Proceedings of the 4th International Conference on Statistics: Theory and Applications (ICSTA'22)
Prague, Czech Republic – July 28- 30, 2022
Paper No. 131
DOI: 10.11159/icsta22.131

Longitudinal Beta GEE Modelling for Analysing Global and Regional Prevalence of Anaemia in Women

Eliana Ibrahimi[1], Jona Shkurti[2], Aldiona Kërri[3], Thao Mai Phuong Tran[4]
[1]Department of Biology, University of Tirana
Bulv Zogu I, 25/1, Tirana, Albania
eliana.ibrahimi@fshn.edu.al
[2]The Netherlands Cancer Institute – Antoni van Leeuwenhoek Hospital
Plesmanlaan 121, 1066 CX, Amsterdam, The Netherlands
j.shkurti@nki.nl
[3]EHESP School of Public Health
20 avenue George Sand – 93 210 La Plaine Saint-Denis, Paris, France
aldiona.kerri@gmail.com
[4] Epidemiology & Pharmacovigilance, P95, Leuven, Belgium.
thao.tran@p-95.com

Abstract- In this study, we use a beta regression approach to model the worldwide longitudinal prevalence of anaemia in pregnant and non-pregnant women. The estimates of anaemia prevalence from 1990 to 2016 are extracted for each country from the WHO Data Repository. Since the data for the subjects (i.e., countries) are clustered within sampling units, and the measurements within the same country are correlated, a beta-distributed Generalized Estimating Equation (GEE) model allowing for a population-averaged interpretation of the regression coefficients is fitted. The analysis is implemented in the SAS GLIMMIX procedure. Regardless, parameter coefficients in the GEE are estimated invariably; even if the covariance structure is miss-specified, a careful selection of the working correlation structure is performed to improve the efficiency of the estimates. Pregnancy and WHO regions had significant effects on the prevalence of anaemia. The significant interaction between pregnancy and time suggested that the decline in prevalence over time was larger in non-pregnant women than in pregnant women.

***Keywords*:** Beta regression, marginal model, the longitudinal prevalence of anaemia, women health

1. Introduction

Anaemia is a condition in which the haemoglobin (Hb) concentration in the blood falls below established cut-off values, thereby compromising the capacity of the blood to deliver oxygen to tissues. It affects approximately one-third of the world's population and accounts for about 9% of the total global disability burden from all conditions [1]. As a result, anaemia has significant consequences for human health, as well as for the socio-economic development of countries. The 2016 estimates indicate that anaemia affects 33% of women of reproductive age globally. In Africa and Asia, the prevalence is the highest at over 35% [1]. Iron deficiency, haemoglobinopathies and malaria are considered the three top causes of anaemia globally, with iron deficiency comprising about 50% of the total number of cases [1]. Women of reproductive age (15–49 years) and pregnant women are among the most vulnerable population groups to develop anaemia due to iron deficiency. Regular blood loss due to menstruation, pregnancy, childbirth bleeding, and diets that are low in bioavailable iron, may cause significant iron deficiency [2]. Pregnant adolescents are at particular risk of developing anaemia [3], not only due to their dual iron requirements but also because they are less likely to access antenatal care. Anaemia in the first or second trimester significantly increases the risk of low birth weight and preterm birth [4]. Postpartum anaemia is associated with decreased quality of life [5] and it subjects women to a greater risk of postpartum depression [6]. Severe anaemia, which is associated with substantially worse mortality, cognitive and functional outcomes, affects 0.8–1.5% of these population groups [7].

Research on anaemia prevalence is done constantly to inform stakeholders and policy decision-makers on the necessary type of measures needed to prevent and control anaemia. The causes of anaemia may vary by country and they need to be accurately identified, to implement tailored prevention and treatment strategies according to population characteristics.

The beta regression models are shown to be a good choice when modelling prevalence data which are within the interval [0-1]. The beta distribution is a continuous probability distribution defined on the interval [0, 1] with density function [8]

$$f(y;\varphi,\mu) = \frac{\Gamma(\varphi)}{\Gamma(\mu,\varphi)\,\Gamma((1-\mu)\varphi)}\, y^{\mu\varphi-1}\,(1-y)^{(1-\mu)\varphi-1}, \qquad (1)$$

$$0 < y < 1,$$

where Γ (.) denotes the gamma function, the parameter μ denotes the expected value of Y, i.e. E(Y) = μ, the parameter φ is a precision parameter.

Through this study, we aim to assess the potential of a beta regression model to estimate the worldwide trends of anaemia prevalence from 1990 to 2016 in women of reproductive age, by pregnancy status and by regions of the World Health Organization (WHO). Since we deal with longitudinal data and we are interested in the population-averaged interpretation of the regression coefficients, we extended the beta regression approach to a beta-distributed GEE model. In section 2.2 we give more details on the Beta GEE model from the theoretical point of view and application.

2. Methods

2.1. Data Description and Exploration

The estimates of the prevalence of anaemia are extracted from the WHO Data Repository; a huge dataset that provides health-related statistics for WHO's member-states [9]. This repository collects data from as many sources as possible, including scientific literature and collaborators. Briefly, collaborators were WHO regional and country offices, UN organizations, not for profits, ministries of health, research and academic institutions [10]. A systematic search of MEDLINE and WHO regional database, as well as hand-searching of grey literature, provided related articles. Inclusion criteria were Hb measured in capillary, venous, or cord blood using quantitative photometric methods or automated cell counters. Data sources were representative of any administrative level within each country. A detailed description of the study design, sampling method and data collection can be found elsewhere [10]. In this study, we use data for the estimated prevalence in pregnant and non-pregnant women, from 1990 to 2016, 27-time points. The data comes from 186 countries in the six WHO regions. In total, we have 10044 observations, 5022 for pregnant and 5022 for non-pregnant women, and 372 observations per year.

Figure 1 shows the observed mean profiles of the prevalence of anaemia by pregnancy and WHO region where we can see the differences between pregnant and non-pregnant women and different WHO regions. The prevalence of anaemia is higher over time for pregnant women compared to non-pregnant women (Figure 1). The European region and the Americas show lower prevalence over time compared to the other WHO regions.

2.2. Beta GEE modelling

Beta regression approach was used to model the worldwide longitudinal prevalence of anaemia. Here, we define $g(\mu_{ij}) = \log(\mu_{ij}/1-\mu_{ij}) = x^T_{ij}\,\beta$, where: μ_{ij} is the expected value for the dependent variable Y_{ij} (i.e., the expected value for the prevalence of anaemia); x^T_{ij} is a vector of covariates (i.e., time, pregnancy, region, time*pregnancy), β denotes the vector of regression coefficients; i=1, 2, …, n denotes the country; j = 1, 2, …, k_i denotes the measurement within a country [11, 12]. Since the data within the same country are correlated, a beta-distributed Generalized Estimating Equation (GEE) model is considered to be appropriate to capture the within-cluster correlation and to allow for a population-averaged interpretation of the regression coefficients [13, 14]. The variance function is defined as $\mathrm{Var}(Y_{ij}|x_{ij}) = \varphi\mu_{ij}(1-\mu_{ij})$ where parameter φ fulfils the definition of a precision parameter. A larger scaling factor indicates substantial heterogeneity which might not be captured if a scaling factor is not used [15, 16]. The most common working correlation matrix used for longitudinal data are compound symmetry, unstructured correlation, and the autoregressive structure [12].

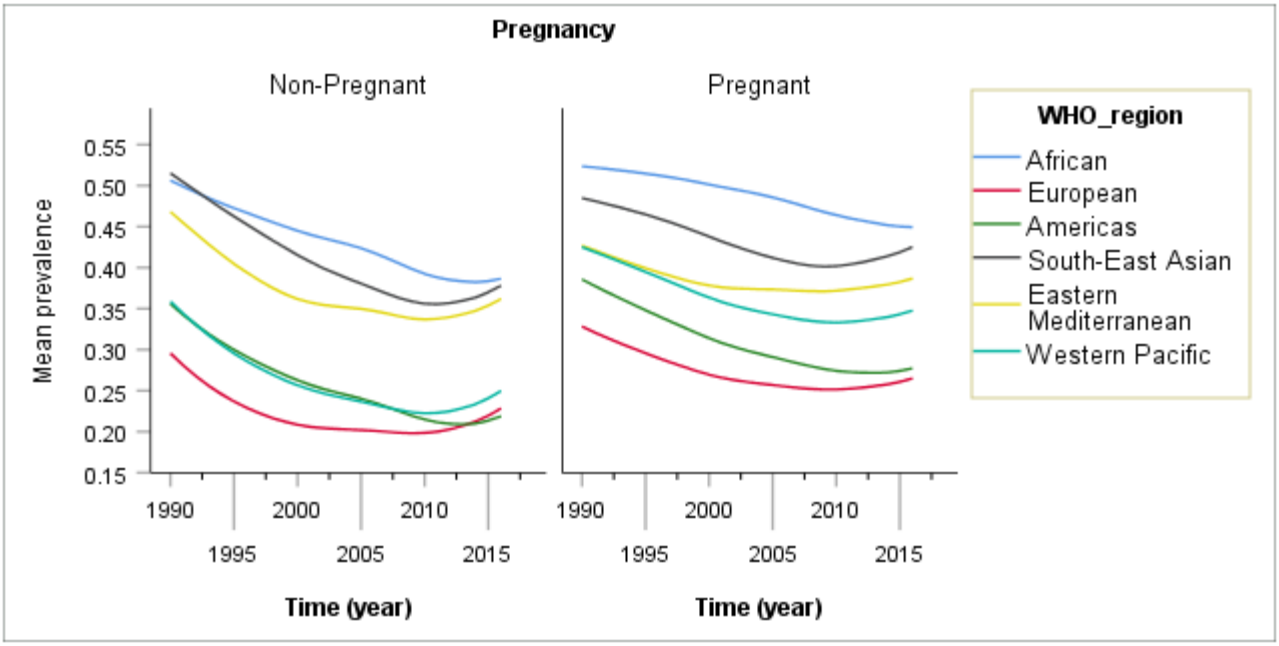

Fig. 1: Global mean profiles of anaemia prevalence by pregnancy and WHO region

The variance function and correlation matrix are then added into a 'working' covariance matrix V_i [12], and the parameter estimates in the marginal model are estimated by solving the generalized estimating equations introduced by Liang and Zeger [17] given in equation 1.

$$\sum_{i=1}^{N} D_i^T V_i^{-1} (Y_i - \mu_i) = 0 \qquad (2)$$

where $Y_i = (Y_{i1}, \ldots, Y_{in})^T$, $\mu_i = (\mu_{i1}, \ldots, \mu_{ini})^T$, $D_i = D_i(\beta) = \partial u_i(\beta) / \partial \beta^T$, and V_i is the working covariance matrix.

For this type of model, there are no closed-form solutions in general, so iterative algorithms are used. The correlation structure is chosen by using the analysis of model-based and empirical standard errors [12]. Nevertheless, valid standard errors for β could be obtained by the sandwich estimator approach [16].

The analysis is implemented in the SAS GLIMMIX procedure [18]. Although parameter coefficients in the GEE are estimated invariably even if the covariance structure is miss-specified, we still carried out a careful selection of the working correlation to improve the efficiency of the estimates. We considered several working correlation structures such as compound-symmetry, unstructured correlation structure, a Toeplitz covariance structure, and the first-order autoregressive (AR (1)) structure [18]. Based on the Hannan-Quinn information criterion (HQIC) we considered the AR (1) structure which has homogeneous variances and correlations that decline exponentially with distance, the best to be used in our final model. This correlation structure implies that the variability is constant over time and that the correlation between measurements decreases when the time lag between them increases.

3. Results

Type III tests of fixed effects indicate that all covariates (pregnancy, region, time, time*pregnancy) effects are highly significant ($p<0.0001$ in all cases). We observe a decline in time for the global prevalence of anaemia from 1990 to 2010, followed by a slight increase from 2011 to 2016. The significant interaction between pregnancy and time suggests that the decline in prevalence over time was more substantial in non-pregnant women compared to pregnant women, as shown by the observed and predicted mean profile plots of the prevalence of anaemia by pregnancy given in Figure 2.

Here we can interpret the parameter estimates like in the case of logistic regression where exponentiated coefficients can be interpreted in terms of odds ratios. The parameter coefficient for non-pregnant women means that for a non-pregnant woman, the ratio between the expected anaemia prevalence μ and the difference 1-μ is about 33% lower than for a pregnant woman with the same set of covariates, exp (-0.2535) = 0.77 (see Table 2).

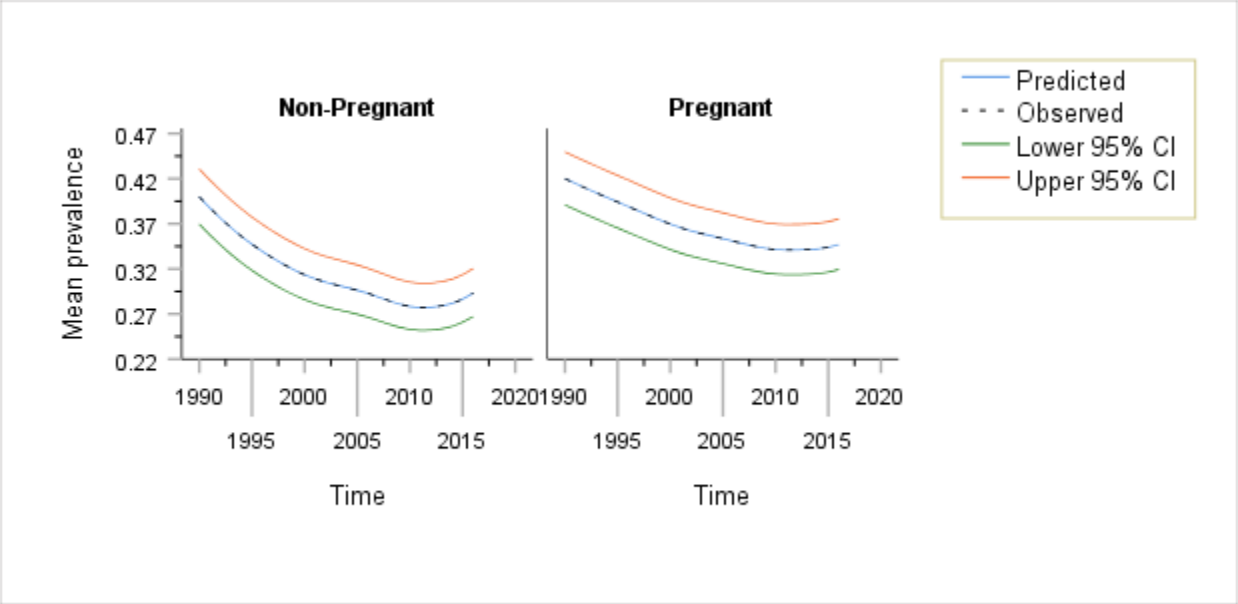

Fig. 2: Observed and predicted mean profile plots of the prevalence of anaemia by pregnancy.

The African region shows the highest prevalence of anaemia over time compared to other regions, with the highest difference observed with the European region. For the African region, the odds to have a higher prevalence are 2.7 times higher compared to Europe and 1.95 times higher compared to the Americas. Results show no significant difference between Africa and South-East Asia (p=0.404), with the same ratio between the expected anaemia prevalence μ and the difference 1-μ. The lowest prevalence of anaemia is observed in Europe (Table 2), where the odds to have a high anaemia prevalence are lower compared to all the other regions (OR<1 in all comparisons).

The observed and predicted mean profile plots of the prevalence of anaemia for pregnant women by WHO region show a slight overestimation for South-East Asian and the Eastern Mediterranean regions. In Eastern Mediterranean region we observe a slight underestimation for non-pregnant women (Figure 3).

4. Discussion

In this study, we examined the potential of beta GEE in analysing the longitudinal prevalence of anaemia in pregnant and non-pregnant women. Our results show that beta regression is a promising method for modelling the longitudinal prevalence of anaemia, confirming the findings of previous studies which have shown this method to be effective for continuous bounded responses with dependent observations [19].

Longitudinal designs typically produce correlated observations, which do not meet the assumptions of ordinary regression methods. Therefore, to model longitudinal data we should apply regression methods that count for this correlation.

Table 2: Comparison parameter estimates by pregnancy status and WHO regions

Comparison	**Estimate**	**Standard error**	**Pr> |t|**	**Odds Ratio**
Non-pregnant Vs. Pregnant	-0.2535	0.01695	<.0001	0.7761
African Vs. European	1.0051	0.08089	<.0001	2.7321
African Vs. Americas	0.6699	0.08448	<.0001	1.9540
African Vs. South-East Asian	0.1311	0.1572	0.4041	1.1401
African Vs. Eastern Mediterranean	0.4027	0.08576	<.0001	1.4958
African Vs. Western Pacific	0.6079	0.1234	<.0001	1.8366
European Vs. Americas	-0.3352	0.08354	<.0001	0.7152
European Vs. South-East Asian	-0.8739	0.1567	<.0001	0.4173
European Vs. Eastern Mediterranean	-0.6024	0.08491	<.0001	0.5475
European Vs. Western Pacific	-0.3972	0.1227	0.0012	0.6722
Americas Vs. South-East Asian	-0.5387	0.1586	0.0007	0.5835
Americas Vs. Eastern Mediterranean	-0.2672	0.08837	0.0025	0.7655
Americas Vs. Western Pacific	-0.06193	0.1251	0.6205	0.9399
South-East Asian Vs. Eastern Mediterranean	-0.5387	0.1586	0.0007	0.5835
South-East Asian Vs. Western Pacific	-0.2672	0.08837	0.0025	0.7655
Eastern Mediterranean Vs. Western Pacific	0.2715	0.1593	0.0883	1.3120

The correlation among repeated observations can be modelled by including random effects as in the linear mixed models, or through a covariance structure as in the marginal models [12]. We emphasize the need to use marginal models, namely beta GEE when dealing with longitudinal data with the aim of comparing populations. For beta regression, the interpretation of the regression coefficients depends on whether a generalized linear mixed model or a marginal model, i.e., beta GEE, is fitted.

It is important to be aware that in contrast to the generalized linear mixed model, in the marginal model the mean response is conditional only on covariates and not on random effects [20]. It should also be mentioned that the beta regression does not contain the boundary points 0 and 1, so in case we have these values in our data, transformation methods need to be applied. However, in our case, there was no need for transformation as the data did not contain 0 and 1 observations and the values fell sufficiently far away from the boundaries.

In the healthcare context, we observed a decline in time for the global prevalence of anaemia from 1990 to 2010, followed by a slight increase from 2011 to 2016. In 1990, the prevalence of anaemia was higher in Africa and South Asia, but with time it decreased in these and other regions, leading to a modest global reduction in anaemia prevalence. Global anaemia estimates reported by WHO with data from 1993 to 2005 have remained consistent with our estimates of the African region, showing the highest prevalence of anaemia over time compared to other regions. Europe has instead surpassed the Americas in reporting the lowest prevalence [10].

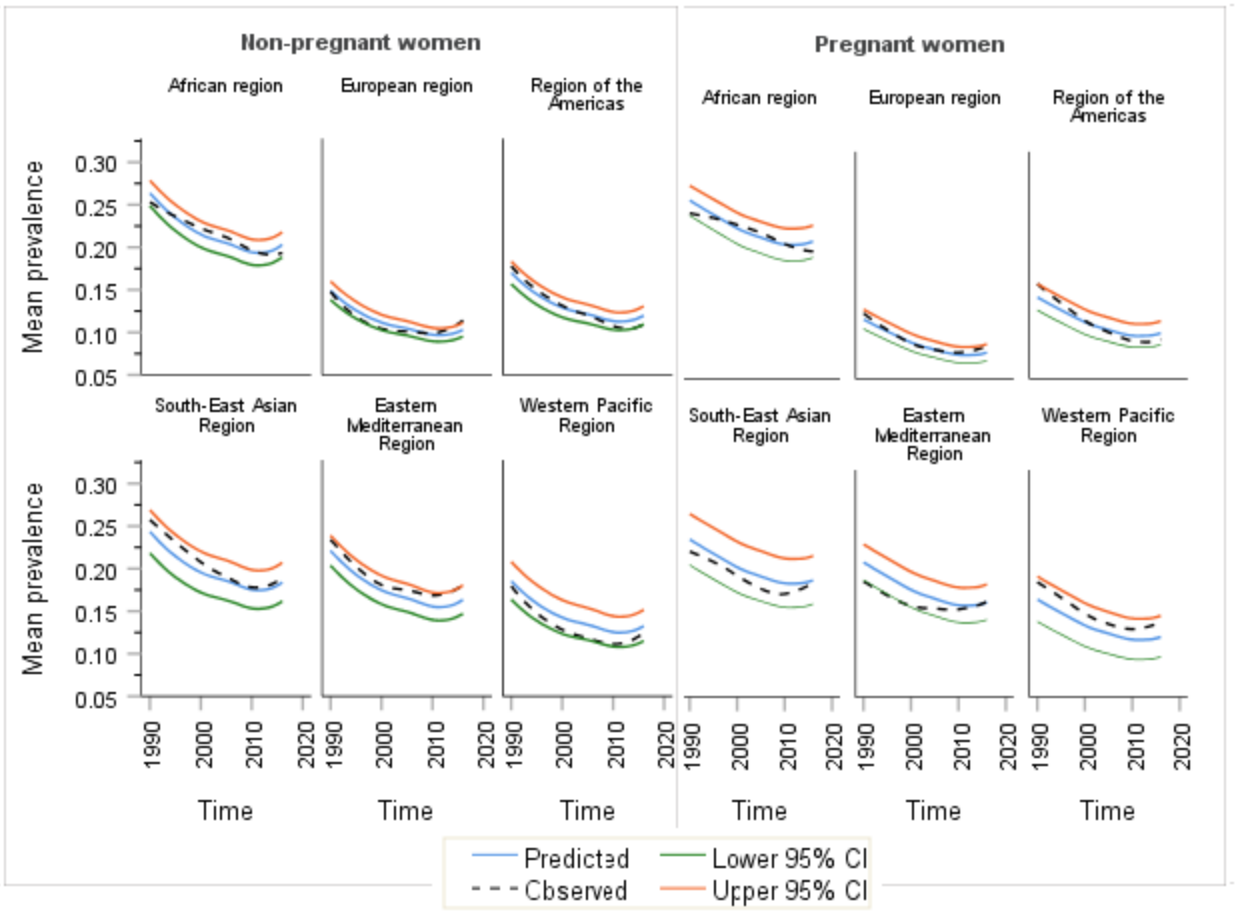

Fig. 3: Observed and predicted mean profile plots of the prevalence of anaemia by pregnancy and WHO region.

The significant interaction between pregnancy and time suggests that the decline in prevalence over time was larger in non-pregnant women compared to pregnant women. Several studies have reported an increased prevalence of anaemia in pregnant women, with the leading factor the iron deficiency [10, 21, 22, 23, 24]. Prevention and control of anaemia require an integrated approach that addresses the primary causes of anaemia in a population. Iron deficiency anaemia should be ideally treated through diet and access to foods that contain a high level of bioavailable iron. To address this critical global health issue in women and children, current strategies should also focus on other contributing factors, i.e., malaria, education, household wealth status [10, 25, 27, 27, 28, 29, 30, 31].

4. Conclusions

In conclusion, beta GEE models can be a natural candidate to model the prevalence of anaemia over time when a population-averaged approach is preferred. These models can deal with the bounded range and the skewed distribution of the response. The global prevalence of anaemia has decreased over time in both pregnant and non-pregnant women but is still high in the poorest regions of the world, being an obstacle to controlling maternal mortality. The slight increase from 2011 to 2016 indicates that further improvements are needed, likely through a combination of programs that address multiple factors contributing to worldwide high levels of anaemia prevalence.

References

[1] FAO, IFAD, UNICEF, WFP, WHO. (2017). The state of food security and nutrition in the world 2017. Building resilience for peace and food security. Rome, FAO [Online]. Available: http://www.fao.org/3/a-I7695e.pdf.

[2] L. E. Torheim, E. L. Ferguson, K. Penrose, M. Arimond, "Women in resource-poor settings are at risk of inadequate intakes of multiple micronutrients, " *J. of Nutrition*., vol. 140, no. 11, pp. 2051S–2058S, 2010.

[3] D. Jeha, I. Usta, L. Ghulmiyyah, A. Nassar, "A review of the risks and consequences of adolescent pregnancy," *J. Neonatal-perinatal med*., vol. 8, no. 1, pp.1-8, 2015.

[4] B. A. Haider, I. Olofin, M. Wang, D. Spiegelman, M. Ezzati, W. W. Fawzi, "Anaemia, prenatal iron use, and risk of adverse pregnancy outcomes: systematic review and meta-analysis," *BMJ*, 346, f3443, 2013.

[5] N. Milman, "Postpartum anaemia: definition, prevalence, causes, and consequences," *J. Ann Hematol*, vol. 90, no. 11, pp. 1247–53, 2011.

[6] E. J. Corwin, L. E. Murray-Kolb, J. L. Beard, "Low haemoglobin level is a risk factor for postpartum depression," *J. Nutrition*, vol. 133, no. 12, pp. 4139–4142, 2003.

[7] G. Stevens, M. Finucane, L. De-Regil, C. Paciorek, S. Flaxman, F. Branca, et al, "Global, regional, and national trends in haemoglobin concentration and prevalence of total and severe anaemia in children and pregnant and non-pregnant women for 1995–2011: a systematic analysis of population-representative data," *J. Lancet Glob Health*, vol.1, no. 1, pp.16–25, 2013.

[8] S. Ferrari, F. Cribari-Neto,"Beta regression or modeling rates and proportions," *J Appl Statist*, vol. 31, no.7, pp.799–815, 2004.

[9] World Health Organization. (2021). Data repository [Online]. Available: www.who.int/data/gho/info/about-the-observatory

[10] World Health Organization. (2008). Worldwide prevalence of anaemia 1993-2005: WHO global database on anaemia/ Edited by Bruno de Benoist, Erin McLean, Ines Egli and Mary Cogswell [Online]. Available: https://apps.who.int/iris/handle/10665/43894.

[11] M. Smithson, J. Verkuilen, "A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables," *J. Psychology Methods*, vol. 11, no. 1, pp. 54–71, 2006.

[12] M. Hunger, A. D¨oring, R. Holle, "Longitudinal beta regression models for analyzing the health-related quality of life scores over time,"*J. BMC Medical Research Methodology*, vol. 12, no. 144, 2012.

[13] J.Verkuilen, M. Smithson, "Mixed and mixture regression models for continuous bounded responses using the beta distribution," *J. Educational and Behavioral Statistics*, vol. 37, no. 1, pp. 82–113, 2012.

[14] D. Zimprich, "Modeling change in skewed variables using mixed beta regression models," *J. Research in Human Development*, vol. 7, no.1, pp. 9–26, 2010.

[15] J. C. Gardiner, Z. Luo, L. A. Roman, "Fixed effects, random effects and GEE: What are the differences*?," J. Statistics in Medicine*, vol. 28, no. 2, pp. 221–239, 2009.

[16] G. Molenberghs, G.Verbeke, "Models for discrete longitudinal data," New York: Springer. 2005.

[17] K.Y. Liang, S. L. Zeger, "Longitudinal data analysis using generalized linear models," *J. Biometrics*, vol. 73, no. 1, pp.13–22, 1986.

[18] SAS Institute Inc. (2013). SAS/STAT® 13.1 User's Guide [Online]. Available: https://documentation.sas.com/api/collections/pgmsascdc/9.4_3.3/docsets/statug/content/glimmix.pdf?locale=en#nameddest=statug_glimmix_syntax

[19] J. Verkuilen, M. Smithson, "Mixed and mixture regression models for continuous bounded responses using the beta distribution," *J. Educ Behav Stat,* vol.37, no. 1, pp. 82–113, 2012.
[20] G. M. Fitzmaurice, L. M. Laird, J.H. Ware, "Applied Longitudinal Analysis". 2nd edition. New York: Wiley. 2011.
[21] C. H. H. Le, "The prevalence of anaemia and moderate-severe anaemia in the US population (NHANES 2003-2012). *J. PLoS ONE,* vol. 11, no.11, e0166635, 2016.
[22] G. A. Stevens, M. M. Finucane, L. M. De-Regil, C. J. Paciorek, S. R. Flaxman, F. Branca, J. P. Peña-Rosas, Z. A. Bhutta, M. Ezzati, & Nutrition Impact Model Study Group (Anaemia), "Global, regional, and national trends in haemoglobin concentration and prevalence of total and severe anaemia in children and pregnant and non-pregnant women for 1995-2011: a systematic analysis of population-representative data,"*J. Lancet. Global health,* vol. 1, no. 1, pp.16–25, 2013.
[23] H. H. Win, M. K. Ko, "Geographical disparities and determinants of anaemia among women of reproductive age in Myanmar: analysis of the 2015-2016 Myanmar demographic and health Survey, "J. *WHO South-East Asia journal of public health,* vol. 7, no. 2, pp.107–113, 2018.
[24] K. Singh, Y. F. Fong, S. Arulkumaran, "Anaemia in pregnancy--a cross-sectional study in Singapore,". *Eur J Clin Nutr,* vol. 52, no. 1, pp. 65–70, 1998.
[25] S. A. Herzog, G. Leikauf, H. Jakse, A. Siebenhofer, M. Haeusler, A. Berghold, "Prevalence of anaemia in pregnant women in Styria, Austria-A retrospective analysis of mother-child examinations 2006-2014", *J PloS one,* vol. 14, no. 7, e0219703, 2019.
[26] C. M. Chaparro, P. S. Suchdev, "Anaemia epidemiology, pathophysiology, and etiology in low- and middle-income countries," *J. Annals of the New York Academy of Sciences,* vol. 1450, no.1, pp. 15–31, 2019
[27] F. Yang, X. Liu, P. Zha, "Trends in socioeconomic inequalities and prevalence of anaemia among children and non-pregnant women in low- and middle-income countries," *J. JAMA network open,* vol. 1, no. 5, e182899, 2018.
[28] Q. Ma, S. Zhang, J. Liu, Q. Wang, H. Shen, Y. Zhang, M. Liu, "Study on the prevalence of severe anaemia among non-pregnant women of reproductive age in rural china: A large population-based cross-sectional study, " *J. Nutrients,* vol. 9, no. 12, pp.1298, 2017
[29] R. E. Ogunsakin, B. T. Babalola, O. Akinyemi, "Statistical modeling of determinants of anaemia prevalence among children aged 6-59 months in Nigeria: A cross-sectional study. *J. Anemia*, 4891965, *2020*.
[30] E.L. Korenromp, J. R. Armstrong-Schellenberg, B. G. Williams, B.L. Nahlen, R. W. Snow, "Impact of malaria control on childhood anaemia in Africa – a quantitative review, ". *J. Trop Med Int Health,* vol. 9, pp. 1050–65, 2004.
[31] A. Neuberger, J. Okebe, D.Yahav, M. Paul, "Oral iron supplements for children in malaria-endemic areas," *J. Cochrane Database Syst Rev*. Vol. 2, no. 2, CD006589, 2016.