International Journal of Statistical Sciences *ISSN 1683-5603*
Vol. 21(2), 2021, pp 225-238

Analysis of Longitudinal Teenage Pregnancy Data of Mpunkunyoni, Kwazulu-Natal using Generalised Linear Mixed Model

Marothi P Letsoalo[1,4*], Yehenew G Kifle[2], Maseka Lesaoana[1] and Christel Faes[3]

[1]University of Limpopo, South Africa
[2]University of Maryland, Baltimore County, USA
[3]Hasselt University, Mathematics and Statistics, Belgium

[*]Correspondence should be addressed to Marothi P Letsoalo
([4]Centre for the AIDS Programme of Research in South Africa, Nelson Mandela School of Medicine, University of KwaZulu-Natal, Durban, South Africa)
(Email: marothi.letsoalo@caprisa.org)

[Received November 2, 2021; Revised November 10, 2021; Accepted November 15, 2021]

Abstract

Teenage pregnancy is an alarming concern in South Africa, as it potentially contributes to the life-threatening disease, HIV, in the young population. Communities in poor villages are hit hard by this problem. Statistics South Africa reported that the total yearly cases of teenage pregnancy has reduced. This work investigated this yearly reduction using 2011-2015 pregnancy census data of female teenagers at Mpunkunyoni, KwaZulu-Natal in South Africa. Apart from the census year effect, this paper accounted for the differences and similarities generated by female teenagers while correcting for the effects of their characteristics. The data were subjected to a generalized linear mixed model, and candidate parsimonious models were attained using the likelihood ratio test and a mixture of chi-squares while the model selection was carried out using the Akaike information criterion. There were no significant differences due to female teenagers while correcting for their age and census year effects. Although there was a tendency for teenage pregnancy risk to decrease over the years, this risk is higher for older teenagers.

Keywords: Cluster-Specific Prediction; Generalized Linear Mixed Model; Generalized Linear Model; Variance Components; Teenage Pregnancy.

AMS Subject Classification: 62H30; 62H15.

0. Tribute to Sinha Brothers

We the authors of this paper feel very much privileged and honored to be able to contribute to this special issue of Statistics and Applications. I, second author Yehenew Kifle, is very much fortunate to have come in contact with the world renowned twin Statisticians, the Sinha brothers, especially Professor Bimal K. Sinha, with whom I share a very close bond. His consistent advice, support and, above all, his unbounded energy have been highly inspirational to me both professionally and in my daily personal life. Thanks to his unreserved support, dedication and encouragement, a number of African Universities have benefited from the great work of Dr. Bimal Sinha via the African International Conferences.

1. Introduction

The choice of teenage pregnancy data is motivated by the fact that teenage pregnancy, defined as pregnancy/birth between ages 13 and 19, inclusive, is an alarming social concern in South Africa (Nguyen et al., 2016). Birdthistle et al. (2019) indicated that HIV among women is increased by young teenagers who contribute 6.5 cases per 100 person-years in 2005. According to Statistics South Africa (StatsSA), previously disadvantaged villages with poor communities are mostly affected (StatsSA, 2014). Between 2018 and 2019, South Africa recorded an increased teenage pregnancy rate in several part of the country. Recent data showed that teenage pregnancy crossed 60% in South Africa; where 17 years and younger contributed 34,587 births of which 688 were girls younger than 10-years (Francke, 2021). Teenage pregnancy is also a burdening issue around the world but mostly in developing countries in Africa (Worldbank, 2019). In 2019, 47 countries have the rate of teenage pregnancy higher than South Africa, with Niger being the highest with 180 per 1,000 women ages 15-19. Other countries above South Africa but outside Africa were, for example, Venezuela, Ecuador, etc.

The problem of teenage pregnancy can be influenced by the families which they live. Therefore, it is of substantial interest to study the impact that families, as well as areas, have on the teenage pregnancy status. Assessing this impact sheds some light on how family practices generate differences in pregnancy statuses between and within families. Moreover, the impact of areas on pregnancy statuses can explain the geographic variation between and within villages. For these

important reasons, teenage pregnancy, like other social and health issues, cannot be addressed by just examining the effects of individual characteristics or by merely averaging within each cluster attribute. In statistical literature, mixed models are recommended to account for the nested structure of the data.

Many data structures in social and health sciences are naturally nested (Zumbo and Chan, 2014). Often they involve longitudinal, nonlinear outcome data or both. Although generalized linear models (GLMs) are known to handle outcome data that follows an exponential-family distribution, these models do not account for correlation among observations data (Molenberghs and Verbeke, 2005). This is because the GLM assumes identically and independently distributed data (Molenberghs and Verbeke, 2005). However, data scientists and researchers commonly neglect clustering during study planning, data collection and analysis (Luke, 2004). This means that cluster-level information is not collected or information is often aggregated at the cluster level for analysis, thereby reducing the information (Luke, 2004). Aggregation of cluster information allows the use of multiple linear regression which assumes that all regression coefficients are equal for all cluster attributes; hence violating the assumption of uncorrelated errors (Hox and Roberts, 46 2011).

Generalised linear mixed models (GLMMs) have presented significant accountability to explore information that comes from populations with nested data structures (Goldstein, 2011). These models have gained popularity since the mid-1980s (Goldstein, 2011) because of their ability to model the effect of individuals and contextual information simultaneously. Although these models were first applied in educational and sociological studies, they can be applied in many study areas (Wang et al., 2011). Other application areas include, but are not confined to, psychology, public health, and economics (Bini et al., 2009). These models are advantageous because they overcome the assumption of independence of observations as well as the correction of overestimation of type-I error (Wang et al., 2011). Statistically, this means that intra-class correlation is non-zero; hence single-level models are inappropriate to analyze nested data (Hox and Roberts, 2011). Moreover, these models can handle unbalanced data.

In the next section, the data used in this study is described and the relationships between teenage pregnancy status and other attributes of the data are explored. After that, we discuss the specification of the model for teenage pregnancy, accounting for nesting of teenagers in families and interpret the results

of model building. Lastly, we present the discussions and conclusions from the analysis.

2. Teenage pregnancy data

In this study, the census data from the health and demographic surveillance system are used. The data are collected and provided by the Africa Health Research Institute (AHRI), KwaZulu-Natal (KZN) in South Africa. The data consist of the population of female teenagers whose pregnancy status was observed during the census years 2011 to 2015 in Mpukunyoni rural area, KZN. These are 11544 females aged 13 to 19, born between 1992 and 2002, inclusive. For each year, it is reducing whether or not the teenager was pregnant. The data are however unbalanced in the sense that we do not have an equal number of observations for all teenagers (Steele, 2008), as e.g. a female who was 19 years old in 2011 would not be included in the remainder of the census years, while a 13-year old female who was observed in 2015 would not have been observed in the previous census years. Furthermore, some female teenagers might not have been observed for some census years because of death, migration, or other reasons. Next, female teenagers (2775) were observed for all the five census years, followed by those observed once (2410), twice (2312), thrice (2075) and four times (1972). In total, there are 35022 measurements of teenage pregnancy for the 11544 female teenagers. Thus, measurements of teenage pregnancy (level 1) are nested within female teenagers at level 2.

These measurements record pregnancy status, denoted by *ps*, which is the response variable of interest. Pregnancy status takes the value 0 (no) or 1 (yes) that respond to whether a female teenager was pregnant. There are four covariates of which one is a measurement of occasion, denoted by *year*, that records values 0, 1, 2, 3 and 4, representing census years 2011, 2012, 2013, 2014 and 2015, respectively. The remaining three covariates are characteristics of female teenagers. The first one is age, denoted by *age* = 0,1,2,3,4,5,6; where 0 is the reference age representing a 13-year old female. The second one is the number of households that a female belongs to, denoted by *hm* = 0,1,2,3; where 0 is the reference category representing one household membership. The third covariate is the number of pregnancies the female had before the time of observation, denoted by *pb* = 0,1,2,3.

The observed proportion of teenage pregnancy is approximately 0.0245 for an average female teenager. For the purpose of visualizing the relationship between *ps* and each covariate, observed proportions were averaged per covariate levels. After that, we computed the logit of the resulting proportions (logit(proportion)), which allows assessing the functional form using scatter plots with a fitted loose curve (Figure 1).

Both *age* in Figure 1B and the number of previous pregnancies *pb* in Figure 1D indicated a positive relationship with the logit of the proportions of *ps*, while *year* in Figure 1A showed a decreasing trend. In addition, Figures A, B, and D show a linear functional form while Figure 1C is non-linear. The relationship between the logit of proportions of *ps* and *hm* seems to be a 2-degree polynomial. For simplicity, the model building will treat *hm* as a binary variable indicating whether a female belongs to more than one household, where 0 indicates *no* and 1 indicates *yes*.

3. Generalized linear mixed models for binary data

3.1. Model specification

Generalized linear mixed models (GLMMs) are extensions of generalized linear models (GLMs) that add a random cluster effect to account for the correlation of the data (Molenberghs and Verbeke, 2005).

In the context of GLMs, the response variables Y_i of measurements from individuals $i(i = 1,2, \dots, N)$ are assumed to be an independent set that is related to a p-dimensional vector of the covariate, x_i. These Y_i's are assumed to have a probability density function (PDF) that belongs to an exponential family, such that $E(y_i|x_i) = \mu = g(x_i'\beta)$, where $E(y_i|x_i)$ is the expected value of y_i given the covariates x_i, β is a p-dimension vector with fixed unknown coefficients and μ is the mean (Molenberghs and Verbeke, 2005). A link function $h(\cdot)$ is chosen such that $h\big(E(y_i|x_i)\big) = g^{-1}\big(E(y_i|x_i)\big) = x_i'\beta$.

In this setting, response variables Y_i's are assumed to be an independent set following a Bernoulli distribution with parameter π_i. Thus, $f(y_i|\pi_i) = \exp\left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i)\right]$ (Czepiel, 2002). The logit link function, $\text{logit}(\pi) = \log\big(\pi/(1-\pi)\big)$ can be utilized for binary outcome data to map the logistic regression model (LRM) to a linear predictor (Rabe-Hesketh and Skrondal, 2008). That is, Model (1a),

$$logit[P(y_i = 1|x_i)] \equiv \log\left[\frac{P(y_i = 1|x_i)}{1 - P(y_i = 1|x_i)}\right] = x_i{}'\beta \tag{1a}$$

where $P(y_i = 1|x_i) = \pi_i$ is the probability of success and $x_{i0} = 1$. The formula for predicting $P(y_i = 1|x_i)$ is then given by

$$P(y_i = 1|\boldsymbol{x_i}\} = \frac{\exp[\boldsymbol{x_i}'\boldsymbol{\beta}]}{1 + \exp[\boldsymbol{x_i}'\boldsymbol{\beta}]}. \tag{1b}$$

Regression coefficients β are estimated using the log likelihood function as

$$l(\boldsymbol{\beta}|y_i, \phi) = \sum_{i=1}^{N} y_i(\boldsymbol{x_i}'\boldsymbol{\beta})\big[log - \big(1 + e^{\boldsymbol{x_i}'\boldsymbol{\beta}}\big)\big]. \tag{1c}$$

In a case of a two-level multilevel data setup where responses vary within a specific cluster, we let $t = 1, \dots, n_i$ represent the level 1 units and $i = 1, \dots, N$ denote the level 2 cluster units.

The GLMM assumes that Y_{ti} are independent, and a cluster-specific regression parameter is assumed to have a PDF, $f\big(y_{ij}\big|\theta_{ti}, \phi\big) = \exp[\phi^{-1}[y_{ti}\theta_{ti} - \psi(\theta_{ti})] + c(y_{ti}, \phi)]$, that belongs to an exponential family. θ_{ti} is a natural parameter that can, through a link function, be represented as a linear predictor η_{ti} while ϕ is a scalar parameter (Czepiel, 2002; Molenberghs and Verbeke, 2005). Functions $\psi(\cdot)$ *and* $c(\cdot,\cdot)$ are all known. Also of interest is the population mean, which is estimated by a linear predictor with both the fixed regression parameters β and cluster-specific random effects υ_i. This equation is written as $h(\mu_{ti}) = x'_{ti\beta} + z'_{ti\upsilon_i} = \eta_{ti}$, where $h(\cdot)$ is some known link function for two vectors x_{ti} and z_{ti} with covariate values. The term υ_i is a vector of random effects that follow a multivariate normal distribution with a vector of zero means and variance-covariance matrix Σ_υ. Likewise, η_{ti} is a linear predictor. The expected value of the response variable given the random effect and the covariates, is $\mu_{ti} = E(Y_{ti}|\upsilon_i, x_{ti})$.

For the teenage pregnancy analysis Y_{ti} represent the pregnancy status of teenager at time *t*. Y_{ti} is assumed to follow a Bernoulli distribution with parameter π_{ti}. A logit link function, $\text{logit}(\pi_{ti}) = \log[\pi_{ti}/(1 - \pi_{ti})]$ isused to map the binary response to a linear predictor function $h(\cdot)$. That is,

$$\text{logit}(\pi_{ti}) \equiv \log[\text{odds}(y_{ti} = 1)] = x'_{ti}\boldsymbol{\beta} + z'_{ti}\boldsymbol{\upsilon_i}, \tag{2a}$$

where π_{ti} in this case, is the probability of teenage pregnancy that are also represented as $E(Y_{ti}|\boldsymbol{\upsilon_i}, \boldsymbol{x_{ti}})$. Model (2a) is a two-level GLMM since the response measurements are clustered within one cluster. The probabilities π_{ti}, given by

$$\pi_{ti} = \frac{\exp\left[x'_{ti\beta} + z'_{ti\upsilon_i}\right]}{1 + \exp\left[x'_{ti\beta} + z'_{ti\upsilon_i}\right]} \tag{2b}$$

are subject-specific estimates of probability of teenage pregnancy. For a random intercepts model, Model (2a) reduces to

$$\text{logit}(\pi_{ti}) = x'_{ti\beta} + \upsilon_i \tag{2c}$$

where $\upsilon_i \sim N\left(0, \sigma^2_{\upsilon(2)}\right)$. The population-averaged probability of teenage pregnancy is then given by

$$\pi_t = \int_{-\infty}^{\infty} \frac{exp\left[\boldsymbol{x'_{ti\beta}} + \boldsymbol{\upsilon_i}\right]}{1 + exp\left[\boldsymbol{x'_{ti\beta}} + \boldsymbol{\upsilon_i}\right]} \phi(\upsilon_i) d\upsilon_i \tag{2d}$$

$$L\left(\beta, \sigma^2_{\upsilon(2)} \middle| Y_{tj}\right) = \prod_{i=1}^{N} \int_{-\infty}^{\infty} \left[\left\{ \prod_{t=1}^{n_i} \frac{exp\left[y_{ti}\left(\boldsymbol{x'_{ti\beta}} + \boldsymbol{\upsilon_i}\right)\right]}{1 + exp\left[y_{ti}\left(\boldsymbol{x'_{ti\beta}} + \boldsymbol{\upsilon_i}\right)\right]} \right\} \frac{exp\left[-\upsilon_i / 2\sigma^2_{\upsilon(2)}\right]}{\left(2\sigma^2_{\upsilon(2)}\right)^{1/2}} \right\} dvi \tag{2e}$$

and the likelihood function used to estimate β and $\sigma^2_{\upsilon(2)}$of the random effect υ_i is given by Equation (2e). This function $L\left(\beta, \sigma^2_{\upsilon(2)} \middle| Y_{tj}\right)$ is evaluated using the adaptive quadrature or Laplace approximation approaches. However, in this paper, we use adaptive quadrature approximation.

3.2. Variance components

Although there are several ways to examine the variance components of mixed models, our study focused on the variance partition coefficient (VPC) and intra-class correlation coefficient (ICC). The VPC reports the proportion of the response variance that lies at each level of the model hierarchy, while the ICC reports the expected degree of similarity between responses within a given cluster. Consider Model (2c) with the fixed intercept (β_0) and m random intercepts $\upsilon_{0i}^{(l)}$. Thus,

$$\eta_{ti} = \beta_0 + \sum_{l=2}^{m} \upsilon_{0i}^{(l)} \tag{3a}$$

where l is the cluster index, for example, $l = 2$ corresponds to teenagers, $l = 3$ to families. The random intercepts $v_{0i}^{(l)}$ are assumed to be normally distributed with mean zero and variance $\sigma_{v0(l)}^2$. The level 1 errors for a logistic regression model are assumed to have variance equal to $\pi^2/3$; hence, the total variance is calculated as $\pi^2/3 + \sum_{l=2}^{m} \sigma_{v0(l)}^2$. The VPCs and ICCs for Model (3a) are respectively calculated as

$$VPC_{\upsilon}^{(L)} = \frac{\sigma_{\upsilon 0(L)}^2}{\sigma_{\upsilon 0(1)}^2 + \sum_{l=2}^{m} \sigma_{\upsilon 0(l)}^2} \tag{3b}$$

$$ICC_{\upsilon}^{(L)} = \frac{\sum_{l=1}^{L} \sigma_{\upsilon 0(l)}^2}{\sigma_{\upsilon 0(1)}^2 + \sum_{l=2}^{m} \sigma_{\upsilon 0(l)}^2} \tag{3c}$$

where $L = 1,2,\ldots,m$ is the hierarchy level and $\sigma_{v0(1)}^2 = \pi^2/3$. In order to evaluate the importance of the female cluster, we used a mixture chi-square (MI-CHI) likelihood ratio test (Verbeke and Molenberghs, 2009, Chapter 6). Model selection for both the GLM and GLMM was made using both the AIC value and the likelihood ratio test (LRT) for nested models. The AIC value is calculated using the formula $AIC = -2ln[l(\beta)] + 2k$, where k is the number of estimated coefficients (Akaike, 2011). A lower AIC value indicates a better fit.

4. Analysis and Interpretation of Results

4.1. Model building and selection

This section presents the model building using both the GLM and GLMM to describe the probability of teenage pregnancy. GLM intends to estimate the effect of covariates on the log-odds of teenage pregnancy, whereas GLMM estimates the effect of both the covariates and account for within female cluster correlation. In both models, we use the notations β_{year}, β_{age}, β_{hm} and β_{pb} to denote the fixed effects of *year*, *age*, *hm* and *pb*. Furthermore, the effects of the interaction of *year* and other covariates (*year* × *hm* and *year* × *pb*) are denoted by β_{yhm} and β_{ypb}, respectively.

To evaluate the importance of the effect of covariates mentioned above, we use LRT on nested models. In the GLMM, we also check whether the effect of the female cluster is significant, using MI-CHI. We then consider the AIC value to decide on a better fit within the fitted model families (GLM or GLMM). A summary of the model building exercises for the GLM and GLMM is given in Table 1, which shows the results of the competing models and the estimates. With

a purpose to achieve a more parsimonious mean structure, the covariates are excluded one by one.

Model (5) is the best model according to the LRT of fixed effects and the AIC value. The inclusion of covariates *hm* or *pb* does not improve the model. Both *hm* and *pb*, and any of the presumed interactions do not significantly affect teenage pregnancy status. For prediction purposes, Model (5) can be written as Equation (4a). Equation (4a) suggests that the probability of teenage pregnancy for a 13-year-old female teenager in 2011 and had no previous pregnancy is $\exp(-5.542)1+\exp(-5.542)=0.0039$, with 95% confidence interval (0.00307,0.00498). The odds of teenage pregnancy for a 13-year-old female teenager in 2011 will increase by $\exp(0.557)$ for each additional year to the female teenager's age. On the other hand, the odds are expected to reduce by exp(0.21) for each additional census year; hence the corresponding probabilities in 2012, 2013, 2014, 2015 are 0.0032, 0.0026, 0.0021 and 0.0017, respectively.

Table 1: Maximum likelihood estimates of GLMs and GLMMs that predict log odds of teenage pregnancy.

		GLM					GLMM				
Covariate	**Parameter**	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Fixed effects											
Intercept	β_0	***-5.521	***-5.523	***-5.525	***-5.528	***-5.542	***-5.521	-5.523	***-5.525	***-5.528	***-5.599
		[0.126]	[0.126]	[0.125]	[0.124]	[0.124]	[0.126]	[0.126]	[0.125]	[0.124]	[0.156]
		(-43.790)	(-43.996)	(-44.197)	(-44.469)	(-44.608)	(-43.790)	(-43.997)	(-44.197)	(-44.469)	(-35.882)
year	β_{year}	***-0.211	***-0.209	***-0.208	***-0.208	***-0.210	***-0.211	***-0.209	***-0.208	***-0.208	***-0.211
		[0.028]	[0.027]	[0.026]	[0.026]	[0.026]	[0.028]	[0.027]	[0.026]	[0.026]	[0.026]
		(-7.521)	(-7.800)	(-8.066)	(-8.072)	(-8.169)	(-7.521)	(-7.801)	(-8.066)	(-8.072)	(-8.146)
age	β_{age}	***0.549	***0.549	***0.549	***0.549	***0.557	***0.549	***0.549	***0.549	***0.549	***0.559
		[0.024]	[0.024]	[0.024]	[0.024]	[0.024]	[0.024]	[0.024]	[0.024]	[0.024]	[0.024]
		(22.730)	(22.730)	(22.731)	(22.731)	(23.419)	(22.730)	(22.731)	(22.731)	(22.731)	(23.161)
pb	β_{pb}	0.175	0.191	0.191	0.191		0.175	0.191	0.191	0.191	
		[0.153]	[0.108]	[0.108]	[0.108]		[0.153]	[0.108]	[0.108]	[0.108]	
		(1.141)	(1.774)	(1.775)	(1.777)		(1.141)	(1.774)	(1.775)	(1.777)	
hm	β_{hm}	-0.041	-0.041	-0.019			-0.041	-0.041	-0.019		
		[0.147]	[0.147]	[0.107]			[0.147]	[0.147]	[0.107]		
		(-0.280)	(-0.279)	(-0.177)			(-0.280)	(-0.279)	(-0.177)		
ypb	β_{ypb}	0.012					0.012				
		[0.082]					[0.082]				
		(0.145)					(0.145)				
yhm	β_{yhm}	0.018	0.018				0.018	0.018			
		[0.082]	[0.082]				[0.082]	[0.082]			
		(0.222)	(0.221)				(0.222)	(0.221)			
Random effects											
		–	–	–	–	–	0.000	0.000	0.000	0.000	0.110
Fit statistics											
	-2ll	7239.0	7239.0	7239.0	7239.1	7242.1	7239.0	7239.0	7239.0	7239.1	7241.7
	AIC	7253.0	7251.0	7249.0	7247.1	7248.1	7255.0	7253.0	7251.0	7249.1	7249.7

*p < 0.05, **p < 0.01, ***p < 0.001; Standard errors in Square brackets; z-value in parentheses.

Table 1 also presents the results from fitting a GLMM to the data, thereby accounting for the longitudinal nature of the measurements. The results show that even in the case of GLMM, at 5% level of significance, the effects of *hm* and *pb* are not significant. We also observe that the estimated standard errors of the fixed effects in GLMs are equal to those of the GLMMs. Comparing with Model (3) with Model (8), it can be observed that the probability of teenage pregnancy does not depend on the female cluster when we correct for all covariates. However, exclusion of hm and both hm and pb in Models (9 and 10), respectively show some variation in the log odds of teenage pregnancy status. In the case of Model (9), the estimates are still similar to the Model (4), also suggesting no variation due to the female cluster. On the contrary, Model (10) shows slight differences in the estimates due to the inclusion of the effect of the female cluster. Using the MI-CHI test, the female cluster random intercepts $\sigma^2_{v0(2)}$ in Model (10) is not significant at 5% level of significance with *p*-value equal to 0.2707, implying that for our data, it is not important to account for the female clustering. Also, the random intercept has not improved the model fit of Model (5), but rather worsened it by a small margin of about 0.6 AIC value.

Given the resulting insignificance of the female cluster, the predictions from Model (10) are equivalent to those of Model (5). Not neglecting the findings on the teenage female cluster, our study further used Model (10) estimate, written as Equation (4b) to make female-specific prediction when both census year and age are corrected for. In Equation (4b), $v_{0i}^{(2)}$ is normally distributed with mean = 0 and variance = 0.11. The estimated fixed intercept of -5.599, deduce that the probability for teenage pregnancy for a 13-year old female in 2012 is $\exp(-5.599)/\big(1+\exp(-5.599)\big) = 0.0037$. The magnitude of the effect of census year is equal to -0.211, which means that for each one year increase to a census year, the odds of teenage pregnancy is expected to decrease by $\exp(-0.211) = 0.8098$ when controlling for the female difference.

We also computed the VPC and ICC. The magnitude of both VPC and ICC suggests that 3.25% of the variation in teenage pregnancy is due to the female cluster. Thus, the correlation of pregnancies within females after correcting for their ages and census year equals to 0.0325, which is very limited.

In addition to the prediction, we further removed age effect from Model (10) in order to illustrate the importance of including the cluster effect when available

covariates do not explain the differences in individuals' responses. The estimates of this model were not presented in this paper; however, its equation together with its counterpart GLM without the random effect were $\hat{\eta}_{ti} = -3.5013 - 0.2027year_{ti} + \upsilon_{0i}^{(2)}$ and $\hat{\eta}_i = -3.3318 - 0.2069year_i$. The variance, $\sigma^2_{\upsilon 0(2)}$, of the random intercept, $\upsilon_{0i}^{(2)}$, was equal to 0.3505 and significant with a *p*-value of 0.033, suggesting that the female cluster is important. That meant 9.63% of the variation in teenage pregnancy is accounted for by the female cluster. While the effect of census year is similar for both the GLM and GLMM, the fixed intercepts were different compared to the models that account for the age. This difference supports the significance of the effect of the female cluster.

Probability predictions

To predict and compare probabilities of the models, we used the two Equations

$$\hat{\eta}_i = -5.542 - 0.21year_i + 0.557age_i \quad (3b)$$

$$\hat{\eta}_{ti} = -5.599 - 0.211year_{ti} + 0.559age_{ti} + \upsilon_{0i}^{(2)} \quad (3c)$$

to produce the plots in Figure 2. For the top two figures, we predicted the probabilities for the whole teenage population and averaged these within census year in Figure 2A and age in Figure 2B. The population averaged lines representing GLM and GLMM probability predictions were similar for both census year and age, justifying that the GLM was sufficient to model teenage pregnancy data for this study. These lines tie with the observed proportions of teenage pregnancy.

5. Discussion and Conclusion

In this work, we observed a hierarchical data structure of teenage pregnancy which was fitted using both the GLM and GLMM with a logit link. This implied that the fitted logit GLM assumed that measurements of teenage pregnancy are independent. The variance component estimates indicated that the female cluster has no effect on teenage pregnancy status after correcting for the effect of census year and age of the female. This means that the female cluster did not generate differences in teenage pregnancy status; hence, it would be sufficient to assume independence of measurements for the data used in our study. This is also observed for the estimated fixed effects which were similar for both the best fit

GLM and GLMM. In addition, this work has shown that the risk of teenage pregnancy reduces by census year.

For demonstration purposes, this work further illustrated the importance of assessing the female cluster by removing the effect of age from the final model. The resulting model indicated that there is a need to consider the effect of the female cluster, thereby suggesting some differences and similarities of the risk of teenage pregnancy due to female cluster. Of course, this is expected since age explains most of that variation, and it is the female demographic characteristic.

Given that this work has not taken into account the missingness of teenage pregnancy status that exists in our data, further studies could be conducted that incorporate reasons for missingness that could be provided by Africa Health Research Institute. Furthermore, missingness reasons are sometimes unknown; hence further studies could also investigate missingness mechanism through sensitivity analysis techniques that are discussed in Molenberghs and Verbeke (2005, chapters 26-32). Figures

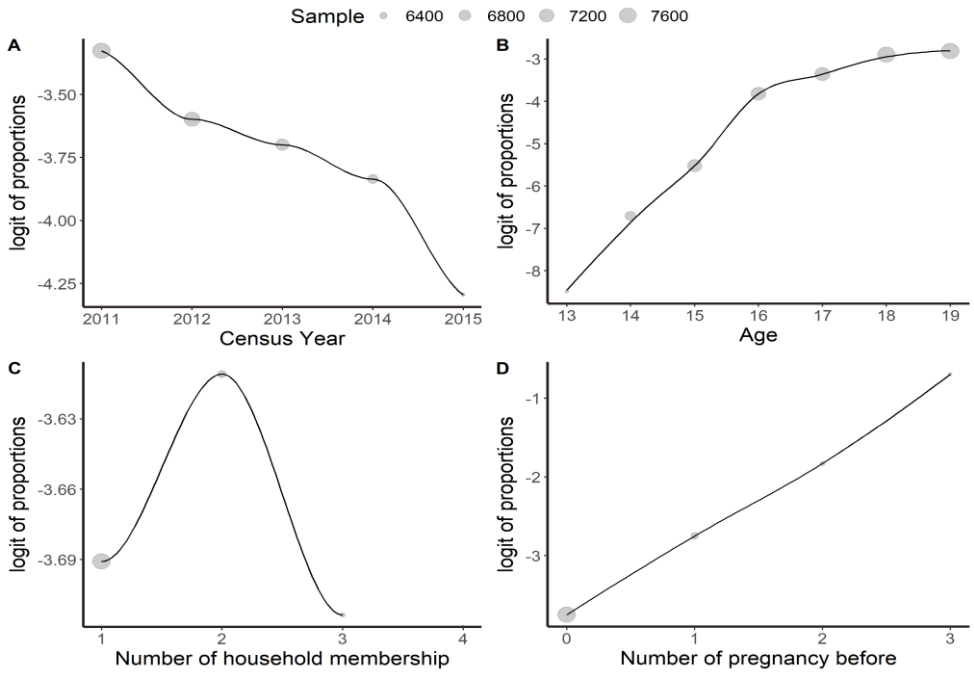

Figure 1: The relationship between the logit of observed teenage pregnancy proportions with covariates

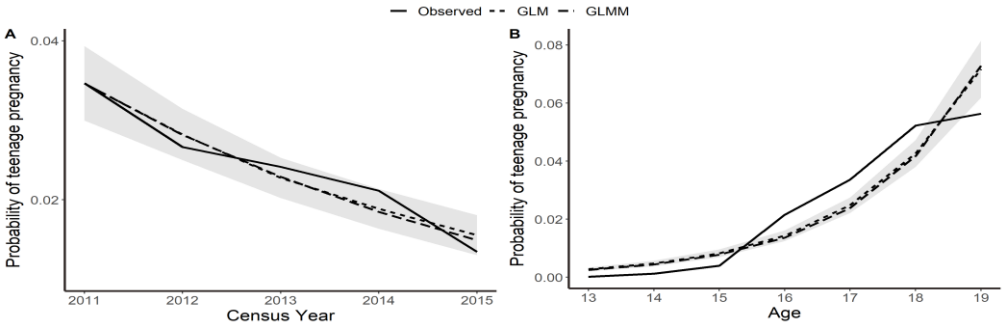

Figure 2: Teenage pregnancy probabilities predictions (Marginal).

6. Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

YGK, ML and CF contributed as promoters of the study. CF supported with the statistical methodology. MPL performed the statistical analysis and wrote the main manuscripts. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by a grant from the Institutional University Cooperation IUC-UL project under the umbrella of the Belgian Flemish Interuniversity Council (VLIR-UOS).

Acknowledgments

The authors gratefully acknowledge Africa Health Research Institute in KwaZulu-Natal for providing teenage pregnancy data free of charge.

References

[1] Akaike, H. (2011). Akaike's Information Criterion. In : International Encyclopedia of Statistical Science. Springer 25–25.

[2] Bini, M., Piccolo, D., Monari, P., Salmaso, L. (2009). Statistical methods for the evaluation of educational services and quality of products. Physica-Verlag.

[3] Czepiel, S. A. (2002). Maximum likelihood estimation of logistic regression models: theory and implementation. Available at czep. net/stat/mlelr. pdf.

[4] Goldstein, H. (2011). Multilevel statistical models 922. John Wiley & Sons.

[5] Hox, J., Roberts, J. K. (2011). Handbook of advanced multilevel analysis. Psychology Press.

[6] Luke, D. A. (2004). Multilevel modeling 143. Sage.

[7] Molenberghs, G., Verbeke, G. (2005). Models for discrete longitudinal data. Springer.

[8] Nguyen, H., Shiu, C., Farber, N. (2016). Prevalence and Factors Associated with Teen Pregnancy in Vietnam: Results from Two National Surveys. Societies 6, 17.

[9] Rabe-Hesketh, S., Skrondal, A. (2008). Multilevel and longitudinal modeling using Stata. STATA press.

[10] Stats S. A. (2014). Recorded Live Births. Tech. rep., Statistics South Africa.

[11] Steele, F. (2008). Multilevel models for longitudinal data. Journal of the Royal Statistical Society: series A (statistics in society) 171, 5–19.

[12] Verbeke, G., Molenberghs, G. (2009). Linear mixed models for longitudinal data. Springer Science & Business Media.

[13] Wang, J., Xie, H., Fisher, J. F. (2011). Multilevel models: applications using SAS®. Walter de Gruyter.

[14] Birdthistle, I., Tanton, C., Tomita, A., de Graaf, K., Schaffnit, S. B., Tanser, F., Slaymaker, E. (2019). Recent levels and trends in HIV incidence rates among adolescent girls and young women in ten high-prevalence African countries: a systematic review and meta-analysis. The Lancet Global Health 7, e1521–e1540.

[15] Worldbank (Accessed December 2019). Adolescent fertility rate (births per 1,000 women ages 15-19) - South Africa. Available at: https://data.worldbank.org/indicator/SP.ADO.TFRT?locations=ZA

[16] Francke, R. L. Shocking Stats SA report shows 33 000 teen mothers in 2020, 660 of them younger than 10 years old.