Contemporary Mathematics
http://ojs.wiserpub.com/index.php/CM/

Research Article

Impact of Imputation on Performance of Goodness-of-Fit Tests for the Logistic Panel Data Model

Opeyo Peter Otieno[1,2*], Cheng Weihu[1], Randa A. Makled[3]

[1] School of Mathematics, Statistics and Mechanics, Beijing University of Technology, Beijing 100124, China
[2] Department of Statistics and Computational Mathematics, The Technical University of Kenya, Nairobi, Kenya
[3] Faculty of Commerce, Damietta University, Damietta 34511, Egypt
E-mail: opeyopeter@gmail.com

Received: 12 March 2024; **Revised:** 15 April 2024; **Accepted:** 17 April 2024

Abstract: Goodness-of-fit tests aim at discerning model misspecification and identifying a model which is poorly fitting a given data set. They are methods used to determine the suitability of the fitted model. The subject of assessment of goodness-of-fit in logistic regression model has attracted the attention of many scientists and researchers. Several methods for assessing how well observed data can fit into logistic regression models have been proposed and discussed where test statistics are functions of the observed data values and their corresponding estimated values after parameter estimation. Considering a correctly specified panel data model with balanced data set, the conditional maximum likelihood estimates of the parameters are less biased and the estimated response variable values are actually in the neighborhood of the observed values. Relative to the induced biases of the parameter estimates resulting from imputation of missing covariates, the performances of the goodness-of-fit tests may be misjudged. This study looks at the susceptibility of the goodness-of-fit tests for logistic panel data models with imputed covariates. Simulation results show that Bayesian imputation impacts less on the goodness-of-fit test statistics and therefore stands out as the better technique against other classical imputation methods. An increased proportion of missingness however appeared to reduce the confidence interval of the test statistics which in turn reduces the chances of adopting the model under study.

***Keywords*:** goodness-of-fit, imputation, conditional maximum likelihood estimator, logistic panel data, Bayesian, Monte Carlo, covariate pattern

MSC: 62F05, 62J20

1. Introduction

For the study of binary data, the logistic regression model generally links the probability of response y to a collection of covariates, X and due to the its popularity, there is a temptation to regularly apply it to binary data without evaluating the model's fit, especially when the initial data set was incomplete.

One general problem for logistic regression models even for complete data sets is the low power of overall goodness-of-fit tests [1]. Such known problem may be aggravated with the fact that missing data points are always encountered which compromise on the models parameter estimates. Most developed goodness-of-fit tests are based on an association

DOI: https://doi.org/10.37256/cm.5420244585

between actual and estimated response values in groups of observations defined by the estimated response probability. Any deviation on the values of the response may yield different results. For specific data sets with missing covariates, such deviations are escalated by the injected bias resulting from various attempts to clean up and complete the data [1].

Recent application studies on statistical methods and models related to goodness-of-fit testing have been conducted. Zamanzade and Mahdizadeh offered a new approach to assess the fit of statistical models using entropy measures derived from ranked set samples and presented a method for entropy estimation from ranked set samples, with an application to goodness-of-fit testing [2]. In their other study, Mahdizadeh and Zamanzade offer goodness-of-fit tests for assessing the fit of data to a Rayleigh distribution using Phi-divergence and they introduce novel statistical tests based on Phi-divergence measures to evaluate the fit of observed data to the Rayleigh distribution [3]. Zamanzade additionally proposes new statistical tests based on empirical distribution functions for evaluating the exponentiality assumption in pair ranked set sampling [4]. Further interest in model diagnosis is seen in studies by Geng et al. and Pho who proposed different goodness-of-fit tests for a parametric mixture cure model with partly interval-censored data and zero-inflated Bernoulli regression models respectively [5, 6]. These studies however assumed applications to complete data sets.

Other studies to compare several goodness of fit tests have assessed the behavior of asymptotic distribution of the test statistics. By simulation, Badi [7] found that correct model specification yielded reasonable power for all test methods while Hosmer-Lemeshow test produced slightly larger variance. Models with missing covariate were also observed to yield smaller variances [7].

Conditional maximum likelihood estimators have been proposed as most efficient for logistic panel data models and they still perform better for unbalanced panels with missing covariates [8]. Imputation, however, induces a bias in the estimates. Relative to these induced biases of the parameter estimates resulting from imputation of missing covariates [8], the performances of the goodness-of-fit tests may be misjudged. This can endorse wrong models and eventual wrong policy actions. We therefore need to look at the susceptibility of the different goodness-of-fit tests for logistic panel data regression models with imputed covariates.

The objective of this present study is to compare the performances of available model diagnostic techniques in cases of imputed covariates for the conditional MLE of binary response panel data regression model. In this section, we begin by highlighting the rationale of using conditional MLE for logistic panel data models and also introduce the literature of model diagnostics. Section 2 outlines the renowned methods for developing the goodness-of-test statistics for logistic regression models and blend them with the considered estimates of the response variables from imputed covariates. To explore the impact of these imputations on the test statistics of the goodness-of-fit, we perform Monte Carlo simulations in section 3 with a hypothetical panel data set for which we compare the different tests across varying sample sizes, imputation techniques and missingness proportions. Finally, in section 4 we give concluding remarks on the findings of the study.

1.1 *The conditional maximum likelihood estimator for logistic panel data model*

1.1.1 *Logit panel data model*

In order to represent dichotomous responses, we use the logistic function of the regression model to develop the logit panel model. This model has found extensive use in virtually all research domains that perform impact analyses to policy changes or product usage. Suppose that all observations are captured for each unit i, $i = 1, \ldots, N$ from a study such that each unit is observed T times. Further, suppose that the response variable is defined as binary. Then, we have a $T \times 1$ vector of the binary response variable Y as $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iT})'$, where $Y_{it} \in \{0, 1\}$, adopting the values 1 or 0 if an event is successful or not. It then follows that $Y_{it} \sim bernoulli(p_{it})$ and $E(Y_{it}) = p_{it}$. Similarly, for each time t, let Y_{it} be predicted as a function of corresponding $1 \times k$ vector of covariates $\mathbf{x_{it}} = (x_{it}^{(1)}, x_{it}^{(2)}, \ldots, x_{it}^{(K)})$ and fixed effects parameter c_i. It is feasible to employ the logistic function to represent the relationship between the response variable (Y_{it}) and the covariates matrix ($\mathbf{x_{it}}$) basing on the fact that the response variable is dichotomous:

$$p_{it} = Pr(Y_{it} = 1) = E(Y_{it} \mid \mathbf{x_{it}}, c_i) = F(\mathbf{x_{it}}'\beta + c_i), \tag{1}$$

where $\beta = (\beta_1, \beta_2, \ldots, \beta_K)'$ denotes the $k \times 1$ vector of k regression slope parameters for the vector $\mathbf{x_{it}}$. This is because the link function $F(\cdot)$ connects the binary regressand to the functional forms of the regressors, making it a probability model. Collating all outcome probabilities for the i^{th} unit gives the vector $p_i = (p_{i1}, p_{i2}, \ldots, p_{iT})'$. The equation (1) can be linearized by finding the logarithm of the odds-ratio so as to get the logit panel data model as

$$log\left(\frac{p_{it}}{1-p_{it}}\right) = \mathbf{x_{it}}'\beta + c_i. \tag{2}$$

Estimating equation (2) for the parameters β and c_i for a panel data set with n subjects imply that we get a total of $k+N$ estimates since c_i's are incidental parameters. As such, we may easily eliminate the c_i's from the estimator by adopting the conditional maximum likelihood estimator. Notice that we have a logit model for a single response only and for a likelihood-based analysis, we must specify the entire joint distribution of the responses. The maximum likelihood estimators of the parameters have been developed for marginal regression models. However, if we consider the various interaction between the responses, we may categorize these estimators as marginal (unconditional) or conditional measures of association. As a matter of fact, Chamberlain [9] shows that the conditional maximum likelihood estimation eliminates the fixed effects from the likelihood function for response variables that are categorical. It does this by conditioning the joint probability of the regressand to the minimal sufficient statistic of the fixed effect model parameters.

1.1.2 *The unconditional likelihood function*

We obtain the likelihood function from all outcomes in the sample as a product of the marginal probabilities of an individual i at a time t. The total sample response values are NT if pooled together and therefore we have the likelihood,

$$L(\beta, c \mid \mathbf{x}; y) = \prod_{i=1}^{N}\prod_{t=1}^{T}\left[\frac{e^{\mathbf{x_{it}}'\beta+c_i}}{1+e^{\mathbf{x_{it}}'\beta+c_i}}\right]^{y_{it}}\left[\frac{1}{1+e^{\mathbf{x_{it}}'\beta+c_i}}\right]^{1-y_{it}}, \tag{3}$$

The maximum likelihood estimators $(\hat{\beta}, \hat{c}_i)$ maximizes the log-likelihood function of (3).

1.1.3 *Conditional likelihood function for logistic panel data model*

A major drawback of estimating equation (3) is the presence of the incidental parameters, c_i. With the logistic functional form we can do away with c_i from the likelihood function by conditioning the joint probabilities on the minimal sufficient statistic for c_i to obtain the conditional likelihood function. Opeyo et al. [8] show that the conditional joint probability when $T = 2$ is

$$Pr(Y_{i1}, Y_{i2} \mid \mathbf{x_{i1}}, \mathbf{x_{i2}}, c_i, Y_{i1}+Y_{i2} = 1) = \begin{cases} 1, & \text{if } (Y_{i1}, Y_{i2}) = (0, 0) \text{ or } (1, 1)\,, \\ \frac{1}{1+e^{(\mathbf{x_{i2}}'-\mathbf{x_{i1}}')\beta}}, & \text{if } (Y_{i1}, Y_{i2}) = (1, 0)\,, \\ \frac{e^{(\mathbf{x_{i2}}'-\mathbf{x_{i1}}')\beta}}{1+e^{(\mathbf{x_{i2}}'-\mathbf{x_{i1}}')\beta}}, & \text{if } (Y_{i1}, Y_{i2}) = (0, 1)\,. \end{cases} \tag{4}$$

We observe that the conditioning is on $Y_{i1}+Y_{i2} = 1$, where Y_{it} only changes in value across the two observation times. This fact that the c_i's are eliminated from the likelihood function makes $\sum_t Y_{it}$ a sufficient statistic for the fixed effects. The conditional log-likelihood function from (4) is then given as

$$\ln L(y \mid \mathbf{x};\, \beta) = \sum_{i=1}^{N} \left\{ h_{01i} \ln \left[\frac{e^{(\mathbf{x_{i2}}' - \mathbf{x_{i1}}')\beta}}{1 + e^{(\mathbf{x_{i2}}' - \mathbf{x_{i1}}')\beta}} \right] + h_{10i} \ln \left[\frac{1}{1 + e^{(\mathbf{x_{i2}}' - \mathbf{x_{i1}}')\beta}} \right] \right\}, \tag{5}$$

where h_{01_i} picks out units for which the response variable changes from 0 to 1 while h_{10_i} picks out changes of the dependent variable from the success value 1 to failure value 0. The generalized conditional joint probability function for T time periods is obtainable using the similar approach by conditioning on $\sum_t Y_{it}$ as.

$$Pr(Y_{i1},\, Y_{i2},\, \ldots,\, Y_{iT} \mid \mathbf{X_i},\, c_i,\, \sum_t y_{it}) = \frac{e^{\sum_t y_{it} \mathbf{x_{it}}' \beta}}{\sum_{h \in B_i} e^{\sum_t h_{it} \mathbf{x_{it}}' \beta}}, \tag{6}$$

where $B_i = \{(h_{i1},\, h_{i2},\, h_{i3},\, \ldots,\, h_{iT}) \mid h_{it} = \{0,\, 1\}$ and $\sum_t h_{it} = \sum_t Y_{it}$.

1.1.4 *Case of imputed covariate matrix*

Suppose the covariate vector $\mathbf{x_{it}}$ has holes due to missingness, it may be partitioned into two sub vectors $\mathbf{x_{it_s}}$ and $\mathbf{x_{it_I}}$ for the sample non-imputed covariate values and the imputed covariate values respectively. When $T = 2$, we have from equation (4) the conditional probabilities expressed as

$$Pr(Y_{i1} = 0,\, Y_{i2} = 1 \mid \mathbf{x_{i1}},\, \mathbf{x_{i2}},\, c_i,\, Y_{i1} + Y_{i2} = 1) = \frac{e^{\Delta \mathbf{x_{i_I}}' \beta}}{e^{-\Delta \mathbf{x_{i_s}}' \beta} + e^{\Delta \mathbf{x_{i_I}}' \beta}} \tag{7}$$

and

$$Pr(Y_{i1} = 1,\, Y_{i2} = 0 \mid \mathbf{x_{i1}},\, \mathbf{x_{i2}},\, c_i,\, Y_{i1} + Y_{i2} = 1) = \frac{e^{-\Delta \mathbf{x_{i_s}}' \beta}}{e^{-\Delta \mathbf{x_{i_s}}' \beta} + e^{\Delta \mathbf{x_{i_I}}' \beta}}, \tag{8}$$

where $\Delta \mathbf{x_{i_I}}' = (\mathbf{x_{i2_I}}' - \mathbf{x_{i1_I}}')$ and $\Delta \mathbf{x_{i_s}}' = (\mathbf{x_{i2_s}}' - \mathbf{x_{i1_s}}')$.

Equations (7) and (8) then yield the conditional log-likelihood function for the imputed covariate matrix as

$$\ln L(y \mid \mathbf{x};\, \beta) = \sum_{i=1}^{N} \left\{ h_{01i} \ln [\frac{e^{\Delta \mathbf{x_{i_I}}' \beta}}{e^{-\Delta \mathbf{x_{i_s}}' \beta} + e^{\Delta \mathbf{x_{i_I}}' \beta}}] + h_{10i} \ln [\frac{e^{-\Delta \mathbf{x_{i_s}}' \beta}}{e^{-\Delta \mathbf{x_{i_s}}' \beta} + e^{\Delta \mathbf{x_{i_I}}' \beta}}] \right\}, \tag{9}$$

The parameters of the nonlinear equation (9) are estimated iteratively using Newton-Raphson algorithm to obtain consistent estimates $\hat{\beta}_{MLE}$ and consequently $\hat{c}_{iMLE}$ which when used back in equation (1) provides the estimated probabilities of a success for the i^{th} unit at a time t as

$$\hat{p}_{it} = \frac{e^{\mathbf{x_{it}}' \hat{\beta}_{\mathbf{MLE}} + \hat{c}_{iMLE}}}{1 + e^{\mathbf{x_{it}}' \hat{\beta}_{\mathbf{MLE}} + \hat{c}_{iMLE}}}. \tag{10}$$

2. The goodness-of-fit test for logistic regression model

When modeling dichotomous response variables and the predicted values obtained from the econometric model fail to accurately portray the observed values, then the model yields large residual variations. Such a model is considered to be a poorly fitting model [10]. The non-linearity of the logistic regression model is among the several reasons why it does not adequately fit the data [11]. Poor model fit may also result from the exclusion of significant covariates that are related to the response variables or higher order terms of covariates from the model. Additionally, presence of outliers may result in a poor fit. Generally, goodness of fit in logistic regression attempts to measure how well a fitted model fits the observed data. Since we may not necessarily get a single model that outshines the rest explicitly, it is reliably wise to fit a series of possible models and evaluate them independently. Classical approaches to goodness-of-fit are anchored on hypothesis testing for which the test statistics are used to establish the fit of the model through the statement:

H_0: There is no significant difference between the observed data and the specified model.

H_1: There is a significant difference between the observed data and the specified model.

The outstanding threat here is that we may wrongfully reject a correctly specified model or fail to reject one which is not well fitting. Therefore, if the failure to reject a null hypothesis is viewed as evidence that the model is correctly specified, which is obviously not always the case, then these tests may not only be ineffective but also dangerous. It therefore suffices to derive a descriptive measure of how well a model fits which may not affect other predictive applications of the model. This bid has become an ongoing research concern and not a single explicitly perfect solution has been mentioned in literature.

Because we cannot explain the magnitude of the residual when the response variable is binary as we can for continuous dependent variables, R^2 or adjusted-R^2 are not viable metric to quantify the predictive power in logistic regression modelling. Even when the model is suitable for the data, Hosmer and Lemeshow [12] observe that the R^2 and/or adjusted-R^2 are frequently low. The traditional Chi-square test by Pearson and the Deviance test perform admirably when the covariates are categorical but yield incorrect p-values when there exists at least one continuous covariate. Additional two very handy criteria to choose a better model are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) respectively developed by Hirotsuge [13] and Schwarz [14]. Generally, BIC offers a more accurate model than AIC which is attributed to its Bayesian perspective. In diagnostic analysis of the logistic regression model, the expressions of AIC and BIC are given as; $AIC = -2 \times lnL + 2k$ and $BIC = -2 \times lnL + k \times ln(n)$, where L is the maximized value of the likelihood function, n is the sample size and k is the number of the parameters in the model [13]. For a set of models, the values of the AIC are compared to establish the better fitting model (that which has the smallest AIC and/or Schwarz's Criterion (SC) value).

For mixed categorical and continuous covariates, several other goodness-of-fit tests for model assessment have been proposed. For instance, Hosmer et al. [15, 16] developed the $\hat{C}$ and $\hat{H}$ tests. Other two tests proposed by Pulksteris et al. [17] also yielded fair results for logistic models with mixed discrete and continuous covariates but performed poorly when the entire covariates vector space is continuous. Tsiatis [18] and Stukel [19] used score tests to evaluate overall goodness-of-fit for the logistic regression model. In multinomial models, goodness-of-fit test statistics that are asymptotically normal for large degrees of freedom were also proposed in studies by Osius et al. [20] and can always be applied to the binary cases as generalizations of the Pearson Chi-square test. Despite the fact that there is a rich literature proposing many goodness-of-fit tests, establishing the most reliable way of assessing the fit of the logistic regression model is still elusive since each of the existing tests have different merits and demerits. The assumption of these tests is that the data set being used is complete and balanced. How then do these goodness-of-fit tests respond to imputed data sets?

In this study, through Monte Carlo simulation, the fit of the logit panel data model is evaluated against data sets with imputed covariates with an aim of establishing the performance of each imputation technique in yielding the best fit model. The aim being to confirm whether the superiority of Bayesian imputation still holds when performing model diagnostics. Although the bias and precision of parameter estimates are comparatively lower for data sets whose missing covariates are imputed by the Full Bayesian model based technique, the need to establish the best fitting imputed model will validate the imputation technique used. The underlying concept of each of the mentioned goodness of fit tests is discussed in the next section and later applied to the simulated logit panel data set.

2.1 *Framework of goodness-of-fit tests in logistic regression model*

This section gives the test specifications and their technical points which bring out their outstanding relevance to specific covariate data sets. Specific to logistic regression, we reflect on available measures of fit that have been proposed thus far which can be clustered into "global" and "local" measures.

2.1.1 *Chi-square goodness-of-fit tests and deviance*

In linear regression, residuals can be defined as $y_i - \hat{y}_i$ where y_i is the observed response variable for the i^{th} unit, and $\hat{y}_i$ the estimate from the model. This is similarly extended to logistic panel data model whose residuals are $y_{it} - \hat{p}_{it}$ with $\hat{p}_{it}$ defined as in equation (10). From these residuals, we define two tests: the chi-square goodness-of-fit test and the deviance test.

2.1.1.1 *The chi-square test*

Pearson's chi-squared goodness-of-fit test statistic is best described from its expression as the sum of the quotient between the squared residuals per covariate pattern and the residual standard error. We also define a covariate pattern as a unique combination of all specified covariates that can be identified to describe a given sample unit. Suppose J denotes the number of such unique covariate patterns from the available covariates, then let m_j be the total number of units belonging to a particular covariate pattern such that $j = 1, 2, \ldots, J$. From this description, we may observe two distinct cases: (1) $J < n$ when several units share covariate patterns leading to fewer number of patterns than the sample size and (2) $J \approx nt$ where the n units are uniquely sampled by the covariates (there are no clusters). Still relating to the logit panel data model, then for all the units within the j^{th} covariate pattern, let $\hat{p}_{it_j}$ be the MLE estimates of p_{it_j}. This estimated probability is the same for all m_j subjects in the covariate pattern group j. We already have that, y_{it} represents the outcome for the i^{th} unit observed at time t. Let $y_j = y_{js} + y_{jf}$ be the sum of the observed outcomes (both successes and fails) in the j^{th} covariate pattern. Average estimated number of successes for the units in the j^{th} group with the j^{th} covariate pattern is obtained as $\hat{y}_{js} = m_j \hat{p}_j$.

The likelihood function and the log-likelihood function for the covariate pattern can be written respectively as $L(\beta) = \prod_{j=1}^{J} \begin{bmatrix} m_j \\ y_{js} \end{bmatrix} {p_{it_j}}^{y_{js}} (1 - p_{it_j})^{(m_j - y_{js})}$ and $\log L(\beta) = \sum_{j=1}^{J} \left\{ \log \begin{bmatrix} m_j \\ y_{js} \end{bmatrix} + y_{js} \log p_{it_j} + (m_j - y_{js}) \log(1 - p_{it_j}) \right\}$. Here p_{it_j} is the probability of success for the unit i at time t and having the j^{th} covariate pattern. It still follows the logistic distribution equation (1) and is a function of parameters β. The values from each covariate pattern can be tabulated as in Table 1 for easy visualization.

Table 1. Pearsons table for covariate patterns

Covariate Pattern j	Number of subjects	No. of Successes	No of Fails	Estimated Prob. of Success	Estimated Prob. of Fail	No of Estimated Success
1	m_1	y_{1s}	y_{1f}	$\hat{p}_{it1}$	$1 - \hat{p}_{it1}$	$m_1 \hat{p}_{it1}$
2	m_2	y_{2s}	y_{2f}	$\hat{p}_{it2}$	$1 - \hat{p}_{it2}$	$m_2 \hat{p}_{it2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
j	m_j	y_{js}	y_{jf}	$\hat{p}_{itj}$	$1 - \hat{p}_{itj}$	$m_j \hat{p}_{itj}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
J	m_J	y_{Js}	y_{Jf}	$\hat{p}_{itJ}$	$1 - \hat{p}_{itJ}$	$m_J \hat{p}_{itJ}$

Consequently, if we suppose that $J < nt$ so that $m_j \hat{p}_{it_j}$ is large enough for every covariate pattern, it suffices to define Pearson's residual as

$$r(y_{js}, \hat{p}_{it_j}) = \frac{(y_{js} - m_j \hat{p}_{it_j})}{\sqrt{\{m_j \hat{p}_{it_j}(1 - \hat{p}_{it_j})\}}} . \tag{11}$$

By definition, we now get the Pearson's chi-squared goodness-of-fit test statistic for the logistic panel data model as the sum of the squared Pearson's residuals, $r(y_{js}, \hat{p}_{it_j})$, as

$$X^2 = \sum_{j=1}^{J} \left(r(y_{js}, \hat{p}_{it_j}) \right)^2 . \tag{12}$$

The statistic (12) is distributed approximately as $\chi^2(J-k)$ where k is the number of independent covariates in the model. Worthy to note is that when there exists some continuous (one or more) covariates in the vector x_{it} of the logistic regression model then the number of covariate patterns is fairly high $(J \approx nt)$. This makes the component $m_j \ \hat{p}_{it_j}$ to be small relative to y_{js} hence Pearson's chi-squared test may be ineffective.

2.1.1.2 *Deviance test*

For the j^{th} covariate pattern, Nelder and Wedderburn [21], used the definition of the deviance residual d to propose another goodness-of-fit measure. For the logistic panel data model, we have d as:

$$d(y_{js}, \hat{p}_{it_j}) = \pm \left\{ 2 \left[y_{js} \ln\left(\frac{y_{js}}{m_j \hat{p}_{it_j}}\right) + (m_j - y_{js}) \ln\left(\frac{(m_j - y_{js})}{m_j(1 - \hat{p}_{it_j})}\right) \right] \right\}^{\frac{1}{2}} , \tag{13}$$

where the parameters are similarly defined as those from the chi-square test. The Deviance test statistic is also obtained from the deviance residual as:

$$D = \sum_{j=1}^{J} \left(d(y_{js}, \hat{p}_{it_j}) \right)^2 , \tag{14}$$

and D is approximately $\chi^2(J-k)$. When there exists numerous continuous covariates in the covariate vector then each covariate pattern uniquely represents a study unit j such that $m_j = 1 \ \forall j$ and equation (13) becomes:

$$d(y_{is}, \hat{p}_{it}) = \pm \left\{ 2 \left[y_{is} \ln\left(\frac{y_{is}}{\hat{p}_{it}}\right) + (1 - y_{is}) \ln \left(\frac{1 - y_{is}}{1 - \hat{p}_{it}} \right) \right] \right\}^{\frac{1}{2}} , \tag{15}$$

with D as $D = \sum_{i=1}^{nt} \left(d(y_{is}, \hat{p}_{it}) \right)^2$. With the assumption that y_{is} is binary, the terms $y_{is} \ln(y_{is})$ and $(1 - y_{is}) \ln(1 - y_{is})$ vanish hence when we use the estimator $\hat{y}_{is} = \hat{p}_{it}$ we have

$$D = -2\sum_{i=1}^{nt}\left\{\hat{p}_{it}\ln(\hat{p}_{it}) + (1-\hat{p}_{it})\ln(1-\hat{p}_{it})\right\}. \tag{16}$$

The test statistic and the corresponding p-value for the Deviance test are relatively easier to calculate compared with those of the Pearson Chi-square test hence giving the Deviance test an analytical preference. However, for these two tests, the p-values are not always correct when the covariate patterns are large such that $J = n$ [12]. Collett [11] justified the prevalence of the Deviance statistic over the Pearson Chi-square statistic by the fact that D is minimized by the maximum likelihood estimates of the success probabilities $\hat{p}_{it}$ of the fitted logistic regression model. Faced with an imputed covariate matrix, the propagation of bias from the MLE $\hat{p}_{it}$ will in turn affect the value of the test statistic.

2.1.2 *Hosmer-lemeshow methods*

Hosmer and Lemeshow continued with the same re-grouping idea as used in deviance tests and developed two widely used tests in which subjects are grouped according to the values of probability estimates of success. These test statistics do not necessarily require the condition $J < n$ as was the necessity for Pearsons Chi-square test and deviance tests [22]. The test statistics developed by Hosmer and Lemeshow are $\hat{H}$ and $\hat{C}$, differentiated according to how the frequencies of the probabilities of success are obtained. $\hat{H}$ statistic is obtained by fixing a pre-determined threshold value of the estimated success probability [15] while $\hat{C}$ is based on the percentage proportions of estimated probabilities [16].

2.1.2.1 *Hosmer and Lemeshow's $\hat{C}$*

To obtain the $\hat{C}$ test statistic we regroup the subjects into $g(\leq 10)$ groups with each group containing $\frac{n\times t}{g}$ subjects. The grouping criterion is such that the first group contains the smallest estimated success probabilities $\hat{p}_{it}$ calculated from the fitted assumed model. As such we have an ordered monotonically increasing set of groups based on the estimated probabilities. Letting $\overline{p}_k$ be the group mean for the success probability estimates in line with the fitted model corresponding to the study units in the k^{th} group with $y_{it} = 1$. Also let o_k be the number of study units in the k^{th} group with $y_{it} = 1$. We generate a two-column frequency table (Table 2) with g columns where the row values correspond to the two recordings (observed and expected) of the response variable and the g columns representing the g groups. A plot of the observed success probabilities against the expected success probabilities for the best fitting model is linear through the origin with slope 1 and diagnoses the overall fit across the spectrum of predicted probabilities.

Table 2. Observed and Expected probabilities of success for each group

	Group				
	1	2	3	…	g
Observed Success probability	o_1	o_2	o_3	…	o_g
Expected Success probability	$n_1\overline{p}_1$	$n_2\overline{p}_2$	$n_3\overline{p}_3$	…	$n_g\overline{p}_g$

From the frequency table the Hosmer and Lemeshow test statistic $\hat{C}$ is obtained as

$$\hat{C} = \sum_{k=1}^{g}\frac{(o_k - n_k\overline{p}_k)^2}{n_k\overline{p}_k(1-\overline{p}_k)}, \tag{17}$$

where $\sum_{k=1}^{g} n_k = nt$ with n_k being the total subjects in the k^{th} group for all $k = 1, 2, \ldots, g$. Here again, $\hat{C} \sim \chi^2(g-2)$ [10]. Limiting $g = 10$ and ensuring that each group has a fair number of subjects makes it difficult to choose an arbitrary $W \in \mathbb{N}$ such that $nt = 10 \times W$. Therefore, $\hat{C}$ is very sensitive to the subjective choice of W during groups formation as evidenced by different statistical softwares having different algorithms for defining the cut-off points. An observation by Bertolini et al. [23] is that Hosmer and Lemeshow goodness-of-fit test results may still be inaccurate when the number of covariate patterns $J < n$.

2.1.2.2 *Hosmer and Lemeshow's $\hat{H}$*

Without fixing the number of subjects per group, $\hat{H}$ is an alternative test statistic also derived by Hosmer and Lemeshow whereby the estimated probabilities are clustered into intervals of 0.1 from 0 to 1 so as to generate the 10 (or less) desired groups. Consequently, the number of study units or subjects may vary across the different groups. Akin to the $g \times 2$ frequency table used in the $\hat{C}$ test, the Hosmer-Lemeshow $\hat{H}$ test statistic is derived from the computation of the Pearson chi-square statistic as:

$$\hat{H} = \sum_{k=1}^{g} \frac{(o_k - n_k \overline{p}_k)^2}{n_k \overline{p}_k (1 - \overline{p}_k)} \,. \tag{18}$$

where o_k, $\overline{p}_k$ and n_k are similarly defined as for the case of $\hat{C}$ for large n, $\hat{H} \sim \chi^2(g-2)$ under the null hypothesis. Comparisons of $\hat{C}$ and $\hat{H}$ tests indicate that $\hat{H}$ test is more powerful than $\hat{C}$ test except when a significant number of estimated probabilities are within the lower two deciles [12].

Summarily, the two Hosmer-Lemeshow's tests can be taken as extensions of the Pearson chi-square test if multiple covariate patterns are merged into one group making the distribution of the Pearson residuals approximately normal. Additionally, other developments on the $\hat{C}$ and $\hat{H}$ have included all probabilities for both success and failure ($Y = 1$ and $Y = 0$) so as to make use of all available data information.

2.1.3 *Classification tables*

A logistic regression model's classification table is used to reveal the model's accuracy, or how well it specifies successes and failures of an event for a given classification cut-off probability. After obtaining all fitted values of the binary response variable, $\hat{p}_{it}$ of a logit panel model, all the fitted values are then classified according to whether they fall above or below a predetermined threshold (cut-off) value, p_0. The classification is such that all $\hat{p}_{it} \geq p_0$ are qualified as successes of an event while $\hat{p}_{it} < p_0$ are failures of the same event. This classification enables us to generate a 2 × 2 table (Table 3) due to the dichotomous nature of the response variable.

Table 3. Classification table from a specified panel data set

			Observed		
			Observed positive $y_{it} = 1$	Observed negative $y_{it} = 0$	
Predicted	Predicted positive	$\hat{p}_{it} \geq p_0;\ y_{it} = 1$	A = TP	B = FP	PP = TP + FP
	Predicted negative	$\hat{p}_{it} < p_0;\ y_{it} = 0$	C = FN	D = TN	PN = FN + TN
			OP = TP + FN	ON = FP + TN	Tot = TP + FP + FN + TN

The Table 3 is referred to as the classification table from which the following counts can be defined:

i. True Positives (TP) = the number of cases that were accurately classified as positive, that is, those that were expected to be successful and were found to be successful,

ii. False Positives (FP) = the number of cases that were incorrectly classified as positive, that is, were predicted to be a success but were actually observed to be a failure,

iii. True Negatives (TN) = the number of cases that were correctly classified to be negative, that is, were predicted to be a failure and were actually observed to be a failure,

iv. False Negatives (FN) = the number of cases that were incorrectly classified as negative, that is, were predicted to be a failure but were actually observed to be a success.

Prior expectation that would indicate a good fit is to have higher counts of TP and TN, and fewer counts of FP and FN. If otherwise, we define analytical test values for a particular model as sensitivity, specificity, precision and Model accuracy.

2.1.3.1 *Sensitivity, specificity, precision and model accuracy*

From the parameters of the specification table, the following four measures are defined:

$$\text{Sensitivity} = \frac{TP}{TP+FN}, \tag{19}$$

$$\text{Specificity} = \frac{TN}{TN+FP}, \tag{20}$$

$$\text{Precision} = \frac{TP}{TP+FP}, \tag{21}$$

$$\text{Model Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)}. \tag{22}$$

In essence the ability of a model to correctly classify successes and failures are evaluated by the sensitivity and specificity, respectively. Precision, however, measures how good the model is at predicting a success of an event. Generally, higher sensitivity and specificity values would indicate a better fit of the model.

2.1.3.2 *Receiver operating characteristic (ROC) curves*

Receiver Operating Characteristic (ROC) curves are graphical representations commonly used in binary classification tasks to assess the performance of a classification model. Such binary classification problems are those where the outcome can be classified into two categories, typically denoted as positive (success of an event) and negative (failure of an event). ROC curves are therefore widely used in medical diagnostics, machine learning, and other fields where binary classification is common, providing a comprehensive evaluation of a model's performance across various thresholds. Graphically, ROC curves are plots of the true positive rate (TPR) against the false positive rate (FPR) such that each point on the curve represents the performance of the model at a particular classification threshold.

Specific to the logistic panel data model and taking into account all possible threshold values $p_0 \in [0, 1]$ that can be used to generate different classification tables, corresponding pairs of specificities and sensitivities can be collated for every chosen cutoff value. A plot of sensitivity against 1-specificity from these paired values provides the Receiver Operating Characteristic (ROC) curve where the domains of the axes are both $[0, 1] \subset \mathbb{R}$. The ROC curve lies between two limiting curves: the random chance diagonal ((Sensitivity = $1-$ Specificity)) and the gold standard lines (Sensitivity = 1 and Specificity = 1). Sampling ROC curves in the Figure 1 shows diagnostic accuracy of the gold standard (lines Q; AUC = 1) on the upper and left axes in the unit square, a typical ROC curve (curve R; AUC = 0.79), and a diagonal line corresponding to random chance (line S; AUC = 0.5). The overall measure of fit of the model is presented by the area under the ROC curve (AUC of the ROC. As such, during diagnosis, a better fitting model yields an AUC value closer to 1 than 0.5.

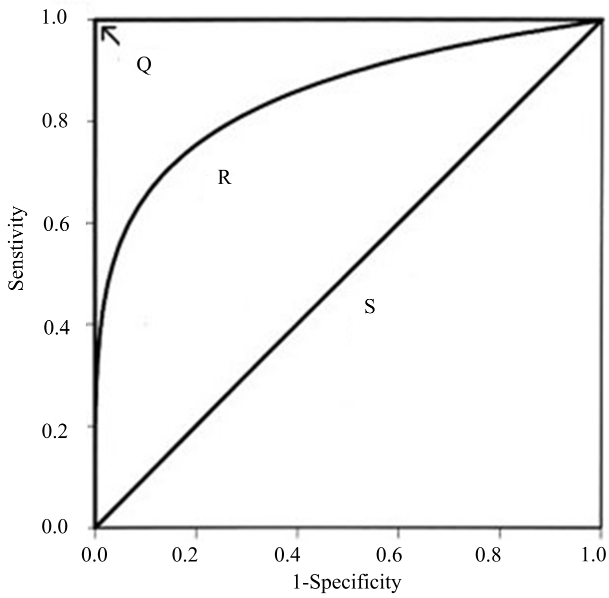


Figure 1. Expected Receiver Operating Characteristic (ROC) curve

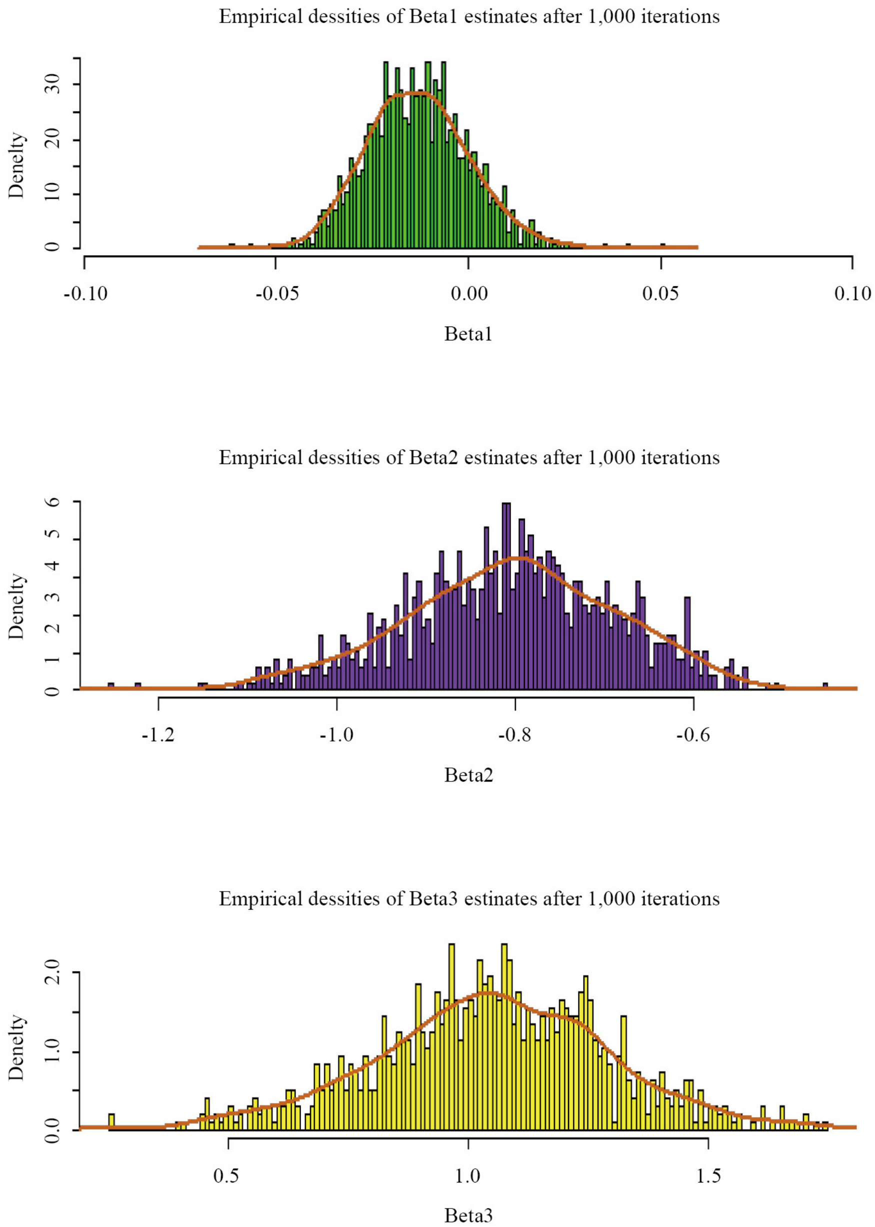
3. Logit panel simulation study

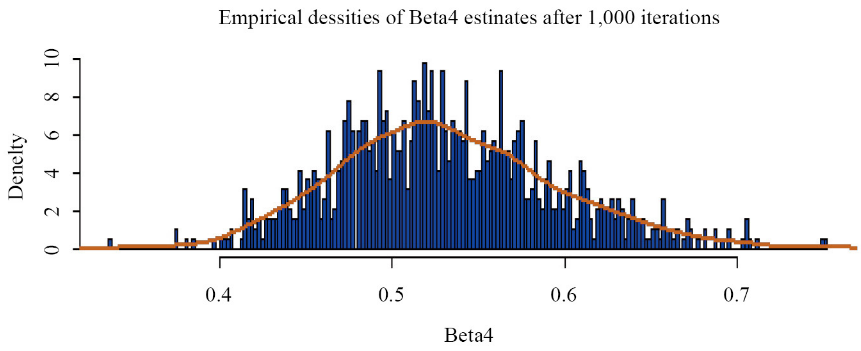
In this section, we assess the impact of covariate imputation to the various goodness-of-fit statistics. Through Monte Carlo simulation, we hypothesize and generate a panel data set with a dichotomous response variable Y_{it} observed in three time periods ($T = 3$) whose probability is predicted from a logistic distribution relating five covariates $X^{(1)}$, $X^{(2)}$, $X^{(3)}$, $X^{(4)}$, and $X^{(5)}$ as

$$Pr(Y_{it} = 1) = \frac{1}{1 + exp{-}(c_i + \beta_1 x_{it}^{(1)} + \beta_2 x_{it}^{(2)} + \beta_3 x_{it}^{(3)} + \beta_4 x_{it}^{(4)} + \beta_5 x_{it}^{(5)})} \tag{23}$$

where c_i is the individual specific fixed effect parameter specified as $c_i = \frac{10 \sum x^{(1)}}{3n} + \alpha_i$ and $\alpha_i \sim N(5,\ 1)$. Using R software, the covariates $X^{(1)}$, $X^{(2)}$, $\ldots$, $X^{(5)}$ are specified and randomly generated as follows: $X^{(1)} = round\,(rnorm(nt,\ 45,\ 15),\ 0)$; $X^{(2)} = rpois(nt,\ 3)$; $X^{(3)} = rbinom(nt,\ 1,\ 0.65)$; $X^{(4)} = round(rnorm(nt,\ 15,\ 8),\ 2)$; $X^{(5)} = bernoulli(nt) = \begin{cases} 1, & \text{if } X^{(1)} + h_{it} \geq 40\ , \\ 0, & \text{Otherwise}\ . \end{cases}$ and $h_{it} = rnorm(nt, 0,\ 1)$.

By randomly deleting proportions of the covariate matrix and imputing the values back, we calculate resultant goodness-of-fit probabilities of the test statistics and the AUC for three different sample sizes ($n = 100$, $n = 200$ and $n = 300$). The proportions of the missingness are fixed as 10% and 30% respectively for each sample size used. Figure 2 shows the empirical densities of the parameters β_1 to β_5 from 1,000 iterations when estimated by conditional MLE. We collated the goodness-of-fit test results when various imputation techniques employed against each sample size and missingness proportions in Tables 4 and 5 below. Across the Tables 4 and 5 we observe that the chi-square (χ^2) statistic values increase with increasing sample size indicating a diminishing chance of rejecting the null hypothesis. This shows that all the goodness-of-fit statistics considered in this study become asymptomatically valid irrespective of the imputation technique used to compensate for the missing covariates. Comparatively, the chi-square statistic values for the Pearson's chi-square, Hosme-Lemeshow's $\hat{C}$ and $\hat{H}$ are seen to be sensitive to missingness proportions and the chi-square values are lower when a bigger proportion of the data need to be imputed. From the ROC curves (Figure 3) and the average area under the curves (AUC), Bayesian imputation though multiple imputation by chained equations (MICE) is seen to perform better than both median and last value carried forward (LVCF) imputations by yielding a higher AUC. The simulation results, however, do not provide much information for mean imputation as most of the test statistic values overlapped with those of the complete data set.





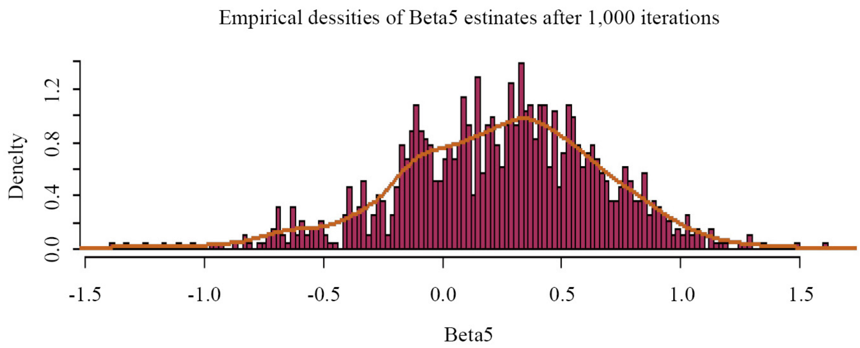


Figure 2. Empirical densities of the parameters β_1 to β_5 from 1,000 iterations when estimated by conditional MLE

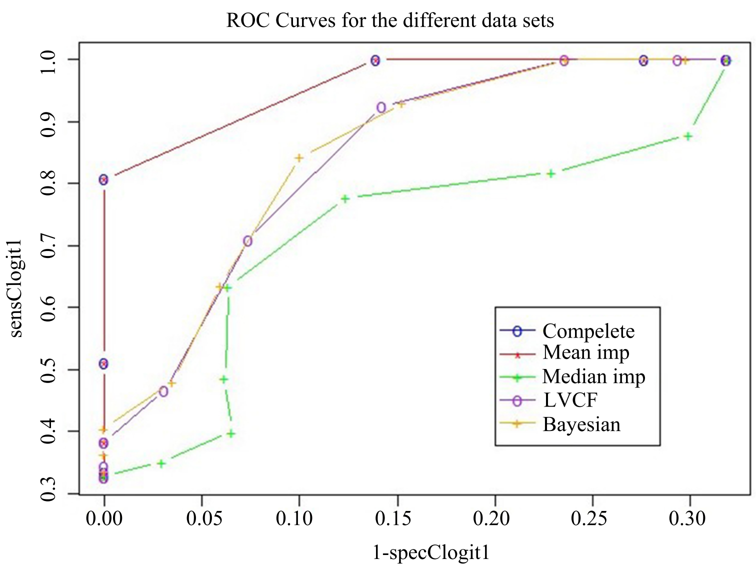


Figure 3. Receiver operating characteristic curves for the different data sets used with a sample size of $N = 300$

Table 4. Goodness-of-fit test statistics for different imputation techniques across varying sample sizes at 10% missingness proportion

	Goodness of Fit Test Statistic (Missingness Proportion = 10%)								
	Chi square Test Statistic p_{value} (X^2)			Hosmer-Lemeshow C Statistic p_{value} $(\hat{C})$			Hosmer-Lemeshow H Statistic p_{value} $(\hat{H})$		
	N = 100	N = 200	N = 300	N = 100	N = 200	N = 300	N = 100	N = 200	N = 300
Complete	1	1	1	9.992×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	3.653×10^{-14}	2.2×10^{-16}	2.2×10^{-16}
	(133.7613)	(249.2736)	(593.5277)	(88.334)	(196.91)	(504.5)	(80.631)	(183.03)	(507.19)
Mean Imp	1	1	1	3.48355×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	3.653×10^{-14}	2.2×10^{-16}	2.2×10^{-16}
	(135.5542)	(249.6281)	(595.2572)	(88.989)	(197.24)	(505.15)	(80.801)	(182.11)	(507.912)
Median Imp	0	0.99971	1	3.038×10^{-13}	2.2×10^{-16}	2.2×10^{-16}	3.348×10^{-8}	2.2×10^{-16}	2.2×10^{-16}
	(2642.748)	(483.4743)	(616.364)	(107.44)	(112.72)	(377.53)	(108.81)	(113.09)	(383.95)
LVCF	0.9657643	1	1	1.176×10^{-10}	2.2×10^{-16}	2.2×10^{-16}	1.787×10^{-13}	2.2×10^{-16}	2.2×10^{-16}
	(252.322)	(281.3253)	(621.1517)	(63.04)	(148.44)	(488)	(77.199)	(146.36)	(423)
Bayesian (MICE Imp)	0.04804199	1	1	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}
	(336.5688)	(342.0526)	(539.783)	(151.68)	(168.1)	(351.02)	(141.93)	(175.53)	(339.91)
	Sensitivity (cut-off p = 0.5)			Specificity (cut-off p = 0.5)			AUC		
Complete	0.736434	0.8689956	0.8215223	1	0.9973046	1	0.9291	0.58	0.5
Mean Imp	0.737841	0.8677265	0.8199215	1	0.9971216	1	0.9279	0.58	0.5
Median Imp	0.72973	0.8136364	0.7615385	0.3274336	0.9447368	0.9686275	0.9411	0.5588	0.5
LVCF	0.701613	0.8916256	0.8214286	0.9545455	0.9521411	0.9738806	0.8769	0.555	0.5
Bayesian (MICE Imp)	0.628571	0.6515342	0.6601732	0.95625	0.974026	0.9817352	0.889	0.78	0.5351

Table 5. Goodness-of-fit test statistics for different imputation techniques across varying sample sizes at 30% missingness proportion

	Goodness of Fit Test Statistic (Missingness Proportion = 30%)								
	Chi square Test Statistic p_{value} (X^2)			Hosmer-Lemeshow C Statistic p_{value} $(\hat{C})$			Hosmer-Lemeshow H Statistic p_{value} $(\hat{H})$		
	N = 100	N = 200	N = 300	N = 100	N = 200	N = 300	N = 100	N = 200	N = 300
Complete	1	1	1	1.176×10^{-8}	2.2×10^{-16}	2.2×10^{-16}	5.624×10^{-11}	2.2×10^{-16}	2.2×10^{-16}
	(83.38594)	(235.3807)	(565.8121)	(52.805)	(178.83)	(492.59)	(64.664)	(178.99)	(456.49)
Mean Imp	1	1	1	1.176×10^{-8}	2.20×10^{-16}	2.2×10^{-16}	5.624×10^{-11}	2.2×10^{-16}	2.2×10^{-16}
	(83.43237)	(234.92732)	(565.7219)	(52.805)	(177.993)	(492.67)	(64.941)	(179.109)	(456.67)
Median Imp	0.0301426	0.002460059	0.9999997	1.21×10^{-5}	2.2×10^{-16}	2.2×10^{-16}	0.0002719	2.2×10^{-16}	2.2×10^{-16}
	(342.2893)	(696.6234)	(698.4148)	(36.825)	(107.44)	(207.82)	(29.38)	(108.81)	(203.26)
LVCF	0.9999997	1	1	0.6477	6.106×10^{-15}	2.2×10^{-16}	0.01652	5.773×10^{-15}	2.2×10^{-16}
	(188.8927)	(344.7959)	(603.6212)	(5.9959)	(84.491)	(275.2)	(18.705)	(84.615)	(277.84)
Bayesian (MICE Imp)	0.3716434	1	1	3.497×10^{-14}	2.2×10^{-16}	2.2×10^{-16}	2.398×10^{-14}	2.2×10^{-16}	2.2×10^{-16}
	(302.3498)	(419.5884)	(594.2662)	(80.721)	(133.07)	(241.11)	(81.539)	(133.48)	(250.28)
	Sensitivity (cut-off p = 0.5)			Specificity (cut-off p = 0.5)			AUC		
Complete	0.8181818	0.8306452	0.8071625	1	1	1	0.8796	0.5752	0.5
Mean Imp	0.8008827	0.831242	0.8061327	1	1	1	0.8724	0.5763	0.5
Median Imp	0.6818182	0.6036364	0.6343826	0.8263158	0.8769231	0.936345	0.5113	0.5751	0.5137
LVCF	0.8811881	0.726087	0.7086835	0.9045226	0.8945946	0.9263352	0.6991	0.6602	0.5154
Bayesian (MICE Imp)	0.6447368	0.6164384	0.6361446	0.9324324	0.9155844	0.9402062	0.8197	0.7767	0.5597

4. Conclusion

In this paper, we carried out a comparative study on the impacts of various imputation techniques on the goodness-of-fit test statistics used when fitting logistic panel data models. The biases introduced to parameter estimates as a result of imputation of missing covariates is in turn propagated into the calculation of the goodness-of-fit test statistics. Our simulation study reveals that for the conditional MLE of the logit panel data model, the model-based imputation (Bayesian, MICE) impacts less on the goodness-of-fit test statistics in comparison with other classical imputation techniques. In addition, the simulation also establishes that larger missingness proportions tended to reduce the confidence interval of the test statistics hence reducing the chances of adopting whichever model under study as it were.

Apart from the measures of goodness-of-fit considered in the study here-in, there exist more openings for other unique proposals for goodness-of-fit test statistics that would be even more robust to imputation and yield plausible model diagnosis results.

Conflict of interest

The authors declare no competing financial interest.

References

[1] Fagerland MW, Hosmer DW, Bofin AM. Multinomial goodness-of-fit tests for logistic regression models. *Statistics in Medicine*. 2008; 27(21): 4238-4253.

[2] Zamanzade E, Mahdizadeh M. Entropy estimation from ranked set samples with application to test of fit. *Colombian Journal of Statistics*. 2017; 40(2): 223-241.

[3] Mahdizadeh M, Zamanzade E. Goodness of fit tests for rayleigh distribution based on phi-divergence. *Colombian Journal of Statistics*. 2017; 40(2): 279-290.

[4] Zamanzade E. Edf-based tests of exponentiality in pair ranked set sampling. *Statistical Papers*. 2019; 60(6): 2141-2159.

[5] Pho KH. Goodness of fit test for a zero-inflated bernoulli regression model. *Communication in Statistics Simulation and Computation*. 2024; 53(2): 756-771.

[6] Geng Z, Li J, Niu Y, Wang X. Goodness-of-fit test for a parametric mixture cure model with partly interval-censored data. *Statistics in Medicine*. 2023; 42(4): 407-421.

[7] Badi NHS. Asymptomatic distribution of goodness-of-fit tests in logistic regression model. *Open Journal of Statistics*. 2017; 7(1): 434-445.

[8] Opeyo PO, Olubusoye OE, Odongo LO. Conditional maximum likelihood estimation for logit panel models with non-responses. *International Journal of Science and Research*. 2014; 3: 2242-2254.

[9] Chamberlain G. Analysis of covariance with qualitative data. *The Review of Economic Studies*. 1980; 47(1): 225-238.

[10] Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*. 1997; 16(9): 965-980.

[11] Collett D. Statistical inference for binary data. In: *Modelling Binary Data*. Berlin, Germany: Springer; 1991. p.17-42.

[12] Hosmer DW, Lemeshow S. *Applied logistic regression*. 2nd ed. Hoboken, NJ, USA: John Wiley & Sons; 2013.

[13] Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*. 1974; 19(6): 716-723.

[14] Schwarz G. Estimating the dimension of a model. *The Annals of Statistics*. 1978; 1978: 461-464.

[15] Hosmer DW, Lemesbow S. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics-Theory and Methods*. 1980; 9(10): 1043-1069.

[16] Lemeshow S, Hosmer DW. A review of goodness of fit statistics for use in the development of logistic regression models. *American Journal of Epidemiology*. 1982; 115(1): 92-106.

[17] Pulkstenis E, Robinson TJ. Two goodness-of-fit tests for logistic regression models with continuous covariates. *Statistics in Medicine*. 2002; 21(1): 79-93.

[18] Tsikriktsis N. A review of techniques for treating missing data in om survey research. *Journal of Operations Management*. 2005; 24(1): 53-62.

[19] Stukel TA. A general model for estimating ED_{100p} for binary response dose-response data. *American Statistician* 1990; 44(1): 19-22.

[20] Osius G, Rojek D. Normal goodness-of-fit tests for multinomial models with large degrees of freedom. *Journal of the American Statistical Association*. 1992; 87(420): 1145-1152.

[21] Nelder JA, Wedderburn RW. Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*. 1972; 135(3): 370-384.

[22] Hosmer DW, Jovanovic B, Lemeshow S. Best subsets logistic regression. *Biometrics*. 1989; 1989: 1265-1270.

[23] Bertolini G, D'Amico R, Nardi D, Tinazzi A, Apolone G. One model, several results: The paradox of the hosmer-lemeshow goodness-of-fit test for the logistic regression model. *Journal of Epidemiology and Biostatistics*. 2000; 5(4): 251-253.