BIG DATA ANALYTICS - MISSING OR MESSY DATA, WHAT NOW?

D. A. HEGER, PhD
CEO/Owner DHTechnologies
& Data Nubes

DHTechnologies, Austin, TX

Introduction. Missing data scenarios are common Big Data problems in domains such as biology, finance, medicine, life-science, research, or climatic science (to name a few). They can arise from different sources such as mishandling of samples, low signal-to-noise ratios, measurement errors, hardware failures, transmission errors, no-response, or deleted aberrant values. Rubin introduced the notion of the distribution of missingness as a way to classify the conditions under which missing data should be treated. Little and Rubin distinguish among data missing completely at random (MCAR), data missing at random (MAR), and data missing not at random (MNAR):

1. Data missing completely at random (MCAR) describes scenarios where the probability of an instance (case) having a missing value for a variable does not depend on either the known values or the missing data, respectively. An example would be a sensor outage that results into missing some measurements.

2. Data missing at random (MAR) describes scenarios where the probability of an instance having a missing value for a variable may depend on the known values but not on the value of the missing data itself. For example, suppose men are less likely (compared to women) to respond to an income question in a survey, but the likelihood of responding is independent of their actual income. In this case, unbiased gender-specific income estimates can be made if one has data on the gender variable (as an example by replacing the missing income value with the gender-specific median income).

3. Data missing not at random (MNAR) describes scenarios where the probability of an instance having a missing value for a variable may depend on the value of that variable. For example, this may occur if either low or high income subjects (or both) are less likely to answer the income question in a survey. This is the most complex type of missing data and in many cases, there is no good value to substitute for the missing one. But, by just dropping these subjects, the results will be biased and hence such an approach is normally not suggested.

It has to be pointed out that missing data introduces an element of ambiguity into the data analysis cycle. Missing data can affect properties of statistical estimators such as the mean, variance, or percentage, resulting in a loss of information power and misleading conclusions. A variety of techniques have been proposed for substituting missing values with statistical predictions, a process that is generally referred to as missing data imputation. To reiterate, it is very often the case that the weakest link in a Big Data analytics study is the quality of the available data sets. It is a fact that the more one can find out about why the data sets are as they are, the more one can develop a case on the pattern of the missing data, as well as on a rationale on why the pattern may or may not matter. It has to be pointed out though that in many studies, just eliminating missing cases prior to the analysis is not viewed as a legitimate solution to the problem.

Some of the rather standard techniques to address missing data (such as list-wise deletion, single

mean imputation, or single regression imputation) may lead to biased estimates of the model parameters. To illustrate, a simple mean substitution method leaves the mean unchanged but decreases the variance! Some of the more appropriate techniques (in many cases) for dealing with missing data are the multiple imputation (MI) and the (full information) maximum likelihood (ML) estimation that incorporate the missing data points in the analysis (see Enders and Enders & Peugh). MI involves imputing a range of random plausible values for missing data, a process that results in several complete data sets that can then be analyzed. One of the advantages of this approach is that the technique introduces variability into the distribution of cases with missing data (simulating the messy world we are living in), a process that is more likely to represent the population than imputing a single value for each missing case. Partial data actually contributes to the estimation of the model's parameters by implying probable values for missing scores via the correlations among the variables. Expectation maximization (EM), a common method for obtaining ML estimates with incomplete data sets, treats the model parameters (rather than the data points themselves) as missing values to be estimated and borrows information from the existing data at successive iterations until differences between successive iterations are minor. Missing data can pose a number of additional problems in multilevel data structures, depending on the sampling design underlying the data set, the extent to which the data are missing at each level, and whether or not the data can be assumed to be missing at random. In some modeling situations, there may be a considerable amount of missing data. Compared with single-level analyses, the difficulties presented by multilevel analyses scenarios concerns the likelihood that the missing data at one level (Level 2) may be linked to the missing data at Level 1.

In any Big Data analytics project, it is paramount to distinguish between data preparation and data analysis. While preparing the data for analysis, it is imperative to determine the amount of missing data, as well as the missing data pattern(s). It is essential to note that the reality is that there is a very small chance to get the missing (real) data back in the first place. Hence, a data scientist always has to deal with the problem of missing information. The quality of the analysis though depends on assumptions one makes on the pattern of the missing data and what is reasonable to conclude about those patterns. Now, what one can do about the missing data becomes the pressing concern!

Definition - Missing Completely at Random (MCAR). Suppose the probability of an observation being missing does not depend on observed or unobserved measurements. In mathematical terms, one can stipulate:

$$\Pr(\mathrm{r} \mid \mathrm{y_o}, \mathrm{y_m}) = \Pr(\mathrm{r}) \tag{1}$$

Then one can state that the observation is missing completely at random (MCAR). Note that in a sample survey, MCAR may be labeled as uniform non-response. If data sets are MCAR, then consistent results with missing data can be obtained by performing the analyses as if there were no missing data. But, doing so will result in some loss of information. So that implies that with MCAR data sets, the analysis only provides valid inferences with complete data sets! So does an analysis based on moment based estimators (for example generalized estimating equations) and other estimators derived from consistent estimating equations. The term consistent estimating equations refers to functions of the data and unknown parameters whose expectation (taken over the complete data at the population parameter values) is 0. Under MCAR, they still have expectation zero and so still lead to valid inferences.

Stating the same mathematically, an estimating equation can be written as $U\ (y, \theta)$ and at the estimate $\hat{\theta}, U\ (y, \hat{\theta}) = 0$. The estimating equation is consistent because $EU\ (Y, \theta) = 0$ (where θ reflects the population parameter value). It remains consistent if the data reflects missing completely at random (MCAR) because $EU\ (Y_o, \theta) = 0$. A simple example of a consistent estimating equation is the sample mean $U\ (y, \theta) = \bar{y} - \theta$. With MCAR a single imputation or a multiple imputation (MI) method can be considered. It has to be pointed out that with a single imputation method, it may be

difficult to generate valid variance estimates though! The author of this report always suggest to at least consider an MI approach with MCAR.

Definition - Missing At Random (MAR). After considering MCAR, a second scenario may come up. That is, what are the most general conditions under which a valid analysis can be done using only the observed data and no information about the missing value mechanism, Pr(r | yo, ym)? The answer here is when, given the observed data, the missingness mechanism does not depend on the unobserved data. Mathematically stated:

$$Pr(r \mid y_o, y_m) = Pr(r \mid y_o). \text{ (MAR)} \quad (2)$$

This is equivalent to stating that the behavior of 2 runs who share observed values have the same statistical behavior on the other observations, whether observed or not. To illustrate:

Table 1. Some Measurements

Measurements	*Features*					
Collection Run	*1*	*2*	*3*	*4*	*5*	*6*
1	**1**	**3**	**4.3**	**3.5**	**1**	**4.6**
2	**1**	**3**	**NA**	**3.5**	**NA**	**NA**

As runs 1 and 2 in Table 1 have the same values (where values for both runs are available), given these observed values (under MAR), features 3, 5 and 6 from run 2 have the same distribution (not the same value!) as features 3, 5 and 6 from run 1. It has to be pointed out that under MAR, the probability of a value being missing will generally depend on observed values, so it does not correspond to the intuitive notion of random. The important idea is that the missing value mechanism can be expressed solely in terms of observations that are observed. Unfortunately, this scenario can hardly ever be definitively determined from the data at hand! An example of a MAR mechanism may be the scenario where 2 measurements of the same variable are made concurrently. If they differ by more than a given amount, a 3d measurement is taken. This 3d measurement is missing for those scenarios that do not differ by the given amount specified for the 1st 2 measurements.

A special case of MAR is known as uniform non-response within classes. To illustrate, one seeks to collect data on income and property tax bands. Typically, those with higher incomes may be less willing to reveal the actual numbers. So a simple average of incomes from people who actually responded will be downwardly biased. However, assuming one has everyone's property tax band and given that the property tax band non-response to the income question is random, then the income data is missing at random. The reason, or mechanism, for it being missing depends on the property band. Given the property band, missingness does not depend on income itself. Therefore, to get an unbiased estimate of income, one has to first average the observed income within each property band. As data is missing at random given a property band, these estimates will be valid. To get an estimate of the overall income, one has to combine these estimates, weighted by the proportion in each property band.

It further has to be pointed out that likelihood methods (such as EM) are valid for MAR as well. However, in general, non-likelihood methods (based on simple completers, moments, estimating equations) are not valid for MAR per se. So ordinary means and other simple summary statistics that are based on the observed data will be biased. Finally, with likelihood, the term ignorable is often used to refer to a MAR mechanism. But it is the mechanism (the model for $Pr(R \mid y_o)$) that is ignorable, not the missing data! To summarize, MAR scenarios can be analyzed via multiple imputation (MI) methods or likelihood-based methods (such as EM).

Definition - Missing Not At Random (MNAR). When neither MCAR nor MAR holds, one states that the data sets are *Missing Not At Random* (MNAR). Now, in a likelihood setting, the *missingness*

mechanism is labeled *non-ignorable*. All this basically implies that while even accounting for all the available observed information, the reason for observations being missing *still depends on the unseen observations themselves*.

In other words, the probability of a data value being missing is related to the unobserved values. To illustrate, a study may focus on analyzing tumors. So that study requires measurements of the actual tumor sizes. The data may show that smaller tumors are less like to have sizes recorded (maybe due to any detection delays). So it is harder to actually measure the size of smaller tumors. Hence, while there may be good data available for larger tumor sizes, the sizing data for smaller tumors may be missing. So with MNAR, the important question to be answered is how is the data related to the unobserved value?

Similar to MAR, MNAR scenarios can be analyzed via Multiple Imputation (MI) methods or likelihood-based methods (such as EM). It has to be pointed out that MNAR is much more complex to deal with and basically requires modeling the process yielding the missing values.

Definition - Multiple Imputation/ Multiple imputation refers to a statistical technique for analyzing incomplete data sets (data sets for which some entries are missing). The actual application of the technique requires 3 steps, imputation, analysis, and pooling.

1. Imputation (impute = fill in). Impute the missing data m times to produce m complete data sets. Regression models, Bayesian ideas (with MCMC), or Fully Conditional Specifications (that works well for categorical data) can be used here.

2. Analysis. Analyze each data set by using a standard statistical procedure. This step results in m analyses.

3. Pooling. Integrate the m analysis results into a final result See Rubin or Schafer.

Note: Rubin has shown that if the method to create imputations is proper, then the resulting inferences will be statistically valid. The real challenge here is in the imputation phase (aka the construction of the m completed data sets). This step accounts for the process that created the missing data. A typical problem here could be that the missing data is actually related to its value (aka wealthy people tend to skip income questions in surveys). It further has to be reemphasized that missing entries

can appear anywhere in the data and that the method used in the imputation step must foresee the intended complete-data analyses. But, the repeated analysis step on the imputed data is actually somewhat simpler than the same analysis without imputation, as there is no need to bother with the missing data per se. The pooling step consists of computing the mean over the (m) repeated analysis, its variance, and its confidence interval or p value. In general, these computation are reasonably simple. Some common data imputation techniques are (to just name a few):

- MI Mean
- K-nearest neighbors (KNN)
- fuzzy K-means (FKM)
- singular value decomposition (SVD)
- Bayesian principal component analysis (BPCA)
- Bayesian ideas (regression, MCMC)
- Multiple imputations by chained equations (MICE)
- Fully conditional specifications (FCS)

To further discuss MI, The MCMC and the FCS methods are elaborated on in more detail here. Both methods are very popular iterative methods for performing multiple imputations for missing data patterns. The Markov Chain Monte Carlo (MCMC) method is widely used for Bayesian inference (Schafer) and is considered as one of the most popular iterative algorithm for multiple imputation scenarios. The basic flow is that one commences with some reasonable starting values for the mean, variance, and the covariance among a given set of variables. Next, one divides the sample into subsamples, each having the same missing data pattern (the same set of variables present and missing). For each missing data pattern, one uses the starting values to construct linear regressions for imputing the missing data, using all the observed variables in that pattern as predictors. One then imputes the

missing values, making random draws from the simulated error distribution, which results into a single completed data set.

Using this data set with missing data imputed, one recalculates the mean, variance and covariance and then makes a random draw from the posterior distribution of these parameters. Finally, one uses these drawn parameter values to update the linear regression equations needed for imputation. This process is typically repeated many times. For example, one may run 200 iterations of the algorithm before selecting the first completed data set, and then allow for another 100 iterations between each successive data set. So producing 5 data sets (as an example) requires 600 iterations (each of which generates a data set). Why so many iterations? The first 200 (burn-in) iterations are designed to ensure that the algorithm has converged to the correct posterior distribution. Then, allowing 100 iterations between successive data sets gives the confidence that the imputed values in the different data sets are statistically independent. If all assumptions are satisfied, the MCMC method produces parameter estimates that are consistent, asymptotically normal, and almost fully efficient. Full efficiency would require an infinite number of data sets, but a relatively small number normally gets one very close.

An alternative algorithm is known as the fully conditional specification (FCS) or multiple imputation by chained equations (MICE) (Brand, Van Buuren, Oudshoorn). This method is attractive because of its ability to impute both quantitative and categorical variables appropriately. It allows one to specify a regression equation for imputing each variable with missing data (usually linear regression for quantitative variables and logistic regression (binary, ordinal, or unordered multinomial) for categorical variables. Under logistic imputation, imputed values for categorical variables will also be categorical. Imputation proceeds sequentially, usually starting from the variable with the least missing data and progressing to the variable with the most missing data. At each step, random draws are made from both the posterior distribution of the parameters and the posterior distribution of the missing values. Imputed values at one step are used as predictors in the imputation equations at subsequent steps (something that never happens in MCMC algorithms). Once all missing values have been imputed, several iterations of the process are repeated before selecting a completed data set. Although attractive, FCS has 2 major disadvantages (compared to the linear MCMC method). First, it is much slower, computationally. Second, FCS itself has no theoretical justification. By contrast, if all assumptions are met, MCMC is guaranteed to converge to the correct posterior distribution of the missing values. FCS carries no such guarantee, although simulation results by Van Buuren are very encouraging.

Summary - Multiple Imputation (MI). Just as the single imputation methods, multiple imputation fills in estimates for the missing data. But to capture the uncertainty in those estimates, MI estimates the values multiple (m) times. As MI utilizes an imputation method that has an error term built in, the multiple estimates should be similar, but not identical. The result basically is multiple data sets (m) with identical values for all of the non-missing values and slightly different values for the imputed values in each data set. The statistical analysis of interest (such as logistic regression) is performed separately on each data set and the results are then consolidated. Because of the variation in the imputed values, there should also be variation in the parameter estimates, leading to appropriate estimates of standard errors and p-values, respectively.

Definition - Maximum Likelihood (ML). Depending on the pattern and the amount of missing data, a potentially legit approach may be to analyze the full, incomplete data set via a maximum likelihood estimation. This method does not impute any data, but rather uses each cases available data to compute maximum likelihood estimates. The maximum likelihood estimate of a parameter equals to the value of the parameter that is most likely to have resulted in the observed data.

When data points are missing, one can factor the likelihood function. The likelihood is computed separately for those cases with complete data on some variables and those with complete data on all variables. These 2 likelihoods are then maximized together to find the estimates. Like multiple imputation, this method provides unbiased parameter estimates and standard errors. One advantage

of ML is that it does not require the careful selection of variables used to impute values on (necessary with MI).

Some Actual Guidelines

In the next few paragraphs, some basic guidelines are provided. It has to be pointed out though that if more than 5% to 10% of the data is missing, a scenario like that is considered a potential major source of a serious bias that has to be addressed accordingly.

– Assuming that the proportion of missing data is ≤ 0.05. If less than 5% of the data is missing, studies have shown that it matters little what imputation method is chosen or whether one adjusts the variance of the regression coefficient estimates for having imputed data in this case. So for continuous variables, imputation via the median value should be adequate. For categorical predictors, the most frequent category can be used. A complete case analysis may be an option here as well.

– Assuming that the proportion of missing data is between **0.05 and 0.15**. In this scenario, if a predictor is unrelated to all of the other predictors, imputations can be done the same way as above (impute a reasonable constant value). If the predictor is correlated with other predictors, develop a customized model (such as via a transcan function - a nonlinear additive transformation and imputation function) to foresee the predictor from all of the other predictors. Then impute missing data with predicted values. For categorical variables, classification trees are good methods for developing customized imputation models. For continuous variables, ordinary regression can be used if the variable in question does not require a non-monotonic transformation to be predicted from the other variables. For either the related or unrelated predictor case, variances may need to be adjusted for imputation. The author of this report suggests multiple imputation or maximum likelihood methods here while single imputation methods may be considered.

– Assuming that the proportion of missing data is **> 0.15**. Such a scenario requires the same considerations as in the previous case and adjusting variances for imputation is even more important. To estimate the strength of the effect of a predictor that is frequently missing, it may be necessary to refit the model on the subject of observations for which that predictor is not missing. In this case either multiple imputation or maximum likelihood methods are preferred for most models.

Summary. Analyzing and cleansing (missing) data is paramount in order to achieve actual, accurate conclusions. To illustrate, while using the same ANN to conduct a regression study, using 2 differently cleansed data-sets as the input, there is a high probability that the 2 different input sets will result into 2 different answers/conclusions. Further, in any data analysis project, greater than 5% to 10% of missing data points is considered a potential source of a very serious bias condition. It is always important to consider and scrutinize the model/environment that produces the missing data.

Many studies have shown that in most scenarios, either a multiple imputation or an ML method does provide excellent results (even for MCAR). Further, if the missing data reflects MNAR, there is a need to consider and scrutinize the model that actually gives rise to the missing data. Plus, if the missingness is strongly related to the value of the variable, the problem becomes rather complex (a fact that cannot just be ignored).

References

[1]. Allison, P. (2002). Missing data. Thousand Oaks, CA: Sage.

[2]. Bodner, T. E. (2008).What improves with missing data imputations? Structural Equation Modeling, 15, 651-675.

[3]. Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. Psychological Methods, 16(1), 1-16.

[4]. Enders, C.K., & Peugh, J. L. (2004). Using an EM covariance matrix to estimate structural equation models with missing data: Choosing an adjusted sample size to improve the accuracy of inferences. Structural Equation Modeling, 11, 1–19.

[5]. Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. Psychological Methods, 2, 64–78.

[6]. Hox, J. (2010). Multilevel applications: Techniques and applications (2nd Edition). New York:

Routledge.

[7]. Kish, L. (1989). Statistical design for research. New York: Wiley.

[8]. Larsen, R. (2011). Missing data imputation versus full information maximum likelihood with second-level dependencies. Structural Equation Modeling, 18(4), 649-662.

[9]. Little, R.J.A & Rubin, D. B. (2002). Statistical analysis with missing data. New York, NY: Wiley.

[10]. Rubin, D.B. (1976). Inference and missing data. Biometrika, 63, 581–592.

[11]. Peugh, J.L. & Enders, C.K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. Review of Educational Research, 74, 525-556.

[12]. Schafer, J. (2005, November). Missing data in longitudinal studies. A review. Paper presented at the Annual Meeting of the American Association of Pharmaceutical Scientists, Nashville, TN.

[13]. Chandola T, Brunner E, Marmot M. (2006). Chronic stress at work and the metabolic syndrome: prospective study. BMJ 332:521-5. PMID:16428252

[14]. Greenland S, Finkle WD. (1995). A critical look at methods for handling missing covariates in epidemiologic regression analysis. Am J Epidemiol 142(12):1255-64.

[15]. Fleiss JL, Levin B, Paik MC. (2003). Statistical Methods for Rates and Proportions, 3rd ed. Hoboken NJ, John Wiley & Sons.

[16]. Harrell Jr FE. (2001). Regression Modeling Strategies With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York, Springer-Verlag.

[17]. Huberman M, Langholz B. (1999). Application of the missing-indicator method in matched case-control studies with incomplete data. Am J Epidemiol 150(12):1340-5.

[18]. Li X, Song X, Gray RH. (2004). Comparison of the missing-indicator method and conditional logistic regression in 1:m matched case-control studies with missing exposure values. Am J Epidemiol 159(6):603-610.

[19]. Moons KG, Grobbee DE. (2002). Diagnostic studies as multivariable, prediction research. J Epidemiol Community Health 56(5):337-8.

[20]. Roth, P. (1994). Missing data: A conceptual review for applied psychologists. Personnel Psychology 47:537-560.

[21]. Royston P. (2004). Multiple imputation of missing values. The Stata Journal 4(3):227-241.

[22]. Royston P. (2005a). Multiple imputation of missing values: update. The Stata Journal 5(2):188-201.

[23]. Royston P. (2005b). Multiple imputation of missing values: update of ice. The Stata Journal 5(4):527-536.

[24]. Royston P. (2007). Multiple imputation of missing vlaues: further update of ice, with an emphasis on interval censoring. The Stata Journal 7(4):445-464.

[25]. Schnonlau M. (2006). Stata software package, hotdeckvar.pkg, for hotdeck imputation. http://www.schonlau.net/stata/.

[26]. Steyerberg EW. (2009). Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York, Springer.

[27]. Twisk JWR. (2003). Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide. Cambridge, Cambridge University Press.

[28]. van Buuren S, Boshuizen HC, Knook DL. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in Medicine 18: 681–694.

[29]. Vandenbrouchke JP, von Elm E, Altman DG, et al. (2007). Strengthening and reporting of observational studies in epidemiology (STROBE): explanation and elaboration. Ann Intern Med 147(8):W-163 to W-194.

[30]. Brand, J.P.L. (1999) Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets. Dissertation, Erasmus University,Rotterdam.

[31]. Cornwell, C. and Rupert, P., "Efficient Estimation with Panel Data: An Empirical Comparison of Instrumental Variable Estimators," Journal of Applied Econometrics, 3, 1988, pp. 149-155.

[32]. Dempster, A. P., Nan M. Laird and Donald B. Rubin. 1977. "Maximum Likelihood Estimation from Incomplete Data Via the EM Algorithm." Journal of the Royal Statistical Society, Series B 39: 1-38.

[33]. Heckman, James J. 1976. "The Common Structure of Statistical Models of Truncated, Sample Selection and Limited Dependent variables, and a Simple Estimator of Such Models." Annals of Economic and Social Measurement 5:475-492.

[34]. Little, Roderick J. A. and Donald B. Rubin. 1987. Statistical Analysis with Missing Data. New York: Wiley. SAS Global Forum 2012 Statistics and Data Analysis

[35]. Molenberghs, G. and Kenward, M.G. (2007) Missing Data in Clinical Studies. Chichester, UK: John

Wiley and Sons Ltd.

[36]. Mroz, T. A. (1987) "The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions." Econometrica 55, 765–799.

[37]. Raghunathan, Trivellore E., James M. Lepkowski, John Van Hoewyk and Peter Solenberger (2001) "A multivariate technique for multiply imputing missing valuesusing a sequence of regression models." Survey Methodology, 27:85-95.

[38]. Rubin, Donald B. 1976. "Inference and Missing Data." Biometrika 63: 581-592.

[39]. Rubin, Donald B. 1987. Multiple Imputation for Nonresponse in Surveys. New York: Wiley.

[40]. Schafer, Joseph L. 1997. Analysis of Incomplete Multivariate Data. London: Chapman & Hall.

[41]. Van Buuren, S., J.P.L. Brand, C.G.M. Groothuis-Oudshoorn and D.B. Rubin (2006) "Fully conditional specification in multivariate imputation." Journal of Statistical Computation and Simulation 76: 1046-1064.

[42]. Van Buuren, S., and C.G.M. Oudshoorn (2000). Multivariate Imputation by Chained Equations: MICE V1.0 User's Manual. Report PG/VGZ/00.038. Leiden: TNO Preventie en Gezondheid.