mathematics

Article

Imputing Missing Data in One-Shot Devices Using Unsupervised Learning Approach

Hon Yiu So [1,*,†], Man Ho Ling [2] and Narayanaswamy Balakrishnan [3]

[1] Department of Mathematics and Statistics, Oakland University, Rochester, MI 48309, USA
[2] Department of Mathematics and Information Technology, The Education University of Hong Kong, Hong Kong, China; amhling@eduhk.hk
[3] Department of Mathematics and Statistics, McMaster University, Hamilton, ON L8S 4K1, Canada; bala@mcmaster.ca
* Correspondence: hso@oakland.edu
† Current Address: 146 Library Drive, Rochester, MI 48309, USA.

Abstract: One-shot devices are products that can only be used once. Typical one-shot devices include airbags, fire extinguishers, inflatable life vests, ammo, and handheld flares. Most of them are life-saving products and should be highly reliable in an emergency. Quality control of those productions and predicting their reliabilities over time is critically important. To assess the reliability of the products, manufacturers usually test them in controlled conditions rather than user conditions. We may rely on public datasets that reflect their reliability in actual use, but the datasets often come with missing observations. The experimenter may lose information on covariate readings due to human errors. Traditional missing-data-handling methods may not work well in handling one-shot device data as they only contain their survival statuses. In this research, we propose Multiple Imputation with Unsupervised Learning (MIUL) to impute the missing data using Hierarchical Clustering, k-prototype, and density-based spatial clustering of applications with noise (DBSCAN). Our simulation study shows that MIUL algorithms have superior performance. We also illustrate the method using datasets from the Crash Report Sampling System (CRSS) of the National Highway Traffic Safety Administration (NHTSA).

Keywords: one-shot devices; missing data; clustering; imputation; inverse probability weighting; unsupervised learning; clustering; k-prototype; DBSCAN

MSC: 62N05; 62F40; 62H30; 62P30

Citation: So, H.Y.; Ling, M.H.; Balakrishnan, N. Imputing Missing Data in One-Shot Devices Using Unsupervised Learning Approach. *Mathematics* **2024**, *12*, 2884. https://doi.org/10.3390/math12182884

Academic Editor: Ioannis Tsoulos

Received: 9 August 2024
Revised: 1 September 2024
Accepted: 3 September 2024
Published: 15 September 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

One-shot devices are products that can only be used once. For example, airbags deploy during collision. They have to be replaced as they cannot deploy again. EpiPens are auto-injectors that are capable of injecting injections of epinephrine into patients who are experiencing life-threatening allergic reactions. These pens need to be replaced because they cannot be deployed again, obviously. Since most one-shot devices are life-saving, they have to be extremely dependable in times of emergency.

As most products come with warranties and insurance, any small systematic errors in one-shot device production would lead to colossal life and financial losses. The recent bankruptcy of the Takata airbag company is an infamous example of this [1]. The quality control of those products and predicting their reliability over time is critically important. Accelerated Life Tests (ALTs) are popular procedures to assess and analyze the quality of one-shot devices, and they depend heavily on the extrapolation of the life model from the high-stress levels during the experiment to the user conditions. Any biases in the model estimates would be amplified during the extrapolation.

Under parametric settings, the one-shot device data with N observations would have the likelihood function,

$$L(\theta;\boldsymbol{x}) = \prod_{i=1}^{N} (F(t_i;\theta,\boldsymbol{x}_i))^{1-\delta_i}(1 - F(t_i;\theta,\boldsymbol{x}_i))^{\delta_i}, \tag{1}$$

where θ is the parameter of the lifetime distribution with cumulative distribution function $F(\cdot)$, $\boldsymbol{x}_i$ is the vector of covariates of the ith subject, $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$, t_i is the observed time when the ith one-shot device is activated and δ_i is the indicator equal to 1 if the device functions correctly.

Manufacturers often carry out these tests before products enter the market, and we rarely test them in actual user conditions. To address this issue, we can refer to publicly available datasets that collect information from actual usage and retrospectively analyze the reliability of those products. For example, the Crash Report Sampling System (CRSS) of the National Highway Traffic Safety Administration (NHTSA) ([2]) has collected national crash data since 2016. It collects details of car crashes, including the status of airbag deployment. It provides a good data source for evaluating different car safety systems during the actual user conditions. This kind of retrospective study is vitally important as it can indicate potential problems in safety devices early and avoid disastrous replacements similar to the Takada airbag recall, see [1].

In most reliability testing, there are seldom any missing data as the experiments are under well-controlled conditions. However, datasets used in retrospective studies often have missing data. For example, in CRSS, the proportion of missing observations in the datasets is substantial (at least 16.43% in 2017, 13.39% in 2018, 10.00% in 2019 and 11.78% in 2020). This is due to various difficulties in data collection, such as lack of human resources, vehicle conditions or the severity of the accidents. Most popular statistical methods, like multiple imputations, to handle missing data often require the missingness not to depend on the response [3–5], which may not be reasonable in a retrospective study. This is because some latent factors not recorded in the datasets may affect both the responses and the missing mechanism. One obvious latent factor is whether the car was parked in a covered space, which is a factor of the airbag lifetime and overall car condition but is not reported in the CRSS. The existence of latent variables leads to heterogeneous data and produces biased imputation results.

Unsupervised learning (UL) techniques such as k-prototype, DBSCAN, and Hierarchical Clustering can effectively discover hidden variables in various domains. In the medical field, refs. [6,7] utilize these methods to classify patients more precisely. They improve the categorization of primary breast cancer and heart failure with maintained ejection fraction by finding separate patient clusters with diverse clinical profiles and outcomes. In engineering, ref. [8] apply the algorithms on railway vibration data to extract useful features for defect detection. UL seems helpful in categorizing the datasets into homogeneous subsets, the observations of which should have similar latent factors.

To analyze the one-shot device data with missing observations and latent factors, we propose using an unsupervised learning algorithm to form clusters, of which the data are considered homogeneous, and a standard statistical imputation technique should be possible. This paper is organized as follows: Section 2 introduces the missing data in one-shot data analysis and the inverse probability weighting (IPW). We review the traditional imputation methods in Section 3. We propose the novel multiple imputations with unsupervised learning (MIUL) in Section 4. We compare the proposed algorithm with some traditional missing data-handling methods in Section 5 and illustrate the usefulness of using CRSS datasets in Section 6. We provide concluding remarks and future research directions in Section 7.

2. Missing Mechanisms and IPW

There are many reasons missing data exist in one-shot device analysis. In particular, when we analyze data from consumers, covariates are often lost for various reasons. For example, customers may not recall certain variables or outcomes, or data entries may be deleted due to human error.

2.1. Missing Data Mechanisms

The missing data mechanism describes how the missingness appears in the datasets. Denote d_i to be the missing indicator for the ith observation, and $d_i = 1$ if all the covariates, age of the devices at the instance and the survival statuses are observed, and $d_i = 0$ otherwise. We denote X^* to be the variables that are completely observed for all data entries, and X_i^* is the corresponding value of the ith observation. If $\Pr(d_i = 0|X_i) = \Pr(d_i = 0)$ which is a constant, the missing mechanism is referred to as missing completely at random (MCAR); if $\Pr(d_i = 0|X_i)$ is not constant but independent of the missing values, it is referred to as missing at random (MAR). Otherwise, we say the mechanism is missing not at random (MNAR).

If data are MCAR, analysis based on complete cases remains valid with respect to yielding unbiased results but loses statistical power [9–11]. The analysis may yield biased results if data are MAR, but this can be overcome using appropriate statistical methods [12]. If data are MNAR, the probability that a variable value is missing depends on the missing value and cannot be fully explained by the remaining observed variables. In that case, analysis tends to yield biased results if missing data are not appropriately handled, and sensitivity analyses are usually recommended to examine the effect of different assumptions about the missing data. In most cases, the missing mechanism is not MCAR, and handling the missing data properly is essential for valid estimation results.

2.2. Inverse Probability Weighting

Dropping the observations with missing information (i.e., complete case analysis) is generally not recommended as it removes much information and may also introduce biases in the model estimation. Therefore, in the analysis of missing data in the one-shot device with completing risks, extra caution is required to assess the extent and types of missing data before analysis, explore potential mechanisms that contribute to the missing data and use appropriate missing-data strategies to handle the missing data and conduct sensitivity analysis to assess the robustness of research findings.

The inverse probability weighting (IPW) method is a popular statistical approach to handling missing data, see [13]. IPW is a technique that can correct the bias resulting from complete data analysis and is also utilized to adjust for unequal fractions in missing data. IPW addresses this issue by assigning weights to each individual in the analysis based on the inverse of the corresponding probability of being a complete case. These weights balance the distribution of observed data to resemble a distribution with no missing data, thus reducing the bias introduced in a complete-case analysis.

Usually, we denote the non-missing probability as $p_i = p(\boldsymbol{x}_i^*)$, where $\boldsymbol{x}^*$ are the covariates observable for all observations and X_i^* are the corresponding values for the ith entry. Under the IPW method, the log-likelihood function of one-shot device data is

$$\ell^{IPW}(\theta;\boldsymbol{x}) = \sum_{i=1}^{N} \frac{d_i}{p_i}((1-\delta_i)\ln(F(t_i;\theta,\boldsymbol{x}_i)) + \delta_i \ln(1-F(t_i;\theta,\boldsymbol{x}_i))), \tag{2}$$

and the score function is

$$S_\theta(\theta) = \frac{\partial \ell^{IPW}(\theta;\boldsymbol{x})}{\partial\theta} = \sum_{i=1}^{N} \frac{d_i}{p_i}\left(\frac{1-\delta_i}{F(t_i;\theta,\boldsymbol{x}_i)} - \frac{\delta_i}{1-F(t_i;\theta,\boldsymbol{x}_i)}\right)\frac{\partial F(t_i;\theta,\boldsymbol{x}_i)}{\partial\theta}. \tag{3}$$

If we assume that a parametric model for the non-missing probability, $p_i = p(\boldsymbol{x}_i^*; \alpha)$, the score function for estimating p_i would be

$$S_\alpha(\alpha) = \sum_{i=1}^{N} \left(\frac{d_i}{p(\boldsymbol{x}_i^*; \alpha)} - \frac{1 - d_i}{1 - p(\boldsymbol{x}_i^*; \alpha)} \right) \frac{\partial p(\boldsymbol{x}_i^*; \alpha)}{\partial \alpha}, \tag{4}$$

where $S_\alpha(\hat{\alpha}) = 0$ and $\hat{p}_i = p(\boldsymbol{x}_i^*, \hat{\alpha})$ is the corresponding estimate.

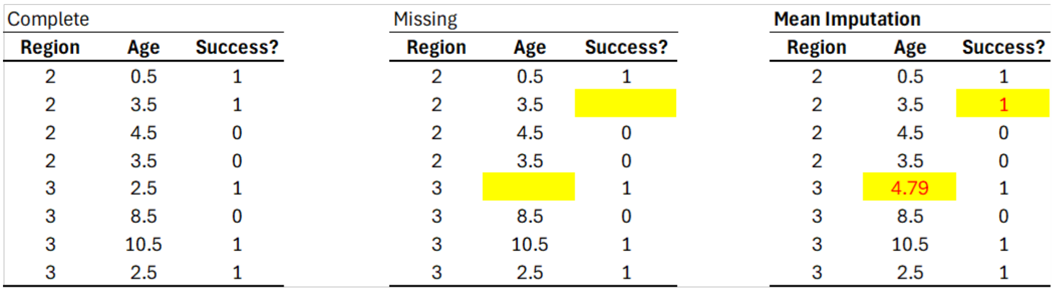
3. Literature Review on Imputation

Besides the IPW method, imputation strategies provide alternative ways to handle missing data. They are more intuitive as they fill out the missing data and give "complete datasets". Researchers can use the datasets directly without complicated mathematical formulas. Here are a few popular methods for imputing missing data.

3.1. Single Imputations

3.1.1. Mean (or Mode) Imputation

Imputation is a common strategy for dealing with missing data. It fills in the blanks with an appropriate value to create a "completed" dataset that can be analyzed using traditional statistical procedures. The most direct and intuitive strategy for continuous variables is to replace missing observations with the mean of the observed sections, known as mean imputation, see [14]. Similarly, unobserved categorical and ordinal variables could be substituted by the mode of the observed ones (mode imputation). However, because this method does not take into account association across variables [15,16], it also produces biased and over-fitting results [17]. In Figure 1, we show how to perform a mean imputation. The variable *Region* represents the region of a car accident, the variable *Age* represents the age of the car involved, and the variable *Success?* represents whether the airbags of the car function properly.



Complete

Region	Age	Success?
2	0.5	1
2	3.5	1
2	4.5	0
2	3.5	0
3	2.5	1
3	8.5	0
3	10.5	1
3	2.5	1

Missing

Region	Age	Success?
2	0.5	1
2	3.5	
2	4.5	0
2	3.5	0
3		1
3	8.5	0
3	10.5	1
3	2.5	1

Mean Imputation

Region	Age	Success?
2	0.5	1
2	3.5	1
2	4.5	0
2	3.5	0
3	4.79	1
3	8.5	0
3	10.5	1
3	2.5	1

Figure 1. Illustration on how to perform a mean imputation. The highlighted entries are the missing observations, and the red numbers represent the values imputed by the mean imputation. The table on the **left** is complete data, and the table in the **middle** is the data with missing observations. The table on the **right** has the missing data imputed by the mean of observed values.

3.1.2. Expectation Maximization Imputations

The expectation maximization (EM) imputation is a type of single imputation technique using the EM algorithm, and it is extensively studied by various literature [5,18]. Refs. [19,20] also consider using EM for the masked causes of one-shot devices. The EM algorithm is an iterative procedure of expectation (E-steps) and maximization steps (M-steps). In one iteration, the expected values of the missing responses or covariates would be updated as the mean conditional on the observed covariates, such as the components' manufacturer and current stress levels (E-step). The complete data formulas will then be applied to the dataset with missing data filled from the E-step to obtain the updated estimates by maximizing the complete likelihood function (M-step). The E-step and M-step will be performed iteratively until the imputed values converge. To illustrate how EM imputation works, we use the data example in Figure 1. We assume the *Age* follows the exponential distribution with rate = λ and the probability of Success, $S = 1$, given the *Age* of the car is

$$P(S = 1) = \begin{cases} \beta, & Age \leq 6 \\ 0.5\beta, & Age > 6 \end{cases}, \text{ for } 0 < \beta < 1.$$

The observed likelihood function is

$$L(\lambda,\ \beta) = 0.5\beta^3(1-\beta)^2(1-0.5\beta)\lambda^7 \exp(-33.5\lambda)(1-0.5\exp(-6\lambda))\beta,$$

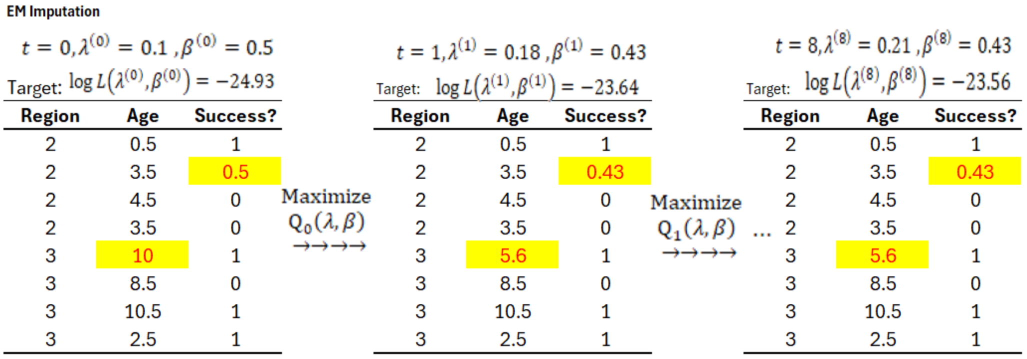
and the complete log-likelihood is

$$\begin{aligned} \ell_c(\lambda,\ \beta) =& (1 + I(Age_5 > 6))\ln 0.5 + 4\beta + 2\ln(1-\beta) + \ln(1-0.5\beta) \\ &+ 8\ln\lambda - (33.5 + Age_5)\lambda \end{aligned}$$

with relevant expected values given the current parameter estimates $\lambda^{(t)}$ and $\beta^{(t)}$ being

$$E(I(Age_5 > 6)) = \exp\left(-6\lambda^{(t)}\right) \text{ and } E(Age_5) = 1/\lambda^{(t)}.$$

We then maximize $Q_t(\lambda,\beta) = E\left(\ell_c(\lambda,\ \beta)|\lambda^{(t)},\beta^{(t)}\right)$ to obtain the next parameter estimates, $\lambda^{(t+1)}$ and $\beta^{(t+1)}$. Figure 2 illustrates how the EM imputation can be implemented.



EM Imputation

$t = 0, \lambda^{(0)} = 0.1, \beta^{(0)} = 0.5$

Target: $\log L(\lambda^{(0)},\beta^{(0)}) = -24.93$

Region	Age	Success?
2	0.5	1
2	3.5	0.5
2	4.5	0
2	3.5	0
3	10	1
3	8.5	0
3	10.5	1
3	2.5	1

Maximize
$Q_0(\lambda,\beta)$
$\rightarrow\rightarrow\rightarrow\rightarrow$

$t = 1, \lambda^{(1)} = 0.18, \beta^{(1)} = 0.43$

Target: $\log L(\lambda^{(1)},\beta^{(1)}) = -23.64$

Region	Age	Success?
2	0.5	1
2	3.5	0.43
2	4.5	0
2	3.5	0
3	5.6	1
3	8.5	0
3	10.5	1
3	2.5	1

Maximize
$Q_1(\lambda,\beta)$ …
$\rightarrow\rightarrow\rightarrow\rightarrow$

$t = 8, \lambda^{(8)} = 0.21, \beta^{(8)} = 0.43$

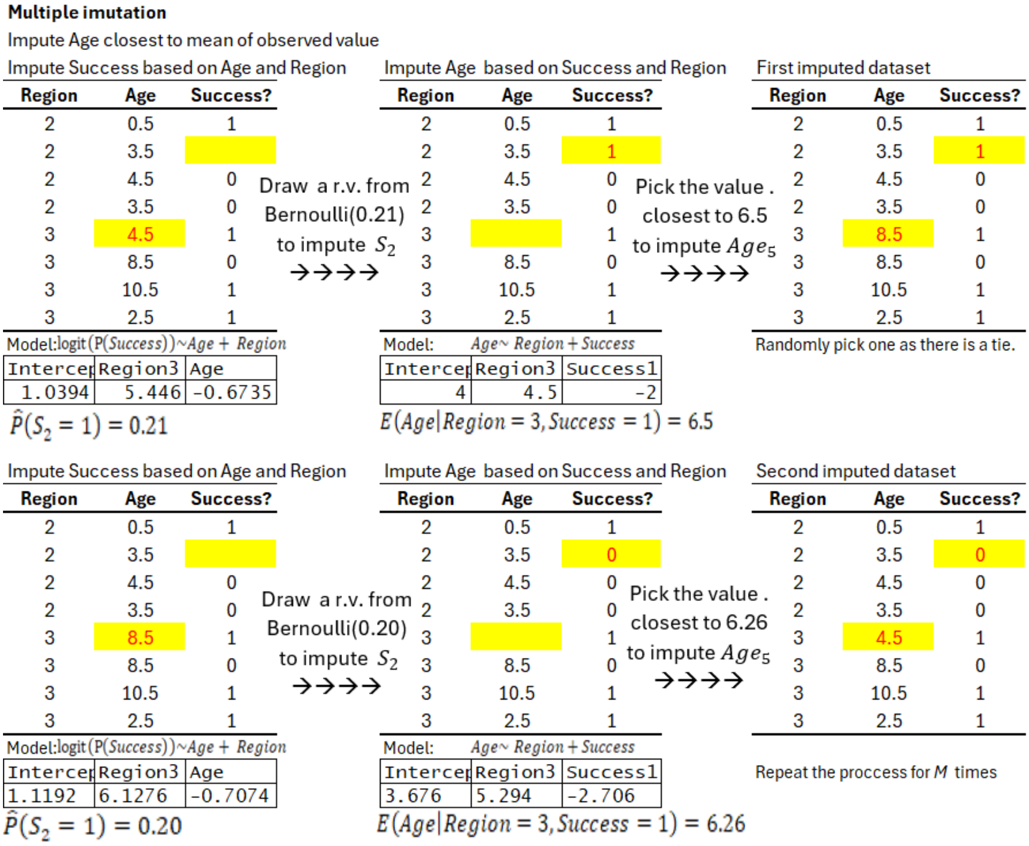
Target: $\log L(\lambda^{(8)},\beta^{(8)}) = -23.56$

Region	Age	Success?
2	0.5	1
2	3.5	0.43
2	4.5	0
2	3.5	0
3	5.6	1
3	8.5	0
3	10.5	1
3	2.5	1

Figure 2. An illustration on how we implement EM algorithm imputation. The highlighted entries are the missing observations, and the red numbers represent the values imputed during the imputation procedure. The initial parameter estimates are $\lambda^{(0)} = 0.1, \beta^{(0)} = 0.5$. We impute the missing data with the expectation based on the current parameter estimates, $\lambda^{(t)}, \beta^{(t)}$.

3.2. Multiple Imputations

Multiple imputations have been considered as a gold standard to account for missing data, and more specifically, the fully conditional specification (FCS) method (also labeled the multiple imputations by chained equations (MICE) algorithm) [21]. The multiple imputations by FCS impute multivariate missing data on a variable-by-variable basis, which is particularly useful for studies with large datasets and complex data structures, see [22]. It requires a specification of an imputation model for each incomplete variable and iteratively creates imputations per variable [23]. Compared to a single imputation, MI replaces each missing value with multiple plausible values, allowing uncertainty about the missing data to be considered. MI consists of two stages. The first stage involves creating multiple imputed datasets. In practice, we can first apply Box-Cox transformation on the numerical covariates such that they follow a multivariate normal distribution approximately. The MI method imputes the missing responses or the covariates stochastically based on the distribution of the observed data. We will repeat the process five to ten times to create ten imputed datasets. Once again, we use the data example in Figure 1 to show how multiple imputation works. We assume the variable **Age** has a linear relationship with the variables **Region** and **Success** and the log-odds of **Success** are linearly related to the variables **Region** and **Age**. We first impute the missing data with the average of observed values. Then, we form a regression model on the variable under imputation and predict the missing value with the model. If the variable is numerical, we often use the predictive mean matching method, which imputes the missing entries with the closest observed values to the predictions. For binary variables, we can perform a logistic regression and draw Bernoulli random variables with the predicted probabilities to impute the missing variables. Figure 3 shows how multiple imputation works step by step. For brevity reasons, we do not consider the model coefficient uncertainty in this example.



Multiple imutation

Impute Age closest to mean of observed value

Impute Success based on Age and Region

Region	Age	Success?
2	0.5	1
2	3.5	
2	4.5	0
2	3.5	0
3	4.5	1
3	8.5	0
3	10.5	1
3	2.5	1

Model:logit(P(*Success*))~*Age* + *Region*

Intercep	Region3	Age
1.0394	5.446	-0.6735

$\hat{P}(S_2 = 1) = 0.21$

Draw a r.v. from Bernoulli(0.21) to impute S_2 →→→→

Impute Age based on Success and Region

Region	Age	Success?
2	0.5	1
2	3.5	1
2	4.5	0
2	3.5	0
3		1
3	8.5	0
3	10.5	1
3	2.5	1

Model: $Age \sim Region + Success$

Intercep	Region3	Success1
4	4.5	-2

$E(Age|Region = 3, Success = 1) = 6.5$

Pick the value . closest to 6.5 to impute Age_5 →→→→

First imputed dataset

Region	Age	Success?
2	0.5	1
2	3.5	1
2	4.5	0
2	3.5	0
3	8.5	1
3	8.5	0
3	10.5	1
3	2.5	1

Randomly pick one as there is a tie.

Impute Success based on Age and Region

Region	Age	Success?
2	0.5	1
2	3.5	
2	4.5	0
2	3.5	0
3	8.5	1
3	8.5	0
3	10.5	1
3	2.5	1

Model:logit(P(*Success*))~*Age* + *Region*

Intercep	Region3	Age
1.1192	6.1276	-0.7074

$\hat{P}(S_2 = 1) = 0.20$

Draw a r.v. from Bernoulli(0.20) to impute S_2 →→→→

Impute Age based on Success and Region

Region	Age	Success?
2	0.5	1
2	3.5	0
2	4.5	0
2	3.5	0
3		1
3	8.5	0
3	10.5	1
3	2.5	1

Model: $Age \sim Region + Success$

Intercep	Region3	Success1
3.676	5.294	-2.706

$E(Age|Region = 3, Success = 1) = 6.26$

Pick the value . closest to 6.26 to impute Age_5 →→→→

Second imputed dataset

Region	Age	Success?
2	0.5	1
2	3.5	0
2	4.5	0
2	3.5	0
3	4.5	1
3	8.5	0
3	10.5	1
3	2.5	1

Repeat the proccess for M times

Figure 3. An illustration of how the multiple imputation procedure works. The highlighted entries are the missing observations, and the red numbers represent the values imputed during the procedure. In this example, we assume there is no model coefficient uncertainty.

In the second stage, we will analyze each of the five or ten imputed datasets separately using standard statistical models and then combine the results from the five or ten analyses to report the conclusion; see [24]. One popular choice is to use Rubin's Rules [25,26] to combine the estimates $\tilde{\beta}^{(m)}, m = 1, \ldots, M$ from multiple imputed datasets, and obtain the the pooled estimate and variance,

$$\hat{\beta} = \frac{1}{M} \sum_{i=1}^{M} \tilde{\beta}^{(m)} \quad \text{and} \quad V_{Total} = V_W + V_B + V_B/M,$$

respectively, where

$$V_W = \frac{1}{M} \sum_{i=1}^{M} \left(SE\left(\tilde{\beta}^{(m)}\right) \right)^2 \quad \text{and} \quad V_B = \frac{1}{M-1} \sum_{i=1}^{M} \left(\tilde{\beta}^{(m)} - \hat{\beta} \right)^2,$$

$SE\left(\tilde{\beta}^{(m)}\right)$ is the standard error of of the estimate from the m imputed dataset.

4. Multiple Imputation with Unsupervised Learning

This research proposes new imputation methods to handle missing data in one-shot devices that hybridize multiple imputations with unsupervised learning (MIUL). Traditional imputation assumes the data are from a homogenous group, which is only sometimes valid in the actual ALTs or observational studies on a one-shot device. The components of devices/systems may be coupled in the manufacturing process or assembly, so the components within the device may have interrelationships, leading to data with latent heterogeneity and dependence, which can be described by frailty models [27,28].

It is well-known that the latent variables will affect the model coefficients when they are correlated with the observed variables. Since the latent variables are not observed, an unsupervised ML algorithm would be a powerful way to discover hidden clusters of the devices based on the observed variable [29]. We cluster one-shot devices into groups with different characteristics using machine-learning techniques. Within each group, their latent factors should be similar to each other. Therefore, the unknown latent factors are "controlled", and the regression model parameters would be correctly adjusted. We can apply several conventional unsupervised clustering techniques to the observed covariates to discover hidden structures. For example, the k-means clustering technique is a popular method for clustering analysis in data mining. It is an unsupervised ML method partitioning the dataset into k clusters according to the distance of each observation to the cluster means. Once we cluster the data, we can impute the data using the following Algorithm 1.

```
Algorithm 1: Multiple imputations with unsupervised learning (MIUL)
Data: Data with missing observations
Result: Multiple imputed datasets
Using unsupervised learning algorithms, divide the input data set into $K$ clusters
  $C_1, C_2, \cdots, C_K$;
for Each cluster $C_k$, $k = 1, 2, \cdots, K$ do
    Multiple impute the data under $C_k$ to obtain the imputed datasets,
      $D_{1k}^{(I)}, D_{2k}^{(I)}, \cdots, D_{Mk}^{(I)}$;
end
for Each Imputed dataset $D_{mk}^{(I)}$, $m = 1, 2, \cdots, M$ do
    Combine the imputed datasets, $D_{m1}^{(I)}, D_{m2}^{(I)}, \cdots, D_{mK}^{(I)}$ as the $m$th imputed
      dataset $D_m^{(I)}$.
end
Treat the imputed datasets $D_1^{(I)}, D_2^{(I)}, \cdots, D_M^{(I)}$ as multiply imputed datasets and
  follow the polling procedure of multiple imputations.
```

This algorithm describes our proposed methods: multiple imputations with unsupervised learning (MIUL). Ref. [30] also discuss similar ideas, but they mainly focus on imputing datasets for clustering, while this research study focuses on clustering the datasets for imputation. There are several popular unsupervised learning algorithms, and they are listed below.

4.1. Hierarchical Clustering

Hierarchical Clustering is a technique that creates an ordered sequence of data clusters using a specific dissimilarity measure [31]. The concept of grouping hierarchically was first introduced by [32]. The term "Hierarchical Clustering" was first coined by [33], who discussed the approaches and the choice of distance measure. Ref. [34] gives further details on how Hierarchical Clustering can be done. It is also an unsupervised ML method that tries to build a hierarchy of clusters. In the agglomerative approach, each observation is first treated as a distinct cluster. These single observation clusters gradually merge depending on the smallest dissimilarity. Given the distance measure between two data points $X_i, X_{i'}$ is $dist(X_i, X_{i'})$, there are three common dissimilarity measures between two clusters [35]: suppose G and H are two groups during the clustering process:

Single linkage (SL) dissimilarity,

$$d_{SL}(G,H) = \min_{\{X_i \in G, X_{i'} \in H\}} dist(X_i, X_{i'}),$$

complete linkage dissimilarity,

$$d_{CL}(G,H) = \max_{\{X_i \in G, X_{i'} \in H\}} dist(X_i, X_{i'}),$$

and group average (GA),

$$d_{GA}(G,H) = \frac{1}{N_G N_H} \sum_{X_i \in G} \sum_{X_{i'} \in H} dist(X_i, X_{i'}),$$

where N_G and N_H are the numbers of data points in Groups G and H, respectively. Similarly, in the divisive approach, all the observations are treated as one cluster and split into clusters based on dissimilarity until each observation is treated as a distinct cluster. Both methods create multi-tiered hierarchies as the procedure goes on. They span the extremes of each observation existing as their cluster and all observations combining into a single, all-encompassing cluster. A dendrogram is a visual representation of the cluster-merging process. The process can stop at k clusters if the average dissimilarity between a pair of clusters changes substantially. Factors like many irrelevant variables may obscure clusters, as they only exist within a small subset of variables. Thus, despite seeming close due to relevant variables, the curse of dimensionality can create a vast distance between observations in a high-dimensional space, which challenges the effectiveness of dissimilarity measures and the accuracy of the resulting clusters.

4.2. K-Prototype

K-prototype was initially proposed by [36]. The term "prototypes" refers to the centroids or representative points of clusters in a dataset. It extends the well-known k-means clustering algorithm to the categorical variables or attributes with mixed numerical and categorical values. The K-means clustering algorithm first normalizes each variable of the ith observation, $X_i^* = (x_{i1}^*, \cdots, x_{ij}^*, \cdots, x_{iJ}^*)$, by

$$x_{ij} = \frac{x_{ij}^* - \min(x_{1j}^*, \cdots, x_{Ij}^*)}{\max(x_{1j}^*, \cdots, x_{Ij}^*) - \min(x_{1j}^*, \cdots, x_{Ij}^*)},$$

where $X_i = (x_{i1}, \cdots, x_{ij}, \cdots, x_{iJ})$ It, then, partitions numerical dataset variables $\boldsymbol{X} = \{X_1, X_2, \cdots, X_n\}$ into k clusters by minimizing the within groups sum of squared error (WGSS),

$$P(W, \boldsymbol{Q}) = \sum_{l=1}^{k} \sum_{i=1}^{n} w_{i,l} dist(X_i, Q_l), \tag{5}$$

where $w_{i,l} \in \{0,1\}, \sum_{l=1}^{k} w_{i,l} = 1, 1 \leq i \leq n$, is the indicator of the ith observation being in the lth cluster, Q_l, $\boldsymbol{Q} = \{Q_1, Q_2, \ldots, Q_k\}$ is a set of observations in the same cluster, and $dist(X_i, Q_l)$ is the squared Euclidean distance between X_i and the cluster center of Q_l,

$$dist_1(X_i, Q_l) = \sum_{j=1}^{p} (x_{i,j} - q_{l,j})^2, \tag{6}$$

where $x_{i,j}$ is the value of the jth variable of the ith observations and there are p numerical variables. We adopt the elbow method to determine k, which gives the reduction of minimized WGSS less than 10%.

However, datasets often contain both numerical and categorical variables. Suppose the first p variables are numerical and the remaining $m - p$ variables are categorical in a dataset. The k-prototype algorithm replaces $d(X_i, Q_l)$ by the following dissimilarity measure,

$$dist_2(X_i, Q_l) = \sum_{j=1}^{p}(x_{i,j} - q_{l,j})^2 + \gamma \sum_{j=p+1}^{m} \delta(x_{i,j}, q_{l,j}), \tag{7}$$

where

$$q_{l,j} = \begin{cases} \frac{1}{N}\sum_{i \in Q_l} x_{i,j}, & \text{for } 1 \leq j \leq p \\ c_{i,j}, & \text{for } p+1 \leq j \leq m \end{cases},$$

and $c_{l,j}$ is the most frequent category of jth variable in the cluster. Q_l and $\delta((x_{i,j}, c_{l,j}) = 1$ if $x_{i,j} = q_{l,j}$ and zero otherwise. The weight γ is chosen to balance the effects of numerical and categorical attributes. Details on how to select γ can be found in [37]. Ref. [38] reduces the misclassification of data points near the boundary region by considering the distribution centroids for the categorical variables in a cluster, and [39] further improves it by proposing a new dissimilarity measure between data objects and cluster centers.

4.3. Density-Based Spatial Clustering of Application with Noise

Density-Based Spatial Clustering of Application with Noise (DBSCAN) was originally proposed by [40]. The algorithm is uniquely designed to discover clusters of arbitrary shape, and it requires minimal domain knowledge to determine input parameters with very general assumptions [41]. It is efficient for large databases, making it a practical choice for various applications. The effectiveness and efficiency of DBSCAN were evaluated using synthetic and real data.

The concept of density-based clusters is central to the operation of DBSCAN. The *Eps-neighborhood* of a point p in the dataset D is

$$N_{Eps}(p) = \{q \in D | dist(p, q) \leq Eps\},$$

where $dist(p, q)$ is a distance measure between points p and q. If, $p \in N_{Eps}(q)$ and the neighborhood size , $|N_{Eps}(q)|$ is greater than a preset minimum number of points, $MinPts$, then p is *directly density-reachable* to q. We also define p as *density-reachable* from q if there is a chain of points $p_1 = p, p_2, \ldots p_n = q$ which are directly density-reachable consecutively. p and q are *density-connected* if they are both density-reachable from a common point o. The cluster here is defined as the set of points that are density-reachable from each other (maximality) and density-connected (connectivity). The set of points not belonging to any cluster are considered to be *noise*. Ref. [42] addresses the issue of detecting the cluster in data of varying densities by assigning the noise points to the closest eligible cluster. Ref. [43] revisits the DBSCAN algorithm and discusses how to obtain the appropriate parameters for DBSCAN. Ref. [44] extends the algorithm to produce DBSCAN clusters in a nested, hierarchical way. Ref. [45] implements the hierarchical DBSCAN in Python that gives the best DBSCAN parameters.

4.4. Gower's Distance

Most dissimilarity measures for numerical variables are based on the Euclidean distance, the square root of (6). If discrete variables exist, the k-prototype algorithm adopts the dissimilarity measure states in (7). However, both of them require the variables to be completely observed. We may use Gower's distance proposed by [46] to handle partially observed data entries. It provides another similarity measure between the ith and kth observations,

$$dist_G(i,k) = \frac{\sum_{j=1}^{m} w_j d_{ik}^{(j)} dist_{ik}^{(j)}}{\sum_{j=1}^{m} w_j d_{ik}^{(j)}}, \tag{8}$$

where $d_{ik}^{(j)}$ is the indicator if the jth variable is observed for both the ith and kth data entries, $dist_{ik}^{(j)}$ is the distance measure of the jth variable between ith and kth observations, $x_i^{(j)}$ and $x_k^{(j)}$. If jth variable is binary or nominal, $dist_{ik}^{(j)} = 1$ if $x_i^{(j)} \neq x_k^{(j)}$ and 0 otherwise. If the jth variable is numerical or ordinal, $dist_{ik}^{(j)}$ would be the absolute difference, standardized by the range $|x_i^{(j)} - x_k^{(j)}|/\text{range}(x_i^{(j)})$. The weight for the jth variable, w_j, indicating the importance of the variable, is usually set to one, based on the Adansonian principle of equal weighting, see [47].

5. Simulation Study

5.1. Simulation Settings

We develop a simulation study to compare various methods when handling missing data in one-shot devices. Suppose we are monitoring the quality of airbags in the United States. For simplicity, we assume that the airbags have exponential lifetimes with the hazard rate, $\lambda(\boldsymbol{x})$,

$$f(t_i; \boldsymbol{x}_i) = \lambda(\boldsymbol{x}_i) \exp(-\lambda(\boldsymbol{x}_i) t_i), \text{ for } t_i > 0,$$

where $\boldsymbol{x}_i$ is the variables that are associated with the airbag quality, the car make (A, B, and C), the car registration region (Northeast, Midwest, South, and West) and parking location (indoor or outdoor). To ensure the value is positive, we adopt a log-linear function.

$$\lambda_{actual}(\boldsymbol{x}_i) = \exp(\beta_0 + \beta_{make} + \beta_{region} + \beta_{park}), \tag{9}$$

where β_0 is the baseline log-hazard for all airbags, equal to -6, -5 and -4 representing high, medium and low reliability (Relb.) levels of the devices, around 62%, 82% and 93%, respectively. $\beta_{make}^{B} = -0.4$ and $\beta_{make}^{C} = -0.8$ correspond to the log-hazard rates of airbags from the car makes B and C relative to A. $(\beta_{region}^{MW}, \beta_{region}^{S}, \beta_{region}^{W}) = (-0.6, 0.4, 0.8)$ represents the log-hazard rates of cars from Midwest, South and West relative to those from Northeast, accordingly.

The effect of outdoor parking is assumed to be $\beta_{park}^{out} = 0.6$ relative to parking inside. Parking location is quite possibly related to various factors, namely, car owner's location (urban or rural), car body type (sedan, Sport Utility Vehicles and others), whether the car is in the West region, and whether the driver is an alcohol drinker. Here, we assume the probability of outdoor parking is linearly related to the factors mentioned. Parking location is unlikely to be recorded, and we consider it a hidden variable. Therefore, the regression equation would be limited to

$$\lambda_{model}(\boldsymbol{x}_i) = \exp(\beta_0 + \beta_{make} + \beta_{region}), \tag{10}$$

which resembles the model misspecification in actual modeling.

Compared to other reliability experiments, this is an observational study, and some data are inevitably missed during the data collection. We simulate two different types of missing indicators. One is unrelated to the one-shot device response, nor any hidden variable corresponding to the MAR mechanism. We also generate another missing indicator related to the hidden variable, parking location, representing the MNAR scenarios. IPW uses logistic regression with all the fully observed variables to model the probability of non-missing. MI and all the MIUL methods use all the variables in the datasets for imputation. Details of the parameter setting can be found in the Appendix A.

In the simulation, we consider the scenarios with 1000, 2000, and 4000 cars we examined, representing small, medium, and large datasets. We also consider three missingness levels: around 10%, 15%, and 25%. The simulation study aims to analyze car makes' coefficients to see which one has potential problems.

5.2. The Competitors' Details

5.2.1. Traditional Methods

In the simulation, we compare six missing handling methods. The first three are mean imputation (Mean.I), IPW and MI, representing traditional and popular missing handling methods.

In Mean.I, we impute the missing numerical observations by the average of the observed variables. If the variables are categorical, it would be intuitive to use the mode (the most frequently observed categorical values). However, it always imputes the survival status as "success" as the one-shot devices are highly reliable. To have a more sensible imputation, we impute categorical variables by linear discrimination analysis, for which linear equations are formed based on other variables to discriminate the categorical responses.

For IPW, we use all the fully observed variables to perform logistic regression on non-missing indicators (1- the missing indicator of the missing variable). We then predict the probabilities of non-missingness, p_i and weight the score function by $1/p_i$ when we estimate the parameters for a one-shot device.

For MI, we use the default setting of `mice()` function from the `mice` package [48] in `R` program [49]. By default, the program creates linear equations for imputing missing data. Numerical variables are imputed by the predictive mean matching method (which gives the closest observed values to the missing value prediction based on a linear model). It also imputes the categorical values by the prediction from logistic regression, multinomial regression and proportional odds model regression for binary, nominal and ordinal variables, respectively. We then estimate the parameters, $\tilde{\beta}^{(m)}_{make}$ for the mth imputed data. We then use Rubin's Rules [25,26] mentioned in Section 3.2 to obtain the estimate and variance.

5.2.2. Our Proposed Methods

The other three methods are based on our proposed method MIUL with the unsupervised learning algorithms, k-prototype (K.MI), DBSCAN with Gower's distance (DB.MI) and Hierarchical Clustering with Gower's distance (HC.MI). Before clustering, we calculate the distance between data points by `daisy` function in the `cluster` package [50].

For k-prototype, we use the `kproto` function from the `clustMixType` package by [51]. We first group the data with the number of clusters $k = 3$ and then keep increasing the cluster numbers until the reduction of the within-cluster sum of squares is less than 10%.

For DBscan, we modify the suggestion by [52] and select the Eps considering the distance from each point to its third-nearest neighbor and define the minimum data points for a cluster, $MinPts$ as 50. We, then, use the `dbscan` function `fpc` package by [53] to perform DBSCAN clustering.

For Hierarchical Clustering, we use the `hclust` from the `stats` package by [49] and we cut trees to form five clusters.

After clustering, we use the MI methods mentioned above to impute data within each cluster. We combine the clusters as imputed datasets and follow Rubin's Rules again to infer the estimation.

5.3. Simulation Results

To evaluate the traditional methods and our proposed methods, we compare the biases and the mean squared error (MSE) of the car make coefficient β_{make}, which are defined as

$$Bias = \frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_{make,i} - \beta_{make}), \quad MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_{make,i} - \beta_{make})^2, \tag{11}$$

for the simulation size n, which is 1000 in our simulation study. We have six scenarios focusing on different missing data cases.

Scenario 1 focuses on the response variable, the survival statuses of one-shot devices, being missing at random. Tables A2–A5 show the bias and mean squared error of $\hat{\beta}^{B}_{make}$ and $\hat{\beta}^{C}_{make}$ when the response variable is missing at random. Although mean imputation generally has lower biases when the one-shot devices have higher reliability, our proposed methods become more accurate when we have more failure responses. The reason is quite apparent: the majority of the devices are successful and the mean imputation imputes the missing response as a success, which is the mode of the response. As a result, the imputed responses are most likely in favor of the success case and result in worse biases. When the reliability is high, the multiple imputations with unsupervised learning (MIUL) are generally better. However, when we consider the mean squared errors of $\hat{\beta}^{B}_{make}$ and $\hat{\beta}^{C}_{make}$, the traditional statistical methods do not have any advantage. The DB.MI performs better when the product reliability is high, while the K.MI works better when the product reliability is high.

Scenario 2 considers the survival statuses of one-shot devices being MNAR due to hidden variables. The simulation results are presented in Tables A6–A9. Considering the biases, they have similar patterns to the previous scenario. The mean imputation has smaller biases with low product reliability, while the MIUL algorithm gains an advantage when product reliability decreases. For MSE, MIUL algorithms generally perform better and, in addition, K.MI has superior performance when product reliability is low.

Scenario 3 considers the age of one-shot devices being missing at random. The simulation results are presented in Tables A10–A13. Similar to Scenarios 1 and 2, Mean.I has the most negligible bias when product reliability is low. However, as the reliability becomes high, DB.MI's performance suppresses other methods and HC.MI comes next. When we look at MSE, DB.MI has the smallest and HC.MI has the second smallest. This indicates that DB.MI and HC.MI outperforms K.MI when the missing variable is numeric instead of categorical.

Scenario 4 considers the age of one-shot devices being missing not at random due to the hidden variable and Tables A14–A17 summarize the results. Here, we have a similar conclusion as Scenario 3, that Mean.I has a low bias when the reliability is low. DB.MI works best regarding bias and MSE when the reliability is high and HC.MI is the second best. Occasionally, MI has less MSE compared to others.

Scenario 5 represents the case when the covariate Region is MAR and the simulation results are shown in Tables A18–A21. The mean imputation usually has the most insignificant biases when the product reliability is low. However, our proposed MIUL methods work relatively well for most cases and HC.MI is slightly better than other MIUL methods.

Scenario 6 represents the case when the covariate Region is missing not at random and we present the results in Tables A22–A25. Again, the Mean.I has the most negligible bias when the reliability of the one-shot device is low, while MIUL methods, especially HC.MI, have better performance in bias for low-reliability products. When we look at the MSE, DB.MI seems the best when the reliability is low to medium, while MI performs pretty well when the reliability is low.

The simulation shows that MIUL algorithms consistently outperform the traditional approach, namely the mean imputation, IPW and MI approaches for products. This tendency is more obvious when the products have a medium or high reliability. This is because high-reliability products produce fewer failure cases. Then, traditional methods tend to weight success cases higher and create biases as they cannot observe the latent variables in the dataset.

With MIUL, the clustering of the observations attempts to "discover" the latent variables and gives more flexible equations for imputing the missing variable. The target parameters β^{B}_{make} and β^{C}_{make} are estimated accurately with different variables under various missing mechanisms. Generally speaking, when the survival statuses of one-shot devices are missing, K.MI is the best; when the ages of one-shot devices are missing, both DB.MI and HC.MI work well; when a covariate is missing, HC.MI works well in most cases. This gives HC.MI the best overall performance among all MIUL algorithms.

To summarize, MIUL algorithms suppress the traditional missing data-handling methods, especially when the product reliability is medium to high, which is valid for most one-shot devices. The missing mechanism does not impact their relative performance much, suggesting that we prefer MIUL most of the time.

6. Applications with the CRSS Data

6.1. CRSS Datasets

In this section, we illustrate the application of the MIUL algorithm on the Crash Report Sampling System (CRSS) datasets 2016–2020. National Highway Traffic Safety Administration (NHTSA) developed and implemented CRSS to reduce the motor vehicle crash experience in the United States [54]. It is an annual survey designed independent of other NHTSA surveys. It is a valuable instrument for comprehending and analyzing crash data, yielding vital insights that can be used to enhance road safety. For simplicity reasons, we ignore the complex sample design and treat the data as independent in this illustration.

6.2. Defining the Airbag Success in CRSS Data

In CRSS, the airbag variable indicates when the airbag deploys or not during the accident. It does not explicitly imply whether the airbags function correctly. Here, we make a few modifications to the variable. First, we define the situation when the airbags should be deployed. It is reasonable to assume that an airbag should be deployed when the driver or passengers are critically injured. After all, airbags are designed to reduce the severity of injury. As the airbag sensor is at the front and due to the size of a car, the airbag should be deployed if the areas are impacted. We also assume the airbag should be deployed if the car has to be towed after the accident, as the collision is severe. We present the details of how such a a situation is exactly defined in Appendix B. Then, we can define the airbags as functioning correctly during the accident if the airbags deploy when they should or they do not deploy when they should not. Figure 4 presents the airbag success rate by manufacturing year for different accident years. The success rates are around 95%, which is reasonable. Each manufacturing year's airbag success rates are consistent with each other for different accident years. Thus, we can conclude that such a definition for airbag success looks reasonable.

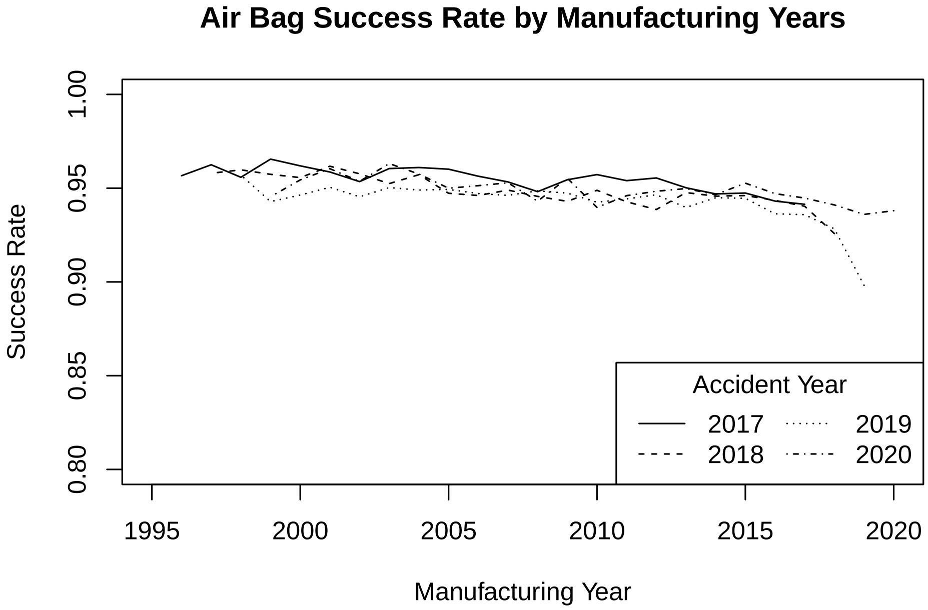


Figure 4. Airbag success rate by manufacturing year for the accidents that happened between 2017 and 2020.

6.3. Data Analysis

To simplify the modeling process, we use the same model (10) in the simulation study and we limit our target to comparing the airbag's reliability of the car makes' origins, namely, America, Asia and Europe. Throughout the accident years 2017–2020, the missing rates are 16.43%, 13.39%, 10.00% and 11.78%, respectively, and they are close to medium to low missing levels in the simulation study. Excluding the observations with missing variables, the success rate of the airbags is about 95%, as shown in Figure 4, confirming that airbags are highly reliable products. The CRSS generally publishes around 95,000 observations each year. Due to technical and resource limitations, we usually work on a dataset with sizes smaller than CRSS. In our illustration, we only sample 1000 (small), 2000 (medium) and 4000 (large) from the CRSS dataset. Then, we repeat this process 1000 times and record the estimates of the log hazard rate of airbags in Asian and European cars relative to American cars. There are two reasons for the sub-sampling. First, as the sample size increases, while there are some hidden variables that we cannot observe, the biases would increase as sample size increases which can be observed in the simulation study above. Therefore, we sample observations from large datasets and see if the estimate averages vary for different sample sizes. Second, it would allow a more straightforward confidence interval calculation using the bootstrap method while fewer computational resources are required. It also resembles the standard airbag tests of particular makes and models, as the sample sizes rarely exceed 4000. We report the means, standard errors (SEs) and the 95% confidence intervals (lower limit, CI.L and upper limit, CI.U) in Tables 1 and 2.

Table 1 presents the estimates of β_{make}^{Asian} under various imputation methods. The estimates are consistent for various sample sizes, indicating that a sample size of 2000 to 4000 should be enough for the parameter estimation. All the $\hat{\beta}_{make}^{Asian}$ are most likely around zero. Therefore, we can safely claim that the airbag motility of Asian brands is pretty much similar to that of American ones.

Next, we present the estimates of $\beta_{make}^{European}$ in Table 2. The estimates increase when the sample size increases for traditional methods, while the estimates from the proposed MIUL methods remain relatively stable. This suggests that the MIUL methods are more robust to the sample sizes. The CIs of $\hat{\beta}_{make}^{European}$ do not cover the zero under K.MI when the sample size is large for accident years 2019 and 2020 with the corresponding Wald statistics $\frac{0.503}{0.231} = 2.175$ (p-value$=$ 0.030) and $\frac{0.529}{0.253} = 2.095$ (p-value$=$ 0.036). Since K.MI usually has the most negligible bias and MSE, especially when the sample size and product reliability are high with medium to low missing levels, the quality of airbags of European manufacturers is likely significantly worse than the American ones, requiring further investigation.

Table 1. The mean, standard error (SE), 95% confidence interval lower and upper limits (CI.L, CI.U) estimates of β_{make}^{Asian} under various imputation methods with bootstrap samples of 1000 (Sml), 2000 (Med) and 4000 (Lrg).

			Methods					
Year	**Size**	**Stat**	**Mean.I**	**IPW**	**MI**	**K.MI**	**DB.MI**	**HC.MI**
2017	Sml	Mean	−0.00520	−0.00883	−0.01023	0.00636	−0.01036	−0.01041
		SE	0.35721	0.35774	0.35517	0.32478	0.35749	0.35748
		CI.L	−0.73138	−0.74240	−0.74146	−0.65902	−0.73862	−0.73884
		CI.U	0.70197	0.68470	0.69647	0.62163	0.68764	0.68775
	Med	Mean	0.01484	0.01049	0.00930	0.01770	0.01030	0.01026
		SE	0.23614	0.24019	0.23639	0.22898	0.23549	0.23552
		CI.L	−0.45937	−0.47115	−0.46515	−0.43956	−0.46083	−0.46396
		CI.U	0.46983	0.46826	0.45999	0.44759	0.46399	0.46300
	Lrg	Mean	0.02493	0.02057	0.01909	0.02714	0.02097	0.02092
		SE	0.16560	0.16903	0.16707	0.16564	0.16536	0.16541
		CI.L	−0.29847	−0.31457	−0.31544	−0.31185	−0.31176	−0.31219
		CI.U	0.33600	0.33924	0.33859	0.34427	0.33320	0.33312

Table 1. *Cont.*

					Methods			
Year	**Size**	**Stat**	**Mean.I**	**IPW**	**MI**	**K.MI**	**DB.MI**	**HC.MI**
2018	Sml	Mean	0.01970	0.02723	0.01685	0.02073	0.01631	0.01623
		SE	0.33288	0.33681	0.33263	0.30876	0.33149	0.33135
		CI.L	−0.63370	−0.62636	−0.63507	−0.57207	−0.63620	−0.63582
		CI.U	0.65528	0.67321	0.66805	0.58658	0.64048	0.64060
	Med	Mean	0.03063	0.03812	0.03144	0.03497	0.02886	0.02874
		SE	0.23144	0.23279	0.22925	0.22562	0.23077	0.23081
		CI.L	−0.41830	−0.40815	−0.42225	−0.41337	−0.41929	−0.41835
		CI.U	0.48970	0.49713	0.48528	0.47663	0.47902	0.47678
	Lrg	Mean	0.01974	0.02741	0.01967	0.02583	0.01745	0.01735
		SE	0.16032	0.16143	0.15951	0.15919	0.15976	0.15975
		CI.L	−0.30033	−0.29550	−0.30266	−0.30400	−0.30021	−0.30070
		CI.U	0.32769	0.34210	0.31599	0.33177	0.32383	0.32292
2019	Sml	Mean	0.02322	0.02289	0.02726	0.02150	0.02531	0.02537
		SE	0.30592	0.30570	0.30348	0.29186	0.30543	0.30532
		CI.L	−0.56986	−0.57100	−0.57995	−0.54529	−0.56919	−0.56984
		CI.U	0.60067	0.60591	0.61962	0.58495	0.60904	0.60940
	Med	Mean	0.03005	0.02988	0.03560	0.03289	0.03252	0.03247
		SE	0.22268	0.22280	0.22270	0.21788	0.22174	0.22169
		CI.L	−0.41365	−0.41014	−0.39257	−0.40485	−0.39461	−0.39340
		CI.U	0.45068	0.44856	0.46157	0.44709	0.45015	0.45006
	Lrg	Mean	0.03713	0.03695	0.04328	0.04274	0.03957	0.03952
		SE	0.14095	0.14145	0.14193	0.14064	0.14053	0.14055
		CI.L	−0.23866	−0.25076	−0.23560	−0.22885	−0.23287	−0.23235
		CI.U	0.31660	0.31528	0.32764	0.32541	0.32208	0.32310
2020	Sml	Mean	−0.05207	−0.05195	−0.04961	−0.04693	−0.04882	−0.04900
		SE	0.32679	0.32686	0.32499	0.31155	0.32559	0.32565
		CI.L	−0.75252	−0.75609	−0.72875	−0.71641	−0.73368	−0.73715
		CI.U	0.62621	0.61703	0.58905	0.55351	0.60640	0.61026
	Med	Mean	−0.05094	−0.05105	−0.04930	−0.04889	−0.04853	−0.04860
		SE	0.23119	0.23157	0.23300	0.22611	0.23053	0.23046
		CI.L	−0.49147	−0.49482	−0.50387	−0.47786	−0.48809	−0.48786
		CI.U	0.39535	0.40019	0.40279	0.40522	0.39801	0.39755
	Lrg	Mean	−0.06173	−0.06173	−0.06056	−0.05789	−0.05855	−0.05867
		SE	0.15746	0.15778	0.15726	0.15485	0.15631	0.15630
		CI.L	−0.40157	−0.40060	−0.39700	−0.39529	−0.38936	−0.38874
		CI.U	0.25195	0.25543	0.24474	0.23698	0.24169	0.23980

Table 2. The mean, standard error (SE), 95% confidence interval lower and upper limits (CI.L, CI.U) estimates of $\beta_{make}^{European}$ under various imputation methods with bootstrap samples of 1000 (Sml), 2000 (Med) and 4000 (Lrg)

					Methods			
Year	**Size**	**Stat**	**Mean.I**	**IPW**	**MI**	**K.MI**	**DB.MI**	**HC.MI**
2017	Sml	Mean	−0.02011	−0.03289	0.31172	0.70079	0.17739	0.17519
		SE	1.80428	1.86420	1.14748	0.56135	1.46658	1.46665
		CL.L	−8.06343	−8.22431	−1.48400	−0.39785	−1.84854	−1.84319
		CL.U	1.42867	1.53936	1.54589	1.62922	1.49423	1.48724
	Med	Mean	0.36758	0.38603	0.49093	0.67242	0.44369	0.44238
		SE	0.53164	0.59659	0.43307	0.37890	0.43006	0.43005
		CL.L	−0.64688	−0.86281	−0.41189	−0.12741	−0.46466	−0.48352
		CL.U	1.20380	1.25691	1.25447	1.34870	1.17280	1.17846
	Lrg	Mean	0.38532	0.41606	0.48753	0.60236	0.45673	0.45434
		SE	0.29940	0.34218	0.28816	0.27733	0.28529	0.28575
		CL.L	−0.24345	−0.32304	−0.13967	−0.01447	−0.15739	−0.15425
		CL.U	0.90793	1.03841	0.97901	1.08562	0.95691	0.95491

Table 2. *Cont.*

					Methods			
Year	**Size**	**Stat**	**Mean.I**	**IPW**	**MI**	**K.MI**	**DB.MI**	**HC.MI**
2018	Sml	Mean	0.16647	0.15027	0.35415	0.62988	0.25679	0.25582
		SE	1.34332	1.37701	0.92730	0.58264	1.13752	1.13778
		CL.L	−1.36720	−1.77539	−1.06060	−0.48064	−1.17999	−1.22472
		CL.U	1.38871	1.41256	1.41676	1.52952	1.36457	1.36863
	Med	Mean	0.36699	0.36244	0.45730	0.60685	0.41407	0.41211
		SE	0.41562	0.43450	0.40177	0.36590	0.39731	0.39717
		CL.L	−0.47720	−0.51849	−0.41055	−0.11368	−0.42388	−0.41376
		CL.U	1.07151	1.07604	1.15827	1.29320	1.10838	1.10896
	Lrg	Mean	0.37497	0.37219	0.45642	0.54508	0.41874	0.41656
		SE	0.28811	0.30315	0.27793	0.26883	0.27054	0.27107
		CL.L	−0.24697	−0.29888	−0.16556	−0.04139	−0.18245	−0.18405
		CL.U	0.86209	0.89625	0.94845	1.01856	0.89849	0.90448
2019	Sml	Mean	0.12799	0.14751	0.31003	0.59799	0.23611	0.23435
		SE	1.12483	1.14024	0.89727	0.50208	0.97898	0.97823
		CL.L	−1.28459	−1.37239	−1.07873	−0.47576	−1.17632	−1.18715
		CL.U	1.24107	1.27065	1.30010	1.43291	1.24504	1.24685
	Med	Mean	0.25313	0.27137	0.36703	0.51753	0.33018	0.32845
		SE	0.40623	0.41212	0.37021	0.34687	0.36915	0.36883
		CL.L	−0.62547	−0.55190	−0.41421	−0.19964	−0.43137	−0.42785
		CL.U	0.93457	0.94905	1.01682	1.14174	0.97396	0.97342
	Lrg	Mean	0.30675	0.32608	0.41438	**0.50307**	0.38252	0.37995
		SE	0.24617	0.24904	0.23975	**0.23126**	0.23244	0.23235
		CL.L	−0.21820	−0.18353	−0.07694	**0.02680**	−0.08056	−0.08262
		CL.U	0.75815	0.79900	0.85556	**0.90354**	0.79991	0.80472
2020	Sml	Mean	0.14460	0.16218	0.36551	0.64016	0.28710	0.28645
		SE	1.07676	1.08771	0.73292	0.56255	0.82508	0.82700
		CL.L	−1.23825	−1.31103	−0.84157	−0.40260	−0.96034	−1.03690
		CL.U	1.31753	1.31334	1.37122	1.51630	1.27246	1.29429
	Med	Mean	0.27679	0.29828	0.40339	0.57873	0.36809	0.36792
		SE	0.41222	0.41826	0.38554	0.36151	0.37847	0.37922
		CL.L	−0.61672	−0.55402	−0.40512	−0.14819	−0.46587	−0.47197
		CL.U	0.97036	0.99896	1.03570	1.20528	1.02671	1.01124
	Lrg	Mean	0.30138	0.32174	0.41495	**0.52914**	0.38813	0.38716
		SE	0.27387	0.27711	0.25597	**0.25260**	0.24996	0.24982
		CL.L	−0.28459	−0.27273	−0.14651	**0.01791**	−0.11541	−0.12547
		CL.U	0.78358	0.81176	0.88905	**0.98116**	0.83944	0.83161

7. Conclusions and Discussions

Retrospective studies are important in reliability analysis as they measure the actual lifetimes of the one-shot devices under the actual users' conditions. They can detect if the one-shot devices are designed and manufactured correctly and give early warnings if there are some systematic defaults in one-shot devices. This is crucial as most safety and life-saving products are one-shot devices. Early detection of potential flaws can save a lot of lives and casualties. In this study, we have proposed a retrospective way to analyze the reliability of one-shot devices through publicly available data.

Different from the usual reliability experiment, those datasets are not collected in a controlled environment, and therefore missing observations are inevitable. The traditional statistical methods may not handle the missing data properly since the observations may not be missing at random. With hidden variables, like parking locations and maintenance habits, those methods may not work well as the model cannot be specified correctly.

When machine learning algorithms are applied to impute the missing data, unsupervised learning may be useful to impute the missing observations. Still, it is not intuitive how the imputation can be carried out. Our study proposed an innovative way for missing data imputation using unsupervised learning, and it works reasonably well when hidden variables are present in the dataset under a missing-not-at-random assumption. With an accurate imputation strategy, retrospective studies on one-shot devices become possible.

Using a simulation study, we have shown that the MIUL methods perform superior to the traditional methods in the context of one-shot device reliability evaluation. We illustrate the methods using the CRSS datasets provided by NHTSA. Under a definition of airbag success, we find out that the airbags made by European cars may be significantly worse than those made by American manufacturers.

The CRSS datasets are collected through a complex survey design, which this research ignores for simplicity reasons. Therefore, in future studies, we could incorporate the survey design structure in the MIUL, enhancing the estimation accuracy. We could also extend the unsupervised learning part to more advanced algorithms like auto-encoder, and the self-organizing map, which may provide more precise results. It would also be interesting to see their estimation performance when the one-shot devices have Weibull and gamma lifetimes with frailty, which are more realistic than the exponential distribution.

This manuscript also provides a potential way to detect possible manufacturing issues with one-shot devices using public data. It would be interesting to regularly track the number of airbag failures and report on any defective models. One method is to monitor failures using control charts. Control charts are commonly used in quality control to identify underlying problems in industrial processes; see [55–57]. Modifying control charts for public datasets would be an intriguing issue worthy of more investigation.

Author Contributions: Writing—original draft, H.Y.S.; Writing—review & editing, M.H.L.; Supervision, N.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the University Research Committee of Oakland University (Grant Type: Faculty Research Fellows).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Datasets are available from the Crash Report Sampling System of National Highway Traffic Safety Administration. https://www.nhtsa.gov/crash-data-systems/crash-report-sampling-system accessed on 20 June 2023.

Acknowledgments: The first author thanks the University Research Committee of Oakland University for supporting this research. The authors appreciate the anonymous referees who provided useful and detailed comments on a previous version of the manuscript. The authors also thank the associate editor and editor for handling our paper.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Details of Simulation Settings

Appendix A.1. One-Shot Device Model Setting

The reliability function of one-shot devices given covariate $\boldsymbol{x}_i$:

$$R(t|\boldsymbol{x}) = \Pr(T > t|\boldsymbol{x}) = \exp(-t\lambda_{actual}(\boldsymbol{x}))$$

The actual hazard rates of one-shot devices:

$$\lambda_{actual}(\boldsymbol{x}) = \exp(\beta_0 + \beta_{make} + \beta_{region} + \beta_{park}),$$

where $\beta_0 = -6, -5$ and -4 represent high, medium and low reliability of the one-shot devices, respectively, $(\beta^B_{make}, \beta^C_{make}) = (-0.4, -0.8)$, $(\beta^{MW}_{region}, \beta^S_{region}, \beta^W_{region}) = (-0.6, 0.4, 0.8)$ and $\beta_{park} = 0.6$.

The probability of parking outside is

$$\Pr(Park = Out|\boldsymbol{x}) = 0.2 + 0.2 * I_{Rural} + 0.2 * I_{West} - 0.1 * I_{Sedan} + 0.2 * I_{Alcohol},$$

where the meanings of indicators are $I_{Rural}, \Pr(I_{Rural} = 1) = 23.5\%$, the car is in a rural area, $I_{West}, \Pr(I_{West} = 1) = 16.9\%$, the car is in the West region, $I_{Sedan}, \Pr(I_{Sedan} = 1) = 47.5\%$, the car is a Sedan, $I_{Alcohol}, \Pr(I_{Alcohol} = 1) = 6\%$, the driver is an alcohol drinker. We also assume that the driver's age following gamma distribution with shape and rate parameters is $(\alpha, \beta) = (20.01, 0.83)$, respectively, if the cars are parked outdoors or $(\alpha, \beta) = (12.04, 0.24)$ otherwise. Finally, *Car_Age*, the car age at the time of the accident, follows the following multinomial distribution.

Table A1. The distribution of car age at the time of the accident in the simulation study.

Car_Age	**0.5**	**1.5**	**2.5**	**3.5**	**4.5**	**5.5**	**6.5**	**7.5**	**8.5**	**9.5**	**10.5**	**11.5**	**12.5**	**13.5**	**14.5**	**15.5**	**16.5**
Prob.	0.36%	5.02%	7.65%	7.39%	7.42%	7.22%	4.94%	4.44%	4.08%	3.95%	4.11%	4.31%	4.61%	4.88%	4.64%	4.24%	3.79%
Car_Age	**17.5**	**18.5**	**19.5**	**20.5**	**21.5**	**22.5**	**23.5**	**24.5**	**25.5**	**26.5**	**27.5**	**28.5**	**29.5**	**30.5**	**31.5**	**32.5**	**33.5**
Prob.	3.37%	2.87%	2.36%	1.93%	1.56%	1.18%	0.88%	0.66%	0.47%	0.33%	0.28%	0.21%	0.16%	0.13%	0.11%	0.07%	0.06%
Car_Age	**34.5**	**35.5**	**36.5**	**37.5**	**38.5**	**39.5**	**40.5**	**41.5**	**42.5**	**43.5**	**44.5**	**45.5**	**46.5**	**47.5**	**48.5**	**49.5**	**50+**
Prob.	0.06%	0.04%	0.03%	0.02%	0.03%	0.02%	0.02%	0.02%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%

Appendix A.2. The Missing Mechanism for Different Scenarios

Scenario 1 in Section 5 stands for the survival statuses of one-shot devices being MAR,

$$\Pr(I_{Miss_i} = 1) = \alpha_0 + 0.05 * I_{Other} + 0.01 * \log(Car\_Age),$$

where $\alpha_0 = 0.05, 0.1$ and 0.2 represent low, medium and high missing levels, respectively, and the variable $I_{Other}, \Pr(I_{Other} = 1) = 32.7\%$ indicates that the car is not a Sedan nor a Utility Vehicle.

Scenario 2 in Section 5 stands for the survival statuses of one-shot devices being MNAR,

$$\Pr(I_{Miss_i} = 1) = \alpha_0 + 0.05 * I_{Other} + 0.01 * \log(Car\_Age) + 0.1 * I_{Out},$$

where $\alpha_0 = 0.05, 0.1$ and 0.2 represent low, medium and high missing levels, respectively, and the indicator $I_{Other} = 1$ if the car is not a Sedan nor a utility vehicle and $I_{Out} = 1$ if the car is parked outdoors.

Scenario 3 in Section 5 stands for the *Car_Age* being MAR,

$$\Pr(I_{Miss_i} = 1) = \alpha_0 + 0.01 * I_{South} + 0.05 * I_{Rural},$$

where $\alpha_0 = 0.05, 0.1$ and 0.2 represent low, medium and high missing levels, respectively, and the indicators, $I_{South}, \Pr(I_{South} = 1) = 53.6\%$, the accident location in the South region and I_{Rural} the car is in a rural area.

Scenario 4 in Section 5 stands for the *Car_Age* being MNAR,

$$\Pr(I_{Miss_i} = 1) = \alpha_0 + 0.01 * I_{South} + 0.05 * I_{Rural} + 0.1 * I_{Out},$$

where $\alpha_0 = 0.05, 0.1$ and 0.2 represent low, medium and high missing levels, respectively, and the indicators $I_{South} = 1$ if the accident is located in the South region, $I_{Rural} = 1$ if the accident location is in a rural area and $I_{Out} = 1$ if the car is parked outdoors.

Scenario 5 in Section 5 stands for the *Region* being MAR,

$$\Pr(I_{Miss_i} = 1) = \alpha_0 + 0.05 * I_{Rural},$$

where $\alpha_0 = 0.05, 0.1$ and 0.2 represent low, medium and high missing levels, respectively, and the indicator $I_{Rural} = 1$ means the car is in a rural area.

Scenario 6 in Section 5 stands for the *Region* being MNAR,

$$\Pr(I_{Miss_i} = 1) = \alpha_0 + 0.01 * I_{South} + 0.05 * I_{Rural} + 0.1 * I_{Out},$$

where $\alpha_0 = 0.05, 0.1$ and 0.2 represent low, medium, and high missing levels, respectively, and the meanings of indicators are $I_{South} = 1$ if the accident location is in the South region, $I_{Rural} = 1$ if the car is in a rural area and $I_{Out} = 1$ if the car is parked outdoors.

Appendix B. Details of the Real Data Analysis

For each year, the CRSS publishes more than 20 datasets, including `PERSON.CSV`, which contains motorists' and passengers' data, `VEHICLE.CSV`, which contains the data of the vehicles involved in the accidents and `VEVENT.CSV`, which contains the harmful and non-harmful events for the vehicles; see [2]. We focus on these three data files for real data analysis.

We define the variable `AirBag_Should` to indicate the situations in which the airbags should deploy which satisfy at least one of the following conditions:

1. If there are any motorists or passengers severely injured or dead (`INJ_SEV` in categories 3 or 4);
2. If the area of impact on the vehicle is not at the back (`AOI1` in categories 1–5 or 7–12);
3. if the car has to be towed (`TOWED` = 1).

The variable `AirBag_Deploy` is an indicator concerning whether there is any airbag deployed during the accident and the variable `AirBag_Success` indicates whether `AirBag_Deploy` is equal to `AirBag_Should`. It shows if the airbags work probably on a vehicle during an accident.

Then, we treat `AirBag_Success` as the one-shot device status, δ_i, and the vehicle age (accident year minus car model year) as the one-shot device's observed time t_i. The rest of the covariates, $\boldsymbol{x}_i$, are the origins of the car makes (America, Europe or Asia), accident region (Northeast, Midwest, South, West) and if it happens in the urban areas (urban or rural), the vehicle body type (Sedans, Sport Utility Vehicle and others) and the driver's age at the accident.

Then, we sample 1000, 2000 or 4000 observations from the dataset to mimic the situation in which smaller data are collected in practice. Then, we apply the traditional missing data-handling methods and our proposed method, MIUL.

The R code, which modifies the CRSS datasets, is posted on GitHub: https://github.com/hso-OU/OneShotMIUL/blob/main/MLOnshotDataAnalysisV2.R accessed on 20 June 2023. The dataset created, `CRSSData.RData`, is uploaded to Kaggle: https://www.kaggle.com/datasets/honyiuso/airbagcrss/data accessed on 20 June 2023.

Appendix C. Tables of Simulation Results

Table A2. Bias of $\hat{\beta}^B_{make}$ when the response variable is missing at random. Bold values represent the best result.

			Methods					
Relb.	**Miss.**	**Size**	**Mean.I**	**IPW**	**MI**	**K.MI**	**DB.MI**	**HC.MI**
Low	Hig	Sml	−**0.0227**	−0.0229	−0.0272	−0.0359	−0.0385	−0.0386
		Med	−**0.0263**	−0.0267	−0.0311	−0.0494	−0.0440	−0.0462
		Lrg	−**0.0245**	−0.0250	−0.0271	−0.0687	−0.0404	−0.0451
	Med	Sml	−**0.0225**	−0.0225	−0.0252	−0.0284	−0.0325	−0.0306
		Med	−**0.0279**	−0.0281	−0.0315	−0.0388	−0.0375	−0.0393
		Lrg	0.0001	**0.0001**	−0.0046	−0.0253	−0.0118	−0.0157
	Low	Sml	−**0.0233**	−0.0238	−0.0258	−0.0284	−0.0301	−0.0290
		Med	−**0.0235**	−0.0236	−0.0253	−0.0297	−0.0296	−0.0304
		Lrg	−0.0188	−**0.0184**	−0.0225	−0.0325	−0.0255	−0.0273
Med	Hig	Sml	−**0.0017**	−0.0022	−0.0064	−0.0417	−0.0215	−0.0268
		Med	0.0151	0.0143	0.0147	−0.0596	−**0.0070**	−0.0213
		Lrg	0.0408	0.0418	0.0285	−0.0982	0.0177	−**0.0079**
	Med	Sml	−**0.0125**	−0.0128	−0.0131	−0.0295	−0.0248	−0.0271
		Med	0.0072	0.0072	0.0034	−0.0323	−**0.0017**	−0.0148
		Lrg	0.0594	0.0594	0.0525	−0.0266	0.0356	**0.0232**
	Low	Sml	−**0.0068**	−0.0071	−0.0084	−0.0196	−0.0149	−0.0155
		Med	0.0216	0.0216	0.0187	−**0.0040**	0.0136	0.0089
		Lrg	0.0360	0.0362	0.0341	−0.0196	0.0252	**0.0131**

Table A2. *Cont.*

			Methods					
Relb.	**Miss.**	**Size**	**Mean.I**	**IPW**	**MI**	**K.MI**	**DB.MI**	**HC.MI**
Hig	Hig	Sml	0.0280	0.0273	0.0218	−0.0876	**0.0008**	−0.0229
		Med	0.0966	0.0965	0.0747	−0.1276	0.0566	**0.0208**
		Lrg	0.4869	0.4870	0.3553	**−0.1799**	0.2995	0.2463
	Med	Sml	0.0218	0.0218	0.0183	−0.0411	0.0098	**−0.0061**
		Med	0.0919	0.0915	0.0884	**−0.0431**	0.0691	0.0522
		Lrg	0.3949	0.3941	0.3289	**−0.0285**	0.3023	0.2683
	Low	Sml	0.0185	0.0186	0.0165	−0.0194	0.0094	**0.0015**
		Med	0.0670	0.0670	0.0622	**−0.0186**	0.0547	0.0406
		Lrg	0.3280	0.3289	0.2902	**0.0756**	0.2836	0.2698

Table A3. Bias of $\hat{\beta}^{C}_{make}$ when the response variable is missing at random. Bold values represent the best result.

			Methods					
Relb.	**Miss.**	**Size**	**Mean.I**	**IPW**	**MI**	**K.MI**	**DB.MI**	**HC.MI**
Low	Hig	Sml	−0.0057	**−0.0056**	−0.0070	−0.0140	−0.0175	−0.0185
		Med	**−0.0077**	−0.0080	−0.0096	−0.0209	−0.0218	−0.0224
		Lrg	−0.0047	−0.0049	**−0.0027**	−0.0318	−0.0149	−0.0185
	Med	Sml	−0.0068	**−0.0067**	−0.0077	−0.0100	−0.0143	−0.0132
		Med	**−0.0179**	−0.0182	−0.0199	−0.0241	−0.0245	−0.0270
		Lrg	0.0099	0.0103	0.0075	−0.0058	**0.0007**	−0.0024
	Low	Sml	**−0.0070**	−0.0073	−0.0076	−0.0106	−0.0120	−0.0111
		Med	**−0.0070**	−0.0072	−0.0076	−0.0103	−0.0117	−0.0123
		Lrg	−0.0020	**−0.0013**	−0.0035	−0.0094	−0.0062	−0.0069
Med	Hig	Sml	0.0046	0.0042	**0.0024**	−0.0235	−0.0170	−0.0192
		Med	0.0211	0.0205	0.0228	−0.0306	**−0.0024**	−0.0121
		Lrg	0.0425	0.0439	0.0355	−0.0550	0.0181	**0.0008**
	Med	Sml	−0.0079	−0.0081	**−0.0075**	−0.0187	−0.0213	−0.0221
		Med	0.0221	0.0221	0.0211	−0.0065	0.0109	**0.0024**
		Lrg	0.0608	0.0613	0.0555	**−0.0028**	0.0367	0.0279
	Low	Sml	−0.0005	−0.0007	**−0.0005**	−0.0096	−0.0090	−0.0090
		Med	0.0243	0.0239	0.0231	**0.0060**	0.0155	0.0125
		Lrg	0.0416	0.0417	0.0408	**−0.0001**	0.0318	0.0213
Hig	Hig	Sml	0.0284	0.0276	0.0257	−0.0552	**−0.0027**	−0.0183
		Med	0.0775	0.0773	0.0646	−0.0859	0.0361	**0.0083**
		Lrg	0.4698	0.4697	0.3459	**−0.0927**	0.2805	0.2407
	Med	Sml	0.0256	0.0255	0.0230	−0.0217	0.0122	**0.0000**
		Med	0.0901	0.0901	0.0916	**−0.0090**	0.0657	0.0553
		Lrg	0.3859	0.3854	0.3255	**0.0213**	0.2906	0.2670
	Low	Sml	0.0134	0.0136	0.0116	−0.0144	**0.0030**	−0.0035
		Med	0.0671	0.0671	0.0641	**0.0025**	0.0535	0.0424
		Lrg	0.3302	0.3312	0.2951	**0.1150**	0.2883	0.2764

Table A4. Mean squared error of $\hat{\beta}^{B}_{make}$ when the response variable is missing at random. Bold values represent the best result.

			Methods					
Relb.	**Miss.**	**Size**	**Mean.I**	**IPW**	**MI**	**K.MI**	**DB.MI**	**HC.MI**
Low	Hig	Sml	**0.0166**	0.0166	0.0174	0.0177	0.0170	0.0179
		Med	0.0344	0.0345	0.0357	0.0359	**0.0341**	0.0358
		Lrg	0.0630	0.0633	0.0654	0.0653	**0.0616**	0.0642
	Med	Sml	**0.0142**	0.0142	0.0146	0.0147	0.0144	0.0144
		Med	0.0273	**0.0272**	0.0280	0.0278	0.0275	0.0275
		Lrg	0.0608	0.0606	0.0628	**0.0583**	0.0589	0.0599
	Low	Sml	0.0140	**0.0139**	0.0142	0.0145	0.0140	0.0143
		Med	0.0305	0.0306	0.0308	0.0311	**0.0305**	0.0311
		Lrg	0.0517	0.0519	0.0517	0.0518	**0.0512**	0.0518

Table A4. *Cont.*

			Methods					
Relb.	**Miss.**	**Size**	**Mean.I**	**IPW**	**MI**	**K.MI**	**DB.MI**	**HC.MI**
Med	Hig	Sml	0.0396	0.0395	0.0411	0.0403	**0.0364**	0.0383
		Med	0.0786	0.0786	0.0808	**0.0710**	0.0711	0.0722
		Lrg	0.1664	0.1665	0.1648	**0.1255**	0.1473	0.1486
	Med	Sml	0.0318	0.0319	0.0332	0.0317	**0.0310**	0.0318
		Med	0.0699	0.0704	0.0695	**0.0648**	0.0665	0.0662
		Lrg	0.2305	0.2317	0.2306	**0.1284**	0.1578	0.2156
	Low	Sml	0.0301	0.0301	0.0303	0.0294	**0.0291**	0.0294
		Med	0.0686	0.0689	0.0687	**0.0649**	0.0665	0.0671
		Lrg	0.1486	0.1492	0.1497	**0.1268**	0.1418	0.1408
Hig	Hig	Sml	0.1048	0.1047	0.1088	**0.0880**	0.0906	0.0917
		Med	0.2947	0.2946	0.2702	**0.1527**	0.2079	0.2103
		Lrg	3.6690	3.6716	2.3723	**0.4207**	1.7349	1.8630
	Med	Sml	0.0855	0.0858	0.0876	**0.0792**	0.0805	0.0805
		Med	0.2686	0.2691	0.2654	**0.1714**	0.2272	0.2320
		Lrg	2.9573	2.9600	2.2744	**0.6249**	1.9409	1.9558
	Low	Sml	0.0837	0.0835	0.0843	**0.0776**	0.0806	0.0818
		Med	0.1916	0.1928	0.1905	**0.1542**	0.1782	0.1782
		Lrg	2.0587	2.0610	1.6831	**0.8159**	1.6829	1.7118

Table A5. Mean squared error of $\hat{\beta}^{B}_{make}$ when the response variable is missing not at random. Bold values represent the best result.

			Methods					
Relb.	**Miss.**	**Size**	**Mean.I**	**IPW**	**MI**	**K.MI**	**DB.MI**	**HC.MI**
Low	Hig	Sml	0.0175	0.0175	0.0181	0.0179	**0.0169**	0.0177
		Med	0.0340	0.0340	0.0350	0.0342	**0.0325**	0.0341
		Lrg	0.0656	0.0659	0.0683	0.0641	**0.0627**	0.0653
	Med	Sml	0.0147	0.0148	0.0152	0.0149	**0.0145**	0.0147
		Med	0.0284	0.0283	0.0293	0.0281	**0.0281**	0.0284
		Lrg	0.0618	0.0616	0.0638	**0.0580**	0.0601	0.0602
	Low	Sml	0.0144	0.0144	0.0146	0.0147	**0.0142**	0.0147
		Med	0.0300	0.0300	0.0299	0.0301	**0.0295**	0.0302
		Lrg	0.0542	0.0543	0.0540	0.0534	**0.0530**	0.0543
Med	Hig	Sml	0.0426	0.0425	0.0435	0.0414	**0.0393**	0.0406
		Med	0.0826	0.0823	0.0846	**0.0722**	0.0739	0.0747
		Lrg	0.1821	0.1822	0.1809	**0.1285**	0.1580	0.1634
	Med	Sml	0.0334	0.0334	0.0349	0.0334	**0.0318**	0.0333
		Med	0.0730	0.0734	0.0726	**0.0663**	0.0690	0.0689
		Lrg	0.2394	0.2403	0.2403	**0.1383**	0.1678	0.2231
	Low	Sml	0.0312	0.0313	0.0316	0.0307	**0.0299**	0.0308
		Med	0.0693	0.0696	0.0700	**0.0656**	0.0669	0.0674
		Lrg	0.1558	0.1563	0.1559	**0.1322**	0.1475	0.1469
Hig	Hig	Sml	0.1138	0.1137	0.1187	**0.0918**	0.0995	0.0997
		Med	0.2884	0.2884	0.2676	**0.1443**	0.2026	0.2119
		Lrg	3.6988	3.7023	2.4099	**0.4005**	1.7617	1.8864
	Med	Sml	0.0952	0.0955	0.0965	**0.0862**	0.0881	0.0889
		Med	0.2747	0.2754	0.2750	**0.1761**	0.2298	0.2366
		Lrg	2.9395	2.9415	2.2649	**0.6245**	1.9271	1.9404
	Low	Sml	0.0875	0.0874	0.0873	**0.0820**	0.0836	0.0853
		Med	0.2065	0.2075	0.2078	**0.1686**	0.1922	0.1918
		Lrg	2.0502	2.0525	1.6797	**0.8199**	1.6807	1.7181

Table A6. Bias of $\hat{\beta}^{B}_{make}$ when the response variable is missing not at random. Bold values represent the best result.

			Methods					
Relb.	**Miss.**	**Size**	**Mean.I**	**IPW**	**MI**	**K.MI**	**DB.MI**	**HC.MI**
Low	Hig	Sml	**−0.0184**	−0.0196	−0.0255	−0.0359	−0.0392	−0.0371
		Med	**−0.0198**	−0.0218	−0.0265	−0.0500	−0.0411	−0.0428
		Lrg	**−0.0181**	−0.0196	−0.0256	−0.0725	−0.0373	−0.0494

Table A6. *Cont.*

					Methods			
Relb.	**Miss.**	**Size**	**Mean.I**	**IPW**	**MI**	**K.MI**	**DB.MI**	**HC.MI**
	Med	Sml	**−0.0225**	−0.0238	−0.0275	−0.0328	−0.0358	−0.0338
		Med	**−0.0252**	−0.0262	−0.0303	−0.0416	−0.0387	−0.0374
		Lrg	0.0005	**−0.0004**	−0.0023	−0.0298	−0.0162	−0.0166
	Low	Sml	**−0.0216**	−0.0228	−0.0256	−0.0295	−0.0317	−0.0296
		Med	**−0.0206**	−0.0214	−0.0242	−0.0317	−0.0313	−0.0305
		Lrg	**−0.0136**	−0.0148	−0.0176	−0.0337	−0.0238	−0.0257
Med	Hig	Sml	0.0031	**0.0028**	−0.0032	−0.0410	−0.0171	−0.0269
		Med	0.0214	0.0208	0.0146	−0.0686	**−0.0052**	−0.0247
		Lrg	0.0601	0.0611	0.0408	−0.1146	0.0232	**−0.0097**
	Med	Sml	**−0.0059**	−0.0066	−0.0069	−0.0315	−0.0192	−0.0218
		Med	0.0116	0.0109	0.0044	−0.0374	**−0.0016**	−0.0170
		Lrg	0.0721	0.0711	0.0635	**−0.0290**	0.0507	0.0296
	Low	Sml	**−0.0033**	−0.0046	−0.0057	−0.0211	−0.0130	−0.0165
		Med	0.0236	0.0230	0.0186	−0.0146	0.0120	**0.0053**
		Lrg	0.0419	0.0415	0.0349	−0.0275	0.0240	**0.0150**
Hig	Hig	Sml	0.0358	0.0349	0.0245	−0.0996	**0.0054**	−0.0231
		Med	0.1227	0.1229	0.0965	−0.1489	0.0680	**0.0281**
		Lrg	0.5194	0.5175	0.3956	**−0.2238**	0.3316	0.2786
	Med	Sml	0.0277	0.0271	0.0194	−0.0475	**0.0089**	−0.0132
		Med	0.0888	0.0877	0.0777	−0.0722	0.0620	**0.0349**
		Lrg	0.4695	0.4695	0.3848	**−0.0871**	0.3268	0.2880
	Low	Sml	0.0166	0.0166	0.0159	−0.0332	**0.0037**	−0.0094
		Med	0.0800	0.0793	0.0705	**−0.0236**	0.0583	0.0399
		Lrg	0.3470	0.3496	0.3154	**0.0241**	0.2812	0.2654

Table A7. Bias of $\hat{\beta}^{C}_{make}$ when the response variable is missing not at random. Bold values represent the best result.

					Methods			
Relb.	**Miss.**	**Size**	**Mean.I**	**IPW**	**MI**	**K.MI**	**DB.MI**	**HC.MI**
Low	Hig	Sml	**−0.0047**	−0.0050	−0.0073	−0.0153	−0.0204	−0.0186
		Med	**−0.0037**	−0.0048	−0.0052	−0.0221	−0.0195	−0.0192
		Lrg	0.0057	**0.0050**	0.0051	−0.0302	−0.0087	−0.0164
	Med	Sml	**−0.0072**	−0.0079	−0.0090	−0.0134	−0.0169	−0.0147
		Med	**−0.0157**	−0.0160	−0.0167	−0.0251	−0.0260	−0.0243
		Lrg	0.0093	0.0089	0.0075	−0.0088	**−0.0014**	−0.0035
	Low	Sml	**−0.0064**	−0.0069	−0.0080	−0.0114	−0.0143	−0.0123
		Med	**−0.0077**	−0.0079	−0.0087	−0.0142	−0.0153	−0.0142
		Lrg	0.0012	**0.0007**	−0.0012	−0.0110	−0.0055	−0.0081
Med	Hig	Sml	0.0083	0.0083	**0.0056**	−0.0224	−0.0137	−0.0193
		Med	0.0290	0.0286	0.0247	−0.0366	**−0.0001**	−0.0130
		Lrg	0.0618	0.0626	0.0490	−0.0651	0.0248	**0.0006**
	Med	Sml	−0.0022	−0.0028	**−0.0018**	−0.0201	−0.0165	−0.0167
		Med	0.0259	0.0259	0.0213	−0.0070	0.0136	**−0.0003**
		Lrg	0.0690	0.0686	0.0656	**−0.0012**	0.0475	0.0329
	Low	Sml	0.0017	**0.0006**	0.0008	−0.0106	−0.0080	−0.0098
		Med	0.0268	0.0266	0.0236	**0.0002**	0.0149	0.0133
		Lrg	0.0455	0.0452	0.0412	**−0.0047**	0.0268	0.0209
Hig	Hig	Sml	0.0364	0.0351	0.0284	−0.0620	**0.0000**	−0.0173
		Med	0.1028	0.1030	0.0829	−0.1029	0.0464	**0.0170**
		Lrg	0.5150	0.5140	0.4107	**−0.1145**	0.3310	0.2895
	Med	Sml	0.0301	0.0297	0.0251	−0.0257	**0.0076**	−0.0080
		Med	0.0904	0.0906	0.0831	**−0.0297**	0.0633	0.0418
		Lrg	0.4556	0.4576	0.3814	**−0.0299**	0.3200	0.2834
	Low	Sml	0.0092	0.0091	0.0102	−0.0265	**−0.0052**	−0.0153
		Med	0.0769	0.0768	0.0709	**0.0012**	0.0542	0.0429
		Lrg	0.3490	0.3523	0.3242	**0.0799**	0.2845	0.2755

Table A8. Mean squared error of $\hat{\beta}^{B}_{make}$ when the covariate variable is missing not at random. Bold values represent the best result.

			Methods					
Relb.	**Miss.**	**Size**	**Mean.I**	**IPW**	**MI**	**K.MI**	**DB.MI**	**HC.MI**
Low	Hig	Sml	**0.0175**	0.0175	0.0183	0.0185	0.0177	0.0184
		Med	0.0358	0.0359	0.0376	0.0359	**0.0352**	0.0368
		Lrg	0.0698	0.0706	0.0716	**0.0674**	0.0682	0.0691
	Med	Sml	**0.0153**	0.0153	0.0159	0.0160	0.0158	0.0158
		Med	**0.0284**	0.0285	0.0299	0.0297	0.0289	0.0292
		Lrg	0.0637	0.0637	0.0641	0.0615	**0.0611**	0.0638
	Low	Sml	0.0140	**0.0140**	0.0142	0.0148	0.0146	0.0145
		Med	0.0310	0.0311	0.0320	0.0315	**0.0307**	0.0315
		Lrg	0.0567	0.0567	0.0583	**0.0547**	0.0551	0.0550
Med	Hig	Sml	0.0432	0.0430	0.0448	0.0411	**0.0397**	0.0414
		Med	0.0854	0.0852	0.0889	0.0752	0.0778	**0.0750**
		Lrg	0.3405	0.3445	0.2925	**0.1530**	0.2312	0.2076
	Med	Sml	0.0327	0.0328	0.0334	0.0327	**0.0321**	0.0325
		Med	0.0700	0.0700	0.0722	0.0652	0.0667	**0.0651**
		Lrg	0.1792	0.1794	0.1825	**0.1357**	0.1593	0.1521
	Low	Sml	0.0324	0.0322	0.0334	0.0325	**0.0314**	0.0317
		Med	0.0749	0.0755	0.0747	**0.0692**	0.0701	0.0730
		Lrg	0.1548	0.1548	0.1525	**0.1308**	0.1430	0.1450
Hig	Hig	Sml	0.1152	0.1153	0.1153	0.0972	**0.0971**	0.0999
		Med	0.5020	0.5045	0.3965	**0.1729**	0.2859	0.3570
		Lrg	3.6951	3.7095	2.3974	**0.3427**	1.8047	1.9310
	Med	Sml	0.0901	0.0900	0.0916	**0.0789**	0.0819	0.0810
		Med	0.2182	0.2185	0.2118	**0.1526**	0.1914	0.1796
		Lrg	3.5867	3.5971	2.6571	**0.4558**	2.1484	2.0664
	Low	Sml	0.0854	0.0855	0.0866	**0.0790**	0.0814	0.0819
		Med	0.2030	0.2026	0.2006	**0.1558**	0.1793	0.1801
		Lrg	2.3210	2.3310	2.0643	**0.7520**	1.7086	1.7976

Table A9. Mean squared error of $\hat{\beta}^{C}_{make}$ when the response variable is missing not at random. Bold values represent the best result.

			Methods					
Relb.	**Miss.**	**Size**	**Mean.I**	**IPW**	**MI**	**K.MI**	**DB.MI**	**HC.MI**
Low	Hig	Sml	0.0182	0.0182	0.0189	0.0187	**0.0175**	0.0184
		Med	0.0364	0.0365	0.0385	0.0354	**0.0350**	0.0367
		Lrg	0.0715	0.0720	0.0727	**0.0657**	0.0673	0.0685
	Med	Sml	0.0149	0.0149	0.0153	0.0150	**0.0147**	0.0149
		Med	0.0303	0.0302	0.0313	0.0308	**0.0296**	0.0303
		Lrg	0.0688	0.0684	0.0689	**0.0645**	0.0652	0.0679
	Low	Sml	0.0151	0.0150	**0.0149**	0.0155	0.0151	0.0152
		Med	0.0305	0.0305	0.0310	0.0305	**0.0297**	0.0309
		Lrg	0.0607	0.0606	0.0622	0.0585	**0.0579**	0.0588
Med	Hig	Sml	0.0440	0.0439	0.0460	0.0410	**0.0396**	0.0418
		Med	0.0927	0.0923	0.0954	**0.0758**	0.0827	0.0800
		Lrg	0.3527	0.3564	0.3072	**0.1518**	0.2463	0.2183
	Med	Sml	0.0352	0.0352	0.0363	0.0348	**0.0339**	0.0345
		Med	0.0750	0.0748	0.0777	**0.0688**	0.0696	0.0694
		Lrg	0.1922	0.1922	0.1978	**0.1461**	0.1696	0.1669
	Low	Sml	0.0336	0.0335	0.0344	0.0334	**0.0323**	0.0327
		Med	0.0751	0.0755	0.0752	**0.0694**	0.0696	0.0738
		Lrg	0.1583	0.1580	0.1570	**0.1335**	0.1456	0.1498
Hig	Hig	Sml	0.1229	0.1230	0.1229	**0.0982**	0.1035	0.1069
		Med	0.5040	0.5076	0.3997	**0.1653**	0.2901	0.3615
		Lrg	3.7086	3.7212	2.4320	**0.3239**	1.8169	1.9330
	Med	Sml	0.0982	0.0981	0.1001	**0.0848**	0.0889	0.0881
		Med	0.2304	0.2309	0.2265	**0.1562**	0.1997	0.1911
		Lrg	3.5898	3.6052	2.6613	**0.4513**	2.1473	2.0800
	Low	Sml	0.0911	0.0911	0.0940	**0.0841**	0.0869	0.0879
		Med	0.2169	0.2167	0.2159	**0.1671**	0.1909	0.1941
		Lrg	2.2912	2.3042	2.0378	**0.7515**	1.6862	1.7798

Table A10. Bias of $\hat{\beta}^{B}_{make}$ when *Car_Age* is missing at random. Bold values represent the best result.

			Methods					
Relb.	**Miss.**	**Size**	**Mean.I**	**IPW**	**MI**	**K.MI**	**DB.MI**	**HC.MI**
Low	Hig	Sml	−0.0088	−0.0088	−0.0095	**−0.0088**	−0.0135	−0.0123
		Med	−0.0097	−0.0101	**−0.0085**	−0.0097	−0.0129	−0.0135
		Lrg	**0.0027**	0.0027	−0.0059	−0.0051	−0.0099	−0.0095
	Med	Sml	**−0.0074**	−0.0075	−0.0090	−0.0089	−0.0107	−0.0113
		Med	**−0.0121**	−0.0122	−0.0173	−0.0168	−0.0196	−0.0192
		Lrg	0.0070	0.0071	0.0006	0.0025	0.0004	**−0.0004**
	Low	Sml	**−0.0053**	−0.0055	−0.0064	−0.0068	−0.0080	−0.0077
		Med	**−0.0056**	−0.0060	−0.0067	−0.0071	−0.0081	−0.0077
		Lrg	−0.0039	−0.0048	−0.0014	**−0.0013**	−0.0024	−0.0020
Med	Hig	Sml	0.0079	0.0080	**0.0005**	0.0008	−0.0011	−0.0011
		Med	0.0232	0.0231	0.0176	0.0189	**0.0163**	0.0164
		Lrg	0.0351	0.0336	0.0172	0.0180	0.0151	**0.0143**
	Med	Sml	−0.0070	−0.0072	−0.0069	**−0.0066**	−0.0078	−0.0075
		Med	0.0193	0.0193	0.0180	0.0186	**0.0163**	0.0169
		Lrg	0.0579	0.0581	0.0507	0.0509	0.0496	**0.0488**
	Low	Sml	0.0011	0.0011	−0.0003	**−0.0002**	−0.0005	−0.0006
		Med	0.0230	0.0231	0.0227	0.0229	0.0224	**0.0224**
		Lrg	0.0393	0.0389	0.0355	0.0361	**0.0349**	0.0349
Hig	Hig	Sml	0.0332	0.0330	0.0240	0.0250	0.0237	**0.0236**
		Med	0.0736	0.0737	0.0602	0.0615	**0.0586**	0.0588
		Lrg	0.4492	0.4517	0.2497	0.2498	**0.2471**	0.2477
	Med	Sml	0.0184	0.0185	0.0180	0.0182	**0.0177**	0.0177
		Med	0.0824	0.0821	0.0717	0.0727	**0.0707**	0.0711
		Lrg	0.3139	0.3147	0.2787	0.2790	**0.2772**	0.2785
	Low	Sml	0.0135	0.0138	0.0095	0.0097	**0.0093**	0.0096
		Med	0.0757	0.0757	0.0623	0.0628	**0.0621**	0.0625
		Lrg	0.3052	0.3055	0.2712	0.2713	0.2707	**0.2703**

Table A11. Bias of $\hat{\beta}^{C}_{make}$ when the *Car_Age* is missing at random. Bold values represent the best result.

			Methods					
Relb.	**Miss.**	**Size**	**Mean.I**	**IPW**	**MI**	**K.MI**	**DB.MI**	**HC.MI**
Low	Hig	Sml	0.0164	0.0163	**0.0131**	0.0131	0.0135	0.0136
		Med	0.0336	0.0338	0.0271	0.0271	**0.0267**	0.0271
		Lrg	0.0664	0.0661	0.0511	0.0507	**0.0502**	0.0505
	Med	Sml	0.0145	0.0145	**0.0130**	0.0131	0.0133	0.0132
		Med	0.0256	0.0255	**0.0228**	0.0228	0.0230	0.0231
		Lrg	0.0587	0.0585	0.0508	0.0510	**0.0500**	0.0505
	Low	Sml	0.0130	0.0129	**0.0124**	0.0124	0.0125	0.0125
		Med	0.0287	0.0285	0.0268	0.0268	0.0268	**0.0267**
		Lrg	0.0497	0.0497	0.0467	**0.0466**	0.0467	0.0467
Med	Hig	Sml	0.0372	0.0372	0.0296	0.0294	0.0294	**0.0292**
		Med	0.0736	0.0737	0.0579	0.0580	**0.0574**	0.0577
		Lrg	0.1588	0.1579	0.1192	**0.1190**	0.1195	0.1190
	Med	Sml	0.0314	0.0314	0.0277	0.0277	0.0275	**0.0275**
		Med	0.0650	0.0648	0.0602	0.0601	**0.0597**	0.0599
		Lrg	0.1417	0.1425	0.1278	0.1280	0.1272	**0.1271**
	Low	Sml	0.0304	0.0305	0.0275	0.0276	0.0276	**0.0275**
		Med	0.0672	0.0670	0.0633	0.0636	**0.0633**	0.0633
		Lrg	0.1387	0.1385	0.1273	0.1278	**0.1270**	0.1271
Hig	Hig	Sml	0.0993	0.0997	0.0747	0.0750	**0.0745**	0.0747
		Med	0.3044	0.3057	0.2353	0.2352	**0.2339**	0.2341
		Lrg	3.2773	3.2961	1.4935	1.4906	**1.4861**	1.4905
	Med	Sml	0.0814	0.0817	0.0697	0.0695	0.0693	**0.0693**
		Med	0.1948	0.1937	0.1695	0.1697	0.1693	**0.1692**
		Lrg	2.3989	2.4021	2.0547	2.0512	2.0458	**2.0457**
	Low	Sml	0.0822	0.0823	0.0729	0.0729	0.0730	**0.0728**
		Med	0.1860	0.1862	0.1678	0.1678	0.1674	**0.1674**
		Lrg	1.9062	1.9011	1.6304	1.6306	1.6281	**1.6278**

Table A12. Mean squared error of $\hat{\beta}^{B}_{make}$ when *Car_Age* is missing at random. Bold values represent the best result.

						Methods		
Relb.	**Miss.**	**Size**	**Mean.I**	**IPW**	**MI**	**K.MI**	**DB.MI**	**HC.MI**
Low	Hig	Sml	0.0164	0.0163	**0.0131**	0.0131	0.0135	0.0136
		Med	0.0336	0.0338	0.0271	0.0271	**0.0267**	0.0271
		Lrg	0.0664	0.0661	0.0511	0.0507	**0.0502**	0.0505
	Med	Sml	0.0145	0.0145	**0.0130**	0.0131	0.0133	0.0132
		Med	0.0256	0.0255	**0.0228**	0.0228	0.0230	0.0231
		Lrg	0.0587	0.0585	0.0508	0.0510	**0.0500**	0.0505
	Low	Sml	0.0130	0.0129	**0.0124**	0.0124	0.0125	0.0125
		Med	0.0287	0.0285	0.0268	0.0268	0.0268	**0.0267**
		Lrg	0.0497	0.0497	0.0467	**0.0466**	0.0467	0.0467
Med	Hig	Sml	0.0372	0.0372	0.0296	0.0294	0.0294	**0.0292**
		Med	0.0736	0.0737	0.0579	0.0580	**0.0574**	0.0577
		Lrg	0.1588	0.1579	0.1192	**0.1190**	0.1195	0.1190
	Med	Sml	0.0314	0.0314	0.0277	0.0277	0.0275	**0.0275**
		Med	0.0650	0.0648	0.0602	0.0601	**0.0597**	0.0599
		Lrg	0.1417	0.1425	0.1278	0.1280	0.1272	**0.1271**
	Low	Sml	0.0304	0.0305	0.0275	0.0276	0.0276	**0.0275**
		Med	0.0672	0.0670	0.0633	0.0636	**0.0633**	0.0633
		Lrg	0.1387	0.1385	0.1273	0.1278	**0.1270**	0.1271
Hig	Hig	Sml	0.0993	0.0997	0.0747	0.0750	**0.0745**	0.0747
		Med	0.3044	0.3057	0.2353	0.2352	**0.2339**	0.2341
		Lrg	3.2773	3.2961	1.4935	1.4906	**1.4861**	1.4905
	Med	Sml	0.0814	0.0817	0.0697	0.0695	0.0693	**0.0693**
		Med	0.1948	0.1937	0.1695	0.1697	0.1693	**0.1692**
		Lrg	2.3989	2.4021	2.0547	2.0512	2.0458	**2.0457**
	Low	Sml	0.0822	0.0823	0.0729	0.0729	0.0730	**0.0728**
		Med	0.1860	0.1862	0.1678	0.1678	0.1674	**0.1674**
		Lrg	1.9062	1.9011	1.6304	1.6306	1.6281	**1.6278**

Table A13. Mean squared error of $\hat{\beta}^{C}_{make}$ when *Car_Age* is missing at random. Bold values represent the best result.

						Methods		
Relb.	**Miss.**	**Size**	**Mean.I**	**IPW**	**MI**	**K.MI**	**DB.MI**	**HC.MI**
Low	Hig	Sml	0.0167	0.0167	0.0133	0.0133	**0.0131**	0.0132
		Med	0.0324	0.0325	0.0259	0.0258	**0.0251**	0.0256
		Lrg	0.0660	0.0659	0.0511	0.0516	**0.0504**	0.0509
	Med	Sml	0.0145	0.0145	0.0130	0.0130	0.0129	**0.0128**
		Med	0.0269	0.0268	**0.0236**	0.0237	0.0236	0.0236
		Lrg	0.0589	0.0587	0.0518	0.0522	**0.0514**	0.0520
	Low	Sml	0.0136	0.0136	0.0131	0.0131	0.0131	**0.0130**
		Med	0.0285	0.0284	0.0263	0.0263	0.0262	**0.0261**
		Lrg	0.0533	0.0531	0.0496	0.0497	**0.0495**	0.0497
Med	Hig	Sml	0.0387	0.0387	0.0316	0.0317	**0.0315**	0.0315
		Med	0.0760	0.0761	0.0608	0.0609	**0.0601**	0.0607
		Lrg	0.1745	0.1738	0.1304	0.1307	0.1303	**0.1299**
	Med	Sml	0.0336	0.0337	0.0292	0.0292	**0.0290**	0.0290
		Med	0.0694	0.0692	0.0638	0.0640	**0.0636**	0.0638
		Lrg	0.1540	0.1544	0.1387	0.1390	**0.1383**	0.1384
	Low	Sml	0.0323	0.0324	0.0289	0.0290	0.0289	**0.0289**
		Med	0.0694	0.0692	0.0645	0.0646	**0.0643**	0.0643
		Lrg	0.1426	0.1425	0.1303	0.1311	0.1300	**0.1300**
Hig	Hig	Sml	0.1087	0.1090	0.0811	0.0814	0.0810	**0.0809**
		Med	0.2988	0.3003	0.2308	0.2308	**0.2293**	0.2298
		Lrg	3.2300	3.2501	1.4881	1.4870	**1.4832**	1.4833
	Med	Sml	0.0886	0.0890	0.0769	0.0770	**0.0766**	0.0768
		Med	0.2025	0.2015	0.1759	0.1764	**0.1756**	0.1756
		Lrg	2.3960	2.4005	2.0551	2.0513	2.0465	**2.0463**
	Low	Sml	0.0852	0.0852	0.0763	0.0763	0.0764	**0.0761**
		Med	0.1972	0.1980	0.1809	0.1808	**0.1805**	0.1805
		Lrg	1.9007	1.8979	1.6243	1.6234	1.6215	**1.6214**

Table A14. Bias of $\hat{\beta}^{B}_{make}$ when *Car_Age* is missing not at random. Bold values represent the best result.

			Methods					
Relb.	**Miss.**	**Size**	**Mean.I**	**IPW**	**MI**	**K.MI**	**DB.MI**	**HC.MI**
Low	Hig	Sml	**−0.0218**	−0.0251	−0.0255	−0.0260	−0.0375	−0.0367
		Med	**−0.0273**	−0.0308	−0.0286	−0.0291	−0.0392	−0.0392
		Lrg	**−0.0204**	−0.0244	−0.0301	−0.0304	−0.0381	−0.0369
	Med	Sml	**−0.0199**	−0.0226	−0.0251	−0.0257	−0.0323	−0.0320
		Med	**−0.0236**	−0.0262	−0.0295	−0.0302	−0.0360	−0.0357
		Lrg	**−0.0027**	−0.0068	−0.0126	−0.0119	−0.0182	−0.0174
	Low	Sml	**−0.0206**	−0.0232	−0.0243	−0.0250	−0.0288	−0.0284
		Med	**−0.0186**	−0.0213	−0.0240	−0.0243	−0.0275	−0.0277
		Lrg	**−0.0165**	−0.0202	−0.0193	−0.0184	−0.0225	−0.0226
Med	Hig	Sml	**−0.0026**	−0.0037	−0.0094	−0.0081	−0.0128	−0.0130
		Med	0.0242	0.0217	0.0082	0.0094	0.0054	**0.0045**
		Lrg	0.0325	0.0292	0.0165	0.0180	**0.0122**	0.0125
	Med	Sml	**−0.0091**	−0.0110	−0.0124	−0.0116	−0.0142	−0.0140
		Med	0.0123	0.0111	0.0017	0.0025	−0.0011	**−0.0005**
		Lrg	0.0642	0.0612	0.0489	0.0497	**0.0457**	0.0459
	Low	Sml	**−0.0066**	−0.0086	−0.0091	−0.0086	−0.0101	−0.0098
		Med	0.0207	0.0194	0.0167	0.0176	**0.0153**	0.0162
		Lrg	0.0352	0.0322	0.0292	0.0305	0.0281	**0.0279**
Hig	Hig	Sml	0.0317	0.0306	0.0203	0.0218	**0.0180**	0.0182
		Med	0.1006	0.0984	0.0750	0.0753	0.0735	**0.0726**
		Lrg	0.4817	0.4820	0.2513	0.2541	**0.2493**	0.2504
	Med	Sml	0.0229	0.0224	0.0147	0.0148	**0.0137**	0.0139
		Med	0.0973	0.0959	0.0719	0.0735	0.0715	**0.0714**
		Lrg	0.4476	0.4495	0.2896	0.2903	**0.2874**	0.2881
	Low	Sml	0.0156	0.0153	**0.0133**	0.0136	0.0134	0.0134
		Med	0.0720	0.0703	0.0609	0.0613	**0.0597**	0.0600
		Lrg	0.3140	0.3170	0.2676	0.2695	0.2674	**0.2672**

Table A15. Bias of $\hat{\beta}^{C}_{make}$ when *Car_Age* is missing not at random. Bold values represent the best result.

			Methods					
Relb.	**Miss.**	**Size**	**Mean.I**	**IPW**	**MI**	**K.MI**	**DB.MI**	**HC.MI**
Low	Hig	Sml	−0.0095	−0.0110	−0.0103	**−0.0095**	−0.0142	−0.0137
		Med	−0.0130	−0.0144	−0.0097	**−0.0096**	−0.0138	−0.0135
		Lrg	**−0.0013**	−0.0035	−0.0060	−0.0068	−0.0092	−0.0085
	Med	Sml	**−0.0049**	−0.0062	−0.0084	−0.0083	−0.0112	−0.0107
		Med	**−0.0134**	−0.0143	−0.0174	−0.0177	−0.0195	−0.0196
		Lrg	0.0062	0.0046	−0.0014	**−0.0011**	−0.0042	−0.0035
	Low	Sml	**−0.0039**	−0.0049	−0.0071	−0.0075	−0.0088	−0.0090
		Med	**−0.0048**	−0.0056	−0.0079	−0.0080	−0.0094	−0.0094
		Lrg	−0.0025	−0.0045	−0.0027	**−0.0017**	−0.0034	−0.0033
Med	Hig	Sml	**0.0003**	0.0003	−0.0021	−0.0007	−0.0026	−0.0033
		Med	0.0295	0.0278	0.0157	0.0163	0.0146	**0.0135**
		Lrg	0.0283	0.0261	0.0200	0.0206	**0.0185**	0.0188
	Med	Sml	**−0.0046**	−0.0057	−0.0074	−0.0068	−0.0081	−0.0075
		Med	0.0249	0.0244	0.0177	0.0185	0.0169	**0.0167**
		Lrg	0.0622	0.0608	0.0506	0.0510	**0.0485**	0.0488
	Low	Sml	**−0.0001**	−0.0010	−0.0014	−0.0010	−0.0015	−0.0012
		Med	0.0244	0.0243	0.0218	0.0224	**0.0211**	0.0217
		Lrg	0.0421	0.0404	0.0351	0.0356	0.0345	**0.0344**
Hig	Hig	Sml	0.0339	0.0332	0.0233	0.0243	**0.0219**	0.0223
		Med	0.0794	0.0780	0.0587	0.0598	0.0586	**0.0576**
		Lrg	0.4818	0.4833	0.2473	0.2492	**0.2462**	0.2469
	Med	Sml	0.0272	0.0271	0.0177	0.0181	0.0176	**0.0174**
		Med	0.0932	0.0922	0.0718	0.0729	**0.0717**	0.0719
		Lrg	0.4359	0.4392	0.2777	0.2777	**0.2760**	0.2767
	Low	Sml	0.0109	0.0112	**0.0102**	0.0103	0.0104	0.0102
		Med	0.0713	0.0705	0.0616	0.0620	**0.0611**	0.0614
		Lrg	0.3170	0.3207	0.2716	0.2725	0.2710	**0.2708**

Table A16. Mean squared error of $\hat{\beta}^{B}_{make}$ when *Car_Age* is missing not at random. Bold values represent the best result.

			Methods					
Relb.	**Miss.**	**Size**	**Mean.I**	**IPW**	**MI**	**K.MI**	**DB.MI**	**HC.MI**
Low	Hig	Sml	0.0158	0.0159	0.0133	**0.0131**	0.0137	0.0137
		Med	0.0331	0.0332	0.0264	**0.0262**	0.0266	0.0268
		Lrg	0.0673	0.0673	0.0489	0.0493	0.0489	**0.0479**
	Med	Sml	0.0153	0.0153	**0.0130**	0.0130	0.0132	0.0132
		Med	0.0278	0.0278	0.0234	**0.0234**	0.0234	0.0234
		Lrg	0.0606	0.0603	0.0513	0.0512	**0.0507**	0.0509
	Low	Sml	0.0136	0.0135	**0.0125**	0.0126	0.0129	0.0127
		Med	0.0296	0.0294	0.0268	0.0266	0.0267	**0.0265**
		Lrg	0.0530	0.0532	0.0468	0.0471	**0.0464**	0.0466
Med	Hig	Sml	0.0390	0.0389	0.0294	0.0294	0.0292	**0.0292**
		Med	0.0787	0.0781	0.0599	0.0594	**0.0592**	0.0592
		Lrg	0.1697	0.1686	0.1201	0.1195	**0.1180**	0.1187
	Med	Sml	0.0333	0.0332	0.0275	0.0276	0.0275	**0.0274**
		Med	0.0735	0.0730	0.0598	**0.0593**	0.0593	0.0594
		Lrg	0.1707	0.1705	0.1277	0.1277	**0.1270**	0.1275
	Low	Sml	0.0305	0.0305	0.0277	**0.0277**	0.0277	0.0277
		Med	0.0692	0.0692	0.0633	0.0634	**0.0631**	0.0633
		Lrg	0.1439	0.1429	0.1267	0.1271	**0.1267**	0.1269
Hig	Hig	Sml	0.1055	0.1048	0.0759	0.0758	0.0757	**0.0756**
		Med	0.3113	0.3117	0.2344	0.2334	**0.2330**	0.2340
		Lrg	3.5278	3.5578	1.4887	1.4911	**1.4824**	1.4866
	Med	Sml	0.0897	0.0898	0.0698	0.0698	**0.0694**	0.0696
		Med	0.3240	0.3264	0.1697	0.1704	**0.1695**	0.1695
		Lrg	3.3592	3.3801	2.0526	2.0500	2.0474	**2.0448**
	Low	Sml	0.0836	0.0835	**0.0734**	0.0735	0.0735	0.0735
		Med	0.1892	0.1887	0.1673	0.1675	0.1672	**0.1671**
		Lrg	2.0090	2.0207	1.6320	1.6313	**1.6305**	1.6307

Table A17. Mean squared error of $\hat{\beta}^{C}_{make}$ when *Car_Age* is missing not at random. Bold values represent the best result.

			Methods					
Relb.	**Miss.**	**Size**	**Mean.I**	**IPW**	**MI**	**K.MI**	**DB.MI**	**HC.MI**
Low	Hig	Sml	0.0159	0.0159	0.0134	**0.0133**	0.0134	0.0133
		Med	0.0326	0.0326	0.0256	0.0255	**0.0251**	0.0253
		Lrg	0.0686	0.0683	0.0502	0.0505	0.0499	**0.0487**
	Med	Sml	0.0152	0.0151	0.0129	0.0128	**0.0128**	0.0129
		Med	0.0292	0.0290	0.0241	0.0240	0.0238	**0.0237**
		Lrg	0.0609	0.0606	0.0525	**0.0520**	0.0522	0.0521
	Low	Sml	0.0143	0.0141	0.0133	0.0133	0.0134	**0.0133**
		Med	0.0290	0.0288	0.0262	0.0260	0.0261	**0.0260**
		Lrg	0.0559	0.0560	0.0501	0.0503	**0.0493**	0.0497
Med	Hig	Sml	0.0409	0.0407	0.0317	0.0318	**0.0316**	0.0317
		Med	0.0845	0.0841	0.0630	0.0625	0.0623	**0.0621**
		Lrg	0.1861	0.1859	0.1304	0.1306	**0.1284**	0.1293
	Med	Sml	0.0349	0.0347	0.0293	0.0294	0.0292	**0.0291**
		Med	0.0777	0.0772	0.0635	0.0633	**0.0631**	0.0632
		Lrg	0.1769	0.1755	0.1386	**0.1380**	0.1385	0.1385
	Low	Sml	0.0327	0.0329	0.0290	0.0290	0.0289	**0.0289**
		Med	0.0698	0.0697	**0.0645**	0.0647	0.0645	0.0646
		Lrg	0.1507	0.1496	0.1304	0.1307	0.1302	**0.1299**
Hig	Hig	Sml	0.1125	0.1119	0.0818	**0.0816**	0.0817	0.0816
		Med	0.3075	0.3081	0.2301	0.2297	**0.2291**	0.2305
		Lrg	3.4957	3.5303	1.4867	1.4878	**1.4766**	1.4827
	Med	Sml	0.0984	0.0983	0.0769	0.0771	**0.0768**	0.0769
		Med	0.3257	0.3281	0.1753	0.1758	**0.1751**	0.1751
		Lrg	3.3488	3.3750	2.0510	2.0471	2.0465	**2.0436**
	Low	Sml	0.0865	0.0863	**0.0766**	0.0768	0.0767	0.0767
		Med	0.2045	0.2044	0.1806	0.1810	**0.1804**	0.1806
		Lrg	2.0213	2.0345	1.6259	1.6250	1.6242	**1.6235**

Table A18. Bias of $\hat{\beta}^{B}_{make}$ when *Region* is missing at random. Bold values represent the best result.

			Methods					
Relb.	**Miss.**	**Size**	**Mean.I**	**IPW**	**MI**	**K.MI**	**DB.MI**	**HC.MI**
Low	Hig	Sml	−**0.0213**	−0.0218	−0.0245	−0.0235	−0.0300	−0.0303
		Med	−**0.0233**	−0.0237	−0.0269	−0.0261	−0.0322	−0.0324
		Lrg	−**0.0232**	−0.0235	−0.0278	−0.0274	−0.0329	−0.0340
	Med	Sml	−0.0245	−0.0249	−0.0250	−**0.0244**	−0.0281	−0.0280
		Med	−0.0286	−0.0289	−0.0287	−**0.0283**	−0.0317	−0.0319
		Lrg	−**0.0073**	−0.0086	−0.0082	−0.0081	−0.0114	−0.0124
	Low	Sml	−**0.0233**	−0.0236	−0.0240	−0.0236	−0.0258	−0.0258
		Med	−**0.0216**	−0.0219	−0.0226	−0.0222	−0.0242	−0.0244
		Lrg	−**0.0153**	−0.0158	−0.0178	−0.0176	−0.0194	−0.0194
Med	Hig	Sml	−**0.0018**	−0.0020	−0.0074	−0.0068	−0.0092	−0.0095
		Med	0.0271	0.0269	0.0100	0.0106	0.0084	**0.0076**
		Lrg	0.0327	0.0326	0.0174	0.0167	0.0150	**0.0138**
	Med	Sml	−**0.0102**	−0.0104	−0.0108	−0.0105	−0.0121	−0.0123
		Med	0.0057	0.0050	0.0020	0.0023	0.0007	**0.0005**
		Lrg	0.0555	0.0548	0.0497	0.0497	0.0481	**0.0478**
	Low	Sml	−0.0080	−0.0082	−0.0080	−**0.0079**	−0.0086	−0.0087
		Med	0.0203	0.0201	0.0181	0.0182	**0.0172**	0.0173
		Lrg	0.0372	0.0361	0.0306	0.0303	0.0300	**0.0293**
Hig	Hig	Sml	0.0227	0.0225	0.0211	0.0213	0.0202	**0.0201**
		Med	0.0926	0.0915	0.0757	0.0754	0.0747	**0.0743**
		Lrg	0.4200	0.4224	0.2515	0.2511	0.2517	**0.2508**
	Med	Sml	0.0207	0.0209	0.0151	0.0152	0.0147	**0.0147**
		Med	0.0801	0.0798	0.0720	0.0720	**0.0713**	0.0716
		Lrg	0.3601	0.3594	0.2913	**0.2910**	0.2916	0.2913
	Low	Sml	0.0181	0.0185	0.0131	0.0132	**0.0129**	0.0130
		Med	0.0712	0.0709	0.0623	0.0622	**0.0618**	0.0620
		Lrg	0.2806	0.2789	0.2684	0.2683	0.2688	**0.2682**

Table A19. Bias of $\hat{\beta}^{C}_{make}$ when *Region* is missing at random. Bold values represent the best result.

			Methods					
Relb.	**Miss.**	**Size**	**Mean.I**	**IPW**	**MI**	**K.MI**	**DB.MI**	**HC.MI**
Low	Hig	Sml	−**0.0062**	−0.0065	−0.0083	−0.0082	−0.0105	−0.0108
		Med	−**0.0064**	−0.0065	−0.0078	−0.0082	−0.0101	−0.0104
		Lrg	0.0017	**0.0016**	−0.0044	−0.0046	−0.0069	−0.0076
	Med	Sml	−0.0092	−0.0094	−0.0085	−**0.0084**	−0.0095	−0.0095
		Med	−0.0186	−0.0187	−0.0174	−**0.0173**	−0.0183	−0.0184
		Lrg	0.0023	0.0014	0.0018	0.0016	0.0005	−**0.0004**
	Low	Sml	−**0.0058**	−0.0060	−0.0070	−0.0069	−0.0078	−0.0078
		Med	−**0.0062**	−0.0063	−0.0069	−0.0067	−0.0076	−0.0075
		Lrg	0.0014	**0.0010**	−0.0017	−0.0018	−0.0025	−0.0023
Med	Hig	Sml	0.0023	0.0023	−0.0005	−**0.0004**	−0.0009	−0.0012
		Med	0.0311	0.0310	0.0163	0.0164	0.0160	**0.0153**
		Lrg	0.0338	0.0335	0.0200	0.0184	0.0192	**0.0179**
	Med	Sml	−**0.0057**	−0.0060	−0.0060	−0.0058	−0.0066	−0.0067
		Med	0.0189	0.0183	0.0177	0.0177	0.0173	**0.0170**
		Lrg	0.0552	0.0548	0.0493	0.0491	0.0487	**0.0484**
	Low	Sml	4×10^{-4}	4×10^{-4}	-4×10^{-4}	-4×10^{-4}	-5×10^{-4}	-7×10^{-4}
		Med	0.0231	0.0228	0.0225	0.0223	**0.0221**	0.0223
		Lrg	0.0407	0.0399	0.0353	0.0347	0.0352	**0.0343**
Hig	Hig	Sml	0.0258	0.0257	0.0235	0.0239	**0.0232**	0.0232
		Med	0.0758	0.0747	0.0607	0.0601	0.0604	**0.0600**
		Lrg	0.4127	0.4162	0.2464	**0.2458**	0.2470	0.2462
	Med	Sml	0.0217	0.0220	0.0183	0.0184	**0.0180**	0.0182
		Med	0.0794	0.0796	0.0721	0.0721	**0.0718**	0.0719
		Lrg	0.3410	0.3416	0.2777	**0.2772**	0.2786	0.2782
	Low	Sml	0.0145	0.0149	0.0094	0.0095	0.0093	**0.0093**
		Med	0.0725	0.0721	0.0624	0.0621	**0.0620**	0.0622
		Lrg	0.2858	0.2849	0.2713	**0.2713**	0.2720	0.2714

Table A20. Mean squared error of $\hat{\beta}^{B}_{make}$ when *Region* is missing at random. Bold values represent the best result.

			Methods					
Relb.	**Miss.**	**Size**	**Mean.I**	**IPW**	**MI**	**K.MI**	**DB.MI**	**HC.MI**
Low	Hig	Sml	0.0155	0.0156	0.0126	**0.0125**	0.0127	0.0127
		Med	0.0332	0.0332	0.0263	0.0263	0.0262	**0.0262**
		Lrg	0.0610	0.0609	0.0485	0.0486	**0.0484**	0.0484
	Med	Sml	0.0146	0.0146	0.0129	**0.0128**	0.0129	0.0129
		Med	0.0251	0.0252	0.0230	0.0229	0.0229	**0.0229**
		Lrg	0.0561	0.0561	0.0504	0.0504	0.0502	**0.0502**
	Low	Sml	0.0133	0.0133	0.0124	**0.0124**	0.0125	0.0125
		Med	0.0276	0.0275	0.0266	**0.0266**	0.0267	0.0267
		Lrg	0.0495	0.0496	0.0463	0.0462	0.0462	**0.0462**
Med	Hig	Sml	0.0365	0.0364	0.0293	0.0293	**0.0291**	0.0291
		Med	0.0875	0.0879	0.0582	0.0583	**0.0578**	0.0580
		Lrg	0.1586	0.1587	0.1180	0.1186	0.1175	**0.1174**
	Med	Sml	0.0314	0.0313	0.0274	0.0274	**0.0273**	0.0274
		Med	0.0694	0.0690	0.0595	0.0596	0.0593	**0.0592**
		Lrg	0.1434	0.1443	0.1276	0.1278	0.1274	**0.1273**
	Low	Sml	0.0290	0.0291	0.0276	0.0276	**0.0275**	0.0275
		Med	0.0667	0.0667	0.0633	0.0634	0.0633	**0.0632**
		Lrg	0.1404	0.1402	0.1268	0.1270	**0.1264**	0.1265
Hig	Hig	Sml	0.1004	0.1001	0.0749	0.0752	**0.0748**	0.0749
		Med	0.3542	0.3532	**0.2340**	0.2348	0.2349	0.2354
		Lrg	3.0461	3.0672	1.4855	**1.4851**	1.4978	1.4996
	Med	Sml	0.0814	0.0815	0.0698	0.0698	0.0697	**0.0697**
		Med	0.1978	0.1973	0.1700	0.1700	0.1695	**0.1694**
		Lrg	2.6731	2.6765	**2.0533**	2.0567	2.0653	2.0650
	Low	Sml	0.0806	0.0809	0.0732	0.0732	0.0731	**0.0730**
		Med	0.1812	0.1813	0.1681	0.1681	**0.1678**	0.1679
		Lrg	1.7210	1.7156	**1.6306**	1.6311	1.6350	1.6328

Table A21. Mean squared error of $\hat{\beta}^{C}_{make}$ when *Region* is missing at random. Bold values represent the best result.

			Methods					
Relb.	**Miss.**	**Size**	**Mean.I**	**IPW**	**MI**	**K.MI**	**DB.MI**	**HC.MI**
Low	Hig	Sml	0.0160	0.0161	0.0130	0.0130	**0.0129**	0.0129
		Med	0.0329	0.0329	0.0251	0.0252	0.0249	**0.0248**
		Lrg	0.0646	0.0643	0.0497	0.0498	**0.0491**	0.0492
	Med	Sml	0.0147	0.0147	0.0126	0.0126	**0.0126**	0.0126
		Med	0.0264	0.0265	0.0236	0.0235	0.0233	**0.0233**
		Lrg	0.0581	0.0579	0.0521	0.0522	0.0518	**0.0517**
	Low	Sml	0.0140	0.0140	0.0131	0.0131	**0.0131**	0.0131
		Med	0.0275	0.0274	0.0262	**0.0262**	0.0262	0.0262
		Lrg	0.0538	0.0536	0.0495	0.0495	0.0494	**0.0494**
Med	Hig	Sml	0.0395	0.0394	0.0315	0.0316	**0.0313**	0.0313
		Med	0.0899	0.0904	0.0615	0.0615	**0.0611**	0.0612
		Lrg	0.1685	0.1684	0.1285	0.1289	0.1283	**0.1282**
	Med	Sml	0.0333	0.0332	0.0291	0.0291	**0.0291**	0.0291
		Med	0.0732	0.0729	0.0636	0.0637	**0.0633**	0.0633
		Lrg	0.1559	0.1564	0.1384	0.1384	0.1381	**0.1381**
	Low	Sml	0.0304	0.0305	**0.0289**	0.0290	0.0289	0.0290
		Med	0.0675	0.0677	**0.0644**	0.0645	0.0645	0.0644
		Lrg	0.1450	0.1450	0.1303	0.1304	**0.1300**	0.1300
Hig	Hig	Sml	0.1097	0.1094	0.0807	0.0811	**0.0807**	0.0808
		Med	0.3447	0.3435	**0.2295**	0.2301	0.2304	0.2309
		Lrg	3.0495	3.0697	1.4834	**1.4833**	1.4959	1.4974
	Med	Sml	0.0893	0.0894	0.0772	0.0772	**0.0772**	0.0772
		Med	0.2021	0.2015	0.1758	0.1758	0.1754	**0.1750**
		Lrg	2.6473	2.6494	**2.0516**	2.0547	2.0649	2.0637
	Low	Sml	0.0840	0.0841	0.0765	0.0765	0.0764	**0.0764**
		Med	0.1956	0.1957	0.1813	0.1812	**0.1808**	0.1809
		Lrg	1.7120	1.7068	**1.6237**	1.6247	1.6289	1.6262

Table A22. Bias of $\hat{\beta}^{B}_{make}$ when *Region* is missing not at random. Bold values represent the best result.

			Methods					
Relb.	**Miss.**	**Size**	**Mean.I**	**IPW**	**MI**	**K.MI**	**DB.MI**	**HC.MI**
Low	Hig	Sml	−**0.0218**	−0.0256	−0.0261	−0.0252	−0.0318	−0.0318
		Med	−0.0273	−0.0313	−0.0269	−**0.0263**	−0.0324	−0.0333
		Lrg	−**0.0204**	−0.0250	−0.0273	−0.0277	−0.0330	−0.0340
	Med	Sml	−**0.0199**	−0.0231	−0.0252	−0.0247	−0.0287	−0.0288
		Med	−**0.0236**	−0.0265	−0.0280	−0.0279	−0.0316	−0.0322
		Lrg	−**0.0027**	−0.0073	−0.0100	−0.0097	−0.0135	−0.0142
	Low	Sml	−**0.0206**	−0.0236	−0.0240	−0.0237	−0.0262	−0.0262
		Med	−**0.0186**	−0.0218	−0.0228	−0.0223	−0.0246	−0.0251
		Lrg	−**0.0165**	−0.0205	−0.0178	−0.0173	−0.0196	−0.0202
Med	Hig	Sml	−**0.0026**	−0.0040	−0.0073	−0.0066	−0.0092	−0.0095
		Med	0.0242	0.0216	0.0106	0.0113	0.0083	**0.0076**
		Lrg	0.0325	0.0296	0.0163	0.0165	0.0142	**0.0132**
	Med	Sml	−**0.0091**	−0.0110	−0.0114	−0.0109	−0.0126	−0.0128
		Med	0.0123	0.0109	0.0021	0.0017	0.0005	**0.0003**
		Lrg	0.0642	0.0610	0.0502	0.0499	0.0491	**0.0480**
	Low	Sml	−**0.0066**	−0.0089	−0.0081	−0.0079	−0.0089	−0.0090
		Med	0.0207	0.0192	0.0181	0.0181	**0.0169**	0.0171
		Lrg	0.0352	0.0321	0.0291	0.0289	0.0284	**0.0278**
Hig	Hig	Sml	0.0317	0.0307	0.0213	0.0215	0.0202	**0.0201**
		Med	0.1006	0.0985	0.0757	0.0753	**0.0741**	0.0741
		Lrg	0.4817	0.4825	0.2503	0.2499	0.2512	**0.2497**
	Med	Sml	0.0229	0.0223	0.0152	0.0152	**0.0144**	0.0144
		Med	0.0973	0.0967	0.0734	0.0732	**0.0721**	0.0724
		Lrg	0.4476	0.4489	**0.2902**	0.2905	0.2913	0.2908
	Low	Sml	0.0156	0.0150	0.0129	0.0133	0.0128	**0.0125**
		Med	0.0720	0.0707	0.0619	0.0617	0.0616	**0.0614**
		Lrg	0.3140	0.3174	0.2675	0.2674	0.2680	**0.2671**

Table A23. Bias of $\hat{\beta}^{C}_{make}$ when *Region* is missing not at random. Bold values represent the best result.

			Methods					
Relb.	**Miss.**	**Size**	**Mean.I**	**IPW**	**MI**	**K.MI**	**DB.MI**	**HC.MI**
Low	Hig	Sml	−0.0095	−0.0112	−**0.0093**	−0.0094	−0.0116	−0.0116
		Med	−0.0130	−0.0147	−**0.0081**	−0.0083	−0.0101	−0.0110
		Lrg	−**0.0013**	−0.0040	−0.0051	−0.0067	−0.0074	−0.0082
	Med	Sml	−**0.0049**	−0.0063	−0.0079	−0.0079	−0.0093	−0.0094
		Med	−**0.0134**	−0.0144	−0.0163	−0.0168	−0.0177	−0.0183
		Lrg	0.0062	0.0042	0.0005	**0.0003**	−0.0010	−0.0017
	Low	Sml	−**0.0039**	−0.0051	−0.0067	−0.0067	−0.0077	−0.0076
		Med	−**0.0048**	−0.0059	−0.0070	−0.0069	−0.0075	−0.0080
		Lrg	−0.0025	−0.0045	−0.0018	−**0.0014**	−0.0023	−0.0030
Med	Hig	Sml	0.0003	0.0001	−0.0002	−**0.0001**	−0.0007	−0.0011
		Med	0.0295	0.0277	0.0174	0.0175	0.0166	**0.0159**
		Lrg	0.0283	0.0266	0.0197	0.0189	0.0191	**0.0183**
	Med	Sml	−**0.0046**	−0.0056	−0.0063	−0.0063	−0.0068	−0.0070
		Med	0.0249	0.0244	0.0182	0.0173	0.0175	**0.0173**
		Lrg	0.0622	0.0608	0.0504	0.0501	0.0503	**0.0493**
	Low	Sml	−**0.0001**	−0.0012	−0.0004	−0.0003	−0.0005	−0.0007
		Med	0.0244	0.0243	0.0226	0.0224	**0.0219**	0.0220
		Lrg	0.0421	0.0404	0.0339	0.0335	0.0341	**0.0335**
Hig	Hig	Sml	0.0339	0.0333	0.0236	0.0236	**0.0229**	0.0230
		Med	0.0794	0.0782	0.0604	**0.0598**	0.0599	0.0599
		Lrg	0.4818	0.4841	0.2457	**0.2446**	0.2466	0.2451
	Med	Sml	0.0272	0.0271	0.0184	0.0183	**0.0178**	0.0179
		Med	0.0932	0.0932	0.0736	0.0732	**0.0724**	0.0729
		Lrg	0.4359	0.4391	**0.2781**	0.2784	0.2794	0.2787
	Low	Sml	0.0109	0.0109	0.0093	0.0094	0.0092	**0.0091**
		Med	0.0713	0.0709	0.0623	0.0621	0.0623	**0.0620**
		Lrg	0.3170	0.3209	0.2714	**0.2711**	0.2717	0.2711

Table A24. Mean squared error of $\hat{\beta}^{B}_{make}$ when *Region* is missing not at random. Bold values represent the best result.

					Methods			
Relb.	**Miss.**	**Size**	**Mean.I**	**IPW**	**MI**	**K.MI**	**DB.MI**	**HC.MI**
Low	Hig	Sml	0.0158	0.0160	0.0128	**0.0127**	0.0130	0.0129
		Med	0.0331	0.0333	0.0264	**0.0263**	0.0263	0.0264
		Lrg	0.0673	0.0672	0.0487	0.0493	**0.0485**	0.0486
	Med	Sml	0.0153	0.0153	**0.0128**	0.0128	0.0129	0.0129
		Med	0.0278	0.0278	**0.0228**	0.0230	0.0229	0.0229
		Lrg	0.0606	0.0600	0.0505	0.0506	**0.0501**	0.0503
	Low	Sml	0.0136	0.0135	**0.0124**	0.0125	0.0125	0.0125
		Med	0.0296	0.0293	0.0267	0.0267	0.0267	**0.0267**
		Lrg	0.0530	0.0529	0.0467	0.0466	**0.0465**	0.0465
Med	Hig	Sml	0.0390	0.0389	0.0292	0.0291	0.0290	**0.0289**
		Med	0.0787	0.0785	0.0584	0.0586	0.0580	**0.0580**
		Lrg	0.1697	0.1690	0.1177	0.1181	**0.1171**	0.1171
	Med	Sml	0.0333	0.0331	0.0275	0.0275	**0.0274**	0.0274
		Med	0.0735	0.0730	0.0596	0.0596	0.0593	**0.0593**
		Lrg	0.1707	0.1705	0.1279	0.1279	**0.1273**	0.1275
	Low	Sml	0.0305	0.0305	0.0274	0.0275	0.0274	**0.0274**
		Med	0.0692	0.0692	0.0637	0.0636	0.0634	**0.0634**
		Lrg	0.1439	0.1425	0.1271	0.1270	**0.1267**	0.1268
Hig	Hig	Sml	0.1055	0.1049	0.0750	0.0749	0.0748	**0.0748**
		Med	0.3113	0.3117	**0.2345**	0.2356	0.2350	0.2354
		Lrg	3.5278	3.5611	**1.4778**	1.4811	1.5035	1.4995
	Med	Sml	0.0897	0.0899	0.0698	0.0700	0.0698	**0.0698**
		Med	0.3240	0.3269	0.1697	0.1697	**0.1693**	0.1695
		Lrg	3.3592	3.3825	**2.0527**	2.0552	2.0690	2.0670
	Low	Sml	0.0836	0.0834	**0.0732**	0.0733	0.0732	0.0732
		Med	0.1892	0.1888	0.1678	0.1679	0.1676	**0.1675**
		Lrg	2.0090	2.0216	**1.6296**	1.6348	1.6385	1.6370

Table A25. Mean squared error of $\hat{\beta}^{C}_{make}$ when *Region* is missing not at random. Bold values represent the best result.

					Methods			
Relb.	**Miss.**	**Size**	**Mean.I**	**IPW**	**MI**	**K.MI**	**DB.MI**	**HC.MI**
Low	Hig	Sml	0.0159	0.0159	0.0131	0.0131	0.0129	**0.0129**
		Med	0.0326	0.0326	0.0255	0.0254	**0.0251**	0.0251
		Lrg	0.0686	0.0682	0.0497	0.0499	**0.0491**	0.0493
	Med	Sml	0.0152	0.0151	0.0127	0.0127	**0.0125**	0.0125
		Med	0.0292	0.0290	0.0237	0.0238	0.0236	**0.0235**
		Lrg	0.0609	0.0603	0.0519	0.0520	**0.0514**	0.0515
	Low	Sml	0.0143	0.0141	0.0131	0.0131	0.0131	**0.0131**
		Med	0.0290	0.0287	0.0263	0.0262	0.0262	**0.0261**
		Lrg	0.0559	0.0558	0.0499	0.0500	**0.0496**	0.0497
Med	Hig	Sml	0.0409	0.0407	0.0314	0.0314	0.0312	**0.0311**
		Med	0.0845	0.0844	0.0612	0.0613	0.0608	**0.0608**
		Lrg	0.1861	0.1860	0.1295	0.1297	0.1286	**0.1286**
	Med	Sml	0.0349	0.0347	0.0291	0.0291	**0.0290**	0.0290
		Med	0.0777	0.0773	0.0637	0.0636	0.0632	**0.0632**
		Lrg	0.1769	0.1757	0.1384	0.1385	**0.1377**	0.1378
	Low	Sml	0.0327	0.0329	0.0289	0.0289	0.0288	**0.0288**
		Med	0.0698	0.0698	0.0648	0.0649	0.0646	**0.0646**
		Lrg	0.1507	0.1490	0.1306	0.1304	**0.1301**	0.1302
Hig	Hig	Sml	0.1125	0.1119	0.0809	0.0809	**0.0807**	0.0808
		Med	0.3075	0.3080	**0.2298**	0.2309	0.2304	0.2307
		Lrg	3.4957	3.5334	**1.4725**	1.4762	1.4996	1.4955
	Med	Sml	0.0984	0.0984	0.0773	0.0775	0.0772	**0.0772**
		Med	0.3257	0.3288	0.1755	0.1755	**0.1753**	0.1755
		Lrg	3.3488	3.3775	**2.0493**	2.0521	2.0666	2.0649
	Low	Sml	0.0865	0.0862	**0.0765**	0.0767	0.0766	0.0766
		Med	0.2045	0.2045	0.1808	0.1810	**0.1805**	0.1805
		Lrg	2.0213	2.0350	**1.6247**	1.6292	1.6328	1.6328

References

1. Tabuchi, H. Air Bag Flaw, Long Known to Honda and Takata, Led to Recalls. *The New York Times*, 11 September 2014.
2. National Center for Statistics and Analysis. *Crash Report Sampling System Analytical User's Manual, 2016–2021*; (Report No. DOT HS 813 436); National Highway Traffic Safety Administration: Washington, DC, USA, 2023.
3. Rubin, D.B. Inference and Missing Data. *Biometrika* **1976**, *63*, 581–592. [CrossRef]
4. Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*; John Wiley & Sons: New York, NY, USA, 1987.
5. Little, R.; Rubin, D. *Statistical Analysis with Missing Data*; John Wiley & Sons: Hoboken, NJ, USA, 2019.
6. Ferro, S.; Bottigliengo, D.; Gregori, D.; Fabricio, A.S.C.; Gion, M.; Baldi, I. Phenomapping of Patients with Primary Breast Cancer Using Machine Learning-Based Unsupervised Cluster Analysis. *J. Pers. Med.* **2021**, *11*, 272. [CrossRef] [PubMed]
7. Nouraei, H.; Nouraei, H.; Rabkin, S.W. Comparison of Unsupervised Machine Learning Approaches for Cluster Analysis to Define Subgroups of Heart Failure with Preserved Ejection Fraction with Different Outcomes. *Bioengineering* **2022**, *9*, 175. [CrossRef] [PubMed]
8. Zuo, Y.; Lundberg, J.; Chandran, P.; Rantatalo, M. Squat Detection and Estimation for Railway Switches and Crossings Utilising Unsupervised Machine Learning. *Appl. Sci.* **2023**, *13*, 5376. [CrossRef]
9. Groenwold, R.H.; White, I.R.; Donders, A.R.; Carpenter, J.R.; Altman, D.G.; Moons, K.G. Missing covariate data in clinical research: When and when not to use the missing-indicator method for analysis. *Can. Med. Assoc. J.* **2012**, *184*, 1265–1269. [CrossRef]
10. Pedersen, A.B.; Mikkelsen, E.M.; Cronin-Fenton, D.; Kristensen, N.R.; Pham, T.M.; Pedersen, L.; Petersen, I. Missing data and multiple imputation in clinical epidemiological research. *Clin. Epidemiol.* **2017**, *15*, 157–166. [CrossRef]
11. Yang, S.; Berdine, G. Missing values in data analysis. *Southwest Respir. Crit. Care Chronicles* **2022**, *10*, 57–60. [CrossRef]
12. Sterne, J.A.; White, I.R.; Carlin, J.B.; Spratt, M.; Royston, P.; Kenward, M.G.; Wood, A.M.; Carpenter, J.R. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Br. Med. J.* **2009**, *29*, b2393. [CrossRef]
13. Seaman, S.R.; White, I.R. Review of inverse probability weighting for dealing with missing data. *Stat. Methods Med. Res.* **2013**, *22*, 278–295. [CrossRef]
14. Jerez, J.M.; Molina, I.; García-Laencina, P.J.; Alba, E.; Ribelles, N.; Martín, M.; Franco, L. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artif. Intell. Med.* **2010**, *50*, 105–115. [CrossRef]
15. García-Laencina, P.J.; Sancho-Gómez, J.L.; Figueiras-Vidal, A.R. Pattern classification with missing data: A review. *Neural Comput. Appl.* **2010**, *19*, 263–282. [CrossRef]
16. Waljee, A.K.; Mukherjee, A.; Singal, A.G.; Zhang, Y.; Warren, J.; Balis, U.; Marrero, J.; Zhu, J.; Higgins, P.D.R. Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open* **2013**, *3*, 1–7. [CrossRef] [PubMed]
17. Barakat, M.S.; Field, M.; Ghose, A.; Stirling, D.; Holloway, L.; Vinod, S.; Dekker, A.; Thwaites, D. The effect of imputing missing clinical attribute values on training lung cancer survival prediction model performance. *Health Inf. Sci. Syst.* **2017**, *5*, 16. [CrossRef] [PubMed]
18. Gmel, G. Imputation of missing values in the case of a multiple item instrument measuring alcohol consumption. *Stat. Med.* **2001**, *20*, 2369–2381. [CrossRef] [PubMed]
19. Balakrishnan, N.; So, H.Y.; Ling, M.H. EM algorithm for one-shot device testing with competing risks under exponential distribution. *Reliab. Eng. Syst. Saf.* **2015**, *137*, 129–140. [CrossRef]
20. Balakrishnan, N.; So, H.Y.; Ling, M.H. EM Algorithm for One-Shot Device Testing with Competing Risks under Weibull Distribution. *IEEE Trans. Reliab.* **2016**, *65*, 973–991. [CrossRef]
21. Azur, M.J.; Stuat, E.A.; Frangakis, C.; Leaf, P.J. Multiple imputation by chained equations: What is it and how does it work? *Int. J. Methods Psychiatr. Res.* **2011**, *20*, 40–49. [CrossRef]
22. Liu, Y.; De, A. Multiple Imputation by Fully Conditional Specification for Dealing with Missing Data in a Large Epidemiologic Study. *Int. J. Stat. Med. Res.* **2015**, *4*, 287–295. [CrossRef]
23. Murray, J.S. Multiple Imputation: A Review of Practical and Theoretical Findings. *Stat. Sci.* **2018**, *33*, 142–159. [CrossRef]
24. Lee, K.; Carlin, J.B. Multiple imputation in the presence of non-normal data. *Stat. Med.* **2016**, *36*, 606–617. [CrossRef]
25. Barnard, J.; Rubin, D.B. Small-Sample Degrees of Freedom with Multiple Imputation. *Biometrika* **1999**, *86*, 948–955. [CrossRef]
26. Heymans, M.W.; Eekhout, I. *Applied Missing Data Analysis with SPSS and (R)Studio*. First Draft. 2019. Available online: https://bookdown.org/mwheymans/bookmi/missing-data-in-questionnaires.html (accessed on 22 February 2024).
27. Ling, M.H.; Balakrishnan, N.; Yu, C.; So, H.Y. Inference for One-Shot Devices with Dependent k-Out-of-M Structured Components under Gamma Frailty. *Mathematics* **2021**, *9*, 3032. [CrossRef]
28. Ling, M.H.; Balakrishnan, N.; Bae, S.J. On the application of inverted Dirichlet distribution for reliability inference of completely censored components with dependent structure. *Comput. Ind. Eng.* **2024**, *196*, 110452. [CrossRef]
29. Hand, D.J.; Bolton, R.J. Pattern discovery and detection: A unified statistical methodology. *J. Appl. Stat.* **2004**, *31*, 885–924. [CrossRef]
30. Aschenbruck, R.; Szepannek, G.; Wilhelm, A.F.X. Imputation Strategies for Clustering Mixed Type Data with Missing Values. *J. Classif.* **2023**, *40*, 2–24. [CrossRef]
31. Agresti, A. *An Introduction to Categorical Data Analysis*, 3rd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2019.
32. Ward, J.H. Herarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244. [CrossRef]
33. Johnson, S.C. Hierarchical Clustering schemes. *Psychometrika* **1967**, *32*, 241–254. [CrossRef]

34. Lance, G.N.; Williams, W.T. A general theory of classificatory sorting strategies: 1. Hierarchical systems. *Comput. J.* **1967**, *9*, 373–380. [CrossRef]
35. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, 2nd ed.; Springer: New York, NY, USA, 2009.
36. Huang, Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Min. Knowl. Discov.* **1998**, *2*, 283–304. [CrossRef]
37. Huang, Z. Clustering large data sets with mixed numeric and categorical values. In Proceedings of the First Pacific Asia Knowledge Discovery and Data Mining Conference, Singapore, 23–24 February 1997; World Scientific: Singapore, 1997; pp. 21–34.
38. Ji, J.; Pang, W.; Zhou, C.; Han, X.; Wang, Z. A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data. *Knowl.-Based Syst.* **2012**, *30*, 129–135. [CrossRef]
39. Ji, J.; Bai, T.; Zhou, C.; Ma, C.; Wang, Z. An improved k-prototypes clustering algorithm for mixed numeric and categorical data. *Neurocomputing* **2010**, *120*, 590–596. [CrossRef]
40. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, OR, USA, 2–4 August 1996.
41. Shukla, N.; Hagenbuchner, M.; Win, K.T.; Yang, J. Breast cancer data analysis for survivability studies and prediction. *Comput. Methods Programs Biomed.* **2018**, *155*, 199–208. [CrossRef] [PubMed]
42. Ankerst, M.; Breunig, M.M.; Kriegel, H.-P.; Sander, J. OPTICS: ordering points to identify the clustering structure. In Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data (SIGMOD '99), Philadelphia, PA, USA, 1–3 June 1999; Association for Computing Machinery: New York, NY, USA, 1999; pp. 49–60.
43. Schubert, E.; Sander, J.; Ester, M.; Kriegel, H.-P. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. *CM Trans. Database Syst.* **2017**, *42*, 19. [CrossRef]
44. Campello, R.J.G.B.; Moulavi, D.; Zimek, A.; Sander, J. Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. *ACM Trans. Knowl. Discov. Data* **2017**, *10*, 1–51. [CrossRef]
45. McInnes, L.; Healy, J.; Astels, S. hdbscan: Hierarchical density-based clustering. *J. Open Source Softw.* **2017**, *2*, 205. [CrossRef]
46. Gower, J.C. A General Coefficient of Similarity and Some of Its Properties. *Biometrics* **1971**, *27*, 857–871. [CrossRef]
47. Gower, J.C. A note on Burnaby's character-weighted similarity coefficient. *J. Int. Assoc. Math. Geol.* **1970**, *2*, 39–45. [CrossRef]
48. van Buuren, S.; Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* **2011**, *45*, 1–67. [CrossRef]
49. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2023.
50. Maechler, M.; Rousseeuw, P.; Struyf, A.; Hubert, M.; Hornik, K. *Cluster: Cluster Analysis Basics and Extensions*. R package version 2.1.4. 2022.
51. Szepannek, G. clustMixType: User-friendly clustering of mixed-type data in R. *R J.* **2018**, *10*, 200–208. [CrossRef]
52. Rubin, D.B. Determination of optimal Epsilon (Eps) value on DBSCAN algorithm to clustering data on peatland hotspots in Sumatra. *IOP Conf. Ser. Earth Environ. Sci.* **1976**, *31*, 012012.
53. Hennig, C. *fpc: Flexible Procedures for Clustering.* R package version 2.2-12. 2024.
54. Zhang, F.; Subramanian, R.; Chen, C.-L.; Noh, E.Y. *Crash Report Sampling System: Design Overview, Analytic Guidance, and FAQs (Report No. DOT HS 812 688)*; National Highway Traffic Safety Administration: Washington, DC, USA, 2019.
55. Uncu, N.; Koyuncu, M. Enhancing Control: Unveiling the Performance of Poisson EWMA Charts through Simulation with Poisson Mixture Data. *Appl. Sci.* **2023**, *13*, 11160. [CrossRef]
56. Vivancos, J.-L.; Buswell, R.A.; Cosar-Jorda, P.; Aparicio-Fernández, C. The application of quality control charts for identifying changes in time-series home energy data. *Energy Build.* **2020**, *215*, 109841. [CrossRef]
57. Yeganeh, A.; Shadman, A. Using evolutionary artificial neural networks in monitoring binary and polytomous logistic profiles. *J. Manuf. Syst.* **2021**, *61*, 546–561. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.