Weighted Least Squares Estimation with Sampling Weights

Hee-Choon Shin

National Center for Health Statistics[1], 3311 Toledo Road, Hyattsville, MD 20782

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 survey data, sampling weights, as relatively important variables, should be used for unbiased and efficient estimates. We will briefly go over the least squares theory and related issues and propose a specific form of "weight" variable when we apply the sampling weights to the weighted equations. 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) will be analyzed to demonstrate the proposed approach.

Key Words: Projection, Regression, 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). In the following, we will briefly review LS and weighted least squares (WLS) methods, and review two types of weights for WLS solutions: 1) inverse of variances, and 2) sampling weights. We will consider specific forms of the "weight" variable when we apply the sampling weights to the weighted equations. We will examine the effects of using various functional forms of the two weight types 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).

Consider a survey of a single response/dependent variable and P explanatory variables for a sample of n eligible individuals/elements. There will be n equations that express the dependent variable as a function of P explanatory variables. In matrix form, they are

$$\boldsymbol{Ax} = \boldsymbol{b},$$

where $\boldsymbol{b}$ is $(n \times 1)$ column vector of dependent variable and $\boldsymbol{A}$ is $(n \times P)$ matrix of explanatory variables. Our objective is to find the unknown $(P \times 1)$ vector of $\boldsymbol{x}$. With a

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

non-singular and square $\boldsymbol{A}$, the solution is simply

$$\boldsymbol{x} = \boldsymbol{A^{-1}b},$$

where $\boldsymbol{A^{-1}}$ is the inverse of $\boldsymbol{A}$. As Simpson demonstrated in his letter to the President of the Royal Society (Simpson, 1755; Plackett, 1958), it has been confirmed since mid-18th century that "the more observations or experiments there are made, the less will the conclusion be liable to err, provided they admit of being repeated under the same circumstances." In usual statistical applications, accordingly, n is much larger than P (i.e., $n \gg P$). In such a case, usually there is no exact solution for the set of equations because $\boldsymbol{b}$ is not in a column space of $\boldsymbol{A}$. We need to project $\boldsymbol{b}$ onto a column space of $\boldsymbol{A}$ at a right angle to find the best approximate solution for estimate $\hat{x}$ of $\boldsymbol{x}$ (Strang, 2009). Utilizing orthogonality between $\boldsymbol{A}$ and the error vector $\boldsymbol{e} = (\boldsymbol{b} - \boldsymbol{A\hat{x}})$, we have the following normal equations for a least squares (LS) solution $\boldsymbol{\hat{x}}$

$$\boldsymbol{A^T A\hat{x} = A^T b},$$

where $\boldsymbol{A^T}$ is the transpose of $\boldsymbol{A}$**.** And the LS solution is

$$\boldsymbol{\hat{x}} = \left(\boldsymbol{A^T A}\right)^{\boldsymbol{-1}} \boldsymbol{A^T b}.$$

Assuming non-stochastic $\boldsymbol{A}$, independent $\boldsymbol{b}$ and a known variance $\boldsymbol{\sigma^2}$ for each component of $\boldsymbol{e}$, the variance of $\boldsymbol{\hat{x}}$ is

$$Var(\boldsymbol{\hat{x}}) = \left(\boldsymbol{A^T A}\right)^{\boldsymbol{-1}} \boldsymbol{\sigma^2}.$$

2. Weighted Least Squares

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 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

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

where $\boldsymbol{x_w}$ is the exact solution vector if a solution exists. And the normal equations from the weighted equations are

$$(\boldsymbol{WA})^{\boldsymbol{T}} \boldsymbol{WA\hat{x}_w} = (\boldsymbol{WA})^{\boldsymbol{T}} \boldsymbol{Wb}.$$

Rearranging terms, the weighted normal equations without parentheses are

$$\boldsymbol{A^T W^T WA\hat{x}_w = A^T W^T Wb}.$$

And the WLS solution is

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

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

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

3. Choice of the Weights

How should we choose the weights? Gauss considered differences in precision of $\boldsymbol{b}$ assuming a known variance σ_i^2 $(i = 1, 2, \ldots n)$ and generalized his method of least squares with weights as inverses of the square root of variances (Plackett, 1949).

$$\boldsymbol{W} = diag\left\{\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \ldots, \frac{1}{\sigma_n}\right\}.$$

Without any covariates, $\boldsymbol{A}$ would be a vector of *1*'s.

$$\widehat{\boldsymbol{x}}_{\boldsymbol{w}} = \frac{\sum_{i=1}^{n} \sigma_i^{-2}\, b_i}{\sum_{i=1}^{n} \sigma_i^{-2}}.$$

$$Var(\widehat{\boldsymbol{x}}_{\boldsymbol{w}}) = 1.$$

In practice, the weights are typically unknown and need to be estimated. In special circumstances, as in laboratory settings, each σ_i^2 could be estimated, based on n_i measurements or observations (Cochran & Carroll, 1953).

4. Sampling Weights

Now consider a survey of a single response/dependent variable and *P* explanatory variables for a sample of n eligible individuals/elements with a column of sampling weight variables. Let m_i $(i = 1, \ldots, n)$ be the normalized sampling weights for the i^{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. 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. Frequently we are interested in estimating unbiased means, proportions, totals, and their variances. For $\boldsymbol{b}$, for example, a desired mean and its variances are

$$\bar{b}_w = \frac{\sum_{i=1}^{n} m_i b_i}{\sum_{i=1}^{n} m_i}$$

$$VAR\left(\bar{b}_w\right) = \frac{\hat{\sigma}^2}{n},$$

where $\hat{\sigma}^2$ is an estimated variance of the variable $\boldsymbol{b}$. Let $\boldsymbol{W}$ be $diag\{m_1, m_2, \ldots, m_n\}$. Without any covariates, $\boldsymbol{A}$ would be a vector of *1*'s. The least square solution $\widehat{\boldsymbol{x}}_{\boldsymbol{w}}$ is simply

$$\hat{\boldsymbol{x}}_{\boldsymbol{w}} = \frac{\sum_{i=1}^{m} m_i^2 \, b_i}{\sum_{i=1}^{m} m_i^2},$$

which is another weighted mean of $\boldsymbol{b}$. However, notice that each b_i in this mean is weighted by the square of the sampling weights, *i.e.*, m_i^2 $(i = 1, 2, \ldots, n)$. To produce $\hat{\boldsymbol{x}}_{\boldsymbol{w}} = \bar{b}$ (the unweighted sample mean of $\boldsymbol{b}$) with $\hat{x}_w$ formulated this way, one should use the positive square roots of the sampling weights ($\sqrt{m_i}$) for the diagonal entries of $\boldsymbol{W}$:

$$\boldsymbol{W} = diag\{\sqrt{m_1}, \sqrt{m_2}, \ldots, \sqrt{m_n}\}.$$

5. Effects of Functional Form of "Weights" on WLS estimation

To evaluate the effects of specific functional forms of the "weights", we investigated body weights of the U.S. population by analyzing data collected in the National Health and Nutrition Examination Survey (NHANES). Specifically, we analyzed the 2009-2010 NHANES public use file (PUF).

Figure 1. Density Function of Body Weights(Kg) of U.S. Population:2009-2010 National Health and Nutrition Examination Survey.

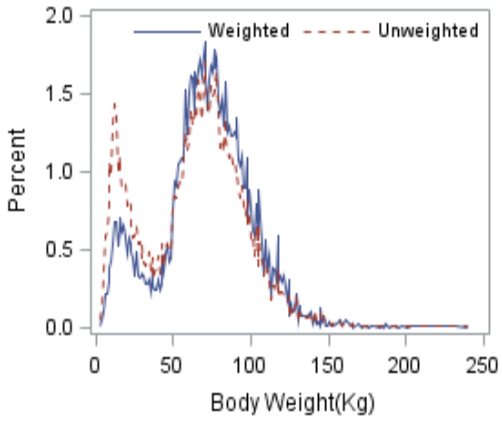


Figure 2. Distribution Function of Body Weights(Kg) of U.S. Population:2009-2010 National Health and Nutrition Examination Survey.

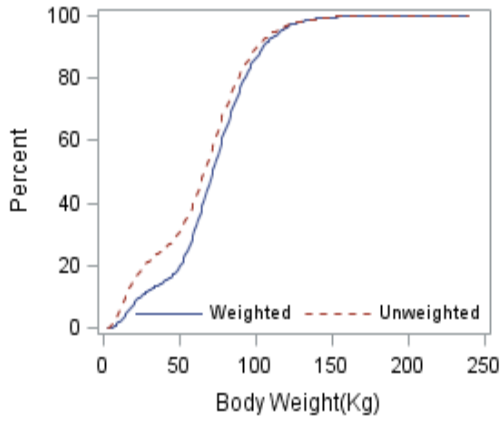


Figure 1 shows unweighted and weighted density functions of body weights in the U.S. Figure 2 shows the corresponding distribution functions. The weighted functions are shown in solid lines and dashed lines are for unweighted functions. Figures 1 and 2 show differences between weighted and unweighted distributions, and indicate an oversampling of children or individuals with less than 40 kg of body weights. It emphasizes a necessity of weighting for unbiased estimates of body weights. Table 1 shows differences between weighted and unweighted arithmetic means (70.86 Kg *vs.* 63.26 Kg), weighted by the sampling weights (m_i). Table 2 shows solutions of weighted least squares with two different functional forms of sampling weights: 1) the sampling weights (m_i), and 2) the positive square root of the sampling weights ($\sqrt{m_i}$). As expected, the correct answer (70.86 Kg) shown in Table 1 is obtained by applying the square root of the sampling weights to the normal equations. Applying sampling weights

(m_i) to the normal equations generated a biased estimate (75.78 Kg) for the average body weight for the U.S. population.

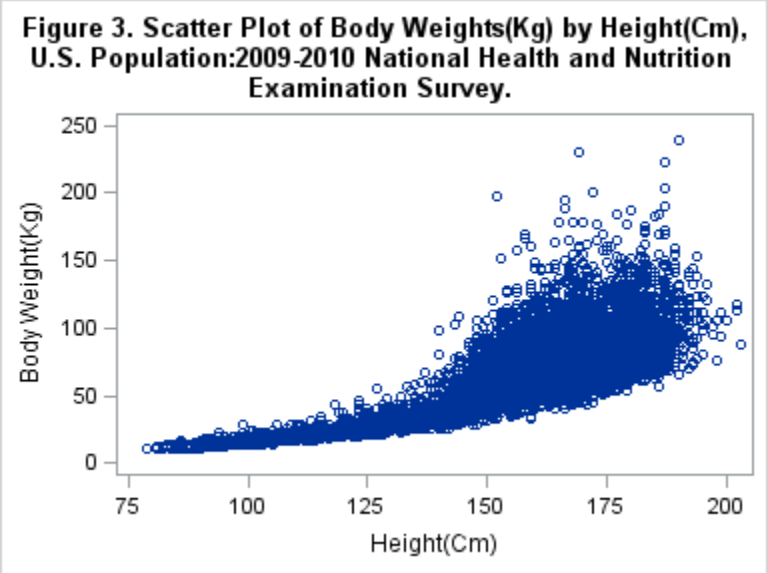
To consider the impact of differing functional forms of weights on the WLS solutions, the following five forms or methods were considered: (1) 1 (constant); (2) m_i (sampling weights); (3) $\sqrt{m_i}$ (positive square root of the sampling weights); (4) $1/\hat{\sigma}_i^2$ (inverse of the estimated variance); and (5) $1/\hat{\sigma}_i$(square root of the inverse of the estimated variance).

Table 1. Unweighted and weighted summary statistics of body weight (Kg)

	Sample Size	Mean	Standard Deviation	Median	Standard Error
Unweighted	10,162	63.29	32.36	66.80	.3210
Weighted	10,162	70.86	29.19	72.10	.2896

Table 2. Estimated means and standard errors of the body weights: Solutions to weighted normal equations $\boldsymbol{WAx_w = Wb}$, where $\boldsymbol{A}$ is a vector of *1*'s.

Weights	Sample Size	Mean	Standard Error
m_i	10,162	75.78	.3440
$\sqrt{m_i}$	10,162	70.86	.2896



Estimation of variance (σ_i^2). As stated, in practice, the variance is not known and needs to be estimated. Figure 3 shows the relationship between body weights (Kg) and height (Cm). We see non-homogeneous variances by height level. 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 of $1/\hat{\sigma}_i^2$ and $1/\hat{\sigma}_i$ were normalized so that their sums were equal to the sample size.

Table 3 shows WLS solutions for 5 methods for each of 4 models. Model I is a WLS model with an intercept. Given the distribution of body weights by heights shown in Figure 3, Model I is improper – it predicts unreasonably large negative weights for shorter persons. Model II is a no-intercept model and does not predict a negative weight. All the solutions or coefficients from Model II are statistically different from each other. For example, the WLS solution for $\boldsymbol{W} = diag\{\sqrt{m_1}, \sqrt{m_2}, \ldots, \sqrt{m_n}\}$ is .46 and .47 for $\boldsymbol{W} = diag\{m_1, m_2, \ldots, m_n\}$. With $\boldsymbol{W} = diag\{1/\sigma_1, 1/\sigma_2, \ldots, 1/\sigma_n\}$, WLS solution is .40.

Model III was obtained by adding gender to Model II. With sampling weights (m_i or $\sqrt{m_i}$) in the equations, the estimated coefficients for men are positive (5.79 and 2.52) so that expected body weights of the males were greater than those of the females at a given height. With the inverse of estimated variances ($1/\hat{\sigma}_i^2$ or $1/\hat{\sigma}_i$), however, the estimated coefficients for men are negative (-9.26 and -6.40) so that expected body weights of the male were lower than those of the female at a given height.

Table 3. 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		m_i		$\sqrt{m_i}$		$1/\hat{\sigma}_i^2$		$1/\hat{\sigma}_i$	
		Coeffi-cent	Standard Error	Coeffi-cent	Standard Error	Coeffi-cent	Standard Error	Coeffi-cent	Standard Error	Coeffi-cent	Standard Error
I	Intercept	-90.14	1.253	-96.8	1.617	-93.41	1.279	-66.7	0.999	-77.8	1.087
	Height (cm)	1.01	0.008	1.04	0.010	1.03	0.008	0.85	0.006	0.93	0.007
II	Height (cm)	0.44	0.002	0.47	0.002	0.46	0.001	0.32	0.002	0.40	0.002
III	Height (cm)	0.45	0.002	0.45	0.002	0.45	0.002	0.36	0.002	0.42	0.002
	Men	-0.94	0.471	5.79	0.548	2.52	0.455	-9.26	0.505	-6.40	0.471
IV	Height (cm)	0.47	0.003	0.45	0.003	0.46	0.003	0.40	0.003	0.46	0.003
	Men	-0.83	0.462	5.80	0.542	2.60	0.449	-8.26	0.483	-5.72	0.455
	Race/Ethnicity [a)]										
	Hispanic	-8.26	0.511	-4.11	0.599	-5.79	0.496	-11.46	0.534	-11.59	0.503
	NH Black	-1.64	0.621	4.55	0.728	2.07	0.603	-7.94	0.649	-5.88	0.611
	NH Other	-13.06	0.966	-9.69	1.133	-10.51	0.939	-14.65	1.010	-15.77	0.952

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 (m_i or m_i') in the equations, the estimated coefficients for non-Hispanic blacks are positive (4.55 and 2.07). Expected body weights of non-Hispanic blacks were higher than those of non-Hispanic whites after controlling for the effects of height and gender, as indicated by positive coefficients. With the inverse of estimated variances ($1/\hat{\sigma}_i^2$ or $1/\hat{\sigma}_i$), however, the expected body weights of non-Hispanic blacks were lower than those of non-Hispanic whites after controlling for the effects of height and gender, as indicated by negative coefficients (-7.94 and -5.88).

The differing solutions shown in Table 3 indicate that choosing a correct form of weights is important in solving linear equations and in measuring the effects of explanatory variables.

6. Concluding Remarks

We showed that the positive square root of sampling weights should be used as weights in the normal equations for weighted least squares solutions with sampling weights. We showed that applying sampling weights to normal equations resulted in biased estimates (75.78 *vs.* 70.86 in Table 2). We also demonstrated the importance of choosing a correct functional form of weights for WLS estimation with sampling weights by analyzing 2009-2010 NHANES data. Estimates resulting from WLS solutions with incorrect functional weight forms are significantly different from those with correct functional weight forms, and led to erroneous research findings.

7. Acknowledgements

We thank Don Malec and Iris Shimizu for helpful comments and suggestions.

References

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.

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

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.

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

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.

Plackett, R. L. (1958). Studies in the History of Probability and Statistics: VII. The Principle of the Arithmetic Mean. *Biometrika, 45*(1/2), 130-135.

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.

Simpson, T. (1755). A letter to the right honorable George Earl of Macclesfield, President of the Royal Society, on the advantage of taking the mean of a number of observations, in practical astronomy. *Philosophical Transactions, 49*(Part 1), 82-93.

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.

Strang, G. (2009). *Introduction to Linear Algebra* (4th ed.). Wellesley, MA: Wellesley-Cambridge Press.

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.