Science Journal of Applied Mathematics and Statistics
2015; 3(6): 243-249
Published online November 20, 2015 (http://www.sciencepublishinggroup.com/j/sjams)
doi: 10.11648/j.sjams.20150306.13
ISSN: 2376-9491 (Print); ISSN: 2376-9513 (Online)

Incorporating Survey Weights into Binary and Multinomial Logistic Regression Models

Kennedy Sakaya Barasa*, Chris Muchwanju

Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Science and Technology, Nairobi, Kenya

Email address:
sakayak33@gmail.com (K. S. Barasa), chrismuchwanju@gmail.com (C. Muchwanju)

To cite this article:
Kennedy Sakaya Barasa, Chris Muchwanju. Incorporating Survey Weights into Binary and Multinomial Logistic Regression Models. *Science Journal of Applied Mathematics and Statistics*. Vol. 3, No. 6, 2015, pp. 243-249. doi: 10.11648/j.sjams.20150306.13

Abstract: Since sampling weights are not simply equal to the reciprocal of selection probabilities its always challenging to incorporate survey weights into likelihood-based analysis. These weights are always adjusted for various characteristics. In cases where logistic regression model is used to predict categorical outcomes with survey data, the sampling weights should be considered if the sampling design does not give each individual an equal chance of being selected in the sample. The weights are rescaled to sum to an equivalent sample size since original weights have small variances. The new weights are called the adjusted weights. Quasi-likelihood maximization is the method that is used to make estimation with the adjusted weights but the other new method that can be created is correct likelihood for logistic regression which included the adjusted weights. Adjusted weights are further used to adjust for both covariates and intercepts when the correct likelihood method was used. We also looked at the differences and similarities between the two methods. Analysis: Both binary logistic regression model and multinomial logistic regression model were used in parameter estimation and we applied the methods to body mass index data from Nairobi Hospital, which is in Nairobi County where a sample of 265 was used. R-software Version 3.0.2 was used in the analysis. Conclusion: The results from the study showed that there were some similarities and differences between the quasi-likelihood and correct likelihood methods in parameter estimates, standard errors and statistical p-values.

Keywords: Binary Logistic Regression, Multinomial Logistic Regression, Adjusted Weights, Correct Likelihood, Quasi-Likelihood, Nairobi

1. Introduction

1.1. Introduction

Logistic regression provides a method for modeling a binary response variable, which takes values 1 and 0. For example, we may wish to investigate how death (1) or survival (0) of patients can be predicted by the level of one or more metabolic markers. When the response variable is binary (e.g. death or survival), then the probability distribution of the number of deaths in a sample of a particular size, for given values of the explanatory variables, is usually assumed to be binomial. (Courvoisier et al, 2011)

Regression models have become an integral component of any data analysis concerned with describing the relationship between the response variable and one or more explanatory variables. It is often the case that the outcome variable is discrete, taking on two or more possible values. (Hosmer and Lemeshow, 2000).

For example, in a study of obesity for adults, selected individuals can have a high body mass index (BMI) or do not have a high BMI. In such a case BMI will be the independent variable while the independent variables gender, age and race. The dependent variable has two possible outcomes: individuals having a high BMI, not having a high BMI. Subsequently, we can code them as 1 and 0, respectively.

Binary logistic regression model can be extended to multinomial logistic regression model, in which the response variable has more than two levels. The simultaneous increases in obesity in almost all countries seem to be driven mainly by changes in the global food system, which is producing more processed, affordable, and effectively marketed food than ever before. This passive overconsumption of energy leading to obesity is a predictable outcome of market economies predicated on consumption-based growth (Swinburn et al, 2011) Using an example, in the study of obesity for adults the BMI value can be divided into four different levels (obese, overweight, normal, and underweight), then we build the multinomial logistic regression model with gender, age and race as covariates. We labeled the levels as 1, 2, 3 and 4, respectively

Weights always make sure the sample is representative of the population of interest and that other objectives are met and are particularly important when over-sampling occurs. The sampling weights should be considered if the sampling design does not give each individual an equal chance of being selected. Sampling weights can be thought as the number of observations represented by a unit in the population if they are scaled to sum to the population size. According to Gelman (2007) sampling weight is a mess. It is not easy to estimate anything more complicated using weights than a simple mean or ratio, and standard errors are tricky even with simple weighted means. Contrary to what is assumed by many researchers, survey weights are not in general equal to the inverse of probabilities selection, but rather are constructed based on a combination of probability calculations and nonresponsive adjustments.

In this research since the variance was too small, we rescaled the sampling weights to sum to an equivalent sample size. These new weights are called the adjusted weights and are the ones that were incorporated into the logistic regression model to estimate the parameters.

1.2. Quasi-Likelihood and Adjusted Weights

Quasi-Likelihood Method can be used to estimate parameters in the logistic regression model with adjusted weights since the variance function and mean function varies independently. This model estimates the variance function from the data directly without normal distributional assumption.

1.2.1. Probability Weights

Weights may vary for several reasons. The estimator of total will be equal to $\hat{y} = \sum_{i=1}^{n} {}^{y_i}/_{p_i}$, where p_i is the overall probability that the ith element is selected. We can define the sampling weight for the i th element as $w_i = {}^{1}/_{p_i}$. Since Smaller selection probabilities may be assigned to the elements with high data collection costs and a high selection probabilities may be assigned to the elements with larger variances. (Kutner and Nachtsheim, 2004)

1.2.2. Adjusted Weights

Here we don't rely on conditioning on model elements such as covariates to adjust for design effects. But we rescale sample weights to sum to the equivalent sample size so as to obtain estimators The equivalent sample size is smaller than the sample size but in other cases the equivalent sample size could be larger, but we restrict attention to simple random sampling

In the super population model, let $y_i$denote the response variable for the ith unit in the sample. Here, $y_i$are assumed to be independent random variables. Let us define the mean for ith unit$m_i = E\ (y_i)$ and variance$v_i = var\ (y_i),\ i = 1, \dots, n$ where$n$ is the sample size. The mean and variance of the super population model are

$$m = \left(\frac{1}{\sum_{i=1}^{n} w_i}\right) \sum_{i=1}^{n} w_i m_i \ v = \left(\frac{1}{\sum_{i=1}^{n} w_i^2}\right) \sum_{i=1}^{n} w_i^2 v_i$$

The estimate of$m$ and variance of mean are

$$\widehat{m} = \left(\frac{1}{\sum_{i=1}^{n} w_i}\right) \sum_{i=1}^{n} w_i y_i \ var(\widehat{m}) = \left[\frac{\sum_{i=1}^{n} w_i^2}{(\sum_{i=1}^{n} w_i)^2}\right] v.$$

Let us consider another set of weights defined by$w_i^* = \hat{n}\left(\frac{w_i}{\sum_{i=1}^{n} w_i}\right)$, where $\hat{n}$ is $\hat{n} = \frac{(\sum_{i=1}^{n} w_i)^2}{\sum_{i=1}^{n} w_i^2}$. We call w_i^* as the adjusted weights and $\hat{n}$ as an equivalent sample size (Potthoff, Woodbury and Manton 1992).

The equivalent sample size is smaller than the population size. We rescale the sampling weights to sum to an equivalent sample size because the original variance is too small to include enough information. These new weights are called the adjusted weights.

We can rewrite the estimators using $\hat{n}$as

$$m = \left(\frac{1}{\hat{n}}\right) \sum w_i m_i \ \ v = \left(\frac{1}{\hat{n}}\right) \sum w_i^2 v_i$$

$$\widehat{m} = \left(\frac{1}{\hat{n}}\right) \sum w_i y_i \ \ var(\widehat{m}) = \left(\frac{1}{\hat{n}}\right) v$$

1.3. Binary Logistic Regression Model

When have a binary output variable Y, and we want to model the conditional probability Pr(Y = 1|X = x) as a function of x; any unknown parameters in the function are to be estimated by maximum likelihood.

1. First let $p(x)$ be a linear function of x. Every increment of a component of x would add or subtract so much to the probability, p must be between 0 and 1, and linear functions are unbounded.
2. Let log p(x) be a linear function of x, so that changing an input variable multiplies the probability by a fixed amount. Remember that logarithms are unbounded in only one direction, whereas linear functions are not.
3. Finally, we make the logistic (or logit) transformation, $\log\frac{p}{1-p}$ (Which has an unbounded range) a linear function of x.

Formally, the model logistic regression model is that

$$\log\frac{p(x)}{1-p(x)} = \beta_0 + x.\beta$$

Solving for p this gives

$$p(\mathrm{Y} \mid X = x) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)}}$$

Where p= the probability of individual i who has a high BMI with x set of predictors,

e = the base of natural logarithms

β_0 = the constant of the equation and,

x = the coefficient of the predictor variables (Hosmer and Lemeshow, 2000)

1.4. Multiple Logistic Regression Model

The simple logistic regression model can be easily extended, for it to have to more than one predictor variable.

Definition,

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_{p-1} \end{bmatrix}_{p\times 1} \quad X = \begin{bmatrix} 1 \\ X_1 \\ \cdots \\ X_{P-1} \end{bmatrix}_{P\times 1} \quad X_i = \begin{bmatrix} 1 \\ x_{i1} \\ \cdots \\ x_{i,p-1} \end{bmatrix}_{p\times 1}$$

We shall have

$$X'\beta = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$$

$$X_i'\beta = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1}$$

Therefore $E\{Y_i\} = \pi_i = \frac{exp(X_i'\beta)}{1+exp(X_i'\beta)}$

1.5. Multinomial Logistic Regression Model

When the response variables have more than two levels, we still use logistic regression model. We divide the response into J response categories, the variables will be$Y_{i1}, \dots,\ Y_{iJ}$. Then, let J be the baseline, the logit for the j^{th} comparison is:

$$\pi'_{ijJ} = log_e\left[\frac{\pi_{ij}}{\pi_{iJ}}\right] = X'_i\beta_{jJ} j = 1, 2, \dots, J-1$$

$$\pi_{ij} = \frac{exp(X_i'\beta_j)}{1 + \sum_{k=1}^{J-1} exp(X_i'\beta_k)} j = 1, 2, \dots, J-1$$

(Andersson et al. 2010)

1.6. Maximum Likelihood

Maximum likelihood, also called the maximum likelihood method, is the procedure of finding the value of one or more parameters for a given statistic which makes the known likelihood distribution a maximum. (Einicke et al, 2012)

Maximum likelihood methods are used to estimate and make inference about the parameters and these methods are efficient and attractive when the model follows the normal distribution assumption. Since not all the distributions are normal, such as a Poisson distribution, in which the variance is same as the mean the method will not always apply. The mean and variance parameters do not vary independently. (Balgobin and Choi, 2010)

We can recall that, the joint probability function for binary logistic regression is:

$$g(Y_1, \dots, Y_n) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$$

$$log_e g(Y_1, \dots, Y_n) = log_e \prod_{i=1}^{n} f_i(Y_i)$$

$$= log_e \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$$

$$= \sum_{i=1}^{n} [Y_i log_e \pi_i + (1-Y_i) log_e(1-\pi_i)]$$

$$= \sum_{i=1}^{n} \left[Y_i log_e\left(\frac{\pi_i}{1-\pi_i}\right)\right] + \sum_{i=1}^{n} log_e(1-\pi_i)$$

Since we know that $1-\pi_i = \frac{1}{1+exp(\beta_{0+}\beta_1 x_i)}$ And $log_e\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i$

Therefore, $log_e L(\beta_0, \beta_1) = \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} log_e[1 + exp(\beta_0 + \beta_1 x_i)]$here we are trying to find β_0 and β_1 to maximize the log-likelihood function:

$$ln = log_e L(\beta_0, \beta_1)$$

$$= \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_i) - \sum_{i=1}^{n} log_e[1 + \exp(\beta_0 + \beta_1 X_i)]$$

Define:

$$\underset{\sim}{y}U = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{bmatrix} X_U = \begin{bmatrix} X_1^T \\ X_2^T \\ \cdots \\ X_N^T \end{bmatrix}$$

The model is $Y = X^T\beta$. The estimator of B is $\underline{\hat{\beta}} = (X_U^T \Sigma_U^{-1} X_U)^{-1} X_U^T \Sigma_U^{-1} y_U$, where $\Sigma_u is$ a diagonal matrix with ith diagonal element σ_i^2. (Nandram and Choi, 2002)

1.7. Quasi-Likelihood

We analyze the binary logistic regression with sampling weights.

The quasi likelihood is $\prod_{i=1}^{n} f_i(y_i)^{w_i} = \prod_{i=1}^{n} \pi_i^{y_i w_i}(1-\pi_i)^{(1-y_i)w_i}$

$$ln = log_e g(Y_1, \dots, Y_n)^w$$

$$= log_e \prod_{i=1}^{n} f_i(y_i)^{w_i} = log_e \prod_{i=1}^{n} \pi_i^{y_i w_i}(1-\pi_i)^{w_i(1-y_i)}$$

$$= \sum_{i=1}^{n} w_i[y_i log_e \pi_i + (1-y_i) log_e(1-\pi_i)]$$

$$= \left(\sum x_i y_i w_i\right)\beta - \sum w_i log_e\left(1 + e^{x_i\beta}\right)$$

Let $\frac{\partial logln}{\partial\beta}$=$(\sum x_i y_i w_i) - \frac{\sum w_i x_i \beta}{1+e^{x_i\beta}} = 0$

Let $\frac{\partial^2 logln}{\partial^2\beta} = -\frac{\sum_{i=1}^{n} w_i\left[x_i\beta\left(1+e^{x_i\beta}\right) - x_i\beta e^{x_i\beta}\right]}{\left(1+e^{x_i\beta}\right)^2}$

We find estimators to maximize the quasi log-likelihood function: $L(y) = \sum_{i=1}^{n} w_i L(y_i)$.

The estimator of $\beta$is$\underline{\hat{\beta}} = (X_U^T W_U \Sigma_U^{-1} X_U)^{-1} X_U^T W_U \Sigma_U^{-1} \underline{y_U}$.

1.8. Correct Likelihood

When we incorporate the weights into the probability distribution function (pdf), in order to keep the new function still to be a pdf, we need to normalize it. For the discrete distribution, the new function becomes$h(x) = \frac{f(x)^w}{\sum f(t)^w}$, and for the continuous distribution, the new probability distribution function becomes$h(x) = \frac{f(x)^w}{\int f(t)^w dt}$. We introduce the sampling weights in the probability distribution function,

$$h(x) = f(x)^w.$$

Here are some of the distributions and their normalization with sampling weights. We compared their similarities and differences using the mean and variance.

Example 1

Let $x \sim N(\mu, \sigma^2)$with sampling weights (w), The density function of normal distribution is:

$$f(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \ , -\infty < x < \infty$$

Introducing the sampling weights we have:

$$f(x_*, w|\ \mu, \sigma^2) = \frac{\left[\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_*-\mu)^2}{2\sigma^2}}\right]^w}{\int_{-\infty}^{+\infty} \left[\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_*-\mu)^2}{2\sigma^2}}\right]^w dx_*}$$

$$= \frac{\left[e^{-\frac{(x_*-\mu)^2}{2\sigma^2}}\right]^w}{\int_{-\infty}^{+\infty} e^{-\frac{w(x_*-\mu)^2}{2\sigma^2}} dx_*} = \frac{e^{-\frac{w(x_*-\mu)^2}{2\sigma^2}}}{\frac{\sigma}{\sqrt{w}} \cdot \sqrt{2\pi}} = \frac{1}{\sqrt{2\pi}} \cdot \frac{\sqrt{w}}{\sigma} \cdot e^{-\frac{w(x_*-\mu)^2}{2\sigma^2}}$$

Here, $x_* \sim N(\mu, \frac{\sigma^2}{w})$

We see that the mean of normal distribution is$E(x) = \mu$, and the mean of normalized distribution with sampling weights is $E(x_*) = \mu$. There, the mean of normal distribution does not change after normalization. Similarly, the variance of normal distribution is $var(x) = \sigma^2$,and the variance of normalized distribution with sampling weights is$var(x_*) = \frac{\sigma^2}{w}$. The variance of the normal distribution changes after normalization.

Example 2

Let $x \sim Bernoulli\ (p)$ with sampling weights (w), The density function of Bernoulli distribution is:

$$P(X = x \mid p) = p^x(1-p)^{1-x}\ x = 0,1\ 0 \le p \le 1$$

Introducing the sampling weights we have: $p(x_*, w|p) = \frac{[p^{x_*}(1-p)^{1-x_*}]^w}{p^w + (1-p)^w} x^* = 0,1$

Here, $x_* \sim Bernoulli\ \left(\frac{p^w}{p^w+(1-p)^w}\right)$

We see that the mean of Bernoulli distribution is$E(x) = p$. The mean of normalized Bernoulli distribution with sampling weights is $E(x_*) = \frac{p^w}{p^w+(1-p)^{1-w}}$. So that $E(x_*) < E(x)$ or $E(x_*) > E(x)$ or $E(x_*) = E(x)$when $w = 1$.

Example 3

Let $x \sim Multinomial\ (p)$ with sampling weights (w), The density function of Multinomial distribution is:

$p(\underset{\sim}{x}*) = \prod_{j=1}^{k} p_j^{X_j}$, $X_j = 1$, where the unit is j, otherwise 0. Introducing the sampling weights we have:

$$p(\underline{X}) = \frac{\left[\prod_{j=1}^{k} p^{X_j}\right]^w}{\left[\sum_{J_X} \prod_{j=1}^{k} p^{X_j}\right]^w}$$

$$\text{z} = \frac{\prod_{j=1}^{k} p^{X_j w}}{\prod_{j=1}^{k} [p_1^w + p_2^w + \cdots + p_k^w]^{X_j}} = \prod_{j=1}^{k} \left[\frac{p^w}{\sum_{j=1}^{k} p_j^w}\right]^{X_j}$$

Here, $\underset{\sim}{X} \sim Mult(1, q)$, where $q = \frac{p_j^w}{\sum_{j=1}^{k} p_j^w} j = 1, 2, \dots k$

The mean of n independent Bernoulli distributions is equal to p without any sampling weights; it has changed to $q = \frac{p_j^w}{\sum_{j=1}^{k} p_j^w} j = 1,2, \dots, k$ in the presence of the sampling weights.

Example 4

Let $y_* \sim Ber\left(p = \frac{e^{x\beta}}{1+e^{x\beta}}\right)$ with sampling weights (w), The density function of binary logistic regression is:

$$p(Y_* = y_*|\beta) = p^{y_*}(1-p)^{1-y_*}$$

$$= \left[\frac{e^{x\beta}}{1+e^{x\beta}}\right]^{y_*} \cdot \left[\frac{1}{1+e^{x\beta}}\right]^{1-y_*}$$

Introducing the sampling weights we have:

$$p(Y = y|\beta) = \frac{\left[\frac{e^{x\beta}}{1+e^{x\beta}}\right]^{wy} \left[\frac{1}{1+e^{x\beta}}\right]^{w(1-y)}}{\left[\frac{e^{x\beta}}{1+e^{x\beta}}\right]^w + \left[\frac{1}{1+e^{x\beta}}\right]^w}$$

$$= \frac{\left[\frac{e^{x\beta}}{1+e^{x\beta}}\right]^{wy} \left[\frac{1}{1+e^{x\beta}}\right]^{w(1-y)}}{\frac{1+e^{wx\beta}}{(1+e^{x\beta})^w}} = \frac{e^{x\beta wy}}{1+e^{x\beta w}}$$

$$= \frac{e^{x\beta wy}}{[1+e^{x\beta w}]^y} \cdot \frac{[1+e^{x\beta w}]^y}{[1+e^{x\beta w}]}$$

$$= \left[\frac{e^{x\beta w}}{1+e^{x\beta w}}\right]^y \cdot \left[\frac{1}{1+e^{x\beta w}}\right]^{1-y}$$

$$= \left[\frac{e^{xw\beta}}{1+e^{xw\beta}}\right]^y \left[\frac{1}{1+e^{xw\beta}}\right]^{1-y} \ , y = 0, 1$$

The mean of binary logistic regression is $p_* = \frac{e^{x\beta}}{1+e^{x\beta}}$. The mean of normalized binary logistic regression with sampling weights is $p = \frac{e^{x\beta w}}{1+e^{x\beta w}}$. Sampling weights will be used to adjust the covariates and intercepts in the normalized binary logistic regression. When we estimate binary logistic regression coefficients, we multiply sampling weights with both covariates and intercepts to create new covariates and new intercepts, and use correct likelihood estimation methods. In the new method, we normalize the binary logistic regression with adjusted weights and use the correct likelihood to make estimation and inferences. Clearly, there are some differences between correct likelihood of normalized binary logistic regression with adjusted weights and quasi-likelihood of binary logistic regression with adjusted weights.

Example 5

Let $y_*\sim Mult\left(p=\frac{e^{\underset{\sim}{x}'\beta_1}}{1+\sum_{s=1}^{s-1}e^{\underset{\sim}{x}'\beta's}},\dots,\frac{e^{\underset{\sim}{x}'\beta_{s-1}}}{1+\sum_{s=1}^{s-1}e^{\underset{\sim}{x}'\beta's}}\right)$ with sampling weights(w), The density function of multinomial logistic regressions is:

$$y_*\sim Mult\left(1,\frac{e^{\underset{\sim}{x}'\beta_1}}{1+\sum_{s=1}^{s-1}e^{\underset{\sim}{x}'\beta's}},\dots,\frac{e^{\underset{\sim}{x}'\beta_{s-1}}}{1+\sum_{s=1}^{s-1}e^{\underset{\sim}{x}'\beta's}},\frac{1}{1+\sum_{s=1}^{s-1}e^{\underset{\sim}{x}'\beta's}}\right)$$

Introducing the sampling weights we have:

$$p(Y=y|\beta)$$

$$=\frac{\left(\frac{1}{\prod_{s=1}^{s-1}y_s!}\right)^w\left\{\prod_{s=1}^{s-1}\left[\frac{e^{\underset{\sim}{x}'\beta}}{1+\sum_{s=1}^{s-1}e^{\underset{\sim}{x}'\beta}}\right]^y\left[\frac{1}{1+\sum_{s=1}^{s-1}e^{\underset{\sim}{x}'\beta}}\right]^{1-\sum_{s=1}^{s-1}y}\right\}^w}{\left[\frac{e^{\underset{\sim}{x}'\beta}}{1+\sum_{s=1}^{s-1}e^{\underset{\sim}{x}'\beta}}\right]^w+\left[\frac{1}{1+\sum_{s=1}^{s-1}e^{\underset{\sim}{x}'\beta}}\right]^w}$$

$$=\frac{\left(\frac{1}{\prod_{s=1}^{s-1}y_s!}\right)^w\cdot\prod_{s=1}^{s-1}e^{\underset{\sim}{x}'\beta yw}}{e^{\underset{\sim}{x}'\beta w}+1}$$

$$=\left(\frac{1}{\prod_{s=1}^{s-1}y_s!}\right)^w\cdot\frac{\prod_{s=1}^{s-1}e^{\underset{\sim}{x}'\beta y_s w}}{\left(1+e^{\underset{\sim}{x}'\beta w}\right)^{\sum y_s}}\cdot\frac{\left(1+e^{\underset{\sim}{x}'\beta w}\right)^{\sum y_s}}{1+e^{\underset{\sim}{x}'\beta w}}$$

$$=\left(\frac{1}{\prod_{s=1}^{s-1}y_s!}\right)^w\prod_{s=1}^{s-1}\left[\frac{e^{\underset{\sim}{x}'\beta_1 w}}{1+\sum_{s=1}^{s-1}e^{\underset{\sim}{x}'\beta'_s w}}\right]^{y_s}\left[\frac{1}{1+\sum_{s=1}^{s-1}e^{\underset{\sim}{x}'\beta'_s w}}\right]^{1-\sum y_s}$$

For multinomial logistic regression, we take one category as the reference category, then compare others with it. We use sampling weights to adjust the covariates and intercepts in the normalized multinomial logistic regressions. When we estimate multinomial logistic regressions coefficients, we multiply both covariates and intercepts with sampling weights to create new covariates and new intercepts, and use correct likelihood estimation methods. In the new method, we normalize multinomial logistic regression with adjusted weights and use the correct likelihood to make estimation and inferences.

There are some differences between correct likelihood of normalized multinomial logistic regressions with adjusted weights and quasi-likelihood of multinomial logistic regressions with adjusted weights. (Grilli and Pratesi, 2004)

1.9. Summary

The sampling weights of the correct likelihood method are the same as those in the quasi-likelihood method but both of them are adjusted weights.

Quasi likelihood method has the intercepts as 1, and covariates are the regular covariates. However, in the correct likelihood method, the sampling weights are further used to adjust for both covariates and intercepts. In correct likelihood method, we multiply adjusted weights with both intercepts and covariates; the intercepts are equal to adjusted and also the covariates are equal to regular covariates.

Table 1. *Differences between quasi-likelihood and correct likelihood.*

	Quasi-likelihood	Correct Likelihood
Covariates	Regular Covariates	Adjusted Covariates
Intercepts	1	Adjusted Weights
Likelihood	$h(x)=f(x)^w$Not Normalized	$h(x)=\frac{f(x)^w}{\sum f(t)^w}$ Normalized

2. Analysis and Interpretation

2.1. Analysis Using Binary Logistic Regression

A binary logistic regression model was used to analyze the data. We had four levels of BMI, 1, 2, 3 and 4 in which we compared underweight with no underweight or normal weight with not normal weight and so on. In the dataset of Nairobi County we took each variable as 1, and call the others 0, to perform the binary regression. We call BMI equals to 1 as 1, and call BMI equals to 2 to 4 as 0 in Nairobi County , did the first binary regression, then we repeated the process, call BMI equals to 2 as 1, and call BMI equals to 1, 3 and 4 as 0 in Nairobi County did the second binary regression. We kept repeating until the fourth binary regression analyses .The following table shows the results of the binary regression. It shows the differences and similarities between the QLM and CLM. In this study we analyzed the binary and the multinomial logistic regression one by one, basing on the two methods to compare their differences and similarities. We included p-values (Pr), estimates, Wald Chi-Square (WS) statistics and standard errors (SE). The variables age, race and gender are the independent variables for regression of BMI.

From table 2 above we were comparing binary logistic regressions in Nairobi County; we can see differences and similarities between quasi-likelihood and correct likelihood methods. Specific about differences, when BMI equals to three compared to the others, the *p-value* of race is different; the quasi-likelihood methods value was 0.4740 (>0.05), but when using the correct likelihood method we obtained a value of 0.0118 (<0.05). Also in this analysis, we found out that, the p-value of gender was also different; the quasi-likelihood methods value was 0.0186 (<0.05), but new methods (correct likelihood) value was 0.0809 (>0.05). When BMI equals to four compared to the others, the *p-value* of age is different; the quasi-likelihood methods value was 0.0377 (<0.05), but the new was 0.6150 (>0.05). The *p-value* of gender is also different; the quasi-likelihood methods value was 0.0857 (>0.05), while the correct likelihoods method value was 0.0330 (<0.05).

Also from table 2, we can see that there were some similarities, for example, when the BMI was equal to two compared to the others, the p-value of age was the same; the quasi-likelihood methods value was 0.6215 and using the correct likelihood method, we obtained 0.5671, a clear indication that both values obtained were greater that the *p-value* (0.05) When BMI equals to three compared to the others, the p-value of age for the quasi-likelihood method and correct likelihood methods were 0.4195 (>0.05), and 0.3789(>0.05) respectively.

Table 2. *Coefficients of binary logistic regression model (n=265).*

Quasi-Likelihood					**Correct-Likelihood**			
	Estimate	**SE**	**WS**	**Pr**	**Estimate**	**SE**	**WS**	**Pr**
1 vs 2 3 4								
Intercept	-4.711	1.3430	10.341	0.0018	-5.375	1.0402	20.3551	0.0001
Age	-0.006	0.0314	0.0241	0.8675	0.0433	0.0311	2.1511	0.1301
Race	0.1811	0.3781	0.2814	0.6058	0.5311	0.3955	1.0161	0.3155
Gender	1.481	0.7431	3.6571	0.0341	0.4261	0.6135	0.8131	0.2101
2 vs 1 3 4								
Intercept	-1.3105	1.0711	1.6151	0.2157	-0.615	0.7151	0.8315	0.3457
Age	-0.0015	0.0121	0.2123	0.6215	-0.0041	0.00817	0.3451	0.5671
Race	0.2101	0.4131	0.7131	0.7157	0.4715	0.2503	3.1131	0.063
Gender	0.2711	0.5104	0.2501	0.6351	-0.2615	0.3218	0.7505	0.3431
3 vs 1 2 4								
Intercept	2.2511	1.0431	4.5631	0.00315	1.6810	0.7850	4.7483	0.0281
Age	-0.0106	0.014	0.6905	0.4195	-0.0075	0.00849	0.7855	0.3789
Race	-0.3141	0.4751	0.4831	0.4740	-0.9581	0.3751	6.3510	0.0118*
Gender	-1.2451	0.5351	5.5201	0.0186	-0.5448	0.3181	3.0465	0.0812*
4 vs 1 2 3								
Intercept	-3.3165	1.2371	7.4851	0.0069	-2.5671	0.7788	11.1657	0.007
Age	0.0241	0.0118	4.2341	0.0377	0.0036	0.00745	0.2525	0.6150*
Race	-0.2370	0.3499	0.433	0.519	0.154	0.2740	0.3701	0.5440
Gender	0.8070	0.4740	2.9151	0.0857	0.7205	0.3405	4.4811	0.0330*

*Values with asterisk indicate differences, values without asterisk indicate similarities

2.2. Analysis Using Multinomial Logistic Regression

We constructed a multinomial logistic regression model to analyze the four levels of BMI together. The four levels were labeled as 1, 2, 3 and 4 where we compared underweight, normal weight, overweight and obese at the same time. We used BMI ='1' as the reference category and compared it with the other three levels of BMI all together. It is the same to use BMI = '2' as the reference category. In the study, normal weights, overweight and obese were compared basing on the underweight.

Table 3. *Coefficients of multinomial logistic regression model (n=265).*

Quasi-Likelihood					**Correct-Likelihood**			
	Estimate	**SE**	**WS**	**Pr**	**Estimate**	**SE**	**WS**	**Pr**
Intercept 2	3.231	1.7401	3.3670	0.0671	4.1280	1.324	9.6420	0.0015*
Intercept 3	5.4941	1.6300	11.4640	0.0006	5.7941	1.3001	19.8751	0.0001
Intercept 4	1.6015	1.7861	0.7811	0.3801	2.6711	1.2831	4.3391	0.0371*
Age 2	-0.00041	0.0340	0.0004	0.9908	-0.041	0.0171	2.1459	0.1410
Age 3	-0.00312	0.0335	0.006	0.9271	-0.301	0.0184	2.1441	0.1408
Age 4	0.0214	0.0346	0.4171	0.5164	-0.0213	0.0181	1.3811	0.2431
Race 2	0.0343	0.4801	0.0041	0.9421	-0.1201	0.4401	0.0845	0.7711
Race 3	-0.4105	0.5083	0.6516	0.4140	-1.1480	0.5270	4.6750	0.0311*
Race 4	-0.3767	0.4034	0.8740	0.3411	-0.2751	0.4431	0.4041	0.5311
Gender 2	-1.1131	0.7501	2.0501	0.1531	-0.6761	0.5731	1.2811	0.2411
Gender 3	-2.0131	0.8165	6.401	0.0145	-0.851	0.7015	2.151	0.1455*
Gender 4	-0.613	0.8001	0.6651	0.4615	0.0311	0.6171	0.0012	0.9134

Values with asterisk indicate differences, values without asterisk indicate similarities

From table 3 above we can see coefficients of multinomial logistic regression, there were differences and similarities between quasi-likelihood and correct likelihood methods. Specific about differences, when BMI equals to three, the p-value of gender is different; the quasi-likelihood methods value was 0.0145 (<0.05), whereas that of correct likelihood method was 0.1455 (>0.05). The p-value of race is different; the quasi-likelihood methods value was 0.4140 (>0.05), but that of correct likelihood method is 0.0311 (<0.05).

The similarities we see from table 3 above shows that when BMI was equals to two, the p-value of age is the same between these two methods, in the sense that they were both greater than 0.05 since the quasi-likelihood methods value was 0.9908 and correct likelihood method value 0.1410. The p-value of gender was 0.1531 and 0.2411 for the quasi-likelihood and correct likelihood respectively.

3. Discussions and Conclusion

In this study we used quasi-likelihood method as the old method for binary logistic regression model and multinomial logistic regression model. The maximum likelihood methods to make estimation and inference are no longer useful especially when the logistic regression fails to meet normal distribution assumption. As Pfeffermann et al (1998) pointed out maximum likelihood estimation will produce some bias.

The contribution of this research is to use the correct likelihood method as the new method for binary logistic regressions model and multinomial logistic regression model. We put weights in the *pdf*, and in order to keep the new function still a *pdf*, we should divide it by the integral or sum of distribution with weights (i.e., we accommodate the weights by normalization). In the new method (correct likelihood), the weights are further used to adjust the covariates and intercepts. The quasi-likelihood method is the un-normalized distribution with sampling weights, but the correct likelihood method is the normalized distribution with sampling weights. By comparing the results of the two methods from the analysis, we conclude that there are similarities and differences.

The practical examples we used was to diagnose overweight and obesity for adults. The dependent variable is BMI with four levels, underweight, normal weight, overweight and obese. We built binary logistic regression models and multinomial logistic regression models to show the differences and similarities between un-normalized distribution with sampling weights and normalized distribution with sampling weights. We believe using the normalized distribution, the correct likelihood, is the right thing to do, although the use of survey weights is a controversial area. Gelman (2007). It would be nice to compare our methods with the method of post stratification as described by Gelman (2007). One may want to post-stratify the survey weights to get approximately equal survey weights within strata.

Nomenclature

BMI	Body Mass Index
CLM	Correct Likelihood Method
Pdf	Probability density function
QLM	Quasi-Likelihood Method
SE	Standard Error
WS	Wald Chi-square

References

[1] Andersson J.C, Verkuilen, J., and Peyton, B. L. (2010).Modeling Polytomous Item Responses using Simultaneously Estimated Multinomial Logistic Regression Model. *Journal of Educational and Behavioral Statistics*, 422.

[2] Balgobin Nandram and Jai Won Choi. (2010). A Bayesian Analysis of Body Mass Index Data from Small Domains Under Nonignorable Nonresponse and Selection. *Journal of American Statistical Association*, 105, 120-135.

[3] Courvoisier, D.S., C. Combescure, T. Agoritsas, A. Gayet-Ageron and T.V. Perneger (2011) "Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure." Journal of Clinical Epidemiology 64: 993-1000.

[4] Einicke, G.A.; Falco, G.; Dunn, M.T.; Reid, D.C. (May 2012). "Iterative Smoother-Based Variance Estimation". *IEEE Signal Processing Letters* 19 (5)

[5] Gelman, Andrew. (2007) Struggles with Survey Weighting and Regression Modeling. *Statistical Science*, 22, 153-164.

[6] Grilli, L., and Pratesi, M. (2004). Weighted Estimation in Multinomial Ordinal and Binary Models in the Presence of Informative Sampling Designs. *Survey Methodology*, 30, 93-103.

[7] Hosmer D.W and Lemeshow S, (2000), *Applied Logistic Regression* 2nd Ed. John Wiley & Sons, Inc. Canada. PP1-17

[8] Michael, H., Kutner, C J., and Nachtsheim, J. N. (2004).*Applied Linear Regression Models.* McGraw-Hill/Irwin

[9] Nandram, B., and Choi, J. W. (2002), A Hierarchical Bayesian Nonresponse Model for Binary Data With Uncertainty About Ignorability, *Journal of the American Statistical Association*, 97, 381-388.

[10] Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., and Rasbash, J. (1998).Weighting for Unequal Selection Probabilities in Multinomial Models. *Journal of the Royal Statistical Society*, 60, 23-40.

[11] Potthoff R, Woodbury, M. A., and Manton, K. G. (1992). Equivalent Sample Size and Equivalent Degrees of Freedom Refinements for Inference using Survey Weights under Super population Models. *Journal of the American Statistical Association*, 87, 383-396.

[12] Swinburn BA, Sacks G, Hall KD, et al. The global obesity pandemic: shaped by global drivers and local environments. *Lancet*. 2011; 378(9793)

[13] Wang YC, McPherson K, Marsh T, Gortmaker SL, Brown M. Health and economic burden of the projected obesity trends in the USA and the UK. *Lancet*. 2011; 378(9793).