Missing Data - Introductory notes

David Pham

Contents

These are my current notes and the themes

1 Wikipedia

Case deletion (CC) The method does not introduce any bias if the missing values are uniformly distributed.

Single imputation If one sort the data matrix according some order, *Last observation carried forward* is the method of replacing the missing value with last valid value.

The missing value can also be replaced with the mean of the other observations, however, correlations are attenuated.

Regression imputation use the other variables as predictors to replace the missing value, although precision is misleadingly augmented, hence does not reflect the statistical errors of the missing data. This problem is partially solved by multiple imputation.

Multiple imputation The multiple imputation [Rubin (1987)] is similar to bootstrapping method: Missing variables are simulated, say B times, and the desired statistics are averaged except for the standard error which is constructed by adding the variance of the imputed data and the within variance of each data set.

2 Matloff Blog post

Complete-case analysis (CC), listwise deletion Delete all record for which at least one variable is missing.

Single and multiple imputation Estimation of the distribution of missing variables conditional on the others and then sampling from that distribution. Multiple alternate matrix are generated without the NAs.

In multiple imputation, the distribution of each variable conditional and the others is fitted and in case of missing value, a sample is drawn from this distribution.

Available cases (AC), pairwise deletion Keep the observation if the missing feature is not retained for the desired measure, for example the correlation (where only 2 variables are needed). It can, nonetheless, produce correlations over 1.

2.1 MCAR: Missing Completely at Random

Let Y the variable of interest, $M \in \{0, 1\}$ denotes if Y is missing, and D the other variables than Y. This is often denoted as

$$P(M = 1|Y = s, D = t) = P(M = 1),$$

or equivalently

$$P(Y = s, D = t|M = i) = P(Y = s, D = t), i \in \{0, 1\}.$$

2.2 MAR: Missing at Random

For multiple imputation, one requires only $M \perp Y|D$, that is

$$P(M = 1|Y = s, D = t) = P(M = 1|D = t),$$

2.2.1 Conditional estimation under MAR

In practice, problems arise as D might not hold any predictive ability of the desired variable and that D might as well contain missing data. Interestingly

$$\begin{aligned}
P(Y = s|D = t, M = i) &= \frac{P(Y = s, D = t, M = i)}{P(D = t, M = i)} \\
&= \frac{P(Y = s, D = t)P(M = i|Y = s, D = t)}{P(D = t, M = i)} \\
&= \frac{P(Y = s|D = t)P(D = t)P(M = i|D = t)}{P(D = t, M = i)} \\
&= P(Y = s|D = t).
\end{aligned}$$

Hence if we are interested in the relationship between Y and D, that is the conditional distribution Y given D, the fact that it is missing or not will not introduce bias, hence CC and AC would perform equally well. This is ironic as MAR is meant to apply where CC and AC should not be used.

2.2.2 Unconditional estimation under MAR

Observe that

$$P(Y = s|M = 0) = \frac{P(M = 0|Y = s)}{P(M = 0)} P(Y = s),$$

hence our estimation of $P(Y = s)$ might still be biased with the factor of $P(M = 0|Y = s)/P(M = 0)$.

3 Missing data (Schafer and Graham 2002)

> With or without missing data, the goal of a statistical procedure should be to make valid and efficient inferences about a population of interest—not to estimate, predict, or recover missing observations nor to obtain the same results that we would have seen with complete data.

Let Y_{com} denote the complete data, and denote its partitions with observed and missing data $Y_{com} = (Y_{obs}, Y_{mis})$. If R is the random variable representing missingnes, then MAR (also called ignorable nonresponse) is defined as

$$P(R|Y_{com}) = P(R|Y_{ob}),$$

and MCAR

$$P(R|Y_{com}) = P(R).$$

Missing not at random (MNAR) or nonignorable nonresponse, is the situation when MAR is violated. Issue with MAR is, it is often unverifiable, however, only little deviation of estimates and standard errors are observed in practice.

$P(Y_{com}; \theta)$ can be interpreted as either the sampling mechanism of Y_{com} with parameter θ or the likelihood function. The following formula

$$P(Y_{obs}; \theta) = \int P(Y_{com}; \theta) dY_{mis}$$

provides a sampling distribution only when MCAR holds and is a valid likelihood function when MAR is assumed (favoring the Bayesian view). For MNAR, R and an additional parameter ξ defining the distribution of R has to be added:

$$P(Y_{obs}, R; \theta, \xi) = \int P(Y_{com}; \theta) P(R; \xi) dY_{mis}.$$

3.1 Older Methods

Listwise and Pairwise deletion Listwise deletion (case deletion or complete-case analysis) dismiss all observation with any missing values and pairwise deletion (available-case analysis) uses different sets of sample units for different parameters. Critics of AC are that the standard errors or other measures of uncertainty are difficult to assess as the parameters are computed from different sets of units.

CC analysis only works with MCAR but even if it holds, MCAR can be inefficient (e.g. with large data matrix with mild rates of missing values.

Reweighting Reweighting can eliminate bias from CC, for more details (Little and Rubin 2002, chap. 4.4). It is easy to use with univariate and monotone missing patterns.

Average imputation It replaces the missing value with the mean of the observation. This introduce bias and underestimate the standard errors. The new value is an artifact of a specific data sets and disturbes the scale of the variables. If MI is not feasible, then averagin is a reasonable choice if

> reliability is high ($\alpha > 0.7$) and each group of items to e averaged seems to form a single, well, defined domain

3.2 Single imputation

Imputation is the process of predicting the missing value conditional on the other values. It has the advantages of sharing the same dataset to all researcher working on a common project. See (Little and Rubin 2002) for shortcomings of single imputation.

Imputing unconditional means Mean substitution consists of replacing the missing value of a variable with the average accross all the other non-missing observations. Weakness are that confidence intervals $\bar{y} \pm z_\alpha \sqrt{S^2/N}$ are narrowed by overstating the number of observation N and the downward bias into S^2. Under MCAR the coverage is only $2\Phi(z_\alpha r) - 1$ where r is the rate of missingness.

Imputing from unconditional distributions Hot deck imputation fills in nonrespondents' data with values from actual respondents, that is we replace with a random draw from the observed values. This methods still distort correlation and standard errors.

Imputing conditional means In the univariate situation (where only one value is of interest), one can fill with a prediction from the other variable using regression methods.

> This is nearly ptimal for a limited class of estimations problem if special correction are made to standard errors.

However, it overstates the correlation and covariance as R^2 for imputed value is 1.00.

Imputing from conditional distribution Under MAR assumption, the weaknesses from the previous methods are overcome by drawing an obesrvation from the fitted regression distribution of Y given X. In general, one has to sample from

$$P(Y_{mis}|Y_{obs}, \theta),$$

where, in practice, we replace θ with its estimated value $\hat{\theta}$ from Y_{obs}. With monotone patterns, one can set a seqence of regression for Y_j given $Y_1, \ldots, Y_{j-1}$, for $j \in 1, \ldots, p$.

Undercoverage and reasonnable application In a simulation exerices, one can deduce that the actual coverae is much loewr thatn 95%. Compared to CC, if the missing rate is low, single imputation might still be a valid method. For example, if $p = 25$ and the missing rate $r = 0.03$, then CC would delete $1 - (1 - r)^p = 0.53$ of the cases, whereas conditional distribution would allow to use all the participants.

3.2.1 Maximum likelihood estimation

One of the advantage of using the MLE $\hat{\theta}$ is hypothesis testing. If $\tilde{\theta}$ is the MLE for the null hypothesis, one could use likelihood-ratio tests and thus compare

$$2[l(\hat{\theta}; Y_{obs}) - l(\tilde{\theta}; Y_{obs})],$$

and the $(1 - \alpha)$-quantile of the χ^2_p distribution. Hence one would not need to compute the second derivative of l in order to get Fisher information (or equivalently the asymptotic standard error of the MLE).

In order to solve the maximization problem, one often resolve to use the EM algorithm. ML still has the problem of undercoverage.

Assumptions Sample size has to be large enough for the ML estimates be approximately unbiased and normally distributed and with missing data, the sample might be larger than usual. Thenlikelihood functions comes from an asusmed parametric model for complete data $P(Y_{obs}, Y_{mis}; \theta)$, hence departure from model assumptions might effect inference. MAR is still assumed.

3.2.2 Multiple imputation

Multiple imputation (MI) solves the problem of understating uncertainty. MI is similar to bootstrapping methods: one make artifical B samples and complete-case analysis. The final estimates (except standard errors) are then the arithmetic mean. Standard errors should reflect missing-data uncertainty and finte-sample variation.

An advantage of MI is the number of need imputation: the efficiency based on m samples relative o an infinite number is $(1 + \lambda/m)^{-1}$, where λ is the rate of missing information, which measures the increase in the large-sample variance of a paramter estimate due to missing values. $m = 20$ is often good in practice.

Combining standard errors In the one-dimensional case, if the sample is large enough so that the estimator Q follows a gaussian distribution, then the estimate $\hat{Q}$ and the standard error T can be computed from the estimates of $(Q^j, U^j)_{j=1}^m$, Q^j, respectively, U^j being the fitted value of Q, respectively the standard error, for data sets j

$$
\begin{aligned}
\hat{Q} &= m^{-1} \sum_{j=1}^{m} Q^j, \\
\hat{U} &= m^{-1} \sum_{j=1}^{m} U^j, \\
B &= (m-1)^{-1} \sum_{j=1}^{m} (Q^j - \hat{Q})^2, \\
T &= \hat{U} + (1 + m^{-1})B.
\end{aligned}
$$

For confidence interval, the Student's t approximation can be used with the degree of freedom given by

$$
\nu = (m-1)\left[1 + \frac{\hat{U}}{(1+m^{-1})B}\right]^2.
$$

The estimated rate of missing information for Q is approximately $\tau/(\tau+1)$ where $\tau = (1 + m^{-1})B/\hat{U}$, the relative increase in variance du to nonresponse. See [schafer1997@multivariate] for more cases.

This model still use the MAR asusmption.

Obviously, the missing values problem is dealt before the analysis with MI, in contrast with ML. The danger from MI is the ability ot use different models for imputation and analysis.

4 Data analysis using regression and multilevel/hierarchical models (Gelman and Hill 2006)

Chapter 25 of contains information about missing values.

5 Statistical analysis with missing data (Little and Rubin 2002)

The monograph describes mechanisms underlying the missingness come in several type (*mi*, *mice*, *Amelia* in R packages).

Bibliography

Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/hierarchical Models.* Cambridge University Press.

Little, RJA, and DB Rubin. 2002. "Statistical Analysis with Missing Data." Wiley.

Schafer, Joseph L, and John W Graham. 2002. "Missing Data: Our View of the State of the Art." *Psychological Methods* 7 (2). American Psychological Association: 147.