

ASAP #10























6/12/2003
























MISSING DATA

April 2003

Paul D. Allison, Instructor

Copyright © 2003 by Paul D. Allison

- 1  Missing Data
- 2  Basics
- 3  Many Methods
- 4  Assumptions
- 5  Assumptions
- 6  Assumptions
- 7  Assumptions
- 8  Listwise Deletion (Complete Case)
- 9  Listwise Deletion (continued)
- 10  Listwise Deletion (continued)
- 11  Pairwise Deletion (Available Case)
- 12  Dummy Variable Adjustment
- 13  Imputation
- 14  Maximum Likelihood
- 15  Properties of Maximum Likelihood
- 16  ML When Data are MAR
- 17  ML for 2 x 2 Contingency Table
- 18  Maximizing the Likelihood with ℓ_{EM}
- 19  ML for Monotonic Missing Data
- 20  Monotone Principle for 2 x 2 Table
- 21  ML for Multivariate Normal Data
- 22  EM Algorithm

- 46  Imputation with the Dependent Variable
- 47  SAS Program (using defaults)
- 48  Results
- 49  PROC MI Options
- 50  PROC MI Options
- 51  PROC MI Options
- 52  MCMC Options
- 53  MCMC Options
- 54  Checking for Convergence
- 55  Checking for Convergence
- 56  Categorical Variables
- 57  Categorical Variables (cont.)
- 58  Transformations for Normality
- 59  Output from Other SAS PROCS
- 60  PROCS GENMOD and MIXED
- 61  PROC GLM
- 62  Multivariate Inference
- 63  Multivariate Inference (cont.)
- 64  Multivariate Inference (cont.)
- 65  Combining Chi-Squares
- 66  Interactions and Nonlinearities
- 67  Interaction Results
- 68  Imputation Model vs. Analysis Model

Contents of Appendix

Output 1. LEM Example	A1
Output 2. EM Algorithm in SAS	A3
Output 3. PROC MI and MIANALYZE	A7
Output 4. Results from IMPUTE	A14
Computer Exercise 1	A17
Computer Exercise 2	A18
Computer Exercise 3	A19
Code for Exercises	A20
Quiz 1	A24
Quiz 2	A25

Missing Data

Paul D. Allison

April 2003

www.ssc.upenn.edu/~allison

allison@ssc.upenn.edu

1

Basics

Definition: Data are missing on some variables for some observations

Problem: How to do statistical analysis when data are missing? Three goals:

- Minimize bias
- Maximize use of available information
- Get good estimates of uncertainty

2

Assumptions

Missing at random (MAR)

Data on Y are missing at random if the probability that Y is missing does not depend on the value of Y , after controlling for other observed variables

$$\Pr(Y \text{ is missing} | X, Y) = \Pr(Y \text{ is missing} | X)$$

E.g., the probability of missing income depends on marital status, but within each marital status, the probability of missing income does not depend on income.

- ❑ Considerably weaker assumption than MCAR
- ❑ Can test whether missingness on Y depends on X
- ❑ Cannot test whether missingness on Y depends on Y

5

Assumptions

Ignorable

- Missing at random and
- Parameters that govern the missing data process are distinct from parameters to be estimated (unlikely to be violated)

-
- In practice, "MAR" and "ignorable" are used interchangeably
 - If MAR but not ignorable (parameters not distinct), methods assuming ignorability would still be good, just not optimal.
 - If missing data are ignorable, no need to model the missing data process.
 - Any general purpose method for handling missing data must assume that the missing data mechanism is ignorable.

6

Listwise Deletion (continued)

Weaknesses

- May introduce bias if MAR but not MCAR
- May delete a large proportion of cases, resulting in loss of statistical power

Robust to NMAR for predictor variables in regression analysis

Let Y be the dependent variable in a regression (any kind) and X one of the predictors. Suppose

$$\Pr(X \text{ missing} | X, Y) = \Pr(X \text{ missing} | X)$$

Then listwise deletion will not introduce bias.

9

Listwise Deletion (continued)

Example: Want to estimate a regression with number of children as dependent variable and income as an independent variable.

- 30% of cases have missing data on income, persons with high or low income are less likely to report income
- But probability of missing income does not depend on number of children
- Then listwise deletion will not introduce any bias into estimates of regression coefficients

For logistic regression, listwise deletion is robust to NMAR on independent OR dependent variable (but not both)

This property of listwise deletion presumes that regression coefficients are invariant across subgroups (no interaction)

10

Imputation

Any method that substitutes estimated values for missing values

- Replacement with means
- Regression imputation (replace with conditional means)
- Hot deck: Divide sample into homogeneous strata on observed variables. Within each stratum pick "donor" units with observed values to fill in missing values for other units.

Problems

- Often leads to biased parameter estimates (e.g., variances)
- Usually leads to standard error estimates that are biased downward
 - Treats imputed data as real data, ignores variability in imputation.

13

Maximum Likelihood

Choose as parameter estimates those values which, if true, would maximize the probability of observing what has, in fact, been observed.

Likelihood function: Expresses the probability of the data as a function of the data and the unknown parameter values.

Example: Let $p(y|\theta)$ be the probability of observing y , given θ . For a sample of n independent observations, the likelihood function is

$$L(\theta) = \prod_{i=1}^n p(y_i | \theta)$$

14

ML for 2 x 2 Contingency Table

	<u>Vote</u>		
	Yes	No	
Male	28	45	Furthermore, voting was missing for 10 males and 15 females.
Female	22	52	

The parameters are p_{11} , p_{12} , p_{21} , p_{22} . If we exclude cases with missing data, the likelihood is

$$(p_{11})^{28}(p_{12})^{45}(p_{21})^{22}(p_{22})^{52}$$

If we allow for missing data, the likelihood is

$$(p_{11})^{28}(p_{12})^{45}(p_{21})^{22}(p_{22})^{52}(p_{11}+p_{12})^{10}(p_{21}+p_{22})^{15}$$

17

Maximizing the Likelihood with ℓ EM

Freeware for DOS and Windows by Jeroen Vermunt:

www.kub.nl/faculteiten/fsw/organisatie/departementen/mto/

<u>Input</u>	<u>Output</u> (see Output 1 for all results)
man 2 (Two observed variables)	
res 1	* P(sv) *
dim 2 2 2	
lab r s v	1 1 0.1851 (0.0311)
sub sv s	1 2 0.2975 (0.0361)
mod sv	2 1 0.1538 (0.0297)
dat [28 45 22 52 10 15]	2 2 0.3636 (0.0384)

response variable only allows on character labeling

ℓ EM fits a large class of models for categorical data, including log-linear, logit, latent class, and event history models.

r=response s=sex v=vote

sv - fits a saturated model allowing for dependence between S V
s v - sex and voting are independent

It can not handle continuous predictors with missing data
see page A1

18

ML for Multivariate Normal Data

Multivariate normality implies

- All variables are normally distributed
- All conditional expectation functions are linear
- All conditional variance functions are homoscedastic

A strong assumption but widely invoked as the basis for multivariate analysis

Several ways to get ML estimates with missing data, based on this assumption

- Factoring the likelihood for monotone missing data
- EM algorithm
- Direct maximization of the likelihood

21

EM Algorithm

A general approach to getting ML estimates with missing data

Two-step procedure

1. Expectation (E): Find the expected value of the log-likelihood for the observed data, based on current parameter values.
2. Maximization (M): Maximize the expected likelihood to get new parameter estimates.

Repeat until convergence. Log-likelihood must increase at each step.

For multivariate normal data, parameters are means, variances, and covariances.

22

College Example

1994 U.S. News Guide to Best Colleges

- 1302 four-year colleges in U.S.
- Goal: estimate a regression model predicting graduation rate (# graduating/#enrolled 4 years earlier x 100)
- 98 colleges have missing data on graduation rate

Independent variables:

- 1st year enrollment (logged, 5 cases missing)
- Room & Board Fees (40% missing)
- Student/Faculty Ratio (2 cases missing)
- Private=1, Public=0
- Mean Combined SAT Score (40% missing)
- Other variable: Mean ACT scores (45% missing)

25

EM with PROC MI in SAS *convert longitudinal data to patient as the unit of analysis*
 PROC MI DATA=my.college NIMPUTE=0; *Do multiple imputation on zero datasets*
 VAR gradrat lenroll rmbd private stufac csat act;
 EM OUTEM=collem; *regress each variable on all other variables*
 RUN;

See Output 2

EM (MLE) Parameter Estimates

TYPE	_NAME_	GRADRAT	CSAT	LENROLL	private	STUFAC	RMBRD	ACT
MEAN		59.861800	957.875547	6.169419	0.639017	14.863722	4.072556	22.219789
COV	.GRADRAT	355.713651	1352.986086	-0.499848	3.608253	-31.141706	10.384738	30.584246
COV	CSAT	1352.986086	14745	23.238090	9.381605	-198.405558	67.120577	298.905769
COV	LENROLL	-0.499848	23.238090	0.993680	-0.296404	1.382231	-0.018849	0.469532
COV	private	3.608253	9.381605	-0.296404	0.230674	-0.915604	0.188534	0.291178
COV	STUFAC	-31.141706	-198.405558	1.382231	-0.915604	26.885548	-1.685419	-4.121744
COV	RMBRD	10.384738	67.120577	-0.018849	0.188534	-1.685419	1.329032	1.514260
COV	ACT	30.584246	298.905769	0.469532	0.291178	-4.121744	1.514260	7.352990

26

Regression Output

Variable	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	-32.39455	5.62396	-5.76	<.0001
LENROLL	2.08321	0.69661	2.99	0.0029
STUFAC	-0.18139	0.10864	-1.67	0.0954
RMBRD	2.40383	0.51673	4.65	<.0001
private	12.91450	1.48078	8.72	<.0001
CSAT	0.06688	0.00504	13.27	<.0001

Coefficients are true ML estimates—consistent & efficient

But standard errors and t-statistics are not valid. There is no sample size specification that will give correct standard errors.

For input to structural equation modeling software (e.g., PROC CALIS or LISREL) there is an additional problem: For overidentified models, the coefficients will not be true ML estimates (they will be consistent but not efficient).

29

Direct ML

Also known as "raw" ML or "full information" ML

Directly maximize the likelihood for the specified model

Five computer packages will do this for any "LISREL" model

- Amos (commercial)
- Mplus (commercial)
- LISREL 8.5 (commercial)
- MX (freeware) views.vcu.edu/mx
- LINC (Gauss module)

With no missing data, the likelihood for multinormal data is

$$L(\mu, \Sigma) = \prod_i f(y_i | \mu, \Sigma)$$

30

Has trouble with categorical data

\$500-600

Amos Results

Same as proc reg

inverse of the information matrix

Regression Weights						
			Estimate	S.E.	C.R.	P
gradrat	<--	lenroll	2.083	0.595	3.499	0.000
gradrat	<--	private	12.914	1.277	10.114	0.000
gradrat	<--	stufac	-0.181	0.092	-1.968	0.049
gradrat	<--	csat	0.067	0.005	13.949	0.000
gradrat	<--	rmbrd	2.404	0.548	4.386	0.000

Compare with Listwise deletion

Variable	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	-35.02840	7.68461	-4.56	<.0001
lenroll	2.41705	0.95904	2.52	0.0121
private	13.58806	1.94612	6.98	<.0001
stufac	-0.12306	0.13188	-0.93	0.3513
csat	0.06727	0.00643	10.47	<.0001
rmbrd	2.16169	0.71364	3.03	0.0026

33

Limitations of Maximum Likelihood

- Requires estimation of a model for the joint distribution of all the variables
 - Often only interested in conditional distributions.
 - May not be robust
- Models and software may not be readily available
 - Good for linear and loglinear models
 - But nothing for Cox regression, Poisson regression or logistic regression (with continuous predictors).
- Sometimes difficult to incorporate auxiliary information.

34

Adding a Random Component

Solution: For generating imputations use $X = a + bY + s_{x,y}E$ where $s_{x,y}$ is the root mean squared error from the regression and E is a random draw from a standard normal distribution.

For this example, use $X = 2.07 + .189*Y + .906*E$ to fill in the missing values. In SAS:

```
DATA impute;
  SET binorm;
  IF x=. THEN x = 2.07 + .189*y + .906*RANNOR(0);
RUN;
```

When I did this, the estimated correlation between Y and X was .384, close to the true value.

37

Multiple, Random Imputations

Problems with single, random imputation

- Not fully efficient because of random variation
- Standard errors not properly estimated

Solution: Do it multiple times *default = 5 times*

- Averaging the parameter estimates dampens the variation
- Variability among the estimates provides information for correcting the standard errors.

Six replications for the correlation example

Obs	r	se	Obs	r	se
1	0.374	0.038466	4	0.378	0.038331
2	0.368	0.038665	5	0.377	0.038365
3	0.403	0.037458	6	0.407	0.037313

38

Random Variation in Parameters

Problem: Method still underestimates the standard errors.

Why? In using $X = a + bY + s_{x,y}E$ to generate imputations, we assume that a , b , and $s_{x,y}$ are the true parameters, when they're only estimates.

Solution: For "proper" imputations, regression parameters should be random draws from their posterior distribution.

Bayesian statistics:

Prior distribution \rightarrow Posterior Distribution \leftarrow Data

Prior reflects our beliefs about the parameters. In practice, we can use a diffuse or "uninformative" prior.

For a and b , an uninformative prior leads to a normal posterior with means a and b and standard deviations given by their estimated standard errors.

41

Data Augmentation (MCMC)

Problem: The method just described won't work for non-monotone missing data

Solution: Use an iterative method, e.g, Markov chain Monte Carlo

1. Based on starting values for the parameters, generate random draws from the implied distribution of the variables with missing data.
2. Based on the filled-in data, generate a random draw from the posterior distribution of the parameters.
3. Repeat step 1 with newly drawn parameter values, and continue iterating through steps 1 and 2, until the distributions converge.

42

Steps for MI (continued)

4. Back transform any normalized variables and round imputations for discrete variables.
5. Use standard software to estimate desired model on each imputed data set.
6. Use PROC MIANALYZE to combine results into a single set of parameter estimates, standard errors and test statistics.

When generating imputed data sets, you may want to produce an extra set for exploratory analysis. Once you've decided on the model, then apply these six steps.

45

Imputation with the Dependent Variable

For multiple imputation, the dependent variable in a regression analysis should always be included. This means that the dependent variable is used to impute missing values of the independent variables.

Won't this create bias?

- Yes, for conventional deterministic, imputation.
- No, for imputation with a random component. In fact, leaving out the dependent variable will cause bias.

Goal of multiple imputation is to reproduce all the relationships in the data as closely as possible. This can only be accomplished if the dependent variable is included in the imputation process.

46

PROC MI Options

Note: I use syntax for version 8.2 which is slightly different from 8.1.

□ Change number of imputed data sets:

```
PROC MI DATA=my.college OUT=miout NIMPUTE=7;
```

The more the better: More data sets gives more stable parameter estimates, and better standard error estimates.

But there's rapidly diminishing returns. With moderate amounts of missing data, 5 is sufficient. But with more missing data, you should have more data sets.

49

PROC MI Options

□ Force MI to produce the same results every time.

```
PROC MI DATA=my.college OUT=miout SEED=512;
```

SEED can be any positive integer, which controls start of the random number generator.

If you use the same SEED value, you get identical results.

□ Impose minima and/or maxima on imputed values

```
PROC MI DATA=my.college OUT=miout SEED=512
```

```
  MINIMUM= 8 600 . 1 . 1200 11
```

```
  MAXIMUM= 100 1410 . 92 . 8700 31;
```

If MI draws outside the range, it rejects and draws another. 50

* can produce bias variances

MCMC Options

- Choose number of iterations

MCMC NBITER=400 NITER=200;

NBITER sets the number of "burn-in" iterations before the first data set is taken. Default is 200.

NITER sets the number of iterations between successive data sets. Default is 100.

Problem: Unlike ML, there is no simple criterion for convergence. How many iterations you need for convergence is unknown. More is better, but there's always a trade-off between more iterations and more data sets.

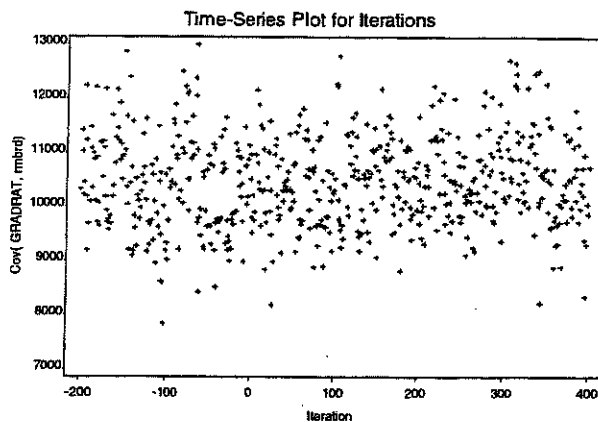
* you may need more if you have lots of missing data.

53

Checking for Convergence

Time series plots for selected parameters

MCMC TIMEPLOT(MEAN(csat rmbd) COV(gradrat*rmbd) WLF);



54

* do not want to see a pattern

Categorical Variables (cont.)

2. Impute N and F in the usual way, with max of 1 and min of 0, but don't round.
3. Calculate $C = 1 - N - F$.
4. Determine which of the three variables has the highest value. Assign that variable a value of 1 and other two a value of 0.

```
DATA miout2;  
  SET miout;  
  c=1-n-f;  
  IF c>n AND c>f THEN DO; n=0; f=0; END;  
  ELSE IF n>f THEN DO; n=1; f=0; END;  
  ELSE DO; n=0; f=1; END;  
RUN;
```

57

Transformations for Normality

Imputations can be improved by transforming variables to achieve approximate normality before imputing, then reversing the transformation after imputation.

In SAS, this can be done in DATA steps, but PROC MI can do many transformations more easily.

For example, RMBRD is somewhat skewed to the right. A logarithmic transformation removes the skewness.

```
PROC MI DATA=my.college OUT=miout;  
  VAR gradrat csat lenroll stufac pubpriv rmbird act;  
  TRANSFORM LOG(rmbird);  
RUN;
```

This applies the transformation, imputes, and back-transforms.
Other available transforms: BOXCOX, EXP, LOGIT, POWER₅₈

* these don't work

PROC GLM

```
PROC GLM DATA=miout;  
  MODEL gradrat=csat lenroll stufac pubpriv rmbrd /  
    INVERSE;  
  BY _IMPUTATION_;  
  ODS OUTPUT PARAMETERESTIMATES=gmparms INVXPX=xpx;  
RUN;  
PROC MIANALYZE PARMS=gmparms XPXI=xpx;  
  VAR INTERCEPT csat lenroll stufac pubpriv rmbrd;  
RUN;
```

What about auxiliary statistics like R^2 or mean squared error?
Just average across the multiple analyses. Use hand calculator, or write to a data set using ODS and apply PROC MEANS.

* do not average test statistics like
F-statistics & ~~intercept~~ t-statistics

61

Multivariate Inference

Suppose you want to test the null hypothesis that all the regression coefficients are 0. Three ways to do it:

1. Wald test using combined covariance matrices--built into MIANALYZE, but based on implausible assumption.
2. Likelihood ratio tests--a better method but awkward to implement.
3. Combining chi-squares--easy but may not be as accurate.

To implement Wald test, simply use the MULT option.

```
PROC MIANALYZE DATA=a MULT EDF=1302;  
  VAR csat lenroll stufac pubpriv rmbrd;  
RUN;
```

* MULT - the equivalent to the overall F-test⁶²
EDF - keeps the degree of freedom estimate
to be below the number of observations

* Note: the INTERCEPT is not included because
we don't want to see if that is different from zero

Combining Chi-Squares

For each completed data set, compute a chi-square to test the null hypothesis of interest.

- This can be Wald, likelihood-ratio or score statistic.
- Convert F-statistic to approximate chi-square by multiplying by its numerator DF (which becomes the DF for the chi-square). Approximation excellent for large denominator DF.

Combine chi-squares with COMBCHI macro (on my web site).

Example: Three completed data sets with chi-squares of 10.1, 14.2, 9.6 and 4 d.f.

```
%combchi(df=4, chi=10.1 14.2 9.6)
```

F	DF	DDF	P
2.3212517	4	21.783421	0.0890181

65

Interactions and Nonlinearities

Because data augmentation is based on linear regression, imputed values will not reflect interactions and nonlinearities. If a variable has a lot of missing data, this may produce misleading results.

If you want to estimate a model with interactions and nonlinearities, these should be built into the imputation process.

Example: In the model for GRADRAT, suppose we hypothesize that the effect of CSAT is different for public and private colleges. Consider the following methods:

1. Impute data as usual. Form the product of CSAT and PRIVATE, and include this in the model. *Not recommended*
2. Divide the data into public and private colleges. Do separate data augmentation in each group. Recombine the two groups, and estimate a model with the product of CSAT and PRIVATE.
3. Form the product of CSAT and PRIVATE. Include that variable in the set of variables for imputation. Then estimate a model that includes the product.

66

Other Parametric Methods

Methods using data augmentation:

Schafer's CAT program (Splus library): For data in which all variables are categorical. Assumes unrestricted multinomial or loglinear model.

Schafer's MIX program (Splus library): For combinations of categorical and quantitative variables. Assumes multinomial model for categorical variables. Within each cell of the table, quantitative variables are assumed to have a multivariate normal distribution with different means but common covariance matrix.

Methods based on multivariate normal model, but using SIR (Sampling Importance/Resampling) algorithm

AMELIA Freeware program developed by King, Honaker, Joseph, Scheve and Singh. (gking.harvard.edu/stats.shtml)

69

Sequential Generalized Regression

Data augmentation requires a comprehensive model for all variables, e.g., multivariate normal. May be unrealistic.

Instead, formulate separate regression model for each variable with missing data (as dependent variable).

- Linear model for quantitative variables.
- Logistic model for categorical variables (2 or more categories).
- Poisson model for count variables.

For each variable, estimate appropriate regression model.
Use model to generate one set of random imputations.

Start with the variable having the least missing data and proceed, in sequence, to the variable with most missing data.

For each regression, use imputed values of predictors based on previous regressions (After first round, all variables will have imputed values.)

Continue through subsequent rounds. Take imputations from every k 'th round.

70

Using IMPUTE

```
OPTIONS SASAUTOS="C:\Program Files\SAS Institute\SAS\VB\SRCLIB"
MAUTOSOURCE;
%IMPUTE(SETUP=new, NAME=mysetup, DIR=c:\My Documents)
DATAIN my.gssmiss;
DATAOUT my.gssout ALL;
CONTINUOUS income;
CATEGORICAL spanking female nodoubt;
BOUNDS income(>0);
DROP age educ region marital region nochild black;
ITERATIONS 10;
MULTIPLES 5; * How many data sets to produce
RUN;

DATA gssout;
SET my.gssout;
RENAME _MULT_=_IMPUTATION_; * this This code must be included
RUN;
PROC SORT DATA=gssout; *
BY _IMPUTATION_;
RUN;
```

73

Using IMPUTE

```
PROC LOGISTIC DATA=gssout OUTEST=a COVOUT;
MODEL spanking=female income nodoubt;
BY _IMPUTATION_;
RUN;

PROC MIANALYZE DATA=a EDF=2991;
VAR INTERCEPT_1 INTERCEPT_2 female income nodoubt;
RUN;
```

See Output 4

Similar methods are also available in MICE library for Splus
www.multiple-imputation.com

Problem with sequential methods: No guarantee that iterations will converge to the posterior distribution of the missing values.

74

Hot Deck Imputation

Let Y be the variable with some missing data. Let X be a set of categorical variables (with no missing data) used to impute Y .

1. Form the contingency table based on X .
2. Identify all cells of the contingency table with some cases missing Y . For each cell, let n_0 be the number of cases missing Y and let n_1 be the number of cases not missing Y .
3. From the set of n_1 cases with complete data, randomly draw n_1 cases (with replacement).
4. From those n_1 cases, randomly draw n_0 cases (with replacement).
5. Randomly assign the n_0 values of Y to the cases with missing data.

Repeat 3-5 for each data set.

77

Hot Deck Imputation (cont.)

This method of sampling is called the ABB method (Approximate Bayesian Bootstrap, Rubin 1987). Unlike other methods, it produces "proper" imputations because it allows for all sources of variability.

McNally describes SAS routines for hot deck multiple imputation, but doesn't use the ABB method.
(www.pstc.brown.edu/papers/wp-1997/97-12ab.html)

Advantages of hot deck: no assumptions about distribution of variable with missing data. Imputed values are just like observed values.

Problems: Can't stratify too finely on X or you may have cells with no donors or a single donor. Not easily generalized to nonmonotonic missing data.

78

Panel Data Example (cont.)

Solution 2 is not always feasible:

- When a variable is missing for everyone at some time point.
- When you have other kinds of clustered data, e.g., students nested in classrooms.

Coefficient Estimates (and Standard Errors) for Fixed-Effects Models Predicting CESD.

	Listwise Deletion by Person	Listwise Deletion by Person-Time	Data Augmentation by Person-Time	Data Augmentation by Person
SRH	2.341 (.586)**	1.641 (.556)**	2.522 (.617)**	1.538 (.501)**
WALK	-1.552 (.771)*	-1.381 (.761)	-1.842 (.960)	-.550 (.825)
ADL	-.676 (.528)	-.335 (.539)	-.385 (.562)	-.410 (.435)
PAIN	.031 (.179)	.215 (.168)	.305 (.180)	.170 (.164)
WAVE 1	8.004 (.650)**	8.787 (.613)**	6.900 (.729)**	9.112 (.615)**
WAVE 2	7.045 (.579)**	7.930 (.520)**	5.808 (.642)**	8.131 (.549)**
N (person-times)	303	453	660	660

* $p < .05$, ** $p < .01$

81

Nonignorable Missing Data

Sometimes we suspect that data are not missing at random:

- people with high incomes may be less likely to report their income
- people who've been arrested may be less likely to report arrest status.

If these are independent variables in a regression model, listwise deletion may do the job.

If data are NMAR, correct inference requires that the missing data mechanism be modeled as part of the inference. Assuming a correct model, both ML and MI can produce optimal inferences in NMAR situations.

Problems:

1. Data contain no information to determine the correct model for the missing data mechanism.
2. Inference may be very sensitive to model choice.

82

Heckman's Model (cont.)

This model implies a likelihood function whose parameters are all identified and which can be maximized by conventional numerical methods.

Problem: ML estimates are extremely sensitive to normality assumption on Y .

Two-step method avoids this problem

1. Estimate probit model without Y . Use predictions from probit model to form "inverse Mills ratio."
2. Estimate linear model with available cases, including inverse Mills ratio as a covariate

But: To achieve identification, the predictors in the probit model can't be exactly the same as in the linear model.

85

Pattern-Mixture Models

Given two variables X and Y , there are four possible patterns:

1. Both variables observed
2. X observed, Y missing
3. Y observed, X missing
4. Both variables missing

Let $R=1,2,3,4$ index the pattern. Pattern-mixture model:

$$f(X,Y,R) = f(X,Y|R)\Pr(R)$$

Might assume that X and Y are bivariate normal with means, standard deviations and a correlation that depend on R . But without further restrictions, this model is highly underidentified.

Little (1993) has proposed several sets of possible restrictions that yield identified models and can be estimated with ML_{ec}

Summary and Review

Among conventional methods, listwise deletion is the least problematic.

- Unbiased if MCAR
- Standard errors good estimates of true standard errors
- Resistant to NMAR for independent variables in regression

All other conventional methods introduce bias into parameter estimates or standard error estimates

By contrast ML and MI have optimal properties under MAR, or under a correctly specified model for missingness

- Parameter estimates approximately unbiased and efficient
- Good estimates of standard errors and test statistics.

89

Summary and Review

ML attractive for linear or loglinear models

- Widely available software
- Simple decision process
- Always produces the same results

For other estimation tasks, consider MI

- Works for any kind of model or data
- May be more robust than ML
- But does not produce a deterministic result
- There are many different ways to do it leading to uncertainty and confusion.

Can also use ML and MI for nonignorable missing data, but

- Requires very good knowledge of missing data process
- Should always be accompanied by a sensitivity analysis

90

Output 1. LEM Example

LEM: log-linear and event history analysis with missing data.
Developed by Jeroen Vermunt (c), Tilburg University, The Netherlands.
Version 1.0 (September 18, 1997).

*** INPUT ***

```
man 2
res 1
dim 2 2 2
lab r s v
sub sv s
mod sv
dat [28 45 22 52
     10 15]
```

*** STATISTICS ***

```
Number of iterations = 6
Converge criterion   = 0.0000001429
Seed random values   = 2248

X-squared            = 0.7985 (0.3716)
L-squared            = 0.8042 (0.3698)
Cressie-Read         = 0.8000 (0.3711)
Dissimilarity index  = 0.0240
Degrees of freedom    = 1
Log-likelihood        = -212.75185
Number of parameters = 4 (+1)
Sample size          = 172.0
BIC(L-squared)       = -4.3433
AIC(L-squared)       = -1.1958
BIC(log-likelihood)  = 446.0937
AIC(log-likelihood)  = 433.5037
```

```
Eigenvalues information matrix
217.3248  135.0234  99.9867
```

*** FREQUENCIES ***

* SUBGROUP r 1 *

s v	observed	estimated	std. res.
1 1	28.000	27.208	0.152
1 2	45.000	43.728	0.192
2 1	22.000	22.614	-0.129
2 2	52.000	53.450	-0.198

X-squared = 0.1161, L-squared = 0.1160

Output 2: EM Algorithm in SAS *For longitudinal data convert to patient as the unit of analysis*
The MI Procedure

Model Information

Data Set	MY.COLLEGE
Method	MCMC
Multiple Imputation Chain	Single Chain
Initial Estimates for MCMC	EM Posterior Mode
Start	Starting Value
Prior	Jeffreys
Number of Imputations	0
Number of Burn-in Iterations	200
Number of Iterations	100
Seed for random number generator	34817

Missing Data Patterns

Group	GRADRAT	lenroll	rmbrd	private	STUFAC	CSAT	ACT	Freq
1	X	X	X	X	X	X	X	297
2	X	X	X	X	X	X	.	158
3	X	X	X	X	X	.	X	123
4	X	X	X	X	X	.	.	156
5	X	X	.	X	X	X	X	158
6	X	X	.	X	X	X	.	119
7	X	X	.	X	X	.	X	81
8	X	X	.	X	X	.	.	110
9	X	.	X	X	X	.	.	1
10	X	.	.	X	X	.	X	1
11	.	X	X	X	X	X	X	16
12	.	X	X	X	X	X	.	5
13	.	X	X	X	X	.	X	11
14	.	X	X	X	X	.	.	16
15	.	X	.	X	X	X	X	16
16	.	X	.	X	X	X	.	8
17	.	X	.	X	X	.	X	9
18	.	X	.	X	X	.	.	13
19	.	X	.	X	.	.	X	1
20	.	.	.	X	X	X	.	1
21	.	.	.	X	X	.	.	1
22	.	.	.	X	.	X	X	1

14	17.225000		
15	14.268750	880.312500	20.375000
16	17.437500	859.125000	
17	16.555556		19.888889
18	14.938462		
19			20.000000
20	16.900000	920.000000	
21	8.100000		
22		890.000000	22.000000

Initial Parameter Estimates for EM

TYPE	_NAME_	GRADRAT	lenroll	rmbrd	private
MEAN		60.405316	6.167520	4.145115	0.639017
COV	GRADRAT	356.796514	0	0	0
COV	lenroll	0	0.994300	0	0
COV	rmbrd	0	0	1.367937	0
COV	private	0	0	0	0.230852
COV	STUFAC	0	0	0	0
COV	CSAT	0	0	0	0
COV	ACT	0	0	0	0

Initial Parameter Estimates for EM

	STUFAC	CSAT	ACT
	14.858769	967.978177	22.120448
	0	0	0
	0	0	0
	0	0	0
	0	0	0
	0	0	0
	26.898730	0	0
	0	15271	0
	0	0	6.655879

EM (MLE) Parameter Estimates

TYPE	_NAME_	GRADRAT	lenroll	rmbrd	private
MEAN		59.861800	6.169419	4.072556	0.639017
COV	GRADRAT	355.713651	-0.499848	10.384738	3.608253
COV	lenroll	-0.499848	0.993680	-0.018849	-0.296404
COV	rmbrd	10.384738	-0.018849	1.329032	0.188534

Output 3: PROC MI and MIANALYZE in SAS

The MI Procedure

Model Information

Data Set	MY.COLLEGE
Method	MCMC
Multiple Imputation Chain	Single Chain
Initial Estimates for MCMC	EM Posterior Mode
Start	Starting Value
Prior	Jeffreys
Number of Imputations	5
Number of Burn-in Iterations	200
Number of Iterations	100
Seed for random number generator	40818 - allows you to duplicate a run.

Missing Data Patterns

Group	GRADRAT	CSAT	lenroll	STUFAC	private	rmbrd	ACT	Freq
1	X	X	X	X	X	X	X	297
2	X	X	X	X	X	X	.	158
3	X	X	X	X	X	.	X	158
4	X	X	X	X	X	.	.	119
5	X	.	X	X	X	X	X	123
6	X	.	X	X	X	X	.	156
7	X	.	X	X	X	.	X	81
8	X	.	X	X	X	.	.	110
9	X	.	.	X	X	X	.	1
10	X	.	.	X	X	.	X	1
11	.	X	X	X	X	X	X	16
12	.	X	X	X	X	X	.	5
13	.	X	X	X	X	.	X	16
14	.	X	X	X	X	.	.	8
15	.	X	.	X	X	.	.	1
16	.	X	.	.	X	.	X	1
17	.	.	X	X	X	X	X	11
18	.	.	X	X	X	X	.	16
19	.	.	X	X	X	.	X	9
20	.	.	X	X	X	.	.	13
21	.	.	X	.	X	.	X	1
22	.	.	.	X	X	.	.	1

14	0.750000	.	.
15	1.000000	.	.
16	0	.	22.000000
17	0.545455	3.220636	20.818182
18	0.687500	3.197125	.
19	0.666667	.	19.888889
20	0.538462	.	.
21	0	.	20.000000
22	0	.	.

EM (Posterior Mode) Estimates

TYPE	_NAME_	GRADRAT	CSAT	lenroll	STUFAC
MEAN		59.861361	957.872020	6.169419	14.863727
COV	GRADRAT	353.443513	1344.749434	-0.496657	-30.948417
COV	CSAT	1344.749434	14636	23.098589	-197.212436
COV	lenroll	-0.496657	23.098589	0.987600	1.373794
COV	STUFAC	-30.948417	-197.212436	1.373794	26.721189
COV	private	3.586141	9.323952	-0.294594	-0.910016
COV	rmbrd	10.321651	66.733784	-0.018747	-1.675178
COV	ACT	30.402550	296.922664	0.466704	-4.097392

EM (Posterior Mode) Estimates

	private	rmbrd	ACT
	0.639017	4.072529	22.219723
	3.586141	10.321651	30.402550
	9.323952	66.733784	296.922664
	-0.294594	-0.018747	0.466704
	-0.910016	-1.675178	-4.097392
	0.229266	0.187376	0.289382
	0.187376	1.317345	1.505729
	0.289382	1.505729	7.298942

Multiple Imputation Variance Information

Variable	-----Variance-----			DF
	Between	Within	Total	
GRADRAT	0.004816	0.272642	0.278421	1118.8
CSAT	3.575124	11.248270	15.538419	49.699
lenroll	0.000003534	0.000764	0.000769	1279.3

$$\text{with} = U = \frac{1}{M} \sum_k S_k^2 \quad B = \frac{1}{M-1} \sum_k (b_k - \bar{b})^2$$

$$r = \left(1 + \frac{1}{M}\right) \frac{B}{U} \text{ "relative increase in variance due to missing data"}$$

A9

* The Fraction Missing Information is not only a function of missing data from the variable but the variables corr@it.
 ----- Imputation Number=1 -----

The REG Procedure
 Model: MODEL1
 Dependent Variable: GRADRAT

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	209473	41895	216.59	<.0001
Error	1296	250687	193.43114		
Corrected Total	1301	460160			

Root MSE	13.90795	R-Square	0.4552
Dependent Mean	59.76663	Adj R-Sq	0.4531
Coeff Var	23.27043		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-31.18091	4.40677	-7.08	<.0001
CSAT	1	0.06092	0.00408	14.95	<.0001
lenroll	1	2.47861	0.54445	4.55	<.0001
STUFAC	1	-0.21523	0.08440	-2.55	0.0109
private	1	13.08602	1.14974	11.38	<.0001
rmbrd	1	2.96712	0.41546	7.14	<.0001

Results for Imputation Numbers 2-5 are omitted here

Multiple Imputation Parameter Estimates

Parameter	Minimum	Maximum
INTERCEPT	-34.543012	-30.022663
csat	0.060922	0.071036
lenroll	1.845776	2.478607
stufac	-0.268629	-0.160517
private	12.283751	13.723749
rmbrd	1.807667	2.967120

Multiple Imputation Parameter Estimates

Parameter	Theta0	t for H0:	
		Parameter=Theta0	Pr > t
INTERCEPT	0	-6.75	<.0001
csat	0	11.62	<.0001
lenroll	0	3.41	0.0010
stufac	0	-2.20	0.0308
private	0	9.71	<.0001
rmbrd	0	3.72	0.0034

The MIANALYZE Procedure

Model Information

Data Set WORK.A
Number of Imputations 5

Multiple Imputation Variance Information

Parameter	-----Variance-----			DF
	Between	Within	Total	
INTERCEPT_1	0.017452	0.018159	0.039101	13.806
INTERCEPT_2	0.013768	0.019100	0.035621	18.381
female	0.001547	0.004903	0.006760	51.763
income	0.000002968	0.000001913	0.000005475	9.3658
nodoubt	0.017698	0.005329	0.026567	6.1948

Multiple Imputation Variance Information

Parameter	Relative Increase in Variance	Fraction Missing Information
INTERCEPT_1	1.153264	0.590405
INTERCEPT_2	0.864995	0.513465
female	0.378635	0.300537
income	1.861904	0.706711
nodoubt	3.984852	0.842722

Multiple Imputation Parameter Estimates

Parameter	Estimate	Std Error	95% Confidence Limits		DF
INTERCEPT_1	-0.454215	0.197741	-0.87889	-0.02954	13.806
INTERCEPT_2	1.638597	0.188735	1.24267	2.03453	18.381
female	-0.477818	0.082217	-0.64282	-0.31282	51.763
income	-0.007701	0.002340	-0.01296	-0.00244	9.3658
nodoubt	0.673868	0.162993	0.27806	1.06968	6.1948

Computer Exercises for Missing Data Course

All exercises will use the data set NLSY (available on my web site), which has records for 581 children who were interviewed in 1990 as part of the National Longitudinal Study of Youth. Here are the variables:

ANTI	antisocial behavior, measured with a scale ranging from 0 to 6.
SELF	self-esteem, measured with a scale ranging from 6 to 24.
POV	poverty status of family, coded 1 for in poverty, otherwise 0.
BLACK	1 if child is black, otherwise 0
HISPANIC	1 if child is Hispanic, otherwise 0
CHILDAGE	child's age in 1990
DIVORCE	1 if mother was divorced in 1990, otherwise 0
GENDER	1 if female, 0 if male
MOMAGE	mother's age at birth of child
MOMWORK	1 if mother was employed in 1990, otherwise 0

There are two versions of this data set:

1. An SPSS data set named NLSY.SAV. This can be read directly by Amos.
2. A SAS data set (for Windows) named NLYS.SAS7BDAT. This can be accessed directly by SAS procedures, but you will have to declare a SAS Library before doing so.

My web site contains a text file of this data set (with SAS code for reading the data).

Exercise 1. EM Algorithm with SAS

1. Use PROC CORR to get the number of non-missing observations for each variable, along with descriptive statistics and correlations for available cases. To do this, you need only submit the statements:

```
proc corr data=nlsy; run;
```

2. Use PROC REG to estimate the regression of ANTI on all the other variables (default is to use listwise deletion). The syntax is

```
proc reg data=nlsy;  
model anti=self pov black hispanic childage divorce gender  
      momage momwork;  
run;
```

Exercise 3. Multiple Imputation with SAS

A. Basic (slide 48)

1. Use PROC MI with default settings to produce five multiply imputed data sets (stacked into one SAS data set).
2. Use PROC REG with a BY statement to estimate five regression models, requesting the coefficients and covariance matrix be written to an output data set.
3. Use PROC MIANALYZE to combine results from the five regression analyses. Compare results with those using listwise deletion and maximum likelihood.

B. Advanced

1. Use PROC MI to produce 10 multiply imputed data sets (slide 50). Use the following options.
 - a. Wherever appropriate set maxima, minima, and rounding units (slides 51-52).
 - b. Change the number of iterations between successive data sets to 100 (slide 54).
 - c. Request the TIMEPLOT and the ACFPLOT for the worst linear function. Evaluate the results (slides 55-56).
2. In a DATA step, modify the imputed values for BLACK and HISPANIC as described in class (slides 57-58).
3. Use PROC REG with a BY statement to estimate ten regression models, requesting the coefficients and covariance matrix be written to an output data set. Use a TEST statement to test the null hypothesis that both BLACK and HISPANIC have coefficients of zero, i.e.,

TEST black=0, hispanic=0;

Compare results with those from previous runs.

4. Use PROC MIANALYZE to combine results from the 10 regression analyses. Set the EDF option to 581 (slides 65-66). Use the MULT option to get a test of the null hypothesis that all coefficients are equal to 0. (You'll have to omit INTERCEPT from the variable list to get this right).
5. Redo PROC MIANALYZE omitting all variables but HISPANIC and BLACK and using the MULT option. This will produce a test of the null hypothesis that both coefficients are equal to 0.
6. As an alternative test, combine the F-statistics produced by the TEST statement, using the COMBCHI macro (p.68). You'll have to multiply the F statistics by 2 (the numerator DF) to convert them to chi-square statistics before putting them in the COMBCHI macro.

Exercise 2

```
Sub Main
Dim Sem As New AmosEngine
Sem.TableOutput
Sem.Smc
Sem.ModelMeansAndIntercepts
Sem.BeginGroup "c:\data\nlsy.sav"
Sem.Mean "self"
Sem.Mean "pov"
Sem.Mean "black"
Sem.Mean "hispanic"
Sem.Mean "childage"
Sem.Mean "divorce"
Sem.Mean "gender"
Sem.Mean "momage"
Sem.Mean "momwork"
Sem.Structure
"anti=()+self+pov+black+hispanic+childage+divorce+gender
    +momage+momwork+(1)other"
Sem.FitModel
End Sub
```

```

PROC REG DATA=modified OUTEST=a COVOUT;
  MODEL anti=self pov black hispanic chldage divorce gender
    momage momwork;
  TEST black=0,hispanic=0;
  BY _IMPUTATION_;
  ODS OUTPUT TESTANOVA=b; /*optional statement*/
RUN;

PROC MIANALYZE DATA=a EDF=581 MULT;
  VAR self pov black hispanic chldage divorce gender
    momage momwork;
RUN;

PROC MIANALYZE DATA=a EDF=581 MULT;
  VAR black hispanic ;
RUN;

DATA test; /*This data step is optional*/
  SET b;
  WHERE df=2;
  fvalue=2*fvalue;
  PUT fvalue;
RUN;

%INCLUDE 'c:\data\combchi.sas';
%combchi(df=2,chi=5.83 6.49 7.84 4.31 3.48 6.39 4.97 8.67 13.07
5.93)

```

Missing Data Quiz 2

Name _____

1. The principal reason for introducing random variation into the imputation process is to avoid bias in parameter estimates.
T F
2. One reason for doing random imputation more than once is to make it possible to get good standard error estimates.
T F
3. In generating random regression imputations, the random component is simply a random draw from a standard normal distribution.
T F
4. For proper multiple imputations, random draws must be made for the parameters in order to avoid bias in the parameter estimates.
T F
5. Like ML, multiple imputation is based on the assumption of multivariate normality.
T F
6. The set of variables used in multiple imputation should NOT include variables that are unrelated to the variables with missing data.
T F
7. It's rare to need more than five completed data sets to do multiple imputation.
T F
8. The problem of getting different results every time you do multiple imputation can be solved by using the same "seed" each time.
T F
9. Unfortunately, there's no good criterion for determining whether the MCMC algorithm has converged.
T F
10. When imputing dummy variables under the multivariate normality assumption, the imputed values should always be rounded to 0 or 1.
T F
11. In version 8 of SAS, any procedure output that would ordinarily be printed can be written to a SAS data set, using the ODS statement.
T F
12. A missing data pattern is simply the set of variables present and the set of variables absent for some subset of the cases.
T F
13. In PROC MIANALYZE, the method used for testing whether several parameters are all equal to zero depends on the assumption that the amount of missing information is the same for all those parameters.
T F
14. If your analysis model contains interactions and nonlinearities, so should your imputation model.
T F
15. Under multivariate normality, multiple imputation is based on linear regression. But if the R^2 's from all regressions are zero, multiple imputation has no advantage over listwise deletion.
T F