Visualization of Complex Survey Data: Regression Diagnostics

Susan Hinkins[1], Edward Mulrow[2], Fritz Scheuren[3]
[1]NORC at the University of Chicago, 1122 South 5th Ave, Bozeman MT 59715
[2] NORC at the University of Chicago, 4350 E-W Highway, Bethesda, MD 20814
[3] NORC at the University of Chicago, 4350 E-W Highway, Bethesda, MD 20814

Abstract
Over the last two decades advances in computing technology have provided tools for data visualization that have changed the way statisticians analyze data. However, this revolution in data visualization has not been effectively incorporated in the analysis of complex survey data. In part, this is due to the fact that many visualization techniques are not designed to incorporate survey weights or multi-stage clustered designs in a way that mimics a conventional simple random sample data set. We explore the possibility of changing the structure of the complex survey data through the use of inverse sampling in order to allow the use of common visualization tools. In particular, we examine regression diagnostic plots.

Key Words: Complex survey design, inverse sampling, regression diagnostic plots

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). Regression diagnostic plots are often scatterplots, but producing survey-weight-modified plots may not be easy.

2. Regression Diagnostics

Regression diagnostics provide visualization techniques for assessing the quality of the fit of the data to a model. Residuals vs. Fit plots help to identify patterns in residuals. Normal Q-Q plots help assess the normality of the residuals. Scale (Spread)-Location plots help detect monotone spread (heteroscedascity). Residuals vs. Leverage plots help identify influential data points.

2.1 A Simple Example
Asabere and Huffman (1996) studied home prices of houses in Mount Laurel, NJ. We explore the relationship between List Price (LPrice) and Sale Price (SPrice). A scatterplot with loess smoother indicates that the relationship may be (as expected) linear.

Figure 1 is a set of diagnostic plots for a simple linear fit (SPrice = a + b·LPrice) that indicate problems with the fit. Some aspects of the Residual vs. Fitted and Normal Q-Q plots indicate a good fit, but the residuals are "v-shaped." The Scale-Location plot confirms the monotone spread of the residuals, and the Residual vs. Leverage plot indicates that there is an overly influential data point. A transformation of the data is probably needed to fit a better model, and influential points should be investigated.

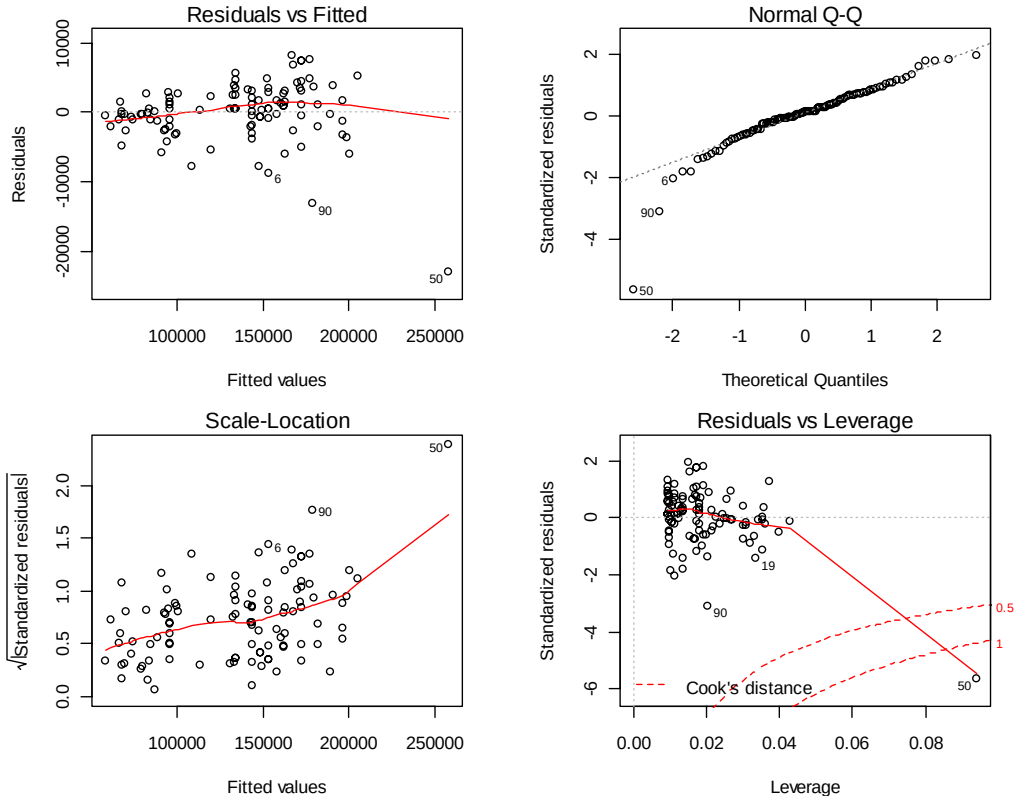

Figure 1: A set of diagnostic plots for a simple linear fit of Sale Price vs. List Price. The diagnostic plots are the default plots given by R for a linear model fit.

2.2 Survey Weights and Regression

The issue of weighting in regressions has long been controversial (Klein and Morgan, 1951; Brewer and Mellor, 1973; DuMouchel and Duncan, 1983). Advocates of ignoring survey weights say that the justification for weighted regression, in terms of adjusting for unequal error variances, is not an issue for estimating model coefficients (DuMouchel and Duncan, 1983). On the contrary, Nathan and Holt (1980) point out that, in general, ordinary least squares (OLS) regression will be biased when using complex survey data—even for large samples. Additionally, when using survey data for analytic purposes, such as regression, appropriate measures must be taken to estimate model parameters (Holt, Smith, and Winter (1980); Pfeffermann and Holmes (1985); Kott 1991).

2.2.1 Ad Hoc Approach

A compromise approach used by some researchers is to fit both an OLS regression and a weighted regression model, and compare the coefficients. If the results are similar, regression diagnostics for the unweighted OLS regression may be appropriate. Let's take a closer look at this approach using survey data from a complex design.

The Survey of Consumer Finances (SCF) is a triennial survey of the balance sheet, pension, income, and other demographic characteristics of U.S. families, sponsored by the Board of Governors of the Federal Reserve System in cooperation with the Internal Revenue Service Statistics of Income Division.

It is a dual-frame sample design—one set of the survey cases was selected from a standard multi-stage area-probability design, the other set of the survey cases was selected as a list sample from the IRS Individual Research Tax File. For disclosure avoidance purposes, the public use data do NOT include most variables related to the sample design (replicate weights are provided to compute reasonable estimates of the sampling variances).

Consider the following two SCF 2007 variables.

- TPAY is the total value of monthly debt payments for a household
- DEBT is the total value of debt held by a household

Is there a simple linear relationship between these variables or a transformation of these variables? An initial analysis indicates that there may be a linear relationship between $TPay^{1/4}$ and $DEBT^{1/4}$. Can we predict the monthly payment from the total debt value using a straight-line model ($TPAY^{1/4} = a + b \cdot DEBT^{1/4}$)? Following the ad hoc regression diagnostics approach for models fit with complex survey data, we compare the coefficients from OLS and weighted regression.

Table 1: OLS fit for the model $TPAY^{1/4} = a + b \cdot DEBT^{1/4}$ using Survey of Consumer Finances 2007 public use data. Fit was obtained using the R function lm().

	Estimate	**Std. Error**	**t value**	**Pr(> |t|)**
Intercept	1.5449	0.0348	44.41	0.000
$DEBT^{1/4}$	0.2462	0.0017	145.96	0.000

Table 2: Weighted regression of the model $TPAY^{1/4} = a + b \cdot DEBT^{1/4}$ using Survey of Consumer Finances 2007 public use data. The fit was obtained using Lumley's R survey package using a replicate weight design and the function svyglm().

	Estimate	**Std. Error**	**t value**	**Pr(> |t|)**
Intercept	1.5106	0.0344	43.90	0.000
$DEBT^{1/4}$	0.2463	0.0019	127.52	0.000

The coefficients are similar, so the ad hoc approach indicates that OLS regression diagnostics may be appropriate. Figure 2 is a set of diagnostic produced by R for the fit of the model $TPAY^{1/4} = a + b \cdot DEBT^{1/4}$ using Survey of Consumer Finances 2007 public use data without the survey weights.

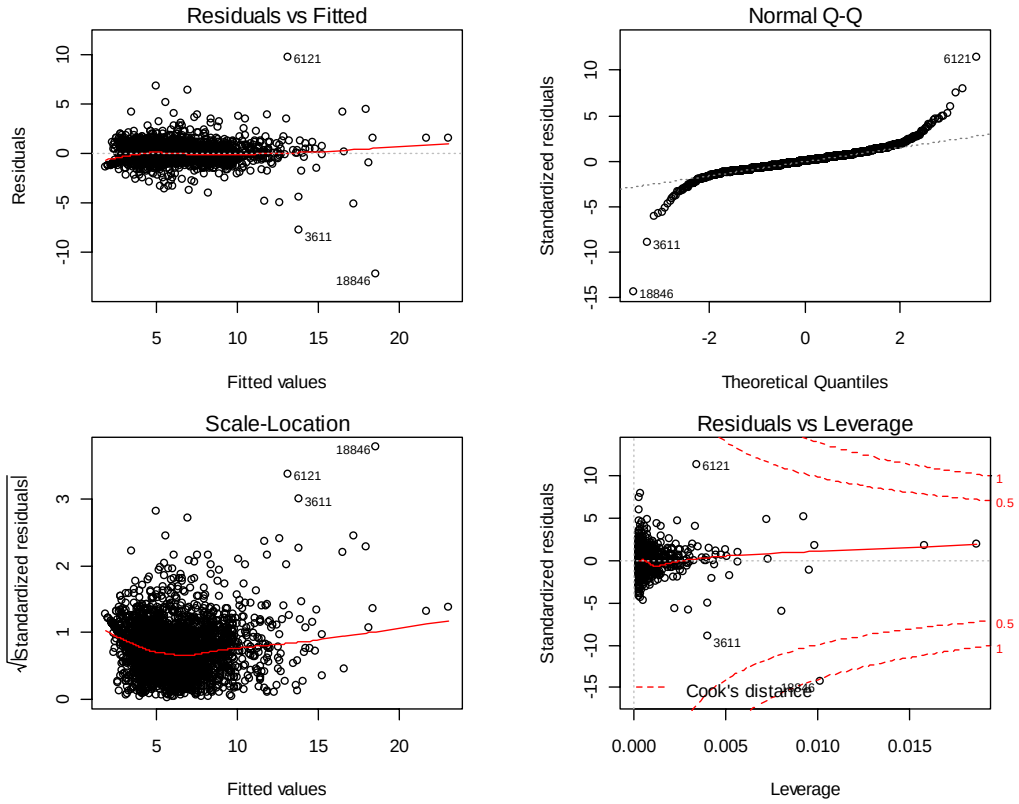

Figure 2: A set of diagnostic plots for the OLS fit of the model $TPAY^{1/4} = a + b \cdot DEBT^{1/4}$ using Survey of Consumer Finances 2007 public use data. The fit was obtained using the R function lm(), and the diagnostic plots are the default plots given by R for a linear model fit.

The diagnostic plots look reasonable. The Scale-Location plot shows some changes in spread, but this might be reasonable. So, using the ad hoc approach for visual regression diagnostics, we might conclude that a straight-line model between $TPay^{1/4}$ and $DEBT^{1/4}$ is adequate.

But are the diagnostic plots affected by the complex survey design? How can we tell? Li and Valliant (2006, 2007, 2009a, 2009b) have done some research on the calculation of regression diagnostic statistics, which would provide a way to redo the diagnostic plots. The R survey package (Lumley, 2009) incorporates survey weights into regression fits, and appropriate diagnostic plots are obtained using the plot() function. Figure 3 is a set of diagnostic plots produced by the output from the R survey package. The Scale-Location plot indicates a clear decrease in the spread of data with the fitted values. So the model may not be appropriate for these data. We should also note that a Scale-Location plot for a weighted regression fit that does not use the replicate weight design for variance calculations looks similar to this plot. However, the Residual vs. Leverage plots differ, so it's better to do the diagnostic plots using a method that incorporates the complex sample design than a method that just incorporates the survey weights.

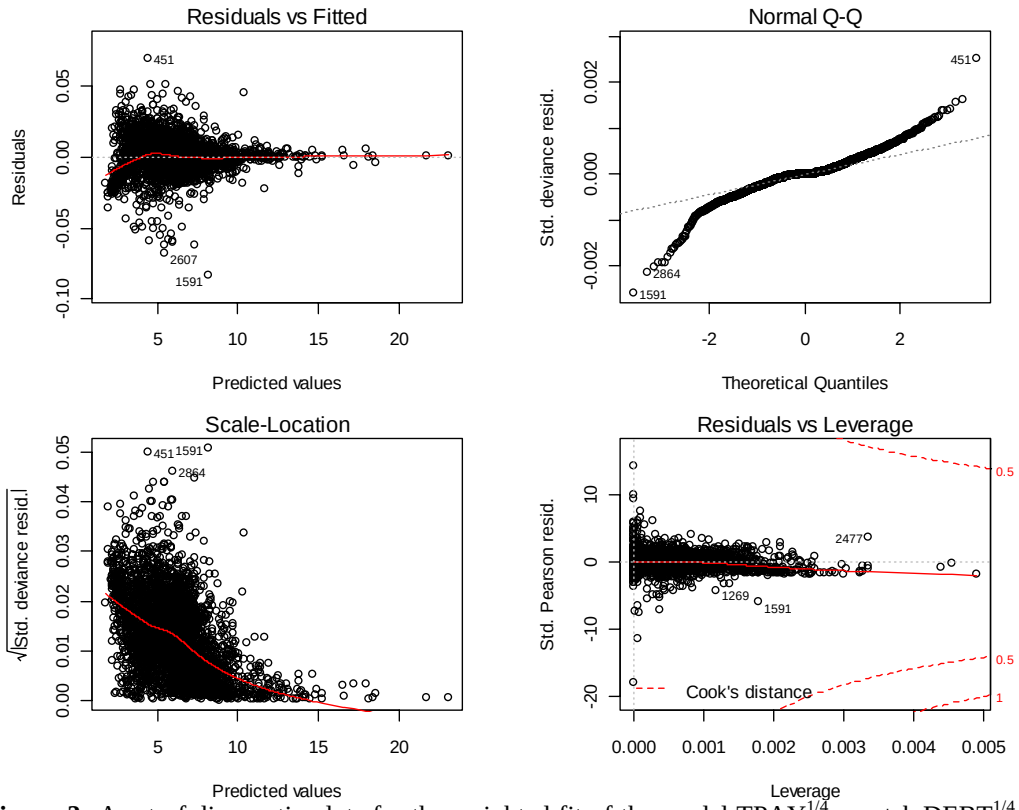

Figure 3: A set of diagnostic plots for the weighted fit of the model $TPAY^{1/4} = a + b \cdot DEBT^{1/4}$ using Survey of Consumer Finances 2007 public use data. The fit was obtained using the R survey package function svyglm(), and the diagnostic plots are the default plots given by R using the plot function with the output of svyglm().

But a researcher interested in analytic modeling of the data that is using the ad hoc approach may not be using a software product like R and its survey package, so we might want a method that does not depend on specialized survey software. It is also the case that many survey data sets are quite large, and it may be hard to interpret scatterplots because of the large amount of symbol over-plotting such as that which occurs in many of the plots in Figure 2 and Figure 3.

3. Reduce the Problem to One That's Solved

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.

3.1 Inverse Samples

An inverse sample is selected by subsampling from the complex sample to obtain a sample equivalent to selecting a simple random sample without replacement (srswor) from the population. An inverse sampling algorithm exists for many types of complex sample designs, as described in Hinkins, Oh and Scheuren (1997) and Rao, Scott and Benhin (2003).

Obviously there are restrictions on the size of the simple random sample that can be selected in this way[1]. For example, in a stratified design where random samples of size n_h are selected from N_h population units in strata $h = 1, 2 \ldots H$, the largest srswor that can be selected using inverse sampling is of size $m = \min\{n_h\}$. A significant loss of information due to the much smaller sample size can be offset by drawing multiple, conditionally independent, inverse samples, conditional on the selected units in the stratified sample. 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.

For regression diagnostic plots, two options could be considered:

- produce separate plots for each inverse sample, or
- combine several inverse samples into one diagnostic plot.

In this paper, panels of diagnostic plots are used, with a separate plot for each sample. Research is needed to investigate the properties of combining the correlated inverse samples for use in regression analysis.

3.1 Panels of Diagnostic Plots

For the analysis described here, it was assumed that the sample structure for the Survey of Consumer Finances 2007 public use data is a stratified design with eight (8) strata and the smallest stratum sample size is $m = 129$. Using the algorithm described in Hinkins, Oh and Scheuren (1997), 20 conditionally independent inverse samples of size 129 were selected.

The same regression model ($\text{TPAY}^{1/4} = a + b \cdot \text{DEBT}^{1/4}$) was fit to each data set and the panels of diagnostic plots are shown in Figure 4, Figure 5, Figure 6, and Figure 7. The first two diagnostic plots correspond reasonably well with the corresponding plots in Figure 2. In the Residual vs. Fitted plot, it is somewhat easier to see a possible change in the spread with the inverse samples than the combined sample.

The Scale-Location plots more clearly indicate that the changes in the spread of the data may be a problem. And some of the Residual vs. Leverage plots show trends that were not as noticeable in the full, complex sample plot. Therefore, the inverse sample plots—similar to Figure 3—suggest that the model is not appropriate.

[1] The inverse algorithm selects a sample of size m so that, unconditionally, every sample of size m in the population is equally likely to be selected.

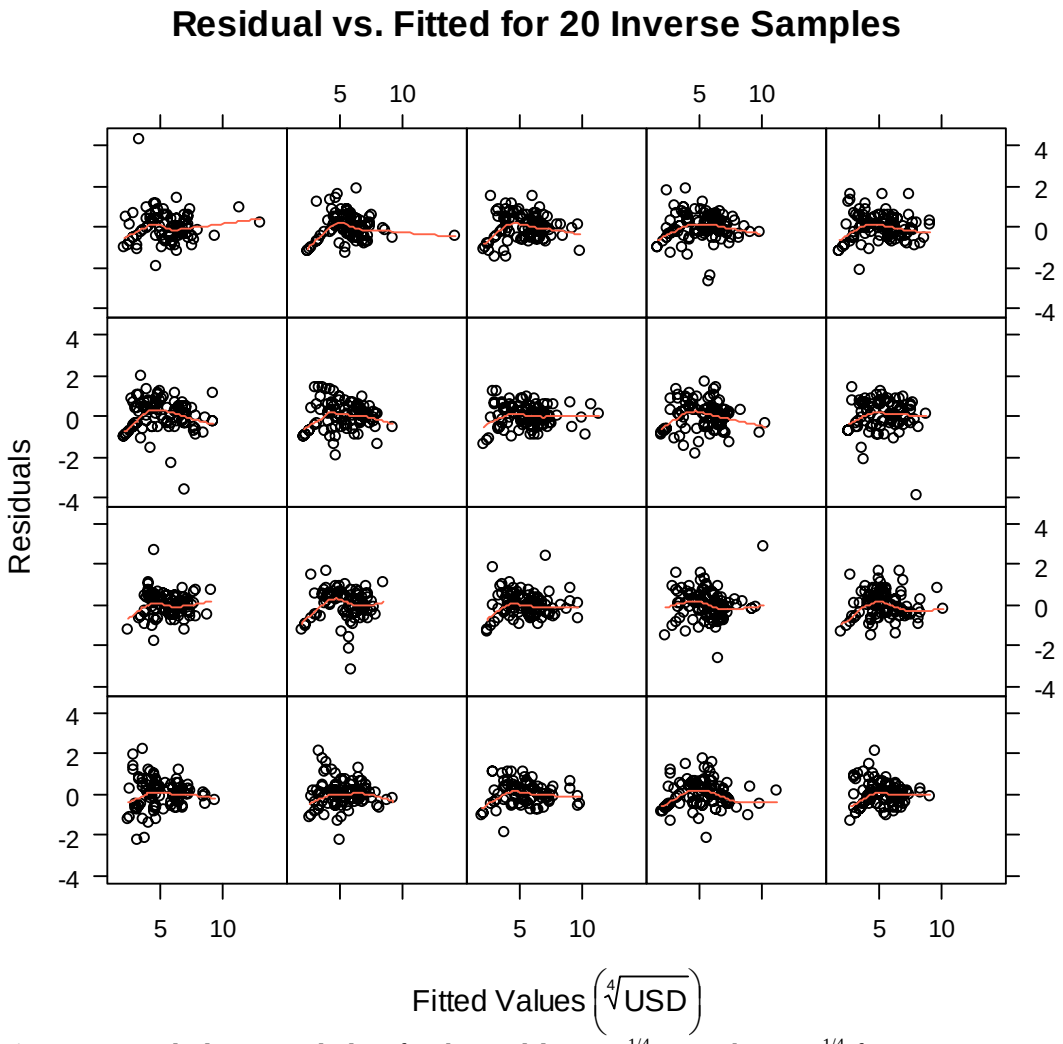

Figure 4: Residual vs. Fitted plots for the model $TPAY^{1/4} = a + b \cdot DEBT^{1/4}$ from 20 inverse samples from the SCF 2007 Public Use data set.

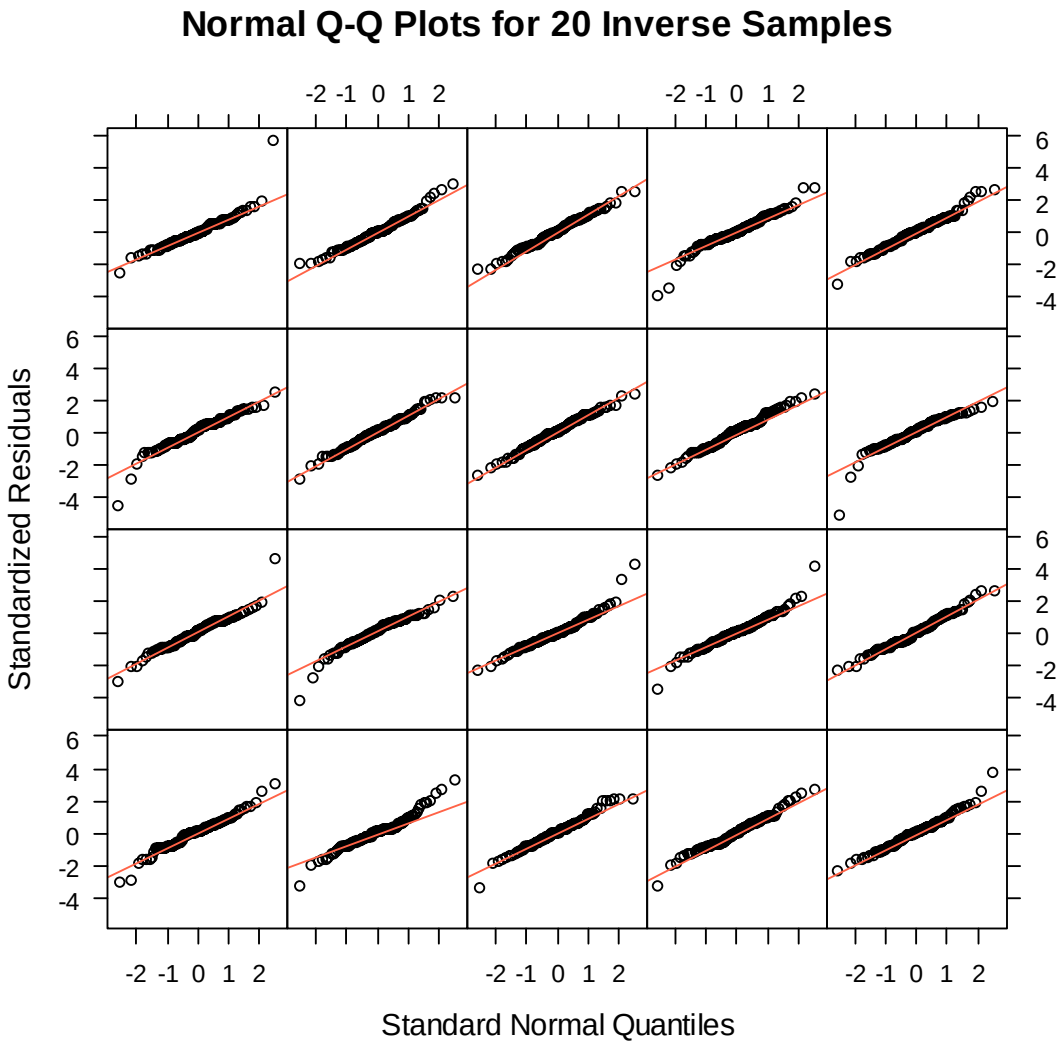

Figure 5: Normal Q-Q plots for the model $TPAY^{1/4} = a + b \cdot DEBT^{1/4}$ for 20 inverse samples from the SCF 2007 Public Use data set.

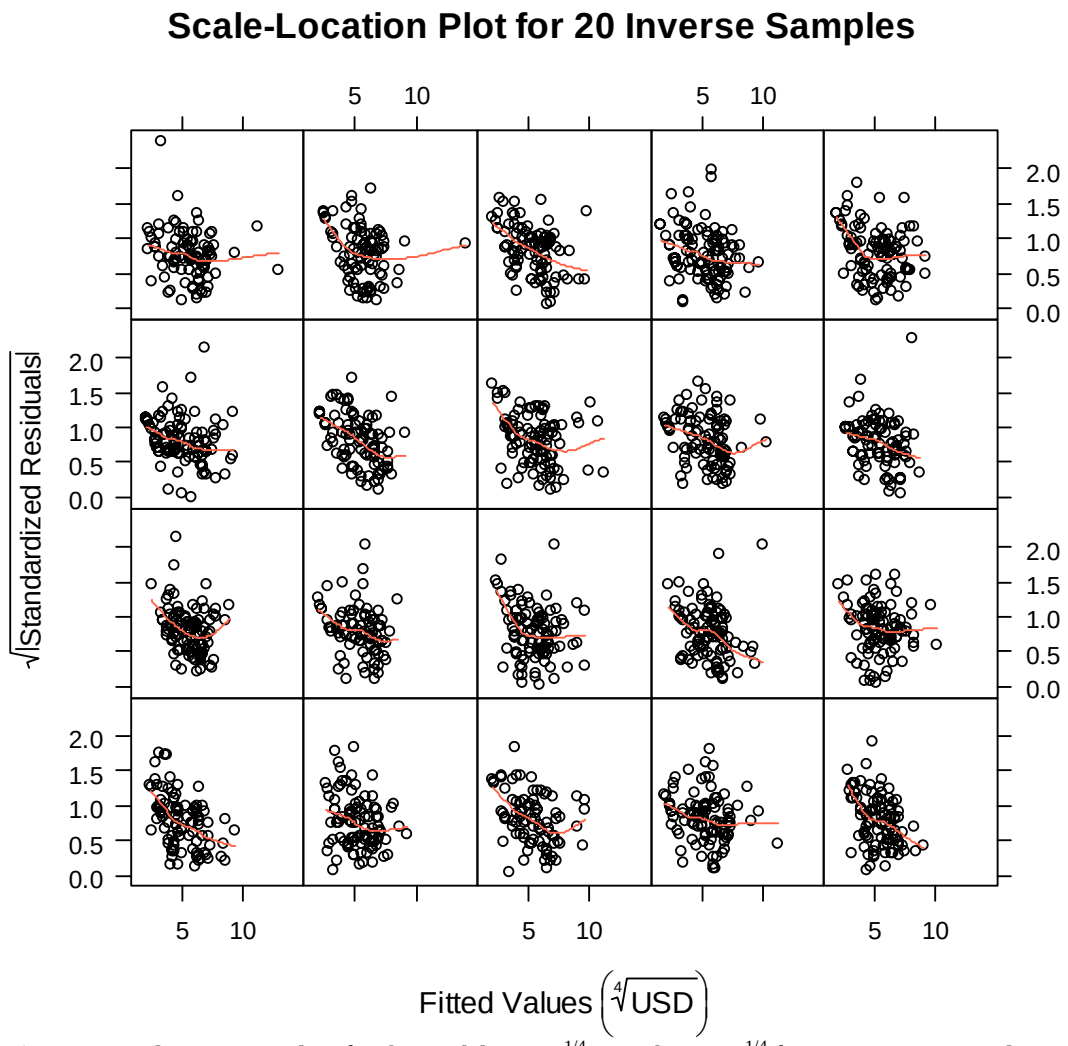

Figure 6: Scale-Location plots for the model $TPAY^{1/4} = a + b \cdot DEBT^{1/4}$ from 20 inverse samples from the SCF 2007 Public Use data set.

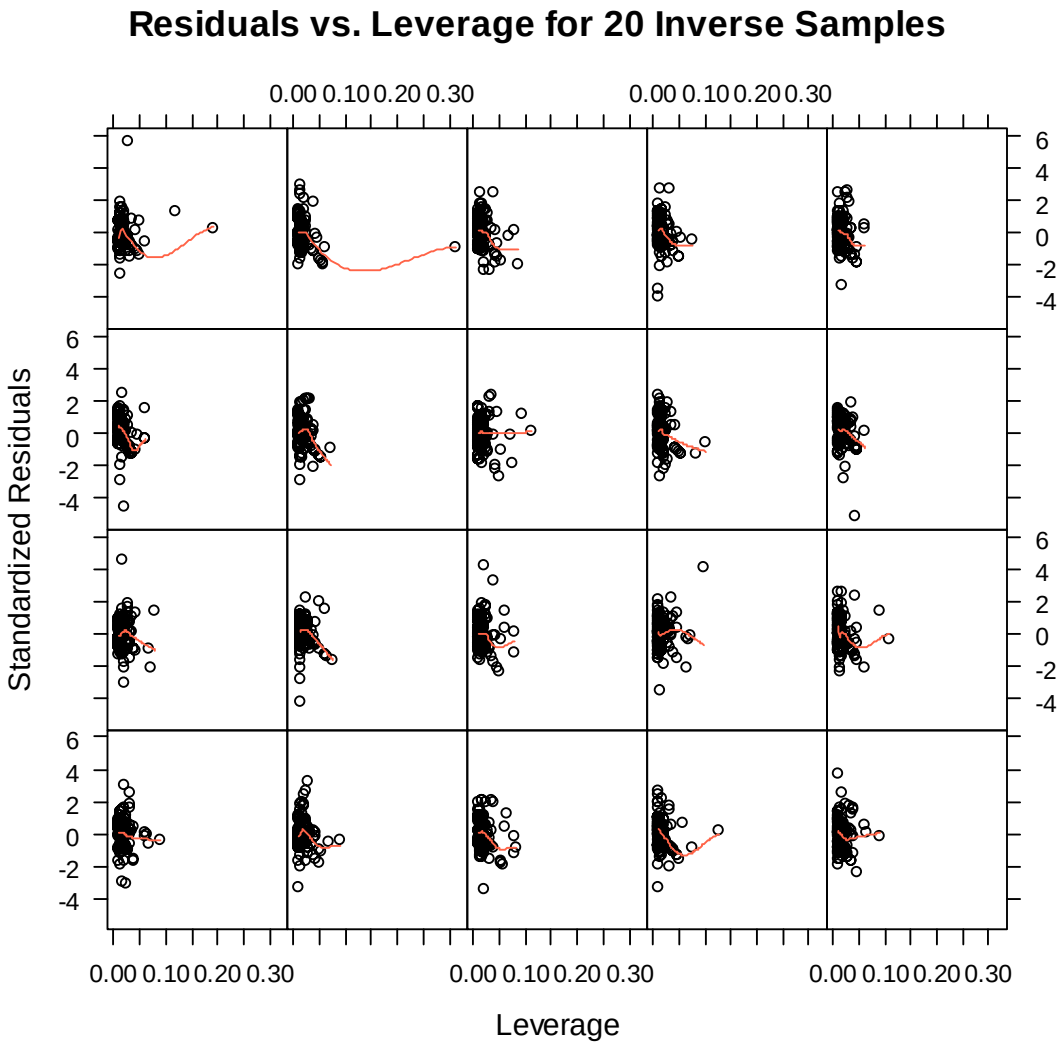

Figure 7: Residual vs. Leverage plots for the model $TPAY^{1/4} = a + b \cdot DEBT^{1/4}$ from 20 inverse samples from the SCF 2007 Public Use data set.

SCF 2007 is a very "rich" data set—if the TPAY\DEBT relationship was truly of interest, a model with additional covariates is probably more appropriate. Also, SCF 2007 is a multiply imputed data set—only the first implicate was used for the illustration. A full analysis would include all five implicates with subsamples from each.

Nevertheless, applying visualization tools using appropriate data—inverse samples with characteristics of simple random samples—shows that the ad hoc approach used by some analysts can lead to fitting inappropriate models. Additionally, diagnostic plots of a model fit to multiple inverse samples gave a similar, overall visualization to the diagnostic plots produced by the R survey package, which incorporated the complex survey design into the background calculations.

4. Conclusions, Caveats, and Future Work

4.1 Conclusion

Correct weighted analysis can be done to calculate regression coefficients and standard errors, and diagnostic plots can be adjusted to include the effect of the sample weights. But from the literature, there is interest in methods of analysis and visualization techniques which can produce valid results without using the sample weights. The user may not have sufficient tools or expertise to use the weighted data correctly or the weighted data may result in visual plots that are difficult to interpret.

One approach is to base the analysis and diagnostic plots on simple random samples drawn from the complex design using inverse sampling methods. For stratified sampling, the inverse algorithm uses straightforward techniques, i.e. selection from a hypergeometric probability distribution and simple random sampling. Alternatively, the data producer could provide such subsamples as part of the data base. This effort may not be any more difficult than producing sets of replicate weights, which is often done for public use files.

In this paper, we have shown an example where this technique provided somewhat different results than the ad hoc technique of comparing OLS and weighted regression. The ad hoc approach indicated that the OLS resulted in the same estimates of coefficients and standard error as the weighted regression. However, the diagnostic plots based on 20 inverse samples showed changes in the spread of the data that were not apparent in the diagnostic plots from the OLS.

Weights matter in regression. They may or may not have an effect on the coefficient estimates and standard error, but we've shown that they are important for judging whether or not the fitted model is appropriate for the data.

4.2 Caveats

We actually used pseudo-inverse samples based on a cluster model[2] of the sample weights because the SCF 2007, like most public use data, does not include all the design information that is needed for inverse sampling. An investigation of this method and alternative synthetic methods are needed.

4.3 Future Work

There is much further work to be done in developing and evaluating visualization and other diagnostic tools for the analysis of complex, weighted sample survey data. The example provided in this paper indicates that the use of inverse samples may be a useful tool. However, there are many open questions to be addressed for this possible methodology. A panel of plots is not as easy to evaluate as a single plot, and in some cases, the sample size of the inverse samples may be too small to make such plots useful. Further work could include a cognitive review of how people interpret scatterplots, and whether multiple plots, such as those shown in this paper, or alternative single scatterplots that incorporate the survey design and weights, e.g. bubble plots, are better for visualizing relationships in complex survey data.

[2] We are not referring to a cluster sample design here, rather a cluster classification algorithm.

We also plan to investigate an aggregation approach where a linear model is fit to a data set created by aggregating multiple inverse samples selected from the complex design. It would be useful to compare this method of aggregating inverse samples to the alternative methodology of creating synthetic data sets by using PPS sampling based on the sample weights to select subsamples from the complex data. In both cases, there are questions of the effect of the correlation due to repeated draws from the original sample. A related issue in both approaches is the determination of the sample size for such a simulated data base.

The application of analytic methods to data obtained from a complex survey design with a primary goal of achieving good enumerative statistic estimates is important in an environment where key policy decisions may be driven by the analysis of such data. It is therefore important that we provide appropriate tools for researchers using these data. Specialized survey software tools are important, but we believe that it is also important to provide users with data that can be analyzed using methods already familiar to the researcher. This includes visualization techniques that are becoming more prevalent given the currently available computing capacity. We hope that the ideas that we have introduced in this paper will bring about more awareness and research in this area.

Acknowledgements

Support for this research was provided by the NORC Center for Excellence in Survey Research (CESR).

References

Asabere , P. K. and Huffman, F. E. (1996). “Negative and positive impacts of golf course proximity on home prices,” *The Appraisal Journal*, pp. 351-355.

Brewer, K. R. W., and Mellor, R. W. (1973), "The Effect of Sample Structure on Analytic Surveys," *Australian Journal of Statistics*, 15, 145-152.

Deepayan Sarkar (2009). lattice: Lattice Graphics. R package version 0.17-25. http://CRAN.R-project.org/package=lattice.

DuMouchel, W., and Duncan, G., (1983). “Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples,” *Journal of the American Statistical Association*, Vol. 78, No. 383, pp. 535-543.

Financial Studies Section, Division of Research and Statistics, (2009). “Codebook for 2007 Survey of Consumer Finances,” Board of Governors of the Federal Reserve System, http://www.federalreserve.gov/pubs/oss/oss2/2007/codebk2007.txt.

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, pp. 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, pp. 11-21.

Holt, D., Smith, T. M. F., and Winter, P. D. (1980), "Regression Analysis of the Data From Complex Surveys," *Journal of the Royal Statistical Society*, Ser. A, 143, 474-487.

Klein, L. and Morgan, J. (1951), “Results of Alternative Statistical Treatments of Sample Survey Data,” *Journal of the American Statistical Association*, 46, pp. 442-460.

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

Kott, P. (1991). "A Model-Based Look at Linear Regression with Survey Data," *The American Statistician*, Vol. 45, No. 2, pp. 107-112 .

Li, J., and Valliant, R. (2006), "Influence Analysis in Linear Regression with Sampling Weights," *2006 Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Li, J., and Valliant, R. (2007), "Linear Regression Diagnostics in Cluster Samples," *2007 Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Li, J., and Valliant, R. (2009a), "Survey Weighted Hat Matrix and Leverages," *Survey Methodology*, to appear.

Li, J., and Valliant, R. (2009b), "Linear Regression Diagnostics for Unclustered Survey Data," Joint Program in Survey Methodology technical report, University of Maryland and University of Michigan.

Lohr, S. (1999). *Sampling: Design and Analysis*. Duxbury Press, ISBN 0-534-35361-4

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

Lumley, T. (2009) "survey: analysis of complex survey samples". R package version 3.16.

Nathan, G. and Holt, D. (1980). "The Effect of Survey Design on Regression Analysis ," *Journal of the Royal Statistical Society*. Series B, Vol. 42, No. 3 (1980), pp. 377-386

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

Pfeffermann, D., and Holmes, D. J. (1985), "Robustness Consider- ations in the Choice of Methods of Inference for Regression Analysis of Survey Data," *Journal of the Royal Statistical Society*, Ser. A, 148, 268-278.

R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, 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.