Journal of the Faculty of Commerce for Scientific Research©
Volum 56 – No. 4 – October 2019

Estimating the Simultaneous Association-Marginal Model for Longitudinal Data with Missingness: *A Simulation Study* [1]

Mahi Mohssen El-Zayat [2] **Prof. Dr. Emtissal Mohamed** [3]
Prof. Dr. Adel Halawa [4] **Dr. Labiba El-Attar** [5]

Faculty of Commerce - Alexandria University

ABSTRACT

This paper introduces and applies a new model that describes simultaneously the association structure (A) with the marginal distributions (M) of the responses for longitudinal data in the presence of missing data (MS) through a composite link. This new model (AM-MS) is of great importance where it is applicable for large and sparse tables. In addition it can also be used for fitting log linear models to contingency tables with missing data (MS), fitting log linear models with some variables more finely categorized for some units than other units (sparse tables) and fitting models with various assumptions about the missing data mechanisms either MCAR, MAR or NMAR. A simulation study is conducted to apply this new idea, under various situations including (missing mechanisms, missing rates and five methods for handling missing data). The goodness-of-fit test statistics and the number of adjusted residuals greater than 2 are used as evaluation criteria. The results showed that after analyzing and

[1] Received in 9/6/2019, accepeted in 13/7/2019. This article is derived from Ph.D. research by Mahi Mohssen and supervised by Prof. Emtissal Mohamed, Prof. Adel Halawa and Dr. Labiba El-Attar.

[2] Mahi Mohssen Younes Mohamed El-Zayat Assistant Lecturer at Statistics, Mathematics and Insurance Department, Faculty of Commerce-Alexandria University.
(E-mail: stat4statalxx@gmail.com)

[3] Professor at Statestics, Mathematics and Insurance Department, Faculty of Commerce, Alexandria University.

[4] Professor at Statestics, Mathematics and Insurance Department, Faculty of Commerce, Alexandria University.

[5] Associate Professor at Statestics, Mathematics and Insurance Department, Faculty of Commerce, Alexandria University.

estimating the AM model with MS for MCAR with low missing rate, the best method for handling MS to estimate the AM model is LOCF while with high missing the best method for handling MS to estimate the AM model is the mode imputation method. For MAR the best method for handling MS is MI. But for NMAR with low missing rate, the best method for handling MS is also the LOCF method while for NMAR with high missing the best method for handling MS is the mode imputation method.

Keywords: Association model (A), Marginal model (M), Simultaneous AM model, Missing data (MS), Ordinal data, Composite link function, Generalized linear models (GLM), CC, mode imputation, LOCF, KNNI, MI, Longitudinal studies.

1- Introduction

Lang and Agresti, (1994) indicated that the analysis of longitudinal multivariate categorical (nominal or ordinal) response data is very common and useful in a variety of applications, especially for social studies. The longitudinal multivariate categorical responses are obtained from repeated measurements taken on subjects over time or occasions. These responses are often inevitably interrelated and the purpose of their modeling is to describe the association structure (changes at the individual level from one point to another) among these responses and also to know the behavior of their marginal distributions (changes within the year for different individuals). The Generalized Linear Models (GLM) are usually used for this analysis. Most of these models allow the researchers to model the association structure among these responses or to model their marginal distributions separately.

The common models which are useful for describing the association structure among the responses are the classical log linear models, (Balagtas, et al., 1995). Bergsma, et al., (2013) presented a second approach for analyzing multivariate categorical response which is to model only the marginal distributions and to ignore the joint distribution structure. Since the simultaneous models for the joint and marginal distributions became useful in a variety of applications, the recent years have seen a rapid development for analyzing and applying these simultaneous models.

Lang, et al., (1997) indicated that when the data are composed of several categorical responses together with categorical or continuous predictors are observed, and it is needed to describe simultaneously the association structure (A) between all these variables with the marginal distribution (M) of the response. Then a link function is used that lie between the two models of the log linear model that describes the association structure between the variables and the logistic regression model that describes the marginal

distributions of the responses.The model derived from these two models is known as the simultaneous association-marginal (AM) model which contains a composite link function that consists of both the log and the logit link. This AM model provides improved model parsimony,one also obtains a single test that summarizes goodness of fit and a single set of fitted values and residuals. Also, estimators of the simultaneous AM model parameters are more efficient than with separate fitting process procedures.

All the researchers whom introduced the AM model conditioned that all the data should be observed without any missing values, but missing data (MS) are often a problem for multivariate response data.

Missing data can be problematic for all researchers and statisticians. They occur when respondents participate in a survey but do not answer a certain question which is known as item nonresponse. Because of it, there are missing data recorded in some variables in a data set. Prior to data analysis, researchers must decide what to do with missing data because removing these observations decreases sample size, and thus decreases statistical power.

Thus, this study is looking for a parsimonious model (AM-MS) that can be used to simultaneously describe the association (A) structure among the responses and the marginal (M) distributions of the responses in the presence of multivariate categorical missing data (MS).

2- Missing Data Mechanisms and Methods for Handling

2-1 Missing Data Mechanisms

Rubin, (1976) defined a clear classification of missingness that has become the standard for any discussion of this topic. This classification depends on the reasons why data are missing. Rubin classified missing data mechanisms into three different types:

a) Missing completely at random (MCAR):

The MCAR assumption is defined as:

$$P(Y \text{ missing} \mid Y, X) = P(Y \text{ missing}).$$

This assumption states that missingness is not related to any factor, known or unknown in the study, (i.e. missingness is unrelated to the data). For example, if any student might have a missing value from any grade of his grades of the years of the faculty because he decided to travel and work in

other country, or because his family transported to other place or the student decided to complete his education abroad.

b) Missing at random (MAR):

It is *a* weaker assumption than MCAR. This assumption states that:

$$P(Y \text{ missing } |Y, X) = P(Y \text{ missing } |X)$$

Horton and Kleinman, (2007) described MAR mechanism and stated that the missingness depends only on observed quantities, which may include outcomes and predictors. For example; any student might have a missing value from any grade of his grades of the years of the faculty because he got a bad grade or failed in the previous year.

c) *Not missing at random (NMAR):*

Rithy, (2016) indicated that this is a case in which the probability of missingness for the variable of interest depends upon the value of that variable itself. For example; there is a high rate of missing data on an item asking about participants' annual income. It may be the case that participants with high rates of income are more likely to omit this item because they are uncomfortable with others knowing their income. The student may have a missing value from any grade of his grades when filling a survey because he does not want to tell his / her bad grade to anyone.

2-2 Methods for Handling Missing Data

2-2-1 Complete Case Analysis (CC)

This method deletes all cases with missing data and then performs statistical analyses on the remaining complete data set (which has a smaller sample size). Since all cases containing missing data have been removed, there is no missing data problem to handle. Therefore, all statistical methods can be used to analyze the smaller data set.

Zhu, (2014); Nakai, et al., (2014); Al-Zahrani, (2018) and Bori, (2013) indicated that one major advantage of this method is its ease of use. In fact, virtually all statistical programs incorporate this method as a default method because it accommodates any type of statistical analysis. The method may be preferred under the situation in which the sample size is large, the proportion of missing data is small, and the missing data mechanism is MCAR. For MCAR missing data, the method will yield unbiased parameter estimates.

While the disadvantages of this approach are that it results in loss of information because a large part of the original sample is excluded and it could possibly lead to losing statistical power due to the reduction of the sample size. Also, complete case techniques decrease the efficiency such that the variation (i.e., the standard error) around the true estimate is too large.

2-2-2 Mode Imputation

Baraldi and Enders, (2010) indicated that mode imputation method replaces missing values of a categorical variable by the mode of non-missing cases of that variable. Mode imputation is used when the missing mechanism is MCAR. It is one of the easiest ways in the case of categorical data is to fill in each missing value with the mode of observed values. This is a common practice; nonetheless, the major disadvantage of mode imputation is that it creates spikes in the distribution by concentrating all the imputed values in the mode. This is a single imputation method, since only one value is used to replace each missing observation.

2-2-3 Last Observation Carried Forward (LOCF)

Al-Zahrani, (2018) and Langkamp, *et al.* (2010) indicated that LOCF method is considered as the simplest imputation approach and can only be applied under a longitudinal study with MCAR mechanism. In this method the missing values are replaced by the last observed value from that variable. The advantage of this method is easy to understand and popular for handling missing data. Also, unlike the listwise deletion method, the sample size does not change. While the disadvantage of this method is that, it can bias results and lead to either overestimation or underestimation of the parameter estimates.

2-2-4 K-Nearest Neighborhood Imputation (KNNI)

Schlomer, et al., (2010) indicated that KNN imputation method uses the K-nearest neighbors approach to impute missing values. What KNN imputation does in simpler terms is as follows: For every observation to be imputed, it identifies 'K' closest observations based on the euclidean distance and computes the weighted average (weighted based on distance) of these 'K' observations. The advantage is that you could impute all the missing values in all variables with one call to the function. It takes the whole data frame as the argument and you don't even have to specify which variable you want to impute.

2-2-5 Multiple Imputation (MI)

Rubin, (1987) was the first to propose multiple imputation to analyze incomplete data under the MAR mechanism. Multiple imputation has one of the main advantages over any of the previous single imputation methods. Instead of replacing a single missing value, MI replaces each missing value multiple times and hence generates multiple (m) data sets. Then, the analyses are carried out using standard analysis procedures on each data set, with the parameter estimates and their standard errors saved for each data set. Finally, the parameter estimates from each imputed data set are combined to get a final set of parameter estimates. In other words, final results are obtained by averaging the parameter estimates across these multiple analyses, which results in an unbiased parameter estimate.

Schafer, (1997) and (1999) indicated that MI is a simulation-based procedure. Its purpose is not only to re-create the individual missing values as close as possible to the true ones, but also to handle missing data to achieve valid statistical inference.

Garg, (2013) and Kombo, et al., (2017) and Nooraee, et al., (2018) indicated to the major advantage of MI is that it allows the use of complete-data methods for data analysis and incorporating random errors in the imputation process. MI can accommodate any model with any data and does not require specialized software. In addition, MI increases the efficiency of the estimates through minimizing the standard errors. Also, the final standard errors of these parameter estimates which are based on the standard errors of the analysis of each data set are used for significance testing and/or construction of confidence intervals around these parameter estimates. Finally, the MI procedure provides accurate standard errors and therefore accurate inferential conclusions. So, the precision of parameter estimates and accuracy of standard errors make MI one of the best options for handling missing data.

Deng, et al. (2016) pointed out that MI needs more effort to create the multiple imputations, more time to run the analyses, and more computer storage space for the imputation-created data sets. Also, MI produces different results every time you use it because the imputed values are random draws rather than deterministic quantities.

3- Modeling the AM Model with MS Model (AM-MS)

In this section the new model will be introduced that simultaneously describes the association structure (A) of the responses with the marginal distributions (M) of these responses when the data contain missing values (MS) through a composite link.

3-1 Modeling Missing Data (MS)

Rindskopf, (1992) introduced, described and illustrated a general approach for analyzing categorical data when there are missing values on one or more observed variables. This approach is based on the GLM of Mccullagh and Nelder, (1989) with composite links to include cases in which expected values corresponding to observed data are composites of elements of m that may not correspond to directly observed values.

The systematic component of any GLMs can be written as:

$$\eta = g(m) = X\beta \qquad (1)$$

where $\eta = g(m)$ is the link function relating m which is the expected value of the dependent variable to the model term $X\beta$, X is the design matrix and β is the vector of parameters. Then $m = h(\eta)$, where $h(.)$ is the inverse of $g(.)$ and $h(.)$ is an exponential function. In the usual GLM, observed data are being modeled and the problems of missing data are beyond the scope of GLMs. But as MS is a problem, it should be under scope and attention. Thus, how to model missing data using GLMs?

Rindskopf, (1992) proposed an extension of the GLM to include cases in which expected values corresponding to observed data are composites of elements of m, which may not correspond to directly observed values but correspond to missing values. He considered linear functions of elements of m, expressed in the form $m^* = Fm$, where the matrix F (consisting of 0's and 1's) tells which elements of the unobserved vector m are summed to result in an estimated observed frequency.

Thus, the GLM with composite links can be expressed as:

$$E(Y) = m^* = Fm, \qquad m = h(\eta) \qquad \eta = X_o\beta_0 \qquad (2)$$

where Y and m^* are $n \times 1$ vectors, F is a $n \times k$ matrix of 0's and 1's, m is a vector of $k \times 1$, $h(\eta)$ is the inverse of the link function, X_0 is a $k \times p$ matrix and β_0 is a $p_0 \times 1$ vector.

Finally, the GLM for missing data which models the frequencies with expected value m in an unobserved table is:

$$m^* = F \exp(X_o\beta_0) \qquad (3)$$

Rindskopf, (1992) indicated that if the categorical data are available on a certain number of variables but some cases have missing values on some of

the variables, then the cases with complete data produce a complete crosstabulation of the variables and the cases with missing values produce the supplemental marginal tables. The supplemental marginal tables are the cases with missing values of certain variables which will produce a marginal table of observed frequencies for the missing variables.

The method presented by Rindskopf, (1992) can be used in many situations including:

- Fitting log linear models to contingency tables with missing data (MS).
- Fitting log linear models with some variables more finely categorized for some units than other units (sparse tables).
- Fitting models with various assumptions about the missing data mechanisms; the data may be MCAR, MAR or NMAR.
- Fitting latent class models with missing data on observed variables.
- Filling in contingency tables with missing data (i.e. contingency tables with supplementary margins).

3-2 Modeling the Association-Marginal Model (AM)

Lang and Eliason, (1997) were the first to simultaneously model the association structure (A) with the marginal distributions (M) using association marginal (AM) model. The AM model's link function is a composite link because it contains both the log and the logit links. This composite link is suitable especially in sparse data situations when there are few covariate patterns and many response profiles.

Lang and Eliason, (1997) presented the multinomial AM model as an intersection of a log linear A model with the form of $\log m = X_1\beta_1$ and a generalized log linear M model with the form $C_2 \log M_2 m = X_2\beta_2$ and the sampling constraint was included. Combining the A and the M models, the multinomial AM model can be written as:

$$\text{AM: } C_2 \log M_2 e^{X_1\beta_1} = X_2\beta_2, \; samp\,(m)=0. \qquad (4)$$

There are many situations in which simultaneous models for joint and marginal distributions may be useful. One situation is the longitudinal data especially in social studies for example, social mobility studies. In the longitudinal study it is required to determine the gross change which is related with the change at the individual level from one point to another through modeling of the joint distribution. In these studies it is also required to determine the net or aggregate change within the year through the modeling of marginal distributions of the responses. Besides the flexibility of the simultaneous AM models, other benefits come in terms of

model parsimony and more efficient estimators of cell expected frequencies and model parameters.These estimators are potentially more efficient than with separate fitting procedures. In addition, a single test that simultaneously summarizes goodness of fit and a single set of the fitted values and residuals can be obtained. Also, the simultaneous AM models are applicable for large and sparse tables, where these models avoid some problems associated with sparse tables and sampling zeros.

Lang and Eliason, (1997) conditioned that for estimating the AM model, that all the data should be complete (fully observed) without any MS, but MS are often a problem for multivariate response data and should be treated in a good way to get valid inferences.

3-3 Modeling simultaneously the AM Model with MS (AM-MS)

In this subsection a new model (AM-MS) will be introduced that simultaneously describes the association structure (A) of the responses with the marginal distributions (M) of the responses when the data contain missing values (MS) through a composite link. This new model will combine the A model with the M model in the presence of MS through composite link. The new model will combine the two models:

$$\text{AM: } C_2 \log M_2 e^{X_1\beta_1} = X_2\beta_2 \text{ , } samp\,(m)=0$$

$$\text{MS: } m^* = F \exp\,(\,X_0\beta_0\,)$$

Thus, by combining the AM model with the MS model the simultaneous AM models will be applicable for large and sparse tables with MS. Also, these simultaneous AM model with MS can be used in fitting log linear models to contingency tables with missing data (MS). Besides the previous advantages, this new model (AM-MS) can be used for fitting AM models with MS by assuming various assumptions about the missing data mechanisms (MCAR, MAR or NMAR) and different missing rates. Also, AM-MS can be used for comparing AM models after applying the different methods for handling MS to choose the best method for treating MS in the AM model in each missing mechanism with each missing rate.

4- The Simulation Study

4-1 Design of the Simulation

To achieve the research's goal, a simulation study was performed to simulate four responses each with three levels (J = 3) using the *SimCorMultRes* package version 1.4.1 in R. Touloumis (2016) and (2018)

indicated that this package is the first R package that targets specifically on the generation of correlated binary, nominal or ordinal responses under marginal model specification.

The *rmul.clm* function in the *SimCorMultRes* package was used to generate ordinal data Y_{it} (*i=1,2,....,N, t=1,... ,T*) for *i*-th subject at *t*-th occasion. The simulation of the data was conducted according to a cumulative logit model:

$$\text{logit}\,[P(Y \leq j)] = \alpha_j + \beta x \qquad (5)$$

where α_j is the intercept for level *j* and β is the slope when using one explanatory variable, *x*. Here in this research each response has J = 3 categories, then there will be 2 intercepts only (α_j = 0.5, 1.5) since models for cumulative probabilities do not use the final one, *P* (*Y* ≤ *J*), since it necessarily equals 1. In this model the parameter β which is the slope (β = 1.5) describes the effect of *X* on the log odds of response in category *j* or below. In the model formula, β does not have a *j* subscript; this means that the model assumes an identical effect of *X* for all *J*-1 logits. The intercepts and the slope are assumed to be constant through the simulation study.The correlation coefficient (*ρ)* between the responses is also assumed to be constant with a positive correlation coefficient of value equal 0.2 .The total number of cases or subjects simulated (N) is 200 and there are 4 responses for each subject. Future researcher will consider other values for α_j and β where the values of them are from the *SimCorMultRes* package.

4-2 Fully Observed Data Generation

Firstly, the AM model will be estimated and analyzed with fully observed data and without any MS using the 200 simulated subjects with the four responses per subject.The A model is expressed as a linear-by-linear model, while the M model is expressed as a proportional odds model. These two models are suitable and parsimonious forms when the study contains ordinal variables, where the linear-by-linear association (A) model and proportional odds marginal (M) model takes into account the ordering of the variables categories. The main maximum-likelihood fitting program, *mph.fit* version 3.1 is used to estimate simultaneously the AM model with fully observed data (Lang, 2009).The contingency table for these simulated 200 subjects is displayed in Table 1 as a 3^4 contingency table.Then the coefficients(*BETA*), their standard errors (*StdErr(BETA), Z-ratio* and the *p-values* for this AM model with fully observed data are obtained and displayed in Table 2 .

Table 1: A contingency table for N=200, J=3and ρ =0.2

Y3			**1**			**2**			**3**	
Y4		**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
Y1	***Y2***									
	1	57	7	5	7	0	2	3	3	0
1	2	9	0	0	1	2	3	1	1	0
	3	6	3	2	2	0	0	0	1	1
	1	4	2	4	4	0	1	1	0	0
2	2	1	0	0	0	0	1	2	0	0
	3	1	2	2	0	0	0	1	3	2
	1	4	0	1	2	1	2	1	2	2
3	2	0	1	1	0	0	3	3	1	4
	3	1	2	2	1	0	1	1	2	15

The first cell of this table contains the count 57, where 57 out of 200 simulated individuals, response one in the first (*Y1*) occasion, one in the second (*Y2*) occasion, one in the third *(Y3*) occasion and one in the fourth (*Y4*) occasion and so on for the rest of counts.

Table 2: Estimates of parameters of the AM model with fully observed data

	BETA	StdErr (BETA)	Z-ratio	p-value
Intercept	-3.5690	0.1578	-22.612	0.000e+00
Y12	-3.3633	0.3130	-10.7443	0.000e+00
Y13	-5.7291	0.7434	-7.7062	1.2879e-4
Y22	-2.7217	0.2602	-10.4619	0.0000e+00
Y23	-4.1238	0.5822	-7.0834	1.4062e-12
Y32	-3.0555	0.2863	-10.6731	0.0000e+00
Y33	-4.9538	0.6638	-7.4623	8.5043e-14
Y42	-2.8726	0.2713	-10.5898	0.0000e+00
Y43	-4.5372	0.6207	-7.3093	2.6845e-13
Y1score:Y2score	0.2785	0.1262	2.2075	2.7277e-02
Y1score:Y3score	0.6029	0.1281	4.7047	2.5430e-06
Y1score:Y4score	0.4880	0.1264	3.8618	1.1257e-04
Y2score:Y3score	0.3016	0.1220	2.4724	1.3423e-02
Y2score:Y4score	0.3484	0.1177	2.9609	3.0674e-03
Y3score:Y4score	0.2263	0.1256	1.8016	7.1616e-02
CUT1	0.3036	0.1370	2.2161	2.6683e-02
CUT2	1.0492	0.1490	7.0426	1.8872e-12
RESPY2	0.0082	0.1521	0.0540	9.5695e-01
RESPY3	0.0449	0.1362	0.3293	7.4194e-01
RESPY4	-0.0361	0.1416	-0.2552	7.9860e-01

MODEL GOODNESS OF FIT: Test of Ho: h(m)=0 vs. Ha: not Ho
Likelihood Ratio Stat (df= 69): Gsq = 85.51392 (pval = 0.08647)
Pearson's Score Stat (df= 69): Xsq = 82.14554 (pval = 0.1333)
Generalized Wald Stat (df= 69): Wsq = 33.75538 (pval = 0.9999)

Adj Resids: -1.862 -1.497 ... 2.422 3.636 , Number |Adj Resid| > 2: 5

Remark: Y12….Y43: the main effects association terms,Y1score: Y2score is a linear-by-linear association term, α_1= CUT1, α_2=CUT2, RESPY2 is the value of the second response in the marginal model, G^2 is the likelihood ratio statistic, χ^2 is Pearson's Score statistic, $W^2$is generalized Wald statistic, (*) is the p-value corresponding each test statistic and Adj Resids are the number of adjusted residuals which are greater than 2.

The analysis of this model showed that G^2 = 85.51 with p-value = 0.09, which mean not to reject H_o and so this model fits the data well and is significant. Also, similar information can be taken from χ^2, which compares the observed cell counts with the expected cell counts to judge whether the data contradicts Ho.

The value of χ^2 = 82.15 with p-value = 0.133, which also means to accept H_o. While Hedeker and Gibbons (2006) indicated that the W^2 is a multi-parameter Wald test which is used to test the joint null hypothesis that the set of β_s' of the more general model but not in the model of interest equal zero. Here, W^2= 33.76 with p-value= 0.999 which also means to accept H_o that the β 's of the more general model equal zero and this model fits well. Thus, the three statistics mean not to reject H_o and so this model fits the data well and is significant. This also implies a good fit and provides no evidence of lack of fit.

The significance of this model can also be obtained by comparing observed and fitted counts individually, using the adjusted residuals for a cell-by-cell, where it is known that an adjusted residuals larger than 2 in absolute value, indicates lack of fit in that cell. Accordingly, this model fits well: there are only 5 out of 81 adjusted residuals having absolute value greater than 2.

A similar conclusion can be obtained from further insight of the estimates of the AM model and their p-values. The p-values of the estimates of the A model including the main effects of the four responses and the linear-by-linear association terms are very small rejecting the null hypothesis of zero valued coefficients of the A model. It should be noted that the maximum likelihood estimates (MLE) of each two adjacent points in time are positive with standard errors not greater than one. This implies a positive relationship among the adjacent responses. Also the estimates of the M model including the intercepts and slopes are highly significant with very small p-values, rejecting the null hypothesis of zero valued coefficients of M model.

Also, adjusted residuals for marginal proportions are studied as inadequacies may result from the marginal model and are displayed in Table 3. Thus, the observed marginal proportions are compared to estimated marginal proportions. This comparison shows no lack of fit as none of the adjusted residuals for marginal proportions exceeds 2. This also proves that this AM model is significant. Finally, this model fits the data well.

Table 3: Marginal Adjusted Residuals for the Simultaneous AM model with fully observed data

		Observed Marginal Proportions	Estimated Marginal Proportions	Adjusted Residuals
Y1	1	0.5753	0.5800	0.4651
	2	0.1653	0.1550	-0.4695
	3	0.2594	0.2650	0.4723
Y2	1	0.5773	0.5750	-0.2389
	2	0.1649	0.1700	0.2345
	3	0.2578	0.2550	0.2310
Y3	1	0.5862	0.5850	-0.1271
	2	0.1629	0.1650	0.0960
	3	0.2509	0.2500	-0.0708
Y4	1	0.5665	0.5650	-0.1453
	2	0.1672	0.1700	0.1293
	3	0.2664	0.2650	-0.1156

4-3 Missing Data Generation

In this subsection MS is inserted and injected in the fully observed data and the performance of five methods (that is, CC analysis, mode imputation, LOCF, KNNI and MI) for handling MS in the AM model was compared. The comparison between the methods was based on the three goodness-of-fit test statistics, and the number of adjusted residuals greater than 2 for each AM model as evaluation criterion. In the simulation, the missing data was considered using the three missing mechanisms (MCAR, MAR and NMAR). In addition, without loss of generality, the missing pattern was assumed to be arbitrary, where missingness can occur at any point in time and to any

subject. The missing rate was assumed to be low missing rate (10%) and high missing rate (50%).

Thus, the AM model will be estimated using the following cases; for each of MCAR and NMAR there will be 10 different cases: 2 (missing rates) × 5 (methods for handling MS). While for MAR mechanism is used without using missing rates as the MS in this mechanism depends on observed values of other variable without MS, thus will be only the 5 methods for handling MS. Here there will be 25 cases; 10 for MCAR, 10 for NMAR and 5 for MAR. Therefore, there will be 25 different cases.

5- Simulation Results

This section reports the results of the simulation study comparing the effect of the methods CC, mode imputation, LOCF, KNNI and MI on handling MS in the AM model.

5-1 Simulation Results for MCAR

Table 4 shows the simulation results for estimating the AM model with MCAR missing mechanism, and 10% missing rate using CC, mode imputation, LOCF, KNNI and MI.

Thus, after handling MS in the AM model and depending on the goodness-of-fit test statistics and the number of adjusted residuals greater than 2 as evaluation criteria, the best method for handling MS with low missing rate in this AM model is LOCF while the worst method for handling MS with low missing rate in this AM model is KNNI.

Table 5 shows the simulation results for estimating the AM model with MCAR missing mechanism, and 50% missing rate using CC, mode imputation, LOCF, KNNI and MI.

Thus, after handling MS in the AM model and depending on the goodness-of-fit test statistics and the number of adjusted residuals greater than 2 as evaluation criteria, the best method for handling MS with high missing rate in this AM model is the mode imputation method while the worst method for handling MS with high missing rate in this AM model is MI method.

Table 4: Goodness-of-Fit test statistics of the AM model with MCAR missing mechanism and 10% missing rate using CC, mode imputation, LOCF, KNNI and MI

	CC	Mode imputation	LOCF	KNNI	MI
G^2	81.429 (0.1454)	71.64 (0.3903)	70.84 (0.42)	85.99 (0.081)	82.141 (0.1334)
χ^2	86.415 (0.0765)	80.542 (0.162)	66.25 (0.57)	88.099 (0.060)	80.045 (0.171)
W^2	34.399 (0.9998)	38.84632 (0.998)	30.193 (1)	39.938 (0.998)	35.587 (0.9997)
adj. resd. > 2	7	4	2	8	4

Remark: G^2 is the likelihood ratio statistic, χ^2 is Pearson's Score statistic, $W^2$is generalized Wald statistic, (*) is the p-value corresponding each test statistic and adj. resd are the number of adjusted residuals which are greater than 2 in each case.

Table 5: Goodness-of-Fit test statistics of the AM model for MCAR missing mechanism and 50% missing rate using CC, mode imputation, LOCF, KNNI and MI

	CC	Mode imputation	LOCF	KNNI	MI
G^2	37.872 (0.9992)	53.27 (0.9191)	105.878 (0.0029)	126.175 (3.244e-05)	197.99 (2.154e-14)
χ^2	95.064 (0.0205)	58.79 (0.80)	127.616 (2.283e-05)	166.606 (4.777e-10)	236.6779 (0)
W^2	15.018 (1)	26.74 (1)	49.939 (0.9594)	48.16621 (0.9734)	69.47938 (0.4612)
adj. resd. > 2	5	4	8	10	14

Remark: G^2 is the likelihood ratio statistic, χ^2 is Pearson's Score statistic, $W^2$is generalized Wald statistic, (*) is the p-value corresponding each test statistic and adj. resd are the number of adjusted residuals which are greater than 2 in each case.

5-2 Simulation Results for MAR

Table 6 shows the simulation results for estimating the AM model with MAR missing mechanism using CC, mode imputation, LOCF, KNNI and MI.

Thus, after handling MS in the AM model and depending on the goodness-of-fit test statistics and the number of adjusted residuals greater than 2 as evaluation criteria, the best method for handling MS in this AM model is MI while the worst method for handling MS in this AM model is KNNI.

Table 6: Goodness-of-Fit test statistics of the AM model for MAR missing mechanism using CC, mode imputation, LOCF, KNNI and MI

	CC	Mode imputation	LOCF	KNNI	MI
G^2	65.657 (0.592)	67.862 (0.516)	76.549 (0.249)	138.129 (1.569e-06)	74.90118 (0.293)
χ^2	68.829 (0.483)	79.627 (0.179)	77.389 (0.229)	202.054 (5.551e-15)	76.64593 (0.247)
W^2	28.503 (1)	36.455 (0.9996)	39.14861 (0.9986)	76.35734 (0.2541)	32.90313 (0.999)
adj. resd. > 2	3	3	6	7	3

Remark: G^2 is the likelihood ratio statistic, χ^2 is Pearson's Score statistic, $W^2$is generalized Wald statistic, (*) is the p-value corresponding each test statistic and adj. resd are the number of adjusted residuals which are greater than 2 in each case.

5-3 Simulation Results for NMAR

Table 7 shows the simulation results for estimating the AM model with NMAR missing mechanism, and 10% missing rate using CC, mode imputation, LOCF, KNNI and MI.

Thus, after handling MS in the AM model and depending on the goodness-of-fit test statistics and the number of adjusted residuals greater than 2 as evaluation criteria, the best method for handling MS with low missing rate

in this AM model is LOCF while the worst methods for handling MS for this AM model are KNNI and MI.

Table 8 shows the simulation results for estimating the AM model with NMAR missing mechanism, and 50% missing rate using CC, mode imputation, LOCF, KNNI and MI.

Thus, after handling MS in the AM model and depending on the goodness-of-fit test statistics and the number of adjusted residuals greater than 2 as evaluation criteria, the best method for handling MS with high missing rate in this AM model is mode imputation method while the worst method for handling MS with high missing rate in this AM model is KNNI.

Table 7: Goodness-of-Fit test statistics of the AM model with NMAR missing mechanism and 10% missing rate using CC, mode imputation, LOCF, KNNI and MI

	CC	**Mode imputation**	**LOCF**	**KNNI**	**MI**
G^2	80.14095 (0.1691)	82.88376 (0.121)	76.36003 (0.254)	96.49399 (0.0161)	97.99125 (0.01244)
χ^2	84.12979 (0.1038)	81.95467 (0.1365)	77.11963 (0.2351)	91.34245 (0.0372)	92.07333 (0.03323)
W^2	35.79652 (0.9997)	35.55548 (0.9997)	37.80 (0.9992)	34.38388 (0.9998)	37.38494 (0.9993)
adj. resd. > 2	6	5	6	8	6

Remark: G^2 is the likelihood ratio statistic, χ^2 is Pearson's Score statistic, $W^2$is generalized Wald statistic, (*) is the p-value corresponding each test statistic and adj. resd are the number of adjusted residuals which are greater than 2 in each case.

Table 8: Goodness-of-Fit test statistics of the AM with NMAR missing mechanism and 50% missing rate using CC, mode imputation, LOCF, KNNI and MI

	CC	Mode imputation	LOCF	KNNI	MI
G^2	67.075 (0.5432)	88.32753 (0.06)	106.159 (0.0027)	153.71 (2.143e-08)	111.4183 (0.00093)
χ^2	70.015 (0.4433)	86.33004 (0.08)	94.6119 (0.0221)	160.58 (2.903e-09)	116.6429 (0.00029)
W^2	22.19 (1)	37.62378 (0.9993)	41.83862 (0.996)	57.67014 (0.832)	56.86323 (0.8516)
adj. resd. > 2	7	7	8	13	6

Remark: G^2 is the likelihood ratio statistic, χ^2 is Pearson's Score statistic, $W^2$is generalized Wald statistic, (*) is the p-value corresponding each test statistic and adj. resd are the number of adjusted residuals which are greater than 2 in each case.

6- Conclusions

In this paper, a longitudinal study was considered with four responses each with three levels and a total of 200subjects. Two possible missing rates and five methods for handling MS were used to detect the effect of the methods of handling MS on the AM model. In addition, three missing mechanisms were considered (that is, MCAR, MAR and NMAR). Based on the simulation results, we have reached the following important conclusions:

1. Although imputation procedures are often useful, in this paper it is noted that no universally best approach to handle missingness exists. Every method suffers from limitations related to the missing data mechanism. Nonetheless, understanding why data are missing can guide the researcher to an appropriate strategy for addressing missingness.

2. In addition, another important outcome of this paper is that it investigated how the performance of the AM model was affected by varying rates of missing data, different missing data mechanisms and the methods used for handling MS.

3. In general, it should be pointed out that MCAR, MAR and MNAR mechanisms led to dissimilar results for the missing rates and a given imputation method.

4. Under the MCAR and NMAR mechanisms with 10% missing rate, LOCF method performed well. This indicates that the LOCF method can be applied if the proportion of missing values is low, as indicated in table 4 and table 7 where LOCF had the smallest goodness of fit test statistics.

5. Also, under the MCAR and NMAR mechanisms but with 50% missing rate, the mode imputation method performed well, as indicated in table 5 and table 8 where mode imputation method had the smallest goodness of fit test statistics and the smallest number of adjusted residuals greater than 2. But for MAR mechanism the MI method performed well as indicated in table 6 where MI was the best method.

6. For the three missing mechanisms either with low or high missing rate, KNNI was the worst method as it had the largest values of test statistics and very small p-values as indicated in table 4, 5, 6, 7 and 8.

7. The results in general revealed that MI is likely to be the best under the MAR

8. mechanism as indicated in table 6.

9. Also, it should be noted that the CC method concludes good imputation method with condition of small missing percentage and large sample size to improve its disadvantages.

7- Recommendations

Finally, as with any study, there are limitations to the current work that must be considered. First, the simulations were based only under the arbitrary missingness pattern. Second, the current paper focused on ordered categorical data which is not common in the surveys. Also, in this paper, MS was only on the response variables. This does not limit the applicability of MS to the AM model. The methods can be extended to situations where data are missing for responses and covariates and will be applicable for the AM model. Other values for J, α_j, β and ρ may be used to detect their effects, also different number of responses with different number of levels together may be used to see the effect of MS on the AM model with different cases.

8. References

Al- Zahrani, H. (2018). Missing Data Analysis for Binary Multivariate Longitudinal Data through a simulation Study, *Biometrics and Biostatistics International Journal,* 7(2), 103-113

Balagtas, C.; Becker, M. and Lang, J. (1995). *Marginal Modeling of Categorical Data from Crossover Experiments,* Applied Statistics, 44, 63-77.

Baraldi, A. and Enders, C. (2010). An Introduction to Modern Missing Data Analyses*, Journal of School Psychology*, 48, 5-37.

Bergsma, W.; Croon, M. and Hagenaars, J. (2013). Advancements in Marginal Modeling for Categorical Data (with discussion), Sociological Methodology, *http://www.stats.lse.ac.uk/bergsma/pdf.*

Bori, M. (2013). Dealing with Missing Data: Key Assumptions and Methods for Applied Analysis, *https://www.bu.edu/sph/files/2014/05/Marina-tech-report.pdf*

Deng, Y.; Chang, C.; Ido, C. and Long, Q. (2016). Multiple Imputation for General Missing Data Patterns in the Presence of High-Dimensional Data, *Scientific Reports*, 6.

Garg, P. (2013). Robustness of Multiple Imputation under Missing At Random (MAR) Mechanism: A simulation Study, Theses and Dissertations, 35, 1-206.

Hedeker, D. and Gibbons, R. (2006). Longitudinal Data Analysis, *New Jersey: Wiley*.

Horton, N. and Kleinman, K. (2007). Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models, *The American Statistician*, 61, 79-90.

Kombo, A.; Mwambi, H. and Molenberghs, G. (2017). Multiple Imputation for Ordinal Longitudinal Data with Nonmonotone Missing Data Patterns. *Journal of applied statistics*, 44 (2), 270-287

Lang, J. and Agresti, A. (1994). Simultaneously Modeling Joint and Marginal Distributions of Multivariate Categorical Responses, *Journal of the American Statistical Association*, 89,625-632.

Lang, J.; Mcdonald, J. and Smith, P. (1997). Association-Marginal Modeling of Multivariate Categorical Responses. A Maximum Likelihood

Approach for Large, Sparse Tables, University of Iowa. *Technical Report* No. 263.

Lang, J. and Eliason, S. (1997). Applications of Association-Marginal Models to the Study of Social Mobility, *Sociological Methods and Research*, 26, 183-212.

Lang, J. (2009). Using *MPH.Fit* to Fit Standard Contingency Table Models, *http://www.stat.uiowa.edu/~jblang/mph.fitting.*

Langkamp, D.; Lehman, A. and Lemeshow, S. (2010). Techniques for Handling Missing Data in Secondary Analyses of Large Surveys, *Academic Pediatrics,* 10, 205-210.

Mccullagh, P. and Nelder, J. (1989). *Generalized Linear Models; Second Edition*, London: Chapman and Hall.

Nakai, M.; Chen, D.; Nishimura, K. and Miyamoto, Y. (2014). Comparative Study of Four Methods in Missing Value Imputations Under Missing Completely At Random Mechanism, *Open Journal of Statistics*, 4, 27-37.

Nooraee, N.; Molenbergh, G.; Ormel, J. and Heuvel, E. (2018). Strategies for Handling Missing Data in Longitudinal Studies with Questionnaires. *https://doi.org/10.1080/00949655.2018.1520854*.

Rindskopf, D. (1992). A General Approach to Categorical Data Analysis with Missing Data, *Using Generalized Linear Models with Composite Links*, Psychometrika, 57, 29- 42.

Rithy, D. (2016). Simulation of Imputation Effects under Different Assumptions, *http://www.digitalcommons.calpoly.edu/cgi/viewcontent. cgi?,1-64*.

Rubin, D. (1976). *Inference and Missing Data, Biometrika*, 63, 581-590.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys,* New York: John Wiley and Sons.

Schafer, J. (1997). *Analysis of Incomplete Multivariate Data*, New York: Chapman and Hall.

Schafer, J. (1999). Multiple Imputation: Primer, *Statistical Methods in Research*, 8, 3–15.

Schafer, J. and Graham, J. (2002). Missing Data: Our View of the State of the Art, *Psychological Methods*, 7, 147-177.

Schlomer, G.; Bauman, S. and Card, N. (2010). Best Practices for Missing Data Management in Counseling Psychology, *Journal of Counseling Psychology*, 57, 1-10.

Touloumis, A. (2016). Simulating Correlated Binary and Multinomial Responses under Marginal Model Specification: The SimCorMultRes Package. The R Journal, 8 (2), 79–91, *https://journal.r-project.org/ archive/ 2016/RJ-2016-034/index.html*.

Touloumis, A. (2018). Simulating Correlated Binary and Multinomial Responses with SimCorMultRes, *http://github.com/Anestis Touloumis/Sim Cor MultRes*.

Zhu, X. (2014). Comparison of Four Methods for Handling Missing Data in Longitudinal Data Analysis Through a Simulation Study, *Open Journal of Statistics*, 4,933-944.

ملخص البحث باللغة العربية

يعتمد النموذج الخطى فى التقدير على التوزيع الطبيعى لمتغيرات الاستجابة ولكن فى كثير من الاحيان قد لا يحدث ذلك كما فى حالة المتغيرات ذات الفئات. فقد تم التوصل فى السنوات الماضية الى فئة النماذج الخطية المعممة (GLMs) التى قدمت العديد من النماذج الهامة التى يمكن ان تستخدم لتحليل المتغيرات ذات الفئات مثل نموذج اللوجت ونموذج اللوغاريتمى الخطى. قدم Rindskopfطريقة لتحليل المتغيرات ذات الفئات فى حالة وجود بيانات مفقودة (MS) فى متغير واحد او أكثر اعتمادا على النماذج الخطية المعممة. كما قدم Lang and Eliason التفاعلات بين متغيرات الاستجابة و التوزيعات الهامشية انيا باستخدام نموذج الاقتران- الهامشى (Association-Marginal). يتكون نموذج الاقتران- الهامشى من نموذجين: نموذج الاقتران (A) لوصف التفاعل و العلاقة بين المتغيرات و له رابطة لوغاريتمية يعتمد النموذج الخطى فى التقدير على التوزيع الطبيعى لمتغيرات الاستجابة. ولكن فى كثير من الاحيان قد لا يحدث ذلك كما فى حالة المتغيرات ذات الفئات. فقد تم التوصل فى السنوات الماضية الى فئة النماذج الخطية المعممة (GLMs) التى قدمت العديد من النماذج الهامة التى يمكن ان تستخدم لتحليل المتغيرات ذات الفئات مثل نموذج اللوجت و والنموذج الهامشى (M) لوصف التوزيعات الهامشية للمتغيرات التابعة و له رابطة اللوجت. بذلك يحتوى نموذج الاقتران- الهامشى (AM) على رابطة مركبة تحتوى على الرابطة اللوغاريتمية و رابطة اللوجت. أوضح Lang and Eliason ان نموذج الاقتران- الهامشى يعتبر مناسبا فى حالة جدول اقتران به العديد من الخلايا الصفرية. تعتبر البيانات المفقودة مشكلة كبرى لكثير من الباحثين ومحللى البيانات. البيانات المفقودة تؤدى الى نقص فى حجم العينة وبالتالى نقص فى الكفاءة الاحصائية للنموذج. ومن هنا تتلخص مشكلة البحث فى التوصل لنموذج لوصف التفاعلات بين المتغيرات التابعة والتوزيعات الهامشية انيا باستخدام نموذج الاقتران- الهامشى و ذلك فى ظل وجود بيانات مفقودة (AM-MS).

الكلمات المفتاحية: نماذج الإقتران (A)، والنماذج الهامشية (M)، ونماذج الإقتران- الهامشية الآنية AM، والبيانات المفقودة (MS)، بيانات الترتيب، والدالة المركبة، والنماذج الخطية المعممة GLM، وتحليل البيانات الكاملة CC،والإحلال بالمنوال، وضع أخر مشاهدة LOCF، وضع أقرب مشاهدة مجاورة KNNI، والإحلال بمشاهدات متعددة MI وبيانات طولية.

Suggested Citation according to APA Style

El-Zayat, M.; Mohamed, E.; Halawa, A. and El-Attar, L. (2019). Estimating the Simultaneous Association-Marginal Model for Longitudinal Data with Missingness "A Simulation Study". *Journal of the Faculty of Comemrce for Scientific Research*, *Faculty of Commerce, Alexandria University*, 56(4), 205 – 228.