African Journal of Mathematics and Statistics Studies

ISSN: 2689-5323

Volume 3, Issue 4, 2020 (pp. 72-77) www.abjournals.org

ON PROBABILITIES OF MISCLASSIFICATION RELATED TO LOGIT-NORMAL DISTRIBUTION

Awogbemi Clement Adeyeye[1] and Urama Kenny Ugochukwu[2]

[1]National Mathematical Centre, Abuja, Nigeria
[2]Michael Okpara University of Agriculture, Umudike, Abia State, Nigeria
Email of corresponding author: awogbemiadeyeye@yahoo.com

ABSTRACT: *The probabilities of misclassification associated with logit-normal distribution are studied in this work by examining the logistic transformation on normally distributed random variables. The distribution theory and rules in respect of logit-normal distribution for conditional and expected probabilities of misclassification are also considered. While the probability density function (pdf) of conditional probability is generated, the expected probability of misclassification is approximately established.*

KEYWORDS: Logit-normal Distribution, Probabilities of Misclassification, Conditional Probability of Misclassification, Expected Probability of Misclassification, Logistic Transformation.

INTRODUCTION

The logit normal distribution used for modeling compositional data involves transforming the composition by taking logs of ratios of the parts. Applying logistic transformation on a normally distributed random variables gives rise to a logit normal distribution (Emmanuel et al, 2007). The Logistic transformation was originally suggested by Johnson (1949) to analyze responses that are restricted to a finite (0,1) called bounded outcome score. Probabilities of misclassification of associated with number of distributions have been studied at different times by various researchers for different purposes (Mahmoud and Moustafa, 1995; Onyeagu and Adegboye; Awogbemi et al, 2017; Awogbemi, 2020).

In this study, the distribution theory and rules; distributions of conditional and expected probabilities of misclassification are generated for Logit-normal distribution. The probability density function for the conditional probability of misclassification for Logit normal distribution is also being formulated.

Preliminaries

Let $Z = \omega + \varphi \log\left(\frac{\theta - m}{n - \theta}\right)$, where $\omega, \varphi, m, n \in R$ and θ is a score on the interval (m.n). This is to ensure that standard normality is established. Assuming $m = 0,\ n = 1, \omega = 0$ and $\varphi = 1$ and logistic transformation transcends to a Normal (μ, σ). If Z consists of the density $f(z)$ which approximates to $f(z, \beta)$, then θ has a density $k(\theta)$which approximates to $k(\theta, \beta) = f[logit(x)]\frac{1}{\theta(1-\theta)}$, where $logit(\theta) = logit\left(\frac{\theta}{1-\theta}\right)$ (1)

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 3, Issue 4, 2020 (pp. 72-77)
www.abjournals.org

Definition 1

A random variable θ is said to be distributed logit normal (μ, σ^2) if its logit transformation, $K = log\left(\frac{\theta}{1-\theta}\right), 0 < \theta < 1$ is distributed $N(\mu, \sigma^2)$.

The density function of a random variable θ is expressed as

$$f(k) = \frac{1}{(\sqrt{2\pi})\sigma_k} \exp\left[-\frac{1}{2\sigma_k^2}(k-\mu_k)^2\right], -\infty < k < \infty \quad (2)$$

The conditions of equation (2) with $-\infty < \mu_k < \infty$, $\sigma_k^2 > 0$ are satisfied by μ_k, σ_k^2.

The density function is obtained of X is obtained by transforming $K = log\left(\frac{\theta}{1-\theta}\right)$ as

$$f(\theta) = \frac{1}{\theta\,(1-\theta\,)\sigma_k(\sqrt{2\pi})} exp\left[-\frac{1}{2\sigma_k^2}\left(log\left\{\frac{\theta}{1-\theta}\right\} - \mu_k\right)^2\right], \; 0 < \theta < 1 \quad (3)$$

METHODS

Conditional Probability of Misclassification

The conditional probability of misclassification when an observation $\theta \in \pi_1$ is misclassified is expressed as

$$e_{12}(\bar{\theta}_1, \bar{\theta}_2) = \int_{\frac{\bar{\theta}_1+\bar{\theta}_2}{2}}^{1} \frac{1}{\theta(1-\theta)\sigma_k(\sqrt{2\pi})} exp\left[-\frac{1}{2\sigma_k^2}\left\{log\left(\frac{\theta}{1-\theta}\right) - \mu_k^2\right\}\right] d\theta, (\bar{\theta}_1 < \bar{\theta}_2) \quad (4)$$

$$= \int_{0}^{\frac{\bar{\theta}_1+\bar{\theta}_2}{2}} \frac{1}{\theta(1-\theta)\sigma_k(\sqrt{2\pi})} exp\left[-\frac{1}{2\sigma_k^2}\left\{log\left(\frac{\theta}{1-\theta}\right) - \mu_k^2\right\}\right] d\theta, (\bar{\theta}_1 \geq \bar{\theta}_2) \quad (5)$$

When the variable of integration is transformed by taking $S = \frac{log\frac{\theta}{1-\theta} - \mu_k}{\sigma_k}$,

$$e_{12}(\bar{\theta}_1, \bar{\theta}_2) = \int_{log\left[\frac{\frac{\bar{\theta}_1+\bar{\theta}_2}{2-\bar{\theta}_1-\bar{\theta}_2}}{\sigma_k}\right] - \mu_{1k}}^{\infty} \frac{1}{\sqrt{2\pi}} exp\left[-\frac{1}{2}s^2\right] ds, \quad (\bar{\theta}_1 < \bar{\theta}_2) \quad (6)$$

$$= \int_{-\infty}^{log\left[\frac{\frac{\bar{\theta}_1+\bar{\theta}_2}{2-\bar{\theta}_1-\bar{\theta}_2}}{\sigma_k}\right] - \mu_{1k}} \frac{1}{\sqrt{2\pi}} exp\left[-\frac{1}{2}s^2\right] ds, \; (\bar{\theta}_1 \geq \bar{\theta}_2) \quad (7)$$

Equation (5) can be expressed as

$$e_{12}(\bar{\theta}_1, \bar{\theta}_2) = 1 - \Phi\left[\log\frac{\left(\frac{\bar{\theta}_1+\bar{\theta}_2}{2-\bar{\theta}_1-\bar{\theta}_2} - \mu_k\right)}{\sigma_k} - \mu_{2k}\right], (\bar{\theta}_1 < \bar{\theta}_2)$$

$$= 1 - \Phi\left[\log\frac{\left(\frac{\bar{\theta}_1+\bar{\theta}_2}{2-\bar{\theta}_1-\bar{\theta}_2}-\mu_k\right)}{\sigma_k} - \mu_{2k}\right], (\bar{\theta}_1 \geq \bar{\theta}_2) \quad (8)$$

Distribution of Conditional Probability of Misclassification

When $\bar{\theta}_1 < \bar{\theta}_2$ and $\frac{\bar{\theta}_1+\bar{\theta}_2}{2-\bar{\theta}_1-\bar{\theta}_2} > exp[\mu_k - \sigma_k\Phi^{-1}(z)]$, $e_{12}(\bar{\theta}_1, \bar{\theta}_2) \leq$ z.

Also, when $\bar{\theta}_1 \geq \bar{\theta}_2$ and $\frac{\bar{\theta}_1+\bar{\theta}_2}{2-\bar{\theta}_1-\bar{\theta}_2} \leq exp[\mu_k + \sigma_k\Phi^{-1}(z)]$, equation (8) holds.

Thus,

$$Pr[e_{12}(\bar{\theta}_1, \bar{\theta}_2) \leq z] = Pr\left\{\bar{\theta}_2 - \bar{\theta}_1 > 0, \frac{\bar{\theta}_1+\bar{\theta}_2}{2-\bar{\theta}_1-\bar{\theta}_2} \geq exp[\mu_k - \sigma_k\Phi^{-1}(z)]\right\}$$
$$+ Pr\left\{\bar{\theta}_2 - \bar{\theta}_1 \leq 0, \frac{\bar{\theta}_1+\bar{\theta}_2}{2-\bar{\theta}_1-\bar{\theta}_2} \leq exp[\mu_k + \sigma_k\Phi^{-1}(z)]\right\} \quad (9)$$

Given that $d_0 = exp[\mu_k - \sigma_k\Phi^{-1}(z)]$ and $d_1 = exp[\mu_k + \sigma_k\Phi^{-1}(z)]$, then

$$\frac{\bar{\theta}_1 + \bar{\theta}_2}{2 - \bar{\theta}_1 - \bar{\theta}_2} \geq d_0 => \bar{\theta}_1 + \bar{\theta}_2 \geq 2d_0 - \bar{\theta}_1 d_0 - \bar{\theta}_2 d_0$$
$$=> \bar{\theta}_1 + \bar{\theta}_1 d_0 + \bar{\theta}_2 + \bar{\theta}_2 d_0 \geq 2d_0$$
$$=> \bar{\theta}_1(1 + d_0) + \bar{\theta}_2(1 + d_0) \geq 2d_0$$
$$=> (\bar{\theta}_1 + \bar{\theta}_2)(1+d_0) \geq 2d_0$$
$$=> (\bar{\theta}_1 + \bar{\theta}_2) \geq \frac{2d_0}{1+d_0}$$
$$=> \bar{\theta}_{1_1} + \bar{\theta}_2 \leq \frac{2d_1}{1+21} \text{ when} \leq d_1 \quad (10)$$

Equation (8) now results to

$$e_{12}[(\bar{\theta}_1, \bar{\theta}_2) \leq z] \text{ when } (\bar{\theta}_1 + \bar{\theta}_2) \geq \frac{2d_0}{1+d_0}, (\bar{\theta}_1 < \bar{\theta}_2)$$
$$\text{or } (\bar{\theta}_1 + \bar{\theta}_2) \leq \frac{2d_1}{1+d_1}, (\bar{\theta}_1 \geq \bar{\theta}_2) \quad (11)$$

$$Pr[e_{12}(\bar{\theta}_1, \bar{\theta}_2) \leq z] = Pr\left\{\bar{\theta}_2 - \bar{\theta}_1 > 0, (\bar{\theta}_1 + \bar{\theta}_2) \geq \frac{2d_0}{1+d_0}\right\}$$
$$+ Pr\left\{\bar{\theta}_2 - \bar{\theta}_1 \leq 0, \leq \frac{2d_1}{1+d_1}\right\} \quad (12)$$

Probability Density Function of $\boldsymbol{e_{12}(\bar{\theta}_1, \bar{\theta}_2)}$

If $A = \bar{\theta}_1 - \bar{\theta}_2$ and $B = \bar{\theta}_1 + \bar{\theta}_2$, then

$$Pr[e_{12}(\bar{\theta}_1, \bar{\theta}_2) \leq z] = Pr\left\{A > 0,\ B \geq \frac{2d_0}{1+d_0}\right\} + Pr\left\{A < 0,\ B \leq \frac{2d_1}{1+d_1}\right\} \tag{13}$$

By Standardizing the variables A and B,

$$Pr[e_{12}(\bar{\theta}_1, \bar{\theta}_2) \leq z] = Pr\left\{\frac{A-\mu_a}{\sigma_a} > \frac{\mu_a}{\sigma_a}, \frac{B-\mu_b}{\sigma_b} \geq \frac{\frac{2d_1}{1+d_1} - \mu_b}{\sigma_b}\right\}$$

$$+Pr\left\{\frac{A-\mu_a}{\sigma_a} < -\frac{\mu_a}{\sigma_a}, \frac{B-\mu_b}{\sigma_b} \leq \frac{\frac{2d_1}{1+d_1} - \mu_b}{\sigma_b}\right\}$$

$$= Pr\left\{z_1 > -\frac{\mu_a}{\sigma_a}, z_2 \geq \frac{\frac{2d_1}{1+d_1} - \mu_b}{\sigma_b}\right\} + Pr\left\{z_1 < -\frac{\mu_a}{\sigma_a}, z_2 \leq \frac{\frac{2d_1}{1+d_1} - \mu_b}{\sigma_b}\right\} \tag{14}$$

Approximating the distributions of $\bar{\theta}_1, \bar{\theta}_2$ by the normal distribution, and considering large sample sizes, equation (14) becomes

$$Pr[e_{12}(\bar{\theta}_1, \bar{\theta}_2) \leq z] = j_{\rho*}\left[\frac{\mu_a}{\sigma_a}, \mu_b - \frac{2d_0}{1+d_0}\right] + j_{\rho*}\left[-\frac{\mu_a}{\sigma_a}, \frac{\frac{2d_1}{1+d_1} - \mu_b}{\sigma_b}\right], \tag{15}$$

where $\rho* = \frac{\sigma_{ab}}{\sigma_a \sigma_b}$, $\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \sim SBVN\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho* \\ \rho* & 1 \end{pmatrix}\right]$

$$\mu_a = \mu_2 - \mu_1,\ \mu_b = \mu_2 + \mu_1,\ \sigma_a^2 = \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) = \sigma_b^2,\ \sigma_{ab}^2 = \left(\frac{\sigma_1^2}{n_1} - \frac{\sigma_2^2}{n_2}\right)$$

Equation (15) results to

$$Pr[e_{12}(\bar{\theta}_1, \bar{\theta}_2) \leq z] = L[j_{11}, j_{21,} \rho^*] + L[j_{12}, j_{22,} \rho^*] \tag{16}$$

where $j_{11} = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_a^2}} = -j_{12}$; $\quad j_{21} = \frac{(\mu_1+\mu_2) - \frac{2exp[\mu_{1k} - \sigma_k \Phi^{-1}(z)]}{\{1+exp(\mu_{1k} - \sigma_k \Phi^{-1}(z))\}}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$;

$$j_{22} = \frac{\frac{2\{2exp[\mu_{1k} + \sigma_k \Phi^{-1}(z)]\}}{1+exp\left(\mu_{1k} - \sigma_k \Phi^{-1}(z)\right) - (\mu_1 + \mu_2)}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

If $L(z_1, z_{2,}\ \rho^*) = J\rho^*(z_1, z_2)$, then the distribution of $e_{12}(\bar{\theta}_1, \bar{\theta}_2)$ can be expressed as

$$Q(z) = \int_{-\infty}^{j_{21}(z)} \int_{-\infty}^{j_{11}} f(z_1, z_2) dz_1 dz_2 + \int_{-\infty}^{j_{22}(z)} \int_{-\infty}^{-j_{11}} f(z_1, z_2) dz_1 dz_2 \tag{17}$$

The pdf of $e_{12}(\bar{\theta}_1, \bar{\theta}_2)$ is now expressed as

$$q(z) = Q^1(z) \tag{18}$$

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 3, Issue 4, 2020 (pp. 72-77)
www.abjournals.org

where $q(z) = \int_{-\infty}^{j_{11}} f[\, z_1, j_{21}(z)] j_{21}^{1}(z) dz_1 + \int_{-\infty}^{-f_{11}} f[z_1, j_{22}(z)] j_{22}^{1} dz_1$

Equation (18) is strictly an intractable function of z.

Expected Probability of Misclassification of $e_{12}(\bar{\theta}_1, \bar{\theta}_2)$

The expected probability of misclassification is expressed as

$$E[e_{12}(\bar{\theta}_1, \bar{\theta}_2)] = \Pr\{(\bar{\theta}_2 - \bar{\theta}_1) > 0, X - \frac{\bar{\theta}_1 + 2}{2} \geq 0/\theta\epsilon\pi_1\}$$

$$+ \Pr\{(\bar{\theta}_2 - \bar{\theta}_1) \leq 0, \theta - \frac{1}{2}\frac{\bar{\theta}_1 + 2}{2} < 0/\theta\epsilon\pi_1 \qquad (19)$$

By Central Limit Theorem, $\bar{\theta}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$, $\bar{\theta}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$ and with

$g(\theta) = \frac{1}{\theta(1-\theta)\sigma_{k\sqrt{2\pi}}} exp\left\{-\frac{1}{2}\sigma_k^2\left[\log\left(\frac{\vartheta}{1-\theta}\right) - \mu_k\right]\right\}$, $0 < \theta < 1$, the distribution of $\bar{\theta}_1, \bar{\theta}_2$ $and\ \ \theta\ are\ independent$. Thus, the first term by the right-hand side of equation (19) becomes

$$Pr\left\{(\bar{\theta}_2 - \bar{\theta}_1) > 0, \theta - \frac{1}{2}(\bar{\theta}_1 + \bar{\theta}_2) \geq 0\right\} = \int_0^1 Pr\{(\bar{\theta}_2 - \bar{\theta}_1) > 0, (\bar{\theta}_1 + \bar{\theta}_2) \leq 2\theta \backslash \theta\} g(\theta)\}\, d\theta \qquad (20)$$

Let the probability in the integrand be denoted by

$$P(\theta) = Pr\{A > 0, B \leq 2\theta\}, \qquad (21)$$

where $\begin{pmatrix} A \\ B \end{pmatrix} \cong BVN\left[\begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix}\right]$

By standardizing the variables, the probability

$$P(\theta) = Pr\left\{z_1 > \frac{\mu_a}{\mu_b}, z_2 < \frac{2\theta - \mu_b}{\sigma_b}\right\} = \int_{-\frac{\mu_a}{\sigma_a}}^{\infty} \int_{-\infty}^{\frac{2\theta - \mu_b}{\sigma_b}} j_\rho(z_1, z_2) \mathrm{d}z_1 \mathrm{d}z_2$$

$$= \int_{-\infty}^{\frac{\mu_a}{\sigma_a}} \int_{-\infty}^{\frac{2\theta - \mu_b}{\sigma_b}} j_{-\rho}(z_1, z_2) \mathrm{d}z_1 \mathrm{d}z_2$$

$$= J_{-\rho}\left[\frac{\mu_a}{\sigma_a}, \frac{2\theta - \mu_b}{\sigma_b}\right] \qquad (22)$$

The last term of equation (19) becomes

$$Pr\{\bar{\theta}_2 - \bar{\theta}_1 \leq 0, \bar{\theta}_2 + \bar{\theta}_1 \geq 2\theta\ \} = \int_0^1 p^1(\theta) g(\theta)\, d\theta, \qquad (23)$$

where $p^1(\theta) = \Pr\{(A) \leq 0, B \leq 2\theta\} = \Pr\{z_1 \leq \frac{\mu_a}{\sigma_a}, z_2 \geq \frac{2\vartheta - \mu_b}{\sigma_b}\}$

$$= \int_{-\infty}^{-\frac{\mu_a}{\sigma_a}} \int_{\frac{2\theta - \mu_b}{\sigma_b}}^{\infty} j_{-\rho}(z_1, z_2) \mathrm{d}z_1 \mathrm{d}z_2$$

$$= \int_{-\infty}^{-\frac{\mu_a}{\sigma_a}} \int_{-\infty}^{-\frac{2\theta-\mu_b}{\sigma_b}} j_{-\rho}\,(z_1, z_2)\mathrm{d}z_1\mathrm{d}z_2$$

$$= J_{-\rho}\left[-\frac{\mu_a}{\sigma_a},\ \frac{\mu_b-2\theta}{\sigma_b}\right] \quad (24)$$

Thus, the approximation for the probability of misclassification is given as

$$E[e_{12}(\bar{\theta}_2 - \bar{\theta}_1)] \approx \int_0^{\infty} g(\theta)\left[J_{-\rho}\,\langle\frac{\mu_a}{\sigma_a}, \frac{2\theta-\mu_b}{\sigma_b}\rangle\right] + J_{-\rho}\left[-\frac{\mu_a}{\sigma_a},\ \frac{\mu_b-2\theta}{\sigma_b}\right] d\theta, \quad (25)$$

where $g(\theta) = \frac{1}{\theta(1-\theta)\sigma_{k\sqrt{2\pi}}} exp\left\{-\frac{1}{2}\sigma_k^2\left[\log\left(\frac{\vartheta}{1-\theta}\right) - \mu_k\right]\right\},\ 0 < \theta < \infty$

CONCLUSION

Probabilities of misclassification associated with logit-normal distribution have been considered in respect of distribution theory, conditional and expected probabilities of misclassification. While the probability density function of conditional probability of misclassification was generated, the approximation of the expected probability of misclassification is effectively established.

REFERENCES

Awogbemi, C.A. (2020). Graphical Evaluation of Probabilities of Misclassifications for Normal and Edgeworth Series Distributions, Int. Journal of Research and Innovation in Applied Science, 5(4), pp 111-121.

Awogbemi, C.A., Urama, K.U. and Alagbe, S.A. (2017). Errors of Misclassification Associated with Log-normal Distribution, *ABACUS MAN*, 44(2), pp 88 - 97.

Emmanuel, L., Dimitris, R. and Roula, T. (2007). The Logistic Transformation for Bounded Outcomes Scores, *Biostatistics,* 2, pp. 72-85.

Johnson, N. (1949). Systems of Frequency Curves Generated by Methods of Translation, *Biometrika*, 36, pp. 149 -176.

Mahmoud, M.A. and Moustafa, H.M. (1995). Errors of Misclassification Associated with Gamma Distribution, *Journal of Mathematical Computing and Modeling*, 22(3), pp. 105-119.

Nwosu, C., Utazi, C. and Afuecheta, E. (2016). *Probability and Distribution Theory*, 2nd ed. Afrika-Link Communication Ltd, Enugu, Nigeria.

Onyeagu, S.I. and Adeboye, O.S. (1996). Some Methods of Estimating Probability of Misclassification in Discriminant Analysis, *ABACUS, MAN*, 24(2), pp 104 – 112.

Steven, E.A. (1990). *Mathematical Statistics*, Prentice Hall Inc. New Jersey.