Overview

Overview

Types of Covariates

There are two types of covariates:

Time-stationary covariates: For example

- gender
- race
- treatment (if fixed for the whole study period)
- study center
- ...

Time-varying covariates: For example

- age
- dietary intake
- bloodmarker
- air pollution
- treatment in certain studies (crossover studies, observational studies)
- ...

Types of Time-Varying Covariates

There are also two types of time-varying covariates:

Fixed by study design: For example

- treatment in a crossover study
- time since baseline (when measurement times fixed by study design)

Stochastic covariates, which vary randomly over time: For example

- blood pressure
- dietary intake
- air pollution exposure
- blood marker

When a covariate is time-varying and stochastic, we may encounter new issues in its interpretation as well as in estimation of regression parameters.

Time-Varying Covariates

In our models, we assumed a relationship for the mean

$$g(\mathrm{E}(Y_{ij}|\boldsymbol{X}_{ij})) = \boldsymbol{X}'_{ij}\boldsymbol{\beta}$$

for a known link function $g(\cdot)$. Implicitly, we are assuming that the conditional mean of Y_{ij} given $\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{in_i}$ only depends on $\boldsymbol{X}_{ij}$,

$$\mathrm{E}(Y_{ij}|\boldsymbol{X}_i) = \mathrm{E}(Y_{ij}|\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{in_i}) = \mathrm{E}(Y_{ij}|\boldsymbol{X}_{ij}).$$

This is of course true for time-invariant variables. It may not hold for time-varying stochastic covariates, however. In this case, preceding or subsequent values of the time-varying covariate can 'confound' the relationship between Y_{ij} and $\boldsymbol{X}_{ij}$, and can lead to biased estimates of parameters of interest.

Time-Varying Covariates

Example: Longitudinal study on the effects of physical activity in reducing blood glucose levels in patients with type 2 diabetes. If patients with high blood glucose levels at one visit increase their physical activity subsequently (feeling guilty! - feedback mechanism), while other patients do not, we will underestimate the relationship between physical activity and blood glucose level.

Here: the current value of Y_{ij}, given X_{ij}, predicts X_{ij+1}, and thus

$$\mathrm{E}(Y_{ij}|X_{i1},\ldots,X_{in_i}) \neq \mathrm{E}(Y_{ij}|X_{ij}).$$

External Covariates

A covariate is called exogenous or external when

$$[X_{i,j+1}|X_{i1},\ldots,X_{ij},Y_{i1},\ldots,Y_{ij}] = [X_{i,j+1}|X_{i1},\ldots,X_{ij}].$$

Otherwise, the covariate is called internal or endogenous.

An example of an external covariate is ambient air pollution. While ambient air pollution is time-varying and stochastic, future values conditional on past values are not predicted by the health outcome studied.
However, personal exposure to air pollution would not be an external covariate if study subjects with poor health outcomes decided to stay indoors to avoid exposure to high levels of air pollution.

For an external covariate,

$$\mathrm{E}(Y_{ij}|\boldsymbol{X}_i) = \mathrm{E}(Y_{ij}|X_{i1},\ldots,X_{in_i}) = \mathrm{E}(Y_{ij}|X_{i1},\ldots,X_{ij}).$$

External Covariates

When variables are external, we can focus on specifying a model for $[Y_{ij}|X_{i1}, \ldots, X_{ij}]$. Possible models include

- concurrent, model $\mathrm{E}(Y_{ij}|X_{ij})$
- lagged, model $\mathrm{E}(Y_{ij}|X_{i,j-k})$ for some k
- cumulative, model $\mathrm{E}(Y_{ij}|\sum_{k=1}^{j} X_{ik})$
- distributed lags, regression coefficients for $X_{ij}, \ldots, X_{i,j-k}$ follow some pre-specified structure (e.g. polynomial)

However, be aware that e.g. modeling $\mathrm{E}(Y_{ij}|X_{ij})$, while Y_{ij} depends on both X_{ij} **and** $X_{i,j-1}$, can still give misleading results.

When variables are internal, we have to think both about meaningful targets of inference and valid methods of inference. Methods include causal inference, and modeling of the joint process $\{Y_j, X_j\}$.

Overview

Missing Data

Missing data in longitudinal studies is a very common phenomenon. Data is missing if a measurement that was intended to be taken is not taken, or not available for another reason. An important question always is why the measurements are missing. For example

- The lab assistant accidentally destroyed the sample to be measured. We probably would not be worried and would simply analyze the available measurements.
- The values are missing because they were below the limit of detection (censored). In this case, leaving out the missing values could mask an important effect.
- The values are missing because the subjects did not show up for their scheduled visits. In this case, we might worry about the tendency of sicker subjects with higher/lower values to have missing data.

Missing Data

Missing data patterns:

- Dropout / loss-to-follow-up: Whenever Y_{ij} is missing, so are Y_{ik} for all $k \geq j$.
- Intermittent missing values

Missing data mechanisms: (Rubin, Biometrika, 1976)
Let $\mathbf{Y}$ indicate the complete data, with $\mathbf{Y}_{obs}$ the observed part and $\mathbf{Y}_{mis}$ the missing part of $\mathbf{Y}$. Define the missing data indicator $R_{ij} = 1$ if Y_{ij} is missing, and $= 0$ otherwise. (Note: sometimes $\mathbf{R}$ is defined the other way around.)

- Missing completely at random (MCAR): $p(\mathbf{R} \mid \mathbf{Y}) = p(\mathbf{R})$, so that the observed data are a completely random sample of the complete data
- Missing at random (MAR): $p(\mathbf{R} \mid \mathbf{Y}) = p(\mathbf{R} \mid \mathbf{Y}_{obs})$, so that the missing data mechanism does not depend on the actual missing values
- Not missing at random (NMAR): $p(\mathbf{R} \mid \mathbf{Y})$ depends on $\mathbf{Y}_{mis}$, so that whether or not an observation is observed depends on the quantities that we were not able to observe. (Also called informative missingness.)

Missing Data

Intermittent:

- Censoring: e.g. all values below a threshold (limit of detection, for example) are missing. In this case, the EM algorithm is a possibility (Dempster, Laird & Rubin, JRSS-B, 1977).
- Other cases: For intermittent missing values, the reason is often known, as subjects remain in the study. One can then find out whether an MCAR or MAR assumption is reasonable.

Dropouts:

For dropout, we usually have to suspect a relation between the dropout and the measurement process. Examples:

- Ethical considerations require a patient to be withdrawn from the study if their condition is not adequately controlled by the treatment they are on.
  $\longrightarrow$ MAR
- Patients stop showing up for visits because they are feeling too sick, which would be reflected in their measured Y value if it could be obtained.
  $\longrightarrow$ NMAR

Missing Data

Simple commonly used strategies to deal with dropout

1. Last observation carried forward: In the simplest case, if y_{ij} is the last observed value, y_{ik} is set to y_{ij} for all $k \geq j$. A refinement is to fit a time-trend and to extrapolate that.
2. Complete case analysis: Non-completers are completely deleted. This is wasteful of data and has the potential to introduce bias.

Neither can be recommended in general.

Testing for MCAR

There are strategies to test for complete randomness of dropout by comparing at each time point the history of y_{ij} values of people "about to drop out" and those not dropping out. See e.g. Diggle, Biometrics, 1989.

Likelihood-Based Inference and Missing Data

For likelihood-based inference, it is most important to distinguish between MCAR/MAR on the one hand, and NMAR on the other hand.
The joint probability density function of $(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{R})$ is

$$f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}, \boldsymbol{r}) = f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}) f(\boldsymbol{r}|\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}).$$

The joint pdf of the observable data then is

$$\begin{aligned} f(\boldsymbol{y}_{obs}, \boldsymbol{r}) &= \int f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}) f(\boldsymbol{r}|\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}) d\boldsymbol{y}_{mis} \\ &\overset{MAR}{=} \int f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}) d\boldsymbol{y}_{mis} f(\boldsymbol{r}|\boldsymbol{y}_{obs}) \\ &= f(\boldsymbol{y}_{obs}) f(\boldsymbol{r}|\boldsymbol{y}_{obs}). \end{aligned}$$

The log-likelihood then is

$$\log \mathcal{L} = \log f(\boldsymbol{y}_{obs}) + \log f(\boldsymbol{r}|\boldsymbol{y}_{obs}).$$

Likelihood-Based Inference and Missing Data

The log-likelihood is

$$\log \mathcal{L} = \log f(\boldsymbol{y}_{obs}) + \log f(\boldsymbol{r}|\boldsymbol{y}_{obs}).$$

It is maximized by maximizing the two terms separately. Since the second term contains no information about $[\boldsymbol{Y}_{obs}]$, we can ignore it for inference on $[\boldsymbol{Y}_{obs}]$. Thus, MCAR/MAR are sometimes jointly referred to as ignorable missingness.

However,

- "ignorability" depends on the likelihood being the basis for inference. GEE is only valid under the stronger assumption of MCAR.
- if $\log f(\boldsymbol{y}_{obs})$ and $\log f(\boldsymbol{r}|\boldsymbol{y}_{obs})$ actually share parameters, ignoring $\log f(\boldsymbol{r}|\boldsymbol{y}_{obs})$ will result in a loss of efficiency.
- this approach assumes that the distribution $[\boldsymbol{Y}_{obs}]$ is the target of inference. Example: A clinical trial for treatment of a potentially life-threatening disease. Missingness is a result of patients' death. It might be more meaningful to draw inference about the distribution of the survival time and the conditional distribution of $\boldsymbol{Y}_{obs}$ given survival, than the unconditional distribution $[\boldsymbol{Y}_{obs}]$.

GEE and Missing Data

- Advantage of GEE is that if interest is in the mean, GEE provides consistent inference if the mean model is correctly specified, even for misspecified variance matrix, and without distributional assumption.
- However, GEE assumes MCAR observations, otherwise we lose consistency.
- The basic GEE was

$$S_\beta(\beta, \alpha) = \sum_{i=1}^{I} \left(\frac{\partial \mu_i}{\partial \beta} \right)' W_i^{-1} (Y_i - \mu_i) = 0.$$

- Robins et al. (JASA, 1995) propose a weighted GEE for MAR, where they assume that an observed measurement is representative of missing observations from subjects with the same history $y_{i1}, \ldots, y_{ij-1}$ and covariate values. Measurements with small probabilities are then upweighted (inverse probability weighting).

GEE and Missing Data

- The weighted GEE then is

$$S_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{I} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)' \boldsymbol{W}_i^{-1} \boldsymbol{P}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0},$$

  where $\boldsymbol{P}_i$ is a diagonal matrix containing the probabilities p_{ij} that subject i has not dropped out by time t_{ij}, given the history $y_{ij-1}, \ldots, y_{i1}$ and covariates.

- This method requires that we can consistently estimate the p_{ij}, and is thus better suited for larger studies. Also, it requires a parametric model for the p_{ij} (while there is sparse information for the dropout process), in a setting where a parametric model for the covariance structure is avoided.

- Scharfstein et al (JASA, 1999) propose an extension to non-ignorable dropout (NMAR), giving a range of inferences based on different parameters for informative dropout processes.

Modeling the Dropout Process

There are also approaches to jointly model the dropout process D and the intended measurements Y^*. Let D be the dropout time, $\boldsymbol{Y}^* = (Y_1^*, \ldots, Y_n^*)$ the intended measurements, and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ the observed measurements (can be NA).

Selection models factorize

$$[\boldsymbol{Y}^*, D] = [\boldsymbol{Y}^*][D|\boldsymbol{Y}^*].$$

Pattern mixture models factorize

$$[\boldsymbol{Y}^*, D] = [D][\boldsymbol{Y}^*|D].$$

Random effects models assume an underlying bivariate random effect $\boldsymbol{U} = (\boldsymbol{U}_1, \boldsymbol{U}_2)$ and factorize

$$[\boldsymbol{Y}^*, D, \boldsymbol{U}] = [\boldsymbol{Y}^*|\boldsymbol{U}_1][D|\boldsymbol{U}_2][\boldsymbol{U}].$$

Modeling the Dropout Process

Models for informative dropout have identifiability issues, and randomness (MAR) of dropout cannot be verified from the data. Thus, analyses of longitudinal data with dropout has always to proceed with caution. However, addressing the dropout problem is often still more reasonable than to simply assume ignorability of dropout. Read more e.g. in Diggle et al (2002), or Molenberghs & Verbeke (2005).

Overview

Joint Modeling of Longitudinal and Survival Data

In many longitudinal studies, we have the following data structure

- I observational units (subjects, animals, ...).
- n_i observations on the ith unit, $i = 1, \ldots, I$.
- The observations Y_{ij} are taken at time points t_{ij}, $t_{i1} < \cdots < t_{in_i}$.
- We also observe times F_i to an event (e.g. death).

Joint modeling:

$$[Y, F | \boldsymbol{x}, \theta]$$

Why joint modeling?

- Interest could be in the time-to-event variable F, but for heavy censoring, joint modeling could improve inference about marginal distribution $[F]$.
  Example: Randomized clinical trial (RCT) for a treatment for liver cirrhosis. F = time of death, Y = prothombin measurements, $\approx$ 30% survival to 96 months

- Interest could be in Y, but informative dropout or death could lead to misleading conclusions.
  Example: RCT for three schizophrenia treatments, interest in reducing PANSS score Y, but drop-out rate is treatment-dependent.

- Interest could be in the relationship between Y and F.
  Example: RCT for two heart surgeries. Y = left-ventricular-mass-index (LVMI), F = time to death. Long-term build-up of LVMI may increase hazard for fatal heart-attack $\Rightarrow$ interested in modeling relationship between survival and subject-level LVMI

Joint Models for Longitudinal and Survival Data

One approach to jointly model longitudinal and survival data:
Random effects model:

- mixed model for Y
- proportional hazards model for $\lambda(t)$ with time-dependent frailty (random effect) for time-to-event F
- submodels linked through shared random effects.

For example (Henderson, Diggle & Dobson, Biostatistics, 2000)

$$\begin{aligned} Y_j &= \boldsymbol{x}_1(t_j)'\beta_1 + W_1(t_j) + Z_t, \quad Z_t \overset{iid}{\sim} N(0, \sigma^2) \\ \lambda(t) &= \lambda_0(t)\exp\{\boldsymbol{x}_2(t)'\beta_2 + W_2(t)\} \\ W_1(t) &= z_1(t)'U_1 + V_1(t) \quad \text{and} \quad W_2(t) = z_1(t)'U_1 + V_2(t), \end{aligned}$$

where U_1 and U_2 are jointly multivariate Gaussian random effects, and the bivariate Gaussian latent stochastic process $W(t)$ has non-zero cross-covariance $\gamma_{12}(u) = \text{Cov}\{W_1(t), W_2(t-u)\}$.

Multivariate Longitudinal Data

Often the outcome Y measured in a longitudinal study is multivariate. Examples: several blood markers or symptom scores which reflect improvement in health, vector of voxel intensities in neuroimaging.

One possible way to model a multivariate outcome (here linear model):

$$Y_{ijk} = \boldsymbol{x}'_{ijk}\boldsymbol{\beta}_k + \varepsilon_{ijk}, \quad k = 1, \ldots, K,$$

where Y_{ijk} is the kth outcome for the ith subject at time t_{ij}.
We would then need to specify a model for

- $\text{Var}(\boldsymbol{\varepsilon}_{ij})$, the variance matrix for the K outcomes for subject i at time t_{ij}
- $\text{Cov}(\boldsymbol{\varepsilon}_{ij}, \boldsymbol{\varepsilon}_{ij'})$, the covariance matrix between times t_{ij} and $t_{ij'}$ for subject i

to obtain the $Kn_i \times Kn_i$ variance matrix $\boldsymbol{V}_i$ for $\boldsymbol{\varepsilon}_i$. Inference can then proceed as it would in the univariate longitudinal case.

Robust standard errors can account for potential misspecification of $\boldsymbol{V}$. The challenge still is to find good approximations for $\boldsymbol{V}_i$, however, so that gains in efficiency from joint modeling of the K outcomes are not lost.

Multivariate Longitudinal Data

Another strategy is to use a hierarchical mixed model with random effects for both subjects and outcome items. We could assume, for example,

$$\begin{aligned} Y_{ijk} &= \boldsymbol{x}'_{ijk}\boldsymbol{\beta}_{ik} + \varepsilon_{ijk} \\ \beta_{ik} &= \bar{\beta} + \delta_k + b_i \end{aligned}$$

$$\delta_k \overset{iid}{\sim} N(\mathbf{0}, \boldsymbol{D_\delta}) \quad \text{independent of} \quad b_i \overset{iid}{\sim} N(\mathbf{0}, \boldsymbol{D_b}).$$

This reduces the number of parameters necessary to specify $\boldsymbol{V}$ dramatically and borrows strength across items. We can proceed with REML-based inference in this linear mixed model, and can use the BLUPs for $\bar{\beta} + \delta_k$ to look at the (population average) item specific coefficients.

Sample Size Calculations

To determine the required sample size of a longitudinal study, we need

- Type I error rate α
- Smallest meaningful difference to be detected d
- Power P
- Measurement variation σ^2
- Number of repeated observations per subject n
- Correlation among repeated observations

Sample size formulas can then be derived, see e.g. Diggle et al (2002).

Sample Size Calculations

For example, consider continuous responses and a comparison of two groups,

$$Y_{ij} = \beta_{0A} + \beta_{1A} x_{ij} + \varepsilon_{ij}, \;\; j = 1, \ldots, n, \;\; i = 1, \ldots, I,$$

and analogous with β_{0B} and β_{1B} for group B, where $x_{ij} = x_j$ for all i, $Var(\varepsilon_{ij}) = \sigma^2$ and $\text{Corr}(Y_{ij}, Y_{ik}) = \rho$, $j \neq k$.

Then, for fixed known n, the number of subjects in each group I needed is

$$I = \frac{2(z_\alpha + z_{1-P})^2 \sigma^2 (1-\rho)}{n s_x^2 d^2},$$

with $d = \beta_{1B} - \beta_{1A}$, z_p the pth quantile of the standard normal distribution, and $s_x^2 = \frac{1}{n}\sum_j (x_j - \bar{x})^2$ the within-subject variance of the x_j.