STAT/Q SCI 403: Introduction to Resampling Methods **Spring 2019**

Lecture 8: Missing Data and Imputation

Instructor: Yen-Chi Chen

Consider a regression problem where we have a binary covariate $X \in \{0,1\}$ and a continuous response $Y \in \mathbb{R}$. However, in our data, some response variables are missing and only the covariates are observed. So our data can be represented as

$$(X_1, Y_1), \cdots, (X_n, Y_n), (X_{n+1}, \star), \cdots, (X_{n+m}, \star).$$

The symbol $\star$ denotes a missing value. Namely, we have n observations that are fully observed while the other m observations that we only observe the covariate, not the response. Suppose that the parameter of interest is the marginal median of the response variable m_Y. How should we estimate the median?

We can introduce an additional variable R to denote the missingness such that $R = 0$ means that Y is not observed whereas $R = 1$ means that Y is observed. Note that R itself is another random variable.

Without any assumptions on the missing data, we are not able to accurately estimate the median consistently. There are two common assumptions people made about the missingness:

1. **MCAR:** missing completely at random. This means that the missingness is independent of any variables. Under the above notations, MCAR means that
$$R \perp X, Y.$$
2. **MAR:** missing at random. Under MAR, the missingness depends only on the observed pattern. In our case.
$$P(R = 0|X, Y) = P(R = 0|X)$$
since Y is not observed when $R = 0$.

When the missingness is neither MCAR nor MAR, it is called MNAR–missing completely at random.

Under MCAR, we can completely ignore the data with missing values and just use the sample median as an estimate of m_Y. However, under MAR, we cannot do such thing because the missingness may depends on X and if the distribution of covariate is different under fully observed data ($R = 1$) and partially observed data ($R = 0$), we will obtain a biased estimate.

While there are other ways to estimate the median under MAR, we will focus on the method of imputation.

8.1 Imputation

The idea of imputation is to impute a value to the missing entry so that after imputing all missing entries, we obtain a data without any missingness. Then we can simply apply a regular estimator (in the above example, sample median) to estimate the parameter of interest.

However, we cannot impute any number to the missing entry because this would cause bias in the estimation. We need to impute the value in a smart way. Generally, we want to impute the value according to the conditional density

$$p(y|x, R = 0),$$

the conditional density of response variable Y given the covariate X and the missing pattern $R = 0$. Namely, for $n + i$-th observation where only X_{n+i} is observed, we want to draw a random number

$$\widetilde{Y}_{n+i} \sim p(y|X_{n+i}, R = 0).$$

If indeed Y_{n+1} is from the above density function, one can show that the sample median

$$\text{median}\{Y_1, \cdots, Y_n, \widetilde{Y}_{n+1}, \cdots, \widetilde{Y}_{n+m}\}$$

is an unbiased estimator of m_Y.

This idea works regardless of what missing assumption is. However, the problem is that the density function $p(y|x, R = 0)$ cannot be estimated using our data because the only case we observed Y is when $R = 1$.

Under this case, MAR implies a powerful result:

$$p(y|x, R = 0) = p(y|x, R = 1). \tag{8.1}$$

Namely, the conditional density of Y given X is independent of the missing indicator R. To see how equation (8.1) is derived, note that MAR implies

$$P(R = 1|X, Y) = 1 - P(R = 0|X, Y) = 1 - P(R = 0|X) = P(R = 1|X).$$

Thus, the conditional density

$$\begin{aligned}
p(y|x, R = 0) &= \frac{p(y, x, R = 0)}{P(x, R = 0)} \\
&= \frac{p(x, y)P(R = 0|x, y)}{P(x, R = 0)} \\
&= p(x, y)\frac{P(R = 0|x)}{P(x, R = 0)} \\
&= p(x, y)\frac{1}{p(x)} \\
&= p(x, y)\frac{P(R = 1|x)}{P(x, R = 1)} \\
&= \frac{p(x, y)P(R = 1|x, y)}{P(x, R = 1)} \\
&= \frac{p(y, x, R = 1)}{P(x, R = 1)} \\
&= p(y|x, R = 1).
\end{aligned}$$

Thus, we obtain equation (8.1).

The power of equation (8.1) is that $p(y|x, R = 1)$ can be estimated by a KDE:

$$\begin{aligned}
\widehat{p}(y|x, R = 1) &= \frac{\frac{1}{nh}\sum_{i=1}^n K\left(\frac{Y_i - y}{h}\right) I(X_i = x)}{\frac{1}{n}\sum_{i=1}^n I(X_i = x)} \\
&= \frac{1}{n_x h}\sum_{i=1}^n K\left(\frac{Y_i - y}{h}\right) I(X_i = x),
\end{aligned}$$

where $n_x = \sum_{i=1}^n I(X_i = x)$ is the number of $X_i = x$ in the completely observed data and $x \in \{0, 1\}$. Namely, $\widehat{p}(y|x, R = 1)$ is the KDE applied to the completely observed data with the covariate $X = x$.

Given an observation $X_{n+i} = x$, how should we sample $\widehat{Y}_{n+i}$ from $\widehat{p}(y|x, R = 1)$? It is very simple. We first sample the index I such that

$$P(I = i|\mathsf{data}) = \frac{1}{n_x} I(X_i = x).$$

Namely, I is chosen from those fully observed data with the covariate $X_i = x$ with equal probability. Given I we then sample Y_{n+i} from the density function

$$q(y) = \frac{1}{h} K\left(\frac{Y_I - y}{h}\right).$$

Although this may look scary, if the kernel function is Gaussian, $q(y)$ is the normal density with mean Y_I and variance h^2. Namely, when K is a Gaussian,

$$\widehat{Y}_{n+i} \sim N(Y_I, h^2).$$

Remark.

- The use of KDE is just one example. You can use any density estimator for $\widehat{p}(y|x, R = 1)$ as long as you are able to sample from it.
- The equation (8.1) relies on the MAR assumption along with the fact that only one variable is subject to missing. When there are more than one variables that can be missing, we no longer have such a simple equivalence.
- The imputed data can be used for other estimators as well, not limited to estimating the median. You may notice that during our imputation process, we do not use any information about the estimator.
- There imputation methods that only imputes a fixed, non-random number for each missing entries. This is often called a deterministic imputation. For certain problem, a deterministic imputation works but in general, it may not work. So a rule thumb is to use a random imputation if possible.

8.2 Multiple imputation

After doing the imputation for all missing entries, we obtain a complete data

$$(X_1, Y_1), \cdots, (X_n, Y_n), (X_{n+1}, \widehat{Y}_{n+1}), \cdots, (X_{n+m}, \widehat{Y}_{n+1}).$$

The estimate of m_Y is just the sample median of this impute dataset. However, there will be Monte Carlo errors in this estimator because every time we do the imputation, we will not get the same number (due to sampling from $p(y|x, R = 0)$). If we just impute the data once (this is often called *single imputation*), we may suffer from the Monte Carlo errors a lot. Thus, a better approach is to perform a *multiple imputation*.

Multiple Imputation.[1] After obtaining a complete data, we do the same imputation procedure again, which gives us another new complete data. Then we keep repeating the above process, leading to several complete data, which can be represented as

$$\begin{aligned}
&(X_1, Y_1), \cdots, (X_n, Y_n), (X_{n+1}, \widehat{Y}^{(1)}_{n+1}), \cdots, (X_{n+m}, \widehat{Y}^{(1)}_{n+1})\\
&(X_1, Y_1), \cdots, (X_n, Y_n), (X_{n+1}, \widehat{Y}^{(2)}_{n+1}), \cdots, (X_{n+m}, \widehat{Y}^{(2)}_{n+1})\\
&\cdots\\
&(X_1, Y_1), \cdots, (X_n, Y_n), (X_{n+1}, \widehat{Y}^{(N)}_{n+1}), \cdots, (X_{n+m}, \widehat{Y}^{(N)}_{n+1}).
\end{aligned}$$

[1] For more introduction on this topic, see https://stats.idre.ucla.edu/stata/seminars/mi_in_stata_pt1_new/

We then combine all these datasets to a huge dataset and compute the estimator of the parameter of interest (in our case, median of the response variable). This estimator has a smaller Monte Carlo error.

8.3 More than one variable missing

When there are more than one variable that are subject to missing, the problem gets a lot more complex. Consider the case where each individual has p variables $X_1, \cdots, X_5$ and all of them may be missing and we may even have many of them missing at the same time. There are two categories of the missing patterns:

1. **Monotone missingness.** In this case, if X_t is missing, then X_s is also missing for any $s > t$. This occurs a lot in medical research due to *dropout* of the individuals. For instance, let X_t denote the BMI of an individual at year t. If this individual left the study at time point τ, then we only observe $X_1, \cdots, X_\tau$ from this individual. Any information beyond year τ is missing.
2. **Non-monotone missingness.** When the missing pattern is not monotone, it is called non-monotone missingness. The non-monotone missing data is a lot more challenging than monotone missing data because there are many possible missing pattern that can occur in the data. If there are p variables, them monotone missing data has p different missing patterns but the non-monotone case may have up to 2^p different missing patterns!

Let $R \in \{0,1\}^5$ be a multi-index set that denotes the observed pattern and we use the notation $X_R = (X_i : R_i = 1)$. For instance, $R = 11001$ means that we observe variable X_1, X_2, and X_5 and $X_{11001} = (X_1, X_2, X_5)$. Under this notation, the MAR assumption can be written as

$$P(R = r|X) = P(R = r|X_r),$$

namely, the probability of seeing a pattern $R = r$ only depends on the observed variable.

8.4 MAR: a deeper look ♠ (optional)

MAR is a very popular assumption that people often assumed in practice (although it may not be reasonable in some cases). However, under the non-monotone case, MAR tells us little about the missingness and it is actually not very to work with. Why is the MAR still so popular in practice?

There are two reasons for why MAR is so popular. The first reason is that under monotone missing data problem, MAR provides an elegant way to identify the entire distribution function. The second reason is that in both monotone and non-monotone case, MAR makes the likelihood inference a lot easier.

8.4.1 MAR under monotone case

Under the monotone missing problem, let T denotes the index of the last observed variable. Namely, the individual dropouts after time point T. We use the notation $X_{\leq t} = (X_1, \cdots, X_t)$. Then the MAR can be written as

$$P(T = t|X) = P(T = t|X_{\leq t}).$$

The above equation gives us a very powerful result–we can estimate the missing probability $P(T = t|X)$ for every $t = 1, \cdots, p$!

To see this, consider the case $t = 1$ so MAR implies

$$P(T = 1|X) = P(T = 1|X_1).$$

Note that $P(T > 1|X) = 1 - P(T = t|X) = P(T = 1|X_1) = P(T > 1|X_1)$. Thus, we can estimate $P(T = 1|X_1)$ by comparing pattern $T = 1$ against $T > 1$ given the variable X_1, which is always observed. Thus, $P(T = 1|X)$ is estimatible. For $t = 2$, the MAR implies

$$P(T = 2|X) = P(T = 2|X_1, X_2).$$

Thus,

$$P(T > 2|X) = P(T = 2|X) + P(T = 1|X) = P(T = 2|X_1, X_2) + P(T = 1|X_1) = P(T > 2|X_1, X_2).$$

Again, we can compare the pattern $T = 2$ against $T > 2$ and estimate the probability $P(T = 2|X)$. We can keep doing this procedure, and eventually all missing probability $P(T = t|X)$ can be estimated.

For instance, if we are interested in estimating the parameter of interest $\rho = \mathbb{E}(\omega(X_1, \cdots, X_p))$, we can then use the famous *inverse probability weighting* estimator[2]:

$$\widehat{\rho} = \frac{1}{n\widehat{P}(T = p|X)} \sum_{i=1}^{n} \omega(X_{i,1}, \cdots, X_{i,p}) I(T_i = p),$$

where $\widehat{P}(T = p|X)$ is an estimate of $P(T = p|X)$. Sometimes, people also called $P(T = p|X)$ as the propensity score.

Actually, MAR under monotone missingness is equivalent to the *available case missing value* (ACMV) assumption:

$$p(x_{t+1}|x_{\leq t}, T = t) = p(x_{t+1}|x_{\leq t}, T > t)$$

for every t. The right-hand side can be estimated by conditional KDE so the density function[3]

$$p(x_{>t}|x_{\leq t}, T = t) = \prod_{s=t}^{p-1} p(x_{s+1}|x_{\leq s}, T = s)$$

can be estimated under ACMV assumption. Why is the above density estimatible so useful? This is because the joint density function has the following pattern mixture model formulation:

$$p(x) = \sum_{t=1}^{p} p(x, t) = \sum_{t=1}^{p} p(x_{>t} \mid x_{\leq t}, T = t) p(x_{\leq t} \mid T = t) p(T = t),$$

where both $p(x_{\leq t} \mid T = t)$ and $P(T = t)$ can be directly estimated using our data so what remains unknown is the density function $p(x_{>t} \mid x_{\leq t}, T = t)$. ACMV implies an estimator of this density function, so the entire joint density function can be estimated.

8.4.2 Likelihood inference with MAR

Another reason why MAR is so popular is its ignorability property, which hold even when the missing is non-monotone. Consider the joint density function $p(x, r)$ of both variable of interest X and the missing pattern R. Recall that $X_R = (X_i : R_i = 1)$ are the observed variables under pattern R. We also denote $X_{\bar{R}} = (X_i : R_i = 0)$ as the missing variables.

[2] See https://en.wikipedia.org/wiki/Inverse_probability_weighting for more details.
[3] Also called the *extrapolation density*.

We can then factorize it into

$$p(x,r) = P(R=r|X=x)p(x).$$

Suppose we use parametric models separately for both $P(R=r|X=x)$ and $p(x)$, leading to

$$p(x,r;\phi,\theta) = P(R=r|X=x;\phi)p(x;\theta) \overset{(MAR)}{=} P(R=r|X_r=x_r;\phi)p(x;\theta).$$

In our data, what we observed are (x_r, r) so we should integrate over the missing variables $x_{\bar{r}}$:

$$p(x_r,r;\phi,\theta) = \int p(x,r;\phi,\theta)dx_{\bar{r}} = P(R=r|X_r=x_r;\phi)\int p(x;\theta)dx_{\bar{r}}.$$

Thus, the log-likelihood function is

$$\begin{aligned}\ell(\theta,\phi|x_r,r) &= \log P(R=r|X_r=x_r;\phi) + \log\int p(x;\theta)dx_{\bar{r}}\\ &= \ell(\phi|x_r,r) + \ell(\theta|x_r,r),\\ \ell(\phi|x_r,r) &= \log P(R=r|X_r=x_r;\phi)\\ \ell(\theta|x_r,r) &= \log\int p(x;\theta)dx_{\bar{r}}.\end{aligned}$$

Since θ is the parameter of the distribution of the variable of interest X, we are often just interested in it. The above log-likelihood shows that maximizing θ and maximizing ϕ can be done separately. So finding the MLE of θ can be done without estimating the parameter ϕ, leading to a simple procedure[4].

[4] Here the likelihood function is defined through $\ell(\theta|x_r,r) = \log\int p(x;\theta)dx_{\bar{r}}$. The maximization of this function is often done through the EM algorithm and is related to problems such as mixture model. See https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm.