.

Review : Logistic Regression with Missing Covariates – Parameter Estimation, Model Selection and Prediction within a Joint-Modeling Framework

Original paper : `https://arxiv.org/pdf/1805.04602.pdf`

Lucas Biéchy & Matisse Roche

Department of Mathematics, University of Paris-Saclay, Orsay Mathematics Laboratory
Email address: `lucas.biechy@universite-paris-saclay.fr`

Department of Mathematics, University of Paris-Saclay, Orsay Mathematics Laboratory
Email address: `matisse.roche@universite-paris-saclay.fr`

GitHub repository : `https://github.com/Biechy/LRwithMissingCovariates`

Abstract

This review delves into the paper titled 'Adapting Logistic Regression for Missing Data: A Stochastic Approximation EM Approach' by Jiang et al., published in 2020. Addressing the challenge of missing data in logistic regression, the paper proposes a comprehensive framework comprising parameter estimation, model selection, and prediction methods and an application to trauma-related health data.

Our review comprises two main parts: an exploration of the mathematical concepts behind the Stochastic Approximation EM (SAEM) algorithm, the model selection and the prediction followed by a practical evaluation through numerical simulations. In the mathematical section, we explain the purpose and workings of the SAEM algorithm, introducing Metropolis Hastings as a component. We also discuss the variance of SAEM's estimator using Fisher's information. Following this, our simulations replicate the experimental setup described in the paper, providing a comparison with different implementations. The results not only validate the original findings but also offer insights into the effectiveness and scalability of the proposed approach.

Keywords. incomplete data • observed likelihood • logistic regression • metropolis-hastings

Contents

1. Introduction

Dealing with missing data is a recurring challenge in statistics, often due to missing responses in a questionnaire or sensor failures, resulting in gaps in data sets. To prevent significant information loss or the introduction of potential biases in statistical models, it is essential to adapt existing estimation methods so they can be applied to incomplete data. Surprisingly little effort has been devoted to adapting logistic regression to solve this problem. This is precisely the aim of the paper by Jiang, W., Josse, J., Lavielle, M., Group, T., others [1], published in 2020. It offers a comprehensive framework comprising parameter estimation, model selection, and prediction method. The paper presents a Stochastic Approximation version of the EM algorithm (SAEM), based on Metropolis-Hastings sampling, to perform statistical inference in the context of logistic regression with incomplete data, where missing values may be distributed among the covariates. Unlike the Monte Carlo EM (MCEM) method, which requires generating a large number of samples, SAEM employs a stochastic approximation approach to estimate the conditional expectation of the likelihood of complete data, thus providing a significant computational advantage, as demonstrated by simulation studies. In addition, this framework enables model selection using a criterion based on a penalized version of the likelihood of observed data, a particularly valuable feature in contexts with incomplete data. Lastly, this framework develops a prediction method on a test set containing missing data. The paper primarily focuses on its application to a dataset that registers trauma-related health data named Traumbase developed by a team of the same name who co-authored this article. Unfortunately, we do not have access to this database, thus we will confine ourselves to verifying the validity of this new framework using simulated data. In addition, this report is part of the course titled *Guidelines in ML - Missing Data* taught by Oliver Coudray, where we will attempt to draw parallels between the advances presented in the article and the course content.

2. Theory

2.1. Assumptions and notation

The paper is set in the context of binary classification and focuses on logistic regression with missing data, aiming to propose a framework to analyze the public health challenge of major trauma.

Consider the observed data as (y, x), where $y = (y_i)_{1 \leq i \leq n}$ taking values in $\{0, 1\}$, and $x = \left(x_{ij}\right)_{1 \leq i \leq n, 1 \leq j \leq p}$ an $n \times p$ matrix of covariates taking values in $\mathbb{R}$. The standard logistic regression model for binary classification can be expressed as:

$$P(y_i = 1|\ x_i; \beta) = \frac{\exp\Big(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\Big)}{1 + \exp\Big(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\Big)}, \quad i = 1, ..., n \tag{1}$$

where $x_i = \big(x_{\text{i1}}, ..., x_{\text{ip}}\big)$ represent the covariates for individual i and $\beta = \big(\beta_0, \beta_1, ..., \beta_p\big) \in \mathbb{R}^p$ is an unknown vector to estimate. Assume that x_i follows a normal distribution with μ and Σ kwon:

$$x_i \overset{\text{iid}}{\sim} \mathcal{N}_p(\mu, \Sigma), \quad i = 1, ..., n$$

Let $\theta = (\mu, \Sigma, \beta)$ denote the set of parameters of the model.
In this document, the goal is to estimate the parameter vector β in the presence of missing values in the matrix x. For each individual i, we distinguish between the observed elements $x_{i,\text{obs}}$ and the missing elements $x_{i,\text{mis}}$. The covariate matrix is decomposed into $x = (x_{\text{obs}}, x_{\text{mis}})$.
Denote the indicator matrix of missing data $M = \big(M_{ij}, 1 \leq j \leq p\big)$, where $M_{ij} = 1$ if x_{ij} is missing and $M_{ij} = 0$ otherwise and the i-th column of the matrix M is M_i.
The missing data mechanism is characterized by the conditional distribution of M given x and y, with parameter ϕ, denoted by $p(M_i|x_i, y_i, \phi)$. The paper adopt the assumption of a missing at random (MAR) mechanism, where the missingness of a data point depends only on the observed covariates, i.e $p(M|x; \phi) = p(M|x_{\text{obs}}; \phi)$.

2.1.1. Reminder : Consequence of the MAR hypothesis on likelihood

In the case of missing data with MAR hypothesis, we have :

$$\begin{aligned} p(x_{\text{obs}}, M; \theta, \phi) &= \int_{x_{\text{miss}}} p(x; \theta) p(M|x; \phi) dx_{\text{miss}} \\ &= \int_{x_{\text{miss}}} p(x; \theta) p(M|x_{\text{obs}}; \phi) dx_{\text{miss}} \text{ (MAR hyptothesis)} \\ &= p(M|x_{\text{obs}}; \phi) \int_{x_{\text{miss}}} p(x; \theta) dx_{\text{miss}} = p(M|x_{\text{obs}}; \phi) p(x_{\text{obs}}; \theta) \end{aligned}$$

So maximizing $p(x_{\text{obs}}, M; \theta, \phi)$ in θ is equivalent to maximizing $p(x_{\text{obs}}; \theta)$ in θ, which is very useful when estimating θ parameters by maximum likelihood.

2.2. Algorithm SAEM

To estimate the parameter θ of the logistic regression model by maximizing the observed log-likelihood $\mathcal{LL}(x_{\text{obs}}, y; \theta)$, it is common to first consider the classical EM formulation. Given an initial value θ_0, each iteration k updates θ_{k-1} to θ_k with the following two steps:

E-step : Evaluate the quantity

$$Q_k(\theta) = \mathbb{E}[\mathcal{LL}(\theta; x, y)|x_{\text{obs}}, y; \theta_{k-1}] = \int \mathcal{LL}(\theta; x, y)p(x_{\text{mis}}|x_{\text{obs}}, y; \theta_{k-1})\mathrm{d}x_{\text{mis}}$$

M-step : Update the estimation of θ : $\theta_k = \arg\max_\theta Q_k(\theta)$

Given the absence of an explicit expression for the expectation in the E-step of the logistic regression model, the Monte Carlo EM (MCEM) algorithm can be utilized [2, 3]. MCEM generates multiple samples of missing data from the target distribution $p(x_{\text{mis}}|x_{\text{obs}}, y; \theta_{k-1})$ and substitutes the expectation of the complete log-likelihood with an empirical mean. Nevertheless, achieving a precise Monte Carlo approximation of the E-step may demand considerable computational resources.

2.2.1. SAEM Algorythm Description

To overcome the computational challenges of MCEM, the paper introduces a derivate of the Stochastic Approximation EM (SAEM) algorithm [4]. SAEM offers a computationally efficient alternative by replacing the E-step with a stochastic approximation. At each iteration, SAEM performs the following steps:

Simulation step : Generate samples of missing data $x_{\text{mis}}^{(k)}$ for each observation y_i according to $p\big(x_{i,\text{mis}}|x_{i,\text{obs}}, y_i; \theta_{k-1}\big)$.

Stochastic update step Update the function Q according to

$$Q_{k+1}(\theta) = Q_k(\theta) + \gamma_k(\mathcal{LL}(\theta; x_{\text{obs}}, x_{\text{mis}}(k), y) - Q_k(\theta))$$

where (γ_k) is a non-increasing sequence of positive numbers.

Maximization step : Update the parameters θ_{k+1} by maximizing $Q_{k+1}(\theta)$ in θ

This formula updates the estimate of the maximum of the log-likelihood expectation [5] by combining the current estimate with an error term representing the difference between the logarithm of the conditional density and the previous estimate of the expectation. The update step size γ_k controls the magnitude of this update. This step aims to gradually adjust the estimate using stochastic observations of the latent variable, facilitating iterative convergence while avoiding the computational costs associated with exact evaluation of this expectation.

2.3. Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm is a tool commonly used in statistics for sampling from complex probability distributions. The Metropolis-Hastings algorithm takes as input a starting point x_0, a function proportional to the density we want to simulate, and produces a list of data points that asymptotically follow the

desired density. It is particularly useful in the context of the SAEM algorithm for generating samples of missing data.
In the context of the SAEM algorithm, we have the proportionality relation

$$p\big(x_{i,\text{miss}}|x_{i,\text{obs}}, y_i; \theta\big) \propto p(y_i|x_i; \beta)p\big(x_{i,\text{mis}}|x_{i,\text{obs}}; \mu, \Sigma\big)$$

so we use $p\big(x_{i,\text{mis}}|x_{i,\text{obs}}; \mu, \Sigma\big)$ as an argument of the Metropolis-Hastings algorithm, where

$$x_{i,\text{mis}} \mid x_{i,\text{obs}} \sim \mathcal{N}_p(\mu_i, \Sigma_i) \tag{2}$$

with

$$\mu_i = \mu_{i,\text{mis}} + \Sigma_{i,\text{mis,obs}}\Sigma^{-1}_{i,\text{obs, obs}}\big(x_{i,\text{obs}} - \mu_{i,\text{obs}}\big)$$

$$\Sigma_i = \Sigma_{i,\text{mis,mis}} - \Sigma_{i,\text{mis,obs}}\Sigma^{-1}_{i,\text{obs,obs}}\Sigma_{i,\text{obs, mis}}$$

as seen in the course.

2.3.1. Metropolis-Hastings algorythm description

Input Initial sample x_0, transition probability $g(x'|x_t)$, target distribution f, number of samples S.
Output : Sample x_s
$x_t \leftarrow x_0$ **for** $s = 1$ **to** S **do**
    Generate candidate sample $x' \sim g(x'|x_t)$;
    Calculate acceptance ratio $\alpha = \frac{f(x')g(x'|x_t)}{f(x_t)g(x_t|x')}$;
    Generate random number $u \sim \mathcal{U}([0, 1])$;
    if $u \leq \alpha$ **then**
        Accept candidate : $x_{t+1} \leftarrow x'$;
    else
        Reject candidate : $x_{t+1} \leftarrow x_t$;
    end $x_t \leftarrow x_{t+1}$;
end
We implemented the Metropolis-Hastings algorithm in Python[1] and visualized its performance by plotting histograms. Testing it on chi square and Gaussian distribution, we simulated a dataset of 60,000 observations.
In our simulations, we chose the transition probability density function $g(x' \mid x_t)$ to be a Gaussian centered at x_t, in this case, we have the equality $g(x' \mid x_t) = g(x_t \mid x')$, simplifying the acceptance ratio to $\alpha = \frac{f(x')}{f(x_t)}$. Additionally, we conducted multiple simulations varying the variance of this transition probability distribution. Graphically, we observed undesired effects when the variance was too high or too low, or when the initial sample x_0 was far from the mean.

[1] https://github.com/Biechy/LRwithMissingCovariates

Finally, when the variance was egal to 1, the results on Figure 1 were quite satisfactory, demonstrating the algorithm's effectiveness in approximating the target distributions.

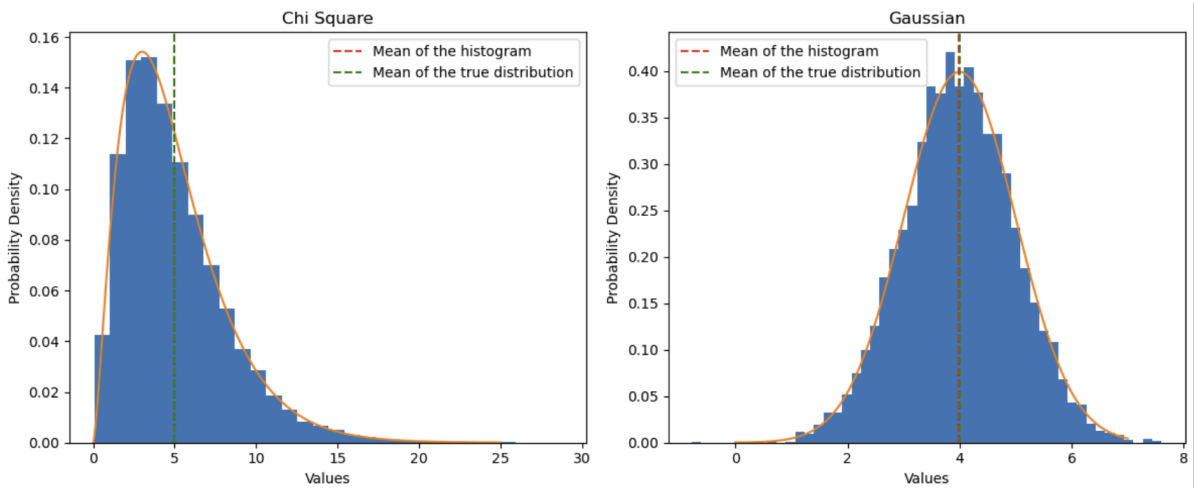

Figure 1. Visualization of Metropolis-Hastings Algorithm Performance

2.3.2. Variance of $\hat{\theta}_{\text{SAEM}}$

In the paper, the properties of the variance of the SAEM estimator $\hat{\theta}_{\text{SAEM}}$ are directly derived from the observed Fisher information $\mathcal{I}$. However, it is important for us to understand why this is correct. Therefore, we will try to briefly explain why $\hat{\theta}_{\text{SAEM}}$ is asymptotically efficient. Firstly, it is worth noting that the almost sure convergence of the SAEM algorithm to a (local) maximum likelihood, under mild assumptions, is demonstrated in several works [5–7].

Secondly, it should be recalled that under certain regularity assumptions on the likelihood and for a regular model, the maximum likelihood estimator $\hat{\theta}_{\text{MLE}}$ is consistent and asymptotically Gaussian by the delta method:

$$\sqrt{n}\left(\hat{\theta}_{\text{MLE}} - \theta_{\text{MLE}}\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \mathcal{I}(\theta_{\text{MLE}})^{-1}\right)$$

Combining these two results, we conclude that $\hat{\theta}_{\text{SAEM}}$ is also asymptotically efficient. Note that this might have been different if the model had not been Gaussian [8]. This efficiency is not strictly that of the Cramer-Rao bound because without knowing the exact likelihood, it is impossible to compute the Fisher information. Thus, the objective is to estimate it to approach the efficiency definition of the Cramer-Rao bound as closely as possible. To achieve this in the paper, Louis' formula [9] is used to decompose the likelihood.

$$\mathcal{I}(\theta) = \mathbb{E}\left(\frac{\partial^2 \mathcal{LL}(\theta; x, y)}{\partial\theta\partial\theta^T}|x_{\text{obs}}, y; \theta\right)\mathbb{E}\left(\frac{\partial \mathcal{LL}(\theta; x, y)}{\partial\theta}\frac{\partial \mathcal{LL}(\theta; x, y)^T}{\partial\theta}|x_{\text{obs}}, y; \theta\right)$$
$$+\mathbb{E}\left(\frac{\partial \mathcal{LL}(\theta; x, y)}{\partial\theta}|x_{\text{obs}}, y; \theta\right)\mathbb{E}\left(\frac{\partial \mathcal{LL}(\theta; x, y)}{\partial\theta}|x_{\text{obs}}, y; \theta\right)^T$$

And the author performs a Monte Carlo estimation by drawing S samples according to the previously specified conditional distribution (**décrit**).

$$\hat{\mathcal{I}}_S\left(\hat{\theta}_{\text{SAEM}}\right) = \sum_{i=1}^{n}\left[-\frac{1}{S}\sum_{s=1}^{S}\frac{\partial^2 \mathcal{LL}\left(\hat{\theta}; x_{i,\text{mis}}^{(s)}, x_{i,\text{obs}}, y_i\right)}{\partial\theta\partial\theta^T}\right.$$
$$-\frac{1}{S}\sum_{s=1}^{S}\left(\frac{\partial \mathcal{LL}\left(\hat{\theta}; x_{i,\text{mis}}^{(s)}, x_{i,\text{obs}}, y_i\right)}{\partial\theta}\right)\left(\frac{\partial \mathcal{LL}\left(\hat{\theta}; x_{i,\text{mis}}^{(s)}, x_{i,\text{obs}}, y_i\right)}{\partial\theta}\right)^T$$
$$\left.+\frac{1}{S^2}\sum_{s=1}^{S}\frac{\partial \mathcal{LL}\left(\hat{\theta}; x_{i,\text{mis}}^{(s)}, x_{i,\text{obs}}, y_i\right)}{\partial\theta}\sum_{s=1}^{S}\frac{\partial \mathcal{LL}\left(\hat{\theta}; x_{i,\text{mis}}^{(s)}, x_{i,\text{obs}}, y_i\right)}{(\partial\theta)^T}\right]$$

This method can be computed algorithmically because the gradient and the Hessian are in closed form.

2.4. Model selection

The framework proposes a model selection method based on penalized log-likelihood. As a reminder, the criteria of this type for a model $\mathcal{M}$ and a maximum likelihood estimator $\hat{\theta}_{\mathcal{M}}$ are generally written as follows:

$$-2\mathcal{LL}\left(\hat{\theta}_{\mathcal{M}}; x, y\right) + \text{pen}(\mathcal{M})$$

The paper suggests to adapt the Bayesian Information Criterion (BIC) for missing data as follows:

$$\text{BIC}(\mathcal{M}) = -2\log\mathcal{LL}\left(\hat{\theta}_{\mathcal{M}}; x_{\text{obs}}, y\right) + \log(n)d(\mathcal{M})$$

Here, $d(\mathcal{M})$ represents the number of estimated parameters in the model $\mathcal{M}$. Under the assumption that all regression models follow the same distribution $\mathcal{N}_p(\mu, \Sigma)$, $d(\mathcal{M})$ is equivalent to the sum of a constant (the number of non-zero parameters of fixed μ and Σ) and a counting function representing the number of non-zero β in the model. Thus, the difficulty lies in calculating $\mathcal{LL}\left(\hat{\theta}_{\mathcal{M}}; x_{\text{obs}}, y\right)$. As a reminder, in the case of missing data for a model $\mathcal{M}$ and a given parameter $\theta_{\mathcal{M}}$, the log-likelihood is written as follows:

$$\mathcal{LL}(\theta_{\mathcal{M}}; x_{\text{obs}}, y) = \sum_{i=1}^{n}\log\left(p\left(y_i, x_{i,\text{obs}}; \theta_{\mathcal{M}}\right)\right)$$

However, the density $p(y_i, x_{i,\text{obs}}; \theta_{\mathcal{M}})$ cannot be expressed in closed form due to missing data. To overcome this problem, the paper uses a Monte Carlo approach with g_i the density defined by Equation 2:

$$\begin{aligned} p(y_i, x_{i,\text{obs}}; \theta_{\mathcal{M}}) &= \int p(y_i, x_{i,\text{obs}} | x_{i,\text{mis}}; \theta_{\mathcal{M}}) p(x_{i,\text{mis}}; \theta_{\mathcal{M}}) \mathrm{d}x_{i,\text{mis}} \\ &= \int p(y_i, x_{i,\text{obs}} | x_{i,\text{mis}}; \theta_{\mathcal{M}}) \frac{p(x_{i,\text{mis}}; \theta_{\mathcal{M}})}{g(x_{i,\text{mis}})} g(x_{i,\text{mis}}) \mathrm{d}x_{i,\text{mis}} \\ &= \mathbb{E}_{g_i} \left[p(y_i, x_{i,\text{obs}} | x_{i,\text{mis}}; \theta_{\mathcal{M}}) \frac{p(x_{i,\text{mis}}; \theta_{\mathcal{M}})}{g(x_{i,\text{mis}})} \right] \end{aligned}$$

If S samples are drawn from the mentioned distribution Equation 2, then:

$$x_{i,\text{mis}}^{(s)} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_i, \Sigma_i), \quad s = 1, 2, ..., S$$

And $p(y_i, x_{i,\text{obs}} | x_{i,\text{mis}}; \theta_{\mathcal{M}})$ can be estimated by:

$$\hat{p}(y_i, x_{i,\text{obs}}; \theta_{\mathcal{M}}) = \frac{1}{S} \sum_{s=1}^{S} p\Big(y_i, x_{i,\text{obs}} | x_{i,\text{mis}}^{(s)}; \theta_{\mathcal{M}}\Big) \frac{p\Big(x_{i,\text{mis}}^{(s)}; \theta_{\mathcal{M}}\Big)}{g\Big(x_{i,\text{mis}}^{(s)}\Big)}$$

Thus $\mathcal{LL}\Big(\hat{\theta}_{\mathcal{M}}; x_{\text{obs}}, y\Big)$ can be deduced as:

$$\sum_{i=1}^{n} \log\big(\hat{p}(y_i, x_{i,\text{obs}}; \theta_{\mathcal{M}})\big) = \sum_{i=1}^{n} \log\left(\frac{1}{S} \sum_{s=1}^{S} p\Big(y_i, x_{i,\text{obs}} | x_{i,\text{mis}}^{(s)}; \theta_{\mathcal{M}}\Big) \frac{p\Big(x_{i,\text{mis}}^{(s)}; \theta_{\mathcal{M}}\Big)}{g\Big(x_{i,\text{mis}}^{(s)}\Big)} \right)$$

We remain sceptical about the practical computation time of this method, as the number of models grows polynomially with $\text{len}(\beta)$, resulting in a large number of samples in total, even BIC computations can be parallelized in practice [10].

2.5. Prediction on test with missing values

Estimating the predictive performance of our model is crucial, yet it is an area that remains underexplored in current literature. Thus, this constitutes the final part of the logistic regression framework. Assume that the training and test sets share a similar distribution with missing data in both. The methodology outlined in the article provides a natural approach to addressing this issue by marginalizing the distribution of missing data with respect to the observed data.
By performing S Monte Carlo samples with $\Big(x_{\text{mis}}^{(s)}, 1 \leq s \leq S\Big) \sim p(x_{\text{mis}} | x_{\text{obs}})$, we then directly obtain the classification by maximum *a posteriori* (MAP) :

$$\begin{aligned} \widehat{y} &= \arg \max_{y \in \{0,1\}^n} p(y | x_{\text{obs}}) = \arg \max_{y \in \{0,1\}^n} \int p(y|x) p(x_{\text{mis}} | x_{\text{obs}}) \mathrm{d}x_{\text{mis}} \\ &= \arg \max_{y \in \{0,1\}^n} \mathbb{E}_{p(x_{\text{mis}} | x_{\text{obs}})}[p(y|x)] = \arg \max_{y \in \{0,1\}^n} \sum_{s=1}^{S} p\Big(y | x_{\text{obs}}, x_{\text{mis}}^{(s)}\Big) \end{aligned}$$

3. Simulations

3.1. Missing clompletly at random

We will now attempt to verify the performance of the framework outlined in this paper, which is contained in the *misaem* package[2], through an implementation in R[3]. To do this, we will adopt the configuration described in the paper. Let x be the design matrix of size $n = 10000$ and $p = 5$, generated by drawing from a normal distribution $\mathcal{N}(\mu, \Sigma)$, to which we associate the target using Equation 1. We then consider the following parameters: $\beta = (-0.2, 0.5, -0.3, 1, 0, -0.6)$, $\mu = (1, 2, 3, 4, 5)$, $\Sigma = \text{diag}(\sigma).C.\text{diag}(\sigma)$ where $\sigma = (1, 2, 3, 4, 5)$ and C is the correlation matrix defined as follows:

$$C = \begin{pmatrix} 1 & 0.8 & 0 & 0 & 0 \\ 0.8 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0.3 & 0.6 \\ 0 & 0 & 0.3 & 1 & 0.7 \\ 0 & 0 & 0.6 & 0.7 & 1 \end{pmatrix}$$

To introduce missing data, we randomly select indices from the design matrix according to a uniform distribution and convert them to NA, thereby creating a mechanism for Missing Completely At Random (MCAR).
Subsequently, we compare SAEM with the complete case (CC), as well as with other methods such as mean imputation (to be avoided as discussed in class) and the MICE method. We estimate the vector β, calculate the observed log-likelihoods, and evaluate two well-known classification metrics: Area Under the Curve (AUC) and F1 score. This is summarized in the following table:

	$\hat{\beta}$	$\mathcal{LL}(\theta; x_{\text{obs}}, y)$	AUC	F1 score
CC	$(-0.2, 0.5, -0.3, 0.9, 0, -0.6)$	-2408	0.78	0.719
SAEM	$(-0.1, 0.4, -0.2, 0.9, 0, -0.5)$	-2802	0.75	0.667
Mean	$(0.1, 0.3, -0.1, 0.5, 0, -0.3)$	-2938	0.74	0.656
MICE	$(0.5, 0.1, -0.1, 0.2, 0, -0.1)$	-3437	0.67	0.516

A significant improvement is observed in all columns with the SAEM algorithm, confirming its suitability for the MCAR logistic regression problem. It is worth noting that the MICE algorithm performs surprisingly poorly in all cases. We

[2]https://cran.r-project.org/web/packages/misaem/misaem.pdf
[3]https://github.com/Biechy/LRwithMissingCovariates

acknowledge that we do not fully understand the reason behind this, which may stem from a coding error.
Additionally, we conducted model selection in the case of a sparse β, and despite the missing data, only the significant β_i were selected without error. As expected, beware of the relatively long execution time, even for small datasets. But as expected, beware of the relatively long execution time, even for small datasets.

3.2. Missing at random

Since we lacked access to the TraumaBase database, which is supposed to adhere to an MAR model, it was crucial for us to assess the framework's efficacy in handling MAR scenarios. In order to do that, we adopt the same configuration as before, but this time adjusting the method of generating missing data. We now select indices of rows where the first column is below a threshold, here 2. Subsequently, for each index, we introduce a Bernoulli random variable with a probability of 0.8 to decide whether to delete data from the second column. We repeat this process for columns 5 and 4, with a threshold of 4 and the same probability. Thus, we obtain a mechanism of Missing At Random (MAR). We chose columns 1, 2, and 5, 4 due to their strong correlation. Furthermore, we had to adjust the vector β to better account for this correlation. Now, $\beta = (-1, 0.5, -0.3, 1, 0.4, -0.6)$.
As before, we estimate the vector β, calculate the observed log-likelihoods, and evaluate AUC and F1 score metrics. This is summarized in the following table:

	$\widehat{\beta}$	$\mathcal{LL}(\theta; x_{\text{obs}}, y)$	AUC	F1 score
CC	$(-0.9, 0.5, -0.3, 1, 0.4, -0.6)$	-2168	0.79	0.727
SAEM	$(-0.5, 0.4, -0.3, 0.9, 0.3, -0.6)$	-2549	0.77	0.693
Mean	$(-0.9, 0.4, -0.2, 0.7, 0.3, -0.4)$	-2556	0.77	0.697
MICE	$(-0.2, 0.3, 0, 0.6, 0, -0.3)$	-2850	0.75	0.664

A significant improvement is observed in all columns with the SAEM algorithm, confirming its suitability for the MAR logistic regression problem. Again, it is worth noting that the MICE algorithm performs surprisingly poorly in all cases. Additionally, we conducted model selection in the case of a sparse β, and despite the missing data, only the significant β_i were selected without error.

4. Discussion

This paper presents a joint-modeling framework for logistic regression in the presence of missing data of MAR type. It outlines a method for estimating the likelihood of observed data, as well as selection and prediction models. After confirming the effectiveness of this framework through simulations, an in-depth study was conducted using the TraumaBase database. While the paper provides a solid foundation for motivating improvements in logistic regression, some readers, like us, may not fully grasp its objective, which appears to focus on synthesizing and implementing existing algorithms to create an R package without really introducing significant research innovations in the context of logistic regression. Although the presentation is clear, we were disappointed by the choice to only simulate missing data mechanisms of the MCAR type for verifying the method's effectiveness, and by the relatively succinct description of some methods. Nevertheless, this paper remains effective and enjoyable to read overall.

References

1. Jiang, W., Josse, J., Lavielle, M., Group, T., others: Logistic regression with missing covariates—Parameter estimation, model selection and prediction within a joint-modeling framework. Computational Statistics & Data Analysis. 145, 106907–106908 (2020)
2. Wei, G. C., Tanner, M. A.: A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. Journal of the American statistical Association. 85, 699–704 (1990)
3. Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R.: Monte Carlo EM for missing covariates in parametric regression models. Biometrics. 55, 591–596 (1999)
4. Lavielle, M.: Mixed effects models for the population approach: models, tasks, methods and tools. CRC press (2014)
5. Delyon, B., Lavielle, M., Moulines, E.: Convergence of a stochastic approximation version of the EM algorithm. Annals of statistics. 94–128 (1999)
6. Kuhn, E., Lavielle, M.: Coupling a stochastic approximation version of EM with an MCMC procedure. ESAIM: Probability and Statistics. 8, 115–131 (2004)
7. Allassonnière, S., Kuhn, E., Trouvé, A.: Construction of bayesian deformable models via a stochastic approximation algorithm: a convergence study. (2010)
8. Jennrich, R. I.: Asymptotic properties of non-linear least squares estimators. The Annals of Mathematical Statistics. 40, 633–643 (1969)
9. Louis, T. A.: Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society Series B: Statistical Methodology. 44, 226–233 (1982)
10. Giraud, C.: Introduction to high-dimensional statistics. Chapman, Hall/CRC (2021)