A Comparison of Three Models in Multivariate Binary Longitudinal Data Analysis: Application to FDCS Study

Hissah Alzahrani

Mathematical Science Department, College of applied sciences, Umm Al-Qura University, Mekkah, Saudi Arabia
Email: hahzahrani@uqu.edu.com.sa

Abstract

Multivariate longitudinal data analysis plays an important role in many biomedical and social problems. In this article, we present three methods for analyzing multiple and correlated binary outcomes; each one can be beneficial for determined aims. We review methods one and two, and we proposed method three. The three methods estimate the marginal means using the GEE approach for multivariate binary longitudinal data. The first method addresses the question of estimating one group of covariate parameters for many binary outcomes while accounting for their multivariate structure. The second method addresses the question of estimating the covariate parameters for each binary outcome separately. The third method is an estimation of the covariate parameters for each combination of outcomes. Our goal is to investigate the differences among the parameter estimations of the three methods. In the application element, we present a follow-up study (Florida Dental Care Study) that measured three binary outcomes and five covariates at four intervals. The FDCS study is useful explanation of the variation between outcomes since the outcomes were highly correlated.

Subject Areas

Statistical Analysis

Keywords

Binary, GEE, Longitudinal, Many Outcomes, Correlation

1. Introduction

Many of the recent medical experiments and social research are characterized by

multiple outcomes. In this research, we focus on the multivariate binary outcomes in longitudinal data. It is a general approach of the univariate longitudinal data when we have more than one outcome. Each individual i has a vector of responses for different outcomes, $k = 1, 2, \cdots, K$, which are measured at different times or occasions, $t = 1, 2, \cdots, T_i$, and has cluster size $T_i = T_i K$. To simplify the notations, we will refer to T_i as T which is the number of occasions or visit numbers over all observations as following:

$$Y_i = \left[Y_{i11}, Y_{i21}, \cdots, Y_{iT1}, Y_{i12}, Y_{i22}, Y_{i32}, \cdots, Y_{iT2}, \cdots, Y_{i1K}, Y_{i2K}, \cdots, Y_{iTK} \right]^{\mathrm{T}}$$

There are multiple methods for analyzing multivariate binary longitudinal data, which are different depending on the research aims. Let's start by referring "outcome" to the longitudinal readings within the same dependent variable and "response" for any longitudinal reading of any dependent variables without specification. If the outcomes are uncorrelated, then we can analyze them as univariate longitudinal data separately. This independence assumption rarely happens in real applications because the outcomes, which are taken from the same observation, are highly probably to be correlated. The best statistical analysis of multiple outcomes should account for the correlation among the outcomes and the occasions since it is longitudinal data [1] [2] [3]. The three methods of analyzing longitudinal data are:

1) Reducing many outcomes into one summary outcome which results in estimating a unique set of regression coefficients [3].

2) Analyzing each outcome separately with account for the correlations, which results in estimating a set of regression coefficients for each outcome [4].

3) Analyzing the outcomes jointly and accounting for the correlations, which results in estimating a set of regression coefficients for each combination case of them (the proposed method).

Our investigation method of the longitudinal data via the marginal models is using GEE approach. The GEE approach of longitudinal data analysis is constructed for discrete and continuous outcomes[1]. Then, applied the alternating logistic regression in multivariate binary model, which is dependent on modeling the association between responses by the odds ratio. The specification of the odds ratio matrix R, is the most important point on the GEE analysis of the marginal models. We need to estimate $\dfrac{(KT)(KT-1)}{2} = \dbinom{KT}{2}$ parameters. It is a large correlation matrix and it will rapidly increase when we have more parameters in R (correlation matrix). Since our interests are in binary multivariate longitudinal outcomes, then it is more natural to use the odds ratio to describe the association between the binary responses [3]. There are two applied methods to reduce the parameter estimations of the within subject association matrix R. The two methods are Kronecker product and regression model for pairwise odds ratio. The two methods are equivalent to each other and have approximately the same results. We will focus on the regression model for pairwise odds ratio to be method 1 in the comparison part.

Also, a SAS macro is built to analyze multivariate binary longitudinal outcomes using Kronecker product [4]. In this paper, we will consider the regression model for pairwise odds ratio to be the statistical tool of the first aim (method 1) [3], and Kronecker product for the second aim (method 2) [4]. In the next section is an overview of the three methods. The application is in Section 3. The last section is about the discussion.

2. Methods

2.1. Method 1: Regression Modeling for Odds Ratio

The GEE model is $g(\mu_i) = X_i\beta$ where $E(Y_i) = \mu_i$, and X_i is the covariate matrix, assuming the vector of responses for subject i is related to the covariates via logit link function. The primary advantage of using GEE in multivariate binary data is working in two-way marginal relationships between the pairs of the responses for any dimension of outcomes and occasions without high order associations. The penalty of this advantage is the cluster size which is $T \times K$. It will cause a larger correlation matrix R since the dimension of R matrix is $TK \times TK$. R is a diagonal and symmetric matrix, thus we need to estimate $\binom{TK}{2}$. For example, if we have 5 outcomes obtained at 5 occasions, then we need to estimate 300 parameters, because that we need to obtain a parsimonious model of working correlation R to be less than $\binom{TK}{2}$.

As we mentioned previously, we can use the odds ratio instead of the correlation since the responses are binary. Then, the matrix R will consist of pairwise log odds ratios [3]. It is based on fitting a regression model for marginal pairwise odds ratio to estimate less parameters than $\binom{TK}{2}$ [3]. Let γ be a vector of size $\binom{TK}{2}$. of all non-redundant pairwise odds ratio in R. Here the goal is reducing the length of the vector γ. Consider the matrix of the odds ratio R consist of three odds ratio types:

1) Let $v_{tk.tk'}$ be the inter-outcome odds ratio which compares the outcome k with the outcome k' at time t:

$$v_{tk.tk'} = \frac{P(Y_{tk}=1, Y_{tk'}=1)P(Y_{tk}=0, Y_{tk'}=0)}{P(Y_{tk}=1, Y_{tk'}=0)P(Y_{tk}=0, Y_{tk'}=1)} \tag{1}$$

2) Let a $\alpha_{tk,t'k}$ be the intra-outcome odds ratio which compares outcome k at time t with the same outcome at time t':

$$\alpha_{tk,t'k} = \frac{Pr(Y_{tk}=1, Y_{t'k}=1)Pr(Y_{tk}=0, Y_{t'k}=0)}{Pr(Y_{tk}=1, Y_{t'k}=0)Pr(Y_{tk}=0, Y_{t'k}=1)} \tag{2}$$

3) Let $T_{tk,t'k}$ be the cross odds ratio which compares the outcome k at time t with outcome k at time t':

$$T_{tk,t'k'} = \frac{Pr(Y_{tk}=1,Y_{t'k'}=1)Pr(Y_{tk}=0,Y_{t'k'}=0)}{Pr(Y_{tk}=1,Y_{t'k'}=0)Pr(Y_{tk}=0,Y_{t'k'}=1)} \tag{3}$$

The general regression model that relates the inter-outcome (ν), the intra-outcome (α) and the cross odds ratio (τ) to matrix z via log link function [2], is:

$$\log(OR) = Z\gamma \tag{4}$$

where is γ a parsimonious vector of the odds ratio parameters that relates the $\binom{KT}{2}$ pairwise log odds ratio to Z. Here Z is a fixed indicator matrix that specifies the outcomes, times and interactions between them which are under consideration. Using the regression model of (4) helps us to model the three types of association in matrix Z.

2.2. Method 2: Kronecker Product

To apply method 2 we used Kronecker Product approach [4], they built a SAS macro to estimate separate sets of regression coefficients for each binary outcome. This method depends on the GEE estimation of the marginal models. The strength of this macro is creating a covariate design to allow for separate regression coefficients for each outcome. The GEE model is $\mathrm{logit}(\mu_i) = X_i\beta$ where $E(Y_i) = \mu_i$ and X_i is the covariate matrix and assuming the vector of responses for subject i is related to the covariate via logit link function.

Using Kronecker product of X matrix and K dimensional identity matrix be used to generate regression coefficients for each outcome. This method is beneficial to fit more than one longitudinal outcome using GEE approach, and accounts for the correlation among the outcomes. The results are separated estimation coefficients for each outcome, which is helpful for the second aim.

2.3. Method 3: The Proposed Approach

Let Y_{itk} denotes the value of kth binary outcome measured at t occasion or visit number for individual i, where $i = 1, 2, \cdots, N$, $k = 1, 2, \cdots, K$, and $t = 1, 2, \cdots, T$. Let X_{itc} denotes the value of the covariate C that is measured at occasion t for individual i where $c = 1, 2, \cdots, C$ (C is the number of covariate variables).

Let $Y_i \sim \mathrm{Bernolli}(\pi_1), Y_2 \sim \mathrm{Bernolli}(\pi_2), \cdots, Y_k \sim \mathrm{Bernolli}(\pi_K)$. Assuming the outcomes are obtained in the same time for same C covariates. We will collapse K outcomes into one outcome Z. Since Multinomial trial process is a simple generation of Bernoulli trial processes, then the new variable Z has multinomial distribution with $J = 2K$ possible outcomes, $Z_1, Z_2, \cdots, Z_J$ Suppose each possible outcome can occur with probability $p_1, p_2, \cdots, p_J$, then the probability of Z_1 occurs m_1 times, Z_2 occurs m_2 times, $\cdots$, Z_J occurs m_J times is following the probability mass function:

$$f(m_1, m_2, \cdots, m_J) = \binom{N}{m_1, m_2, \cdots, m_J} p_1^{m_1} p_2^{m_2} \cdots p_J^{m_J} \tag{5}$$

where $\sum_{j=1}^{J} p_j = 1$, $\sum_{j=1}^{J} m_j = N$. Then, $Z = (Z_1, Z_2, \cdots, Z_J)^{\mathrm{T}}$ has a multinomial distribution with $p = (p_1, p_2, \cdots, p_J)$ parameters. To collapse many binary vectors into one multinomial vector, we suggest to do transformation from a binary coding system to a decimal coding system. Then we analyze the new outcome Z instead of K binary vectors. Since the data is longitudinal, we consider a correlation among the new variable Z_{it}. Then we use an appropriate exciting method for analyzing correlated nominal multinomial responses.

For correlated multinomial responses, there are many methods using random effects model and marginal models. For random effect models, Hedeker adopts a method in Bock's model [5]. For marginal models, the GEE approach avoids specification of the distribution of the multinomial outcome by adopting a "working" correlation matrix. However, a GEE approach using a local odds ratio parameterization is one appropriate method to analyze the correlated data [6]. In the multinomial responses, we set up one of the possible outcomes to be a reference level and estimate the parameter in term of $J - 1$ remaining levels compared to the reference level.

After estimating the parameters of nominal multinomial responses Z_{it}, we must move to the "decoding" step. The decoding step is to transform each category in Z_{itj} response to its original definition. For example, the parameter estimation $\hat{\beta}_1$ for category $z = 3$ is for the original code "011". Then, expresses the estimated parameter of the logit probabilities of accruing outcome 1 and outcome 2 simultaneously. In conclusion, using the GEE approach for nominal multinomial correlated responses could be helpful to analyze the joint distribution of many binary longitudinal outcomes.

3. Application

The Florida Dental Care Study (FDCS) is a longitudinal study of oral health and dental service utilization. It is conducted in four centers in North Florida in the United States [7]. The sample size is 873 subjects who had baseline interview exams and four clinical examination interviews at 6, 12, 18 and 24 months past baseline visit. By the end of 24 months, 87.5% of 873 remained in the study. 3.3% refused to participate, 1.1% unable to participate due to medical issues, 3.3% were deceased and 4% were dropout. The issue of bias was discussed by comparing the characteristics of the patients who remained in the study at 24 months, with those who did not for any reason [8]. Also, there are three binary outcomes that were measured over four time intervals (0 - 6, 6 - 12, 12 - 18, 18 - 24 months). In this application, a subset of five covariates were chosen in Table 1. The three binary outcomes measured in this study are "problem oriented visit", "dental cleaning" and "dental check up". At the end of each interval, each subject was asked whether he had visited a dentist within the past 6 months "problem oriented visit" and whether this dental visit was for a "dental cleaning" or "check up". The three binary outcomes are coded as (0 = "No", 1 = "Yes"). First, we consider the following three models.

Model 1:

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 X_{ira} + \beta_2 X_{gender} + \beta_3 X_{cavit} + \beta_4 X_{loose} + \beta_5 X_{able}$$

Model 2:

$$\text{logit}(\mu_i) = \sum_{k=1}^{3} \left(\beta_{k0} + \beta_{k1} X_{ira} + \beta_{k2} X_{gender} + \beta_{k3} X_{cavit} + \beta_{k4} X_{loose} + \beta_{k5} X_{able} \right)$$

Model 3:

$$\text{logit}(\mu_i) = \sum_{j=1}^{7} \left(\beta_{j0} + \beta_{j1} X_{ira} + \beta_{j2} X_{gender} + \beta_{j3} X_{cavit} + \beta_{j4} X_{loose} + \beta_{j5} X_{able} \right)$$

$$j = \begin{cases} 1 & \text{when} \quad (Y_{i1} = 0) \,\&\, (Y_{i2} = 0) \,\&\, (Y_{i3} = 1) \\ 2 & \text{when} \quad (Y_{i1} = 0) \,\&\, (Y_{i2} = 1) \,\&\, (Y_{i3} = 0) \\ 3 & \text{when} \quad (Y_{i1} = 0) \,\&\, (Y_{i2} = 1) \,\&\, (Y_{i3} = 1) \\ 4 & \text{when} \quad (Y_{i1} = 1) \,\&\, (Y_{i2} = 0) \,\&\, (Y_{i3} = 0) \\ 5 & \text{when} \quad (Y_{i1} = 1) \,\&\, (Y_{i2} = 0) \,\&\, (Y_{i3} = 1) \\ 6 & \text{when} \quad (Y_{i1} = 1) \,\&\, (Y_{i2} = 1) \,\&\, (Y_{i3} = 0) \\ 7 & \text{when} \quad (Y_{i1} = 1) \,\&\, (Y_{i2} = 1) \,\&\, (Y_{i3} = 1) \end{cases}$$

The reference category is 0 when $(Y_{i1} = 0)$ & $(Y_{i2} = 0)$ & $(Y_{i3} = 0)$. Before running the three models, we measured the association between the binary responses using the odds ratio instead of the correlation. The odds ratio and the 95% confidence intervals are in Table 2. The association parameters are larger than one and significant for all the odds ratios because the 95% confidence intervals exclude the value one. Then, we expect a large differences between the three models. The strongest association is between "cleaning" and "check up" outcomes.

We ran the three models with no constraints in the log odds ratio, then we got 66 unique estimated odds ratio ($\#\alpha = 12$, $\#\upsilon = 18$, $\#\tau = 36$) in models 1 and 2. The results of the three models are in Tables 3-5. The covariates Cavit and Able are significant in model 1 while Gender and IRA are close and Loose is not. In model 2, the covariates are estimated for each outcome separately. The Gender covariate still not significant for all. However, The Loose parameter estimation is changed completely from model 1 to model 2 to be significant for all. This change

Table 1. Covariates variables.

Covariate	Definition
IRA	(1) The subject goes to a dentist regularly or occasionally whether or not has a problem, (0) The subject did go to a dental check-up once a year or more often in the previous 5 years.
Gender	(1) Female, (0) Male.
Cavit	(1) The subject reported having cavities (tooth decay) in the previous 6 months, (0) if not.
Loose	(l) The subject had a loose tooth, (0) had not
Able	(1) The subject able to pay unexpected US$ 500 dental bill, but with difficulty, (0) not able to pay

Table 2. Estimated odds ratio.

Odds Ratio	Estimate	95% CI
OR (Y_1, Y_2)	1.977	(0.512, 0.851)
OR (Y_1, Y_3)	2.081	(0.551, 0.915)
OR (Y_2, Y_3)	42.147	(3.497, 3.985)

Table 3. Model 1 results.

Parameter	Estimate	Stderr	z	P-value
Intercept	−1.1302	0.0651	−17.36	**<0.0001**
IRA	0.2379	0.1273	1.87	0.0616
Gender	−0.1649	0.0901	−1.83	0.0673
Cavit	0.6235	0.0795	7.84	**<0.0001**
Loose	0.0818	0.081	1.01	0.3126
Able	−0.2715	0.086	−3.16	**0.0016**

Table 4. Model 2 Results.

Parameter	Estimate	Stderr	z	P-value
Intercept_1	−1.306	0.0827	−15.79	**<0.0001**
IRA_y_1	0.1447	0.1768	0.82	0.4129
Gender_y_1	−0.1904	0.116	−1.64	0.1006
Cavit_y_1	1.0137	0.1059	9.58	**<0.0001**
Loose_y_1	0.5367	0.1191	4.51	**<0.0001**
Able_y_1	−0.1008	0.1077	−0.94	0.3491
Intercept_2	−1.1788	0.0979	−12.04	**<0.0001**
IRA_y_2	0.4656	0.1836	2.54	**0.0112**
Gender_y_2	−0.1652	0.1323	−1.25	0.2118
Cavit_y_2	0.3919	0.1113	3.52	**0.0004**
Loose_y_2	−0.4316	0.1245	−3.47	**0.0005**
Able_y_2	−0.4329	0.1257	−3.44	**0.0006**
Intercept_3	−0.8009	O. 0945	− 8.48	**<0.0001**
IRA_y_3	0.328	0.1898	1.73	0.0839
Gender_y_3	−0.1609	0.1277	−1.26	**0.2075**
Cavit_y_3	0.2894	0.1065	2.72	**0.0066**
Loose_y_3	−0.4101	0.117	−3.51	**0.0005**
Able_y_3	−0.5959	0.1245	−4.79	**<0.0001**

Table 5. Model 3 results.

Parameter	Estimate	Stderr	z	P-value
$Intercept_1$	2.31644	0.15412	15.0297	**$<2e^{-16}$**
IRA_y_1	−0.44468	0.28851	−1.5413	0.12325
$Gender\_y_1$	0.25399	0.20448	I. 2421	0.2142
$Cavit\_y_1$	−1.59776	0.16413	−9.7347	**$<2e^{-16}$**
$Loose\_y_1$	0.30088	0.22217	1.3543	0.17565
$Able\_y_1$	0.56503	0.19391	2.9139	**0.00375**
$Intercept_2$	0.2618	0.20957	1.2493	0.21157
IRA_y_2	−0.31291	0.41567	−0.7528	0.45157
$Gender\_y_2$	0. 32744	0.26896	1.2174	0.22343
$Cavit\_y_2$	−2.22835	0.32981	−6.7565	**$<2e^{-16}$**
$Loose\_y_2$	0.40086	0.27408	1.4626	0.14359
$Able\_y_2$	0.08664	0.2673	0. 3241	0.74585
$Intercept_3$	−1.64295	0.3346	−4.9101	**$<2e^{-16}$**
IRA_y_3	0. 34044	0.53157	0.6404	0.52188
$Gender\_y_3$	0.09152	0.38695	0.2365	0.81304
$Cavit\_y_3$	−1.67004	0. 46435	−3.5965	**0.00032**
$Loose\_y_3$	0.4977	0.41352	1.2036	0.22876
$Able\_y_3$	0.89498	0.36628	2.4434	**0.01455**
$Intercept_4$	1.08689	0.14938	7.2758	**$<2e^{-16}$**
IRA_y_4	0.02381	0.3449	0.069	0. 94496
$Gender\_y_4$	0.06488	0.21685	0.2992	0.76478
$Cavit\_y_4$	−2.02666	0.21162	−9.577	**$<2e^{-16}$**
$Loose\_y_4$	−0.51225	0.25708	−1.9926	**0.0463**
$Able\_y_4$	−0.07702	0.20244	−0.3805	0.7036
$Intercept_5$	0.7074	0.17148	4.1254	**$<4e^{-5}$**
IRA_y_5	−0.22746	0.32465	−0.7006	0.48353
$Gender\_y_5$	0.09493	0.2191	0.4333	0.66481
$Cavit\_y_5$	−0.92957	0.19173	−4.8484	**$<2e^{-16}$**
$Loose\_y_5$	0.98945	0.22902	4.3204	**$<2e^{-5}$**
$Able\_y_5$	0.54333	0.20724	2.6218	**0.00875**
$Intercept_6$	−0.26746	0.214	−1.2498	0.21136
IRA_y_6	−0.19855	0.43232	−0.4593	0.64604
$Gender\_y_6$	−0.20593	0.29128	−0.707	0.47957
$Cavit\_y_6$	−0.77173	0.26226	−2.9426	**0.00325**
$Loose\_y_6$	0.68872	0.29773	2.3133	**0.02071**
$Able\_y_6$	−0.06312	0.26756	−0.2359	0.81351
$Intercept_7$	−1.41215	0.25849	−15.463	**$<2e^{-16}$**
IRA_y_7	0.10559	0.46596	0.2266	0.82072
$Gender\_y_7$	0.1786	0.32314	0.5527	0.58046
$Cavit\_y_7$	−0.83001	0.33215	−12.4989	**0.01246**
$Loose\_y_7$	0.94458	0.35612	2.6524	**0.00799**
$Able\_y_7$	0.53696	0.30734	1.7471	0.08062

due to separate the effect of the covariate for each outcome, when they are highly correlated. In model 3, we find Gender parameter estimation still not significant for the all. Loose parameter estimation becomes significant only for outcome $Y_1, Y_{1\&2}, Y_{1\&3}$ and $Y_{1,2\&3}$ which means the high association between the "problem visit" outcome and the others hides the effect of Loose covariate in the outcome in model 1. To get a closer look at the parameters' differences in the three models, we aggregate the parameter estimations of all models for each covariate in Figure 1.

The dotplot in Figure 1 indicates many points. First, the variation of the parameter estimation between model 1, model 2 and model 3 are relatively. The variation in IRA and Gender parameter estimations over the three models is limited due to insignificant results. IRA parameter estimation is close to being significant in model 1, then model 2 explains the source of the effect which is the second outcome. The third model clarifies the effect after breaking up the responses for each joint case and finding that there is no significant estimations. In a contrary case, the parameter estimation of Able covariate starts significant in model 1 and 2 expect for the first outcome, then model 3 illustrates some significant parameters related to the significant association between outcomes 1 and 3. The Loose variable begins as not significant in model 1 but it is significant in model 2 for all outcomes while model 3 clarifies the sources of effect which is the associations between the three outcomes. The case of the intercept is clearly starting significant in model 1 and model 2, but model 3 explains the cause of the significance which are outcomes 1 and 3. The parameter of Cavit covariate shows how the third model changes the sign of the parameter to negative after using the coding method.
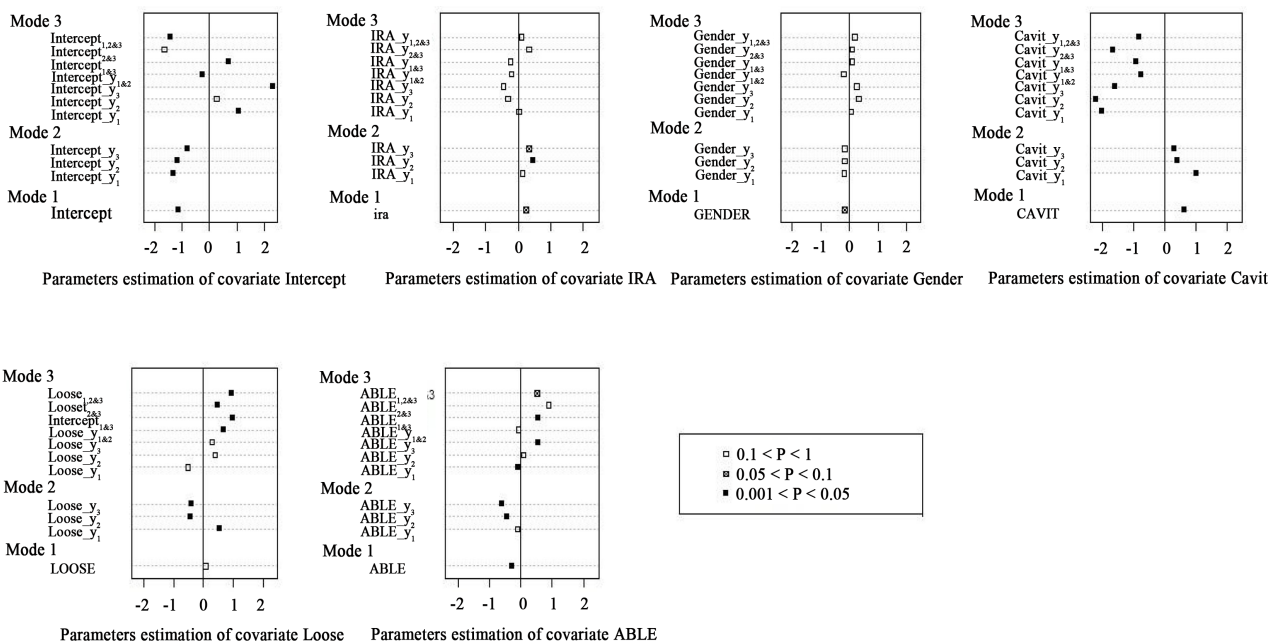

Figure 1. The parameter estimations of each covariant through the three models.

In summary, The FSCS study of three outcomes "problem visit", "cleaning" and "check up" and four intervals is a good example of highly correlated responses that are used to investigate parameter estimation differences of the three models. Over this example, we see that the effect of each covariate on the responses varies over the three models. The most significant covariates over the three models are Cavity, Loose tooth, and the ability to pay an unexpected US$500 dental bill, but with difficulty.

4. Discussion

In this article, we review two existing methods for analyzing binary multivariate longitudinal data and proposed a third method. In most longitudinal experiments, there are more than one outcomes are obtained from the subjects over many occasions. Usually, the researcher is interested in investigating the effect of many independent variables on these outcomes. We applied three different methods to be beneficial in this case. The aim of this article is to give a more general overview of analyzing binary multivariate longitudinal data in case of expecting a correlation among the outcomes other than the correlation among the repeated measurements.

If the researcher is interested in estimating one group the estimated effects of covariates for all outcomes and accounting of the multivariate structure, then the first method is the most appropriate one. However, if the researcher is eager to separate the effects of covariates on the outcomes, then the second method is appropriate. The joint analysis of all of the outcomes in the third method is beneficial to investigate the relationship between the correlated outcomes and how the covariates are affecting on them. The third method consists of encoding and decoding stages. The encoding stage is a step of defining each case of associations between the outcomes into simple number. It is an idea of store of all the possible combinations of the outcomes into one outcome. We converted three binary outcomes into one multinomial outcome. The decoding stage is the last step of returning each code to its original definition.

Then, as a result, the joint analysis in method 3 is limited to small number of outcomes because the number of parameters of estimation is rapidly increasing when the number of outcomes is increased. It would be beneficial for future work to improve the third method to reduce the dimension of the parameters estimations by adding a filtering stage between the encoding and decoding steps. We presented a simulation part for many scenarios and application. The FCSD example is applied to the three methods and we saw how the parameter estimations of the three methods were changed over covariates. Each method explains more details starting from model 1, and moving to models 2 and 3; each method answers a different question. In conclusion, the ultimate chosen model depends on the research goal.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Liang, K.Y. and Zeger, S.L. (1986) Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, **73**, 13-22. https://doi.org/10.1093/biomet/73.1.13

[2] Carey, V., Zeger, S. and Diggle, P. (1993) Modelling Multivariate Binary Data with Alternating Logistic Regressions. *Biometrika*, **80**, 517-526. https://doi.org/10.1093/biomet/80.3.517

[3] O'Brien, L. and Fitzmaurice, G. (2004) Analysis of Longitudinal Multiple-Source Binary Data Using Generalized Estimating Equations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **53**, 177-193. https://doi.org/10.1046/j.0035-9254.2003.05296.x

[4] Shelton, B., Gilbert, G., Liu, B. and Fisher, M. (2004) A SAS Macro for the Analysis of Multivariate Longitudinal Binary Outcomes. *Computer Methods and Programs in Biomedicine*, **76**, 163-175. https://doi.org/10.1016/j.cmpb.2004.05.005

[5] Hedeker, D. (2003) A Mixed-Effects Multinomial Logistic Regression Model. *Statistics in Medicine*, **22**, 1433-1446. https://doi.org/10.1002/sim.1522

[6] Touloumis, A., Agresti, A. and Kateri, M. (2013) GEE for Multinomial Responses Using a Local Odds Ratios Parameterization. *Biometrics*, **69**, 633-640. https://doi.org/10.1111/biom.12054

[7] Gilbert, G., Duncan, R.P., Kulley, A., Coward, T. and Heft, M. (1997) Evaluation of Bias and Logistics in a Survey of Adults at Increased Risk for Oral Health Decrements. *Journal of Public Health Dentistry*, **57**, 48-58. https://doi.org/10.1111/j.1752-7325.1997.tb02472.x

[8] Gilbert, G., Duncan, R. and Vogel, W. (1998) Determinants of Dental Care Use in Dentate Adults: Six-Monthly Use during a 24-Month Period in the Florida Dental Care Study. *Social Science U Medicine*, **47**, 727-737. https://doi.org/10.1016/S0277-9536(98)00148-8