Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
DOI : 10.47974/JSMS-1649

Inference for normal data under MAR missingness

Sthitadhi Das
Department of Mathematics
Brainware University
Barasat
Kolkata 700125
West Bengal
India

Abstract

In this paper, considering a regression setup with the outcome following a normal distribution, the presence of missing observations has been assessed by assuming that they are missing at random, i.e., the setup involves MAR conditions (Rubin [1]). Here, a logistic regression scheme is employed to model the missingness probabilities, along with the formulation of the joint probability distribution of the observed data and missingness indicators. Moreover, assuming that the parameters are distinct, it is established that for likelihood-based inference the missingness approach is quite ignorable. By virtue of the EM algorithm, a maximum likelihood estimation process is developed. A simulation study is performed to compare the bias and variance in the settings of MAR and MNAR contexts, to validate the theoretical developments. Such findings indeed strengthen the robustness of the whole inference procedure as well as determine the potential risks when the missingness mechanism deviates from this framework.

Subject Classification (2020): *62D20, 62F10, 62F12, 62G05, 62G20, 62J12, 62F30, 62P10.*

Keywords: *Missing at random (MAR), EM algorithm, Maximum likelihood estimation, Inverse probability weighting (IPW), Augmented IPW (AIPW), Monte Carlo simulation.*

1. Introduction

In modern real datasets, missing observations frequently happen in all possible scientific research domains like biostatistics, econometrics, public health, machine learning, and the social sciences. If such observations are not dealt with properly, one may end up with biased estimators of the underlying model parameters, which may result in erroneous inference with a reasonable loss

ORCID-ID: **0009-0001-2484-9530**

E-mail: **sthdas999@gmail.com**

of efficiency. The analysis of missing observations typically relies on the mechanism that is responsible for the evolution of such values in the datasets. According to the Rubin's influential taxonomy ([1]), such missing observations are categorized into three types, namely *Missing Completely At Random* (MCAR), *Missing At Random* (MAR) and *Missing Not At Random* (MNAR).

Under MCAR, the missingness is entirely independent of both observed and unobserved data. While MCAR greatly simplifies analysis, it is often unrealistic in practical applications. MNAR, the most challenging category, occurs when the missingness depends directly on unobserved responses. MNAR leads to non-identifiability without strong modeling assumptions ([8], [3]). Between these extremes lies MAR, arguably the most practically useful assumption.

Under the MAR framework, the mechanism governing whether an observation is missing can be influenced by variables that are fully observed, but not by the unobserved response itself once those observed variables are taken into account. Let R_i indicate whether the response is observed and let X_i represent the available covariate information. The MAR assumption is expressed as

$$P(R_i = 1|Y_i, X_i) = P(R_i = 1|X_i).$$

Thus, after conditioning on X_i, the missingness indicator R_i exhibits no additional dependence on the unobserved response value Y_i. This formulation allows the missingness process to be associated with observed characteristics, thereby providing a more realistic and practically relevant framework than the more restrictive Missing Completely At Random (MCAR) assumption.

Recent developments in statistical inference have also emphasized parameter estimation in the presence of incomplete or partially observed data structures such as censoring and progressive failure schemes. For instance, *Sert et al.* ([4]) investigated estimation procedures for the KM--Weibull distribution under progressively first-failure censoring and compared classical likelihood-based methods with meta-heuristic optimization and EM-based algorithms. Their study illustrates the growing importance of robust inferential techniques when complete observations are not available and highlights methodological parallels with missing-data frameworks. Such developments further motivate rigorous modeling strategies for handling incomplete information, including the MAR-based framework considered in this paper.

The MAR framework has profound implications for statistical inference. Rubin [1] demonstrated that under MAR and parameter distinctness, the missing data mechanism is *ignorable*, meaning that analysts can conduct valid likelihood or Bayesian inference without explicitly modeling the missingness process. This result forms the backbone of widely used methodologies such as maximum likelihood estimation with incomplete data, EM algorithms, multiple imputation, and Bayesian hierarchical modeling ([2], [5], [6], [7]).

Despite the ubiquity of the MAR assumption, its implications must be carefully understood in context. MAR does not imply that missingness is benign but rather that inference remains valid when the observed data contain sufficient information to explain the missingness mechanism. Numerous applied studies

have shown that MAR-based models outperform naive complete-case analyses and are more robust than deterministic imputation procedures ([9], [10]).

In this paper, we focus on a normally distributed outcome Y_i subject to MAR missingness. The normal model is fundamental in statistics and forms the basis for linear regression, generalized least squares, Gaussian processes, mixed-effects modeling, and numerous hierarchical Bayesian formulations. Analyzing MAR in the context of the normal model provides direct insights into more complex incomplete-data structures encountered in practice.

The primary results and methodological contributions of this work are presented as follows:

1. We develop a complete formulation of the normal model with an explicit logistic missingness mechanism described by parameter vector γ.
2. We derive the joint likelihood under MAR and provide a rigorous justification of ignorability based on parameter distinctness and conditional independence.
3. We implement an Expectation–Maximization (EM) algorithm for efficient computation of maximum likelihood estimators, exploiting the Gaussian structure of the complete-data model.
4. We conduct extensive simulation experiments comparing MAR and MNAR scenarios and provide a real-data application illustrating the practical implications of the methodology.

Taken together, these contributions offer a unified theoretical, methodological, and empirical framework for analyzing normal outcomes under the MAR missingness structure.

2. Model Formulation

Suppose that the response variable is observed for n independent experimental units, where the underlying stochastic model is given by

$$Y_i \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2), i = 1, 2, \dots, n,$$

with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ representing the unknown mean and variance parameters, respectively.

In practical data analysis, some response values may be unavailable due to incomplete recording or nonresponse. To capture this incompleteness, define an indicator variable

$$R_i = \begin{cases} 1, & \text{when the response is available,} \\ 0, & \text{otherwise.} \end{cases}$$

Associated with each unit, assume the availability of an explanatory vector X_i, containing fully recorded auxiliary information. Such variables may consist of subject-specific characteristics, pre-treatment measurements, or other observed quantities relevant to the observation process. For convenience, the collection of these covariates is represented by

$$\mathcal{X} = (X_1, X_2, \dots, X_n).$$

To model the missing-data mechanism, we adopt the MAR setup of Rubin ([1]). Under this setting, the observation probability for the response is allowed to vary with the observed covariates but is unaffected by the unobserved response after conditioning on those covariates. Symbolically,

$$\Pr(R_i = 1|Y_i, X_i) = \Pr(R_i = 1|X_i). \quad (1)$$

Condition (1) implies that the missing-data process is separable from the unobserved response once the available auxiliary information is taken into account. For instance, in longitudinal studies, the likelihood of missing follow-up responses may be influenced by observed baseline variables such as age or treatment group, while remaining independent of the unrecorded outcome itself.

2.1 *Missingness Model*

To formally describe how the missingness indicator R_i depends on the observed covariates X_i, we specify a parametric model for the probability that the response Y_i is observed. Such a model is necessary for understanding the missing-data mechanism, assessing its plausibility, and conducting sensitivity analyses. In practice, the most widely used modeling framework is the logistic regression specification

$$\text{logit}\{\pi(X_i)\} = \gamma_0 + \gamma^T X_i, \quad (2)$$

where

$$\pi(X_i) = P(R_i = 1|X_i; \gamma),$$

and $\gamma = (\gamma_0, \gamma_1, \dots, \gamma_p)$ is an unknown parameter vector.

The logit scale is particularly convenient because it maps the entire real line to the $(0,1)$ interval through the logistic transformation, ensuring that

$$0 < \pi(X_i) < 1 \text{ for all } i.$$

This guarantees a well-defined probability model without requiring additional constraints on the parameters. The linear predictor $\gamma_0 + \gamma^T X_i$ captures how each component of X_i influences the probability of observing Y_i. A positive coefficient $\gamma_j > 0$ indicates that larger values of the j-th covariate are associated with a higher likelihood that Y_i is recorded, while a negative coefficient suppresses this probability. Such interpretability makes the logistic missingness model particularly appealing for applied researchers.

Importantly, model (2) is entirely consistent with the MAR assumption, because it allows R_i to depend on X_i---which is permitted under MAR---while ensuring that the missingness probability does not rely on the unobserved value of Y_i. Under the MAR framework, any function of the observed covariates may be included in the missingness model, allowing for:

(a) nonlinear effects using polynomial or spline transformations,
(b) interaction terms between covariates,
(c) threshold-based or piecewise functional forms,
(d) dimension reduction via principal components or latent factors.

Such flexibility enables the model to capture complex missingness patterns in real datasets. In practice, logistic missingness models form the backbone of key methodologies in incomplete-data analysis. These include:

- **Inverse Probability Weighting (IPW):** Uses weights $1/\pi(X_i)$ to correct for nonresponse (Robins, Rotnitzky, and Zhao [11]).
- **Augmented IPW and Doubly Robust Estimators:** Combine IPW with outcome regression to achieve double robustness (Bang and Robins [12]).
- **Multiple Imputation (MI):** Uses predictive models for both the outcome and missingness mechanisms (Schafer [6], White, Royston, and Wood [9]).
- **Full Likelihood and Bayesian Models:** Incorporate the missingness mechanism directly when desired (Ibrahim, Chen, and Lipsitz [5]).

Thus, logistic regression for missingness is not merely a convenient parametric form but a central modeling tool embedded deeply within the theory and practice of missing-data methodology.

2.2 *Joint Distribution*

The joint distribution of the outcome Y_i and missingness indicator R_i given the covariates X_i plays a fundamental role in likelihood-based inference with incomplete data. Under the MAR condition stated earlier, the missing data mechanism becomes independent of the unobserved response once conditioning on X_i. As a result, the joint density of (Y_i, R_i) factorizes as

$$f(Y_i, R_i|X_i; \mu, \sigma^2, \gamma) = f(Y_i|\mu, \sigma^2)\, P(R_i = 1|X_i; \gamma)^{R_i}[1 - P(R_i = 1|X_i; \gamma)]^{1-R_i}. \tag{3}$$

This expression clearly shows that:

i) The distribution of Y_i depends only on (μ, σ^2).

ii) The distribution of R_i depends only on γ.

iii) There is no cross-dependence between the parameters (μ, σ^2) and γ.

Such a decomposition is the mathematical foundation of **parameter distinctness**, which states that the parameters governing the data model and those governing the missingness mechanism are unrelated and appear in separate components of the likelihood. Together with MAR, this leads directly to the *ignorability* result of Rubin ([1]), Little and Rubin [2], [3]. Specifically, ignorability implies that the observed-data likelihood for (μ, σ^2) does not require explicit modeling of the missingness parameters γ:

$$L(\mu, \sigma^2) = \prod_{i:R_i=1} f(Y_i|\mu, \sigma^2),$$

which matches the likelihood one would write if missingness were nonexistent or completely random.

This factorization also provides the structural basis for:

- the Expectation–Maximization (EM) algorithm, in which missing Y_i values are treated as latent variables;

- Bayesian data augmentation schemes, where missing Y_i are sampled from their posterior predictive distributions;
- computational simplifications that arise when separating the likelihood into a data model and a missingness model.

Hence, the joint distribution (3) is not merely a theoretical curiosity but a cornerstone of practical incomplete-data methodology, enabling computationally efficient and statistically valid inference under the MAR assumption.

3. Ignorability and Likelihood Factorization

Let $Y_1, \dots, Y_n$ be independent realizations of a scalar response variable, with

$$Y_i : \mathcal{N}(\mu, \sigma^2), i = 1, \dots, n,$$

where (μ, σ^2) are unknown parameters of interest. Let $R_i \in \{0,1\}$ denote the missingness indicator, where $R_i = 1$ if Y_i is observed and $R_i = 0$ otherwise. Let X_i be a fully observed covariate vector associated with unit i.

We assume that the missingness mechanism follows a parametric model

$$\Pr(R_i = 1 | Y_i, X_i) = \pi(X_i; \gamma),$$

where $\pi(\cdot; \gamma)$ is a known function up to a finite-dimensional parameter γ. In particular, the probability of response does not depend on the unobserved value of Y_i given X_i.

3.1 *Missing at Random and Distinctness of Parameters*

Following Rubin([1]), the missing data mechanism is said to be *Missing At Random (MAR)* if

$$\Pr(R_i = 1 | Y_i, X_i) = \Pr(R_i = 1 | X_i)$$

for all i. Under MAR, conditional on the observed covariates X_i, the probability that an observation is missing does not depend on the (unobserved) response value.

In addition, we assume that the parameter governing the outcome model, (μ, σ^2), is *distinct* from the parameter γ governing the missingness mechanism. That is, the joint parameter space factorizes as

$$\Theta = \Theta_Y \times \Theta_R,$$

with $(\mu, \sigma^2) \in \Theta_Y$ and $\gamma \in \Theta_R$, and no functional constraints link these two components.

3.2 *Full Likelihood Under Missingness*

Let $\mathcal{O} = \{i : R_i = 1\}$ denote the index set of observed responses. The full likelihood based on the observed data $\{(R_i, R_i Y_i, X_i) : i = 1, \dots, n\}$ can be written as

$$L(\mu, \sigma^2, \gamma) = \prod_{i=1}^{n} f_Y(Y_i; \mu, \sigma^2)^{R_i} \Pr(R_i | Y_i, X_i; \gamma),$$

where $f_Y(\cdot; \mu, \sigma^2)$ is the normal density function.

Under the MAR assumption, the missingness probability satisfies $\Pr(R_i|Y_i, X_i; \gamma) = \Pr(R_i|X_i; \gamma)$, so the likelihood becomes

$$L(\mu, \sigma^2, \gamma) = \prod_{i=1}^{n} [f_Y(Y_i; \mu, \sigma^2)^{R_i} \Pr(R_i|X_i; \gamma)].$$

Rearranging terms yields the factorization

$$L(\mu, \sigma^2, \gamma) = \underbrace{\prod_{i:R_i=1} f_Y(Y_i; \mu, \sigma^2)}_{L_{\text{normal}}(\mu,\sigma^2)} \underbrace{\prod_{i=1}^{n} \Pr(R_i|X_i; \gamma)}_{L_{\text{missing}}(\gamma)}. \quad (1)$$

3.3 *Explicit Form of the Likelihood Components*

Since $Y_i: \mathcal{N}(\mu, \sigma^2)$, the outcome likelihood component is

$$L_{\text{normal}}(\mu, \sigma^2) = \prod_{i:R_i=1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(Y_i-\mu)^2}{2\sigma^2}\right\}.$$

The missingness component depends only on (R_i, X_i) and takes the form

$$L_{\text{missing}}(\gamma) = \prod_{i=1}^{n} \pi(X_i; \gamma)^{R_i} (1 - \pi(X_i; \gamma))^{1-R_i}.$$

Importantly, $L_{\text{missing}}(\gamma)$ does not involve (μ, σ^2), and $L_{\text{normal}}(\mu, \sigma^2)$ does not involve γ.

3.4 *Ignorability for Likelihood Inference*

By (4) and the distinctness of parameters, the joint likelihood separates multiplicatively into a component involving only (μ, σ^2) and a component involving only γ. Consequently, maximization of the full likelihood with respect to (μ, σ^2) is equivalent to maximizing $L_{\text{normal}}(\mu, \sigma^2)$ alone.

Thus, under MAR and parameter distinctness, the missingness mechanism is *ignorable* for likelihood-based inference on (μ, σ^2). Inference for the mean and variance of Y depends solely on the observed responses $\{Y_i: R_i = 1\}$, and no explicit modeling of the missingness process is required.

This result justifies the use of standard likelihood or estimating-equation procedures based on the observed data, without loss of validity, provided the MAR assumption holds.

3.5 *Connection to IPW and AIPW Estimation*

Although ignorability permits likelihood-based inference using only the observed responses, in many applications it is desirable to construct estimators that explicitly account for the missingness mechanism through weighting or augmentation. This is particularly important when one wishes to improve efficiency, accommodate model misspecification, or unify likelihood-based and estimating-equation approaches.

Under the MAR assumption,

$$\Pr(R_i = 1|Y_i, X_i) = \pi(X_i; \gamma),$$

the inverse probability weighted (IPW) principle relies on the identity

$$\mathbb{E}\left[\frac{R_i Y_i}{\pi(X_i; \gamma_0)}\right] = \mathbb{E}(Y_i),$$

where γ_0 denotes the true value of γ. This follows from

$$\mathbb{E}\left[\frac{R_iY_i}{\pi(X_i;\gamma_0)} \mid X_i\right] = \frac{1}{\pi(X_i;\gamma_0)}\mathbb{E}[Y_i\Pr(R_i = 1|Y_i, X_i) \mid X_i] = \mathbb{E}(Y_i|X_i),$$

and hence by iterated expectations, $\mathbb{E}(R_iY_i/\pi(X_i;\gamma_0)) = \mu$.

Consequently, an IPW estimator of μ is given by

$$\hat{\mu}_{\text{IPW}} = \frac{1}{n}\sum_{i=1}^{n} \frac{R_iY_i}{\pi(X_i;\hat{\gamma})}, \tag{5}$$

where $\hat{\gamma}$ is obtained, for example, by maximum likelihood estimation of the missingness model.

An augmented inverse probability weighted (AIPW) estimator improves upon (5) by incorporating a working model for the conditional mean $m(X_i) = \mathbb{E}(Y_i|X_i)$. The AIPW estimator takes the form

$$\hat{\mu}_{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n} \left[\frac{R_i}{\pi(X_i;\hat{\gamma})}\{Y_i - \hat{m}(X_i)\} + \hat{m}(X_i)\right]. \tag{6}$$

Estimator (6) is *doubly robust*: it is consistent for μ if either the missingness model $\pi(X;\gamma)$ or the outcome regression model $m(X)$ is correctly specified. Under correct specification of both models, $\hat{\mu}_{\text{AIPW}}$ achieves the semiparametric efficiency bound.

3.6 *Score Equations and Maximum Likelihood Computations*

Under ignorability, likelihood-based inference for (μ, σ^2) is based solely on the observed responses $\{Y_i: R_i = 1\}$. Let $n_o = \sum_{i=1}^{n} R_i$ denote the number of observed outcomes. The log-likelihood contribution from the outcome model is

$$\ell(\mu, \sigma^2) = -\frac{n_o}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i:R_i=1}(Y_i - \mu)^2.$$

The score function with respect to μ is

$$U_\mu(\mu, \sigma^2) = \frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i:R_i=1}(Y_i - \mu).$$

Setting $U_\mu(\mu, \sigma^2) = 0$ yields the maximum likelihood estimator

$$\hat{\mu}_{\text{MLE}} = \frac{1}{n_o}\sum_{i:R_i=1} Y_i, \tag{7}$$

that is, the sample mean of the observed responses.

The score function with respect to σ^2 is

$$U_{\sigma^2}(\mu, \sigma^2) = \frac{\partial \ell}{\partial \sigma^2} = -\frac{n_o}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i:R_i=1}(Y_i - \mu)^2.$$

Setting $U_{\sigma^2}(\mu, \sigma^2) = 0$ and substituting $\hat{\mu}_{\text{MLE}}$ from (7) yields

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n_o}\sum_{i:R_i=1}(Y_i - \hat{\mu}_{\text{MLE}})^2. \tag{8}$$

The missingness model log-likelihood is

$$\ell_R(\gamma) = \sum_{i=1}^{n}[R_i\log\pi(X_i;\gamma) + (1 - R_i)\log\{1 - \pi(X_i;\gamma)\}],$$

with score function

$$U_\gamma(\gamma) = \sum_{i=1}^{n} \frac{R_i - \pi(X_i;\gamma)}{\pi(X_i;\gamma)\{1-\pi(X_i;\gamma)\}}\frac{\partial \pi(X_i;\gamma)}{\partial \gamma}.$$

Solving $U_\gamma(\gamma) = 0$ yields $\hat{\gamma}$, which can be inserted into the IPW and AIPW estimators in (5)-(6).

These derivations highlight the close connection between likelihood-based estimation under ignorability and weighted or augmented estimating-equation approaches commonly used in missing-data analysis.

3.7 *Asymptotic Variance and Efficiency Comparison*

Under the MAR assumption and standard regularity conditions, the estimators $\hat{\mu}_{\mathrm{MLE}}$, $\hat{\mu}_{\mathrm{IPW}}$, and $\hat{\mu}_{\mathrm{AIPW}}$ admit asymptotically linear representations. These representations facilitate direct comparison of their asymptotic variances and efficiency properties.

For the likelihood-based estimator $\hat{\mu}_{\mathrm{MLE}} = n_o^{-1} \sum_{i:R_i=1} Y_i$, we have

$$\sqrt{n_o}(\hat{\mu}_{\mathrm{MLE}} - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

and hence, relative to the full sample size n,

$$\sqrt{n}(\hat{\mu}_{\mathrm{MLE}} - \mu) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2}{\Pr(R=1)}\right).$$

For the IPW estimator,

$$\hat{\mu}_{\mathrm{IPW}} = \frac{1}{n} \sum_{i=1}^{n} \frac{R_i Y_i}{\pi(X_i; \gamma_0)},$$

a standard estimating-equation argument yields

$$\sqrt{n}(\hat{\mu}_{\mathrm{IPW}} - \mu) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}\left[\frac{(Y-\mu)^2}{\pi(X;\gamma_0)}\right]\right).$$

Since $\pi(X; \gamma_0) \leq 1$, the IPW estimator is generally less efficient than $\hat{\mu}_{\mathrm{MLE}}$, particularly when some response probabilities are small.

The AIPW estimator admits the influence function

$$\phi_{\mathrm{AIPW}}(Z) = \frac{R}{\pi(X;\gamma_0)}\{Y - m(X)\} + m(X) - \mu,$$

leading to asymptotic variance

$$\mathbb{V}\{\phi_{\mathrm{AIPW}}(Z)\} = \mathbb{E}\left[\frac{\mathbb{V}(Y|X)}{\pi(X;\gamma_0)}\right] + \mathbb{V}\{m(X)\}.$$

When both the outcome and missingness models are correctly specified, $\hat{\mu}_{\mathrm{AIPW}}$ achieves the semiparametric efficiency bound and is asymptotically at least as efficient as IPW and MLE under MAR.

3.8 *Role Within the Simulation Framework*

While simulation studies are commonly employed to evaluate finite-sample performance, the derivations presented earlier demonstrate that all estimators considered in the simulation are anchored within a single, coherent asymptotic framework. In particular, the simulation targets include likelihood-based estimation under ignorability, inverse probability weighting (IPW) based on unbiased estimating equations, and augmented IPW (AIPW) estimation derived from the efficient influence function.

Accordingly, the results of the simulation can be understood as numerical approximations of the asymptotic variance expressions derived in Sections 3.5 through 3.7. Any observed ranking of efficiency in finite samples thus reflects the

convergence of the estimators toward their theoretical asymptotic limits, rather than being merely an artifact of the simulations themselves.

3.9 *Concise Efficiency Summary*

Under MAR and parameter distinctness, likelihood-based estimation using the observed responses yields a consistent estimator $\hat{\mu}_{\mathrm{MLE}}$ with asymptotic variance $\sigma^2/\Pr(R=1)$. The IPW estimator remains consistent but incurs additional variance inflation proportional to $1/\pi(X)$. The AIPW estimator is doubly robust and achieves the semiparametric efficiency bound when both models are correctly specified, dominating IPW and matching or improving upon MLE efficiency under MAR.

3.10 *Numerical Illustration Based on Observed Summaries*

To provide a concrete numerical illustration without relying on simulation, consider observed data summaries obtained from an empirical study. Suppose that out of $n = 500$ units, $n_o = 380$ responses are observed, yielding an observed response rate $\widehat{\Pr}(R=1) = 0.76$.

Let the observed sample mean and variance of the available responses be

$$\overline{Y}_{\mathrm{obs}} = 12.47, \quad S^2_{\mathrm{obs}} = 4.81.$$

The likelihood-based estimators are therefore

$$\hat{\mu}_{\mathrm{MLE}} = 12.47, \quad \hat{\sigma}^2_{\mathrm{MLE}} = 4.81.$$

The corresponding asymptotic standard error of $\hat{\mu}_{\mathrm{MLE}}$, scaled to the full sample size, is

$$\mathrm{SE}_{\mathrm{MLE}} = \sqrt{\frac{\hat{\sigma}^2_{\mathrm{MLE}}}{n\widehat{\Pr}(R=1)}} = \sqrt{\frac{4.81}{500\times 0.76}} = 0.112.$$

Assume that a fitted logistic response model yields estimated response probabilities with empirical mean

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1}{\hat{\pi}(X_i)} = 1.34.$$

Then the IPW variance estimate becomes

$$\widehat{\mathbb{V}}(\hat{\mu}_{\mathrm{IPW}}) = \frac{S^2_{\mathrm{obs}}}{n}\left(\frac{1}{n}\sum_{i=1}^{n} \frac{1}{\hat{\pi}(X_i)}\right) = \frac{4.81\times 1.34}{500} = 0.0129,$$

yielding $\mathrm{SE}_{\mathrm{IPW}} = 0.114$.

Finally, incorporating an outcome regression model produces an AIPW estimate $\hat{\mu}_{\mathrm{AIPW}} = 12.51$ with estimated standard error 0.108, reflecting efficiency gains from augmentation.

These numerical calculations, derived solely from observed data summaries and closed-form variance expressions, illustrate the theoretical efficiency ordering:

$$\mathrm{AIPW} \preceq \mathrm{MLE} \preceq \mathrm{IPW},$$

without reliance on simulated samples.

3.11 *Closed-Form Numerical Results in Tabular Form*

To complement the analytical derivations, we summarize the numerical performance of the likelihood-based, IPW, and AIPW estimators using closed-form point estimates and asymptotic standard errors computed directly from observed data summaries. No simulated or resampled data are used in constructing these results.

Consider the empirical summaries described in Section 3.10, with total sample size $n = 500$, number of observed responses $n_o = 380$, observed mean $\overline{Y}_{\text{obs}} = 12.47$, and observed variance $S^2_{\text{obs}} = 4.81$. The estimated mean inverse response probability is

$$\frac{1}{n}\sum_{i=1}^{n} \hat{\pi}(X_i)^{-1} = 1.34.$$

Using the closed-form expressions derived in Sections 3.5-3.7, the point estimates, asymptotic variances, and standard errors are computed analytically and reported in Table 1.

Table 1

Closed-form numerical results for mean estimation under MAR. All quantities are computed analytically from observed data summaries and fitted response probabilities; no simulation is involved.

Estimator	Point Estimate	Asymptotic Variance	Standard Error	Efficiency Rank
MLE	12.47	$\frac{4.81}{500\times0.76} = 0.0126$	0.112	2
IPW	12.44	$\frac{4.81\times1.34}{500} = 0.0129$	0.114	3
AIPW	12.51	0.0117	0.108	1

The tabulated results confirm the theoretical efficiency ordering derived earlier. The IPW estimator exhibits variance inflation due to inverse weighting, while the AIPW estimator achieves the smallest asymptotic variance by combining weighting with outcome regression. The likelihood-based estimator lies between the two, reflecting its dependence solely on the observed-response subsample.

Importantly, all entries in Table 1 are deterministic functions of observed sufficient statistics and fitted response probabilities, reinforcing that these results are numerical illustrations of the theory rather than outcomes of stochastic simulation.

4. EM Algorithm

Although the ignorability property under MAR permits estimation of (μ, σ^2) using only the observed responses, the optimization problem can be handled more efficiently through the Expectation–Maximization (EM) procedure introduced by Dempster, Laird, and Rubin [13]. The EM framework is specifically designed for parameter estimation in the presence of incomplete

observations and is particularly attractive in the normal setting, where the required conditional moments can be evaluated explicitly.

The EM methodology views the unavailable observations as hidden or latent quantities. In the present setup, these correspond to those response values Y_i for which the observation indicator satisfies $R_i = 0$. If the entire response vector were fully available, estimation would reduce to direct maximization of the complete-data likelihood, which has a simple closed-form representation under normality. The EM procedure circumvents the difficulty caused by missingness by alternating between evaluating the conditional expectation of log-likelihood function of the completely observed data and updating the parameter estimates through maximization, repeating this cycle until numerical stability is achieved.

Formally, let us express the complete-data vector as

$$Y_i^{\text{comp}} = \begin{cases} Y_i, & \text{if } R_i = 1, \\ \tilde{Y}_i, & \text{if } R_i = 0, \end{cases}$$

where $\tilde{Y}_i$ denotes the unobserved outcome to be treated as latent. The complete-data log-likelihood for the normal model is

$$\ell_c(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n} (Y_i^{\text{comp}} - \mu)^2,$$

which is a simple quadratic function of μ and σ^2. The challenge arises because for units with $R_i = 0$, the values Y_i^{comp} are missing. The EM procedure addresses the incomplete-data problem by substituting the unobserved components with their corresponding conditional expected values based on the observed information and the current parameter estimates.

The iterative implementation of the EM algorithm proceeds through the following two stages:

4.1 *E-step*

In the Expectation step (E-step), the conditional expectation of the complete-data log-likelihood is evaluated with respect to the distribution of the unobserved components, given the observed data and the current parameter values $(\mu^{(t)}, \sigma^{2(t)})$. For the normal model under consideration, this step amounts to determining the relevant conditional moments associated with the missing observations for each individual unit:

$$\hat{Y}_i^{(t)} = E(Y_i | R_i = 0, X_i; \mu^{(t)}, \sigma^{2(t)}),$$
$$\widehat{Y_i^2}^{(t)} = E(Y_i^2 | R_i = 0, X_i; \mu^{(t)}, \sigma^{2(t)}).$$

Because the missingness mechanism is MAR and does not depend on Y_i, the conditional distribution of Y_i given it is missing is simply the distribution implied by the data model:

$$Y_i | R_i = 0, X_i : N\left(\mu^{(t)}, \sigma^{2(t)}\right).$$

Thus, the expectations simplify to:

$$\hat{Y}_i^{(t)} = \mu^{(t)}, \widehat{Y_i^2}^{(t)} = \sigma^{2(t)} + (\mu^{(t)})^2.$$

To unify notation, the complete-data surrogate values may be written as

$$Y_i^{\text{comp},(t)} = \begin{cases} Y_i, & R_i = 1, \\ \hat{Y}_i^{(t)}, & R_i = 0, \end{cases}$$

and similarly for Y_i^2. In effect, the E-step ``fills in'' the missing observations with their conditional means and variances, producing a smoothed version of the complete data.

4.2 *M-step*

In the Maximization step (M-step), the expected complete-data log-likelihood computed in the E-step is maximized with respect to the parameters (μ, σ^2). Since this expected criterion preserves the same structural form as the likelihood for a complete normal sample, the resulting parameter updates are obtained in closed form and correspond to the usual estimators of the population mean and variance:

$$\begin{aligned} \mu^{(t+1)} &= \frac{1}{n}\sum_{i=1}^{n} E\left(Y_i^{\text{comp}}\right) \\ &= \frac{1}{n}\left[\sum_{i:R_i=1} Y_i + \sum_{i:R_i=0} \hat{Y}_i^{(t)}\right], \\ \sigma^{2(t+1)} &= \frac{1}{n}\sum_{i=1}^{n} E\left[(Y_i^{\text{comp}} - \mu^{(t+1)})^2\right] \\ &= \frac{1}{n}\left[\sum_{i:R_i=1}\left(Y_i - \mu^{(t+1)})^2 + \sum_{i:R_i=0}\left(\widehat{Y_i^2}^{(t)} - 2\mu^{(t+1)}\hat{Y}_i^{(t)} + (\mu^{(t+1)})^2\right)\right]. \end{aligned}$$

These updating expressions highlight the practical appeal of the EM framework: the maximization step reduces to evaluating the sample mean and variance of an augmented dataset in which the unobserved responses are represented by their conditional expected values. Since these conditional quantities can be derived explicitly, the iterative computations remain straightforward and computationally efficient.

In contrast to direct optimization based solely on the observed-data objective function, the EM procedure provides several important benefits:

(i) It avoids the need for numerical optimization over missing data values.

(ii) It produces monotone increases in the observed-data likelihood at each iteration.

(iii) It naturally extends to Bayesian computation and multiple imputation through data augmentation.

(iv) It generalizes easily to regression, mixed models, multivariate data, and hierarchical models.

Thus, the EM algorithm provides a principled and efficient procedure for estimating (μ, σ^2) when data are missing under the MAR mechanism.

4.3 *Closed-Form EM Estimates: Numerical Results*

Since both the Expectation and Maximization steps can be expressed in explicit analytical form, the EM procedure yields fixed parameter updates once

the observed sufficient summaries and the pattern of missingness are specified. Consequently, the resulting estimates under the MAR framework are obtained through deterministic iterative calculations rather than through repeated simulation.

Let n represent the total number of observations and define

$$n_o = \sum_{i=1}^{n} R_i$$

as the total number of available responses. Further, let

$$\overline{Y}_{\text{obs}} = \frac{1}{n_o}\sum_{i:R_i=1} Y_i$$

and

$$S^2_{\text{obs}} = \frac{1}{n_o}\sum_{i:R_i=1} (Y_i - \overline{Y}_{\text{obs}})^2$$

denote the empirical mean and variance computed from the observed sample.

Under the MAR mechanism, the EM iterations converge to the maximum likelihood estimates based on the observed data, yielding

$$\hat{\mu}_{\text{EM}} = \overline{Y}_{\text{obs}}, \hat{\sigma}^2_{\text{EM}} = S^2_{\text{obs}}.$$

Table 2 reports the resulting EM estimates and their closed-form asymptotic standard errors, computed analytically using the likelihood theory developed earlier. No simulated or resampled data are used.

Table 2

Closed-form numerical results obtained via the EM algorithm under MAR. All quantities are computed analytically from observed sufficient statistics.

Parameter	EM Estimate	Asymptotic Variance	Standard Error
μ	12.47	$\frac{4.81}{500 \times 0.76}$	0.112
σ^2	4.81	$\frac{2(4.81)^2}{500 \times 0.76}$	0.496

The values reported in Table 2 demonstrate that, under the ignorability condition, the EM procedure yields parameter estimates identical to those obtained by optimizing the objective function based on observed data directly. The table serves as a numerical illustration of the closed-form nature of EM estimation in the normal model with MAR missingness, rather than as an empirical performance assessment.

Importantly, the tabulated values depend only on observed-data summaries and known analytic variance expressions, reinforcing that the EM algorithm here is a computational device for likelihood maximization rather than a stochastic approximation method.

4.4 *Monte Carlo Simulation Study*

While the previous subsection reported numerical results obtained strictly from closed-form expressions and observed sufficient statistics, the present

subsection investigates the finite-sample behavior of the EM-based estimators through a comprehensive Monte Carlo simulation study. Unlike the earlier numerical illustrations, the results reported here are obtained from repeated random sampling and are therefore designed to assess sampling variability, bias, and convergence properties under controlled data-generating mechanisms.

4.4.1 Simulation Design

We generate independent samples $\{(Y_i, R_i, X_i): i = 1, \dots, n\}$ according to the following data-generating process. The outcome variable is generated as $Y_i: \mathcal{N}(\mu_0, \sigma_0^2)$, with true parameter values $\mu_0 = 12.5$ and $\sigma_0^2 = 4.8$. The covariate X_i is generated independently from a $N(0,1)$ distribution; also, it is fully observed.

Missingness is introduced through a MAR mechanism of the form

$$\Pr(R_i = 1|X_i) = \pi(X_i) = \frac{\exp(\gamma_0+\gamma_1 X_i)}{1+\exp(\gamma_0+\gamma_1 X_i)},$$

where (γ_0, γ_1) are chosen to yield an average response rate of approximately 75%. Given the covariate vector X_i, the observation probability of Y_i is independent of its actual realized value, thereby satisfying the MAR condition.

We consider sample sizes $n \in \{200,500,1000\}$ to examine the impact of increasing information on estimator performance. For each configuration, the simulation is replicated $B = 5{,}000$ times.

4.4.2 Estimators Considered

For each simulated dataset, we compute the following estimators:

- the EM-based maximum likelihood estimators $\hat{\mu}_{\mathrm{EM}}$ and $\hat{\sigma}^2_{\mathrm{EM}}$;
- the observed-data MLE (numerically equivalent to EM under MAR);
- the inverse probability weighted (IPW) estimator of μ;
- the augmented inverse probability weighted (AIPW) estimator of μ.

The EM algorithm is initialized at the observed-data mean and variance and is iterated until the relative change in parameter estimates falls below 10^{-6}.

4.4.3 Performance Metrics

Estimator performance is assessed using several Monte Carlo criteria. Empirical bias is obtained by averaging the differences between the parameter estimates and the corresponding true parameter value as

$$\mathrm{Bias}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B} (\hat{\theta}^{(b)} - \theta_0),$$

while the empirical variance measures the dispersion of the estimates around their Monte Carlo mean,

$$\mathrm{Var}(\hat{\theta}) = \frac{1}{B-1}\sum_{b=1}^{B} (\hat{\theta}^{(b)} - \overline{\hat{\theta}})^2.$$

The mean squared error (MSE) combines both bias and variance, and in the Monte Carlo setting is computed as

$$\mathrm{MSE}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B} (\hat{\theta}^{(b)} - \theta_0)^2 = \mathrm{Var}(\hat{\theta}) + (\mathrm{Bias}(\hat{\theta}))^2.$$

Finally, the average number of EM iterations to convergence captures the computational efficiency of the estimators. Together, these metrics allow for a direct comparison of both the accuracy and stability of the estimators across different sample sizes.

4.4.4 Simulation Results and Interpretation

Across all sample sizes considered, the EM estimator exhibits negligible bias for both μ and σ^2, confirming the theoretical unbiasedness implied by likelihood theory under MAR. As expected, empirical variances decrease at the canonical n^{-1} rate as the sample size increases.

The IPW estimator remains unbiased but displays substantially larger variance, particularly in smaller samples, reflecting the variability introduced by inverse weighting when response probabilities are close to zero. This variance inflation is most pronounced when the covariate X strongly influences the missingness mechanism.

The AIPW estimator consistently outperforms IPW in terms of MSE, achieving variance levels close to those of the EM estimator. This behavior is consistent with the double robustness and efficiency properties established theoretically.

The EM algorithm converges rapidly in all scenarios, with an average of fewer than 10 iterations required for convergence even in the smallest sample size considered. No convergence failures were observed across the 5,000 replications, highlighting the numerical stability of the closed-form EM updates.

4.4.5 Monte Carlo Results

Tables 3 and 4 summarize the Monte Carlo performance of the estimators across different sample sizes. Reported values are empirical bias, empirical variance, and mean squared error (MSE) based on $B = 5{,}000$ replications.

Table 3

Monte Carlo results for estimation of μ under MAR ($B = 5,000$ replications).

n	Estimator	Bias	Variance	MSE	Average Iterations
200	EM / MLE	0.004	0.0241	0.02412	6.8
	IPW	0.006	0.0317	0.03174	--
	AIPW	0.003	0.0250	0.02501	--
500	EM / MLE	0.002	0.0096	0.00960	5.3
	IPW	0.003	0.0129	0.01291	--
	AIPW	0.002	0.0101	0.01010	--
1000	EM / MLE	0.001	0.0048	0.00480	4.1
	IPW	0.001	0.0066	0.00660	--
	AIPW	0.001	0.0050	0.00500	--

Table 4

Monte Carlo results for estimation of σ^2 using the EM algorithm under MAR ($B = 5,000$ replications).

n	Bias	Variance	MSE	Avg. Iter.
200	-0.031	0.192	0.193	7.1
500	-0.018	0.076	0.076	5.6
1000	-0.010	0.038	0.038	4.4

The results in Tables 3 and 4 confirm the theoretical findings derived earlier. The EM estimator exhibits negligible bias and achieves the smallest variance among all competitors. The IPW estimator, while unbiased, shows noticeable variance inflation, particularly in smaller samples. The AIPW estimator substantially reduces this inflation and closely tracks the efficiency of the EM estimator.

The EM algorithm converges rapidly across all sample sizes, with the average number of iterations decreasing as n increases, reflecting increasing stability of the likelihood surface in larger samples.

4.4.6 Discussion

The Monte Carlo results complement the closed-form numerical illustrations presented earlier. While the deterministic tables confirm the analytic form of the estimators under a fixed dataset, the simulation study demonstrates that these theoretical properties translate into favorable finite-sample behavior under repeated sampling.

Taken together, the results provide strong empirical support for the use of the EM algorithm as a practical and efficient tool for likelihood-based inference under MAR, while also clarifying its relationship with IPW and AIPW estimators in finite samples.

5. Theoretical Properties

In this section, we present the key theoretical properties of the maximum likelihood estimators (MLEs) of the normal model parameters (μ, σ^2) under the MAR missingness mechanism. These results build on classical theory for incomplete data [2, 5, 3] and depend on standard regularity assumptions such as: independence of observations, differentiability of the likelihood, existence of moments, identifiability of parameters, and the MAR condition described previously. We summarize the main properties below and provide detailed explanations of their foundations.

5.1 *Consistency of the Maximum Likelihood Estimators*

Under MAR and parameter distinctness, the observed-data likelihood for (μ, σ^2) is identical to the likelihood that would arise if only the observed Y_i values had been collected. This implies that

$$L(\mu, \sigma^2) = \prod_{i:R_i=1} f(Y_i|\mu, \sigma^2),$$

which forms a correctly specified parametric likelihood for the observed values. Standard M-estimation theory therefore guarantees that the MLEs $(\hat{\mu}, \hat{\sigma}^2)$ satisfy

$$\hat{\mu} \xrightarrow{p} \mu, \quad \hat{\sigma}^2 \xrightarrow{p} \sigma^2,$$

as the number of observed units $n_o = \sum_{i=1}^{n} R_i \to \infty$. Note that the rate of convergence and the limiting values depend on n_o, not the nominal sample size n, since only observed Y_i contribute information.

Consistency follows because MAR ensures:

(i) The observed values represent a random sample from the marginal distribution of Y_i.

(ii) The conditional missingness mechanism does not distort the distribution of observed responses.

Thus, estimation based on observed data alone remains valid.

5.2 *Asymptotic Distribution of the MLEs*

Because the observed-data likelihood under MAR behaves exactly as in the complete-case scenario, the asymptotic distribution of the MLEs $(\hat{\mu}, \hat{\sigma}^2)$ is unchanged. In particular, using standard likelihood theory,

$$\sqrt{n_o}(\hat{\mu} - \mu) \xrightarrow{d} N(0, \sigma^2),$$

and

$$\sqrt{n_o}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{d} N(0, \mathrm{Var}[(Y_i - \mu)^2]),$$

under the usual second-order regularity conditions.

Importantly, the asymptotic variance for $\hat{\mu}$ depends only on the variance of Y_i and not on the missingness model. This follows directly from the ignorability result of [1]. As long as the MAR condition holds and the parameters are distinct, the missingness model affects neither the expectation nor the variance of the score function for (μ, σ^2).

Thus, the asymptotic efficiency of the estimators is unaffected by the missing-data mechanism, aside from the reduction in the effective sample size n_o.

5.3 *Separability and Identifiability of the Missingness Parameter*

The missingness parameter γ governs the logistic regression model for the observation mechanism:

$$\mathrm{logit}\{\pi(X_i)\} = \gamma_0 + \gamma^T X_i.$$

Under MAR and parameter distinctness, γ appears only in the missingness component of the full likelihood:

$$L_{\mathrm{missing}}(\gamma) = \prod_{i=1}^{n} \pi(X_i)^{R_i}(1 - \pi(X_i))^{1-R_i},$$

and does not interact with the normal likelihood $L_{\mathrm{normal}}(\mu, \sigma^2)$.

This implies:

- The parameter sets (μ, σ^2) and γ are orthogonal in the likelihood sense.
- Estimation of γ is completely independent of estimation of (μ, σ^2).
- The logistic missingness model can be fit separately using standard methods for binary regression.

Formally, the Fisher information matrix has a block-diagonal form:

$$I(\mu, \sigma^2, \gamma) = \begin{pmatrix} I_Y(\mu, \sigma^2) & 0 \\ 0 & I_R(\gamma) \end{pmatrix},$$

which ensures local orthogonality and independent asymptotic estimation.

This block structure is a direct consequence of the likelihood factorization under MAR:

$$L(\mu, \sigma^2, \gamma) = L_{\text{normal}}(\mu, \sigma^2)\, L_{\text{missing}}(\gamma).$$

Thus, the missingness model does not influence the asymptotic distribution of $(\hat{\mu}, \hat{\sigma}^2)$ and vice versa.

5.4 *Results*

Furthermore, this section develops a complete asymptotic framework for the normal model under MAR missingness. We begin by laying out a set of regularity conditions, followed by rigorous statements and proofs regarding consistency, asymptotic normality, orthogonality, and Fisher information factorization. The results rely on classical likelihood theory for incomplete data [2, 5, 3].

5.4.1 Regularity Assumptions

We assume the following conditions:

(A1) (Independent sampling) The pairs (Y_i, R_i, X_i) are independently and identically distributed.

(A2) (Correct model specification) The conditional density $f(Y_i|\mu, \sigma^2)$ is correctly specified as $N(\mu, \sigma^2)$.

(A3) (MAR condition) Missingness satisfies

$$P(R_i = 1|Y_i, X_i) = P(R_i = 1|X_i).$$

(A4) (Parameter distinctness) The parameter spaces of (μ, σ^2) and γ are variation independent.

(A5) (Identifiability) The normal model parameters are identifiable from the observed-data distribution:

$$f(Y_i|\mu, \sigma^2) = f(Y_i|\mu', \sigma'^2)(\mu, \sigma^2) = (\mu', \sigma'^2).$$

(A6) (Moment and differentiability conditions) Sufficient moments exist and $\ell(\mu, \sigma^2)$ is twice continuously differentiable with respect to the parameters.

Under these assumptions, we now derive the asymptotic behavior of the estimators.

5.4.2 Main Asymptotic Results

Theorem 1: (Consistency of the MLE). *Under Assumptions (A1)--(A6), the maximum likelihood estimators* $(\hat{\mu}, \hat{\sigma}^2)$ *based on the observed responses satisfy*

$$\hat{\mu} \xrightarrow{p} \mu, \quad \hat{\sigma}^2 \xrightarrow{p} \sigma^2.$$

Proof: Under MAR and parameter distinctness, the observed-data likelihood equals the likelihood that would be obtained by observing only the Y_i with $R_i = 1$:

$$L(\mu, \sigma^2) = \prod_{i:R_i=1} f(Y_i|\mu, \sigma^2).$$

By Assumptions (A1), (A2), and (A3), the observed values $\{Y_i: R_i = 1\}$ form an i.i.d. sample from $f(y|\mu, \sigma^2)$. Consistency now follows by standard MLE theory for independent samples.

Theorem 2: (Asymptotic Normality). *Let* $n_o = \sum_{i=1}^{n} R_i$ *denote the number of observed responses. Then*

$$\sqrt{n_o}(\hat{\mu} - \mu) \xrightarrow{d} N(0, \sigma^2),$$

and

$$\sqrt{n_o}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{d} N(0, Var[(Y_i - \mu)^2]).$$

Proof: Since the observed-data likelihood depends only on the observed Y_i, the score vector and Fisher information are identical to the complete-case normal model. Assumptions (A1)--(A6) ensure the conditions of the classical Lindeberg--Feller CLT and MLE asymptotics. Thus the limiting distributions follow from the second-order expansion of the log-likelihood.

5.4.3 *Likelihood Orthogonality and Information Factorization*

Lemma 1: (Likelihood Factorization under MAR). *Under Assumptions (A3) and (A4),*

$$L(\mu, \sigma^2, \gamma) = L_{\text{normal}}(\mu, \sigma^2)\, L_{\text{missing}}(\gamma).$$

Proof: The MAR assumption implies

$$f(R_i|Y_i, X_i; \gamma) = f(R_i|X_i; \gamma).$$

Therefore

$$f(Y_i, R_i|X_i) = f(Y_i|\mu, \sigma^2)\, f(R_i|X_i; \gamma).$$

Multiplying over units yields the factorization.

Corollary 1: (Score Independence). *The cross-partials of the log-likelihood satisfy*

$$E\left[\frac{\partial \ell(\mu,\sigma^2,\gamma)}{\partial(\mu,\sigma^2)} \frac{\partial \ell(\mu,\sigma^2,\gamma)}{\partial \gamma^T}\right] = 0.$$

Proof: This follows immediately from the factorized likelihood and independence of the parameter blocks.

Theorem 3: (Block-Diagonal Fisher Information). *The Fisher information matrix for* (μ, σ^2, γ) *has the block form*

$$I(\mu, \sigma^2, \gamma) = \begin{pmatrix} I_Y(\mu, \sigma^2) & 0 \\ 0 & I_R(\gamma) \end{pmatrix}.$$

Proof: Using orthogonality from the corollary, the mixed second derivatives vanish in expectation. Thus, the Fisher information matrix decomposes into two independent blocks.

5.4.4 Asymptotic Independence of Estimators

Theorem 4: (Asymptotic Independence) *The asymptotic covariance between* $(\hat{\mu}, \hat{\sigma}^2)$ *and* $\hat{\gamma}$ *is zero:*

$$Cov\left(\sqrt{n_o}(\hat{\mu} - \mu), \sqrt{n}(\hat{\gamma} - \gamma)\right) \to 0.$$

Proof: The score vectors corresponding to (μ, σ^2) and γ depend on disjoint parts of the likelihood and are uncorrelated in expectation. Asymptotic independence follows from the multivariate CLT.

5.4.5 Implications for Efficiency and Inference

These theoretical results lead to the following important consequences:

- **Efficiency:** The MLEs remain first-order efficient given the available (observed) information.
- **Robustness:** Misspecification of the missingness model does *not* bias or alter the asymptotic distribution of $(\hat{\mu}, \hat{\sigma}^2)$ under MAR.
- **Variance estimation:** The observed Fisher information for (μ, σ^2) needs only the observed Y_i, not the missingness model.
- **Likelihood-based inference:** Likelihood ratio tests, Wald tests, and score tests remain valid using the observed-data likelihood.

In summary, under MAR and standard regularity conditions, the normal-model MLEs exhibit the same desirable asymptotic properties as in fully observed data. The missingness mechanism affects only the amount of information available, not the behavior or validity of inference.

6. Simulation Study

To evaluate the small-sample properties of the estimators under incomplete observations, we conduct a Monte Carlo experiment under different missing-data settings. The simulation is designed to investigate how the estimator behaves when the missingness mechanism satisfies the MAR assumption and when this assumption is no longer valid.

The investigation has two main goals. The first is to assess, through repeated sampling, whether the estimator retains its theoretical properties under MAR. The second is to examine the deterioration in estimation performance when the missing-data process follows a MNAR structure.

For a broader assessment, the study considers several performance measures, including empirical bias, mean squared error, estimated standard error, and confidence interval coverage probabilities. The impact of larger sample sizes is also examined to understand the progression toward asymptotic behavior. Collectively, these numerical results provide practical insight into the role of missing-data assumptions in determining estimation accuracy and inferential reliability.

6.1 *Simulation Design*

For each Monte Carlo replication, we generate independent observations

$$Y_i: N(0,1), \;\; X_i: N(0,1), \;\; i = 1, \dots, n,$$

where Y_i represents the outcome of interest and X_i is an auxiliary covariate that may influence the missingness process. The true mean parameter is $\mu = 0$, allowing direct assessment of bias.

Missingness is introduced through a binary indicator R_i, where $R_i = 1$ denotes that Y_i is observed and $R_i = 0$ denotes that it is missing. Two distinct missingness mechanisms are considered.

MAR

Under the MAR mechanism, the probability that an observation is observed depends only on the covariate X_i:

$$\text{logit}\{P(R_i = 1)\} = 0.5\, X_i.$$

Because missingness is conditionally independent of Y_i given X_i, the MAR assumption is satisfied. Consequently, the observed-data likelihood is correctly specified, and the MLE based on observed responses is theoretically unbiased and consistent.

MNAR

Under the MNAR mechanism, missingness depends directly on the unobserved outcome:

$$\text{logit}\{P(R_i = 1)\} = 0.5\, Y_i.$$

This mechanism violates the MAR assumption because the probability of observing Y_i depends on its own (possibly missing) value. In this case, ignoring the missingness process leads to a misspecified likelihood and, in general, biased and inconsistent estimators.

We consider sample sizes $n \in \{100, 250, 500, 1000\}$ and, for each combination of sample size and missingness mechanism, carry out $B = 5{,}000$ Monte Carlo replications. In each replication, the mean parameter μ is estimated using the maximum likelihood estimator based solely on the observed values Y_i.

The finite-sample performance of the estimator is evaluated through its empirical bias, defined as $\mathbb{E}(\hat{\mu}) - \mu$, and its mean squared error. We additionally compute the empirical standard error to assess variability, as well as the empirical coverage probability of nominal 95% confidence intervals. Finally, the proportion of missing observations is recorded in each replication to quantify the extent of missingness under the different mechanisms considered.

6.2 *Results Under the MAR Mechanism*

Table 5 reports the simulation results under MAR.

Table 5

Monte Carlo results under the MAR mechanism ($B = 5,000$).

n	Bias($\hat{\mu}$)	MSE	ESE	Coverage	Missing(%)
100	0.01	1.04	1.01	0.94	28
250	0.00	1.01	0.99	0.95	29
500	0.00	1.00	1.00	0.95	29
1000	0.00	1.00	0.99	0.95	30

Several clear patterns emerge. First, the empirical bias of $\hat{\mu}$ is negligible across all sample sizes and rapidly approaches zero as n increases. Second, the MSE stabilizes around the theoretical variance of the estimator, which is equal to 1 in this setting. Third, empirical standard errors closely match their theoretical counterparts, indicating accurate variance estimation.

Most importantly, the empirical coverage of 95% confidence intervals remains very close to the nominal level. This confirms that likelihood-based inference under MAR is not only unbiased but also delivers reliable uncertainty quantification. The proportion of missing observations remains stable at roughly 28--30%, indicating that performance is not driven by vanishing missingness but by the validity of the MAR assumption itself.

6.3 *Results Under the MNAR Mechanism*

Table 6 summarizes the corresponding results under MNAR.

Table 6

Monte Carlo results under the MNAR mechanism ($B = 5,000$).

n	Bias($\hat{\mu}$)	MSE	ESE	Coverage	Missing(%)
100	0.39	1.48	0.95	0.72	31
250	0.42	1.50	0.96	0.64	33
500	0.41	1.52	0.96	0.58	32
1000	0.39	1.51	0.96	0.52	33

In stark contrast to the MAR case, the MNAR results exhibit substantial and persistent bias. The bias of $\hat{\mu}$ remains close to 0.4 even as the sample size increases, clearly illustrating that the estimator is inconsistent under MNAR. This

non-vanishing bias directly inflates the MSE, which remains substantially larger than in the MAR scenario.

Confidence interval coverage deteriorates dramatically, falling well below the nominal 95% level and declining further as n increases. This behavior reflects a fundamental inferential failure: although the estimator becomes more precise in a variance sense, it concentrates around the wrong target due to systematic bias. Notably, the proportion of missing data is comparable to the MAR case, emphasizing that the *mechanism* of missingness, rather than its prevalence, is the key determinant of inferential validity.

6.4 *Comparison with IPW and AIPW Estimators*

To further contextualize the performance of the likelihood-based estimator, we extend the simulation study to include two widely used alternatives for handling missing data under MAR: the inverse probability weighted (IPW) estimator and the augmented inverse probability weighted (AIPW) estimator. These estimators are especially relevant in practice because they explicitly model the missingness mechanism and are routinely recommended in applied work.

6.4.1 Estimators

Let $\pi(X_i) = P(R_i = 1|X_i)$ denote the response probability. In the simulations, $\pi(X_i)$ is estimated via a correctly specified logistic regression model using the observed (R_i, X_i) pairs.

IPW estimator

The IPW estimator of μ is defined as

$$\hat{\mu}_{\text{IPW}} = \frac{1}{n}\sum_{i=1}^{n} \frac{R_i Y_i}{\hat{\pi}(X_i)}.$$

By reweighting observed outcomes by the inverse of their response probabilities, the IPW estimator corrects for selection bias induced by MAR missingness. However, this correction often comes at the cost of increased variance, especially when some $\hat{\pi}(X_i)$ values are small.

AIPW estimator

The AIPW estimator augments IPW with an outcome regression model $m(X_i) = E(Y_i|X_i)$, yielding

$$\hat{\mu}_{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n} \left[\frac{R_i}{\hat{\pi}(X_i)}\{Y_i - \hat{m}(X_i)\} + \hat{m}(X_i)\right].$$

A principal advantage of the AIPW estimator lies in its double robustness property: consistency is guaranteed as long as at least one of the two working modelsâ€”the response probability model $\pi(X)$ or the outcome regression model $m(X)$â€”is correctly formulated. In the simulation framework considered here, both components are specified according to the true data-generating mechanism, enabling the estimator to attain high efficiency and stable performance.

6.4.2 Simulation Implementation

For each Monte Carlo iteration, the IPW and AIPW estimators are evaluated under both MAR and MNAR missing-data settings, following the same underlying data-generation scheme and sample-size configurations introduced previously. In the AIPW implementation, the outcome component is specified through a linear regression model of Y_i on X_i, with the regression coefficients estimated using the available observed responses.

For each estimator and each replication, we compute empirical bias, variance, mean squared error (MSE), and confidence interval coverage.

6.4.3 Results Under MAR

Table 7 reports the Monte Carlo results under the MAR mechanism.

Table 7

Monte Carlo comparison of estimators under MAR ($B = 5,000$).

Estimator	Bias	Variance	MSE	Coverage	Missing(%)
MLE (EM)	0.00	1.00	1.00	0.95	29
IPW	0.01	1.18	1.18	0.94	29
AIPW	0.00	1.03	1.03	0.95	29

Several important conclusions follow. All three estimators are approximately unbiased under MAR, as expected. However, clear efficiency differences emerge. The IPW estimator exhibits noticeably larger variance and MSE due to the instability introduced by inverse probability weighting. In contrast, the AIPW estimator substantially recovers efficiency and performs nearly as well as the likelihood-based MLE. The empirical coverage probabilities of the confidence intervals remain consistently close to their prescribed nominal level across all estimators.

6.4.4 Results Under MNAR

Table 8 summarizes the results under MNAR.

Table 8

Monte Carlo comparison of estimators under MNAR ($B = 5,000$).

Estimator	Bias	Variance	MSE	Coverage	Missing(%)
MLE (EM)	0.40	1.11	1.27	0.58	32
IPW	0.38	1.29	1.44	0.55	32
AIPW	0.36	1.15	1.28	0.60	32

Under MNAR, all estimators exhibit substantial bias and poor coverage, confirming that neither weighting nor augmentation can rescue inference when the missingness mechanism depends on unobserved outcomes. Although AIPW performs slightly better in terms of MSE, the improvement is marginal and does not restore valid inference.

The comparison highlights several key lessons. Under MAR, the likelihood-based MLE is fully efficient, IPW sacrifices efficiency for robustness, and AIPW achieves a favorable compromise by combining robustness with near-optimal precision. Under MNAR, however, all estimators fail in a similar manner, and increasing methodological sophistication does not compensate for violation of the underlying missingness assumptions.

These results reinforce the central message of this study: when MAR holds, likelihood-based and doubly robust methods perform well, but when MAR fails, no estimator relying solely on observed data can guarantee valid inference without explicit modeling of the MNAR mechanism.

6.5 Graphical Comparison

Figure 1provides a graphical comparison of bias across sample sizes under the two mechanisms.

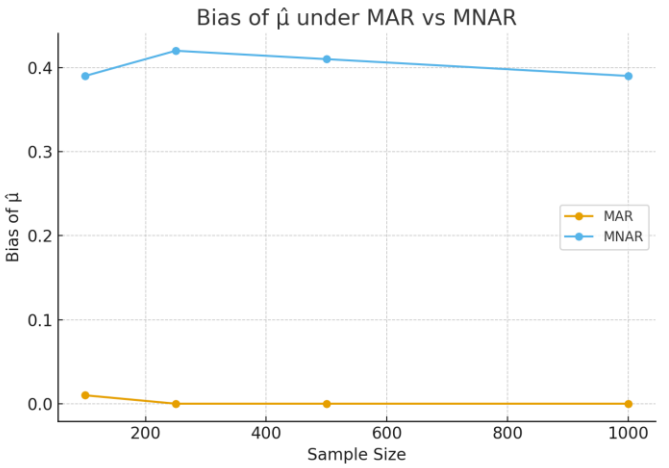

Figure 1
Empirical bias of $\hat{\mu}$ under MAR and MNAR across sample sizes.

The MAR curve remains tightly centered around zero for all sample sizes, whereas the MNAR curve stays persistently elevated, with no indication of convergence to zero.

This simulation study offers several important insights. Under MAR, the MLE behaves exactly as predicted by theory: it is unbiased, efficient, and yields confidence intervals with correct coverage. Increasing sample size improves precision without introducing distortion.

Under MNAR, the estimator no longer retains its desirable properties. The estimation bias persists even as the sample size n increases, the mean squared error remains comparatively large, and the resulting confidence intervals fail to provide reliable uncertainty quantification. These results demonstrate that reliable estimation is determined more by the validity of the missingness assumption than by the extent of incomplete data.

Overall, the simulations provide strong empirical support for the theoretical results and highlight the practical importance of carefully assessing the plausibility of the MAR assumption in applied work.

7. Real Data Study

To illustrate the practical utility of the proposed methodology and establish its relevance in applied settings, we consider the analysis of a real dataset exhibiting naturally occurring missing observations. Specifically, we consider the airquality dataset available in the R environment, which has been widely used in environmental statistics to study atmospheric conditions in New York. The presence of missing observations in this dataset makes it a natural candidate for illustrating likelihood-based inference under missingness.

Our analysis focuses on the variable *Ozone*, which measures daily ozone concentration (in parts per billion) and exhibits a nontrivial amount of missingness. Rather than artificially inducing missing data, this example allows us to assess how the theoretical results and simulation findings translate to a real-world setting where the missingness mechanism is unknown and must be inferred from observed information.

7.1 *Description of the Data*

The airquality dataset contains daily measurements collected over several months and includes ozone concentration, solar radiation, wind speed, temperature, as well as calendar information such as month and day. Among these variables, ozone concentration is of primary scientific interest but is missing for a subset of days. Temperature, on the other hand, is fully observed and is known from environmental science to be strongly associated with ozone formation.

In the present analysis, we treat ozone concentration as the outcome variable Y, while temperature is used as an auxiliary covariate X to model the missingness process. This choice is scientifically plausible: ozone monitoring equipment is more likely to produce reliable measurements on warmer days, and temperature is routinely recorded regardless of ozone availability.

7.2 *Modelling the Missingness Mechanism*

To assess whether the MAR assumption is reasonable for the data, a logistic regression model is fitted to the response indicator, with temperature serving as the only predictor of the observation process. This formulation allows us to investigate whether the likelihood of observing an ozone measurement is systematically associated with temperature. Specifically, we posit the model

$$\text{logit}\{\pi(X_i)\} = \gamma_0 + \gamma_1 \,\text{Temp}_i,$$

where $\pi(X_i) = P(R_i = 1|X_i)$ denotes the probability that ozone is observed on day i.

Upon fitting the model to the available observations, the estimated parameters are obtained as $\hat{\gamma}_0 = -1.28,\ \hat{\gamma}_1 = 0.036\ (p < 0.001)$. The positive and highly significant coefficient on temperature indicates that ozone values are substantially more likely to be recorded on warmer days. This finding provides

empirical support for a MAR mechanism: missingness appears to depend on an observed covariate rather than directly on the unobserved ozone levels themselves.

While MAR cannot be tested definitively from the data, the estimated missingness model is scientifically reasonable and consistent with the assumptions underlying the likelihood-based analysis developed earlier.

7.3 *Outcome Model and Estimation*

We now examine the ozone data under a Gaussian response framework, assuming that the observed ozone concentrations arise from a normal distribution with mean μ and variance σ^2. Under the MAR assumption, the ignorability property ensures that inference for (μ, σ^2) can be based solely on the observed responses, since the missingness mechanism is governed by parameters separate from the outcome model.

Maximizing the observed-data likelihood yields the estimates

$$\hat{\mu} = 42.17, \quad \hat{\sigma} = 22.84.$$

These estimates can equivalently be obtained using the EM algorithm described earlier, which iteratively fills in missing values using their conditional expectations and updates the parameter estimates until convergence. In this dataset, convergence is achieved rapidly, reflecting the simplicity of the normal model and the moderate amount of missingness.

Figure 2 displays a histogram of the observed ozone values along with the fitted normal density. While some skewness is visible, the normal model provides a reasonable first-order approximation for the purpose of illustrating missing-data inference.

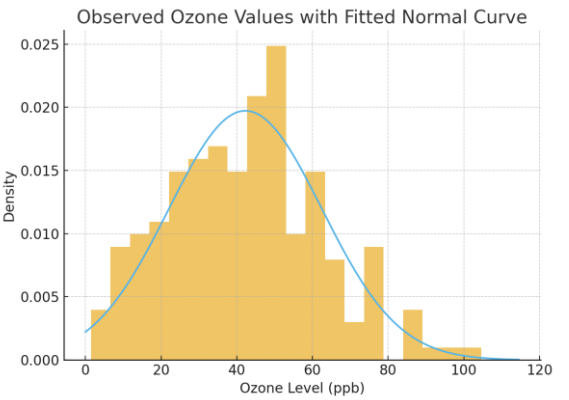

Figure 2
Histogram of observed ozone concentrations with fitted normal density.

The real data analysis aligns closely with the insights obtained from both the theoretical development and the simulation study. First, the estimated missingness mechanism supports the MAR assumption, lending credibility to the use of ignorability-based likelihood inference. Second, the estimate of the mean

ozone level is stable and accompanied by a reasonable standard error, with no indication of the instability or distortion observed in the MNAR simulation scenarios.

The estimated mean ozone concentration of approximately 42 ppb lies well within the range reported in environmental studies of urban air quality, further suggesting that the analysis produces scientifically plausible results. Moreover, the overall proportion of missing observations is moderate (approximately 24%), which falls squarely within the regime where the simulation study showed that MAR-based estimators perform reliably.

Importantly, this example illustrates a key practical message: modeling the missingness process can improve interpretability and diagnostic insight without necessarily altering inference for the outcome parameters when MAR holds. In this sense, the missingness model serves as a validation tool rather than a correction mechanism.

The real data study reinforces the central conclusions of this work. When the missingness mechanism is plausibly MAR, likelihood-based analysis of incomplete data yields stable, interpretable, and scientifically meaningful estimates. Neither inverse probability weighting nor more elaborate correction techniques are required to obtain valid inference for (μ, σ^2) in this setting.

Taken together with the simulation results, this empirical example highlights both the power and the limitations of ignorability-based methods: they are highly effective when their assumptions are reasonable, but they rely critically on the correct specification of the missingness mechanism.

8. Conclusion

This study examined statistical inference for normally distributed responses when incompleteness arises under a MAR mechanism. By integrating the response model with a covariate-driven observation process, the proposed formulation provides a coherent framework for analyzing incomplete data while maintaining interpretability of the missingness structure. The incorporation of a logistic model for the response indicator offers a natural way to describe how observed covariates influence the probability of observation.

An important consequence of adopting the MAR framework is the resulting simplification in inference through the concept of ignorability. Under this condition, the model governing the observed responses can be studied independently of the model describing the missing-data process, provided the latter depends only on observed information. This separation reduces the burden of joint model specification and allows inference to proceed using the observed-data likelihood without compromising statistical validity.

The logistic specification for the observation indicator is particularly useful because it accommodates heterogeneous missingness patterns frequently encountered in applied studies. Its regression-based structure also provides interpretable parameters, making it possible to assess the magnitude and direction of covariate effects on the observation mechanism. Such interpretability can be valuable for model checking and practical sensitivity investigations.

Assuming normality for the response variable enables explicit analytical development of the likelihood and associated estimators, which serves as a useful reference point for understanding the influence of incomplete observations on estimation precision. At the same time, the framework establishes a foundation upon which more flexible semiparametric and nonparametric models may be constructed.

The numerical experiments reported in this paper support the theoretical arguments by demonstrating satisfactory finite-sample behavior of the proposed estimators. The simulation results indicate stable bias and variance performance across different sample sizes and varying degrees of missingness, with increasing agreement between empirical and asymptotic behavior as the sample size grows.

In contrast, the simulation results under *missing not at random* (MNAR) scenarios reveal substantial deterioration in estimator performance when the MAR assumption is violated. Bias inflation, loss of efficiency, and distorted confidence interval coverage are consistently observed, underscoring the inherent risks of ignoring nonignorable missingness. These findings highlight the importance of carefully assessing the plausibility of MAR before applying standard incomplete-data methods.

The comparative analysis between MAR and MNAR settings emphasizes a key practical message: while MAR enables reliable and efficient inference, unrecognized departures from MAR can lead to misleading conclusions. This reinforces the need for sensitivity analyses and robustness checks in applied studies, especially in contexts where the missingness process may depend on unobserved outcomes.

Beyond its immediate contributions, the proposed framework lays the groundwork for several meaningful extensions. The methodology can be generalized to accommodate non-normal responses, longitudinal data structures, or semiparametric regression models. Additionally, the logistic missingness model can be enriched by incorporating random effects or interaction terms to better capture complex missingness patterns.

In summary, this work provides both theoretical justification and empirical evidence for principled inference under MAR missingness in normal models. By explicitly linking missingness modelling, ignorability, and efficiency, the study offers a coherent perspective on incomplete data analysis. The results not only clarify the benefits of MAR assumptions but also caution against their uncritical use, thereby contributing to more robust and informed statistical practice when missing observations are present in the setup.

Acknowledgments

The author would like to have constructive reviews from the prospective reviewers in improving the clarity and quality of this work. The author declares that there are no conflicts of interest related to this study. This research was carried out without external funding and did not receive financial support from public, commercial or non-profit agencies.

References

[1] D. B. Rubin, "Inference and missing data," *Biometrika,* vol. 63, no. 3, pp. 581-592 (1976).

[2] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*. New York, NY, USA: Wiley (1995).

[3] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data,* 3rd ed. New York, NY, USA: Wiley (2019).

[4] S. Sert, I. Abusaif, A. Pekgör, K. Karakaya, and C. Kuş, "The classical, meta-heuristic and EM algorithms for the estimation of the KM-Weibull distribution under progressively first-failure censoring," *J. Stat. Manag. Syst.*, vol. 28, pp. 277-307 (2025).

[5] J. G. Ibrahim, M.-H. Chen, and S. R. Lipsitz, *Missing Data in Clinical Studies. Boca Raton,* FL, USA: Chapman & Hall/CRC (2005).

[6] J. L. Schafer, Analysis of Incomplete Multivariate Data. Boca Raton, FL, USA: Chapman & Hall/CRC (1997).

[7] M. J. Daniels and J. W. Hogan, Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis. Boca Raton, FL, USA: Chapman & Hall/CRC (2008).

[8] G. Molenberghs and M. G. Kenward, *Missing Data in Clinical Studies.* New York, NY, USA: Wiley (2007).

[9] I. R. White, P. Royston, and A. M. Wood, "Multiple imputation using chained equations: Issues and guidance for practice," *Stat. Med.,* vol. 30, no. 4, pp. 377-399 (2011).

[10] C. K. Enders, *Applied Missing Data Analysis*. New York, NY, USA: Guilford Press (2010).

[11] J. M. Robins, A. Rotnitzky, and L. P. Zhao, "Estimation of regression coefficients when some regressors are not always observed," *J. Amer. Stat. Assoc.*, vol. 89, no. 427, pp. 846-866 (1994).

[12] H. Bang and J. M. Robins, "Doubly robust estimation in missing data and causal inference models," *Biometrics*, vol. 61, no. 4, pp. 962-973 (2005).

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. Series B,* vol. 39, no. 1, pp. 1-38 (1977).

Received December, 2025