Ordered Sample Scatterplots for Displaying Survey Data

Edward Mulrow
NORC at the University of Chicago, 4350 E-W Highway, Bethesda, MD 20814

Abstract

Korn and Graubard (1998) suggested various modifications to make scatterplots more informative for sample survey data. These included the notion of sampled scatterplots—plotting a subsample of the data such that the resulting subsample is approximately a simple random sample. Hinkins et al (2009) used inverse sampling to construct regression diagnostic scatterplots, and illustrated how multi-panel plots of many inverse samples could be used to account for the loss of information when subsampling. One criticism of this approach is that the multi-panel displays may be harder to view; a smaller set of plots, or ideally a single plot, may be better. We explore using the ideas of Jones and Rice (1992) to order a large collection of curves. Using this technique on smoothers through scatterplots of survey data subsamples will provide a way to order scatterplots and therefore reduce the number, e.g. a minimum, median, and maximum plot or some other alternative based on order statistics.

Key Words: Complex survey design, inverse sampling, sampled scatterplots

1. Introduction

Many survey designs concentrate on obtaining samples that will produce precise “enumerative” estimates, e.g. population totals, and probability samples that minimize the variance of important population quantities are desired. Most visualization techniques are not designed for complex samples; simple random samples are more appropriate. But some visualization techniques can be used to produce “population” visualizations, e.g. box plots (Lumley, 2007).

It’s possible to modify scatterplots by incorporating survey weights into the plots, e.g. bubble plots (Korn and Graubard, 1998; Lohr, 1999; Lumley, 2007), or “population” scatterplots via hex binning (Lumley, 2007). However, these methods may make it hard to visualize trends in the data, which is a primary purpose of the scatterplot.

Methods have been proposed to display or analyze complex data without using the sample weights. Hinkins, Oh, and Scheuren (1994, 1997) proposed using inverse samples—subsamples of the complex survey sample that have the features of a simple random sample—for a variety of analytic problems with survey data. Korn and Graubard (1998) and Lumley (2007) suggest similar “synthetic” approaches, that is, a "sampled scatterplot." Korn and Graubard note that while a sampled scatterplot is preferred over a bubble plot for a good visual display of the population, there is a loss of information when choosing a display of a subsample over a display of all the data. Additionally, atypical points may be lost in the sampling process.

Hinkins et al (2009) used inverse sampling to construct regression diagnostic scatterplots, and illustrated how multi-panel plots of many sampled scatterplots could be used to account for the loss of information when subsampling. One criticism of this approach is that the multi-panel displays may be harder to view; a smaller set of plots, or ideally a single plot, may be better.

In this paper we explore ordering a large set of sampled scatterplots in order to reduce the number of displayed plots to a small but representative subset, such as a minimum, median and maximum sampled scatterplot, or some other alternative based on order statistics. In Section 2 we provide examples of different styles of survey data scatterplot, and discuss the strengths and weaknesses of the displays. In Section 3, we review the ideas of Jones and Rice (1992) to order a large collection of curves. In section 4, we apply this technique to smoothers through a collection of sampled scatterplots. Section 5 provides some observations on ordered sampled scatterplots.

2. Survey Data Scatterplots

Korn and Graubard (1998) have noted that 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. However, using a scatterplot with survey data collected under a complex sample design can be misleading. Survey designs concentrate on obtaining samples that will produce precise “enumerative” estimates such as population totals or percentages. Probability samples that minimize the variance of important population quantities are desired. Most visualization techniques are not designed for complex samples; simple random samples are more appropriate.

To illustrate the issues, we follow Korn and Graubard (1998), using data from the 1988 National Maternal and Infant Health Survey for our illustrations. **Figure 1** is a scatterplot of daughter's birth weight versus mother's birth weight for mothers aged 30-39 years at the time of birth. A number of authors have shown that survey weights cannot be ignored when modeling data from complex sample designs (e.g. Little, 2008, Hinkins, Mulrow, and Scheuren, 2009). Because one of the primary purposes of a scatterplot is to assess whether or not a modeling relationship exists between the variable, the scatterplot in **Figure 1** is misleading. There is no attempt to account for the complex sample design in the plot, and it should not be relied upon to assess a modeling relationship.

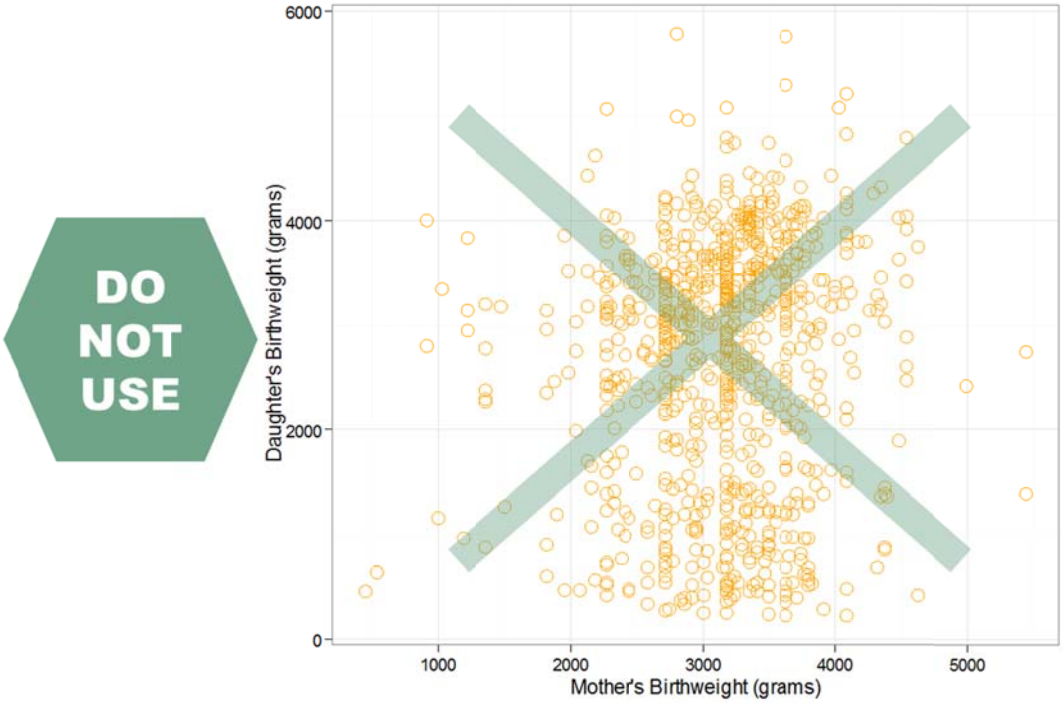

Figure 1: Simple scatterplot on data from Mothers aged 30-39 surveyed in the 1988 National Maternal and Infant Health Survey (n=879). The scatterplot is misleading because it does not account for the complex sample design.

2.1. Incorporate Weights into the Scatterplot

It's possible to modify scatterplots by incorporating survey weights into the plots: Bubble plots (Korn and Graubard, 1998; Lohr, 2009; Lumley, 2010), and "population" scatterplots via hex binning (Lumley, 2010) are possibilities. These methods may make it hard to visualize trends in the data, though.

Figure 2 illustrates a bubble plot. It is scatterplot of daughter's birthweight versus mother's birthweight for mothers aged 30-39 years at the time of birth—the same data plotted in **Figure 1**. The complex sample design is incorporated in the plot by making the area of each plotting symbols proportional to the sample weight. Even though the complex design is incorporated into the plot, the overall trend is not apparent in the bubble plot. On the plus side, atypical points may be noticeable.

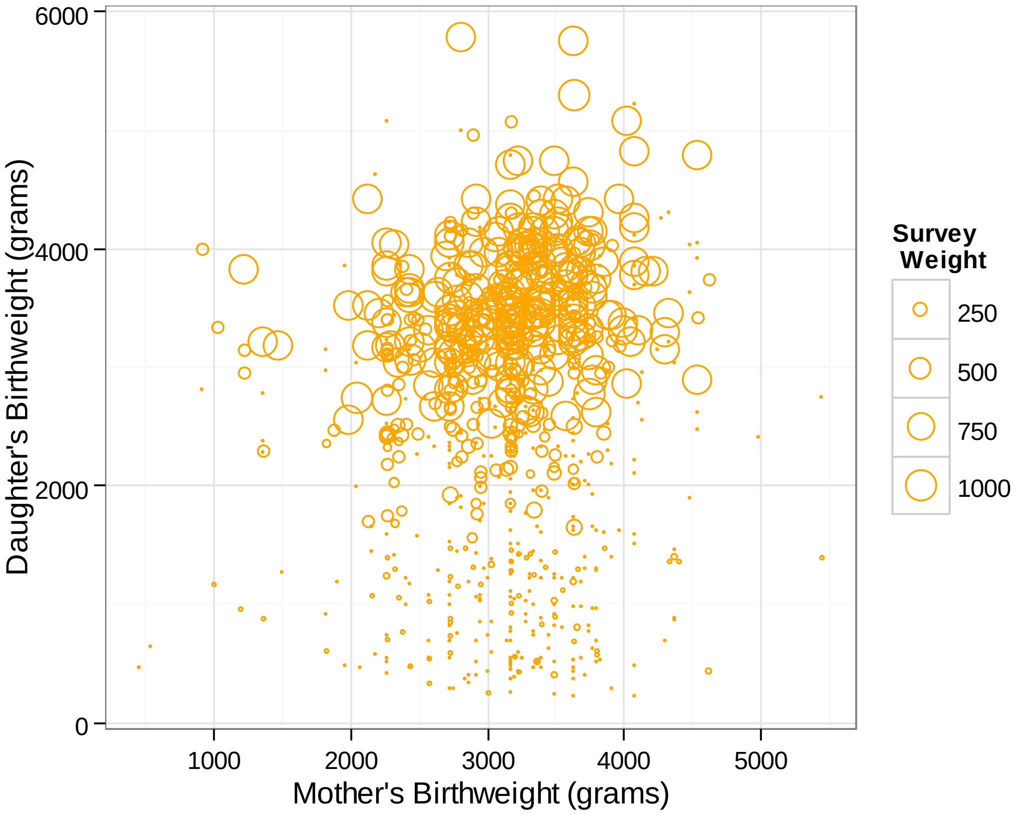

Figure 2: Bubble plot of data plotted in **Figure 1**. Areas of circles are proportional to the sample weights.

2.2. Compensate for the Weights via Inverse Sampling

Hinkins, Oh, and Scheuren (1994, 1997) proposed using inverse samples—subsamples of the complex survey sample that have the features of a simple random sample—for a variety of analytic problems. Korn and Graubard (1998) and Lumley (2007) suggest similar “synthetic” approaches, that is, a "sampled scatterplot."

Figure 3 is a sampled scatterplot of the data plotted in **Figure 1**. Instead of plotting all the points, a subsample was selected using probability proportional to size with replacement sampling where the measure of size is the survey weight. The sample size for the subsample is 325. This value was chosen because it is the effective sample size under the sample design of the 879 points plotted in **Figure 1**. Because points with high weights are selected many times, the plotting symbols are jittered to avoid over plotting. An overall trend is now easier to discern. However, atypical points could go unnoticed because only a subset of the data is plotted. There is an overall loss of information.

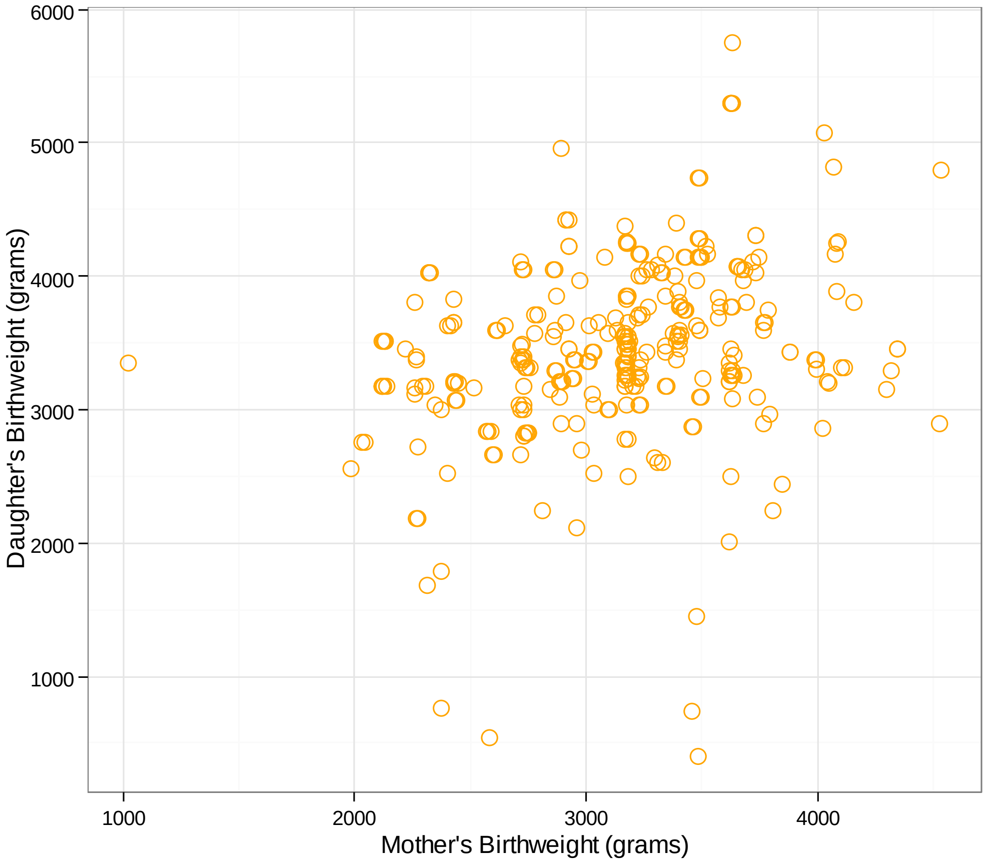

Figure 3: A sampled scatterplot of the data plotted in **Figure 1** based on a probability proportional to the survey weight subsample of 325. Plotting symbols are jittered to avoid over plotting.

2.2.1. Panel of Sampled Scatterplots

The loss of information due to the much smaller sample size can be offset by drawing multiple, conditionally independent, inverse samples.

- For estimation of means and totals, aggregating multiple inverse subsamples can achieve nearly the efficiency of the original design and unbiased estimates of the standard errors can be calculated from the aggregate.
- Similarly, viewing multiple sampled scatterplots can reduce the visual loss of information. Hinkins, Mulrow, and Scheuren (2009) illustrated how multi-panel plots of many sampled scatterplots could be used to account for the loss of information when subsampling.

Figure 4 is a panel of 20 sampled scatterplots. Each individual plot is a scatterplot of a different subsample of 325 from the 879 data points shown in **Figure 1**. By viewing a collection of sampled scatterplots, we reduce the loss of information.

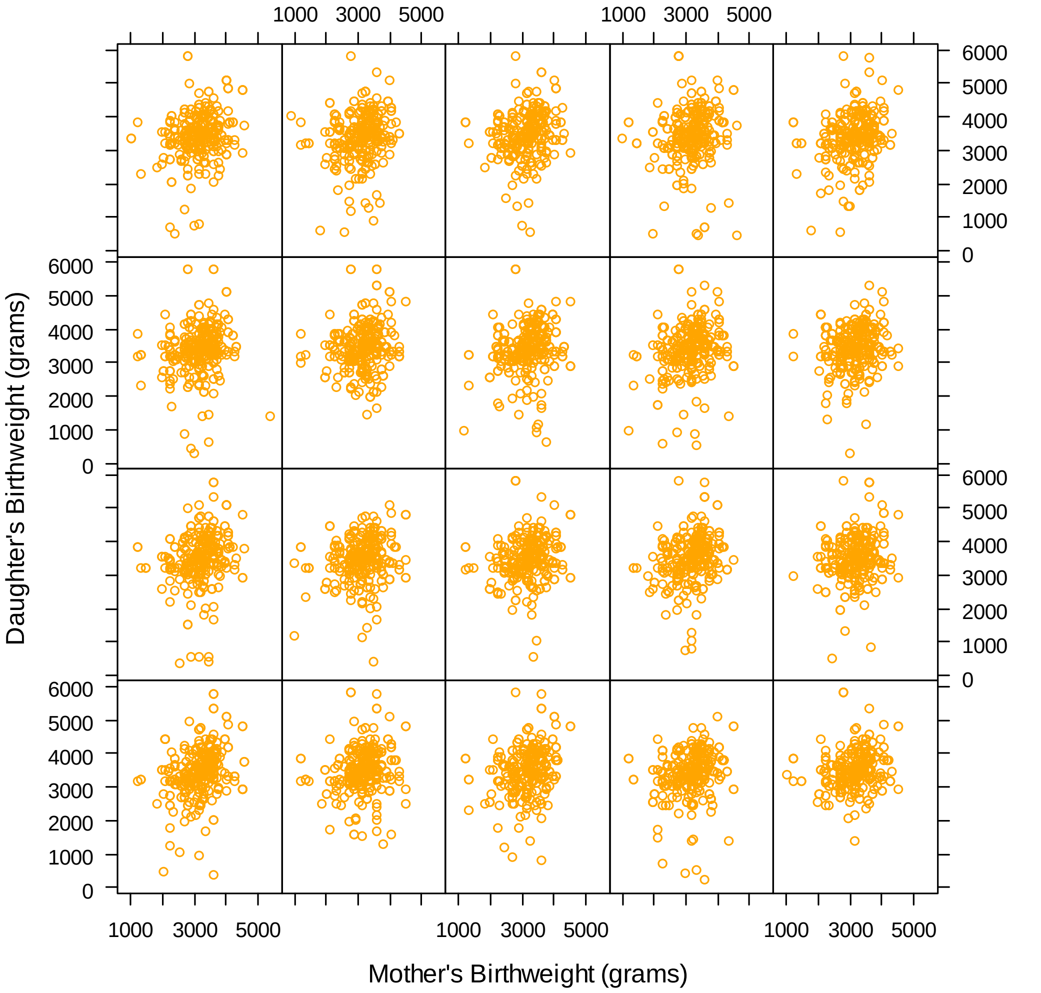

Figure 4: Twenty sampled scatterplot of data plotted in Figure 1 based on a probability proportional to the survey weight subsample of 325. Plotting symbols are jittered to avoid over plotting.

The plots are similar in terms of assessing a trend. Additionally, atypical points can be seen in a few of the plots. One criticism of this approach is that the multi-panel displays may be hard to view, and there is no guarantee that atypical points will turn up if a very large number of plots is not viewed.

The method would be more useful if we could reduce the number of plots without losing too much information. Is there a way to reorganize the display or reduce the number of sampled scatterplots?

Guha, Kidwell, Hafen, and Cleveland (2009) discuss how a large display of ordered graphs can be effectively explored in a sequential manner. Thus, if a collection of sampled scatterplots can be sorted it will be possible to reduce the number of plots to a small set based on key order statistics.

3. Ordering a Large Collection of Curves

Jones and Rice (1992) suggest using principle component analysis to display only those few curves out of a large collection that best reflect the important modes of variation present in that generally similar collection of curves. For a large collection of curves, which can be thought of as N functions evaluated at K points of a common equispaced grid, standard principal component analysis can be used to explore important modes of variation. The projections of the data points onto the principal component directions—the principal component scores—can be used for selecting important order statistics. For the principle component that account for the largest proportion of the overall variation, the curves corresponding to principal component score order statistics of interest (e.g. minimum, median, maximum) can be plotted to illustrate the mode of variation.

Exhibit 1 comes from Jones and Rice (1992) and shows three density estimate curves out of a large collection of density curve estimates generated from a simulation. The plot shows the curves associated with the minimum, median and maximum principle component scores for the first principle component.

Exhibit 1: Jones and Rice (1992) display of important modes of variation**.** *Figure 1* mentioned in the caption below this exhibit is a plot of 1,000 density estimates from simulated data.

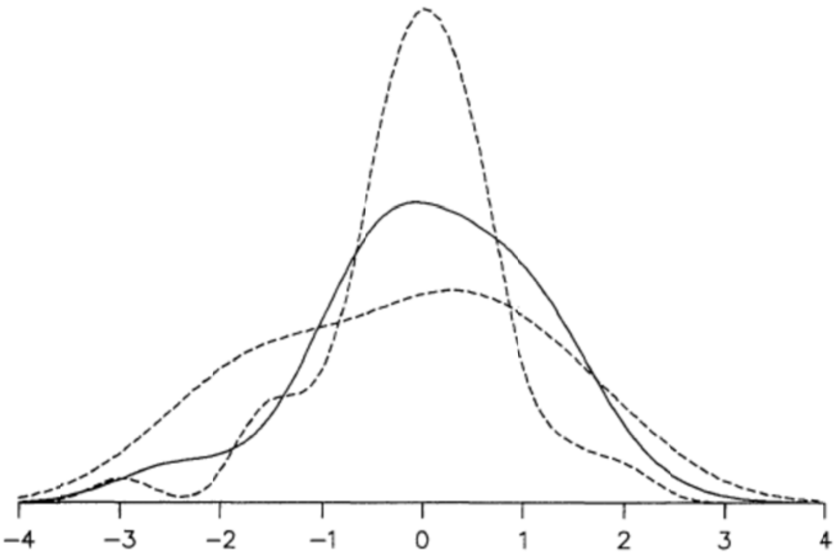

Figure 2. Curves Representative of the First Mode of Variation of the Density Estimates. The curves correspond to the median (solid line), minimum, and maximum (dashed lines) scores on the first principal component, extracted from Figure 1.

We now explore using Jones' and Rice's methodology to order sampled scatterplots.

4. Ordering Sampled Scatterplots

One of the primary purposes for using a scatterplot is to visualize trends. An aid to doing so is smoothers such as loess or Friedman's super smoother. We propose using Jones and

Rice (1992) on smoother curves through each of the sampled scatterplots as a way to order them. We proceeded as follows.

- Select 1,000 inverse sample of size n=325 from the set of mothers aged 30-39 at the time of birth in the 1988 National Maternal and Infant Health Survey.
    - Use probability proportional to size with replacement sampling where the measure of size is the survey weight.
- Fit a loess smoother to the mother's\daughter's birthweight pairs for each inverse sample.
- Find the predicted value of each loess smoother for each mother's birthweight on an equispaced grid between the minimum and maximum mother's birthweight from the full sample.
- Determine the principal components that account for at least 95% of the total variation in the loess curves.
    - Robustness features are turned off when fitting the loess curves with the hope that atypical points will influence the fit.
- For each principal component, plot the daughter's versus mother's birthweight pairs for the inverse samples that correspond to the minimum, median, and maximum principal component scores.
    - The results are shown in **Figure 5**.
    - The first principle component (PC1) and the second principle component (PC2) account 98% of the total variation (61% and 37%, respectively).

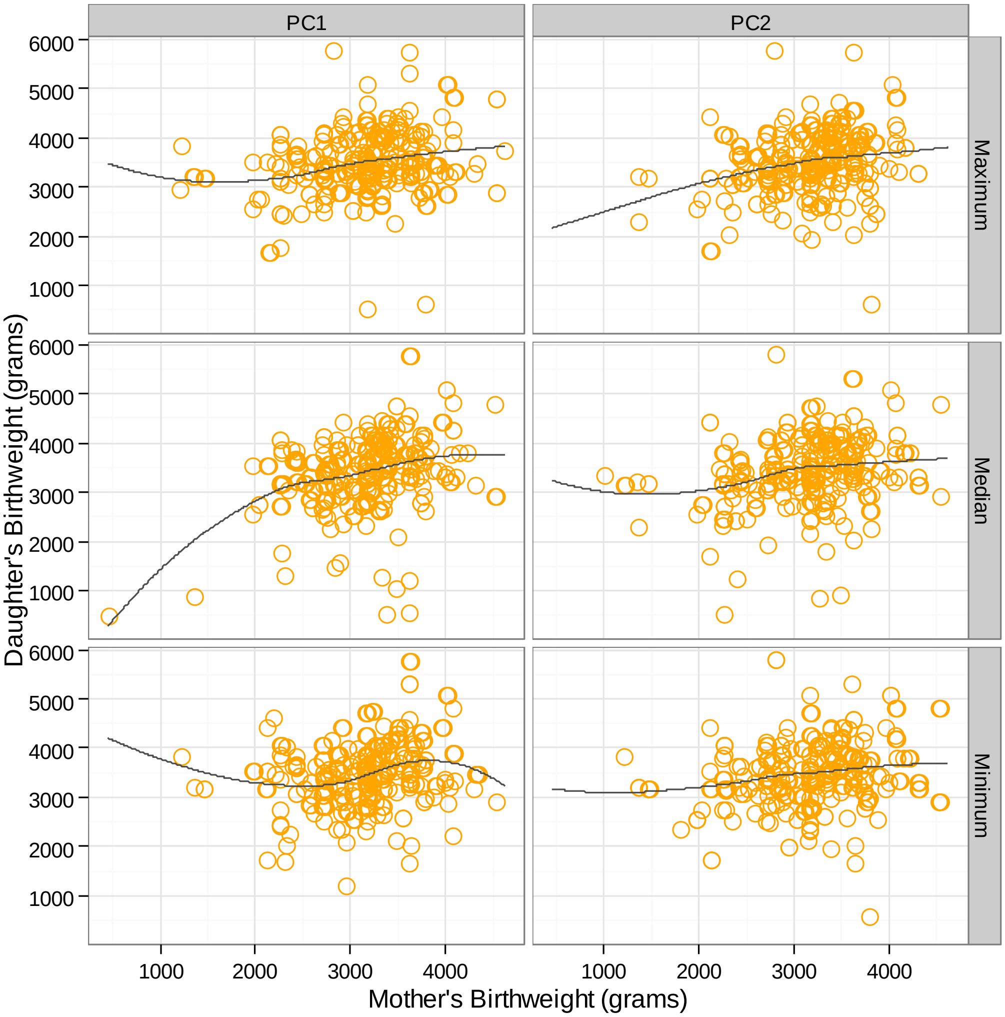

Figure 5: Sampled scatterplots of data plotted in Figure 1 based on principle component scores of the first two principle components of loess smoothers from 1,000 probability proportional to the survey weight with replacement subsamples of 325. Plotting symbols are jittered to avoid over plotting.

In using this approach, we hope that a large number of inverse samples will produce an array of trends (as determined by the loess smoother). Additionally, we assume that the large number of samples will include some atypical points that affect trend. By looking at a few plots based on order statistics within principle components of the smoothers variation, we hope to see plots with different features that overcome the loss of information when looking at only one sampled scatterplot. In Figure 5, we do notice some differences in trends for each sampled scatterplot, and we have an idea of how the trend varies across plots. Additionally, some atypical points are noticeable. So, we have successfully reduced the number of sampled scatterplots to an easy-to-view, small set of plots that help explore relationship with data collected under a complex survey design.

4. Observations

As Korn and Graubard (1998) note, there is not one plot that tells a complete story. However, displays of a small, strategically chosen, collection of sampled scatterplots are useful. Ordering sampled scatterplots provides better insight into bivariate relationships. General trends are noticeable, and, assuming a large enough collection of sampled scatterplots is analyzed, atypical points are noticeable as well.

The choice of an ordering method is somewhat arbitrary. We chose to use a method based on Jones' and Rice's method for studying modes of variation across a large collection of curves. But other methods of ordering may be possible as well. For example, a scatterplot's convex hull area could be used. This might place more emphasis on finding atypical points and less on the trend.

An interactive graphic, or an animation, based on ordered sampled scatterplots would enable analysts to explore relationships and find atypical points. Features such as choice of smoothing method, the number of principle components, or the choice of order statistics could be turned on/off by an analyst allowing a more complete view of the data. We intend to explore this more in the future

Acknowledgements

The author would like to thank his NORC colleagues Fritz Scheuren and Susan Hinkins for their help and advice. It was their work on inverse sampling that influenced this paper.

References

Flury, B., D. and Tarpey, T. (1993), “Representing a Large Collection of Curves: A Case for Principal Points,” *The American Statistician*, 47, 304-306

Hinkins, S., Oh, H. L., and Scheuren, F. (1994), “Inverse Sampling Design Algorithms,” *1994 Proceedings of the Section on Survey Research Methods*, American Statistical Association, 626-631.

Hinkins, S., Oh, H. L., and Scheuren, F. (1997), “Inverse Sampling Design Algorithms,” *Survey Methodology*, Statistics Canada, June 1997, Vol. 23, No. 1, 11-21.

Hinkins, S., Mulrow, E., and Scheuren, F. (2009), “Visualization of Complex Survey Data: Regression Diagnostics,” *2009 Proceedings of the Section on Survey Research Methods*, American Statistical Association, 2206-2218.

Jones, M. C., and Rice, J. A. (1992), "Displaying the Important Features of Large Collections of Similar Curves," *The American Statistician*, 46, 140-145.

Korn, E. L., and Graubard B. I. (1998), “Scatterplots with Survey Data,” *The American Statistician*, Vol. 52, No. 1, 58-69.

Little, R. (2008). “Weighting and Prediction in Sample Surveys,” *Calcutta Statistical Association Bulletin*, Vol. 60, Nos. 239-240.

Lohr, S. (2009). *Sampling: Design and Analysis*, 2nd Edition. Cengage Learning, ISBN 0495105279.

Lumley, T. (2007). “Complex survey samples in R,” http://faculty.washington.edu/tlumley/survey/survey-wss.pdf.

Lumley, T. (2012) "survey: analysis of complex survey samples". R package version 3.28-2.

Pfeffermann , D. (1993). “The Role of Sampling Weights When Modeling Survey Data”, *International Statistical Review*, Vol. 61, No. 2, 317-337.

R Development Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

Rao, J.N.K., Scott, A.J. and Benhin, E. (2003). “Undoing Complex Survey Data Structures: Some Theory and Applications of Inverse Sampling,” *Survey Methodology*, Vol. 29, No. 2, 107-128.

Sarkar, Deepayan (2008) Lattice: Multivariate Data Visualization with R. Springer, New York. ISBN 978-0-387-75968-5

Wickham, H. ggplot2: Elegant Graphics for Data Analysis. Springer New York, 2009.