WEIGHTING IN REGRESSION FOR USE IN SURVEY METHODOLOGY

James R. Knaub, Jr., Energy Information Administration
U.S. Department of Energy, EI-524, Washington, DC 20585

Key Words:
heteroscedasticity, variance of a variance function parameter, establishment surveys, model-based inference, stratification, sampling, imputation, estimation, prediction

Abstract:
Weighted linear regression models have been developed for use in the estimation of totals and variances for survey data. (Consider works by Brewer, and by Royall and Cumberland, et.al.) Weighted linear regression models have also been developed for prediction and variance studies in analyses of physical and biological data. (Consider works by Carroll and Ruppert, et.al.) There are similarities and differences between these approaches. This paper considers this and deduces further implications for survey methodology.

Both real and artificial test data sets are used in the analyses. The artificial data are the result of a process that allows examination of error structures without randomization. This simplifies comparisons and reduces the number of observations needed for testing, thus making the testing process more efficient.

Introduction:
When dealing with highly skewed data, a cutoff, model-based sample may be used to avoid substantial nonsampling error which could have been generated by the smallest entities. Also, a model automatically provides an indication (variance) as to whether imputation may be allowed.

A longer version of this paper is found in the Internet statistics journal, *InterStat*. To obtain that article, enter **http://interstat.stat.vt.edu** and proceed as indicated. This article is found in the area for April 1997. **Note that there is an "errata" file.**

Goals of this paper:
The productive application of weighting in regression to survey methodology is considered here. Two parts to this goal are pursued: (1) an investigation of the accuracy of the estimated degree of heteroscedasticity, and (2) the establishment of guidelines for the practical implementation of weighted regression estimation to survey methodology. To meet the goals of this article, three cases will be described. The first two cases use artificial data, generated to exhibit specific properties relevant to the goals of this article. The third case uses real data. The accuracy with which we measure heteroscedasticity is an integral part of all three of the cases.

Background:
A) Method for Estimating γ, When the Nonrandom Factor of 'Error' is x^{γ} -
Consider the model $y_i - \beta x_i = e_{0_i} x_i^{\gamma}$.

(Note that this methodology may be modified to accommodate other models, including other formats for the nonrandom factor of the residual.) The method below is an alternative to the Iterated Reweighted Least Squares (IRLS) method. The IRLS method can be found in Carroll and Ruppert (1988).

The comparative usefulness of the method below is discussed in Knaub(1993). One of the graphs generated will take on different appearances when some model failures such as nonlinearity are present. (See Knaub(1993) for details.) However, the first two examples in a following section of this paper show that even when this graph indicates no problem, there may still be important, hidden characteristics about the data.

1) Each error is considered as a product of a random factor, e_0, and a nonrandom factor, x^{γ}.

2) Assuming this linear, zero intercept, heteroscedastic model, find $\gamma = w$ where $(y_i - bx_i)/x_i^{w} = e_{0_i}$ are nearly homoscedastic.

3) With the original regressor, x, still on the horizontal axis and $|e_{0_i}|$ on the vertical axis, a fitted homoscedastic, linear regression should have a slope near zero, as there should be no growth trend as x increases.

4) Next, in a new graph, if the absolute values of these slopes are plotted on the vertical axis against gamma values on the horizontal axis, points where the plotted line contacts the horizontal axis would correspond to values for γ that make the model consistent with the data.

B) A Method for Estimating the Standard Error of the Estimate of γ -
Methods employing the logarithm of absolute residuals are among those described in Carroll and Ruppert(1988). This was also one of the methods suggested to me by Ken Brewer. It seems aesthetically pleasing here:

Consider the model $y_i - \beta x_i = e_{0_i} x_i^{\gamma}$.

Taking the logarithm of the absolute values of the residuals, when x is always positive, and writing this as a regression equation, we obtain

$$\log|y_i - \beta x_i| = \log(E|e_{0_i}|) + \gamma \log x_i + e_{0_i}'(\log x_i)^{\gamma'},$$

which we will rewrite as

$$y'_i = \alpha' + \beta' x'_i + e_{0_i}' x'^{\gamma'}_i .$$

If we plot $\log|y_i - \beta x_i|$, *or* y_i', on the vertical axis, and $\log x_i$, *or* x_i', on the horizontal axis, then the slope is an estimate of γ. The advantage of obtaining this estimate of γ, is that it is then easy to obtain an estimate of the standard error for γ. ***Any estimated standard error for the estimate of*** β' ***now becomes an estimated standard error for the estimate of*** γ, ***because*** $\beta' = \gamma$.

C) Method Used to Construct Artificial Data:
Points symmetric in the dimension of the y-axis about a straight line are used to represent the increase in variance associated with a given nonrandom factor of error (i.e., nonrandom factor of the residual), of the form x^{γ} (or any other form chosen). This methodology functions as if one were to have drawn confidence interval contours about a line, considering a given degree of heteroscedasticity. (See Knaub(1995a).)

D) Another Form for the Nonrandom Factor of the Residuals:
Real establishment survey data investigated by this author have never been shown to support a more elaborate model of error than the standard one in which the nonrandom factor of the residuals is of the form x^{γ}. Perhaps this is because the data sets used were generally not large. Note, however, that in Steel and Fay (1995), such a model was found for their data. To use the methodology of A) above, instead of looking for the one value w, three values would need to be found simultaneously for their model. The graphs described in A) above could be made by holding two values constant and searching for a better value for the third, given the other two, in an iterative process, looking for the 'best' set of values to use.

Examples Illustrating Possible Usefulness of Proposing Guidelines:
The following examples illustrate that although x^{γ} is a very useful form for the nonrandom factor of the residuals, and we may find a value for γ that appears to perform very well, there may be some improvements one can make. In the first example, we see that in a case where we could have stratified had we known more about the data, there may be no indication of this when we treat the data as a homogeneous group with inflated heteroscedasticity. Resulting estimates of total and its variance may still be very good, but perhaps results would be better had we known how to stratify those data. In the second example, we see a case where the value of γ is a step function of x. Perhaps this might perform well in the case of the data used in Steel and Fay(1995). Similarly, in Knaub(1995b): "Dr. Nancy Kirkendall, EIA/OSS, once noted an article by Karmel and Jain (1987), which suggests modeling strata delineated by size of the regressor values."

The first two examples used artificial data to isolate specific phenomena. In example 3, real data are used. These data constitute a strata of data found in Knaub(1996). Stratification criteria were thus employed, and the estimated value of γ for this more homogeneous set of data was found to be smaller, as example 1 would indicate. With such 'real' data, however, other factors could be present.

Example 1:
This example uses artificial data with a single regressor, and zero-intercept. The data set actually consists of three sets of data, each having $\gamma = 0$, but each with a different value for β. These values for β are 0.8, 0.85 and 0.9. See Figure 1.1. This represents a case where stratification should have been used, but perhaps there was too little information to do so.

Figure 1.1

Example Using Artificial 'Test' Data

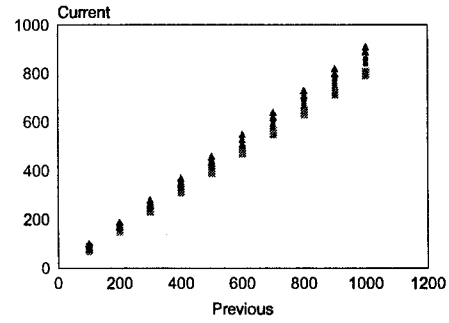
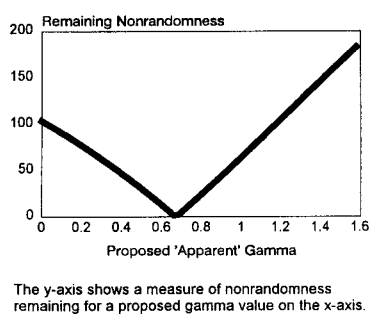


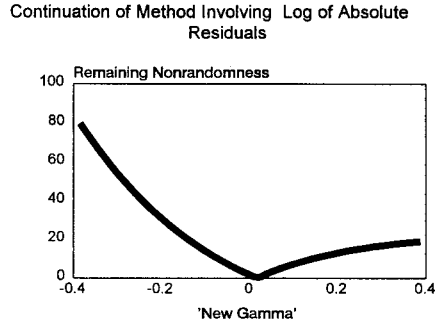
Figure 1.2 shows an approximation for γ. Here, $\gamma^* = 0.67$. (See "Background," section A) 4) above.)

Figure 1.2
Heterogeneous 'Test' Population



The y-axis shows a measure of nonrandomness remaining for a proposed gamma value on the x-axis.

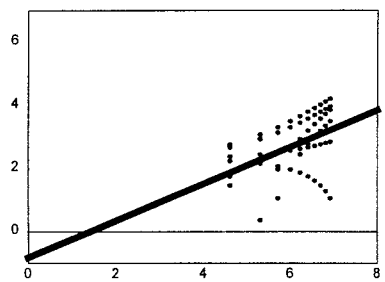
Figure 1.4
Heterogeneous Test Population



This is to investigate heteroscedasticity in the model used to study variance of the 'original' gamma.

Figures 1.3 and 1.4 illustrate the method for estimating the standard error of an estimate of γ by first (Figure 1.3) plotting $\log|y_i - \beta x_i|$ on the vertical axis and $\log x_i$ on the horizontal axis.

Figure 1.3
Heterogeneous Test Population
Method Involving Logarithm of Absolute Residuals



The slope here corresponds to gamma in Figure 1.2.

The slope in Figure 1.3 is an estimate of γ. In this case, the estimate of γ and its standard error are 0.62 and 0.10, respectively.

Notice from Figure 1.4 that we have a nearly homoscedastic situation in this instance.

A cursory look at a few examples seems to indicate that γ' is often near zero or negative. However, results may not be highly dependent upon choice of γ', at any rate.

Example 2:
This example also uses artificial data with a single regressor and zero-intercept. The data set actually consists of two sets of data, each with the same value for β. One data subset was designed with $\gamma = 1$, and the other with $\gamma = 0.5$. (This is shown in Knaub(1997), Figure 2.0 and Figure 2.1.) This represents a case where the nonrandom factor of "error" is not represented best by x_i^{γ}, using only one value of γ. However, we can estimate a value for γ that ***appears*** to be indicated by the data, overall. Figure 2.2 in Knaub(1997) shows an approximation for γ to be 0.85. Figures 2.3 and 2.4 in that article illustrate the method for estimating the standard error of an estimate of γ. First, a plot of $\log|y_i - \beta x_i|$ (vertical axis) and $\log x_i$ (horizontal axis) is made. The slope is an estimate of γ. Estimates of γ and its standard error follow as 0.88 and 0.01, respectively. ***One might then be quite convinced of the correctness of the estimate of gamma, not suspecting the true nature of the data.***

Example 3:
The final example employs naturally occurring data with two regressors. The model, $y_i = \beta_1 x_i + \beta_2 c_i + e_{0_i} c_i^{\gamma}$, was used. (For graphs, again see Knaub(1997).) The size of the data set is n = 642. (This is unusually large in this author's experience.) An approximate value for γ was found to be $\overset{*}{\gamma} = 0.82$. This value for γ is indicated for heteroscedasticity ***with respect to*** x.

For heteroscedasticity with respect to any regressor, first estimate the values for the proposed random factor of the residuals by dividing the residuals by the proposed nonrandom factor. Next check whether the resulting estimated e_{0_i} values actually are nearly random, with respect to that regressor. This would be covered in A) 3) under "Background" earlier in this paper. In this example, for heteroscedasticity with respect to x, $\gamma^* = 0.82$. For heteroscedasticity with respect to the other regressor, c, $\gamma^* = 0.96$. Here, x is the more influential regressor with regard to estimating population totals, as, in general, $\beta_1 x_i > \beta_2 c_i$. With respect to x, estimates of γ and its standard error are 0.82 and 0.03, respectively.

Concluding remarks for this example:
Consider: "Are there enough data to decide we should use $\gamma = 0.82$, or should we use something more robust?"

For highly skewed, stratified, establishment surveys, the least aggregate numbers estimated may be represented by a few of the largest entities. If so, using $x_i^{1/2}$ as the nonrandom factor of the residuals may often perform well. This is often the case when the regressor is the same variate that is represented by y in the sample, but from a previous census. In example 3, however, there is a second regressor, and this may change the situation appreciably. Using $\gamma = 0.5$ yields a larger effective sample size than does using $\gamma = 0.82$. (Only in the case where $\gamma = 0$ (ordinary least squares) are all data points weighted equally.)

In this two-regressor example there appear to be enough 'well-behaved' data (n = 642) that we may use the estimated value, $\gamma = 0.82$. Consider the following: The estimated standard error for the estimate of γ is small. The estimation of γ based on the methodology found in "A)" under "Background" indicates no problem, and stratification has been applied to some extent, the data consisting of one of those strata. Thus the value for γ estimated from these data may be useful.

To review:
Examples 1 and 2, illustrate that there may be considerations other than the standard error of the estimate of γ, and the graphical analysis of Knaub(1993) as explained under "Background" above, when contemplating regression weights for use in survey methodology. It may well be that such stratification and model failure considerations as were shown in examples 1 and 2 may not greatly impact on the estimation of totals and their variances if we were to remain ignorant of these phenomena. For example, using $\gamma = 0.5$ may be very robust for estimating totals; using the estimated γ may be excellent for imputation in general; but for a better localized imputation, and for data analyses as in Carroll and Ruppert(1988), the more detailed examination may be imperative.

In example 3, perhaps further stratification would be desirable, but the mechanism for doing this may be unknown. Perhaps γ could vary as x and/or c increase(s). In the case of cutoff model-based sampling, for possibly combined strata which we do not know how to separate, perhaps $\gamma = 1$ is a better default value than $\gamma = 0.5$. That would, for instance, prevent one strata with the largest few data points from unduly influencing the estimation of a large number of smaller data points. If several strata have been combined, then perhaps there would be enough data points to allow the use of the larger value of γ, which affectively reduces the sample size. Because example 3 actually represents a case where some stratification had already been accomplished, one may decide to use $\gamma = 0.8$.

Survey Methodology vs. Analytical Use of Weighting in Regression:
The use of weighted linear regression models in survey methodology may generally, to date, have been more conservative than in data analyses, such as those found in Carroll and Ruppert(1988). In survey methodology, models have often been used less for exploring survey data and more for summarizing. However, model use in imputation helps 'bridge that gap,' and there is some history of exploratory use, as may be seen in Cochran(1977). Sarndal, Swensson and Wretman (1992) exemplifies some more recent analytical use of models in survey methodology as does Sweet and Sigman (1995). Steel and Fay (1995), in their survey methodology, use a relatively elaborate model, which may be more consistent with Carroll and Ruppert (1988) than with Brewer (1963), Royall (1970), Royall and Cumberland (1981), or even more recent survey methodology works by these and/or other authors either. Still, as indicated above, more attention seems to be paid to models in survey methodology in recent years. Note, for example, that Chaudhuri and Stenger (1992) gives a fairly balanced view of survey sampling (ie, design-based and model-based survey sampling), and Brewer (1995) provides more insight.

Although the goals of survey work and analytical work may keep the methods somewhat distinct, more seems to be borrowed from analytical work as the field progresses.

Still, cutoff, model-based sampling for relatively small establishment surveys would seem to generally require less elaborate models than perhaps some imputations would, and analytical work may make use of even more elaborate models.

Some Guidelines/Considerations:
In light of the examples given, when applying weighted regression estimation for survey sampling, consider the following:

1) Could stratification be needed?
2) Could gamma vary, or is x^{γ} even a useful format to apply as the nonrandom factor of the residuals?
3) Are there enough data to estimate the nonrandom factor of 'error,' or should we use a robust default such as $x^{0.5}$?

Acknowledgment:
Thanks to K.R.W. Brewer for helpful discussions, although he can in no way be considered responsible for any errors or inadequacies in this paper.

References:

Brewer, K.R.W. (1963), "Ratio Estimation in Finite Populations: Some Results Deducible from the Assumption of an Underlying Stochastic Process," Australian Journal of Statistics, 5, pp. 93-105.

Brewer, K.R.W. (1995), "Combining Design-Based and Model-Based Inference," Business Survey Methods, ed. by B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott, John Wiley & Sons, pp. 589-606.

Carroll, R.J., and Ruppert, D. (1988), Transformation and Weighting in Regression, Chapman & Hall.

Cochran, W.G. (1977), Sampling Techniques, 3rd ed., John Wiley & Sons.

Chaudhuri, A. and Stenger, H. (1992), Survey Sampling: Theory and Methods, Marcel Dekker, Inc.

Karmel, T.S., and Jain, M. (1987), "Comparison of Purposive and Random Sampling Schemes for Estimating Capital Expenditure," Journal of the American Statistical Association, American Statistical Association, 82, pp. 52-57.

Knaub, J.R., Jr. (1992), "More Model Sampling and Analyses Applied to Electric Power Data," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 876-881.

Knaub, J.R., Jr. (1993), "Alternative to the Iterated Reweighted Least Squares Method: Apparent Heteroscedasticity and Linear Regression Model Sampling," Proceedings of the International Conference on Establishment Surveys, American Statistical Association, pp. 520-525.

Knaub, J.R., Jr. (1994), "Relative Standard Error for a Ratio of Variables at an Aggregate Level Under Model Sampling," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 310-312.

Knaub, J.R., Jr. (1995a), "A New Look at 'Portability' for Survey Model Sampling and Imputation," Proceedings of the Section on Survey Research Methods, Vol. II, American Statistical Association, pp. 701-705.

Knaub, J.R., Jr. (1995b), "Planning Monthly Sampling of Electric Power Data for a Restructured Electric Power Industry," Data Quality, 1, pp. 13-20.

Knaub, J.R., Jr. (1996), "Weighted Multiple Regression Estimation for Survey Model Sampling," InterStat, May 1996, http://interstat.stat.vt.edu/InterStat. (Note shorter, more recent version in ASA Survey Research Methods Section proceedings, 1996.)

Knaub, J.R., Jr. (1997), "Weighting in Regression for Use in Survey Methodology," InterStat, April 1997, http://interstat.stat.vt.edu/InterStat. (Note that this article is a longer version of the current paper.)

Royall, R.M. (1970), "On Finite Population Sampling Theory Under Certain Linear Regression Models," Biometrika, 57, pp. 377-387.

Royall, R.M. and Cumberland, W.G. (1981), "An Empirical Study of the Ratio Estimator and Estimators of its Variance," Journal of the American Statistical Association, 76, pp.66-88.

Sarndal, C.-E., Swensson, B. and Wretman, J. (1992), Model Assisted Survey Sampling, Springer-Verlag.

Steel, P. and Fay, R.E. (1995), "Variance Estimation for Finite Populations with Imputed Data," Proceedings of the Section on Survey Research Methods, Vol. I, American Statistical Association, pp. 374-379.

Sweet, E.M. and Sigman, R.S. (1995), "Evaluation of Model-Assisted Procedures for Stratifying Skewed Populations Using Auxiliary Data," Proceedings of the Section on Survey Research Methods, Vol. I, American Statistical Association, pp. 491-496.