Scatterplots With Survey Data

Edward L. KORN and Barry I. GRAUBARD

We suggest various modifications to make scatterplots more informative when used with data obtained from a sample survey. Aspects of survey data leading to the plot modifications include the sample weights associated with the observations, imputed data for item nonresponse, and large sample sizes. Examples are given using data from the 1988 National Maternal Infant and Health Survey, the second National Health and Nutrition Examination Survey, and the epidemiologic follow-up of the first National Health and Nutrition Examination Survey.

KEY WORDS: Added variable plot; Conditional percentile; Graphical methods; Imputation; Influential points; Kernel smoothing; Nonparametric regression; Partial residual plot; Sample weights; Survey methods.

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 differing 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. As will be shown in the following, scatterplots 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 article 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.

Edward L. Korn is Head, Clinical Trials Section, Biometric Research Branch, National Cancer Institute, Bethesda, MD 20892 (E-mail: korne@ctep.nci.nih.gov). Barry I. Graubard is Acting Head, Biostatistical Methodology and Cancer Control Epidemiology Section, Biometry Branch, National Cancer Institute, Bethesda, MD 20892. The authors thank Douglas Midthune for his help with the computer programming and a referee for helpful comments.

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 represent 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. The last modification to the scatterplot we consider is using conditional mean and percentile curves constructed using kernel smoothing for displaying the relationship between Y and X when the sample sizes are large. Examples of this are given in section 2.5.

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, Placek, and Keppel

 © 1998 American Statistical Association

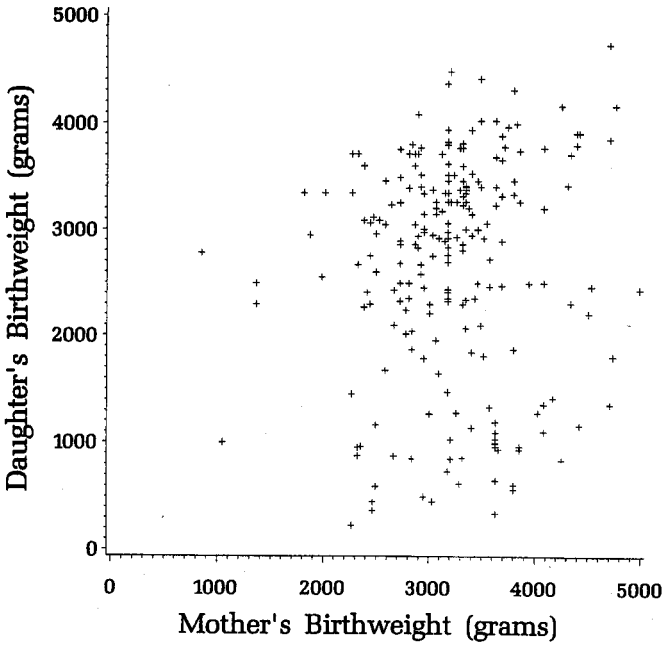

Figure 1. Simple Scatterplot Based on Data from Mothers aged 30–39 Surveyed in the 1988 National Maternal and Infant Health Survey.

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, Zuckerman, Coffman, and Corwin 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 (Tab. 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 Figure 2; the areas of the circles are proportional to the sample weights.

Another reason to use 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 association was also studied by Korn and Graubard (1995) and others (e.g., Stevens, Jones, Micozzi, and Taylor 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 5,073 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.

Table 1. Sampling Strata and Sampling Rates of 1988 National Maternal and Infant Health Survey

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

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, ch 12.1; Atkinson 1985, ch 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) suggested 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

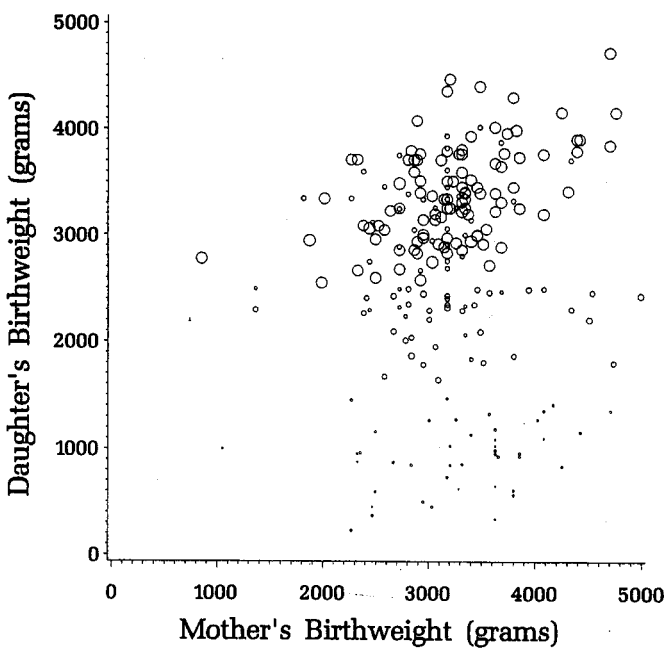

Figure 2. Bubble Plot of Data Plotted in Figure 1; Areas of Circles are Proportional to the Sample Weights.

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[1], dropping certain data points

Point[2] dropped from the analysis:	Sample size	$\beta \pm SE$[3]
None	5270	.025 ± .014
Point A	5269	.009 ± .009
Point B	5269	.024 ± .014
Point C	5269	.028 ± .014

[1] 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 ($<$\$3,000, \$3,000–6,999, \$7,000–9,999, \$10,000–14,999, and $\geq$ \$15,000).
[2] Points are designated in Figure 3.
[3] To account for the complex sampling design, the computer program SUDAAN (Shah, Barnwell, and Bieler 1995) was used to calculate the standard errors.

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 because 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.

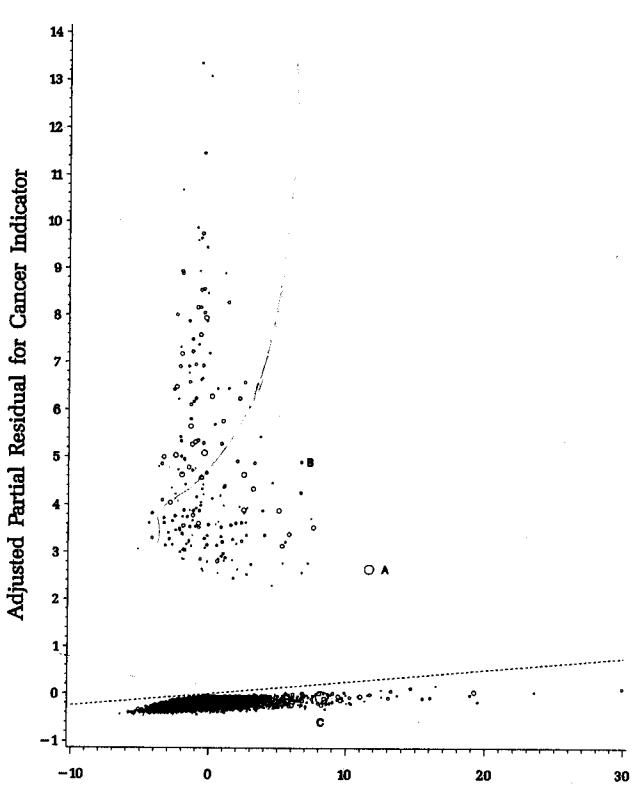

Figure 3. Added Variable Plot for Transferrin Saturation Based on Weighted Multiple Logistic Regression Described in Table 2. Dashed line is weighted least-squares line; labeled points are described in the text.

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 Figure 3 is the weighted least-squares line; its slope is .025, the same as the logistic regression coefficient for transferrin saturation (Tab. 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 (Tab. 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 Figure 3 in identifying influential points. For example, without the circles, the point labeled B might appear about as influential as point A. However, because of its small sample weight, it has very little influence on the coefficient (Tab. 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 Figure 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 (Tab. 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 Figure 2. The ith observation from the original data set was included in Figure 4 if a uniform (pseudo-) random number was less than $w_i/w_{\max}$, where w_i is the sample weight of the ith observation and $w_{\max}(= 1008.515)$ is the largest sample weight of the 225 observations in Figure 2. In general, one samples the ith 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. 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, Oh, and Scheuren (1994) in order to use conventional nonsurvey methods of analysis for survey data.

There is no question that there is a loss of information in going from Figure 2 to Figure 4. Therefore, Figure 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 Figure 4 to Figure 2, and this preference would become stronger if the sample

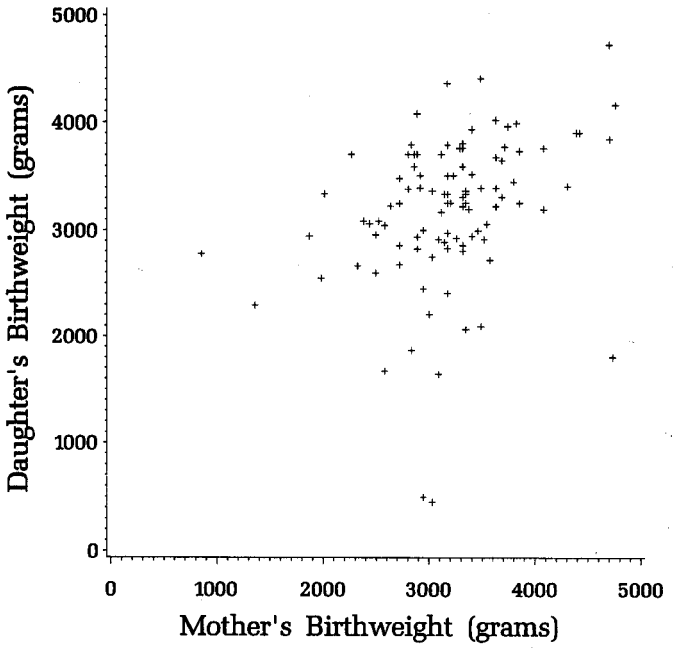

Figure 4. Sampled Scatterplot of Data Plotted in Figure 2. Points were chosen for plotting with probability proportional to their sample weights.

size were larger; see the height/age example given in the following.

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 Figure 4 correspond to black daughters; this is reflective of the population. Because black babies were oversampled in the survey, there is a lot 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 Figure 5 is not representative of the population, it is representative of the black and nonblack populations separately. It appears from Figure 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: .32 ($n = 170$) versus .07 ($n = 55$), respectively.

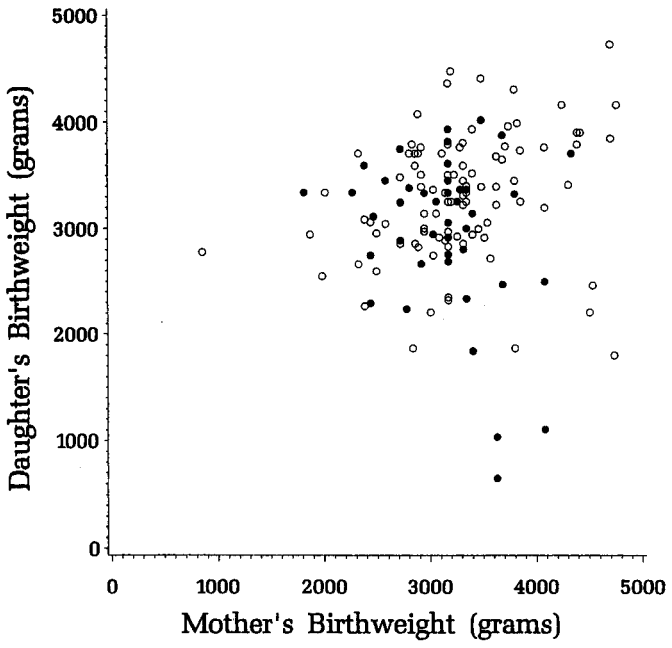

Figure 5. Sampled Scatterplot of Data Plotted in Figure 2. Black daughters (filled-in circles) were sampled for plotting at approximately six times the rate as nonblack daughters (open circles).

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

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 3,667 boys aged 2 to 19 years sampled in the second National Health and Nutrition Examination Survey. The sample weights for these boys ranged

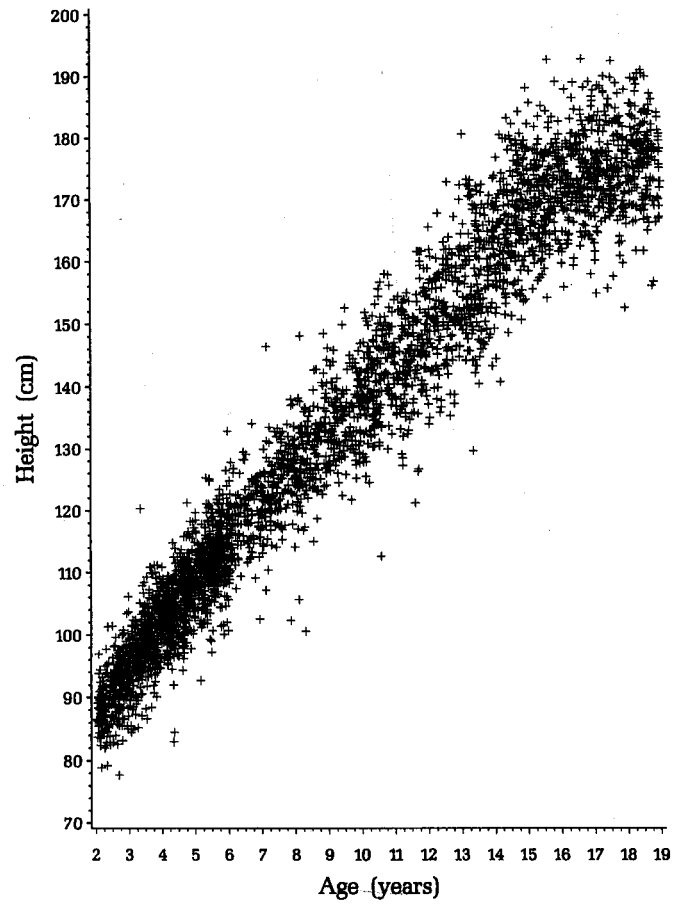

Figure 6. Simple Scatterplot of Height Versus Age for Boys Aged Less Than 19 Years Sampled in the Second National Health and Nutrition Examination Survey.

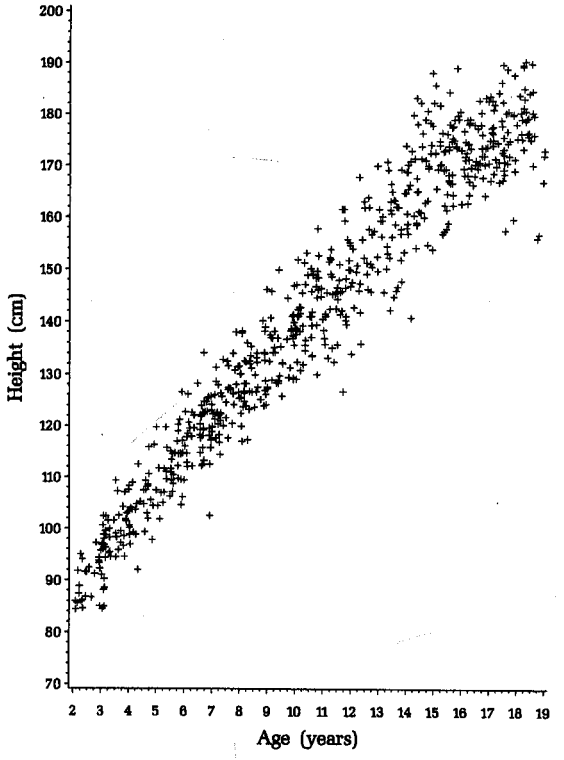

Figure 7. Sampled Scatterplot of Data Plotted in Figure 6. Points were chosen for plotting with probability proportional to their sample weights.

from 1,359 to 47,385, with a coefficient of variation of 71%; see McDowell, Engel, Massey, and Maurer (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 Figure 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 Figure 6 would not be useful. We can solve the two problems of excessive density and representativeness at once by using a sampled scatterplot; see Figure 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) so that their plotted points are 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, Figure 8 is a bubble plot of systolic blood pressure versus the logarithm of blood lead values for 595 white 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 to 79176, coefficient of variation = 41%). The relationship of blood pressure and lead levels has been previously studied using these data by Pirkle, Schwartz, Landis, and Harlan (1985). The lattice pattern of Figure 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

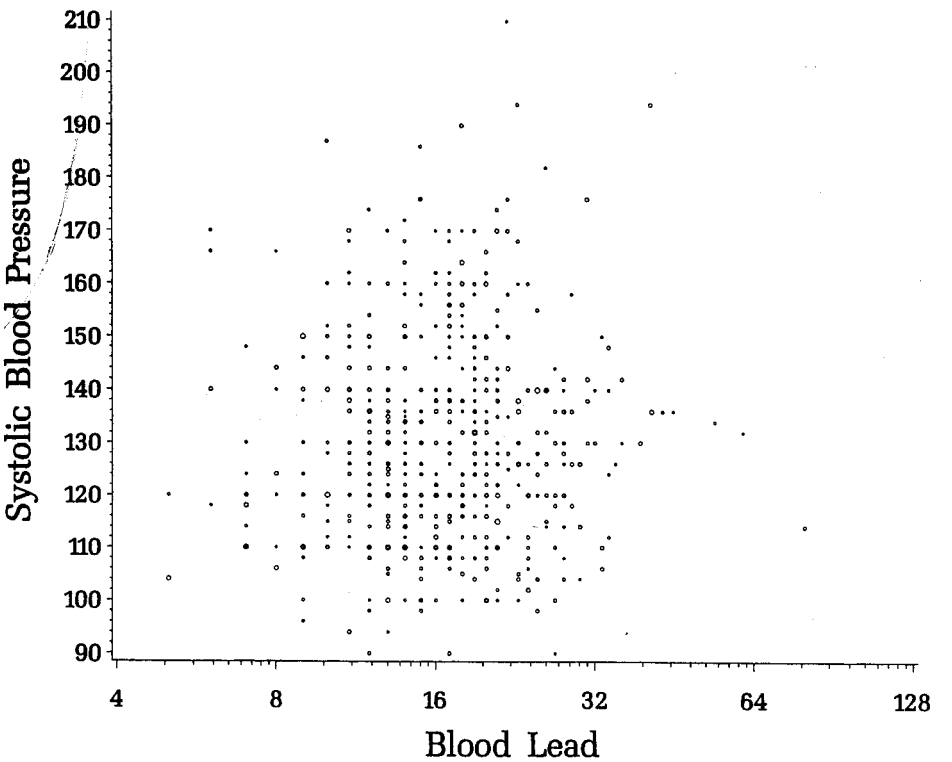

Figure 8. Bubble Plot Based on Data From White Males Aged 40–59 Years Sampled in the Second National Health and Nutrition Examination Survey; Areas of Circles are Proportional to the Sample Weights. There are many overlapping circles in this plot.

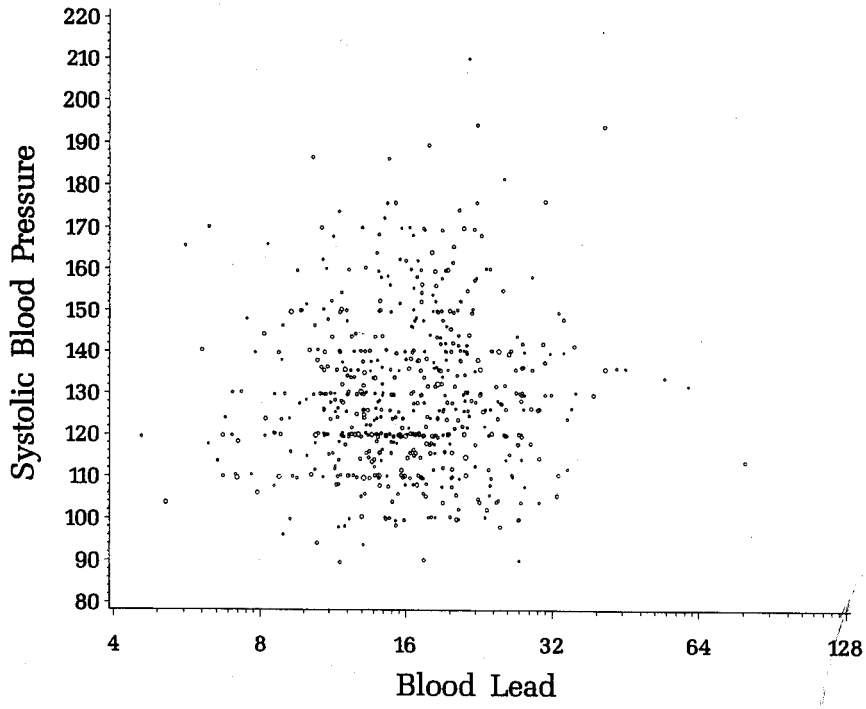

Figure 9. Jittered Bubble Plot of Data Plotted in Figure 8.

(Chambers, Cleveland, Kleiner, and Tukey 1983, pp 107–107): 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 assume that the observed values had been rounded to the nearest integer from the true values. The jittered plot displayed in Figure 9 not only avoids the overlap of plotted points, but also gives a better representation of the pre-rounded 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 non-survey 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 ver-

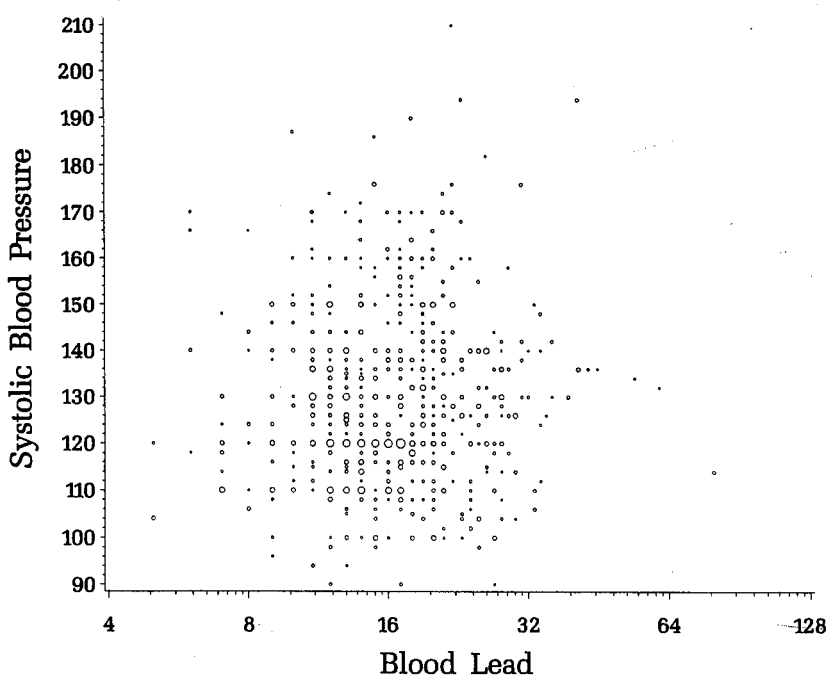

Figure 10. Summed Bubble Plot of Data Plotted in Figure 8. Areas of circles are proportional to the sum of the sample weights of the individuals with the same data values to be plotted.

sus many individuals with small sample weights plotted at the same location.

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 2.1–2.2. Item nonresponse is sometimes handled by imputing values for the missing values. There are many ways to do this (Little and Rubin 1987, ch 4.5), one of which is described in the following.

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 Figure 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 side of Figure 11. The 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 .18. An alternative to using the box plots in Figure 11 would be to display weighted histograms of the distributions.

As mentioned previously, 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 on the National Center for Health Statistics data tapes for mother's birthweight). We use the regression model

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

where X_{M-HT} and $X_{M-\text{RACE}}$ denote mother's height and race (1=nonblack, 2=black), and X_{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 missing daughter's birthweight was assigned the mean daughter's birthweight). The fitted regression was

$$\text{predicted mother's birthweight} = -202 + 37.3X_{M-HT} + 123X_{M-\text{RACE}} + .270X_{D-BW} \quad (2)$$

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-proportional-to-size sampling, where the inclusion proba-

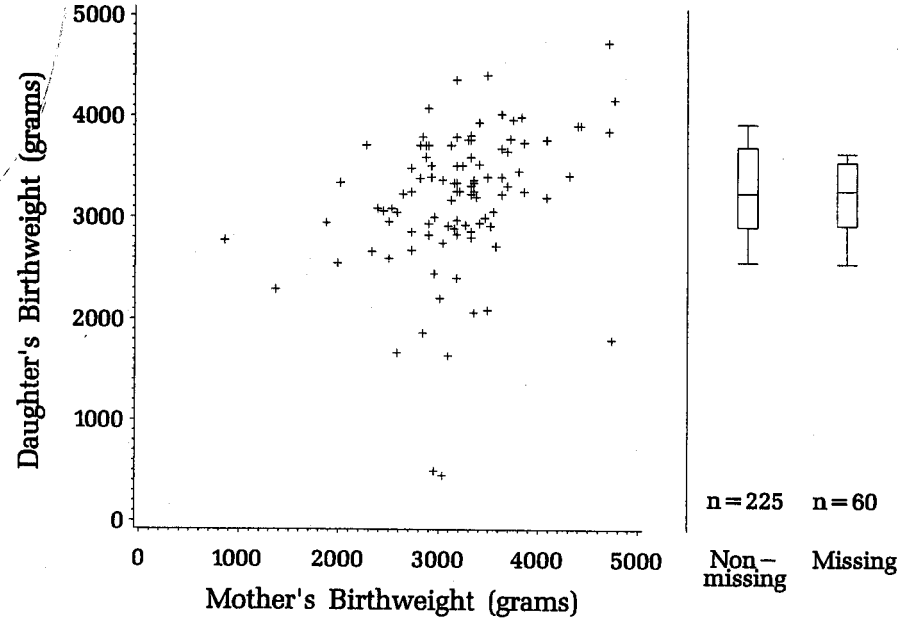

Figure 11. Sampled Scatterplot of Nonmissing Data With Weighted Box Plots of Nonmissing and Missing Data. Data are from mothers aged 30–39 surveyed in the 1988 National Maternal and Infant Health Survey.

bilities 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's and the nonimputed values by +'s. If one used for the imputed values the predicted mothers' birthweights from (2) without adding the error term, the sampled scatterplot would be Figure 13. The spread of the imputed values is misleadingly small in Figure 13, demonstrating the importance of including an error term in the imputed values.

In both Figures 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's instead of o's to denote the imputed values in Figure 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, Figure 14 is a strip box plot (Chambers et al. 1983, pp 87–91) of height as a function of age; see Figure 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 range from 144 to 429. Figure 14 is not a particularly pleasing display

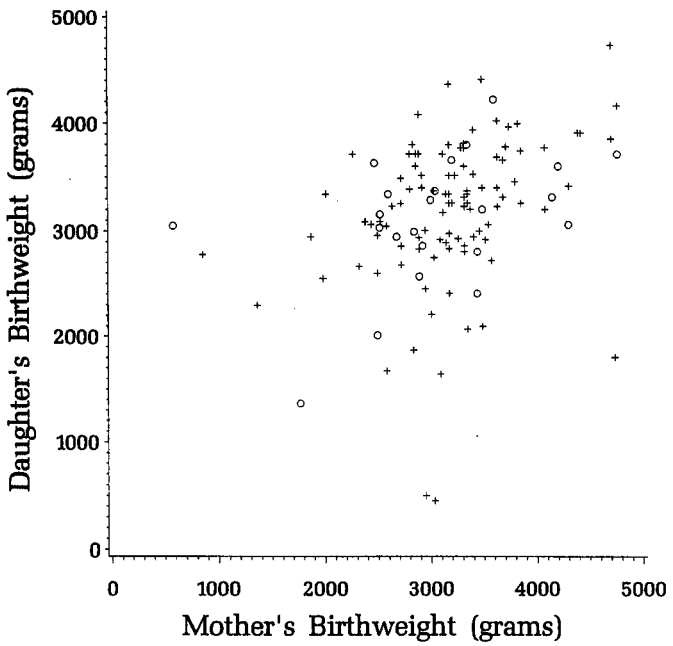

Figure 12. Sampled Scatterplot Based on Data From Mothers Aged 30–39 Surveyed in the 1988 National Maternal and Infant Health Survey (circles = imputed values, +'s = nonimputed values).

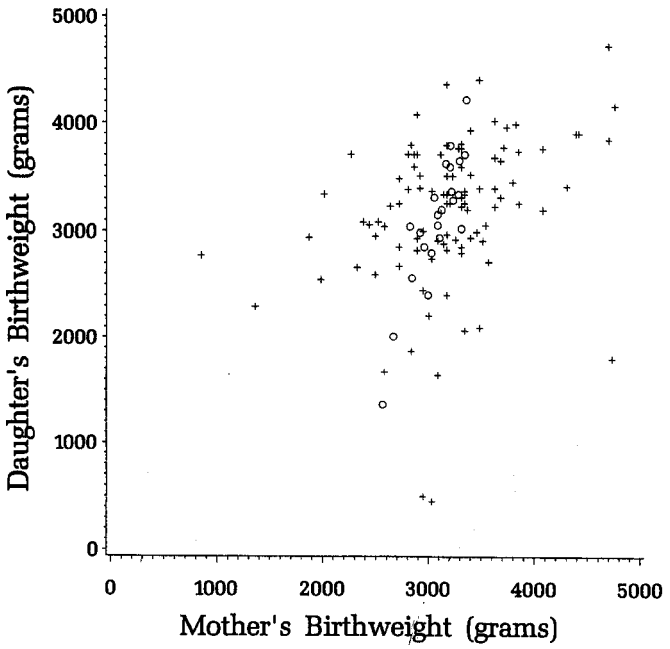

Figure 13. Sampled Scatterplot Based on Data From Mothers Aged 30–39 Surveyed in the 1988 National Maternal and Infant Health Survey (circles = imputed values without error included, +'s = nonimputed values).

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, Figure 15 displays a cubic spline through the percentiles (SAS 1990); this was the approach used in an early presentation of growth curves by the National Center for Health Statistics (1976). Guo et al. (1990) discussed 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 $\{(x_i, y_i, w_i) | i = 1, \ldots, n\}$ 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. We describe in the Appendix 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

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

where the kernel weights w_i^F 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 sampled scatterplot of Figure 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 Appendix for details.

A benefit of the development of the conditional mean estimator (3) as a weighted mean of the $y_i's$ is that the ap-

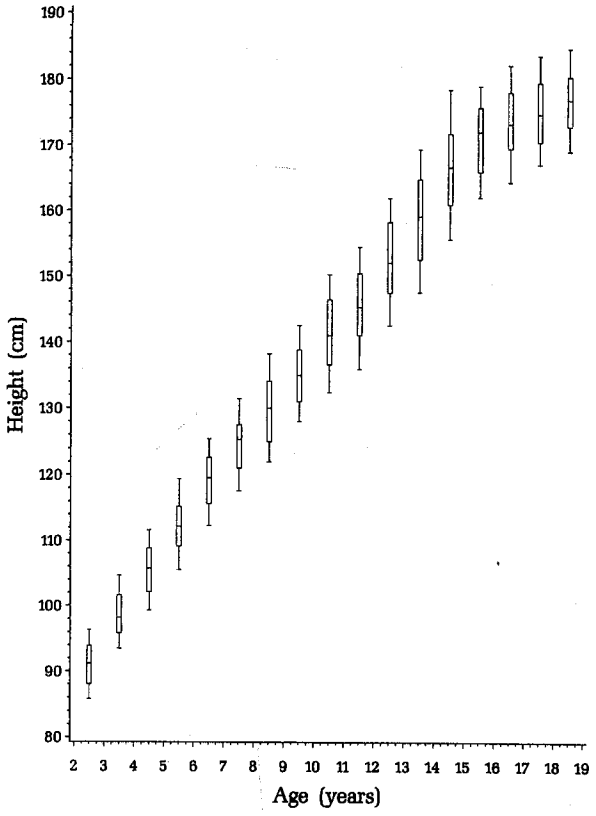

Figure 14. Strip Box Plot of Height Versus Age for Data Plotted in Figure 6. Box plots show weighted 10th, 25th, 50th, 75th, and 90th percentiles for each year of age.

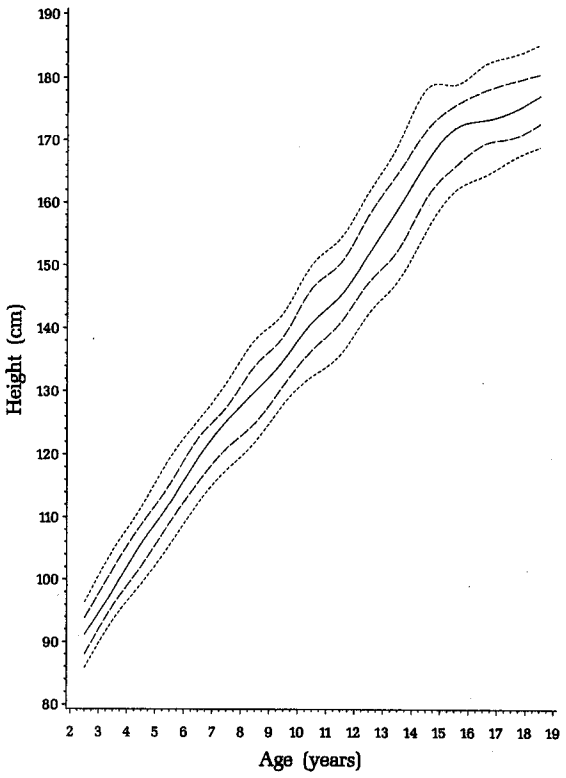

Figure 15. Cubic Spline Interpolation of Weighted Percentiles Shown in Figure 14. Solid line is the median, dashed lines are the quartiles, and dotted lines are the 10th and 90th percentiles.

proach 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 cumulative distribution function (CDF) of the y_i using the w_i^F weights. Unfortunately, this approach has a serious drawback for quantiles other than the median: Even if the relationship of the quantiles and x were linear (but not horizontal), the larger the bandwidth the more the 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 previously; denote it by $\text{med}(y|x)$ and let $z_i = y_i - \text{med}(y|x_i)$. 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 z's given x using only the data points for which $z_i > 0$. If we denote this conditional 80th percentile by $h80(z|x)$, then the desired conditional 90th percentile is

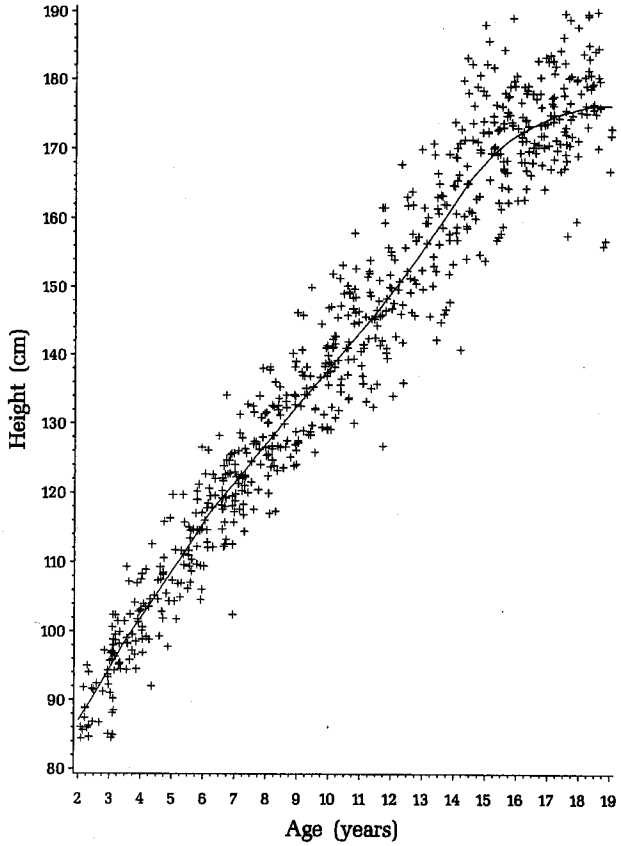

Figure 16. Replot of Figure 7 With the Local Linear Kernel Estimator of the Conditional Mean Using a Triangular Kernel With a Bandwidth Determined by a One-Sided Sample Size of 350.

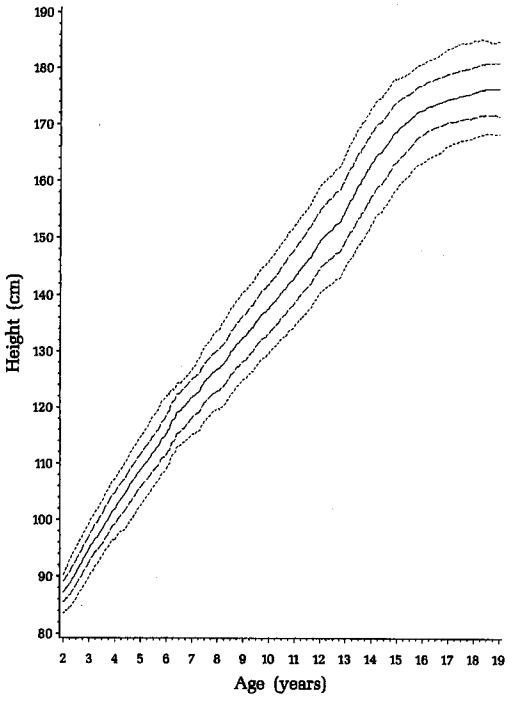

Figure 17. Weighted Conditional Percentiles of Height as a Function of Age of Data Plotted in Figure 6. Solid line is the median, dashed lines are the quartiles, and dotted lines are the 10th and 90th percentiles. Conditional percentiles are estimated using a local linear kernel estimator using a triangular kernel with a bandwidth determined by a one-sided sample size of 350 (see text).

estimated by $\text{med}(y|x)+h80(z|x)$. 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 to the nearest integer (plot not shown). If this is a problem, the weighted empirical CDF calculated at each can itself be smoothed before estimating the percentiles; see Woodruff (1952) and Korn, Midthune, and Graubard (1997) for some simple methods of doing this.

We end this section with an example showing how to examine whether a smoothed conditional mean or percentile curves is reflecting a property of the underlying distributions rather than just noise. As an example, Figure 18 is a partial residual plot for the logarithm of blood lead from a (sample-) weighted linear regression of systolic blood pressure on loglead, age, and body mass index using the data described previously. Partial residual plots for an independent variable x, also known as component-plus-residual plots, are plots of the residuals plus x times the estimated regression coefficient of x versus the independent variable (Atkinson 1985, chap. 5.4; Cook and Weisberg 1994, chap. 9). These plots are useful for examining possible needed transformations of the independent variable. The dotted line in Figure 18 is the weighted least-squares line; its slope is identical to the estimated regression coefficient of loglead in the weighted multiple linear regression.

The smooth curve in Figure 18 is a local linear kernel estimator of the conditional mean using a triangular kernel with the fixed bandwidth of ± 1.5 units of loglead. The curve shows no great nonlinearity, although there is the suggestion of a rise and then fall of the curve for loglead values greater than 3.5. To check the reality of this nonlinearity,

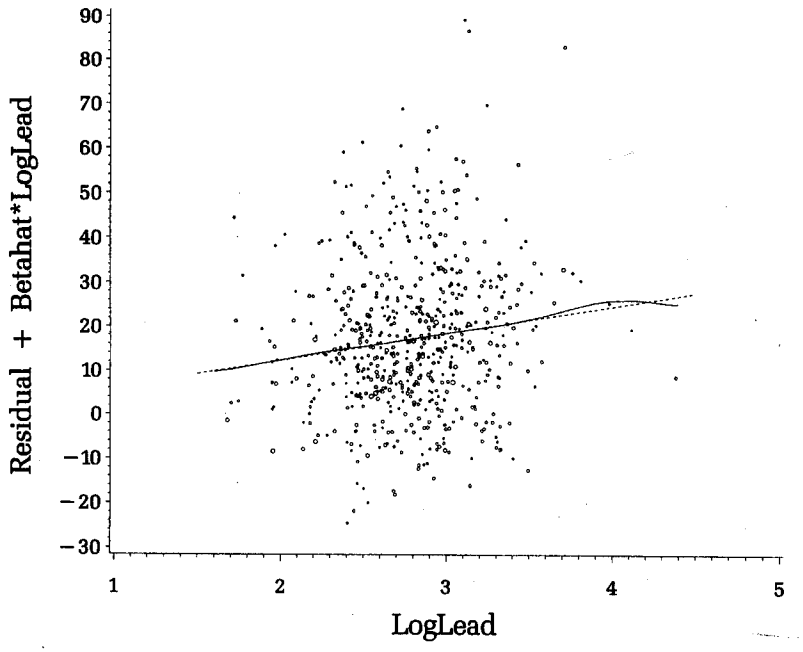

Figure 18. Partial Residual Plot of the Logarithm of Blood Lead (loglead) From a Weighted Regression of Systolic Blood Pressure on Loglead, Age, and Body Mass Index Using Data From 595 White Males Aged 40–59 Sampled in the Second National Health and Nutrition Examination Survey. Areas of circles are proportional to the sample weights. Dotted 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 loglead.

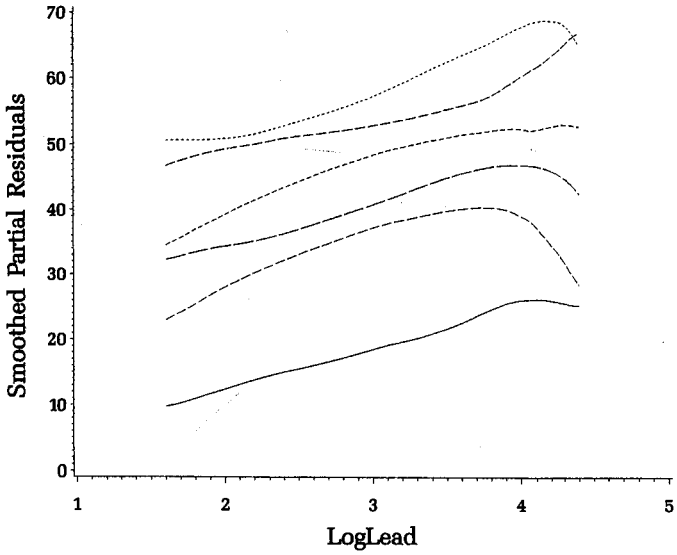

Figure 19. Replot of the Kernel Estimator of the Conditional Mean From Figure 18 (solid curve) With Kernel Estimators of the Conditional Mean Based on Five Simulated Data Sets for Which the Conditional Mean Should be Linear (dashed and dotted curves, translated in the vertical direction to avoid overlapping curves).

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 values were simulated with normal distributions around the predicted values (with standard deviation equal to the residual standard deviation from the observed data set). There should be no structure in the residuals from the weighted linear regressions using these simulated data sets. The top five curves in Figure 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 Figure 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.

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 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 article 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.

APPENDIX

Let the kernel function $K(u)$ be a nonnegative symmetric function that integrates to one; for example, the triangular kernel $K(u) = (1 - |u|)I(|u| \leq 1)$. One possible kernel estimator of the conditional mean is given by

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

where $w_i^K = K\left(\frac{x-x_i}{h_x}\right) \Big/ \sum_{j=1}^{n} K\left(\frac{x-x_j}{h_x}\right)$ 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. To avoid this problem, one can use a locally weighted regression (Cleveland 1979), with a local linear smoother being a special case: Instead of using the weighted mean (A.1), one fits a weighted linear regression to the data around x using the w_i^K weights. Then, one defines $\text{mean}^L(y|x)$ to be the predicted value of Y at $X = x$ from this regression. The estimator $\text{mean}^L(y|x)$ can still be defined as a weighted mean, namely,

$$\text{mean}^L(y|x) = \sum_{i=1}^{n} w_i^L y_i \tag{A.2}$$

with weights equal to

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

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.

To account for the sample weights (w_i), one lets

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

and

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

where $\overline{x}^{KS} = \sum_{j=1}^{n} w_j^{KS} x_j$. The local linear smoother is then defined by (3). The use of the sample weights implies that (3) is estimating what (A.2) would be estimating if all the population values were available and used for the estimation.

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, chap. 5); 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. 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.) We shall refer to this as a bandwidth determined by a one-sided sample size of 50.

[Received May 1996. Revised January 1997.]

REFERENCES

Atkinson, A.C. (1985), *Plots, Transformations, and Regression*, Oxford: Clarendon Press.

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983), *Graphical Methods for Data Analysis*, Belmont, CA: Wadsworth International Group.

Cleveland, W. S. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, 829–836.

Cleveland, W. S., and McGill, R. (1984), "The Many Faces of a Scatterplot," *Journal of the American Statistical Association*, 79, 807–822.

Cook, R. D., and Weisberg, S. (1994), *An Introduction to Regression Graphics*. New York: Wiley.

Guo, S., Roche, A. F., Baumgartner, R. N., Chumlea, W. C., and Ryan, A. S. (1990), "Kernel Regression for Smoothing Percentile Curves: Reference Data for Calf and Subscapular Skinfold Thicknesses in Mexican Americans," *American Journal of Clinical Nutrition*, 51, 908S–916S.

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

Hinkins, S., Oh, H. L., and Scheuren, F. (1994), "Inverse Sampling Design Algorithms," in *1994 Proceedings of the Section on Survey Research Methods*, Alexandria, VA: American Statistical Association, pp 626–631.

Korn, E. L., and Graubard, B. I. (1995), "Analysis of Large Health Surveys: Accounting for the Sample Design," *Journal of the Royal Statistical Society*, Ser. A., 158, 263–295.

Korn, E. L., Midthune, D., and Graubard, B. I. (1997), "Estimating Interpolated Percentiles from Grouped Data with Large Samples," *Journal of Official Statistics*, in press.

Little, R. J. A., and Rubin, D B. (1987), *Statistical Analysis with Missing Data*, New York: Wiley.

McDowell, A., Engel, A., Massey, J. T., and Maurer, K. (1981), "Plan and Operation of the Second National Health and Nutrition Examination Survey, 1976-80," *Vital and Health Statistics*, Series 11, No. 15, Washington, DC: National Center for Health Statistics.

Murthy, M. M., and Sethi, V. K. (1965), "Self-Weighting Design at Tabulation Stage," *Sankhya*, Series B, 27, 201–210.

National Center for Health Statistics (1976), "NCHS Growth Charts, 1976," in *Monthly Vital Statistics Report*, vol. 25, no. 3, Suppl. (HRA) 76-1120. Rockville, MD: Health Resources Administration.

National Center for Health Statistics, Annest, J. L., and Mahaffey, K. (1984), "Blood Lead Levels for Persons Ages 6 months-74 years, United States, 1976-80," in *Vital and Health Statistics*, Series 11, No. 223 (DHHS pub. no. PHS 84-1683), Washington, DC: U.S. Government Printing Office.

National Center for Health Statistics, Cohen, B. B., Barbano, H. E., Cox, C. S., et al. (1987), "Plan and Operation of NHANES I Epidemiologic Followup Study, 1982-84," in *Vital and Health Statistics*, Series 1, No. 22 (DHHS pub. no. PHS 87-1324). Washington, DC: U.S. Government Printing Office.

O'Hara Hines, R. J., and Carter, E. M. (1993), "Improved Added Variable and Partial Residual Plots for the Detection of Influential Observations in Generalized Linear Models" (with discussion), *Applied Statistics*, 42, 3–20.

Owen, A. B. (1987), *Nonparametric Conditional Estimation*, Technical Report No. 265, Stanford University, Dept. of Statistics.

Pirkle, J. L., Schwartz, J., Landis, J. R., and Harlan, W. R. (1985), "The Relationship Between Blood Lead Levels and Blood Pressure and Its Cardiovascular Risk Implications," *American Journal of Epidemiology*, 121, 246–258.

Sanderson, M., Placek, P. J., and Keppel, K. G. (1991), "The 1988 National Maternal and Infant Health Survey: Design, Content, and Data Availability," *Birth*, 18, 26–32.

SAS (1990), *SAS/GRAPH Software: Reference*, Version 6, First Edition, Volume 1, Cary, NC: SAS Institute, Inc.

Shah, B. V., Barnwell, B. G., and Bieler, G. S. (1995), *SUDAAN User's Manual*, Research Triangle Park, NC: Research Triangle Institute.

Stevens, R. G., Jones, D. Y., Micozzi, M. S., and Taylor, P. R. (1988), "Body Iron Stores and the Risk of Cancer," *New England Journal of Medicine*, 319, 1047–1052.

Stone, C. J. (1977), "Consistent Nonparametric Regression" (with discussion), *Annals of Statistics*, 5, 595–645.

Wang, X., Zuckerman, B., Coffman, G.A., and Corwin, M. J. (1995), "Familial Aggregation of Low Birth Weight Among Whites and Blacks in the United States," *New England Journal of Medicine*, 333, 1744–1749.

Woodruff, R. S. (1952), "Confidence Intervals for Medians and Other Position Measures," *Journal of the American Statistical Association*, 47, 635–646.