BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 19

Missing Data and Dropout

Missing Data and Dropout

In longitudinal studies missing data are the rule not the exception.

The term "missing data" is used to indicate that an intended measurement could not be obtained.

With missing data there must necessarily be some loss of information.

Of greater concern, missing data can introduce bias and result in misleading inferences about change over time.

When data are missing we must carefully consider the reasons for missingness.

Unequal n_i per Subject or Unbalanced Designs

Basically, the methods that we have discussed so far can handle situations in which $n_i \neq n$ for all i (i.e., unequal number of observations per subject).

That is, modern regression methods can handle unbalanced longitudinal designs with relative ease.

However, we do need to be more careful when $n_i \neq n$ for all i due to missingness.

Missing Data

Why might we have $n_i \neq n$ for all i?

For most designed studies, we *plan* on measuring the same number of outcomes, so if $n_i \neq n$ for all i, then some outcomes are *missing*.

Let Y denote the complete response vector which can be partitioned into two sub-vectors:

(i) Y^O the measurements observed

(ii) Y^M the measurements that are missing

If there were no missing data, we would have observed the complete response vector Y.

Instead, we get to observe Y^O.

The main problem with missing data is that distribution of the observed data may not be the same as distribution of the complete data.

Consider the following simple illustration:

Suppose we intend to measure subjects at 6 months (Y_1) and 12 months (Y_2) post treatment.

All of the subjects return for measurement at 6 months, but many do not return at 12 months.

If subjects fail to return at 12 months because they are not well (say, values of Y_2 are low), then distribution of observed Y_2's will be positively skewed compared to distribution of Y_2's in the population of interest.

When data are missing we must carefully consider the reasons for missingness.

Estimation of β with missing data depends on the missing data mechanism.

The missing data mechanism is a probability model for missingness:

- Missing Completely at Random (MCAR)
- Missing at Random (MAR)
- Not Missing at Random (NMAR)

Missing Completely at Random (MCAR)

MCAR: probability that responses are missing is unrelated to either the specific values that, in principle, should have been obtained (the *missing* responses) or the set of *observed* responses.

MCAR: probability responses are missing is independent of Y^O and Y^M.

Missingness is simply the result of a chance mechanism that is unrelated to either observed or unobserved components of the outcome vector.

Consequently, observed data can by thought of as a random sample of the complete data.

Covariate-Dependent Missingness

If missingness depends only on X, then technically it is MCAR. However, sometimes this is referred to as *covariate dependent* non-response.

In general, if non-response depends on covariates, X, it is harmless and same as MCAR provided you always condition on the covariates (i.e., incorporate the covariate in the analysis).
This type of missingness is only a problem if you do not condition on X.

Example 1: Consider the case where missingness depends on treatment group. Then the observed means in each treatment group are unbiased estimates of the population means.

However, the marginal response mean, averaged over the treatment groups, is not unbiased for the corresponding mean in the population (the latter, though, is usually not of subject-matter interest).

Sometimes it may be necessary to introduce additional covariates, or stratifying variables, into the analysis to control for potential bias due to missingness.

Example 2: Suppose the response Y is some measure of health, and X_1 is an indicator of treatment, and X_2 is an indicator of side-effects. Suppose missingness depends on side-effects.

If side-effects and outcome are uncorrelated, then there will be no bias.

If side-effects and outcome are correlated, then there will be bias unless you stratify the analysis on both treatment and side-effects (analogous to confounding).

MCAR:

- The means, variances, and covariances are preserved.

- Can use ML/REML estimators for β

- More generally, we can use GLS or GEE estimator with any "working" assumption for the covariance; normality assumption for Y_{ij} (for continuous outcomes) is not necessary

- If we use GLS or GEE estimator with incorrect "working" assumption for the covariance, then must use "empirical" or "sandwich" variance estimator for $\text{Cov}(\widehat{\beta})$

Any method of analysis that yields valid inferences in absence of missing data is also valid when missing data are MCAR and analysis is based on all available data, or even when restricted to so-called "completers".

Given that valid estimates of the means, variances, and covariances can be obtained, GLS or GEE provides valid estimates of β without requiring any distributional assumptions for Y_i.

The GLS or GEE estimator of β is valid provided the model for the mean response has been correctly specified; it does not require any assumptions about the joint distribution of the longitudinal responses.

$\Longrightarrow$ With complete data or data MCAR, distributional assumption is not required.

Missing at Random (MAR)

MAR: probability that responses are missing depends on the set of observed responses, but is unrelated to the specific missing values that, in principle, should have been obtained.

MAR: probability that responses are missing depends on Y^O, but is conditionally independent of Y^M.

Note 1: If subjects are stratified on the basis of similar values for the responses that have been observed, then within strata missingness is simply the result of a chance mechanism unrelated to unobserved responses.

Note 2: Because missingness depends on observed responses, the distribution of Y_i in each of the distinct strata defined by the patterns of missingness is not the same as the distribution of Y_i in the target population.

The "completers" are a biased sample from the target population.

Features of MAR

Means, variances, and covariances are <u>not</u> preserved:

So, if

$$\underset{n\times 1}{E\ (Y_i)} \ = \ \underset{n\times p}{X_i\beta} \quad \text{with complete data}$$

In general

$$\underset{n_i\times 1}{E\ (Y_i)} \ \neq \ \underset{n_i\times p}{X_i\,\beta}, \ \ \mathrm{Cov}(Y_i) \neq \Sigma_i$$

This implies that sample means, variances, and covariances based on either the "completers" or the available data are biased estimates of the corresponding parameters in the target population.

However, the likelihood is preserved.

ML estimation (e.g., PROC MIXED) of β is valid when data are MAR provided the multivariate normal distribution has been correctly specified.

This requires correct specification of not only the model for the mean response, but also the model for the covariance among the responses.

In a sense, ML estimation allows the missing values to be validly "predicted" or "imputed" using the observed data and a correct model for the joint distribution of the responses.

Not Missing at Random (NMAR)

NMAR: probability that responses are missing is related to the specific values that should have been obtained.

An NMAR mechanism is often referred to as "non-ignorable" missingness.

Challenging problem and requires modelling of missing data mechanism; moreover, specific model chosen can drive results of analysis.

Sensitivity analyses are recommended.

Dropout

Longitudinal studies often suffer from problem of attrition; i.e., some individuals "drop out" of the study prematurely.

This is where an individual is observed from baseline up until a certain point in time, thereafter no more measurements are made.

Term *dropout* refers to special case where if Y_{ik} is missing, then $Y_{ik+1}, ..., Y_{in}$ are also missing.

This gives rise to so-called "monotone" missing data pattern displayed in Figure 1.

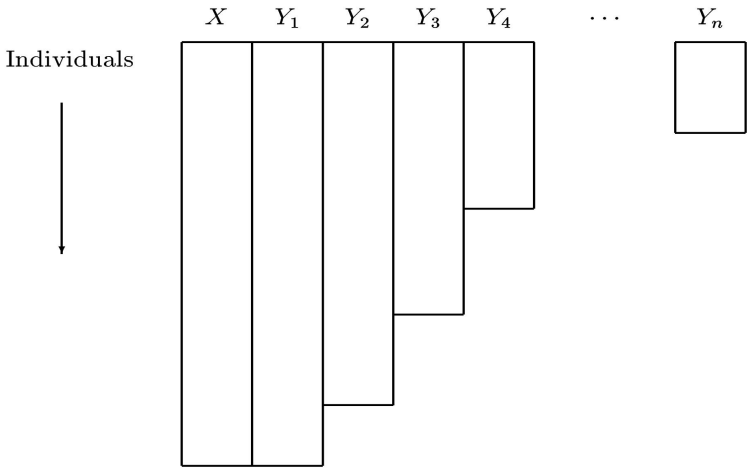

Figure 1: Schematic representation of a monotone missing data pattern for dropout, with Y_j more often observed than Y_{j+1} for $j = 1, ..., n-1$.

Possible reasons for dropout:

1. Recovery

2. Lack of improvement or failure

3. Undesirable side effects

4. External reasons unrelated to specific treatment or outcome

5. Death

When there is dropout, key issue is whether those who "drop out" and those who remain in the study differ in any further relevant way.

If they do differ, then there is potential for bias.

The taxonomy of missing data mechanisms (MCAR, MAR, and NMAR) discussed earlier can be applied to dropout.

Common Approaches for Handling Dropout

Complete-Case Analysis:

Exclude all data from the analysis on any subject who drops out.

That is, a so-called "complete-case" analysis can be performed by excluding any subjects that do not have data at all intended measurement occasions.

This method is very problematic and is rarely an acceptable approach to the analysis.

It will yield unbiased estimates of mean response trends only when dropout is MCAR.

Even when MCAR assumption is tenable, complete-case analysis can be immensely inefficient.

Available-Data Analysis:

General term that refers to a wide collection of techniques that can readily incorporate vectors of repeated measures of unequal length in the analysis.

Standard applications of GLS or GEE are available-data methods.

In general, available-data methods are more efficient than complete-case methods.

Drawbacks of available-data methods:

(i) Sample base of cases changes over measurement occasions.

(ii) Pairwise available-data estimates of correlations can lie outside (-1, 1).

(iii) Many available-data methods yield biased estimates of mean response trends unless dropout is MCAR.

Imputation

Imputation: substitute or fill-in the values that were not recorded with imputed values.

Once a filled-in data set has been constructed, standard methods for complete data can be applied.

Validity of method depends on how imputation is done.

Methods that rely on just a single imputation fail to acknowledge the uncertainty inherent in the imputation of the unobserved responses.

"Multiple imputation" circumvents this difficulty.

Multiple Imputation (MI): Missing values are replaced by a set of m plausible values, thereby acknowledging uncertainty about what values to impute.

Typically, a small number of imputations, for instance, $5 \leq m \leq 10$, is sufficient.

The m filled-in data sets produce m different sets of parameter estimates and their standard errors.

These are then combined to provide a single estimate of the parameters of interest, together with standard errors that reflect the uncertainty inherent in the imputation.

"Last Value Carried Forward" (LVCF):

One widely used imputation method, especially in clinical trials, is LVCF.

Regulatory agencies such as FDA seem to encourage the continuing use of LVCF.

LVCF makes a strong, and often very unrealistic, assumption that the responses following dropout remain constant at the last observed value prior to dropout.

There appears to be some statistical folklore that LVCF yields a *conservative* estimate of the comparison of an active treatment versus the control.

This is a gross misconception!

Except in very rare cases, we do not recommend the use of LVCF as a method for handling dropout.

Model-Based Imputation:

There is a related form of "imputation" where missing responses are *implicitly* imputed by modelling joint distribution of Y_i, $f(Y_i|X_i)$.

When dropout is MCAR or MAR, likelihood-based methods can be used based solely on the marginal distribution of the observed data.

In a certain sense, the missing values are validly predicted by the observed data via the model for the conditional mean of the missing responses given the observed responses (and covariates).

However, likelihood-based approaches require that model for $f(Y_i|X_i)$ is correctly specified (e.g., any misspecification of the covariance will, in general, yield biased estimates of the mean response trend).

Weighting Methods

In weighting methods, under-representation of certain response profiles in the observed data is taken into account and corrected.

These approaches are often called "propensity weighted" or "inverse probability weighted" methods.

Basic Idea: Base estimation on the observed responses but weight them to account for the probability of remaining in the study.

Intuition: Each subject's contribution to the weighted analysis is replicated to count both for that subject and for those subjects with the same history of responses and covariates, but who dropout.

Propensities for dropout can be estimated as a function of observed responses prior to dropout and covariates.

Inverse probability weighted methods were first proposed in sample survey literature, where the weights are known.

In contrast, with dropout the weights are not known, but must be estimated from the observed data.

In general, weighting methods are valid provided model that produces the estimated weights is correctly specified.

Inverse Probability Weighted Methods

Let $R_i = \left(R_{i1}, \ldots, R_{in)}\right)'$ denote a vector of "response indicators":

$$R_{i1} = 1 \text{ if } Y_{ij} \text{ is observed, } R_{i1} = 0 \text{ if not observed.}$$

Let $D_i = 1 + \sum_{j=1}^{n} R_{ij}$ denote the occasion dropout occurs.

Then let π_{ij} denote the *conditional* probability of the i^{th} individual being observed (not dropping out) at the j^{th} occasion, given that this individual was observed at the prior occasions ($R_{i1} = 1$, or person i not in study).

$$\begin{aligned} \pi_{ij} &= P(D_i > j | D_i \geq j, X_i, Y_i) \\ &= P(R_{ij} = 1 | R_{i1} = \ldots, R_{i,j-1} = 1, , X_i, Y_i) \\ &= P(R_{ij} = 1 | R_{i1} = \ldots, R_{i,j-1} = 1, , X_i, Y_{i1}, \ldots, Y_{i,j-1}) \end{aligned}$$

The unconditional probability of being observed at the j^{th} occasion can be expressed as the cumulative product of conditional probabilities:

$$\pi_{i1} \times \pi_{i2} \times \ldots, \pi_{ij}$$

The appropriate weight for Y_{ij} in the analysis is the inverse of this probability:

$$w_{ij} = \frac{1}{\pi_{i1} \times \pi_{i2} \times \ldots, \pi_{ij}}$$

Technical Notes:

1. Beware of small cumulative probabilities $\Rightarrow$ large weights. Leads to an analysis unduly influenced by just a few points.

2. An IPW-GEE analysis should use the "working independence" correlation structure. This approach uses the robust standard errors to guard against misspecification of this independence assumption.

Case Study

We consider data from a longitudinal clinical trial of contracepting women.

Women received an injection of either 100 mg or 150 mg of depot-medroxyprogesterone acetate (DMPA) on the day of randomization three additional injections at 90-day intervals, and a final follow-up visit 90 days after the fourth injection.

Menstrual diary data were used to determine whether a women experienced amenorrhea, the absence of menstrual bleeding for a specified number of days, in each of the four successive 3-month intervals.

A total of 1151 women completed the menstrual diaries.

A feature of this clinical trial is that there was substantial dropout. More than one third of the women dropped out before the completion of the trial.

Interest focuses on a logistic regression model for the marginal probability of amenorrhea:

$$\text{logit}\left(\mu_{ij}\right) = \beta_1 + \beta_2 t_{ij} + \beta_3 t_{ij}^2 + \beta_4 \text{Dose}_i + \beta_5 \left(t_{ij} \times \text{Dose}_i\right) + \beta_6 \left(t_{ij}^2 \times \text{Dose}_i\right)$$

The following fits an all available data (AD-GEE) and an IPW-GEE to the data.

MAR model for missingness:

$$\text{logit}\left(\pi_{ij}\right) = \theta_1 + \theta_2 I(t=2) + \theta_3 I(t=3) + \theta_4 \text{Dose}_i + \theta_5 Y_{i,j-1} + \theta_6 \left(\text{Dose}_i \times Y_{i,j-1}\right)$$

Notes:

1. The estimated weights ranged from 1.0 to 2.1, so there was no concern that a small set of observations would unduly influence the IPW-GEE results.

2. Both analyses (AD and IPW) use a "working independence" assumption for the correlation.

Note: Programs and data available on FLW book website (Section 18.4)

SAS Code

```
data contracep;
     infile 'contracep.dat';
     input id dose time y prevy r;

proc sort data=contracep;
     by id time;
run;

proc genmod descending;
     class time (param=ref ref="1");
     model r = time dose prevy dose*prevy / dist=bin link=logit;
     where time ne 0;
     output out=predict p=probs;
run;
```

```
proc sort data=predict;
     by id time;

data wgt (keep=id time cumprobs probs);
     set predict;
     by id time;
     retain cumprobs;
     if first.id then cumprobs=probs;
     else cumprobs=cumprobs*probs;

data combine;
     merge contracep wgt;
     by id time;
     if (time=0) then ipw=1;
     else ipw=1/cumprobs;
```



```
proc genmod descending data=combine;
     weight ipw;
     class id;
     model y = dose time time*time dose*time dose*time*time /
                       dist=bin link=logit;
     repeated subject=id / type=ind;
run;
```

Estimated marginal rates of amenorrhea for quadratic trend model using GEE with all available data (AD) and inverse probability weighting (IPW)

Method	Time	100 mg	150 mg	Diff	SE	Z
AD	3 mo.	0.184	0.201	0.017	0.023	0.73
	6 mo.	0.274	0.363	0.089	0.025	3.55
	9 mo.	0.388	0.499	0.111	0.030	3.68
	12 mo.	0.517	0.572	0.055	0.036	1.52
IPW	3 mo.	0.183	0.200	0.017	0.023	0.73
	6 mo.	0.276	0.361	0.084	0.026	3.29
	9 mo.	0.393	0.496	0.104	0.031	3.34
	12 mo.	0.521	0.570	0.049	0.037	1.33

Summary

In longitudinal studies missing data are the rule not the exception.

Missing data have two important implications:
(i) loss of information, and
(ii) validity of analysis.

The loss of information is directly related to the amount of missing data; it will lead to reduced precision (e.g., larger SEs, wider CIs) and reduced statistical power (e.g., larger p-values).

The validity of the analysis depends on assumptions about the missing data mechanism.

Likelihood-based methods (e.g., PROC MIXED) are valid under MAR or MCAR.

The distinction between MAR and MCAR determines the appropriateness of ML estimation under the assumption of normality (for continuous outcomes) versus GLS or GEE estimation without requiring distributional assumptions.

With complete data or data MCAR, full distributional assumption is not required.

With data MAR, full distributional assumption is required (e.g. multivariate normality, GLMM formulation for binary outcomes) and correct models for both the mean response and the covariance.