# ONE-STEP WEIGHTED LINEAR REGRESSION FOR DICHOTOMOUS DEPENDENT VARIABLE

Wen-Fu P. Shih, Florida Atlantic University

## Introduction

A general linear approach to the analysis of qualitative data was developed by Gizzle, Stamer and Koch (GSK) in 1969.[3] Lehnen and Koch also applied this technique to the field of political science in 1974.[4]

Based on the measurement of dependent variables and the hypothesis of interest, both dummy variable regression and analysis of variance approaches have been applied to categroical data. The analysis of variance apporach requires the assumption of homogeneity of variance. However, when the dependent variable is nominal or ordinal this assumption needs some modification. Also the homogeneity of variance is not adjusted by the dummy variable regression technique. Thus, the GSK approach uses the analysis of variance type of application to nominal data without the homogeneity assumption. This method is based on the general weighted least squares to estimate appropriate functions of the cell proportions in the complex contingency table.

Most social science data are either from sample surveys with large sample sizes or from government or private institutions with mass data stored on the tapes. To apply the GSK method, researchers have to start from contingency table analysis. Then, based on this table output, a design matrix for regression models are set up. It requires tremendous man power and computer time to handle large samples. Especially, when setting up the design matrix for a saturated model for a large number of independent variables, the computation of interaction terms become very complicated and likely causes errors.

Therefore, the purpose of this study is to demonstrate a weighted linear regression analysis for a dichotomous dependent variable without going through contingency table analysis. It is called a one-step weighted linear regression (OWLR). Based on this OWLR technique, a user can directly apply regression analysis to the raw data and the unequal variance of the dependent variable can be simultaneously adjusted. This approach also can be generalized to the case of a dependent variable with more than two levels by adjusting the variance-covariance matrix.

## Analysis of Technique

The social behavior relationship in some cases can be formulated as a dichotomous dependent variable. For example, a person can send his children to either public school or private school; or, he either protests school desegregation or does not protest. According to Godberger[2], this type of variable is naturally formulated into a regression equation. The dependent variable of this regression $Y_i$ has only two values, which can be coded as 0 and 1 without losing generality.

I.e.: $Y_i = \begin{cases} 0 \text{ if negative} \\ 1 \text{ if positive} \end{cases}$

The general linear regression model can be written as

$$Y = X\beta + \varepsilon \tag{1}$$

where Y is an n x 1 vector of observations on a dependent variable, n is a dimension of the vector, X is an n x p matrix of nonstochastic regressors with rank p, $\beta$ is a p x 1 vector of unknown regression coefficients, and $\varepsilon$ is an n x 1 vector of unknown disturbances.

To illustrate, let behavior toward school desegregation be a dependent variable Y, and let income and education be independent variables $X_1$ and $X_2$, respectively. The model can be written as

$$Y_i = \mu + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \tag{2}$$

where

$$Y_i = \begin{cases} 1 \text{ if protest} \\ 0 \text{ if not protest} \end{cases}$$

$$X_{1i} = \begin{cases} 1 \text{ if high income} \\ -1 \text{ if low income} \end{cases}$$

$$X_{2i} = \begin{cases} 1 \text{ if education high} \\ -1 \text{ if education low} \end{cases}$$

i = 1, 2, ..., n and n is the number of observations. $\mu$, $\beta_1$, and $\beta_2$ are regression coefficients, and $\varepsilon_i$ are unknown disturbances.

This model can also be considered as a conditional probability function $P_r$ (Protest given $X_1$, $X_2$) = $P(Y_i = 1 \mid X_1, X_2)$ (3)

Denote $P(Y_i = 1 \mid X_1, X_2) = p$ for all i (4)

then

$$E(Y_i) = 0\cdot(1-p) + 1 \cdot p = p \tag{5}$$

$$E(Y_i^2) = 0^2\cdot(1-p) + 1^2\cdot p = p \tag{6}$$

$$Var(Y) = E(Y_i^2) - (E(Y))^2 = p(1-p) \tag{7}$$

$$Cov(Y_j, Y_j') = E(Y_j Y_j') - E(Y_j)E(Y_j') = 0 \tag{8}$$

The dependent variable $Y_i$ is a Bernouli trial under the condition of $X_1 = x_{1i}$, $X_2 = x_{2i}$. According to Feller[1] the probability function is

$$P(Y=y) = p^y(1-p)^{1-y}, \ y=0 \text{ and } 1 \text{ for each trial.} \tag{9}$$

The likelihood function of p is

$$L(p) = P(Y = y_1, Y = y_2, \ldots, Y = y_n) = p^{\sum_i y_i}(1 - p)^{n - \sum_i y_i} \tag{10}$$

To maximize likelihood, take log of L

$$\ln L = \sum_i y_i \ln p + (n - \sum y_i)\ln(1-p) \tag{11}$$

and take derivative of L with respect to p,

$$\frac{\partial \ln L}{\partial p} = \frac{\sum_i y_i}{p} - \frac{n - \sum_i y_i}{1-p} = 0 \tag{12}$$

The maximum likelihood estimator of p is

$$\hat{p} = \frac{\sum_i y_i}{n} = \bar{y} \tag{13}$$

From (7) and (13) the estimator of Var(Y), $s^2$, is obtained as

$$s^2 = \hat{p}\ (1 - \hat{p}) \tag{14}$$

Given $X_1 = x_{1i}$ and $X_2 = x_{2i}$, where $x_{1i}$, $x_{2i}$ = -1 or 1, the maximum likelihood estimators of p and Var (Y) are noted as

$$\hat{p}_j = \frac{\sum_{i=1}^{n_j} y_{ij}}{n_j} \tag{15}$$

$$\text{and} \quad s_j^2 = \hat{p}_j(1 - \hat{p}_j) \tag{16}$$

where $n_j$ equals the total frequencies in the subgroup j, with ($X_1$ = -1 or 1 and $X_2$ = -1 or 1).

$\Sigma y_{ij}$ equals the total number of 1's in the jth subgroup.

In this particular case there are four subgroups, for j = 1, Group 1 = ($X_1$ = -1, $X_2$ = -1) with $n_1$ observations, for j = 2, Group 2 = ($X_1$ = -1, $X_2$ = 1) with $n_2$ observations; for j = 3, Group 3 = ($X_1$ = 1, $X_2$ = -1) with $n_3$ observations; for j = 4, Group 4 = ($X_1$ = 1, $X_2$ = 1) with $n_4$ observations. The total number of subgroups equal $\prod_k \lambda_k$, where $\lambda_k$ equals the levels of $X_k$.

Therefore, the variance-covariance matrix $\Sigma$ becomes

$$\left[\begin{array}{c|c|c|c} \begin{matrix} s_1^2 & & \\ & \ddots & \\ & & s_1^2 \end{matrix} & & & \\ \hline & \begin{matrix} s_2^2 & & \\ & \ddots & \\ & & s_2^2 \end{matrix} & & \\ \hline & & \begin{matrix} s_3^2 & & \\ & \ddots & \\ & & s_3^2 \end{matrix} & \\ \hline & & & \begin{matrix} s_4^2 & & \\ & \ddots & \\ & & s_4^2 \end{matrix} \end{array}\right] = {}_n\Sigma_n \qquad (17)$$

where $n = n_1 + n_2 + n_3 + n_4$.

As can be seen from equation 17, the disturbance belongs to heteroskedastic, and the weighted linear regression coefficient $\beta$ can be estimated by

$$\hat{\beta} = (X'\Sigma^{-1}X)^{-1}\ X'\Sigma^{-1}Y \qquad (18)$$

the equation 18 is equivalent to a function used to solve the linear regression coefficient $\beta$, i.e., a linear regression equation is expressed as

$$Y^* = X^*\beta + \Sigma \qquad (19)$$

in which

$$\hat{\beta} = (X^{*\prime}X^*)^{-1}X^{*\prime}Y^*, \qquad (20)$$

$$X^* = \Sigma^{-\frac{1}{2}}X, \qquad (21)$$

and

$$Y^* = \Sigma^{-\frac{1}{2}}Y. \qquad (22)$$

In practical application, the $X^*$ and $Y^*$ are replaced as $x_{ij}^*$ and $y_{ij}^*$, respectively; and the $x_{ij}^*$ and $y_{ij}^*$ are expressed as

$$x_{ij}^* = \frac{x_{ij}}{s_j} \qquad (23)$$

$$y_{ij}^* = \frac{y_{ij}}{s_j} \qquad (24)$$

where

$$s_j = (\hat{p}_j(1 - \hat{p}_j))^{\frac{1}{2}} \qquad (25)$$

Based on the equations 19, 20, 23, 24, and 25, a systematic flow chart as shown in Figure 1 is developed for this OWLR technique. This flow chart has been also converted to a computer program with Fortran IV language. The input data of this program are only raw dependent variables and dichotomous independent variables. All interaction terms will be generated by the program itself. The results of the program output are the estimators of weighted linear regression coefficients, the $\chi^2$ test for each regression, and the goodness of fit of the model.

Another computer program with the same language is also developed to generate the independent variables with more than two levels.

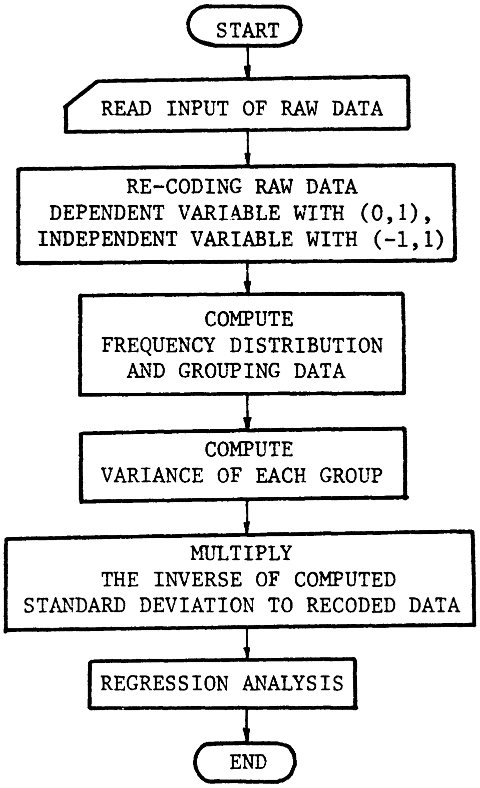

Figure 1: Systematic flow chart of OWLR technique.

## Example of Application

Data used in this example are obtained from a survey of parents in Duval County, Florida.

On the basis of responses to questionnaire items by parents whose children attend desegregated public schools, two groups are classified:

(1) Those who did not protest against desgregation: Y = 0

(2) Those who did protest: Y = 1.

The goal is to study the impact on protest of income, education, the percent of black change in assigned school from 1971-72 to 1972-73, and racial prejudice. The regression model for the ith observation can be written as:

$$\text{(A)}\quad Y_i = \mu + \beta_1X_{1i} + \beta_2X_{2i} + \beta_3X_{3i} + \beta_4X_{4i} + \varepsilon_i \qquad (26)$$

$$\text{or}\quad P(Y_i/X_1,X_2,X_3,X_4) = \mu + \beta_1X_{1i} + \beta_2X_{2i} + \beta_3X_{3i} + \beta_4X_{4i} + \varepsilon_i \qquad (27)$$

i = 1, 2, ..., n

where $X_{1i} = \begin{cases} 1 \text{ for high income} \\ -1 \text{ for low income} \end{cases}$

$$X_{2i}=\begin{cases} 1 & \text{for high education} \\ -1 & \text{for low education} \end{cases}$$

$$X_{3i}=\begin{cases} 1 & \text{for high racial prejudice} \\ -1 & \text{for low racial prejudice} \end{cases}$$

$$X_{4i}=\begin{cases} 1 & \text{for the school with black ratio increase} \\ -1 & \text{for the school that did not change or decreased the black ratio} \end{cases}$$

There are sixteen possible combinations of X's. Hence, sixteen subgroups are used in the sample. The frequency in each subgroup of data are shown in Table 1. For convenience, the equation 27 can be expressed as

$$s_j = (\hat{p}_j\ (1 - \hat{p}_j)\ )^{\frac{1}{2}}$$

$$= (\frac{\sum_i y_{ij}}{n_j}\ (1 - \frac{\sum_i y_{ij}}{n_j})\ )^{\frac{1}{2}}$$

$$= \frac{1}{n_j}\ (\sum_i y_{ij}\ (n_j - \sum_i y_{ij})\ )^{\frac{1}{2}} \qquad (28)$$

As can be seen from Table 1, the GSK method requires all these Y's and X's values for the analysis of the weighted linear regression coefficient. However, on this study, the values of the weighted linear regression can be calculated directly from raw data by using the OWLR program, and the Y is a dichotomous dependent variable instead of using all frequencies as shown in Table 1. The result of regression obtained by both GSK method and OWLR techniques is the same and is expressed as

$$Y = 0.2841 + 0.0332X_1 + 0.0700X_2{}^{**} + 0.0724X_3{}^{**} + 0.0339X_4 \qquad (29)$$

**significant at the $\alpha$ = .01 level.

As can be seen from equation 29, the education $X_1$ and changing black ratio $X_4$ are nonsignificant, but income $X_2$ and racial prejudice $X_3$ are highly significant. These imply that the proportion of protest of school desegregation is highly affected by the parent's income and racial prejudice, but is not affected by the parent's education and black ratio change. The code system (-1, 1) is used in the study. Based on the previous study reported by Shih, et. al.[5] (1975), the regression coefficients are twice larger than the percentage wise effect on the dependent variable which is based on the (0, 1) code system.

TABLE 1
Protest Cross-Classified by
Education, Income, Racism and Percent of Black Change

| Subgroup | Non-protesters | | Protesters | Education | Income | Racism | % Black Change | $(P\ (1 - P)\ )^{\frac{1}{2}}$ |
|---|---|---|---|---|---|---|---|---|
| | 0 | Y | 1 | $X_1$ | $X_2$ | $X_3$ | $X_4$ | |
| 1 | 1 | | 1 | -1 | -1 | -1 | -1 | .5000 |
| 2 | 32 | | 6 | 1 | -1 | -1 | -1 | .3646 |
| 3 | 2 | | 1 | -1 | 1 | -1 | -1 | .4714 |
| 4 | 30 | | 8 | 1 | 1 | -1 | -1 | .4077 |
| 5 | 6 | | 2 | -1 | -1 | 1 | -1 | .4330 |
| 6 | 33 | | 14 | 1 | -1 | 1 | -1 | .4573 |
| 7 | 6 | | 2 | -1 | 1 | 1 | -1 | .4330 |
| 8 | 27 | | 23 | 1 | 1 | 1 | -1 | .4984 |
| 9 | 2 | | 1 | -1 | -1 | -1 | 1 | .4714 |
| 10 | 23 | | 6 | 1 | -1 | -1 | 1 | .4051 |
| 11 | 1 | | 1 | -1 | 1 | -1 | 1 | .5000 |
| 12 | 12 | | 8 | 1 | 1 | -1 | 1 | .4099 |
| 13 | 7 | | 2 | -1 | -1 | 1 | 1 | .4157 |
| 14 | 19 | | 7 | 1 | -1 | 1 | 1 | .4436 |
| 15 | 3 | | 2 | -1 | 1 | 1 | 1 | .4899 |
| 16 | 11 | | 15 | 1 | 1 | 1 | 1 | .4940 |

Thus, the result $b_2$ = .07 can be interpreted to mean that the group with the higher income is 3.5% more likely to protest school desegregation, or that the proportion of protestors is increased 3.5% from lower income to higher income. The value of $b_3$ = .0724 implies that the proportion of protestors is 3.62% higher due to the higher racial prejudice.

The model with two factor interactions of this data can be written as

$$y_j = \mu + \sum_{i=1}^{4} \beta_i X_{ij} + \sum_{\substack{i=1 \\ i' \neq i}}^{4} \beta_{ii'} X_{ij} X_{i'j} + \varepsilon_j$$

The estimated regression coefficients and their $\chi^2$ tests are shown in Table 2. As Table 2 shows, further analysis is probably meaningless because all b's are nonsignificant.

## Summary and Conclusions

The Grizzle, Starmer and Koch (GSK) approach has been widely applied to qualitative (ordinal) data in the social sciences to perform general weighted linear regression. The application of the GSK method to most social science data requires two steps: the first is to find the cell frequencies for each subgroup; the second is to construct a design matrix and apply the method. When the sample size is very large, as in most social science data, this approach is too tedious and cumbersome to use because there are too many interaction terms and much computing time is involved. Therefore, a one-step procedure without going through the procedures of finding cell frequencies and constructing the design matrix has been mathematically modified in this study to provide the weighted linear regression. This

TABLE 2

| Independent Variable | b | $X^2$ |
|---|---|---|
| Education | -.0039 | .0062 |
| Income | .0510 | 1.4700 |
| Racism | .0130 | .0707 |
| % Black Change | .0332 | .6329 |
| Education x Income | .0385 | .8594 |
| Education x Racism | .0668 | 1.9232 |
| Education x % Black | .0081 | .0383 |
| Income x Racism | .0278 | 1.1462 |
| Income x % Black | .0363 | 1.8352 |
| Racism x % Black | -.0147 | .2971 |

All $X^2$'s are nonsignificant at $\alpha = .05$.

one-step weighted linear regression (OWLR) for dichotomous dependent variables has been computerized. The imput data of this program require only a raw dependent variable and dichotomous independent variables. All interaction terms are generated by the computer itself. The result of the program output are the estimators of weighted linear regression coefficients, $X^2$ test for each regressor, and the goodness of fit of the model.

Responses to questionnaire items by parents whose children attend desegregated public school in Duval county, Florida, are used to exemplify the application of OWLR technique. Comparisons of these OWLR results with the GSK approach indicate that OWLR not only is applicable but also that the computing time can be reduced significantly.

## References

1. Feller, W. (1964). "An Introduction to Probability Theory and Its Application," John Wiley & Sons, Inc., New York.

2. Goldberger, A.S. (1964). "Econometric Theory." John Wiley & Sons, Inc., New York.

3. Grizzle, J.E., Stamer, C.F., and Koch, G.G. (1969). "Analysis of Categorical Data by Linear Model," Biometrics, Vol. 25: pp 489-504.

4. Lehnen, R.G. and Koch, G.G. (1974), "A "A General Linear Approach to the Analysis of Nonmetric Data: Application for Political Science." American Journal of Political Science, May 1974, pp 283-313.

5. Shih, W. F. P., Giles, M. W., Cataldo, E. F. and D. S. Gatlin (1975). "A Contrast of Two Systems of Contrast Coding," Social Statistics Section, Proceeding of the American Statistical Association, pp 640-645.