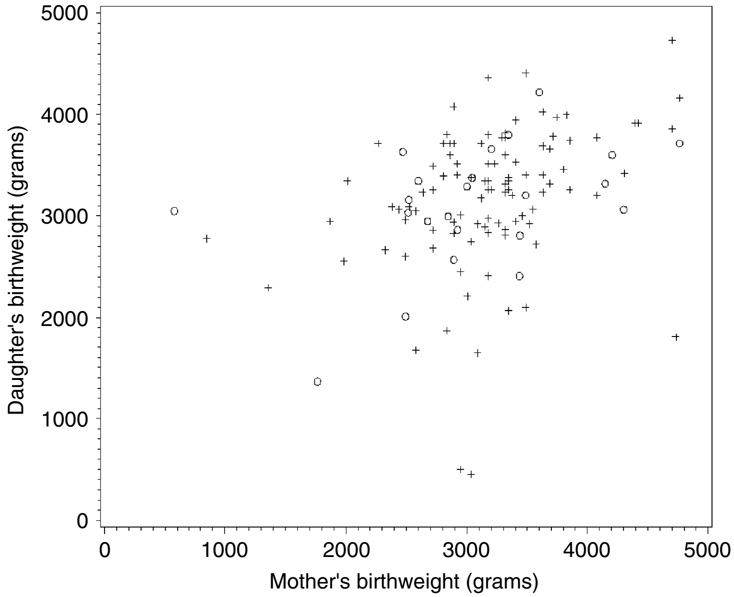

Fig. 12. Sampled scatterplot based on data from mothers aged 30–39 surveyed in the 1988 National Maternal and Infant Health Survey (circles = imputed values, + = nonimputed values).

proportional-to-size sampling, where the inclusion probabilities were proportional to the sampling weights. Figure 12 is a sampled scatterplot of the mother-daughter pairs in which the pairs with imputed mothers' birthweights are designated by o and the nonimputed values by +. If one used for the imputed values the predicted mothers' birthweights from (2) without adding the error term, the sampled scatterplot would be Fig. 13. The spread of the imputed values is misleadingly small in Fig. 13, demonstrating the importance of including an error term in the imputed values.

In both Figs. 12 and 13, the imputed values were highlighted by using a dramatically different symbol in the plots. For many applications, we may want the distinction between imputed and nonimputed values to be visible but not to overpower the display. This can be accomplished by using different symbols that are somewhat similar. For example, one could use x instead of o to denote the imputed values in Fig. 12.

2.5. *Conditional mean and percentile curves: kernel smoothing*

Although one might typically use a polynomial regression to display the X-Y relationship on a scatterplot of a small-to-moderate number of observations, the large number of observations sometimes available with survey data allows for the consideration of less model-dependent approaches. As a simple example, Fig. 14 is a strip box plot (Chambers et al., 1983, pp. 87–91) of height as a function of age for boys aged 2–19 years sampled in the second National Health and Nutrition Examination Survey, see Fig. 7 for a sampled scatterplot of this data. Each box plot displays the sample-weighted 10th, 25th, 50th, 75th, and 90th percentiles of height of those individuals at a particular year of age at the time of examination. The number of observations included for each year of age

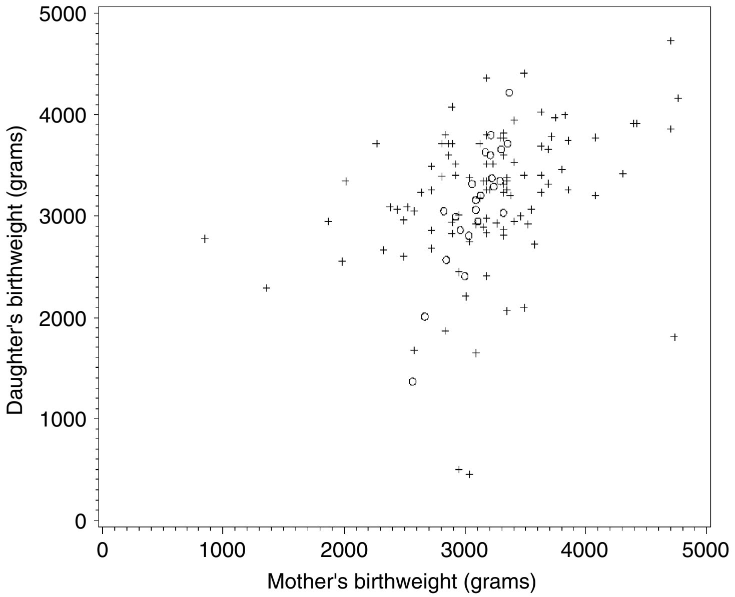

Fig. 13. Sampled scatterplot based on data from mothers aged 30–39 surveyed in the 1988 National Maternal and Infant Health Survey (circles = imputed values without error included, + = nonimputed values).

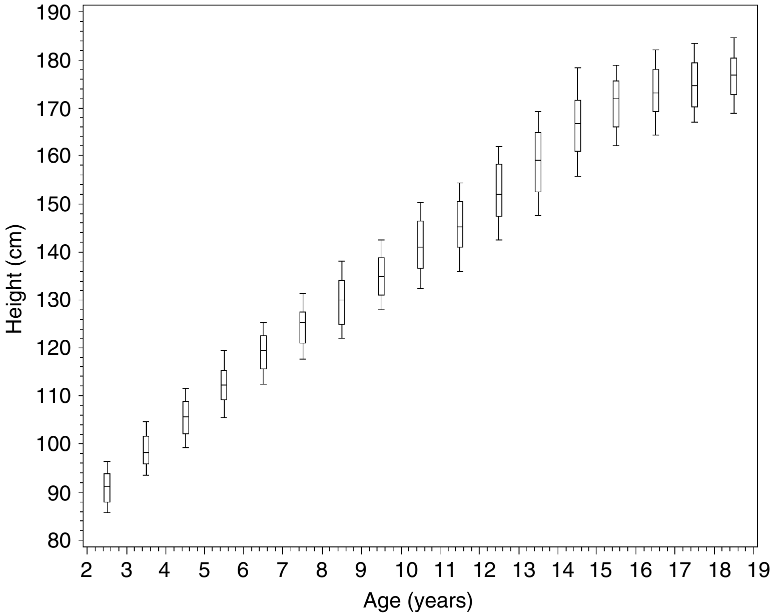

Fig. 14. Strip box plot of height versus age for data plotted in Fig. 6. Box plots show weighted 10th, 25th, 50th, 75th, and 90th percentiles for each year of age.

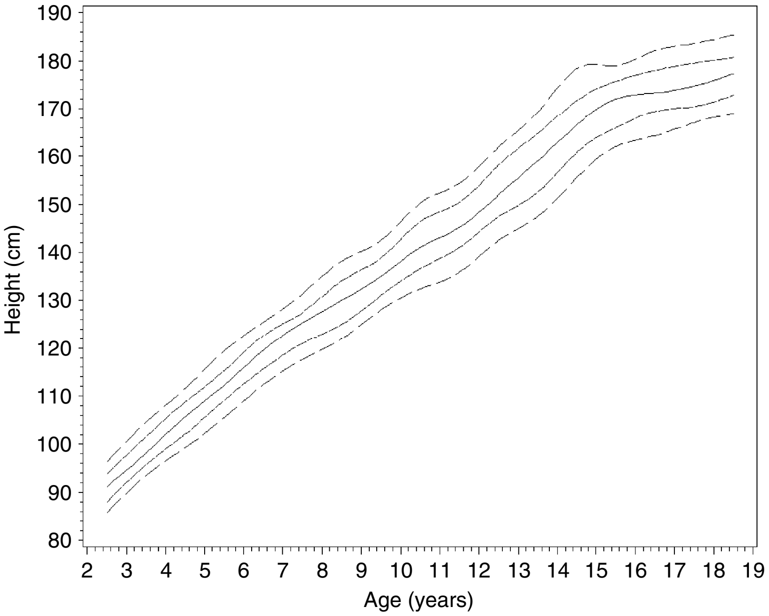

Fig. 15. Cubic spline interpolation of weighted percentiles are shown in Fig. 14. Solid line is the median, dashed lines are the quartiles, and outer dashed lines are the 10th and 90th percentiles.

range from 144 to 429. Figure 14 is not a particularly pleasing display of the percentiles as a function of age. One can remove the boxes and generate smooth curves through the percentiles for the different ages for a better plot. For example, Fig. 15 displays a piecewise cubic spline (SAS, 1990) that fits third-degree polynomials through the percentiles between adjacent years of age; this was the type of approach used in an early presentation of growth curves by the National Center for Health Statistics (1976). Guo et al. (1990) discuss alternative methods for smoothing percentiles for this type of grouped data.

More direct approaches to estimating smooth conditional percentile or mean curves are possible using the original ungrouped data. There are many different ways to do this (Härdle, 1990); we briefly describe a kernel method. Let $\{(x_i, y_i, w_i) \mid i = 1, \ldots, n\}$ be the sampled (X, Y) data with their corresponding sample weights. The idea behind a kernel estimator of the conditional mean of Y given $X = x$ is to evaluate the weighted mean of the y_i whose corresponding x_i are near x. The weights used for this weighted mean incorporate the sample weights and can also weight points with x_i close to x more than points x_i further from x by the choice of a "kernel function." We describe in the next section how to incorporate the sample weights into a particular kernel smoother. The end result is that one can express an estimator of the conditional mean as

$$\text{mean}(y|x) = \sum_{i=1}^{n} w_i^{\text{LS}} y_i, \tag{3}$$

where the weights w_i^{LS} incorporate the sample weights as well as the choice of the kernel function, local regression smoothing, and bandwidth. Figure 16 is a replot of the

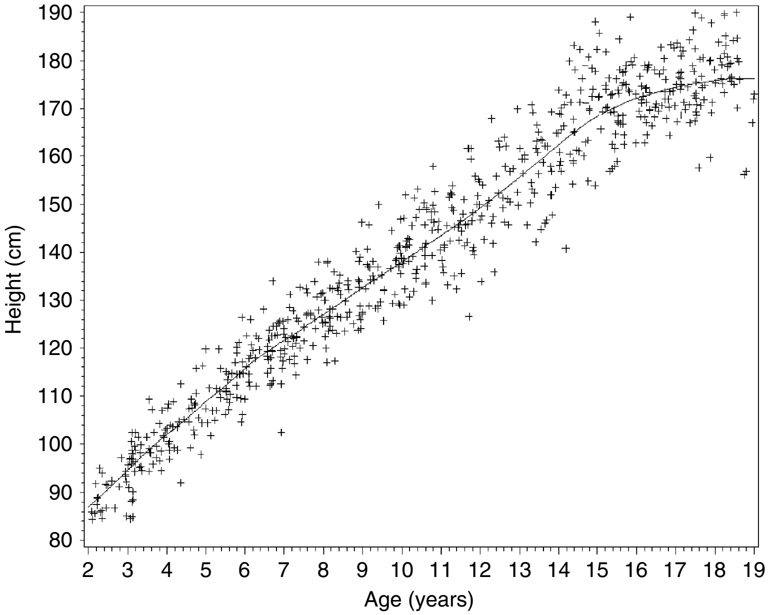

Fig. 16. Replot of Fig. 7 with the local-linear kernel estimator of the conditional mean using a triangular kernel with a bandwidth determined by a one-sided sample size of 350.

sampled scatterplot of Fig. 7 with the local-linear kernel estimator of the conditional mean using a triangular kernel with a bandwidth determined by a one-sided sample size of 350 (see the next section) (The conditional mean estimator uses the full sample of size 3667 and not just the points plotted in Fig. 7.).

2.5.1. *Details of kernel smoothing*

Let the kernel function $K(u)$ be a non-negative symmetric function that integrates to one, for example, the triangular kernel $K(u) = 1 - |u|$ for $|u| \leq 1$ and 0 otherwise. In the nonsurvey setting, one possible kernel estimator of the conditional mean is given by

$$\text{mean}^K(y|x) = \sum_{i=1}^{n} w_i^K y_i, \tag{4}$$

where $w_i^K = K\left(\dfrac{x - x_i}{h_x}\right) \Big/ \sum_{j=1}^{n} K\left(\dfrac{x - x_j}{h_x}\right)$ and h_x is the bandwidth that essentially determines how far the x_i can be from x and still be included in the estimator $\text{mean}^K(y|x)$. A potential problem with the curve $\text{mean}^K(y|x)$ is at the boundaries of the X data. One way to avoid this problem is to use a locally weighted regression (Cleveland, 1979), with a local-linear smoother being a special case: instead of using the weighted mean (4), one fits a weighted linear regression to the data around x using the w_i^K weights. Then, one defines $\text{mean}^L(y \mid x)$ to be the predicted value of Y at $X = x$ from this regression. The estimator $\text{mean}^L(y \mid x)$ can still be defined as a weighted mean, namely,

$$\text{mean}^L(y|x) = \sum_{i=1}^{n} w_i^L y_i \tag{5}$$

with weights equal to

$$w_i^L = w_i^K \left[1 + \frac{\left(x_i - \overline{x}^K\right)\left(x - \overline{x}^K\right)}{\sum_{j=1}^{n} w_j^K \left(x_j - \overline{x}^K\right)^2} \right],$$

where $\overline{x}^K = \sum_{j=1}^{n} w_j^K x_j$. The additional possibility of downweighting points with large residuals ("lowess," Cleveland, 1979) is not pursued here.

In the survey setting, to account for the sample weights (w_i), one let

$$w_i^{\text{KS}} = w_i K\left(\frac{x - x_i}{h_x}\right) \Big/ \sum\nolimits_{j=1}^{n} w_j K\left(\frac{x - x_j}{h_x}\right)$$

and

$$w_i^{\text{LS}} = w_i^{\text{KS}} \left[1 + \frac{\left(x_i - \overline{x}^{\text{KS}}\right)\left(x - \overline{x}^{\text{KS}}\right)}{\sum_{j=1}^{n} w_j^{\text{KS}} \left(x_j - \overline{x}^{\text{KS}}\right)^2} \right],$$

where $\overline{x}^{\text{KS}} = \sum_{j=1}^{n} w_j^{\text{KS}} x_j$. The local-linear smoother is then defined by (3). The use of the sample weights implies that (3) is estimating what (5) would be estimating if all the population values were available and used for the estimation.

Bellhouse and Stafford (2001, 2003) have proposed using local-polynomial regression to estimate a smooth conditional mean curve and have given asymptotic bias and variance properties of their estimators. This approach is a generalization to the local-linear smoother described above, where a weighted polynomial regression is used instead of the weighted simple linear regression.

The choice of the bandwidth is critical in determining how smooth the resulting conditional mean curve will be. There are various ways to choose the bandwidth (Härdle, 1990, Chapter 5; Ruppert et al., 1995). We describe two simple approaches here: one approach is to fix h_x to be a constant that is meaningful to the scale of the data at hand and a second approach is to choose h_x so that a certain minimum sample size is contained in $x \pm h_x$, for example, 100 observations. A modification of this second approach, which we prefer, is to choose h_x so that a certain minimum sample size is contained in either $[x, x - h_x]$ or $[x, x + h_x]$, for example, 50 observations. Without this modification, h_x will tend to increase as x approaches a boundary of the data.

A benefit of the development of the conditional mean estimator (3) as a weighted mean of the $y_i s$ is that the approach extends naturally to other functionals of the conditional distribution of Y given X, for example, percentiles. This was suggested by Stone (1977) and studied extensively by Owen (1987). The idea is to estimate the cumulative distribution function (CDF) for Y using the y_i, whose x_i are near x. In the present context, to estimate the conditional percentiles, one can use for each x the (weighted) percentile estimated from the weighted empirical CDF of the y_i using the w_i^{LS} weights. Unfortunately, this approach has a serious drawback for quantiles other than the median: even if the relationship of the quantiles and x was linear (but not horizontal), the larger the bandwidth, the more the estimated quantiles will be biased away from the median. This is because the changing values of the conditional percentiles as a function of x causes the spread of y values to be larger when a larger bandwidth is considered.

To avoid this bias in the estimated conditional percentiles other than the median, we modify the approach analogously to that used for estimating "upper and lower smoothings" based on conditional means (Cleveland and McGill, 1984). We first estimate the conditional median using the weighted CDF as described above, denote it by $q_{50}(y|x)$ and let $z_i = y_i - q_{50}(y|x_i)$. To estimate a conditional percentile greater than the median, say the 90th percentile, use the weighted CDF approach to estimate the conditional 80th percentile of the zs given x using only the data points for which $z_i > 0$. If we denote this conditional 80th percentile by $q_{80}(z|x)$, then the desired conditional 90th percentile is estimated by $q_{50}(y|x) + q_{80}(z|x)$. In general, one estimates the conditional ηth percentile for $\eta > 50$ by $q_{50}(y|x) + q_{\gamma}(z|x)$, where $\gamma = 2\eta - 100$. This modification works for conditional percentiles less than the median in the obvious fashion. Figure 17 displays selected conditional percentiles for the height/age data using a local-linear kernel estimator using a triangular kernel with a bandwidth determined by a one-sided sample size of 350.

With large data sets, the discreteness of the scale of the measurement of Y can sometimes become noticeable in the conditional percentile curves. For example, consider the blood lead data described in Section 2.3. A plot of the smoothed conditional percentiles of blood lead versus age will take on only integer values since blood lead is recorded

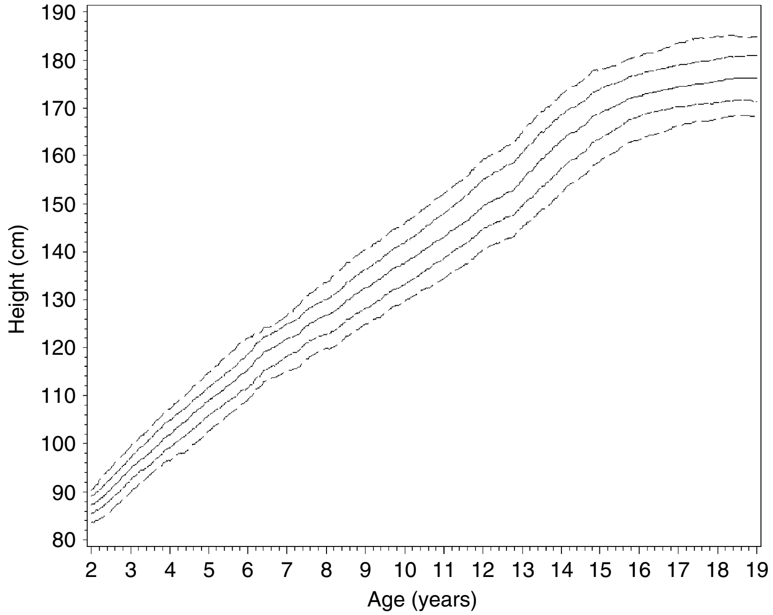

Fig. 17. Weighted conditional percentiles of height as a function of age of data plotted in Fig. 6. Solid line is the median, dashed lines are the quartiles, and outer dashed lines are the 10th and 90th percentiles. Conditional percentiles are estimated using a local-linear kernel estimator using a triangular kernel with a bandwidth determined by a one-sided sample size of 350 (see text).

to the nearest integer (plot not shown). If this is a problem, the weighted empirical CDF calculated at each x can itself be smoothed before estimating the percentiles, see Woodruff (1952) and Korn et al. (1997) for some simple methods of doing this.

The last issue we address is the calculation of standard errors for kernel estimators. A simple approach is to use one of the replication methods of variance estimation (Korn and Graubard, 1999, pp. 29–35). For example, with a balanced half-sample replication, the kernel estimator would be calculated using the data from each half-sample of primary sampling units. The bandwidth used for the replicated kernel estimators should be approximately the same as the bandwidth used for the kernel estimator for the original data (at each x). In particular, if a variable-length bandwidth involving a minimum sample size was used for the original data, you should *not* use a variable-length bandwidth involving the same minimum sample size for the replicates for a jackknife or a balanced half-sample replication. Instead, for example, you should, for a balanced half-sample replication, use a variable-length bandwidth involving half the minimum sample size used for the estimator on the original data or fix the bandwidth for the replicates at the value used for the original data. The rationale for this is that balanced half-sample replication or the jackknife is derived assuming that fewer observations are used in the replicate estimators. For example, balanced half-sample replication yields a reasonable variance estimator because the variability of the half-sample estimators is about twice that of the full-sample estimator. It should be noted that jackknife variance estimators should not be used for conditional percentile curves because they are not differentiable functions of the data. However, jackknife estimators can be used for conditional mean curves.

We caution the reader when using standard errors for kernel estimators. Although they can be interpreted as representing the variability one would see in the estimators if they were calculated from repeated independent surveys of the population, they cannot automatically be used to derive confidence intervals. This is because the smoothed estimators are biased (This bias is hard to quantify because it depends on the amount of smoothing and the curvature of the true curves.). This problem can become especially noticeable when a variable-width bandwidth is used and the data are scarce in a region of the horizontal axis. The bandwidth will be large in this region to capture a sufficient sample size. Therefore, the replicated standard errors of the smoothed curve will be no larger than at other regions of the curve where the data density is higher, presenting a potentially misleading picture. With cautious interpretation, however, we still believe that the presentation of the standard errors of kernel estimator is worthwhile. For example, if they are large, then the kernel estimators are not useful no matter what the size of the bias. If the sample size is so large that the bandwidth is quite small, then the bias of the smoothed estimators will be small.

As an alternative to presenting standard errors, we present a different method for examining whether a smoothed conditional mean or percentile curves is reflecting a property of the underlying distributions rather than just noise. As an example, Fig. 18 is a partial residual plot for the logarithm of blood lead from a (sample-) weighted multiple linear regression of systolic blood pressure on log lead, age, and body mass index using the data described previously. Partial residual plots for a particular independent variable x_1, also known as component-plus-residual plots, are plots of $r_i + \hat{\beta}_1 x_{i1}$ versus x_1, where $\hat{\beta}_1$ and the residuals r_i are estimated from the multiple linear regression model (Atkinson, 1985, Chapter 5.4; Cook and Weisberg, 1994, Chapter 9). These plots are useful for examining possible needed transformations of the independent variable. The

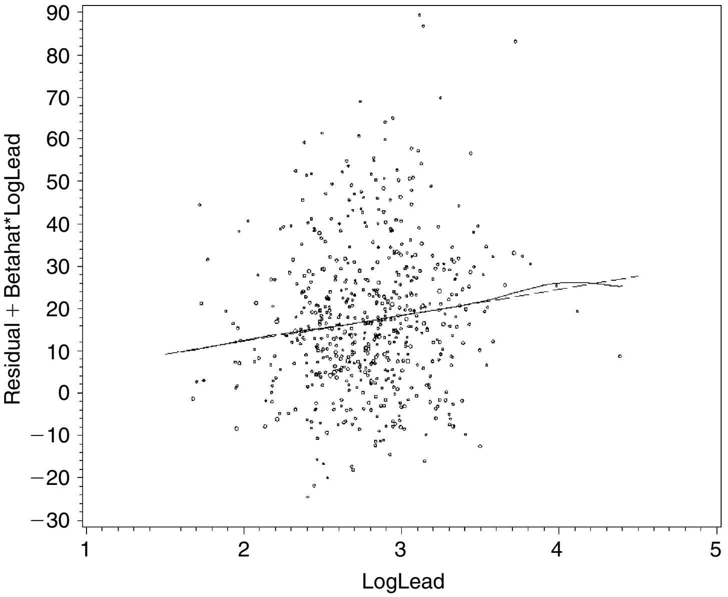

Fig. 18. Partial residual plot of the logarithm of blood lead (log lead) from a weighted regression of systolic blood pressure on log lead, age, and body mass index using data from 595 white males aged 40–59 sampled in the second National Health and Nutrition Examination Survey. Areas of circles are proportional to the sample weights. Dashed line is the weighted least-squares line. Solid line is the local-linear kernel estimator of the conditional mean using a triangular kernel with a fixed bandwidth of ±1.5 units of log lead.

dashed line in Fig. 18 is the weighted least-squares line; its slope is identical to the estimated regression coefficient of log lead in the weighted multiple linear regression. Analogous partial residual plots can also be constructed for other types of regression analyses of complex survey data, including logistic regression and proportional hazard regression (Korn and Graubard, 1999, pp. 111–113, 124–126).

The smooth curve in Fig. 18 is a local-linear kernel estimator of the conditional mean using a triangular kernel with the fixed bandwidth of ±1.5 units of log lead. The curve shows no great nonlinearity although there is the suggestion of a rise and then fall of the curve for log lead values greater than 3.5. As an ad hoc check of the reality of this nonlinearity, we simulated five data sets in which the linear regression model holds exactly—the values of the independent variables and the sample weights were taken as in the observed data set, and Y values were simulated with normal distributions around the predicted values (with standard deviation equal to the residual standard deviation from the observed data set computed without regard to possible sample weighting and correlation from cluster sampling). There should be no structure in the residuals from the weighted linear regressions using these simulated data sets. The top five curves in Fig. 19 are the estimated conditional mean plots from the partial residuals from these five simulated data sets; the bottom curve is a replot of the conditional mean curve from Fig. 18. The structure seen in these curves is at least as great as that seen in the curve calculated from the actual data, suggesting that the structure seen in the curve based on the actual data can be safely ignored.

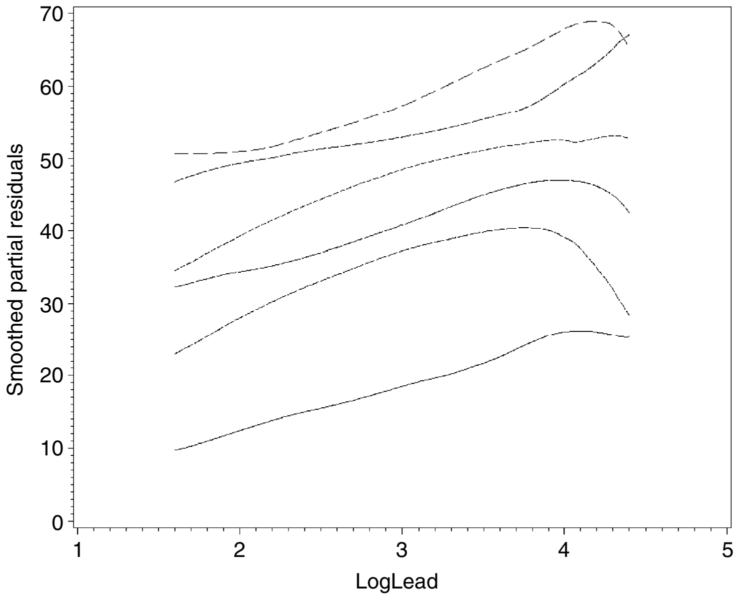

Fig. 19. Replot of the kernel estimator of the conditional mean from Fig. 18 (solid curve) with kernel estimators of the conditional mean based on five simulated data sets for which the conditional mean should be linear (dashed and dotted curves, translated in the vertical direction to avoid overlapping curves).

Although this section has focused on using kernel smoothing to obtain conditional means or quantiles in scatterplots with survey data, we note that there has been work in kernel smooth methods for density estimation with survey data (Bellhouse and Stafford, 1999, 2003; Buskirk and Lohr, 2005). The density estimation work may help to frame the theoretical basis for the kernel smoothing methods that we have presented. Finally, in addition to Härdle (1990) other references on topic of kernel smoothing for simple random samples are Wand and Jones (1995), Eubank (1999), and Simonoff (1996).

2.6. *Regression splines*

Regression splines is an alternative approach to kernel smoothing that use modeling when investigating the functional relationship between an outcome Y and a variable X. In the context of linear regression, the question is whether the simple inclusion of X as an independent variable is sufficient to model the relationship. A common modeling approach to this problem is to include powers of X as additional independent variables to allow for a polynomial relationship. For example, inclusion of X and X^2 in the model allows for Y to be a quadratic function of X and inclusion of X, X^2, and X^3 allows for Y to be a cubic function of X, etc. Frequently, however, this approach does not work because to adequately fit the data over the whole range of X may require a high-degree polynomial. This would lead to a nonparsimonious model with many independent variables involving X.

An alternative approach is to use a regression spline, which involves a fixed set of "knots" $t_1 < t_2 < \cdots < t_K$ in the range of X. The spline function of X is a piecewise polynomial that is smoothly joined at the knots. Spline functions can be fit to the data by adding a small number of independent variables to the (linear) regression model. A very convenient type of spline (and the one we will discuss) is called "restricted cubic regression splines" (Stone and Koo, 1985), which are defined as follows: in between each pair of adjacent knots, the function is a cubic polynomial, with possibly different cubic functions between the different knot pairs. To the left of t_1 and to the right of t_K, the spline is straight lines (These linear constraints rather than allowing cubic functions in the tails are those that distinguish *restricted* cubic regression splines from ordinary cubic regression splines.). The cubic and linear functions are constrained so that the functional values and their first and second derivatives coincide at each knot; this ensures that the spline is a continuous smooth function.

Typically, a small number of knots (e.g., 3–5) are sufficient to model most data. Durrelman and Simon (1989) use knots at the following percentiles of the X data: $\{5, 50, 95\}$, $\{5, 25, 75, 95\}$, and $\{5, 25, 50, 75, 95\}$ for 3, 4, and 5 knots, respectively. In survey data with sample weights, the knots can be placed at the weighted percentiles.

It is simple to express a restricted cubic regression spline in terms of functions of X that are included as independent variables in a regression. Details of the construction of these independent variables are given in Durrelman and Simon (1989) or Korn and Graubard (1999, Appendix C). With 3 knots, besides X, one need only to include one additional independent variable; with 4 knots, two additional independent variables; etc (For details about other approaches to estimating splines in simple random samples see Eubank, 1999.)

In most applications, there will be other independent variables in the model in addition to X and the spline variables that are functions of X. The interpretation of the spline function in these situations is the usual conditional one for a regression. A nice feature of using regression splines is that one can easily test whether a linear relationship is adequate by testing whether the estimated regression coefficients of the spline variables are significantly different from zero. For the restricted cubic regression spline with 3 knots, this involves testing if one regression coefficient equals zero; with 4 knots, a simultaneous test of whether two regression coefficients equal zero is used; etc. For survey data, these tests can be performed by estimating the coefficients using a sample-weighted regression and using a Wald statistic that incorporates the survey design. Standard linear regression software for survey data can be used for the analysis (although the analyst may be required to generate the spline variables).

When a linear relationship is inadequate to model the data, plotting the regression spline can suggest alternative nonlinear relationships. With no other independent variables in the model, a plot of the predicted values versus X can be overlaid on a scatterplot of the data. When there are other independent variables in the model $(Z_1, \ldots, Z_p)$, the following procedure can be applied. For example, for multiple linear regression modeling, suppose there are K knots with the corresponding $K - 2$ spline variables in X being $S_{K,1}, S_{K,2}, \ldots, S_{K,K-2}$. Plot

$$\hat{\alpha} + \hat{\beta}_X X + \hat{\gamma}_{K,1} S_{K,1} + \cdots + \hat{\gamma}_{K,K-2} S_{K,K-2} + \hat{\beta}_{Z_1} c_1 + \cdots + \hat{\beta}_{Z_p} c_p$$

versus X, where $\{\hat{\alpha}, \hat{\beta}_X, \hat{\gamma}_{K,1}, \ldots, \hat{\gamma}_{K,K-2}, \hat{\beta}_{Z_1}, \ldots, \hat{\beta}_{Z_p}\}$ are the estimated regression coefficients from the model including the spline variables, and $c_1, \ldots, c_p$ are a set of constants representing possible values of $Z_1, \ldots, Z_p$. This plot is interpreted as the predicted value for an individual with covariate values $Z_1 = c_1, \ldots, Z_p = c_p$ and $X = x$, as a function of x. A confidence band for this plot can be calculated using the estimated covariance matrix of the estimated regression coefficients (The width of the confidence band will depend on the particular values $c_1, \ldots, c_p$ chosen.). This plot can be overlaid with a plot of $\tilde{\alpha} + \tilde{\beta}_X X + \tilde{\beta}_{Z_1} c_1 + \cdots + \tilde{\beta}_{Z_p} c_p$ versus X (which is a straight line), where $\{\tilde{\alpha}, \tilde{\beta}_X, \tilde{\beta}_{Z_1}, \cdots, \tilde{\beta}_{Z_p}\}$ are the estimated coefficients from the linear regression model without the spline variables. The resulting plot allows for a comparison of the linear and spline-modeled associations of Y and X.

3. Discussion

In the nonsurvey setting, the simple scatterplot is an excellent overall graphical display of bivariate data. In the survey setting, different purposes may be best suited by different plots. For example, is the plot to describe the sample for data cleaning purposes or is to describe the population for population inference? With large sample sizes, is the plot to describe general trends or is to identify possible outliers and influential points? We have given examples in this chapter of some modifications of the simple scatterplot that we have found useful for displaying survey data. Other modifications are possible, and may be advisable, depending on the survey and the purpose of the display.