# Weighted Least Squares Estimation with Simultaneous Consideration of Variances and Sampling Weights

**Hee-Choon Shin[1] and Jibum Kim[2]**

[1] National Center for Health Statistics[3], 3311 Toledo Road, Hyattsville, MD 20782
[2] Sungkyunkwan University, Faculty Hall, #513, 53 Myeongnyun-dong 3-ga, Jongno-gu, Seoul, 110-745, Korea

## Abstract

A set of unweighted normal equations for a least squares solution assumes that the response variable of each equation is equally reliable and should be treated equally. When there is a reason to expect higher reliability in the response variable in some equations, we use weighted least squares (WLS) to give more weight to those equations. For an analysis of experimental or observational data, an inverse of variance is typically used for efficient estimates. For an analysis of survey data, sampling weights are typically used for unbiased and efficient estimates. There might be reasons for deviating from these weights – e.g., heteroscedasticity or extreme weights. Different weights can yield different point and interval estimates of the coefficients, affecting the interpretation of results. In other work, we considered the impact of different functional forms of weights on the WLS solutions. In the current work, we simultaneously consider sampling weights and inverses of variance for the WLS solutions, using data from the 2009-2010 National Health and Nutrition Examination Survey  a periodic survey conducted by the National Center for Health Statistics, Centers for Disease Control and Prevention . Results of using both are compared to results using the following three weights: (1) Constant; (2) Sampling weights only; and (3) Inverse of the estimated variance only. We model body weights (Kg) as a function of heights and other explanatory variables including sex and race/ethnicity, and demonstrate the effects of using both sampling weights and inverse of variance on the regression coefficients.

**Key Words: Projection, Regression, Dispersion Matrix, Variances, Weights, Least Squares, Weighted Least Squares.**

## 1. Introduction

Ever since the seminal publications of Legendre (1805) and Gauss (1809), the method of least squares (LS) has been a main tool or approach of modern statistical analysis (Celmins, 1998; Kalman, 1960; Plackett, 1949; Plackett, 1950; Seal, 1967; Sprott, 1978; Stigler, 1981; Young, 1974).

A set of unweighted normal equations assumes that the response variables in the equations are equally reliable and should be treated equally. When there is a reason to

[3] The findings and conclusions stated in this manuscript are solely those of the authors. They do not necessarily reflect the views of the National Center for Health Statistics or the Centers for Disease Control and Prevention.

expect higher reliability in the response variable in some equations, we use weighted least squares (WLS) to give more weight to those equations.

Now let $\boldsymbol{W}$ be a $(n \times n)$ diagonal matrix with weights. A set of weighted equations can be expressed as

$$\boldsymbol{WAx_w} = \boldsymbol{Wb},$$

where $\boldsymbol{b}$ is $(n \times 1)$ column vector of dependent variable and $\boldsymbol{A}$ is $(n \times P)$ matrix of explanatory variables. The $\boldsymbol{x_w}$ is the exact solution vector if a solution exists. And the normal equations from the weighted equations are

$$(\boldsymbol{WA})^{\boldsymbol{T}}\boldsymbol{WA}\widehat{\boldsymbol{x}}_{\boldsymbol{w}} = (\boldsymbol{WA})^{\boldsymbol{T}}\boldsymbol{Wb}.$$

Rearranging terms, the weighted normal equations without parentheses are

$$\boldsymbol{A^T W^T WA}\widehat{\boldsymbol{x}}_{\boldsymbol{w}} = \boldsymbol{A^T W^T Wb}.$$

And the WLS solution is

$$\widehat{\boldsymbol{x}}_{\boldsymbol{w}} = \left(\boldsymbol{A^T W^T WA}\right)^{\boldsymbol{-1}}\boldsymbol{A^T W^T Wb}.$$

If we assume non-stochastic $\boldsymbol{A}$ and $\boldsymbol{W}$, independent $\boldsymbol{b}$, and a known variance $\boldsymbol{\sigma_i^2}$ for each component of error vector, $\boldsymbol{e}$, the variance of $\widehat{\boldsymbol{x}}_{\boldsymbol{w}}$ would be

$$Var(\widehat{\boldsymbol{x}}_{\boldsymbol{w}}) = \left(\boldsymbol{A^T W^T WA}\right)^{\boldsymbol{-1}}\boldsymbol{\sigma_i^2}.$$

Our earlier works (Shin, 2013; Shin & Jibum, 2014) demonstrated the effects of differing functional forms of weights on the WLS solutions, the following five forms or methods were considered: (1) 1 (constant); (2) $m_i$ (sampling weights); (3) $\sqrt{m_i}$ (positive square root of the sampling weights); (4) $1/\hat{\sigma}_i^2$ (inverse of the estimated variance); and (5) $1/\hat{\sigma}_i$(square root of the inverse of the estimated variance). We demonstrated the importance of choosing a correct functional form of weights for WLS estimation. Estimates resulting from WLS solutions with differing functional weight forms led to conflicting research findings.

## 2. Two Approaches using Both Variances and Sampling Weights

Initially the weight for WLS methods was considered to exclusively handle heterogeneous variances (Aitken, 1935; Cochran & Carroll, 1953; Harter, 1974). Sampling weights were introduced to WLS methods to consider mainly the impact of sample selection probability on the LS estimates (Horvitz & Thompson, 1952; Neyman, 1934). In the following, we will briefly review two main WLS approaches using both variance and sampling weights: 1) Fuller & Rao approach (1978); and 2) S$a$rndal, Swensson, and Wretman (1992).

*Fuller & Rao approach* (1978). Consider a problem of estimating the unknown parameter, $\boldsymbol{x}$, in the linear regression model with heteroscedastic error variances ($\boldsymbol{V}$)

$$E[\boldsymbol{b}] = \boldsymbol{Ax},$$

where $\boldsymbol{b}$ is an $(n \times 1)$ vector of observations $b_{ij}$ $\left(i = 1,\cdots,k; j = 1,\cdots,n_i \; and \; \sum_{i=1}^{k} n_i = n\right)$ and $E$ indicates an expectation. The matrix $\boldsymbol{V}$ is defined as

$$\boldsymbol{V} = block\ diag.\left\{\sigma_1^{-1}\boldsymbol{I}_{n_1},\cdots,\sigma_k^{-1}\boldsymbol{I}_{n_k}\right\},$$

where $\sigma_i^2$ is the known error variance and $\boldsymbol{I}_{n_i}$ is the $(n_i \times n_i)$ identity matrix. The OLS estimator($\widehat{\boldsymbol{x}}$) of $\boldsymbol{x}$ is

$$\widehat{\boldsymbol{x}} = \left(\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{A}\right)^{-\mathbf{1}}\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{b},$$

And aWLS estimator of $\boldsymbol{x}$ is

$$\widetilde{\boldsymbol{x}} = \left(\boldsymbol{A}^{\boldsymbol{T}}\widehat{\boldsymbol{V}}^{\boldsymbol{T}}\widehat{\boldsymbol{V}}\boldsymbol{A}\right)^{-\mathbf{1}}\boldsymbol{A}^{\boldsymbol{T}}\widehat{\boldsymbol{V}}^{\boldsymbol{T}}\widehat{\boldsymbol{V}}\boldsymbol{b},$$

where

$$\widehat{\boldsymbol{V}} = block\ diag.\left\{\hat{\sigma}_1^{-1}\boldsymbol{I}_{n_1},\cdots,\hat{\sigma}_k^{-1}\boldsymbol{I}_{n_k}\right\}.$$

The estimator $\widetilde{\boldsymbol{x}}$ is obtained in two steps. Before calculating the estimator $\widetilde{\boldsymbol{x}}$, an estimator $\widehat{\boldsymbol{V}}$ is obtained in the first step.

The (two-step) WLS estimator of $\boldsymbol{x}$ considering both of variances and sampling weights is

$$\widetilde{\boldsymbol{x}}_{\boldsymbol{w}} = \left(\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{W}^{\boldsymbol{T}}\widehat{\boldsymbol{V}}^{\boldsymbol{T}}\widehat{\boldsymbol{V}}\boldsymbol{W}\boldsymbol{A}\right)^{-\mathbf{1}}\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{W}^{\boldsymbol{T}}\widehat{\boldsymbol{V}}^{\boldsymbol{T}}\widehat{\boldsymbol{V}}\boldsymbol{W}\boldsymbol{b},$$

where

$$\boldsymbol{W} = block\ diag.\left\{w_1^{1/2}\boldsymbol{I}_{n_1},\cdots,w_k^{1/2}\boldsymbol{I}_{n_k}\right\}.$$

and $w_k$ is the sampling weight for the $k^{th}$ group of observations.

*Sarndal, Swensson, and Wretman* (1992). Consider a WLS solution vector ($\widetilde{\boldsymbol{x}}$) for a known finite population of $N$ observations:

$$\widetilde{\boldsymbol{x}} = \left(\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{V}^{\boldsymbol{T}}\boldsymbol{V}\boldsymbol{A}\right)^{-\mathbf{1}}\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{V}^{\boldsymbol{T}}\boldsymbol{V}\boldsymbol{b},$$

where $\boldsymbol{b}$ is a $(N \times 1)$ response vector and $\boldsymbol{A}$ is a $(N \times P)$ matrix of given explanatory variables. $\boldsymbol{V}$ is a $(N \times N)$ diagonal matrix, i.e.

$$\boldsymbol{V} = diag.\{\sigma_1^{-1} \;\cdots\; \sigma_N^{-1}\},$$

where $\sigma_i^2$ is the known error variance.

We can rewrite the equation for $\widetilde{\boldsymbol{x}}$ as

$$\widetilde{\boldsymbol{x}} = \boldsymbol{C}^{-\mathbf{1}}\boldsymbol{c},$$

where $\boldsymbol{C}$ is a $(P \times P)$ matrix and $\boldsymbol{c}$ is a $(P \times 1)$ vector. Using summation notation ($\sum$), $\boldsymbol{C}$ and $\boldsymbol{c}$ can be denoted as

$$\boldsymbol{C} = \sum_U \frac{\boldsymbol{a_k a_k^T}}{\boldsymbol{\sigma_k^2}};\ \boldsymbol{c} = \sum_U \frac{\boldsymbol{a_k b_k^T}}{\boldsymbol{\sigma_k^2}},$$

where $\sum_U$ indicates a summation over the whole universe or population.

Now, the Horvitz-Thompson estimators (Horvitz & Thompson, 1952) for $\boldsymbol{C}$ and $\boldsymbol{c}$ using sample data are

$$\widehat{\boldsymbol{C}} = \sum_s \frac{\boldsymbol{a_k a_k^T}}{\boldsymbol{\sigma_k^2 \pi_k}};\ \widehat{\boldsymbol{c}} = \sum_s \frac{\boldsymbol{a_k b_k^T}}{\boldsymbol{\sigma_k^2 \pi_k}},$$

where $\boldsymbol{\pi_k}$ is inclusion probability and $\sum_s$ indicates a summation over the sample. Using matrix notation, the WLS estimator of $\boldsymbol{x}$ considering both variances (estimated variances) and sampling weights (inverses of inclusion probabilities) is

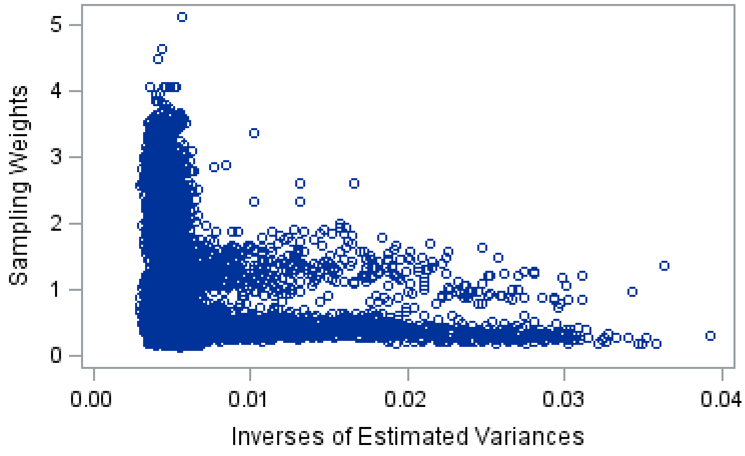
$$\widetilde{\boldsymbol{x}}_{\boldsymbol{w}} = \left(\boldsymbol{A^T W^T \widehat{V}^T \widehat{V} W A}\right)^{-1} \boldsymbol{A^T W^T \widehat{V}^T \widehat{V} W b}\,.$$

We will examine the effects of using both variances and sampling weights in WLS estimation by analyzing data from the National Health and Nutrition Examination Survey (NHANES), a periodic survey conducted by the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC). Specifically, the “examination” sample weight is used for weighted estimation.

## 3. Relationship between Inverses of the Estimated Variances and Sampling Weights

We estimated the variance in the following way (Eicker, 1963; White, 1980). First, we model the body weights as a linear function of standing heights and obtained absolute values of residuals. Second, we model the absolute values of the residuals as a function of heights. Our estimated variance ($\hat{\sigma}_i^2$) is the square of the predicted residuals in the second step. When used as the weights for WLS estimation, the values $1/\hat{\sigma}_i$ were normalized so that their sums were equal to the sample size.

Figure 1. Scatter Plot of Inverses of the Estimated Variances by Sampling Weights: 2009-2010 National Health and Nutrition Examination Survey.



For estimation purposes in an actual survey, the sampling weights would be transformed into survey weights by adjustments for non-response and coverage errors. However, we assume that $m_i$ is a fixed and non-random variable. Let $m_i\ (i = 1, \ldots, n)$ be the normalized sampling weight for the $i^{\text{th}}$ element, i.e., $\sum_{i=1}^{n} m_i = n$. For simplicity, the term "normalized" will be omitted hereafter when referring to sampling weights.

Figure 1 shows a scatter plot between the inverses of the estimated variances and the sampling weights. As we see, the points tend to be away from the 45° line and negatively correlated $(r = -.2458, p < .0001)$. The inverses of the estimated variances and the sampling weights are two distinct measures. Accordingly, it is possible to obtain differing WLS solutions or regression coefficients depending on the specific form of weights used.

## 4. Results

Table 1 shows WLS solutions for four methods for each of four models. The right-most two columns show the coefficients and their significance for simultaneous use of the estimated variances and sampling weights. Model I is a WLS model with an intercept. Model II is a no-intercept model. All the solutions or coefficients from Model II are different depending on the weights used. With constant weights in the equations, the estimated coefficient of height on body weight is .44. With square roots of sampling weights ($\sqrt{m_i}$) in the equations, the estimated coefficient for height is .46. With square roots of the inverses of estimated variances ($1/\hat{\sigma}_i$), the estimated coefficient for height is .40. With simultaneous consideration of variances and sampling weights, i.e., with $\sqrt{m_i}/\hat{\sigma}_i$, the estimated coefficient for height is .43, a value between .46 and .40. Model III is obtained by adding gender to Model II. With square roots of sampling weights ($\sqrt{m_i}$) in the equations, the estimated coefficients for men is positive (2.52) so that expected body weights of the males are greater than those of the females at a given height. With square roots of the inverses of estimated variances ($1/\hat{\sigma}_i$), however, the estimated coefficients for men is negative (-6.40) so that expected body weights of the male are lower than those of the female at a given height. With simultaneous consideration of variances and sampling weights, i.e., with $\sqrt{m_i}/\hat{\sigma}_i$, the estimated coefficient for men is -2.59, a value between 2.52 and -6.40.

Table 1. Effects of height (Cm), sex, and race/ethnicity on body weights (Kg): Solutions (coefficients) to weighted normal equations.

| Model | Variable | Functional Form of Weights for Normal Equations | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1 | | $\sqrt{m_i}$ | | $1/\hat{\sigma}_i^2$ | | $\sqrt{m_i}/\hat{\sigma}_i$ | |
| | | Coeffi-cent | Standard Error | Coeffi-cent | Standard Error | Coeffi-cent | Coeffi-cent | Coeffi-cent | Standard Error |
| I | Intercept | -90.14 | 1.253 | -93.41 | 1.279 | -77.8 | 1.087 | -82.08 | 1.161 |
| | Height (cm) | 1.01 | 0.008 | 1.03 | 0.008 | 0.93 | 0.007 | 0.95 | 0.007 |
| II | Height (cm) | 0.44 | 0.002 | 0.46 | 0.001 | 0.40 | 0.002 | 0.43 | 0.001 |
| III | Height (cm) | 0.45 | 0.002 | 0.45 | 0.002 | 0.42 | 0.002 | 0.44 | 0.002 |
| | Men | -0.94 | 0.471 | 2.52 | 0.455 | -6.40 | 0.471 | -2.85 | 0.470 |
| IV | Height (cm) | 0.47 | 0.003 | 0.46 | 0.003 | 0.46 | 0.003 | 0.45 | 0.003 |
| | Men | -0.83 | 0.462 | 2.60 | 0.449 | -5.72 | 0.455 | -2.59 | 0.460 |
| | Race/Ethnicity [a)] | | | | | | | | |
| | Hispanic | -8.26 | 0.511 | -5.79 | 0.496 | -11.59 | 0.503 | -9.05 | 0.509 |
| | NH Black | -1.64 | 0.621 | 2.07 | 0.603 | -5.88 | 0.611 | -1.47 | 0.618 |
| | NH Other | -13.06 | 0.966 | -10.51 | 0.939 | -15.77 | 0.952 | -12.58 | 0.963 |

Notes: [a)] Reference category is non-Hispanic (NH) White.

Model IV includes race/ethnicity as explanatory variables in addition to the ones in Model III. With sampling weights ($\sqrt{m_i}$) in the equations, the estimated coefficients for non-Hispanic blacks is positive (2.07). Expected body weights of non-Hispanic blacks are higher than those of non-Hispanic whites after controlling for the effects of height and gender, as indicated by positive coefficients. With the inverses of estimated variances ($1/\hat{\sigma}_i$), however, the expected body weights of non-Hispanic blacks are lower than those of non-Hispanic whites after controlling for the effects of height and gender, as indicated by negative coefficients (-5.88). With a simultaneous consideration of variances and sampling weights, i.e., with $\sqrt{m_i}/\hat{\sigma}_i$, the estimated coefficient for non-Hispanic blacks is -1.47, a value between 2.07 and -5.88. Results in Table 1 indicate that a simultaneous consideration of variances and sampling weights could be an important factor in finding a "correct" WLS solution. Applying simultaneously square roots of inverses of the estimated variances and square roots of sampling weights to a system of linear equations generates an intermediate estimate (i.e. an estimate that is between those obtained when of applying the two weights separately).

## 5. Concluding Remarks

Initially the weight for WLS methods was considered to exclusively handle heterogeneous variances. Sampling weights were introduced to WLS methods to consider mainly the impact of sample selection probability on the LS estimates. We demonstrated the effects of simultaneously considering variance and sampling weights on WLS estimation by analyzing 2009-2010 NHANES public use data. Estimates resulting from WLS when applying the square roots of both the inverses of estimated variances and the sampling weights are between estimates obtained when applying the two weights separately In the current work, the variance or dispersion matrix, $\boldsymbol{V}$**,** is diagonal. To analyze complex survey data, we could extend our analysis by specifying non-diagonal $\boldsymbol{V}$ by considering sampling design (Kendall & Stuart, 1968).

## References

Aitken, A. C. (1935). On Least Squares and Linear Combination of Observations. *Proceedings of the Royal Society of Edinburgh, 55*, 42-48.

Celmins, A. (1998). The method of Gauss in 1799. *Statistical Science, 13*(2), 123-135.

Cochran, W. G., & Carroll, S. P. (1953). A Sampling Investigation of the Efficiency of Weighting Inversely as the Estimated Variance. *Biometrics, 9*(4), 447-459.

Eicker, F. (1963). Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *The Annals of Mathematical Statistics, 34*(2), 447-456.

Fuller, W. A., & Rao, J. N. (1978). Estimation for a linear regression model with unknown diagonal covariance matrix. *The Annals of Statistics, 6*(5), 1149-1158.

Gauss, C. F. (1809). *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium* (1857 ed.). (C. H. Davis, Trans.) Boston: Little, Brown & Co.

Harter, W. L. (1974). The method of least squares and some alternatives: Part I. *International Statistical Review, 42*(2), 147-174.

Horvitz, D. G., & Thompson, D. J. (1952). A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association, 47*(260), 663-685.

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME-Journal of Basic Engineering, 82 (Series D)*, 35-45.

Kendall, M. G., & Stuart, A. (1968). *The Advanced Theory of Statistics* (Second ed., Vol. 3). New York: Hafner.

Legendre, A. M. (1805). *Nouvelles Methodes pour la Determination des Orbites des Cometes.* Paris: Courcier.

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society, 97*(4), 558-625.

Plackett, R. L. (1949). A historical note on the method of least squares. *Biometrika, 36(3/4)*, 458-460.

Plackett, R. L. (1950). Some Theorems in Least Squares. *Biometrika, 1/2*, 149-157.

Sarndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model Assisted Survey Sampling.* New York: Springer.

Seal, H. L. (1967). Studies in the History of Probability and Statistics. XV: The Historical Development of the Gauss Linear Model. *Biometrika, 54*, 1-24.

Shin, H.-C. (2013). Weighted least squares estimation with sampling weights. *Proceedings of the Survey Research Methods Section.* Alexandria, VA: the American Statistical Association.

Shin, H.-C., & Jibum, K. (2014). Effects of differing weights on regression coefficients, Paper presented at the 69th Annual Conference of the American Association for Public Opinion Research. Anaheim, CA.

Sprott, D. A. (1978). Gauss's contributions to statistics. *Historia Mathematica, 5*, 183-203.

Stigler, S. M. (1981). Gauss and the invention of Least Squares. *The Annals of Statistics, 9*(3), 465-474.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica, 48*(4), 817-838.

Young, P. (1974). Recursive Approaches to Time Series Analysis. *Bulletin / Institute of Mathematics and its Applications, 10*(May/June), 209-224.