Missing Data

Biostatistics 653

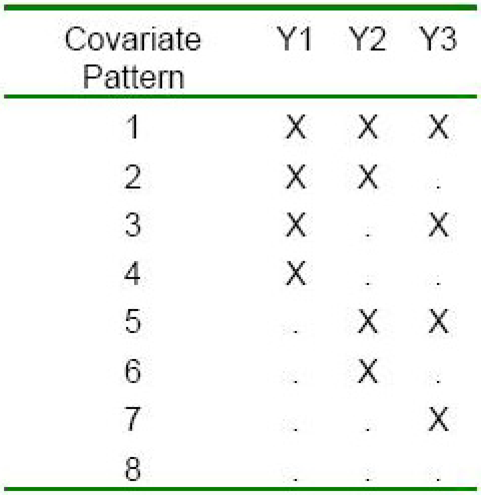
Applied Statistics III: Longitudinal Data Analysis

Pattern of Missing Data

- Missing data patterns may be monotone or non-monotone. In a monotone missing data pattern, observations missing on one variable are a subset of those missing on another variable. That is, missingness is nested. One example of monotone missing data is study dropout. If a subject drops out of a study at time t, then their observations will also be missing at times t + 1, t + 2, and so forth. When missing data follow such a pattern, the group of responses is never larger at a later follow-up time than it is at an earlier time. Missing data are non-monotone when missingness is not nested in this manner, or is intermittent.

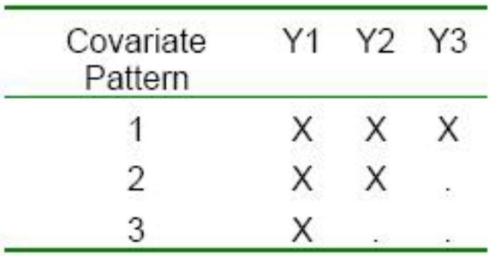
Pattern of Missing Data

Non-monotone



Covariate Pattern	Y1	Y2	Y3
1	X	X	X
2	X	X	.
3	X	.	X
4	X	.	.
5	.	X	X
6	.	X	.
7	.	.	X
8	.	.	.

Monotone



Covariate Pattern	Y1	Y2	Y3
1	X	X	X
2	X	X	.
3	X	.	.

Missing Data Mechanism

- Let Y indicate the complete data, with Y_{obs} representing the observed part of Y and Y_{mis} representing the missing part of Y. Similarly, dene the missing data indicator R_{ij}, which takes value 1 if R_{ij} is observed and takes value 0 otherwise. (We note that sometimes R is given the opposite definition.) So the observed data are ($Y_{i,obs}, R_i$), i = 1, …, N.

Missing Data Mechanism

The missing data mechanism concerns the distribution of R given Y.

- Missing completely at random (MCAR): p(R | Y) = p(R), so that the observed data are a completely random sample of the complete data, and does not depend on outcomes and covariates.
- Missing at random (MAR): p(R | Y) = p(R | Y_{obs}), so that the missing data mechanism does not depend on the actual missing values, but depend on the observed outcomes and covariates.
- Not missing at random (NMAR): p(R | Y) depends on Y_{mis}, so that whether or not an observation is observed depends on the quantities that you were not able to observe (unobserved outcomes or unobserved covariates).

Examples

- MCAR: patients miss a scheduled visit because of bad weather or car out of service.
- MAR: older people may have a higher chance of dropping out of a study (suppose age is observable).
- NMAR: subjects drop out because they have poor treatment outcomes or they die.

Missing Data Mechanism

- We note that when conducting studies, it is very important to do everything possible to collect data on the reasons for missing values or dropouts, so that the investigator can determine the missing data mechanism so that the decision can be made whether it should be accounted for in the analysis, and analysis can properly account for the missing data mechanism if necessary.

Modeling Missing Data

Likelihood Function:

$$P(Y_{obs}, R) = \int P(Y_{obs}, Y_{mis}, R) dY_{mis}$$

Partition the likelihood with two classes of models:

1. Selection model (Diggle and Kenward, 1994)

$$P(Y_{obs}, Y_{mis}, R) = P(Y_{obs}, Y_{mis}) P(R | Y_{obs}, Y_{mis})$$

2. Pattern Mixture Model (Little, 1993, 1994)

$$P(Y_{obs}, Y_{mis}, R) = P(R) P(Y_{obs}, Y_{mis} | R)$$

Selection Models

- The selection model models the marginal distribution of the outcome and models the conditional distribution of the dropout on the outcome.
- Selection models are nice because they directly model $\mathrm{P}(Y|X, \beta)$, the target of our inference. The regression coefficients from Y model hence have attractive population interpretation in practice.
- However, they can be computationally intractable (often involve difficult integrals and need complex versions of EM).
- Results may depend heavily on modeling assumptions, and identifiability can be difficult to characterize. NOTE: complete case analysis assumptions are also usually unverifiable.
- Selection models can be fit easily in specialized software (like WinBUGS) though not necessarily in other software.

Pattern-Mixture Models

- The pattern mixture model models the marginal distribution of the dropout and the conditional distributions of the outcome on each dropout pattern. Hence the interpretation of regression coefficients is conditional on the dropout pattern and has less attractive interpretation in practice.
- However, pattern mixture models help researchers better understand the missing data mechanisms and assumptions.
- The distribution of outcomes given nonresponse patterns is not completely identifiable, because certain outcomes are not observed. These models are also heavily dependent on modeling assumptions.

Inference in the Selection Model

Likelihood Function:

$$P(Y_{obs}, R | X, \theta, \psi)$$
$$= \int P(Y_{obs}, Y_{mis} | X, \theta) P(R | Y_{obs}, Y_{mis}, X, \psi) dY_{mis}$$

1. Likelihood-based inference: specify the full likelihoods $P(Y_{obs}, Y_{mis} | X, \theta)$, e.g. using GLMMs, and $P(R | Y_{obs}, Y_{mis}, X, \psi)$, e.g. using logistic regression. SAS PROC NLMIXED can be used for model fitting.
2. Estimating equation based inference: only specify the first two moments of $P(Y_{obs}, Y_{mis} | X, \theta)$, and specify $P(R | Y_{obs}, Y_{mis}, X, \psi)$, e.g. using logistic regression. Estimate model parameters using modified GEEs, e.g. Inverse-probability weighted (IPW) GEEs and augmented inverse probability weighted (AIPW) GEEs.

Inference in the Selection Model

- Under MAR or MCAR, likelihood-based inference is valid using only the observed data, i.e. SAS PROC MIXED and PROC NLMIXED are valid using all the observed data.

- Standard GEEs require MCAR to hold for regression coefficient estimators to be consistent.

- IPW GEEs and AIPW GEEs are valid under MAR, and AIPW improves robustness and efficiency of the IPW estimators. SAS PROC GENMOD can be used to calculate the IPW estimators.

MAR and MCAR in Likelihood Inference

- MAR and MCAR are ignorable for likelihood-based inference. This is, likelihood-based inference using only the observed data is valid under MAR and MCAR.

- Rationale:

$$P(Y_{obs}, R|\theta, \psi) = \int P(Y_{obs}, Y_{mis}|\theta)P(R|Y_{obs}, Y_{mis}, \psi)dY_{mis}$$

Under MAR

$$P(R|Y_{obs}, Y_{mis}, \psi) = P(R|Y_{obs}, \psi)$$

We have

$$P(Y_{obs}, R|\theta, \psi) = \int P(Y_{obs}, Y_{mis}|\theta)P(R|Y_{obs}, \psi)dY_{mis}$$

$$= P(Y_{obs}|\theta)P(R|Y_{obs}, \psi)$$

MAR and MCAR in Likelihood Inference

$$P(Y_{obs}, R|\theta, \psi) = P(Y_{obs}|\theta)P(R|Y_{obs}, \psi)$$

- If θ, ψ do not overlap, then we can ignore $P(R|Y_{obs}, \psi)$ and maximize the observed Y data likelihood $P(Y_{obs}|\theta)$ for inference about θ. Thus, MCAR and MAR are referred to as ignorable if the likelihood-based methods are used for inference.
- If θ, ψ overlap, then using $P(Y_{obs}|\theta)$ is still fine, but will lead to loss of efficiency.

GEE under MCAR

Under MCAR, GEE based methods are valid, because

$$E\left(Y_{i,obs}\right) = E\left(Y_{i,obs} \middle| R_i\right) = \mu_{i,obs}$$

In general, we can write

$$Y_{i,obs} = M_i Y_i$$

where M_i is a matrix of indicators for the observed responses.

In general, M_i is a random matrix (depending on R_i), but if R_i depends only on X_i, then

$$E\left(Y_{i,obs}\right) = E\left(Y_{i,obs} \middle| R_i\right) = M_i \mu_i = \mu_{i,obs}$$

and

$$V\left(Y_{i,obs}\right) = V\left(Y_{i,obs} \middle| R_i\right) = M_i \Sigma M_i^T = \Sigma_{i,obs}$$

Under this assumption, we get valid, consistent estimates for GEE's.

GEE under MAR

What can go wrong with GEE estimators in the MAR case?

Suppose that

$$Y_i \sim N_2\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{12} & \sigma_{22}\end{pmatrix}\right)$$

Suppose that $P(R_{i1} = 1) = 1$ so that Y_{i1} is always observed but that

$$P(R_{i2} = 1 | Y_{i1}, Y_{i2}) = \begin{cases} 1, & Y_{i1} < c \\ 0, & Y_{i1} \geq c \end{cases}$$

GEE under MAR

- In this case, Y_{i2} is observed when Y_{i1} is small (< c) but not when it is large. This could occur by design if a study plans to follow subjects further based on their previous outcomes (for example, in a study of weight loss after pregnancy, investigators might stop following women once they have lost a fixed amount of weight c).

- Here, missingness is MAR (does not depend on Y_{i2}) but is not ignorable, because

$$P(R_{i2} = 1) = 1 * P(Y_{i1} < c) + 0 * P(Y_{i1} \geq c) = P(Y_{i1} < c)$$

which depends on μ and Σ.

GEE under MAR

- Consider the complete case (CC) estimate of μ_2, given by

$$\hat{\mu}_2 = \frac{\sum R_{i2} Y_{i2}}{\sum R_{i2}}$$

$$E(\hat{\mu}_2 | R) = \frac{\sum R_{i2} E(Y_{i2} | R_{i2})}{\sum R_{i2}} = E(Y_{i2} | R_{i2})$$

GEE under MAR

- Now, using the fact that the mean of the conditional distribution of $Y_2|Y_1$ is $\mu_2 + \frac{\sigma_{12}}{\sigma_{11}}(Y_1 - \mu_1)$ one can show that

$$E(Y_{i2}|R_{i2} = 1) = E(Y_{i2}|Y_{i1} < c)$$

$$= \mu_2 + \frac{\sigma_{12}}{\sigma_{11}}(E(Y_{i1}|Y_{i1} < c) - \mu_1)$$

Truncated Normal Distribution

- Suppose that $y \sim N(\mu, \sigma^2)$. Then

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2)$$

$$-\infty < y < \infty$$

- Now, suppose that we condition on $y \in A = [a_1, a_2]$, where $-\infty < a_1 < a_2 < \infty$. The probability of y falling into this interval is $\Phi\left(\frac{a_2-\mu}{\sigma}\right) - \Phi(\frac{a_1-\mu}{\sigma})$. For example, for $y \sim N(0,1)$ and $a_1 = -1.96, a_2 = 1.96$, then
$$P(Y \in [-1.96, 1.96]) = \Phi(1.96) - \Phi(-1.96) = 0.95$$

Truncated Normal Distribution

- You can show the conditional density of $y, a_1 \leq y \leq a_2$, is given by

$$\frac{\frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2)}{\Phi\left(\frac{a_2-\mu}{\sigma}\right)-\Phi(\frac{a_1-\mu}{\sigma})}$$

- Using the MGF of the normal distribution, you can also show that $E(y|y \in A)$

$$= \mu - \frac{\frac{1}{\sigma\sqrt{2\pi}}\left(\exp\left(-\frac{1}{2}\left(\frac{a_2-\mu}{\sigma}\right)^2\right)-\exp\left(-\frac{1}{2}\left(\frac{a_1-\mu}{\sigma}\right)^2\right)\right)}{\Phi\left(\frac{a_2-\mu}{\sigma}\right)-\Phi(\frac{a_1-\mu}{\sigma})}$$

Truncated Normal Distribution

- In our case, $a_2 = c$ and $a_1 = -\infty$, so

$$E(Y_{i1}|Y_{i1} < c) = \mu_1 - \frac{\frac{\sqrt{\sigma_{11}}}{\sqrt{2\pi}}\exp(-\frac{1}{2}\left(\frac{c-\mu_1}{\sqrt{\sigma_{11}}}\right)^2)}{\Phi\left(\frac{c-\mu_1}{\sqrt{\sigma_{11}}}\right)}$$

Truncated Normal Distribution

- Thus

$$E(Y_{i2}|Y_{i1} < c) = \mu_2 + \frac{\sigma_{12}}{\sigma_{11}}\left(E(Y_{i1}|Y_{i1} < c) - \mu_1\ \right)$$

$$= \mu_2 + \frac{\sigma_{12}}{\sigma_{11}}\left(\mu_1 - \frac{\frac{\sqrt{\sigma_{11}}}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{c-\mu_1}{\sqrt{\sigma_{11}}}\right)^2\right)}{\Phi\left(\frac{c-\mu_1}{\sqrt{\sigma_{11}}}\right)} - \mu_1\right)$$

$$= \mu_2 + \frac{\sigma_{12}}{\sqrt{\sigma_{11}}}\left(\frac{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{c-\mu_1}{\sqrt{\sigma_{11}}}\right)^2\right)}{\Phi\left(\frac{c-\mu_1}{\sqrt{\sigma_{11}}}\right)}\right) \neq \mu_2$$

so that the complete case estimate of μ_2 is biased unless $\sigma_{12} = 0.$

Truncated Normal Distribution

- Also, the complete case estimate of μ_1,

$$\hat{\mu}_1 = \frac{\sum R_{i2} Y_{i1}}{\sum R_{i2}}$$

is also biased, as

$$E(\hat{\mu}_1) = E(Y_{i1} | Y_{i1} < c) = \mu_1 - \frac{\frac{\sqrt{\sigma_{11}}}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{c - \mu_1}{\sqrt{\sigma_{11}}}\right)^2\right)}{\Phi\left(\frac{c - \mu_1}{\sqrt{\sigma_{11}}}\right)} \neq \mu_1$$

NMAR: Non-ignorable Nonresponse

- If $P(R_i|Y_i, X_i, \psi)$ is a function of $Y_{i,mis}$, then the missing data mechanism is always non-ignorable (but this is not the only setting in which we can have non-ignorable missing data). In this setting, for modeling we need to specify the joint distribution
$$f(Y_i, R_i|X_i, \beta, \psi)$$
- for inference. This can be problematic because it is often hard to estimate a missing data mechanism that depends on values that are not even observed. Results in this case often depend strongly on the assumed model, and sensitivity analyses are a useful tool for determining the consequences if your assumed model is not correct.

Non-ignorable Nonresponse

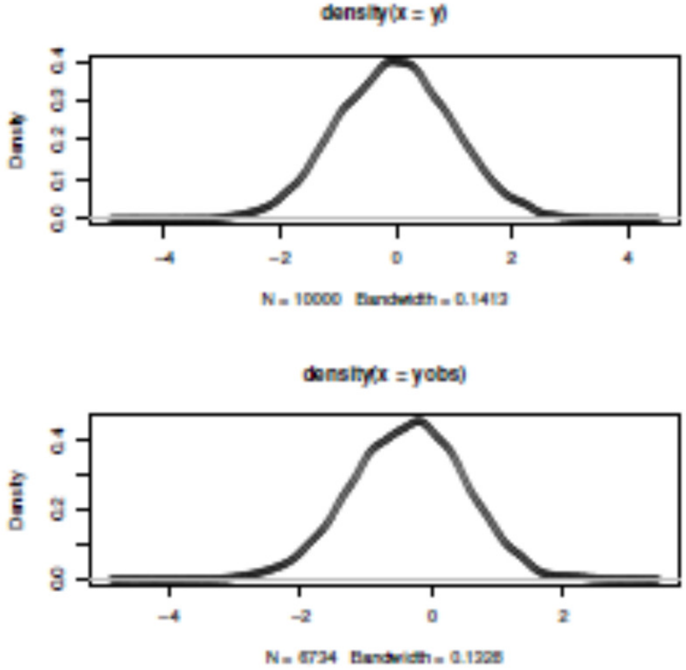
- Example 1: Suppose you are estimating mean income as a function of various covariates. Missingness may be more likely for people with very high or very low incomes, in addition to a subset of people who consider their income “none of your business” regardless of their income level.
- Example 2: Consider a longitudinal clinical trial with interest in modeling health-related quality of life, which is measured every three months by self-report on a detailed multiple-item questionnaire (items might include ability to carry out everyday activities, outlook, daily pain, etc.). There may be a lot of missing data, even on subjects who do not drop out. For example, if subjects who are sicker, or who are in more pain, do not respond, then we may have non-ignorable nonresponse. In particular, nonresponse at time j is likely to be related to quality of life at time j, even conditional on quality of life at times 1, …, j-1

Non-ignorable Nonresponse

- With monotone missingness, if P(dropout after time j) depends only on covariates and responses through time j, then dropout may be ignorable. However, if the probability depends on the unobserved Y_{j+1}, then dropout is non-ignorable.
- Often, the term non-informative is used to mean ignorable, and the term informative is used to mean non-ignorable when talking of dropouts.

Non-ignorable Nonresponse

- Now, consider a univariate case and suppose that $Y_i \sim N(0,1)$ and that $logit\big(P(R_i = 1|Y_i)\big) = -1 + 1.5Y_i$. (This gives us a probability of missing data around 33% on average.) In this case, the observed distribution of Y is skewed, as seen in the figure.



Non-ignorable Nonresponse

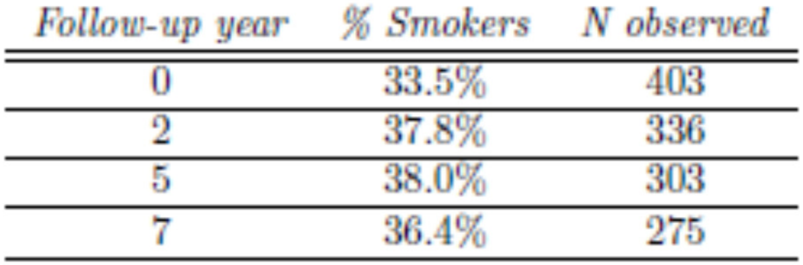
- We note that we will not be able to distinguish between non-ignorable nonresponse for normal Y and MCAR with non-normal Y.
- In addition, it is hard to estimate $P(R_i = 1|Y_i)$ because you only observe Y_i when $R_i = 1$. The distribution you assume for Y_i is going to play a big role in the inferences you obtain.

Example: Smoking Data

- A longitudinal study planned to collect smoking status data on 403 young adults at baseline and follow-up 2, 5, and 7 years after baseline. Investigators wished to estimate the pattern of smoking prevalence over time in this cohort.
- The response variable, Y_{ij}, takes value 1 if young adult i is a smoker at time j (j = 0, 2, 5, 7) and takes value 0 otherwise. Predictors of interest include age, x_{i1}, at baseline (centered at 20 years) and highest educational attainment (a categorical variable with college graduate as the referent and indicators of less than high school, x_{i2}, and some college completed, x_{i3}).

Observed Prevalences

- The following observed proportions of smoking at each follow-up time were observed.



Follow-up year	*% Smokers*	*N observed*
0	33.5%	403
2	37.8%	336
5	38.0%	303
7	36.4%	275

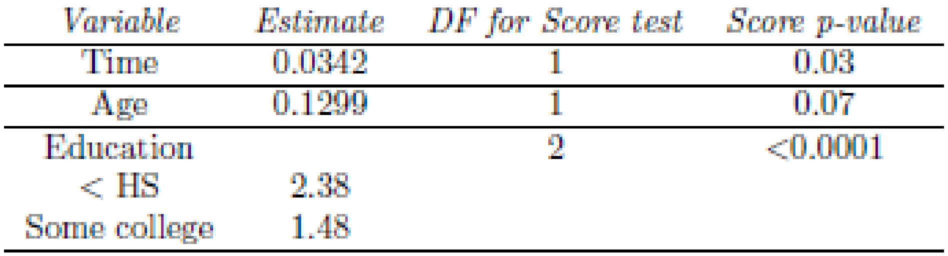
Observed Prevalences

- By the end of follow up, almost 32% of the data are missing. Do we feel comfortable reporting the results of the complete case analysis?
- Consider the following mean model for the smoking data.

$$logit\left(P\left(Y_{ij} = 1\right)\right) = \beta_0 + \beta_1 t_j + \beta_2 x_{i1} + \beta_3 x_{i2} + \beta_4 x_{i3}$$

GEE under MCAR

- We begin by fitting a GEE under the MCAR assumption using an unstructured working correlation structure. We obtain the following parameter estimates.

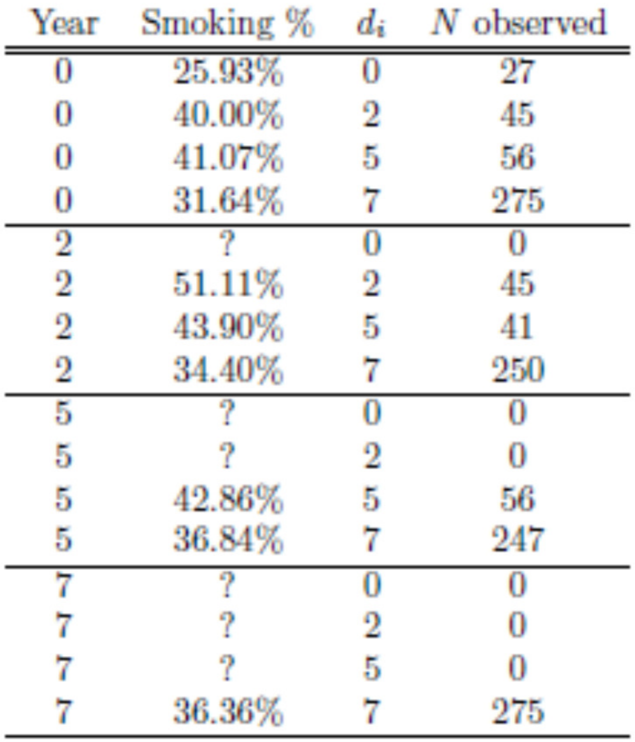


Variable	*Estimate*	*DF for Score test*	*Score p-value*
Time	0.0342	1	0.03
Age	0.1299	1	0.07
Education		2	<0.0001
< HS	2.38		
Some college	1.48		

GEE under MCAR

- Can we believe these results?
- Define four subgroups of young adults, indexed by $d_i = 0{,}2{,}5{,}7$, where d_i indicates the time of the last observation from each youth. These subgroups are defined only in terms of the last measurement time; that is, $d_i = 7$ indicates that the subject was observed at time 7, but tells us nothing about whether the subject was observed at times 2 and 5. However, we know subjects with $d_i = 2$ were dropouts after time 2.

Smoking Prevalences by Dropout Pattern



Year	Smoking %	d_i	N observed
0	25.93%	0	27
0	40.00%	2	45
0	41.07%	5	56
0	31.64%	7	275
2	?	0	0
2	51.11%	2	45
2	43.90%	5	41
2	34.40%	7	250
5	?	0	0
5	?	2	0
5	42.86%	5	56
5	36.84%	7	247
7	?	0	0
7	?	2	0
7	?	5	0
7	36.36%	7	275

Smoking Prevalences by Dropout Pattern

- Based on the table, we see that subjects who were observed only at baseline ($d_i = 0$) have a lower smoking prevalence at baseline than subjects with later follow-up measures. At year 2, however, those subjects who drop out immediately after that measurement had the highest smoking prevalence. Do we think the data are MCAR?

Pattern Mixture GEE

- In order to fit a pattern mixture GEE model, we will first make a couple of assumptions. Given the relatively small number of subjects dropping out at each stage, we will assume the effects of age and education do not depend on the dropout pattern. In addition, we note that we cannot estimate the effect of time within the group of subjects who drop out at time 0; we will combine dropouts at 0 and time 2 for estimation of the time slope.

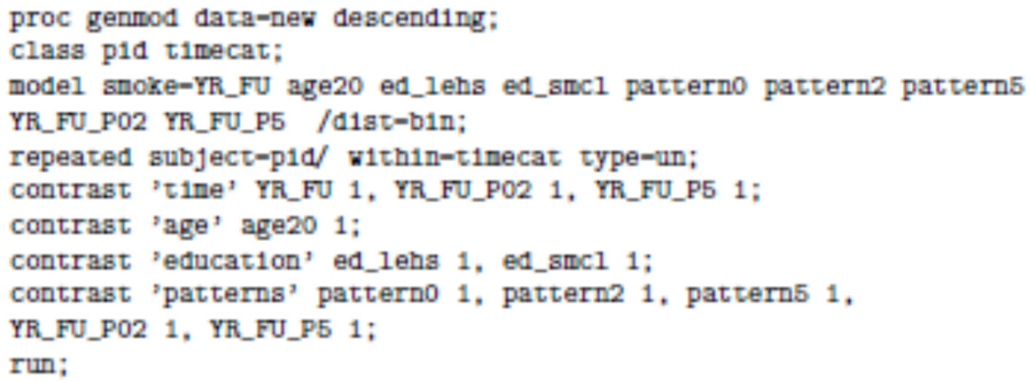
Pattern Mixture GEE

- Then our mean model is given by
$$\begin{aligned} &logit\left(P\left(Y_{ij}=1\right)\right) \\ &= \beta_0+\beta_1 I(d_i=0)+\beta_2 I(d_i=2)+\beta_3 I(d_i=5)+\beta_4 t_j \\ &+ \beta_5 t_j I(d_i=0\ or\ 2)+\beta_6 t_j I(d_i=5)+\beta_7 x_{i1}+\beta_8 x_{i2}+\beta_9 x_{i3} \end{aligned}$$

and again we fit the model using unstructured working correlation.

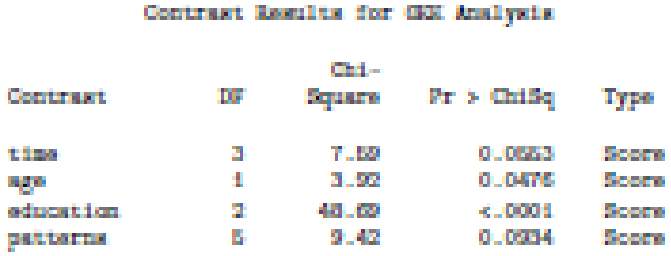
- This model fits separate baseline smoking probabilities for each unique value of d_i. It estimates three slopes for time: one for early dropouts ($\beta_4+\beta_5$), one for dropouts after time 5 ($\beta_4+\beta_6$), and one for those measured at the last follow-up time (β_4). It also assumes the effects of age and education do not depend on dropout time.

SAS Code



```
proc genmod data=new descending;
class pid timecat;
model smoke=YR_FU age20 ed_lehs ed_smcl pattern0 pattern2 pattern5
YR_FU_P02 YR_FU_P5  /dist=bin;
repeated subject=pid/ within=timecat type=un;
contrast ’time’ YR_FU 1, YR_FU_P02 1, YR_FU_P5 1;
contrast ’age’ age20 1;
contrast ’education’ ed_lehs 1, ed_smcl 1;
contrast ’patterns’ pattern0 1, pattern2 1, pattern5 1,
YR_FU_P02 1, YR_FU_P5 1;
run;
```

Results

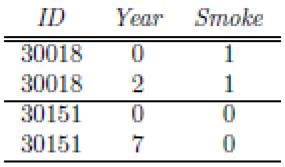


```
                 Contrast Results for GEE Analysis

                                Chi-
Contrast            DF        Square    Pr > ChiSq    Type

time                 3          7.59        0.0553    Score
age                  1          3.92        0.0476    Score
education            2         48.69        <.0001    Score
patterns             5          9.42        0.0934    Score
```

Based on the score test for the pattern mixture terms, we see the pattern terms are not highly significant. However, the score test for the time terms indicates that there are no longer significant changes in prevalences over time once we have conducted the pattern mixture analysis.

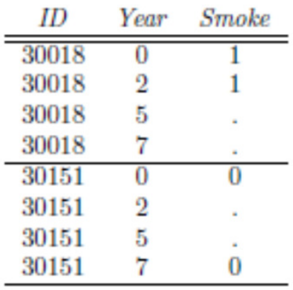
Multiple Imputation

- In order to carry out multiple imputation in SAS, missing data points should be represented by the '.' symbol in the dataset. For the smoking data, the smoking outcomes are the only missing data, and they are simply omitted from the dataset (see following slide). We need to insert the '.' measures for the missing data before we can proceed with multiple imputation.



ID	*Year*	*Smoke*
30018	0	1
30018	2	1
30151	0	0
30151	7	0

current format



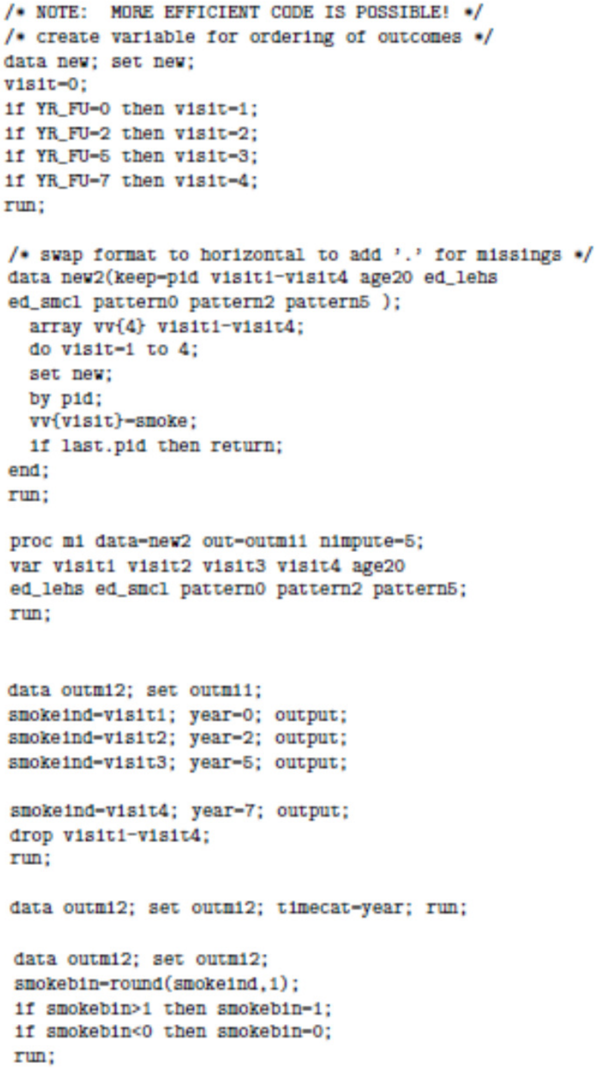
ID	*Year*	*Smoke*
30018	0	1
30018	2	1
30018	5	.
30018	7	.
30151	0	0
30151	2	.
30151	5	.
30151	7	0

desired format

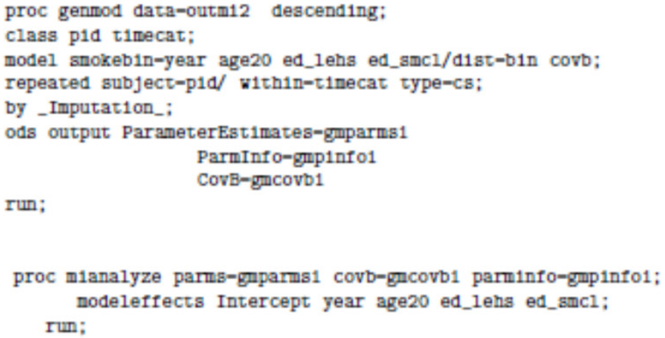
Multiple Imputation

- Because our missing data are not monotone, our only option for imputation in SAS is to impute the missing data from the multivariate normal distribution.
- That is, we will impute values of 0.2, 1.36, etc. for the smoking indicator. If smoking were an exposure, we would get better results by leaving these values 'as-is'; however, we cannot analyze them in PROC GENMOD unless they are rounded.

SAS Code

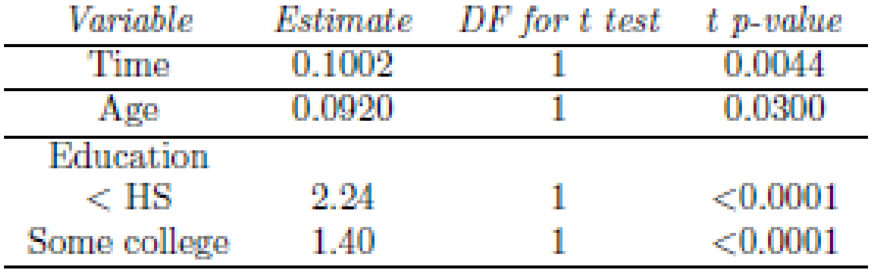


```
/* NOTE:  MORE EFFICIENT CODE IS POSSIBLE! */
/* create variable for ordering of outcomes */
data new; set new;
visit=0;
if YR_FU=0 then visit=1;
if YR_FU=2 then visit=2;
if YR_FU=5 then visit=3;
if YR_FU=7 then visit=4;
run;

/* swap format to horizontal to add '.' for missings */
data new2(keep=pid visit1-visit4 age20 ed_lehs
ed_smcl pattern0 pattern2 pattern5 );
  array vv{4} visit1-visit4;
  do visit=1 to 4;
  set new;
  by pid;
  vv{visit}=smoke;
  if last.pid then return;
end;
run;

proc mi data=new2 out=outmi1 nimpute=5;
var visit1 visit2 visit3 visit4 age20
ed_lehs ed_smcl pattern0 pattern2 pattern5;
run;


data outmi2; set outmi1;
smokeind=visit1; year=0; output;
smokeind=visit2; year=2; output;
smokeind=visit3; year=5; output;

smokeind=visit4; year=7; output;
drop visit1-visit4;
run;

data outmi2; set outmi2; timecat=year; run;

data outmi2; set outmi2;
smokebin=round(smokeind,1);
if smokebin>1 then smokebin=1;
if smokebin<0 then smokebin=0;
run;
```



```
proc genmod data=outmi2  descending;
class pid timecat;
model smokebin=year age20 ed_lehs ed_smcl/dist=bin covb;
repeated subject=pid/ within=timecat type=cs;
by _Imputation_;
ods output ParameterEstimates=gmparms1
                ParmInfo=gmpinfo1
                CovB=gmcovb1
run;


 proc mianalyze parms=gmparms1 covb=gmcovb1 parminfo=gmpinfo1;
      modeleffects Intercept year age20 ed_lehs ed_smcl;
   run;
```

Results



Variable	*Estimate*	*DF for t test*	*t p-value*
Time	0.1002	1	0.0044
Age	0.0920	1	0.0300
Education			
< HS	2.24	1	<0.0001
Some college	1.40	1	<0.0001

Multiple Imputation

- We also note that SAS has two experimental procedures, PROC MI and PROC MIANALYZE, available for multiple imputation. These procedures are currently somewhat restrictive (especially if you have non-normal longitudinal data subject to missingness) but have the potential to be quite useful in the future. For a MAR covariate that can assumed to follow a normal distribution, for example, this could be very useful.

Multiple Imputation

- Create multiple "complete" datasets by filling in values for the missing data
- Analyze each filled-in dataset as if it were the complete data
- Combine separate inferences into one overall result

Rubin (1978), other papers, and book.

Multiple Imputation

- Obtain $\hat{\beta}^{(m)}$ for the m’th imputed dataset, m=1,…,M
- Do this M times to construct M "complete" datasets, and then use the M datasets to estimate variability
- Parameter estimate $\hat{\beta} = \sum_{m=1}^{M} \frac{\widehat{\beta}^{(m)}}{M}$
- Variance estimate straightforward

Multiple Imputation

- Except in special cases (for example, monotone missing data), SAS imputes missing data from a multivariate normal distribution. In particular, SAS assumes that the missing data mechanism is ignorable, and imputes the missing data using a MCMC (Markov chain Monte Carlo) scheme as follows. To obtain each imputed data set, SAS does the following.
- *Imputation step*: Starting with a given prior mean vector μ and prior covariance matrix Σ, the imputation step draws values for the missing data from the conditional distribution of the missing data given all observed data. The joint distribution of all data is assumed to be multivariate normal (regardless of whether observed and missing variables are counts, nominal, continuous, etc.).

Multiple Imputation

- *Bayesian estimation step*: Using the augmented data, calculate the updated posterior distributions of μ and Σ
- *Posterior step*: Draw a new μ and Σ from the posterior distributions
- Iterate until convergence

Multiple Imputation

- $\hat{V}^{(m)}$: Variance estimate of $\hat{\beta}$ from m'th imputed dataset
- Define $\widehat{\mathrm{V}}$ to be the average variance estimate. That is, $\hat{V} = \frac{1}{M}\sum_{m=1}^{M} \hat{V}^{(m)}$. This is the average "within" imputation variance.
- Define $\hat{B} = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\beta}^{(m)} - \hat{\beta}\right)\left(\hat{\beta}^{(m)} - \hat{\beta}\right)^T$. This is the "between" imputation variance.
- The variance estimate is given by

$$\hat{V}_{MI} = \hat{V} + \left(1 + \frac{1}{M}\right)\hat{B}$$

Multiple Imputation

- Depends on correct specification of $p(Y_{mis}|Y_{obs}, X, \beta)$.
- Can improve specification through iterative procedure (can become very sophisticated and similar to EM).
- Allows better estimate of variability than simple imputation with slightly more effort.
- Can be extended to handle non-ignorable nonresponse (but not as implemented in SAS currently)

Summary

- Models for nonignorable nonresponse are typically fundamentally non-identifiable. Inference is possible after making (typically unverifiable!) assumptions about the nonresponse process and distributions of the missing responses. In addition, it may be very difficult to determine whether missing data are MAR or nonignorable.
- Although sensitivity analysis (fitting multiple models under different assumptions, including different assumptions about the nonresponse process) can help quantify the degree to which your inferences depend on assumptions, the final model chosen should depend on subjective knowledge about the missing data mechanism and not on a consensus of results from multiple models.