



Log-linear Models and Closed Form Estimates for Missing Values in Two Dimensional Contingency Tables

Emine Öçal^{1,2,a,*}, Ayfer Ezgi Yılmaz Çakıroğlu^{3,b}

¹ Department of Statistics, Dokuz Eylül University, İzmir, Türkiye.

² Graduate School of Science and Engineering, Hacettepe University, Ankara, Türkiye.

³ Department of Statistics, Hacettepe University, Ankara, Türkiye.

*Corresponding author e-mail address: emine.ocal@deu.edu.tr

Research Article

History

Received: 21.12.2024

Accepted: 26.03.2026

ABSTRACT

The problem of missing data is frequently encountered in scientific research due to various reasons such as nonresponse in surveys, data recording errors, data loss, or limitations inherent in the study design. Missing data mechanisms are classified into three categories: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). In the categorical data analysis, in contingency tables, the direct application of log-linear models in the presence of missing observations in one or more variables may lead to biased or misleading results. Therefore, in order to obtain valid statistical inferences, the missing data problem must be addressed using appropriate methodological approaches prior to analysis. In this study, log-linear models and their closed-form estimators are examined for two-dimensional contingency tables under scenarios where missing data occur in one variable as well as in both variables simultaneously. An illustrative example is conducted using the *Myocardial Infarction Complications* dataset, and the results are evaluated. The findings demonstrate that closed-form estimators provide an effective and interpretable framework for analyzing contingency tables with missing data, enabling reliable inference under different missing data mechanisms.



This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

Keywords: Categorical data, Closed-form estimates, Contingency tables, Log-linear models, Missing data.

^a 0009-0004-5657-8458

^b 0000-0002-6214-8014

1. Introduction

Missing data refer to the absence of data in one or more cells of a dataset and are a common issue in scientific research. They can occur due to incomplete measurements by devices, partially answered surveys, data deletion, or data loss during preprocessing. Missing data are frequently encountered in many scientific fields where data collection is difficult, such as health sciences, ecology, economics, and the social sciences. Missing data can disrupt the randomness of the sample, lead to biased estimates, and reduce the statistical power of analyses [1]. Moreover, a reduction in sample size increases standard error, compromising the accuracy of the analysis results. Therefore, it is essential to address the problem of missing data before starting the analysis.

The most comprehensive studies on missing data were conducted by Donald B. Rubin. In his study, classified missing data into two categories: ignorable and non-ignorable missing data [2]. Ignorable missing data are further divided into missing at random (MAR) and missing completely at random (MCAR), while non-ignorable missing values are defined as not missing at random (NMAR) [2,3]. MAR refers to cases where missingness depends on other observed variables rather than the

variable itself. In this scenario, statistical inferences can be made based only on observed data [3]. For the MAR condition to hold, related variables must be present and fully observed. MCAR occurs when missingness is entirely random and unrelated to any variable [3]. MCAR is a special case of MAR. If the type of missingness in a dataset is MCAR, it can also be classified as MAR [4]. If missing data occur systematically and follow a specific pattern rather than randomly, it is considered an NMAR situation. This typically results from design, implementation, or measurement errors during the research process. In NMAR cases, missingness is related either to the variable itself or to other variables not included in the dataset. NMAR mechanism cannot be ignored in studies. In the NMAR case, in order to deal with the problem of missingness, it is necessary to construct an appropriate model representing the missing values, which can often be challenging [5].

The literature includes various studies on modeling methods addressing the missing data problem. Schafer [6] proposed several algorithms for estimating parameters in contingency tables with missing data under log-linear models. Baker, Rosenberger and Dersimonian presented

closed-form estimators for log-linear models in two-dimensional contingency tables with missing data and discussed their effects on model fit [7]. Molenberghs et al. demonstrated that every NMAR model has an equally well-fitting MAR counterpart [8]. Kim, Park and Kim discussed MAR and NMAR models in multidimensional contingency tables with missing data [9]. Kim, Jeon, and Kim examined log-linear models applicable to two-dimensional contingency tables with MAR and NMAR missing mechanisms in one variable [10]. Ghosh and Vellaisamy explored closed-form estimation and boundary solutions for log-linear models in two- and three-dimensional contingency tables with missing data [11-14]. Popovich compared deletion and imputation methods for handling missing data [15]. In addition, recent studies have explored methods for handling missing data under MNAR assumptions using multiple imputation approaches. For example, Fischer, Little, and West incorporated response indicators into sequential imputation procedures to address MNAR mechanisms in predictive modeling settings [16].

To address the missing data problem, it is necessary to identify the cause of the missingness and quantify the amount of missing data [2]. Based on this information, the mechanism of missing data can be identified, and appropriate solutions can then be applied.

The objective of this study is to investigate log-linear models and closed-form estimates for handling missing data in two-dimensional contingency tables. Although extensive studies on missing data exist in the literature, the present study examines the missing data problem in

the context of categorical data. Specifically, it aims to guide researchers in determining the missing data mechanism in two-dimensional contingency tables and in applying appropriate methods to effectively address missing data, thereby enabling statistical inference from contingency tables containing missing values. Section 2 presents log-linear models and closed-form estimators for handling missing data in one or both variables in two-dimensional contingency tables. Section 3 provides an illustrative example based on the *Myocardial Infarction Complications* dataset, covering scenarios with missing data in a single variable and in both variables. In addition, the effect of varying missingness rates in the single-variable setting is examined. Finally, Section 4 provides the conclusions and discussion.

2. Methods

Consider two categorical variables, Y_1 and Y_2 , with R and C levels, respectively. Two scenarios are considered: missingness in a single variable (either the row or the column variable) and missingness in both variables [7].

2.1 Case of Missingness in a Single Variable

Suppose there are missing values in the Y_2 (column) variable. The missingness status can be represented by a new variable M , where $M = 1$ if Y_2 is observed and $M = 2$ if Y_2 is missing. In this case, the contingency table can be structured as shown in Table 1 [9].

Table 1. Two-dimensional contingency table with one missing variable

	M = 1			...	M = 2	
	Y ₂ = 1	Y ₂ = 2	Y ₂ = C		Missing	
Y ₁ = 1	y_{111}	y_{121}			y_{1C1}	y_{1+2}
Y ₁ = 2	y_{211}	y_{221}			y_{2C1}	y_{2+2}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Y ₁ = R	y_{R11}	y_{R21}			y_{RC1}	y_{R+2}

The definitions of the contingency table in Table 1 are as follows [9,14]:

For $i = 1, 2, \dots, R$, $j = 1, 2, \dots, C$ and $x = 1, 2$ the cell frequencies are denoted as $\{y_{ijx}\}$. The vector of observed frequencies is defined as $y_{obs} = (\{y_{ij1}\}, \{y_{i+2}\})$ where $\{y_{ij1}\}$ represents the fully observed frequencies, and $\{y_{i+2}\}$ corresponds to the values for cases where Y_2 is missing. The symbol "+" denotes the sum over the corresponding index due to the missingness in the Y_2 variable. The vector of sample proportions is denoted by $\pi = \{\pi_{ijx}\}$ and the vector of expected frequencies is denoted as $\mu = \{\mu_{ijx}\}$. The total sample size is given by $n = \sum_{i,j,x} y_{ijx}$.

For the Table 1, the log-linear model is defined for $i = 1, 2, \dots, R$, $j = 1, 2, \dots, C$ and $x = 1, 2$ as shown in Equation (1):

$$\log \mu_{ijx} = \lambda + \lambda_i^{Y_1} + \lambda_j^{Y_2} + \lambda_x^M + \lambda_{ij}^{Y_1 Y_2} + \lambda_{ix}^{Y_1 M} + \lambda_{jx}^{Y_2 M} \quad (1)$$

In this model, the parameters are subject to the usual sum-to-zero constraints. For example, $\sum_i \lambda_{Y_1 Y_2(i,j)} = \sum_j \lambda_{Y_1 Y_2(i,j)} = 0$.

To determine the missing data mechanism in the Y_2 variable, the odds are defined as shown in Equation (2):

$$b_{ij} = \frac{P(M = 2 | Y_1 = i, Y_2 = j)}{P(M = 1 | Y_1 = i, Y_2 = j)} = \frac{\pi_{ij2}}{\pi_{ij1}} = \frac{\mu_{ij2}}{\mu_{ij1}} \quad (2)$$

The expected frequencies μ_{ij2} can be calculated using the equation $\mu_{ij2} = b_{ij} \mu_{ij1}$ and they satisfy the condition $\sum_{ij} \mu_{ij1} (1 + b_{ij}) = n$. The joint probabilities π_{ij+} can be obtained from the equation $\pi_{ij+} = \frac{\mu_{ij1}(1+b_{ij})}{n}$, from which marginal probabilities can be derived.

Using Equation (1), the following expression for b_{ij} can be obtained:

$$b_{ij} = \exp[-2\{\lambda_M(1) + \lambda_{Y_1 M}(i, 1) + \lambda_{Y_2 M}(j, 1)\}] \quad (3)$$

The parameter b_{ij} can be examined under the following three cases corresponding to the missing data mechanisms [9,17].

- 1) $b_{ij} = b_j$ NMAR
- 2) $b_{ij} = b_i$ MAR

3) $b_{ij} = b_{..}$ MCAR

Treating Δ as a constant, the log-likelihood function of μ under Poisson sampling is given in Equation (4):

$$l(\mu; y_{g\ddot{o}z}) = \sum_{i,j} y_{ij1} \log \mu_{ij1} + \sum_i y_{i+2} \log \mu_{i+2} - \sum_{i,j,x} \mu_{ijx} + \Delta \tag{4}$$

In the saturated log-linear models used to describe the missing data mechanisms, the inclusion of parameters associated with the missingness indicators increases the total number of model parameters. In certain MAR and NMAR structures, the number of parameters may exceed the available degrees of freedom, which leads to an identifiability problem. In such situations, different parameter combinations may produce the same likelihood, and therefore unique maximum likelihood estimates cannot be obtained. This issue is commonly referred to in the literature as non-identifiability or under

identification [7]. To address this problem, the models can be reparametrized under appropriate constraints, and closed-form representations of the parameters can be derived. Consequently, the constrained structures proposed for the missing data mechanisms facilitate both the computation of parameter estimates and the identifiability of the models.

The closed-form estimates of the models for the three missing data mechanisms described above (NMAR, MAR, and MCAR) are provided in Table 2 [7,18].

Table 2. Proposed models and closed-form estimates for two-dimensional contingency tables with missing data in a single variable

Model	Closed-Form Estimates
1) $b_{ij} = b_j$ (Y_2 NMAR)	For $i \geq 1$ the parameter $\hat{\beta}_j$ satisfies $\sum_j \hat{\mu}_{ij1} \hat{\beta}_j = y_{i+2}$ $\hat{\mu}_{ij1} = y_{ij1}$
2) $b_{ij} = b_i$ (Y_2 MAR)	$\hat{\beta}_i = \frac{y_{i+2}}{y_{i+1}}$, $i = 1, 2, \dots, R$ $\hat{\mu}_{ij1} = \frac{y_{ij1} y_{i+2}}{y_{i+1} y_{i+2}} = y_{ij1}$
3) $b_{ij} = b_{..}$ (Y_2 MCAR)	$\hat{\beta}_{..} = \frac{y_{++2}}{y_{++1}}$ $\hat{\mu}_{ij1} = \frac{y_{ij1} y_{++2}}{y_{i+1} y_{++1}}$

The hypotheses for the three models described above are formulated as follows:

H_0 : The model fits the data.

H_1 : The model does not fit the data.

To test these hypotheses, the likelihood ratio statistic is defined in Equation (5) [7,18].

$$G^2 = -2 \left[\sum_{i,j} y_{ij1} \ln \left(\frac{\hat{\mu}_{ij1}}{y_{ij1}} \right) + \sum_i y_{i+2} \ln \left(\frac{\sum_j \hat{\mu}_{ij1} \hat{\beta}_j}{y_{i+2}} \right) - \sum_{i,j} \hat{\mu}_{ij1} (1 + \hat{\beta}_j) + n \right] \tag{5}$$

The degrees of freedom of the model are determined by the difference between the number of independent observed cell counts and the number of independently estimated parameters. In the case of missingness in a single variable, each row contains an additional “missing” category, resulting in a total of $R(C + 1)$ cells. Considering the model constraints and the number of estimated parameters p , the degrees of freedom can be expressed as:

$$df = R(C + 1) - p$$

This expression follows the general principle that the degrees of freedom correspond to the amount of independent information in the observed data after

accounting for the parameters estimated in the model. The G^2 statistic is compared with the critical value obtained from the $\chi^2_{(a,df)}$ distribution, a significance level of a . The null hypothesis is rejected if G^2 exceeds the corresponding critical value.

2.2 Case of Missingness in Two Variables

Suppose there are missing values in both the Y_1 (row) and Y_2 (column) variables. The missingness in Y_1 variable can be represented by a new variable M_1 , where $M_1 = 1$ if Y_1 is observed and $M_1 = 2$ if Y_1 is missing. Similarly, the missingness in the Y_2 variable can be represented by a new variable M_2 where $M_2 = 1$ if Y_2 is observed and $M_2 = 2$ if Y_2 is missing. In this case, the contingency table can be structured as shown in Table 3 [7].

Table 3. Two-dimensional contingency table with missing data in two variables

		M ₂ = 1			M ₂ = 2	
		Y ₂ = 1	Y ₂ = 2	...	Y ₂ = C	Missing
M ₁ = 1	Y ₁ = 1	y ₁₁₁	y ₁₂₁	...	y _{1C1}	y ₁₊₂
	Y ₁ = 2	y ₂₁₁	y ₂₂₁	...	y _{2C1}	y ₂₊₂
	⋮	⋮	⋮	⋮	⋮	⋮
	Y ₁ = R	y _{R11}	y _{R21}	...	y _{RC1}	y _{R+2}
M ₁ = 2	Missing	y ₊₁₂₁	y ₊₂₂₁	...	y _{+C21}	y ₊₊₂₂

The definitions for the contingency table provided in Table 3 are as follows [7,14]:

For $i = 1, 2, \dots, R, j = 1, 2, \dots, C$ and $x, s = 1, 2$ the cell frequencies are denoted by $\{y_{ijxs}\}$. The vector of observed frequencies is defined as $y_{obs} = (\{y_{ij11}\}, \{y_{i+12}\}, \{y_{+j21}\}, y_{++22})$, where $\{y_{ij11}\}$ represents the fully observed frequencies.

$\{y_{+j21}\}$: corresponds to the frequencies for cases where the Y₁ variable is missing.

$\{y_{i+12}\}$: corresponds to the frequencies for cases where the Y₂ variable is missing.

$\{y_{++22}\}$: corresponds to the frequencies for cases where both Y₁ and Y₂ variables are missing.

The symbol "+" denotes the sum over the corresponding index due to missingness in the variables.

The vector of sample proportions is denoted by $\pi = \{\pi_{ijxs}\}$ and the vector of expected frequencies is denoted by $\mu = \{\mu_{ijxs}\}$. The total sample size is represented by n .

For $i = 1, 2, \dots, R, j = 1, 2, \dots, C$ and $x, s = 1, 2$ the log-linear model for this contingency table is defined as shown in Equation (6):

$$\log \mu_{ijxs} = \lambda + \lambda_i^{Y_1} + \lambda_j^{Y_2} + \lambda_x^{M_1} + \lambda_s^{M_2} + \lambda_{ij}^{Y_1 Y_2} + \lambda_{ix}^{Y_1 M_1} + \lambda_{jx}^{Y_2 M_1} + \lambda_{is}^{Y_1 M_2} + \lambda_{js}^{Y_2 M_2} + \lambda_{xs}^{M_1 M_2} \tag{6}$$

In this model, all parameters are subject to the usual sum-to-zero constraints. For example

$$\sum_i \lambda_{Y_1 Y_2}(i, j) = \sum_j \lambda_{Y_1 Y_2}(i, j) = 0.$$

The odds for missing values in the Y₁ and Y₂ variables are denoted as α_{ij} and b_{ij} respectively. When Y₂ is observed the odds of missingness in the Y₁ variable can be calculated as shown in Equation (7):

$$\alpha_{ij} = \frac{P(M_1 = 2, M_2 = 1 | Y_1 = i, Y_2 = j)}{P(M_1 = 1, M_2 = 1 | Y_1 = i, Y_2 = j)} = \frac{\pi_{ij21}}{\pi_{ij11}} = \frac{\mu_{ij21}}{\mu_{ij11}} \tag{7}$$

When Y₁ is observed the odds of missingness in the Y₂ variable can be calculated as shown in Equation (8):

$$b_{ij} = \frac{P(M_1 = 1, M_2 = 2 | Y_1 = i, Y_2 = j)}{P(M_1 = 1, M_2 = 1 | Y_1 = i, Y_2 = j)} = \frac{\pi_{ij12}}{\pi_{ij11}} = \frac{\mu_{ij12}}{\mu_{ij11}} \tag{8}$$

The expected frequencies satisfy the relation $\mu_{ij11} = n\pi_{ij11}$. The odds ratio (θ) between the missingness indicators M₁ and M₂ can be obtained from Equation (9).

$$\theta = \frac{P(M_1 = 1, M_2 = 1 | Y_1 = i, Y_2 = j)P(M_1 = 2, M_2 = 2 | Y_1 = i, Y_2 = j)}{P(M_1 = 1, M_2 = 2 | Y_1 = i, Y_2 = j)P(M_1 = 2, M_2 = 1 | Y_1 = i, Y_2 = j)} = \frac{\pi_{ij11}\pi_{ij22}}{\pi_{ij12}\pi_{ij21}} = \frac{\mu_{ij11}\mu_{ij22}}{\mu_{ij12}\mu_{ij21}} \tag{9}$$

If $\theta = 1$, the missingness indicators M₁ and M₂ are independent.

The expected frequencies can be expressed as $\mu_{ij21} = \alpha_{ij}\mu_{ij11}$, $\mu_{ij12} = b_{ij}\mu_{ij11}$ and $\mu_{ij22} = \mu_{ij11}\alpha_{ij}b_{ij}\theta$. Using these expressions, the total sample size can be written as $n = \sum_{i,j} \mu_{ij11}(1 + \alpha_{ij} + b_{ij} + \alpha_{ij}b_{ij}\theta)$.

The joint probability can therefore be written as $\pi_{ij++} = \frac{\mu_{ij11}(1 + \alpha_{ij} + b_{ij} + \alpha_{ij}b_{ij}\theta)}{n}$, from which marginal probabilities can be derived.

The probability of missingness in Y₁ while Y₂ is observed can be calculated using Equation (10), and the probability of missingness in Y₂ while Y₁ is observed can be calculated using Equation (11).

$$\phi_{1|2}(i, j) = P(M_1 = 2, M_2 = 1 | Y_1 = i, Y_2 = j) = \frac{\alpha_{ij}}{1 + \alpha_{ij}} \tag{10}$$

$$\phi_{2|1}(i, j) = P(M_1 = 2, M_2 = 1 | Y_1 = i, Y_2 = j) = \frac{b_{ij}}{1 + b_{ij}} \tag{11}$$

Using the log-linear model given in Equation (6), the expressions for α_{ij} and b_{ij} can be derived as shown in Equations (12) and (13).

$$\alpha_{ij} = \exp[-2\{\lambda_{M_1}(1) + \lambda_{Y_1M_1}(i, 1) + \lambda_{Y_2M_1}(j, 1) + \lambda_{M_1M_2}(1,1)\}] \tag{12}$$

$$b_{ij} = \exp[-2\{\lambda_{M_2}(1) + \lambda_{Y_1M_2}(i, 1) + \lambda_{Y_2M_2}(j, 1) + \lambda_{M_1M_2}(1,1)\}] \tag{13}$$

The odds ratio (θ) between the missingness indicators can be obtained as $\theta = \exp[4\lambda_{M_1M_2}(1,1)]$.

If α_{ij} and b_{ij} depend only on one of the indices i or j , or on neither of them, they are represented as $\alpha_{ij} \in \{\alpha_i, \alpha_j, \alpha_{..}\}$ and $b_{ij} \in \{\beta_i, \beta_j, \beta_{..}\}$.

The parameters α_{ij} and b_{ij} can be analyzed under the following cases according to the missing data mechanisms [7]:

- 1) $\alpha_{ij} = \alpha_i$: Missingness in Y_1 is NMAR
- 2) $\alpha_{ij} = \alpha_j$: Missingness in Y_1 is MAR
- 3) $\alpha_{ij} = \alpha_{..}$: Missingness in Y_1 is MCAR
- 4) $b_{ij} = \beta_i$: Missingness in Y_2 is MAR
- 5) $b_{ij} = \beta_j$: Missingness in Y_2 is NMAR
- 6) $b_{ij} = \beta_{..}$: Missingness in Y_2 is MCAR

Treating Δ as a constant independent of μ_{ij} the log-likelihood function of μ under Poisson sampling is given in Equation (14).

$$l(\mu; y_{g\ddot{o}z}) = \sum_{i,j} y_{ij11} \log \mu_{ij11} + \sum_i y_{i+12} \log \mu_{i+12} + \sum_j y_{+j21} \log \mu_{+j21} + y_{++22} \log \mu_{++22} - \mu_{++++} + \Delta \tag{14}$$

Based on the six scenarios defined above, eight models can be specified. The closed-form estimates for these models are presented in Table 4 [7].

The hypotheses for the eight models presented in Table 4 are formulated as follows:

- H_0 : The model fits the data.
- H_1 : The model does not fit the data.

To test these hypotheses, the likelihood ratio statistic is given in Equation (15) [7].

$$G^2 = -2 \left[\sum_{i,j} y_{ij11} \ln \left(\frac{\hat{\mu}_{ij11}}{y_{ij11}} \right) + \sum_i y_{i+12} \ln \left(\frac{\sum_j \hat{\mu}_{ij11} \hat{b}_{ij}}{y_{i+12}} \right) + \sum_j y_{+j21} \ln \left(\frac{\sum_i \hat{\mu}_{ij11} \hat{\alpha}_{ij}}{y_{+j21}} \right) + y_{++22} \ln \left(\frac{\sum_{i,j} \hat{\mu}_{ij11} \hat{\alpha}_{ij} \hat{b}_{ij} \hat{\theta}}{y_{++22}} \right) - \sum_{i,j} \hat{\mu}_{ij11} (1 + \hat{\alpha}_{ij} + \hat{b}_{ij} + \hat{\alpha}_{ij} \hat{b}_{ij} \hat{\theta}) + n \right] \tag{15}$$

The degrees of freedom for these eight models are calculated as:

$$df = (R + 1)(C + 1) - p$$

where p denotes the number of independently estimated parameters in the model. The G^2 statistic is compared with the critical value obtained from the $\chi^2_{(a;df)}$ distribution, a significance level of a . The null hypothesis is rejected if G^2 exceeds the corresponding critical value.

Table 4. Proposed models and closed-form estimates for two-dimensional contingency tables with missing data in two variables

Model		Closed-Form Estimators	
M1	$\alpha_{..}, \beta_{i.}$ (Y ₁ MCAR and Y ₂ MAR)	$\hat{\alpha}_{..} = \frac{y_{++21}}{y_{++11}}$	$\hat{\beta}_{i.} = \frac{y_{i+12}}{\hat{\mu}_{i+}}$, $i = 1, 2, \dots, R$
		$\hat{\theta} = \frac{y_{++11}y_{++22}}{y_{++12}y_{++11}}$	$\hat{\mu}_{ij11} = \frac{y_{ij11}y_{+j+1}y_{++11}}{y_{+j11}y_{++11}}$
M2	$\alpha_{..}, \beta_{.j}$ (Y ₁ MCAR and Y ₂ NMAR)	$\hat{\alpha}_{..} = \frac{y_{++21}}{y_{++11}}$	$\hat{\beta}_{.j}$ satisfies $\sum_j \hat{\mu}_{ij11} \hat{\beta}_{.j} = y_{i+12}$
		$\hat{\theta} = \frac{y_{++11}y_{++22}}{y_{++12}y_{++11}}$	$\hat{\mu}_{ij11} = \frac{y_{ij11}y_{+j+1}y_{++11}}{y_{+j11}y_{++11}}$
M3	$\alpha_{i.}, \beta_{..}$ (Y ₁ NMAR and Y ₂ MCAR)	$\hat{\alpha}_{i.}$ satisfies $\sum_i \hat{\mu}_{ij11} \hat{\alpha}_{i.} = y_{+j21}$	$\hat{\beta}_{..} = \frac{y_{++12}}{y_{++11}}$
		$\hat{\theta} = \frac{y_{++11}y_{++22}}{y_{++12}y_{++11}}$	$\hat{\mu}_{ij11} = \frac{y_{ij11}y_{i+1}y_{++11}}{y_{i+11}y_{++11}}$
M4	$\alpha_{i.}, \beta_{i.}$ (Y ₁ NMAR and Y ₂ MAR)	$\hat{\alpha}_{i.}$ satisfies $\sum_i \hat{\mu}_{ij11} \hat{\alpha}_{i.} = y_{+j21}$	$\hat{\beta}_{i.} = \frac{y_{i+12}}{y_{i+11}}$, $i = 1, 2, \dots, R$
		$\hat{\theta} = \frac{y_{++22}}{\sum_i y_{i+} \hat{\alpha}_{i.} \beta_{i.}}$	$\hat{\mu}_{ij11} = y_{ij11}$
M5	$\alpha_{i.}, \beta_{.j}$ (Y ₁ and Y ₂ NMAR)	$\hat{\alpha}_{i.}$ satisfies $\sum_i \hat{\mu}_{ij11} \hat{\alpha}_{i.} = y_{+j21}$	$\hat{\beta}_{.j}$ satisfies $\sum_j \hat{\mu}_{ij11} \hat{\beta}_{.j} = y_{i+12}$
		$\hat{\theta} = \frac{y_{++22}}{\sum_{i,j} y_{ij} \hat{\alpha}_{i.} \hat{\beta}_{.j}}$	$\hat{\mu}_{ij11} = y_{ij11}$
M6	$\alpha_{.j}, \beta_{..}$ (Y ₁ MAR and Y ₂ MCAR)	$\hat{\alpha}_{.j} = \frac{y_{+j21}}{\hat{\mu}_{+j}}$, $j = 1, 2, \dots, C$	$\hat{\beta}_{..} = \frac{y_{++12}}{y_{++11}}$
		$\hat{\theta} = \frac{y_{++11}y_{++22}}{y_{++12}y_{++11}}$	$\hat{\mu}_{ij11} = \frac{y_{ij11}y_{i+1}y_{++11}}{y_{i+11}y_{++11}}$
M7	$\alpha_{.j}, \beta_{i.}$ (Y ₁ and Y ₂ MAR)	$\hat{\alpha}_{.j} = \frac{y_{+j21}}{\hat{\mu}_{+j11}}$, $j = 1, 2, \dots, C$	$\hat{\beta}_{i.} = \frac{y_{i+12}}{y_{i+11}}$, $i = 1, 2, \dots, R$
		$\hat{\theta} = \frac{y_{++22}}{\sum_{i,j} y_{ij+11} \hat{\alpha}_{.j} \hat{\beta}_{i.}}$	$\hat{\mu}_{ij11} = y_{ij11}$
M8	$\alpha_{.j}, \beta_{.j}$ (Y ₁ MAR and Y ₂ NMAR)	$\hat{\alpha}_{.j} = \frac{y_{+j21}}{\hat{\mu}_{+j11}}$, $j = 1, 2, \dots, C$	$\hat{\beta}_{.j}$ satisfies $\sum_j \hat{\mu}_{ij11} \hat{\beta}_{.j} = y_{i+12}$
		$\hat{\theta} = \frac{y_{++22}}{\sum_j y_{+j} \hat{\alpha}_{.j} \hat{\beta}_{.j}}$	$\hat{\mu}_{ij11} = y_{ij11}$

3. Illustrative Examples

In this section, the log-linear models introduced in Section 2 for analyzing two-dimensional contingency tables with missing values are illustrated using a real dataset. In health studies, missing data frequently arise for various reasons, such as patients' unwillingness to share information, incomplete reporting, failure to attend follow-up appointments, or interruptions in treatment or monitoring. For this study, a health-related dataset was chosen to illustrate the methodology.

The *Myocardial Infarction Complications* dataset, which includes missing values in different variables, was obtained from the University of California, Irvine (UCI) Machine Learning Repository [19]. The dataset contains information from 1700 patients. Using the variables myocardial infarction and hypertension, a contingency table was constructed and is presented in Table 5. It was observed that some patients had missing values for myocardial infarction, some for hypertension, and others for both variables.

Table 5. Myocardial Infarction x Hypertension table

Myocardial Infarction	Hypertension		
	Present	Absent	Missing
Present	446	187	3
Absent	640	416	4
Missing	0	2	2

3.1. Case of Missingness in a Single Variable

In the Myocardial Infarction × Hypertension contingency table, the case where missingness occurs only in the hypertension variable is considered. The three models proposed for two-dimensional contingency tables with missing values in a single variable (see Table 2) were applied to the data, and the results are summarized in Table 6.

Table 6. Model results for the Myocardial Infarction x Hypertension table with missing values in the hypertension variable

Model	Parameter Estimates	G ²	df	p
(b _{i.})	$\hat{\beta}_{1.} = 0.0047$ $\hat{\beta}_{2.} = 0.0038$	0	0	-
(b _{.j})	$\hat{\beta}_{.1} = 0.0076$ $\hat{\beta}_{.2} = -0.0021$	0.0003	0	-
(b _{..})	$\hat{\beta}_{..} = 0.0041$	0.0850	1	0.770

The (b_{..}) (Y₁ MCAR) model presented in Table 6 was found to fit the data ($G^2 = 0.085$; $df = 1$; $p = 0.770$). According to this model, the missingness in the hypertension is completely independent. The expected frequencies for the (b_{..}) model are presented in Table 7.

Table 7. Expected frequencies for the (b_{..}) model

Myocardial Infarction	Hypertension			
	Observed		Missing	
	Present	Absent	Present	Absent
Present	446.2642	187.1108	1.8495	0.7754
Absent	640.3791	416.2464	2.6540	1.7251

The expected frequencies for the final Myocardial Infarction x Hypertension table is presented in Table 8.

Table 8. The expected frequencies for the final Myocardial Infarction x Hypertension table

Myocardial Infarction	Hypertension	
	Present	Absent
Present	448.1137	187.8862
Absent	643.0331	417.9715

The odds ratio derived from the expected frequencies is:

$$\hat{\theta} = \frac{448.1137 \times 417.9715}{187.8862 \times 643.0331} = 1.55$$

Indicating that the odds of myocardial infarction are approximately 1.55 times higher among individuals with hypertension compared to those without hypertension.

3.2. Case of Missingness in Two Variables

In the Myocardial Infarction × Hypertension contingency table, the structure that accounts for missing values in both variables is considered. To compute the G² statistics for the models, the cells corresponding to cases without missing values were coded as "2" following the approach in [14]. The eight models proposed for two-dimensional contingency tables with missing values in both variables (see Table 4) were applied to the data presented in Table 5, and the results are summarized in Table 9.

Table 9. Model results for the Myocardial Infarction x Hypertension table with missing values in both variables

	Model	Parameter Estimates	G ²	df	p	AIC	BIC
M1	($\alpha_{..}, \beta_{i.}$)	$\hat{\alpha}_{..} = 0.0024$ $\hat{\beta}_{1.} = 0.0047$ $\hat{\beta}_{2.} = 0.0038$ $\hat{\theta} = 120.6429$	0.340	1	0.560	-1.660	-7.100
M2	($\alpha_{..}, \beta_{.j}$)	$\hat{\alpha}_{..} = 0.0024$ $\hat{\beta}_{.1} = 0.0076$ $\hat{\beta}_{.2} = -0.0021$ $\hat{\theta} = 120.6429$	0.340	1	0.560	-1.660	-7.100
M3	($\alpha_{i.}, \beta_{..}$)	$\hat{\alpha}_{1.} = -0.0068$ $\hat{\alpha}_{2.} = 0.0079$ $\hat{\beta}_{..} = 0.0041$ $\hat{\theta} = 120.6429$	2.094	1	0.148	0.094	-5.346
M4	($\alpha_{.j}, \beta_{..}$)	$\hat{\alpha}_{.1} = 0.0018$ $\hat{\alpha}_{.2} = 0.0033$ $\hat{\beta}_{..} = 0.0041$ $\hat{\theta} = 120.6429$	0.085	1	0.771	-1.915	-7.355
M5	($\alpha_{i.}, \beta_{i.}$)	$\hat{\alpha}_{1.} = 0.0079$ $\hat{\alpha}_{2.} = -0.0068$ $\hat{\beta}_{1.} = 0.0038$ $\hat{\beta}_{2.} = 0.0047$ $\hat{\theta} = 178.5714$	0	0	-		
M6	($\alpha_{.j}, \beta_{.j}$)	$\hat{\alpha}_{.1} = 0.0033$ $\hat{\alpha}_{.2} = 0.0018$ $\hat{\beta}_{.1} = -0.0021$ $\hat{\beta}_{.2} = 0.0076$ $\hat{\theta} = 181.8182$	0	0	-		
M7	($\alpha_{i.}, \beta_{.j}$)	$\hat{\alpha}_{1.} = 0.0079$ $\hat{\alpha}_{2.} = -0.0068$ $\hat{\beta}_{.1} = -0.0021$ $\hat{\beta}_{.2} = 0.0076$ $\hat{\theta} = 179.4489$	0	0	-		
M8	($\alpha_{.j}, \beta_{i.}$)	$\hat{\alpha}_{.1} = 0.0033$ $\hat{\alpha}_{.2} = 0.0018$ $\hat{\beta}_{1.} = 0.0038$ $\hat{\beta}_{2.} = 0.0047$ $\hat{\theta} = 121.0424$	0	0	-		

The $(\alpha_j, \beta_{..})$ (Y_1 MAR, Y_2 MCAR) model, which has the smallest *AIC* and *BIC* values, is identified as the best-fitting model ($G^2 = 0.085$; $df = 1$; $p = 0.771$). According to this model, the missingness in the myocardial infarction variable depends on the hypertension status, whereas the missingness in the hypertension variable is completely independent. The expected frequencies for the best-fitting model are presented in Table 10.

Table 10. Expected frequencies for the $(\alpha_j, \beta_{..})$ model

Myocardial Infarction		Hypertension			
		Observed		Missing	
		Present	Absent	Present	Absent
Observed	Present	446.2642	187.1107	1.8495	0.7755
	Absent	639.7727	415.8522	2.6515	1.7235
Missing	Present	0.8218	0.6206	0.4109	0.3103
	Absent	1.1782	1.3794	0.5891	0.6897

Expected frequencies for the final Myocardial Infarction x Hypertension table, considering missing values in both variables, is presented in Table 11.

Table 11. The expected frequencies of final Myocardial Infarction x Hypertension table with missing values in both variables

Myocardial Infarction		Hypertension	
		Present	Absent
Present		449.3465	188.8172
Absent		644.1915	419.6448

The odds ratio computed from the expected frequencies is

$$\hat{\theta} = \frac{449.3465 \times 419.6448}{188.8172 \times 644.1915} = 1.55$$

This result indicates that the odds of myocardial infarction are approximately 1.55 times higher among individuals with hypertension than among those without hypertension.

3.3. Comparison of Different Missing Value Rates

When missing values were ignored, the estimated odds ratio was identical to the odds ratio obtained from the log-linear model. This may be due to the relatively small number of missing values compared to the total number of observations in the dataset. To investigate whether increasing the proportion of missing values would affect the results, hypothetical datasets were generated with missing value percentages of 5%, 10%, 15%, and 20% in the hypertension variable of the Myocardial Infarction x Hypertension contingency table, and the models were analyzed. The hypothetical datasets were summarized in Table 12.

Table 12. Myocardial Infarction x Hypertension table with different percentages of missing values in the hypertension variable

Percentage of Missing Values	Myocardial Infarction	Hypertension		
		Present	Absent	Missing
5%	Present	446	187	38
	Absent	640	416	51
10%	Present	446	187	81
	Absent	640	416	107
15%	Present	446	187	128
	Absent	640	416	170
20%	Present	446	187	181
	Absent	640	416	241

The three models recommended for cases with missing values only in the column variable in two-dimensional tables (see Table 2) were applied to the datasets presented in Table 12, and the model results for each percentage of missing values are summarized in Table 13.

For the tables with 5% and 10% of missing values, the $(b_{..})$ (Y_1 MCAR) model was found to fit the data ($p = 0.293; p = 0.089$). According to these models, the missingness mechanism in the hypertension variable can be considered completely independent. For the tables with 15% and 20% of missing values, none of the three

models provided an adequate fit to the data. The deterioration in model fit as the proportion of missing values increases can be explained by the reduction of available information in the observed data. As the number of missing observations grows, the direct information about the joint distribution of variables decreases, leading to greater uncertainty in parameter estimation. Consequently, the expected cell frequencies deviate more from the observed counts, which results in larger values of the G^2 statistic and poorer model fit [18].

Table 13. Model results for the Myocardial Infarction x Hypertension table under different percentages of missing values in the hypertension variable

Percentage of Missing Values	Model	Parameter Estimates	G^2	df	p
5%	$(b_{i.})$	$\hat{\beta}_{1.} = 0.0600$ $\hat{\beta}_{2.} = 0.0483$	0	0	-
	$(b_{.j})$	$\hat{\beta}_{.1} = 0.0952$ $\hat{\beta}_{.2} = -0.0024$	1.806	0	-
	$(b_{..})$	$\hat{\beta}_{..} = 0.0527$	1.103	1	0.293
10%	$(b_{i.})$	$\hat{\beta}_{1.} = 0.1279$ $\hat{\beta}_{2.} = 0.1013$	0	0	-
	$(b_{.j})$	$\hat{\beta}_{.1} = 0.2078$ $\hat{\beta}_{.2} = -0.0625$	0	0	-
	$(b_{..})$	$\hat{\beta}_{..} = 0.1113$	2.896	1	0.089
15%	$(b_{i.})$	$\hat{\beta}_{1.} = 0.2022$ $\hat{\beta}_{2.} = 0.1609$	0	0	-
	$(b_{.j})$	$\hat{\beta}_{.1} = 0.3258$ $\hat{\beta}_{.2} = -0.0926$	0	0	-
	$(b_{..})$	$\hat{\beta}_{..} = 0.1764$	4.688	1	0.030
20%	$(b_{i.})$	$\hat{\beta}_{1.} = 0.2859$ $\hat{\beta}_{2.} = 0.2282$	0	0	-
	$(b_{.j})$	$\hat{\beta}_{.1} = 0.4590$ $\hat{\beta}_{.2} = -0.1268$	0	0	-
	$(b_{..})$	$\hat{\beta}_{..} = 0.2498$	6.915	1	0.009

To examine how increasing the proportion of missing values affects the distribution of cell counts in the contingency table, the expected cell frequencies were computed under the $(b_{..})$ (Y_1 MCAR) model for each hypothetical dataset. The model redistributes the missing observations across the observed categories according to the estimated parameters. The resulting expected frequencies for both observed and missing categories are presented in Table 14.

Table 14. Expected frequencies under the $(b_{..})$ model for different percentages of missing values in the hypertension variable

Percentage of Missing Values	Myocardial Infarction	Hypertension			
		Observed		Missing	
		Present	Absent	Present	Absent
5%	Present	449.1088	188.3034	23.6653	9.92244
	Absent	644.4610	418.8997	33.9591	22.0734
10%	Present	452.6835	189.8023	50.3875	21.1266
	Absent	649.5907	422.2340	72.3049	46.9982
15%	Present	455.7719	191.0972	80.4144	33.7163
	Absent	654.0225	425.1146	115.3929	75.0054
20%	Present	458.8777	192.3994	114.6515	48.0713
	Absent	658.4792	428.0115	164.5223	106.9395

After allocating the missing observations according to the estimated parameters of the ($b_{..}$) model, the completed Myocardial Infarction × Hypertension contingency tables were obtained. The expected frequencies of these completed tables for different missing value rates are presented in Table 15.

Table 15. The expected frequencies of the final Myocardial Infarction x Hypertension table obtained under the ($b_{..}$) model for different missing value percentages

Percentage of Missing Values	Myocardial Infarction	Hypertension		$\hat{\theta}$
		Present	Absent	
5%	Present	472.7741	198.2259	1.55
	Absent	678.4202	440.9731	
10%	Present	503.0711	210.9289	1.55
	Absent	721.8957	469.2322	
15%	Present	536.1864	224.8135	1.55
	Absent	769.4154	500.1201	
20%	Present	573.5292	240.4707	1.55
	Absent	823.0015	534.9510	

For each percentage of missing values, the odds ratio computed from the expected frequencies remains 1.55. These results indicate that the odds of myocardial infarction are approximately 1.55 times higher among individuals with hypertension compared to those without hypertension.

4. Conclusion and Discussion

In this study, the log-linear modeling approach was examined for analyzing two-dimensional contingency tables with missing values. The models were applied to the *Myocardial Infarction Complications* dataset, and the best-fitting models were determined using information criteria, and the corresponding expected frequencies were calculated.

In the two-dimensional contingency table created as Myocardial Infarctions × Hypertension, where missing values occurred only in the hypertension variable, the model that fit the data indicated that the missingness in the hypertension variable is completely independent (MCAR). When the missingness was MAR or NMAR, it was not possible to test model fit because the number of parameters in the model equaled the number of cells, resulting in zero degrees of freedom. Therefore, determining the most appropriate model among the three types relies on the researcher's expertise in identifying the type of missingness.

In the case where both the myocardial infarctions and hypertension variables had missing values, the best-fitting model indicated that the missingness in the myocardial infarction variable depended on the hypertension variable (MAR), while the missingness in the hypertension variable was completely independent (MCAR). In both cases, the model in which the missingness in the hypertension variable was completely independent emerged as the best-fitting model, and the odds ratios derived from the summary tables yielded the same results.

To investigate the effect of the missing value rate on model results, hypothetical tables were created by keeping the observed frequencies in the Myocardial Infarctions x Hypertension table constant and varying the missing value rate in the hypertension variable to 5%,

10%, 15%, and 20%. For the contingency tables with 5% and 10% missing values, the model assuming that the missingness in the hypertension variable is completely independent (MCAR) was found to fit the data. However, when the missing value rates increased to 15% and 20%, this model no longer fit the data. The model fit results for the ($b_{..}$) model across different missing value rates are summarized in Table 16. It can be concluded that as the missing value rate increases, the model fit decreases, and for missing value rates above 15%, the model fit disappears.

Table 16. Comparison of model fit for the ($b_{..}$) model across different missing value rates

Missing Value Rate	G ²	p
%5	1.103	0.293
%10	2.896	0.089
%15	4.688	0.030
%20	6.915	0.009

Although the model fits may vary, the odds ratios calculated under the model yield the same results as the odds ratios calculated using observed frequencies in all cases. This result can be explained by the missingness mechanism corresponds to MCAR, which represents ignorable missing data.

The models used for cases with missingness in one or two variables are based on the types of missing values. If the researcher can independently determine the type of missingness, they can directly apply the model that fits the table structure and the type of missingness, rather than applying all models. If the type of missingness cannot be

determined, as in this study, the researcher can apply models suitable for the table structure and use information criteria to select the best-fitting model among those that fit. The structure of the best-fitting model will also provide the researcher with insights into the type of missingness.

Despite its contributions, this study has several limitations. First, the analysis is restricted to two-dimensional contingency tables, and the direct extension of the proposed closed-form estimates to higher-dimensional tables may not always be straightforward. Second, the models rely on structural assumptions regarding the missing data mechanism (MCAR, MAR, NMAR), which cannot be definitively verified in real datasets. In addition, higher proportions of missing data may increase the uncertainty of parameter estimates and weaken model fit. Future research may extend the analysis to higher-dimensional contingency tables and compare the proposed approach with alternative estimation methods such as the EM algorithm or Bayesian techniques.

As a continuation of this study, a comprehensive simulation study could be conducted, where tables with different structures containing missing values are generated, and model fit is examined.

Conflict of Interest

There are no conflicts of interest in this work.

Acknowledgments

This research is a part of Emine Öçal's master's thesis on "Log-Linear Models for Contingency Tables with Missing Data".

References

- [1] Peng, C. Y., Harwell, M., Liou, S. M., & Ehman, L. H. (2006). Advances in missing data methods and implications for educational research. *Real Data Analysis*, 3178, 102. <https://api.semanticscholar.org/CorpusID:14341113>
- [2] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592. <https://doi.org/10.1093/biomet/63.3.581>
- [3] Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. John Wiley & Sons.
- [4] Allison, P. D. (2001). Missing data. In *Quantitative applications in the social sciences* (pp. 72–89). SAGE.
- [5] Howell, D. C. (2007). The treatment of missing data. In *The Sage handbook of social science methodology* (pp. 208–224).
- [6] Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC Press.
- [7] Baker, S. G., Rosenberger, W. F., & Dersimonian, R. (1992). Closed-form estimates for missing counts in two-way contingency tables. *Statistics in Medicine*, 11(5), 643–657. <https://doi.org/10.1002/sim.4780110509>
- [8] Molenberghs, G., Beunckens, C., Sotito, C., & Kenward, M. G. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society: Series B*, 70(2), 371–388. <https://doi.org/10.1111/j.1467-9868.2007.00640.x>
- [9] Kim, S., Park, Y., & Kim, D. (2015). On missing-at-random mechanism in two-way incomplete contingency tables. *Statistics & Probability Letters*, 96, 196–203. <https://doi.org/10.1016/j.spl.2014.09.016>
- [10] Kim, S., Jeon, S., & Kim, D. (2020). On log-linear modeling for an incomplete two-way contingency table with one variable subject to nonresponse. *Communications in Statistics—Simulation and Computation*, 49(4), 973–988. <https://doi.org/10.1080/03610918.2018.1441415>
- [11] Ghosh, S., & Vellaisamy, P. (2016). On the occurrence of boundary solutions in multidimensional incomplete tables. *Statistics & Probability Letters*, 119, 63–75. <https://doi.org/10.1016/j.spl.2016.07.015>
- [12] Ghosh, S., & Vellaisamy, P. (2019). Evaluation of missing data mechanisms in two- and three-dimensional incomplete tables. *Journal of the Korean Statistical Society*, 48(2), 297–313. <https://doi.org/10.1016/j.jkss.2018.09.002>
- [13] Ghosh, S., & Vellaisamy, P. (2020). On the occurrence of boundary solutions in two-way incomplete tables. *REVSTAT—Statistical Journal*, 18(1), 89–108.
- [14] Ghosh, S., & Vellaisamy, P. (2024). Closed-form estimates for missing counts in multidimensional incomplete tables. *Hacettepe Journal of Mathematics and Statistics*, 53(3), 803–822. <https://doi.org/10.15672/hujms.1216921>
- [15] Popovic, D. (2025). How to treat missing data in survey research. *Journal of Marketing Theory and Practice*, 33(1), 43–59.
- [16] Fischer, M., Little, R. J., & West, B. T. (2026). Multiple imputation under missing not at random: Incorporating response indicators into sequential imputation. *Journal of Statistical Computation and Simulation*, 96(1), 232–256. <https://doi.org/10.1080/00949655.2025.2558859>
- [17] Baker, S. G., & Laird, N. M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, 83(401), 62–69. <https://doi.org/10.1080/01621459.1988.10478568>
- [18] Öçal, E. (2024). Log-linear models for contingency tables with missing data (Master's thesis). Hacettepe University. <https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.js>
- [19] Golovenkin, S. E., Shulman, V. A., Rossiev, D. A., Shesternya, P. A., Nikulina, S. Y., Orlova, Y. V., & Voyno-Yasenetsky, V. F. (2020). Myocardial infarction complications. <https://doi.org/10.24432/C53P5M>