DOCUMENT RESUME

| | | |
|---|---|---|
| ED 282 906 | | TM 870 317 |


AUTHOR Willett, John B.; Singer, Judith D.

TITLE Interpreting the Results of Weighted Least-Squares Regression: Caveats for the Statistical Consumer.

PUB DATE Apr 87

NOTE 29p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987).

PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS Error of Measurement; Estimation (Mathematics); *Goodness of Fit; *Least Squares Statistics; *Mathematical Models; *Predictor Variables; *Regression (Statistics); Scaling; Statistical Studies

IDENTIFIERS *SPSS Computer Program; *Weighted Data



ABSTRACT

In research, data sets often occur in which the variance of the distribution of the dependent variable at given levels of the predictors is a function of the values of the predictors. In this situation, the use of weighted least-squares (WLS) or techniques is required. Weights suitable for use in a WLS regression analysis must be estimated. A variety of techniques have been proposed for the empirical selection of weights with the ultimate objective being a better "fit." The outcomes of the analysis must be interpreted once the fitting is complete. Problems can arise in the interpretation of some of the statistics when using a computer package. In this paper, such problems in the application and interpretation of WLS regression using the SPSS statistical package are demonstrated, both algebraically and by example. For the purposes of the example, an artificial data set (whose underlying parametric structure is known) has been created. Each of the statistics commonly reported in the WLS regression analysis of such a data set are isolated and their interpretation discussed. Where necessary, adjusted statistics that more reasonably represent the outcomes of the analysis are proposed and their use illustrated. (BAE)

Working paper -- not for quotation

Comments welcomed



# INTERPRETING THE RESULTS OF WEIGHTED LEAST-SQUARES REGRESSION: CAVEATS FOR THE STATISTICAL CONSUMER



John B. Willett and Judith D. Singer*

Harvard University Graduate School of Education

March, 1987

* The order of the authors has been determined by randomization.

Paper presented at the American Educational Research Association Conference, Washington, D.C., April 1987.

## INTERPRETING THE RESULTS OF WEIGHTED LEAST-SQUARES REGRESSION: CAVEATS FOR THE STATISTICAL CONSUMER

Quite frequently in educational, psychological and sociological research, datasets occur in which the variance of the distribution of the dependent variable at given levels of the predictors is a function of the values of the predictors. In this situation, the use of weighted, rather than ordinary, least-squares techniques is required in the fitting of regression models (Draper & Smith, 1981).

Typically, weights suitable for use in a weighted least-squares (WLS) regression analysis are not known in advance and must be estimated _in situ_ by "a combination of prior knowledge, intuition and evidence" (Chatterjee & Price, 1977, p. 101). A variety of techniques have been proposed in the statistical literature 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) to two-stage strategies in which an initial unweighted (OLS) analysis is used to inform the selection of weights (for instance, biweighting in Mosteller & Tukey, 1977). Whatever the selected approach, the ultimate objective is to achieve a "better" 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).

Of course, regardless of the manner in which the empirical weights have been selected, there remains the question of interpreting the outcomes of the analysis once the fitting is complete. It is during this interpretation that the consumer can be lead wildly astray by the output from a computer package such as SPSS$^X$. For some of the computed statistics (such as the estimated slopes) there is no problem. However, pitfalls can arise in the interpretation of other statistics for three reasons. First, by virtue of the manner in which the empirical weighting must be applied in the WLS regression by SPPS$^X$, several important statistics (i.e., the standard errors associated with the slope estimates, elements of the regression ANOVA table, the root mean-square error and related statistics) are likely to be incorrect. Second, because the regression statistics created in a WLS analysis are expressed in the metric of the weighted variates, it is not immediately obvious how even those statistics which have been computed correctly (i.e., the coefficient of determination) should be interpreted. Third, even though an optimal set of weights may have been selected, many of the important statistics and diagnostics (i.e., the standard errors and associated t-statistics, the regression ANOVA sums-of-squares, mean-squares, F-statistic and other related statistics) may not be invariant under multiplication of the weights by a constant -- a process which modifies the measurement metric in the weighted world and raises questions about how the empirical weights might optimally be scaled.

In this paper, such pitfalls in the application and interpretation of WLS regression using the SPSS[x] statistical package are demonstrated, both algebraically and by example. For the purposes of the example, an artifical dataset (whose underlying parametric structure is known) has been created. Each of the statistics commonly reported in the WLS regression analysis of such a dataset are isolated and their interpretation discussed. Where necessary, adjusted statistics that more reasonably represent the outcomes of the analysis are proposed and their use illustrated.

## WEIGHTED LEAST-SQUARES

As Mosteller & 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. Whether the investigator wishes to downplay the importance of datapoints that are intrinsically more variable at specific levels of the predictor variables, or simply to decrease the effect on the fit of remote datapoints, the strategy is the same.

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 a dependent variable and a single predictor. We will assume that observations or measures on two related variables, Y and X, have been obtained from a random sample of n independent subjects and that the relationship between these two variables in the population is given by:

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

where $\beta_0$ and $\beta_1$ are the unknown intercept and slope parameters to be estimated. Furthermore, we will assume that the $\epsilon_i$ are unobserved random errors which are normally distributed with zero mean and variance $\sigma_\epsilon^2 X^2$. Thus, the random errors are heteroscedastic and the typical OLS strategy for estimating $\beta_0$, $\beta_1$ and $\sigma_\epsilon^2$ is inefficient (Neter et al., 1985).

Typically, an empirical response to the inefficiency of the OLS estimation involves the creation of a set of weights, $w_i$, which are inversely proportional to the squared magnitudes of the observed $X_i$. These $w_i$ are then applied in the re-fitting of the regression model by weighted least-squares. Of course, in practice, it is unlikely that the functional dependence of the heteroscedastic error variance on the $X_i$ will be known exactly. However, in an empirical analysis, the error structure is usually inferred from a "combination of prior knowledge, intuition, and evidence" (Chatterjee & Price, 1977, p. 101). Often, the required evidence is obtained by inspection of residuals created in an initial unweighted (OLS) regression analysis (Neter et al., 1985, p. 170).

**Equations for the WLS estimates**

Providing the $w_i$ are known, the more efficient WLS estimates of $\beta_0$ and $\beta_1$, their sampling variances, and $\sigma_\epsilon^2$ can be estimated by direct minimization of the sum of the squared weighted residuals (for instance, see Neter et al., 1985, pp. 167-170). Equivalent results can also be obtained by transformation, in which the original variates are multiplied by the square-roots of the $w_i$ (Neter et al., 1985, pp. 171-172).

Whatever the method of estimation, of particular interest are the estimates of $\beta_0$ and $\beta_1$ (all summations taken over the index $i=1, \ldots, n$):

$$\hat{\beta}_1 = \frac{(\Sigma w_i)(\Sigma w_i X_i Y_i) - (\Sigma w_i X_i)(\Sigma w_i Y_i)}{(\Sigma w_i)(\Sigma w_i X_i^2) - (\Sigma w_i X_i)^2} \quad [2]$$

$$\hat{\beta}_0 = \frac{(\Sigma w_i Y_i) - \hat{\beta}_1 (\Sigma w_i X_i)}{(\Sigma w_i)} \quad , \quad [3]$$

and their estimated sampling variances (standard errors):

$$\text{s.e.}(\hat{\beta}_1) = \hat{\sigma}_\epsilon \left[ \frac{(\Sigma w_i)}{(\Sigma w_i)(\Sigma w_i X_i^2) - (\Sigma w_i X_i)^2} \right]^{1/2} \quad [4]$$

$$\text{s.e.}(\hat{\beta}_0) = \hat{\sigma}_\epsilon \left[ \frac{(\Sigma w_i X_i^2)}{(\Sigma w_i)(\Sigma w_i X_i^2) - (\Sigma w_i X_i)^2} \right]^{1/2} , \qquad [5]$$

where the mean-square error $\hat{\sigma}_\epsilon^2$, an unbiased estimate of the variance of the $\epsilon_i$, is estimated from the sum of the weighted squared residuals:

$$\hat{\sigma}_\epsilon^2 = \frac{\Sigma w_i (Y_i - \hat{Y}_i)^2}{n - 2} , \qquad [6]$$

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

The coefficient of determination, $R^2$, is also estimated in the transformed world. It is a measure of the proportion of the variation in weighted Y that can be accounted for by weighted X. Its estimation is based on the sum of the weighted squared residuals in comparison to the sum of the squared deviations of the weighted $Y_i$ about their (weighted) mean:

$$R^2 = 1 - \left[ \frac{\Sigma w_i (Y_i - \hat{Y}_i)}{\Sigma w_i \left[ Y_i - \frac{\Sigma w_i Y_i}{\Sigma w_i} \right]^2} \right] . \qquad [7]$$

Notice that each of these WLS estimators in Equations [2] through [7] is essentially equivalent to the corresponding OLS estimator, except that the WLS results have been obtained in a transformed "world" in which each point in the dataset has been weighted by the appropriate $w_i$. If all the $w_i$ are set equal to 1, then the simpler OLS estimators can easily be recovered.

**Scaling the weights**

Other than simply choosing the form of the weights by examining an empirical residual plot (perhaps obtained in an initial OLS regression analysis of the same data) and estimating the functional dependence of the $w_i$ on X, the absolute magnitude or scale of the weights must also be decided. At first glance, simple logic might suggest that the multiplication of all the $w_i$ simultaneously by the same numerical constant would not influence the outcomes of the analysis. Notice, for instance, that in the estimation of $\hat{\beta}_0$, $\hat{\beta}_1$ and $R^2$ (Equations [2] [3] and [7]) the multiplication of the $w_i$ by such an arbitrary constant does not influence the estimates obtained in the WLS regression because of cancellation of the constant in the numerators and the denominators of these equations. Thus it seems that, given the necessary functional dependence of the $w_i$ on $X_i$, the absolute magnitude of the weights is unimportant.

However, in the estimation of the mean-square error (Equation [6]), no such cancellation occurs and the estimation depends upon the scaling of the weights. In particular, if all the weights are doubled then the mean-square error is quadrupled (and the root mean-square error is doubled). This is not entirely unexpected, since the mean-square error is being computed in the metric of the 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 -- $\sigma^2_\epsilon$, a population parameter of fixed value -- is not fluctuating with the selection of arbitrary global magnitudes for the weights.

On the other hand, 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 latter estimates of precision are not affected by the rescaling of the $w_i$. Even though $\hat{\sigma}_\epsilon$ may double when the weights are arbitrarily doubled, inspection of Equations [4] and [5] in conjuction with Equation [6] 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 $\sigma^2_\epsilon$ for some reason, then the standard errors in Equations [4] and [5] will also be incorrect -- as is revealed later, this is exactly what happens when SPSS[x] REGRESSION is used to fit the model in Equation [1] using the WLS approach.

In essence, viable weighting schemes act to downplay the effect of remote datapoints in the estimation process, and therefore it is as though "outliers" are being "removed" (or at least "diluted") by the weighting. Intuitively, this constitutes a narrowing or focusing of the point cloud around the regression line. Consequently, we would expect a reduction in the magnitudes of the standard errors associated with the parameter estimates (i.e., an increase in the precision of the estimation) under the WLS regression strategy. This is exactly the effect desired of the

WLS fitting process and, providing the selected weights have the appropriate dependence on the $X_i$, is independent of the scaling of the $w_i$. However, the change in $\hat{\sigma}_\epsilon$ as a consequence of multiplying the $w_i$ by a constant is somewhat disconcerting; the root mean-square error is intended to estimate $\sigma_\epsilon$, a parameter that is fixed in the population! Thus it seems that, although the scaling of the $w_i$ is unimportant in the estimation of $\beta_0$, $\beta_1$ and their standard errors, it is only when $w_i$ equals $1/X_i^2$ that $\sigma_\epsilon^2$ is estimated appropriately. As is revealed later, this conflicts with a strategy (of multiplying the $w_i$ by $n/\Sigma w_i$) that will be proposed in order to rectify other problems arising when SPSS$^X$ REGRESSION is used for the estimation.

Finally, since $R^2$ is estimated in a transformed dataset in which the effects of remote datapoints have been "diluted" in the estimation process, the obtained coefficient of determination is bound to rise when WLS fitting is used. Thus, the estimate of $R^2$ obtained unthinkingly under WLS regression is frequently much larger than the value obtained under the corresponding OLS fit. To the naive consumer of computer output, this apparent increment-to-$R^2$ can represent a considerable improvement in fit and tends to be prominently displayed in any account of the analysis, whereas closer inspection reveals that the increment simply reflects the extent to which outlying datapoints have been "trimmed" from the dataset during weighting. In reality, in terms of the original point cloud, $R^2$ cannot increase in a transition from OLS to WLS regression analysis because the act of fitting by OLS serves to minimize the sum of squared distances (parallel to the ordinate) of the observed datapoints from the fitted line with consequent maximization of $R^2$. In order that confusion be dispelled, the data analyst should re-interpret the goodness-of-fit of the WLS regression in the original metric, not in the transformed world. A suitable technique is described subsequently.

## WEIGHTED LEAST-SQUARES REGRESSION USING SPSS$^X$

In this section, the SPSS$^X$ REGRESSION procedure is used to analyze a sample of artificial data whose parametric structure is known. First, the structure and creation of the sample of artificial data is described. Second, the fitting of the statistical model in Equation [1] using SPSS$^X$ REGRESSION is outlined. Third, the outcomes of the various OLS and WLS analyses are contrasted, and specific miscalculations and inaccuracies are noted and suitable adjustments proposed.

### The data

For the purposes of this paper, a bivariate sample of 50 observations on the pair of variables [$Y_i$ , $X_i$] were randomly generated such that:

$$Y_i = 2 + .5X_i + \epsilon_i \quad ,$$

where the $\epsilon_i$ were drawn from a normal distribution with zero mean and variance $0.04X_i^2$. Thus, in the hypothetical population from which this sample was drawn, the functional relationship between $Y_i$ and $X_i$ has an intercept $\beta_0$ of magnitude 2, a slope $\beta_1$ of magnitude .5, and the random errors are heteroscedastic with variance $.04X_i^2$. The sample data are displayed in Figure [1], where a fan-shaped scatterplot typical of this type of heteroscedasticity is evident.

---

Insert Figure [1] about here

---

**Fitting the statistical model**

SPSS$^X$ REGRESSION was used to fit the statistical model in Equation [1] to the data displayed in Figure [1]. Both OLS and WLS regression strategies were applied. An additional weighting variable was created with the COMPUTE statement to contain the $w_i$ (see below), and the WEIGHT command was used to indicate this variable to SPSS$^X$. This approach, which can be used to weight almost any statistical procedure in the SPSS$^X$ package, causes individual cases in the dataset to be arithmetically replicated. Then, rather than performing a WLS regression by applying Equations [2] through [7] in the original dataset, the package simply runs an OLS regression on the new arithmetically-modified dataset and assumes that appropriate estimates will be produced. As is described below, this assumption is largely unjustified.

For the purposes of the current demonstration, three sets of (supposedly equivalent) weights were created. Each of these sets of weights is proportional to the squared inverse of the value of the independent variable. The three sets of weights differ only in their scale -- any given set of weights being simply a constant multiple of any other set of weights. Commonsense might lead us to believe that the arbitrary choice of scaling factor would make no difference to the outcomes of a particular WLS regression. However, as is shown below, this is not the case -- the specific choice of the constant used as multiplier to create a set of weights is of crucial importance to the correct interpretation of the findings of the WLS regression. Thus, the weighting schemes included the basic set of weights:

$$w_{1i} = \left[\frac{1}{X_i^2}\right] \quad . \qquad [8]$$

A set in which each weight was double the corresponding weight in the basic set above:

$$w_{2i} = 2\ w_{1i} \quad , \qquad [9]$$

and a set for which the sum of the weighted number of cases equals the original sample size (Moser & Kalton, 1972):

$$w_{3i} = \left[ \frac{n}{\sum_{i=1}^{n} w_{1i}} \right] w_{1i} \quad . \qquad [10]$$

The fitting of the statistical model in Equation [1] was carried out four times: once using OLS regression, and three times using WLS regression (once for each of the sets of weights presented in Equations [8] through [10]). Excerpts from the obtained regression results are presented in Exhibits [1] through [4].

---

Insert Exhibits [1]-[4] about here

---

**Summarizing and comparing the obtained fits**

The fits obtained in Exhibits [1] through [4] are summarized in Table [1], also included are hand-calculated estimates obtained by applying Equations [2] through [7] directly. All estimates which have been computed correctly, according to Equations [2] through [7], have been printed in boldface in Table [1]. What is immediately obvious (and rather alarming!) is that there is very little agreement between the estimates obtained by SPSS[x] and the correct estimates obtained by hand. Estimation of each of the parameters is discussed briefly below.

---

Insert Table [1] about here

---

Estimated intercept and slope. From Table [1] note that all three of the SPSS$^{X}$-computed WLS estimates of $\beta_0$ are equal to the hand-computed estimate, regardless of the particular set of weights applied. The four WLS estimates of $\beta_1$ also agree exactly. In addition, the OLS estimates of $\beta_0$ and $\beta_1$ are arithmetically very close to the obtained WLS estimates and neither set of estimates is very far from the known underlying population values. This is not en[illegible]ly unexpected as the OLS and WLS estimators are both unbiased.

Estimated standard errors. 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, it is disturbing to report that the standard errors appear to depend upon which particular set of weights was applied. Notice that SPSS$^{X}$ REGRESSION was unable to obtain a correct estimate of the standard errors under neither of the first and second sets of weights, the correct estimates being obtained only under the third set of weights and by hand-calculation. This is particularly disconcerting because it is the first set of weights, the $w_{1i}$, that are the natural first choice of the data-analyst in a situation such as this.

This fluctuation of the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$ as the regression weights are rescaled is doubly disturbing when the earlier argument (centering on Equations [4] and [5]) is recalled. Earlier it was argued that the estimation of precision would be independent of any re-scaling of the weights because the $w_i$ appeared equally in both the denominators and the numerators of Equations [4] and [5] (by virtue of appearing in the numerator of $\hat{\sigma}_\epsilon$). And yet, in Table [1], we see quite clearly and unexpectedly that the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$ are doubling when the $w_{1i}$ are replaced by the $w_{2i}$. The reason for this peculiar and unexpected fluctuation is largely dependent upon the failure of SPSS$^X$ REGRESSION to estimate $\sigma_\epsilon^2$ correctly.

Estimated error variance. Notice that, in the regression ANOVA tables of Exhibit [2] through [4], the degrees-of-freedom associated with both the error and total sums-of-squares vary with the set of weights applied. Thus, in Exhibit [2], the estimation appears to have been performed under the mis-apprehension that there were 306 subjects in the sample rather than 50, and in Exhibit [3] more than one thousand additional datapoints have apparently joined the existing point cloud! It is only when the third set of weights, the $w_{3i}$, are applied in Exhibit [4] that the degrees-of-freedom are correct. This unlikely fluctuation of the degrees-of-freedom with the selection of different sets of weights is a consequence of the algorithmic strategy used by SPSS$^X$ to fit the WLS regressions, in which individual cases in the dataset were arithmetically replicated rather than applying Equations [2] through [7] directly.

The principal failure of the arithmetic replication strategy is apparent when $\sigma_{\epsilon}^2$ is estimated. Thus, rather than correctly applying Equation [6], SPSS$^X$ has based its estimation on the equation below:

$$\hat{\sigma}_{\epsilon}^2 = \frac{\Sigma w_i (Y_i - \hat{Y}_i)^2}{\Sigma w_i - 2} , \qquad [11]$$

where the sum of the weights has replaced the sample size in the denominator of Equation [6]. The numerator of this new estimator can only be computed appropriately when the first set of weights, the $w_{1i}$, are applied, whereas the denominator is only correct when the third set of weights, the $w_{3i}$, are applied. Consequently, as is evident in Table [1], $\sigma_{\epsilon}^2$ is never estimated correctly by SPSS$^X$ regardless of the set of weights selected! This failure however, can be rectified by adjusting the estimate of $\sigma_{\epsilon}^2$ obtained under the $w_{3i}$. In this case, an approriate estimator of the error variance is given by:

$$\hat{\sigma}_{\epsilon}^2 = \left[\frac{\Sigma w_{1i}}{n}\right] \times \left[\frac{\Sigma w_{3i} (Y_i - \hat{Y}_i)^2}{\Sigma w_{3i} - 2}\right] , \qquad [12]$$

and therefore the estimate of $\sigma_{\epsilon}^2$ obtained under $w_{3i}$ can be corrected by multiplying by $(\Sigma w_{1i}/n)=(306/50)$ to give .0381, a value which equals the value of the hand-computed estimate in Table [1].

Estimating the coefficient of determination. As noted earlier, although all of the WLS estimates of $R^2$ in Table [1] agree and are all correct according to Equation [7], none of these estimates are truly appropriate for describing the empirical goodness-of-fit. Recall that $R^2$ has been estimated in the transformed dataset in which the effect of remote datapoints has been "diluted" during estimation, and therefore the obtained coefficient of determination is necessarily inflated. A more informative measure of empirical goodness-of-fit can be computed by comparing the $\hat{Y}_i$ predicted under the WLS fit and the observed $Y_i$ in the original metric, not in the transformed world. An equation suitable for computing such a pseudo-$R^2$ estimate can be obtained by a simple adjustment of Equation [7]:

$$R^2 = 1 - \left[ \frac{\Sigma(Y_i - \hat{Y}_i)}{\Sigma\left[Y_i - \frac{\Sigma Y_i}{n}\right]^2} \right] , \qquad [13]$$

where the $\hat{Y}_i$ are the predicted values of the dependent variable obtained under the WLS fit, and are independent of which of the three sets of weights in Equations [8] through [10] are applied. In the current application the value of this pseudo-$R^2$ statistic is .5108 -- slightly less than the OLS estimate of .5120 as we would have expected, given the maximization of $R^2$ in an OLS fit.

## RECOMMENDATIONS

As is evident in Table [1], the SPSS$^X$ REGRESSION procedure is spectacularly incorrect in its fitting of a simple linear regression model by weighted least-squares. The magnitudes of many of the obtained estimates depend strongly upon the absolute magnitudes of the weights used in the WLS fit and, in addition, several of the crucial reported outcomes are just plain wrong. This paper has explored these inaccuracies, both algebraically and by example, and has suggested a variety of fix-ups that can be easily applied in practice.

In particular, in selecting suitable weights for application in a WLS regression with SPSS$^X$, the most successful weights are those presented in Equation [10]. These latter weights have been adjusted prior to application by taking the theoretically-appropriate weights of Equation [8] and re-scaling them so that their sum is equal to the original sample-size. However, even the application of these re-scaled weights is not entirely without problem. Specifically, the estimation of $\sigma^2_\epsilon$ continues to be incorrect and the estimation of $R^2$, while not incorrect, leads to an inflated representation of the empirical goodness-of-fit which is misleading at best. Simple and easily-applied adjustments to correct both of these estimators are presented in Equations [12] and [13] respectively.

Finally, this paper has considered only a few of the statistics that are commonly interpreted in a typical regression analysis. Furthermore, although many of our results are easily generalizable to the case of multiple linear regression using WLS, we would advise empirical researchers to be very cautious in all of their interpretations in this latter instance. In particular, although we have not investigated the manner in which more complex and sophisticated statistics such as Mallow's $C_p$, Cook's D and the Hat matrix are affected by an arbitrary re-scaling of regression weights, it would certainly seem appropriate to advise great caution in their interpretation too. The empirical application of weighted least-squares regression analysis using SPSS$^x$ would certainly seem to be a case of "caveat emptor"!!

## REFERENCES

Chatterjee, S. and Price, B. Regression Analysis by Example. New York, NY: Wiley, 1977.

Cox, D.R. and Snell, E.J. Applied Statistics: Principles and Examples. London: Chapman & Hall, 1981.

Draper, N.R. and Smith, H. Applied Regression Analysis, 2nd ed. New York, NY: Wiley, 1981.

Miller, R.G. Jr. Beyond ANOVA, Basics of Applied Statistics. New York, NY: Wiley, 1986.

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

Mosteller, F. and Tukey, J.W. Data Analysis and Regression. Reading, MA: Addison-Wesley, 1977.

Neter, J., Wasserman, W. & Kutner, M.H. Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs. Homewood, IL: Richard D. Irwin, Inc., 1985.

SPSS$^X$ Inc., SPSS$^X$ User's Guide, 2nd. ed. New York, NY: McGraw-Hill, 1986.

## EXHIBIT ONE
## UNWEIGHTED (OLS) REGRESSION

| | |
|---|---|
| Coefficient of Determination, $R^2$ | .5120 |
| Root Mean-Square Error | .1421 |

**Analysis of Variance**

| Source | df | Sum of Squares | Mean Square | F |
|---|---|---|---|---|
| Model | 1 | 1.0161 | 1.0161 | 50.351 |
| Error | 48 | .9686 | .0202 | |
| Total | 49 | 1.9847 | | |

**Variables in the Equation**

| Parameter | Estimate | Standard Error | t-statistic ($H_0$: $\beta=0$) |
|---|---|---|---|
| $\beta_0$ | 1.9853 | .0505 | 39.305 |
| $\beta_1$ | .4974 | .0701 | 7.096 |

## EXHIBIT TWO
## WEIGHTED (WLS) REGRESSION

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

| | |
|---|---|
| Coefficient of Determination, $R^2$ | .6737 |
| Root Mean-Square Error | .0776 |

### Analysis of Variance

| Source | df | Sum of Squares | Mean Square | F |
|---|---|---|---|---|
| Model | 1 | 3.7747 | 3.7747 | 627.667 |
| Error | 304 | 1.8282 | .0060 | |
| Total | 305 | 5.6029 | | |

### Variables in the Equation

| Parameter | Estimate | Standard Error | t-statistic ($H_0$: $\beta$=0) |
|---|---|---|---|
| $\beta_0$ | 1.9977 | .0077 | 260.630 |
| $\beta_1$ | .4751 | .0190 | 25.053 |

EXHIBIT THREE
WEIGHTED (WLS) REGRESSION

$$w_{2i} = 2w_{1i}$$

| | |
|---|---|
| Coefficient of Determination, $R^2$ | .6737 |
| Root Mean-Square Error | .0774 |

Analysis of Variance

| Source | df | Sum of Squares | Mean Square | F |
|---|---|---|---|---|
| Model | 1 | 15.0987 | 15.0987 | 2523.056 |
| Error | 1222 | 7.3128 | .0060 | |
| Total | 1223 | 22.4115 | | |

Variables in the Equation

| Parameter | Estimate | Standard Error | t-statistic ($H_0$: $\beta=0$) |
|---|---|---|---|
| $\beta_0$ | 1.9977 | .0038 | 522.543 |
| $\beta_1$ | .4751 | .0095 | 50.230 |

## EXHIBIT FOUR
## WEIGHTED (WLS) REGRESSION

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

Coefficient of Determination, $R^2$ .6737
Root Mean-Square Error .0790

**Analysis of Variance**

| Source | df | Sum of Squares | Mean Square | F |
|---|---|---|---|---|
| Model | 1 | .6168 | .6168 | 99.105 |
| Error | 48 | .2987 | .0062 | |
| Total | 49 | 1.9155 | | |

**Variables in the Equation**

| Parameter | Estimate | Standard Error | t-statistic ($H_0$: $\beta$=0) |
|---|---|---|---|
| $\beta_0$ | 1.9977 | .0193 | 103.564 |
| $\beta_1$ | .4751 | .0477 | 9.955 |

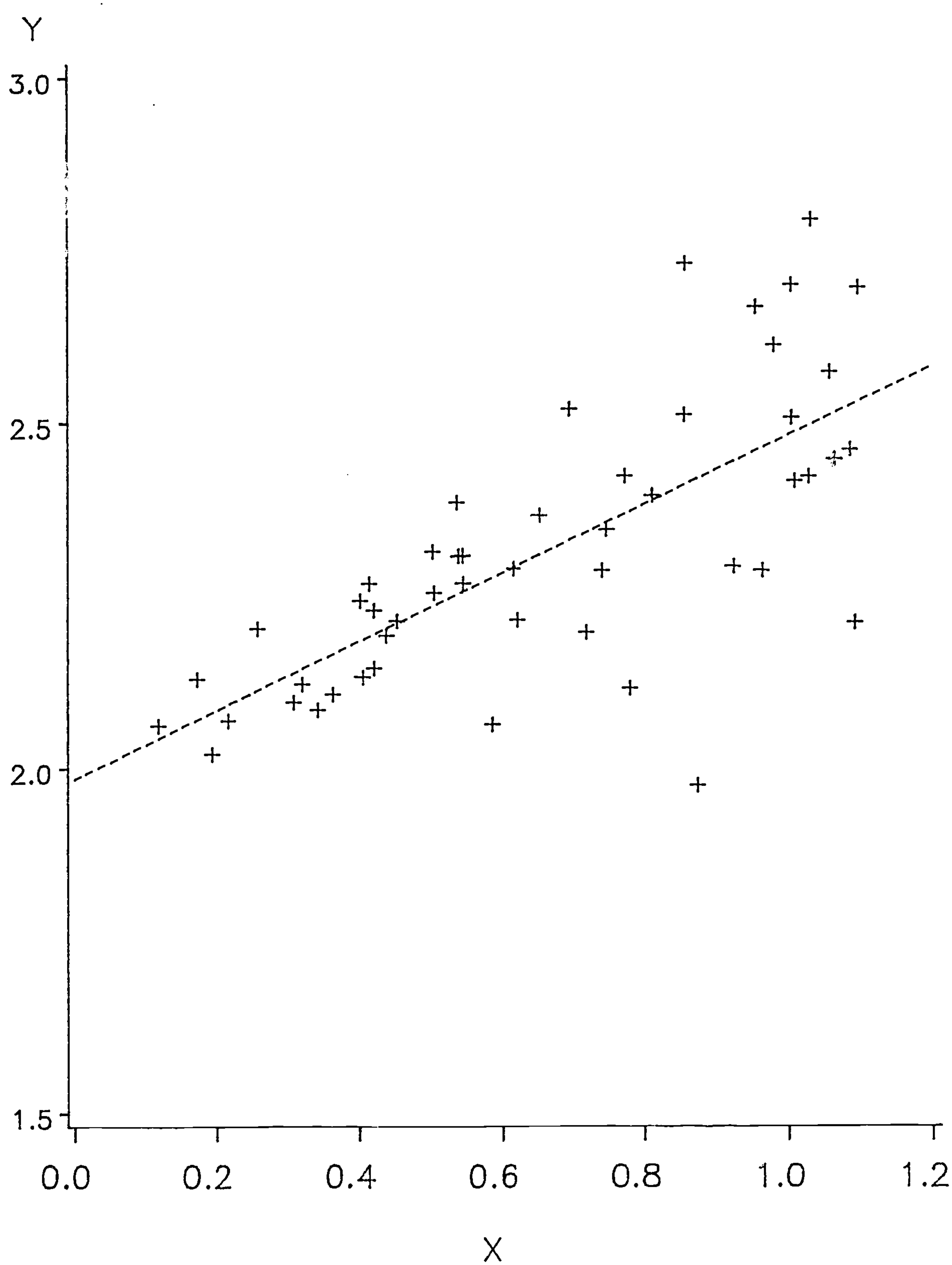
**Table 1:** Summary statistics from the four OLS and WLS regressions estimated by SPSS$^X$ REGRESSION in Exhibits 1 through 4, with accompanying correct estimates obtained by hand-calculation using Equations [2] through [7].

| Estimate* | OLS | WLS | | | |
|---|---|---|---|---|---|
| | | SPSS$^X$-calculated | | | Hand Calc |
| | | $w_1$ | $w_2$ | $w_3$ | $w_1$ |
| $R^2$ | .5120 | .6737 | .6737 | .6737 | .6737 |
| $\hat{\sigma}^2_\epsilon$ | .0202 | .0060 | .0060 | .0062 | .0381 |
| $\hat{\beta}_0$ | 1.9853 | 1.9977 | 1.9977 | 1.9977 | 1.9977 |
| s.e. $(\hat{\beta}_0)$ | .0505 | .0077 | .0038 | .0193 | .0193 |
| t | 39.305 | 26[illegible] 630 | 522.543 | 103.564 | 103.564 |
| $\hat{\beta}_1$ | .4974 | .4751 | .4751 | .4751 | .4751 |
| s.e. $(\hat{\beta}_1)$ | .0701 | .0190 | .0095 | .0477 | .0477 |
| t | 7.096 | 25.053 | 50.231 | 9.955 | 9.955 |

*Known parameter values of $\sigma^2_\epsilon$, $\beta_0$ and $\beta_1$ are .04, 2, and .5 respectively.

# FIGURE CAPTIONS

**Figure 1: Bivariate scatterplot of the artificial dataset.** Values of the dependent variable $Y_i$ plotted against values of the independent variable $X_i$, for $i = 1, \ldots, n$.


Y
3.0
2.5
2.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0
1.2
X