Multiple Imputation

If no missing data is present our statistical methods provide valid inference only if the following assumptions are met:

- For Generalized Estimating Equations, the mean function is correctly specified
- For likelihood-based methods, the probability density function including the mean and variance are correctly specified

Missing data can seriously compromise inferences from randomized clinical trials, especially when handled incorrectly, but inference is still possible with the correct methods.

Missing values in longitudinal studies may occur intermittently when individuals miss one or more planned visits, or drop out early.

Types of Missing Data

- Missing Completely at Random (MCAR) - Missingness is independent of both observed and unobserved data. More formally, the probability of missing data in Y is unrelated to the **value** of Y itself or any other variable X. However, it does allow for the possibility that missingness is Y is related to missingness in some other variable X.
  - Ex. In determining predictors of income, MCAR assumption would be violated if people who reported income were on average younger than the people who did report it.
- Missing at Random (MAR) - Missingness is independent of missing responses after controlling for other variables X. Formally: P(Y missing | Y,X) = P(Y missing | X)
  - Ex. The MAR assumption is satisfied if the probability of missing data on income depended on a person's age, but within each age group the probability of missing income was unrelated to income. Obviously, this cannot be tested as we do not know the missing values of the data.
- Missing Not at Random (MNAR) - Missing value depend on unobserved values.
  - Ex. High income people are less likely to report their income.
  - Also referred to as *non-ignorable* missing or *informative dropout*

Multiple Imputation

Imputation is substituting each missing value with a reasonable guess, which can be done using a variety of methods. In multiple imputation, imputed values are drawn from a distribution so they inherently contain some variation. Thus, it addresses shortcomings of single imputation by introducing an additional form of error based on variation in the parameter estimates across imputation called between imputation error. Since this is a simulation-based procedure, the purpose is **not** to re-create the individual missing values as close as possible to the true ones, but

to handle missing data to achieve valid inference.

It involves 3 steps:

1. Run an imputation model defined by the chosen variables to create imputed data sets. In other words, the missing values are filled in *m* times to generate *m* complete data sets.
   - The standard is m = 10
   - Choosing the correct model requires considering:
     - Which variables have missing values?
     - Which has the largest proportion of missing values?
     - Are there patterns to the missingness?
       - Monotone (dropouts in longitudinal studies) or arbitrary
2. Perform an analysis on each of the *m* completed data sets by using a BY statement in conjunction with an appropriate analytic procedure (MIXED or GENMOD in SAS)
   - Parameter estimates, standard errors, etc. should be considered
3. The parameter estimates from each imputed data set is combined to get a final set of parameter estimates

Pros: Same properties as ML but removes limitations and can be used with any kind of data or software. When the data is MAR, multiple imputation can lead to consistent, asymptotically efficient and asymptotically normal estimates.

Cons: It is challenging to use successfully. It produces different estimates every time.

Use multiple imputation when:

1. When there are covariates associated with the missingness of the response but not normally used in the analysis model.
   - Ex. In a clinical trial missingness could be related to a side effect which is not a variable in the analysis
2. When there are missing covariates; as likelihood-based methods with incomplete covariates are not normally implemented in statistical software and omitted by default.
3. When full likelihood methods are not straightforward as in the case of discrete outcomes where GEE methods are often used, although GEE methods are only valid under MCAR and sometimes MAR

Regression-Based Imputation

Particularly with monotone missingness, we can fit a linear regression model to predict missing values Y.

1. Randomly draw from a chi-squared distribution with (N_j - q) degrees of freedom where N_j is the number of subjects who haven't dropped out at the jth occasion and q is the number of covariates used to predict Y.
2. Calculate the residual variance of the kth draw:
   $$\sigma^2 = (N_j - q) \hat\sigma^2 / \chi^2 $$

3. Randomly draw regression parameters γ from a multivariate distribution N(γ, Cov(γ)) where:
   $$ Cov(\hat\gamma) = \sigma^2 (\sum_{i=1}^{N_j} Z_{ij} Z'_{ij})^{-1} $$
4. Draw *e* from N(0, σ^2), where σ^2 is the estimate of residual variance
5. Calculate $Y_{ij} = Z'_{ij}\gamma + e$
6. Repeat 1-5 *m* times

Predictive Mean Matching

This method is very similar to regression based imputation. This is more robust against misspecification of the regression model and ensures all imputed values are plausible.

1. See step 1 above
2. See step 2 above
3. See step 3 above
4. Calculate $Y_{ij} = Z'_{ij}\gamma$
5. Select a subset of K observations whose predicted values are closest to Y_{ij}
6. Impute the missing value by randomly drawing from these K observed values
7. Repeat step 1-6 *m* times.

Bayesian Principals of Imputation

Y^{obs} = Observed (vector of) quantities
Y^{mis} = Missing (vector of) quantities
θ = Parameter of interest (unobserved)
R = Indicator variable which takes the value 1 for observed part of Y and 0 elsewhere (observed)
τ = Parameter (vector) to describe missing data mechanism (unobserved)

Assume our data has a prior distribution: $\pi(Y_i | X_i, \tau)$ where $\tau = (\beta, \theta)$

The predictive posterior:
$\pi(Y^{mis}_i | Y^{obs}_i, X_i) = \int \pi(Y^{mis}_i | Y^{obs}_i, X_i, \tau) \pi(\tau | Y^{obs}_i, X_i) d\tau$

And the observed-data posterior is closely related:
$\pi(\tau | Y^{obs}_i, X_i) = \int \pi(\tau | Y^{obs}_i, Y^{mis}_i, X_i) \pi(Y^{mis}_i | Y^{obs}_i, X_i) dY_{mis}$
$= E_{Y^{mis}_i | Y^{obs}_i} (\pi(\tau | Y^{obs}_i, Y^{mis}_i, X_i)$

Markov Chain Monte Carlo for Multiple Imputation

1. Imputation step: Given a current estimate $\hat\tau^k$ of the parameters, first simulate a draw from the conditional predictive distribution of $Y^{mis^{k + 1}}_i$ conditional on the observed values and tau:
   $$ Y^{mis^{k + 1}}_i \sim \pi(Y^{mis}_i | Y^{obs}_i, X_i, \hat\tau^k) $$

2. Posterior P-step: Given a complete sample \((Y^{obs^k}_i , Y^{mis^{k+1}}_i)\) take a random draw from the complete-data posterior:
   $$ \hat\tau^{k + 1} \sim \pi(\tau | Y^{obs}_i, Y^{mis^{k+1}}_i, X_i) $$
3. Repeat these two steps starting from \(\hat\tau^0\), create a Markov chain, {\(\hat\tau^k, Y^{mis^k}_i, k = 1, 2 ... \)} whose stationary distribution is \(\tau, Y^{mis}_i. X_i \) with stationary distributions:
   \(\hat\tau^k (k = 1,2,...) \sim \pi(\tau | Y^{obs}_i, X_i)\) and
   \(Y^{mis^k}_i (k = 1,2,...) \sim \pi(Y^{mis}_i | Y^{obs}_i, X_i)\)

SAS Code

```
/*We have created 25 versions of the same dataset
with no missing values. We can run proc mixed
for each version seperately...*/
proc mixed data=MITLC_long;
where _imputation_=2;
class TRT TIME;
model y=time time*trt/s covb;
repeated time/type=un subject=id;
run;

/*
...or run all 25 in one run using a by statement
and saving the solutions using an ods output statement
*/
proc mixed data=MITLC_long;
class TRT TIME;
model y=time time*trt/s covb;
repeated time/type=un subject=id;
by _IMPUTATION_;
ods output solutionf=beta covb=varbeta;
run;

proc mianalyze parms=beta;
class TRT TIME;
modeleffects intercept time TRT*time;
run;

/*Using MCMC for imputation*/
proc sort data=TLC_missing;
by TRT;
```

```
run;
proc mi data=TLC_missing seed=364865 nimpute=25 out=miTLC_MCMC;
var y4 y6;
by TRT;
mcmc chain=multiple displayinit initial=em(itprint);
run;

data MITLC_MCMC_long;
set MITLC_MCMC;
y=y0;time=1;OUTPUT;
y=y1;time=2;OUTPUT;
y=y4;time=4;OUTPUT;
y=y6;time=6;OUTPUT;
drop y0 y1 y4 y6;
run;

proc sort data=MITLC_MCMC_long;
by _IMPUTATION_;
run;

proc mixed data=MITLC_MCMC_long;
class TRT TIME;
model y=time time*trt/s covb;
repeated time/type=un subject=id;
by _IMPUTATION_;
ods output solutionf=beta_mcmc covb=varbeta_mcmc;
run;

proc mianalyze parms=beta_mcmc;
class TRT TIME;
modeleffects intercept time TRT*time;
run;
```

Revision #49
Created 25 March 2023 20:16:53 by Elkip
Updated 29 March 2023 02:37:30 by Elkip