Caveats for the Statistical Consumer: A Cautionary Note on Case Weighting as a Method of Implementing Weighted Least-Squares Regression

JOHN B. WILLETT
JUDITH D. SINGER
Harvard University

ABSTRACT. When no dedicated weighted least-squares procedure is available, misapplying a case-weighting strategy as a method of implementing a weighted least-squares regression analysis can lead to gross inaccuracies. The magnitudes of many of the obtained estimates depend strongly on the absolute magnitudes of the weights applied during fitting and, in addition, several of the crucial regression estimates are incorrect. Among all possible rescalings, the most successful weights are those that have been rescaled so that they sum to the original sample size. However, even with the application of these rescaled weights, the estimation of the error (residual) variance continues to be incorrect. A simple and easily applied adjustment to rectify this problem is presented.

AS MOSTELLER AND TUKEY (1977, p. 346) suggest, the action of assigning "different weights to different observations, either for objective reasons or as a matter of judgement" in order to recognize "some observations as 'better' or 'stronger' than others" has an extensive history. In a regression analysis, whether the investigator wishes to downplay the importance of data points that are intrinsically more variable at specific levels of the predictors or simply to decrease the effect on the fit of remote data points, the use of a weighted, rather than ordinary, least-squares fit is recommended. The ultimate objective of this procedure is to achieve a superior fit in that "while the [ordinary] least-squares estimates and fit may be satisfactory, the precision of the [ordinary] least-squares estimates may be different from that indicated under standard assumptions" (Cox & Snell, 1981, p. 83).

In this paper, we discuss the serious problems of estimation and interpretation that arise when a statistical package, which does not incorporate a

dedicated weighted least-squares (WLS) routine, is tricked into performing WLS regression by the misapplication of a case-weighting strategy. When the case-weighting strategy is used, we show that important regression estimates may be incorrect. We also show that these estimates are not invariant under multiplication of the weights by a constant. This latter problem raises questions as to how the case weights should be selected in practice and how one particular scaling of the weights can be considered optimal. We demonstrate that, among all possible scalings, the performance of one particular set of case weights is superior, although not perfect. Finally, a simple strategy is offered for adjusting those WLS regression estimates that remain incorrect after the "optimal" case-weighting strategy has been applied.

Weighted Least-Squares Regression Analysis

Consider that observations Y_i and X_i on two related variables have been obtained for a random sample of n independent subjects and that the population relationship between these two variables is given by:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad [1]$$

where values of X are fixed, and β_0 and β_1 are the unknown intercept and slope parameters that are to be estimated.[1] Furthermore, assume that the unobserved random errors ϵ_i are normally distributed with zero mean and variance $\sigma_\epsilon^2 k_i$. Thus, providing that the k_i are not all equal, the random errors are heteroscedastic and an ordinary least-squares (OLS) regression strategy for estimating β_0, β_1, and σ_ϵ^2 will be inefficient. Efficient estimation requires the use of WLS regression with weights w_i, which are the inverse of k_i.

Providing the w_i are known, the more efficient WLS estimates of β_0 and β_1, their sampling variances, and σ_ϵ^2 can be obtained by direct minimization of the sum of the squared weighted residuals (for instance, see Neter et al., 1985, pp. 167–170). Equations for these estimates are presented in Table 1. Notice that the WLS estimators in Table 1 are equivalent to the corresponding OLS estimators, except that the WLS results have been obtained in a transformed "world" in which each point in the data set has been weighted by the appropriate w_i.

In practice, the k_i (and hence the w_i) are usually not known but can be inferred from a "combination of prior knowledge, intuition, and evidence" (Chatterjee & Price, 1977, p. 101). A variety of techniques have been proposed for the empirical selection of the weights, ranging from strategies that incorporate substantive knowledge of the form of the residual variance as a function of the predictors (Miller, 1986, pp. 207–214) to two-stage strategies in which an initial unweighted analysis is used to inform the selection of weights (for instance, biweighting in Mosteller & Tukey, 1977, pp. 346–371).

TABLE 1
Selected Weighted Least-Squares Regression Estimators (All Summations Taken Over the Index $i = 1, \ldots, n$)

Estimate	Weighted least-squares estimator	Equation number
$\hat{\beta}_0$	$\dfrac{(\Sigma w_i \hat{Y}_i) - \hat{\beta}_1(\Sigma w_i X_i)}{(\Sigma w_i)}$	a
$\hat{\beta}_1$	$\dfrac{(\Sigma w_i)(\Sigma w_i X_i Y_i) - (\Sigma w_i X_i)(\Sigma w_i Y_i)}{(\Sigma w_i)(\Sigma w_i X_i^2) - (\Sigma w_i X_i)^2}$	b
R^2	$1 - \left[\dfrac{\Sigma w_i (Y_i - \hat{Y}_i)^2}{\Sigma w_i \quad Y_i - \left[\dfrac{\Sigma w_i Y_i}{\Sigma w_i}\right]^2} \right]$	c
$\hat{\sigma}_\epsilon^2$	$\dfrac{\Sigma w_i (Y_i - \hat{Y}_i)^2}{n - 2}$	d
$SE(\hat{\beta}_0)$	$\hat{\sigma}_\epsilon \left[\dfrac{(\Sigma w_i X_i^2)}{(\Sigma w_i)(\Sigma w_i X_i^2) - (\Sigma w_i X_i)^2} \right]^{1/2}$	e
$SE(\hat{\beta}_1)$	$\hat{\sigma}_\epsilon \left[\dfrac{(\Sigma w_i)}{(\Sigma w_i)(\Sigma w_i X_i^2) - (\Sigma w_i X_i)^2} \right]^{1/2}$	f

NOTE: $\hat{Y}_i$ is the predicted value of Y_i obtained in the WLS analysis.

Multiplying the Weights by a Constant

Some of the WLS regression estimates are invariant under multiplication of all the w_i simultaneously by the same numerical constant. For instance, the estimation of $\hat{\beta}_0$, $\hat{\beta}_1$, and R^2 (Equations a, b, and c of Table 1) is not affected by the multiplication of the w_i by an arbitrary constant because of cancellation of the constant in the numerators and denominators of these equations. For these estimators, the absolute magnitude of the weights is unimportant.

On the other hand, in the estimation of the mean-square error (Equation d of Table 1), no such cancellation occurs, and the estimation depends on the scaling of the weights. If all the weights are doubled, then the estimated mean-square error is quadrupled. This is not entirely unexpected because the mean-square error is being estimated in the metric of a transformed world, and this metric is affected by the application of an arbitrary multiplier. Nevertheless, the ad hoc inflation of the error estimates by the arbitrary manipulation of scale is somewhat disconcerting in the sense that

what is being estimated here—σ_ϵ^2, a population parameter of fixed value—is not fluctuating with the selection of arbitrary global magnitudes for the weights.

Finally, although the root mean-square error appears as a multiplier in expressions for the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$, these estimates of precision are not affected by the rescaling of the w_i. Even though σ_ϵ^2 may double when the weights are arbitrarily doubled, inspection of Equations e, f, and d reveals that the multiplying constant cancels out, leaving the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$ unchanged. Notice, however, that if there is a failure of the estimation of σ_ϵ^2 for some reason, then the standard errors will also be incorrect.

Misusing Case Weights to Implement WLS Regression Analysis

Some computer packages (e.g., SAS) possess dedicated procedures for performing WLS regression by applying the equations of Table 1 directly. If you use such packages in your work, then you need read no further because the estimates you obtain will always be correct.

Other computer packages (e.g., SPSSX) do not possess dedicated WLS routines but do possess "case-weighting" commands to facilitate the statistical analysis of frequency (grouped) data by arithmetically replicating individual cases in the data set (SPSSX, 1986, p. 185). If you make a habit of misusing a case-weighting facility to implement WLS regression, then you may be in serious trouble because many of your regression statistics will be incorrect. In the following we illustrate some of the pitfalls that await you.

The Empirical Weights

Using case weighting (and SPSSX regression), the statistical model in Equation 1 was fitted to the heteroscedastic data in Figure 1.[2] These data display classic heteroscedasticity, the ϵ_i being drawn from a population in which the residual variance is directly proportional to X_i^2. An additional variable was created with the COMPUTE statement to contain the regression weights (see below), and the WEIGHT command was used to misinform SPSSX that this variable contained case weights. Then, a case-weighted OLS regression was performed on the new data set.

For the purposes of the current demonstration, three sets of weights were created. Each of these sets of weights was appropriately proportional to the squared inverse of the value of the independent variable. The three sets of weights differed only in their scale—any one set of weights being simply a constant multiple of either of the other sets of weights. Thus, the weighting schemes included the basic set of weights:

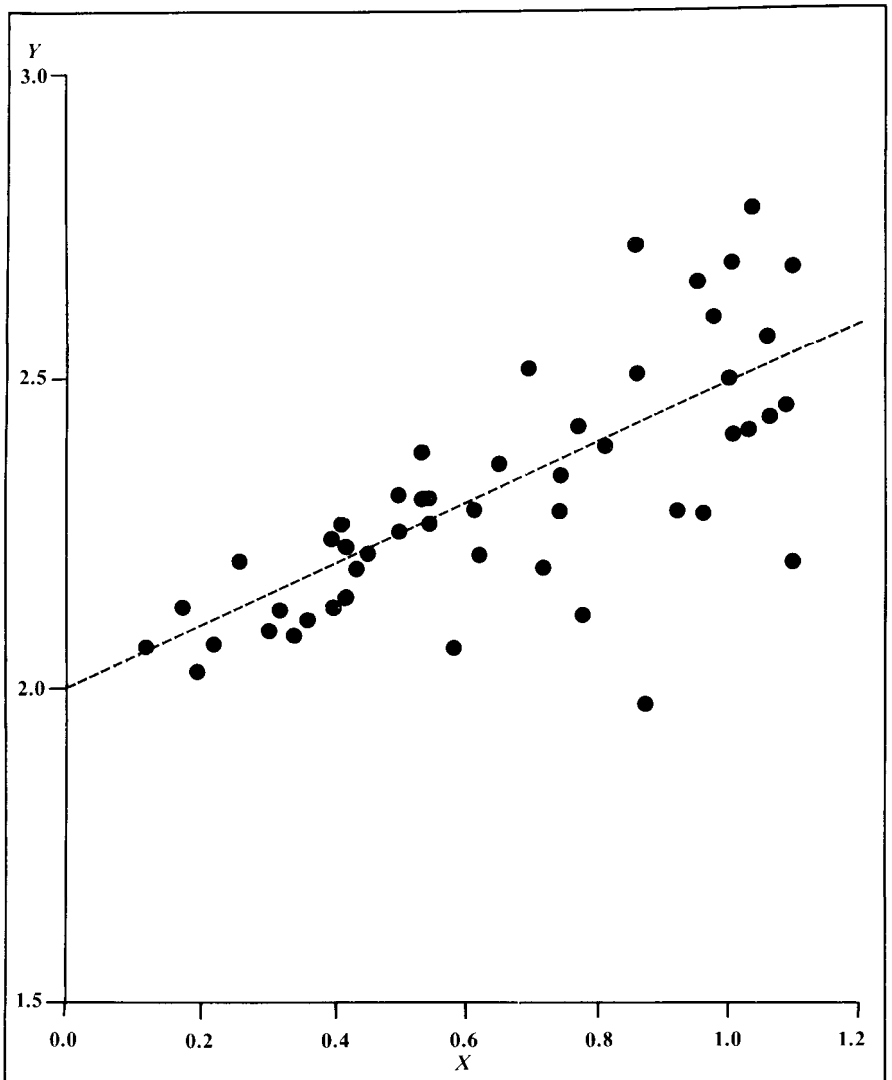

FIGURE 1. Bivariate scatterplot of the artificial data set. Values of the dependent variable Y_i plotted against values of the independent variable X_i, for $i = 1, \ldots, n$. The location of the underlying population regression line is shown as a broken line.

$$w_{1i} = \left[\frac{1}{X_i^2}\right]. \qquad [2]$$

A set in which each weight was double the corresponding weight in Equation 2:

$$w_{2i} = 2w_{1i} \qquad [3]$$

and a set for which the sum of the weighted number of cases equals the original sample size:

$$w_{3i} = \left[\frac{n}{\sum_{i=1}^{n} w_{1i}} \right] w_{1i}. \qquad [4]$$

This latter strategy is suggested by the *SPSSX User's Guide* as an appropriate way of avoiding the artificial inflation (or deflation) of significance tests when the weighted number of cases exceeds (or is less than) the original sample size (1986, p. 186; see also Moser & Kalton, 1972). Unfortunately, it is a strategy that does not ensure appropriate estimation when case weighting is being misapplied in the implementation of WLS regression.

Examining the Estimates

The fitting of the statistical model in Equation 1 was carried out three times, once for each of the three sets of weights. The obtained fits are summarized in Table 2. Also included in this table are estimates obtained by applying the equations of Table 1 directly.

Notice that, even though all three sets of weights are equally acceptable theoretically, there is considerable disagreement among the outcomes of the parallel analyses. Under any of the sets of case weights, the estimated intercept, slope, and coefficient of determination are always correct,[3] but other estimates—including the standard errors of the intercept and slope and the estimated error variance—are less fortunate. Only under the third set of weights is a reasonable level of success achieved and, even then, there is a serious problem with the computation of the residual variance (shown in italics in Table 2).

The principal objective of WLS regression, applied in the context of heteroscedastic errors, is to obtain superior estimates of the precisions of $\hat{\beta}_0$ and $\hat{\beta}_1$. In this context, although the correct estimates were obtained under the w_{3i}, it is disturbing that the magnitude of the standard errors does depend on which particular set of weights was applied. Obviously, in practice, if case weighting is applied, then it is only the third set of case weights that has any validity. This is particularly disconcerting because it is the w_{1i} that would be the natural first choice of the data analyst. This fluctuation in the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$ is due to the failure of the case-weighting strategy to correctly enumerate the error degrees of freedom in the estimation of σ_ε^2.

In Table 2, the degrees of freedom associated with the error sums of squares varies with the particular set of weights applied. Thus, with the w_{1i}, the estimation has been performed under the misapprehension that there were 306 subjects in the sample rather than 50 (2 degrees of freedom being

TABLE 2
Summary Statistics Obtained in the Case-Weighted Regressions

	Estimates obtained under case-weighting strategy			Correct estimate
	w_1	w_2	w_3	
$\hat{\beta}_0$	**1.9977**	**1.9977**	**1.9977**	**1.9977**
$\hat{\beta}_1$	**.4751**	**.4751**	**.4751**	**.4751**
R^2	**.6737**	**.6737**	**.6737**	**.6737**
$SE(\hat{\beta}_0)$	.0077	.0038	**.0193**	**.0193**
$SE(\hat{\beta}_1)$	.0190	.0095	**.0477**	**.0477**
dfE	304	1222	**48**	**48**
$\hat{\sigma}_\epsilon^2$	.0060	.0060	*.0062*	**.0381**

NOTE: Correct values are printed in boldface.

"added back in" to account for the two regression coefficients that have been estimated). With the w_{2i}, more than 1,000 additional data points have apparently joined the existing point cloud. It is only when the third set of weights—which have been constrained to sum to the original sample size—is applied that the error degrees of freedom become correct. This fluctuation in the degrees of freedom under different weight rescalings is a direct consequence of the analyst's misuse of case weighting, and it is a fluctuation that, except in this application, is both intended and required.

In fact, rather than Equation d of Table 1, the case-weighted estimator being applied in the estimation of σ_ϵ^2 is:

$$\hat{\sigma}_\epsilon^2 = \frac{\Sigma w_i (Y_i - \hat{Y}_i)^2}{\Sigma w_i - 2}, \qquad [5]$$

where the sum of the weights has replaced true sample size in the denominator. Under the case-weighting strategy, the numerator of this new estimator will only be computed appropriately when the first set of weights (w_{1i}) is applied, whereas the denominator will only be correct when the third set of weights (w_{3i}) is applied. Hence, the numerator and the denominator can never be correct simultaneously, and whenever WLS regression is being performed by the misapplication of case weighting, σ_ϵ^2 will never be estimated correctly.

Nevertheless, this failure can be easily rectified by simply adjusting the estimate of σ_ϵ^2 obtained under the w_{3i}. Then, an unbiased estimator of the error variance is given by:

$$\hat{\sigma}_\epsilon^2 = \left[\frac{\Sigma w_{1i}}{n}\right] \times \left[\frac{\Sigma w_{3i}(Y_i - \hat{Y}_i)^2}{\Sigma w_{3i} - 2}\right], \qquad [6]$$

and the estimate of σ_ϵ^2 obtained under w_{3i} can be corrected by multiplying by the factor $(\Sigma w_{1i}/n) = (306/50)$ to give .0381, a correct value.

Recommendations

When no dedicated WLS procedure is available, misapplying a case-weighting strategy as a method of implementing a weighted least-squares regression analysis can lead to gross inaccuracies. The magnitudes of many of the obtained estimates depend strongly on the absolute magnitudes of the weights applied during fitting, and, in addition, several of the crucial regression estimates may be incorrect. Nevertheless, among all possible rescalings, the most successful weights are those that have been rescaled so that they sum to the original sample size. However, even the application of these rescaled weights is not entirely without problem because the estimation of σ_ϵ^2 continues to be flawed. To rectify this problem, the simple and easily applied adjustment presented in the paper should be applied.

We would like to stress that we have considered relatively few of the statistics that are commonly interpreted in a typical regression analysis and only one of the many statistical software packages available. Nevertheless, the findings are symptomatic of similar problems that are likely to occur when more complex models are fit and interpreted and other computer packages are used. In particular, the investigator must validate the WLS algorithm of the specific software being applied—a validation that may be of particular importance when one of the newer microcomputer statistical packages is the software of choice. Furthermore, although we have not investigated the manner in which more sophisticated statistics such as Mallow's C_p, Cook's D, and other measures of influence are affected by an arbitrary rescaling of the weights, it is appropriate to advise great caution in interpretation here too. Attempting to carry out weighted least-squares regression analysis by the blind application of case weighting would certainly seem to be a case of *caveat emptor.*

AUTHORS' NOTE

The order of the authors has been determined by randomization. An earlier version of this paper was presented at the American Educational Research Association Conference, Washington, D.C., April 1987. The authors would like to thank an anonymous reviewer for careful and focused comments made on that version.

NOTES

1. Although the results of this paper are easily generalizable to the multiple predictor case, the discussion presented here deals with the estimation of the relationship between the dependent variable and a single predictor.

2. A bivariate sample of 50 observations on the pair of variables (Y_i, X_i) generated such that, in the population, $\beta_0 = 2$ and $\beta_i = .5$ with heteroscedastic random errors ϵ_i drawn from a normal distribution with zero mean and variance $0.04X_i^2$.

3. For other reasons, although the R^2 estimates in Table 2 all agree and are computationally correct regardless of the particular weighting scheme applied, the obtained value may not be appropriate for the interpretation of empirical goodness of fit (see Willett & Singer, 1988).

REFERENCES

Chatterjee, S., & Price, B. (1977). *Regression analysis by example.* New York: Wiley.

Cox, D. R., & Snell, E. J. (1981). *Applied statistics: Principles and examples.* London: Chapman & Hall.

Miller, R. G., Jr. (1986). *Beyond ANOVA, basics of applied statistics.* New York: Wiley.

Moser, C. A., & Kalton, G. (1972). *Survey methods in social investigation* (2nd ed.). New York: Basic Books.

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression.* Reading, MA: Addison-Wesley.

Neter, J., Wasserman, W., & Kutner, M. H. (1985). *Applied linear statistical models: Regression, analysis of variance, and experimental designs.* Homewood, IL: Richard D. Irwin, Inc.

SPSSX Inc. (1986). *SPSSX user's guide* (2nd ed.). New York: McGraw-Hill.

Willett, J. B., & Singer, J. D. (1988). Another cautionary note about R^2: Its use in weighted least-squares regression analysis. *The American Statistician, 42,* 236–238.