STATISTICAL HORIZONS

Missing Data Using Stata

Paul Allison, Ph.D.

Upcoming Seminar:
August 15-16, 2017, Stockholm, Sweden

Missing Data Using Stata

Paul D. Allison, Ph.D.

February 2016

www.StatisticalHorizons.com

Basics

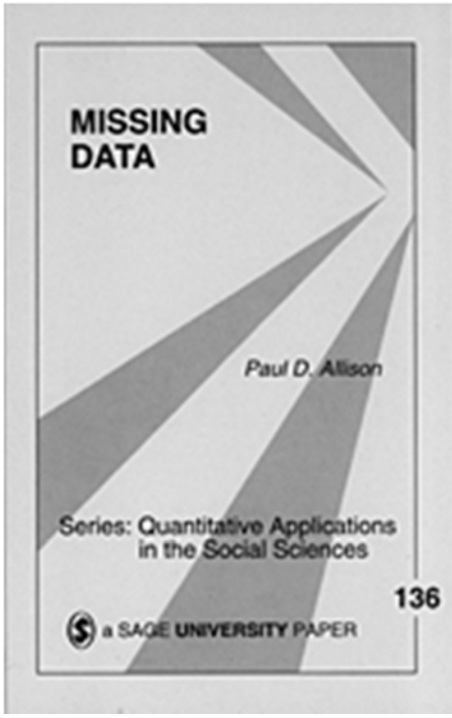
Definition: Data are missing on some variables for some observations

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

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

NOT a goal: imputed values "close" to real values.

For Further Reading



Also:

Allison, Paul D. (2009) "Missing Data." Pp. 72-89 in The SAGE Handbook of Quantitative Methods in Psychology, edited by Roger E. Millsap and Alberto Maydeu-Olivares. Thousand Oaks, CA: Sage Publications Inc. http://statisticalhorizons.com/wp-content/uploads/2012/01/Milsap-Allison.pdf

Many Methods

- Conventional
  - Listwise deletion (complete case analysis)
  - Pairwise deletion (available case analysis)
  - Dummy variable adjustment
  - Imputation
    - Replacement with means
    - Regression
    - Hot deck
- Novel
  - Maximum likelihood
  - Multiple imputation
  - Inverse probability weighting (not discussed here)

Assumptions

Missing completely at random (MCAR)

Suppose some data are missing on *Y*. These data are said to be MCAR if the probability that *Y* is missing is unrelated to *Y* or other variables *X* (where *X* is a vector of observed variables).

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

- MCAR is the ideal situation.
- What variables must be in the *X* vector? Only variables in the model of interest.
- If data are MCAR, complete data subsample is a random sample from original target sample.
- MCAR allows for the possibility that *missingness* on one variable may be related to *missingness* on another
  - e.g., sets of variables may always be missing together

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 observed variables

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

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
- Only *X*'s in the model must be considered. *But*, including other *X*'s (correlated with *Y*) can make MAR more plausible.
- *Can* test whether missingness on *Y* depends on *X*
- *Cannot* test whether missingness on *Y* depends on *Y*

Ignorability

The missing data mechanism is said to be ignorable if

- The data are missing at random *and*
- Parameters that govern the missing data mechanism 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 the missing data mechanism is ignorable, there is no need to model it.
- Any general purpose method for handling missing data must assume that the missing data mechanism is ignorable.

Assumptions

Not missing at random (NMAR)

If the MAR assumption is violated, the missing data mechanism must be modeled to get good parameter estimates.

Heckman's regression model for sample selection bias is a good example.

Effective estimation for NMAR missing data requires very good prior knowledge about missing data mechanism.

- Data contain no information about what models would be appropriate
- No way to test goodness of fit of missing data model
- Results often very sensitive to choice of model
- Listwise deletion able to handle one important kind of NMAR

Listwise Deletion (Complete Case)

Delete any unit with any missing data (only use complete cases)

Strengths

- Easy to implement
- Works for any kind of statistical analysis
- If data are MCAR, does not introduce any bias in parameter estimates
- Standard error estimates are appropriate

Listwise Deletion (continued)

Weaknesses

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

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

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

Then listwise deletion will not introduce bias.

Listwise Deletion (continued)

Example: 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 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)

Caveat: This property of listwise deletion presumes that regression coefficients are invariant across subgroups (no omitted interactions).

Pairwise Deletion (Available Case)

- For linear models, parameters are functions of means, variances and covariances (moments)
- Estimate each moment with all available nonmissing cases
- Plug moment estimates into formulas for parameters

Strengths:
- Approximately unbiased if MCAR
- Uses all available information

Weaknesses:
- Standard errors incorrect (no appropriate sample size)
- Biased if MAR but not MCAR
- May break down (correlation matrix not positive definite)

Dummy Variable Adjustment

A popular method for handling missing data on predictors in regression analysis (Cohen and Cohen 1985)

In a regression predicting *Y*, suppose there is missing data on a predictor *X*.

1. Create a new variable *D*=1 if *X* is missing and *D*=0 if *X* is present.
2. When *X* is missing, set *X*=*c* where *c* is some constant (e.g., the mean of *X*).
3. Regress *Y* on both *X* and *D* (and any other variables)

- Produces biased coefficient estimates (Jones, JASA, 1996)
- So does a related method: For categorical variables, create a separate missing data category
- But may be appropriate for "doesn't apply" missing data
- May also be useful for predictive modeling with missing data.

Imputation

Any method that substitutes estimated values for missing values

- Replacement with means
- Regression imputation (replace with conditional means)

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 inherent uncertainty in imputed values.

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 density for y, given θ (a vector of parameters). For a sample of n independent observations, the likelihood function is

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

Properties of Maximum Likelihood

To get ML estimates, we find the value of θ that maximizes the likelihood function.

Under usual conditions, ML estimates have the following properties:

- Consistent (implies approximately unbiased in large samples)
- Asymptotically efficient
- Asymptotically normal

ML with Ignorable Missing Data

Suppose we have 2 discrete variables X and Y, and there is ignorable missing data on X. Let $p(x,y|\theta)$ be the joint probability function.

For a single observation with X missing, the likelihood is

$$g(y \mid \theta) = \sum_{x} p(x, y \mid \theta) = E_x[p(y \mid x)]$$

The likelihood for the entire sample with m complete cases is

$$L(\theta) = \prod_{i=1}^{m} p(x_i, y_i \mid \theta) \prod_{i=m+1}^{n} g(y_i \mid \theta)$$

This likelihood may be maximized like any other.

ML for 2 x 2 Contingency Table

	Vote: Yes	No
Male	36	37
Female	22	52

Furthermore, voting was missing for 10 males and 15 females.

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

$$(p_{11})^{36}(p_{12})^{37}(p_{21})^{22}(p_{22})^{52}$$

If we allow for missing data, the likelihood is

$$(p_{11})^{36}(p_{12})^{37}(p_{21})^{22}(p_{22})^{52}(p_{11}+p_{12})^{10}(p_{21}+p_{22})^{15}$$

Maximizing the Likelihood with ℓEM

Freeware for Windows by Jeroen Vermunt:
http://members.home.nl/jeroenvermunt/

Input

```
man 2
res 1
dim 2 2 2
lab r s v
sub sv s
mod sv
dat [36 37 22 52 10 15]
```

Output

```
* P(sv) *
 1 1   0.2380 (0.0339)
 1 2   0.2446 (0.0342)
 2 1   0.1538 (0.0297)
 2 2   0.3636 (0.0384)
```

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

ML for Quantitative Variables

Assume multivariate normality, which 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 patterns
- EM algorithm
- Direct maximization of the likelihood

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 log-likelihood to get new parameter estimates.

Repeat until convergence.

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

EM for Multivariate Normal Data

1. Choose starting values for means and covariance matrix.
2. If data are missing on *x*, use current values of parameters to calculate the linear regression of *x* on all variables present for each case.
3. Use linear regressions to impute values of *x*. (E-step)
4. After all data have been imputed, recalculate means and covariance matrix, with corrections for variances and covariances (*see next slide*). (M-step)
5. Repeat steps 2-4 until convergence.

EM for Multivariate Normal Data

Correction: Suppose *X* was imputed using variables *W* and *Z.*

Let $S^2_{x.wz}$ be the residual variance from that regression. Then, in calculating the variance for *X,* wherever you would use x^2_i , substitute $x^2_i + S^2_{x.wz}$

For covariances between two variables with missing values, there's a similar correction in which you add the residual covariance.

EM algorithm for multivariate normal data is available in many commercial software packages: SPSS, Systat, SAS, Splus, Stata

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)
- Auxiliary variable: Mean ACT scores (45% missing)

Preliminary Analysis 1

```
use c:\data\college.dta, clear
mi set wide
```

This declares the data to be a missing data set. It also specifies that imputed data are to be stored in the wide format. The are four different storage formats. But how it's stored usually doesn't matter, and we're not imputing yet anyway.

```
mi misstable summarize
```

This requests basic descriptive statistics.

Preliminary Analysis 2

Missing

Not missing

```
                                                  Obs<.
                                        +------------------------------
          |                             | Unique
 Variable |   Obs=.   Obs>.   Obs<.     | values       Min        Max
----------+-----------------------------+------------------------------
  gradrat |     98            1,204     |     89         8        118
  lenroll |      5            1,297     |   >500  2.890372   8.912608
    rmbrd |    519              783     |   >500      1.26        8.7
   stufac |      2            1,300     |    208       2.3       91.8
     csat |    523              779     |    339       600       1410
      act |    588              714     |     17        11         31
-----------------------------------------------------------------------
```

Preliminary Analysis 3

`mi misstab patterns`

Missing-value patterns (1 means complete)

Percent	1	2	3	4	5	6
23%	1	1	1	1	1	1
12	1	1	1	0	1	1
12	1	1	1	1	1	0
12	1	1	1	1	0	0
9	1	1	1	1	0	1
9	1	1	1	0	1	0
8	1	1	1	0	0	0
6	1	1	1	0	0	1
1	1	1	0	0	1	1
1	1	1	0	1	0	0
1	1	1	0	1	1	1
<1	1	1	0	0	0	0
<1	1	1	0	1	0	1
<1	1	1	0	0	0	1
<1	1	1	0	0	1	0
<1	1	1	0	1	1	0
<1	0	0	0	0	1	1
<1	0	1	0	0	0	1
<1	1	0	0	0	0	0
<1	1	0	0	0	1	0
<1	1	0	1	0	0	1
<1	1	0	1	1	0	0
100%						

Variables are (1) stufac (2) lenroll (3) gradrat (4) rmbrd (5) csat (6) act

EM in Stata

```
mi register impute gradrat lenroll rmbrd stufac csat
  act private
mi impute mvn gradrat lenroll rmbrd stufac csat act
  private, emonly
```

	gradrat	lenroll	rmbrd	stufac	csat	act	private
_cons	59.8618	6.169419	4.072555	14.86372	957.8762	22.2198	.6390169
Sigma							
gradrat	355.7137	-.4998451	10.38471	-31.14171	1352.981	30.58451	3.608253
lenroll	-.4998451	.9936801	-.0188409	1.382231	23.23804	.4695323	-.2964039
rmbrd	10.38471	-.0188409	1.32903	-1.685404	67.11875	1.514341	.1885311
stufac	-31.14171	1.382231	-1.685404	26.88555	-198.4039	-4.121786	-.9156043
csat	1352.981	23.23804	67.11875	-198.4039	14745.07	298.9068	9.381542
act	30.58451	.4695323	1.514341	-4.121786	298.9068	7.353064	.29118
private	3.608253	-.2964039	.1885311	-.9156043	9.381542	.29118	.2306743

These are the maximum likelihood estimates of the means and the covariance matrix.

Convert Covariances to Correlations

ML covariance matrix → ML correlation matrix

```
matrix Sigma=r(Sigma_em)
matrix M=r(Beta_em)
*we'll need these means later
_getcovcorr Sigma, corr
matrix C = r(C)
matlist C
```

	gradrat	lenroll	rmbrd	stufac	csat	act	private
gradrat	1						
lenroll	-.0265865	1					
rmbrd	.4776137	-.016395	1				
stufac	-.3184437	.2674224	-.2819532	1			
csat	.5907693	.1919786	.4794608	-.3151137	1		
act	.598022	.1737033	.4844202	-.2931513	.907775	1	
private	.3983337	-.6191004	.3404992	-.367662	.1608612	.2235773	1

EM As Input to **regress**

```
corr2data gradrat lenroll rmbrd stufac csat act
  private, cov(Sigma) mean(M) clear
regress gradrat lenroll rmbrd stufac csat private
```

This produces ML estimates of the regression coefficients. But standard errors and associated statistics are incorrect because the sample size is taken to be 1302.

gradrat	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lenroll	2.083176	.5393847	3.86	0.000	1.025013	3.141339
rmbrd	2.403941	.4000983	6.01	0.000	1.61903	3.188852
stufac	-.1813901	.0841226	-2.16	0.031	-.3464216	-.0163587
csat	.066875	.0039007	17.14	0.000	.0592227	.0745273
private	12.91442	1.146564	11.26	0.000	10.66509	15.16374
_cons	-32.39475	4.354628	-7.44	0.000	-40.93764	-23.85186

These are ML estimates

These are biased estimates

Direct ML

Also known as “raw ML” or “full information ML” (FIML)

Directly maximize the likelihood for the specified model

Several structural equation modeling (SEM) packages can do this for a large class of linear models.

- Amos (www-03.ibm.com/software/products/en/spss-amos)
- Mplus (www.statmodel.com)
- LISREL (www.ssicentral.com/lisrel)
- OpenMX (R package) (openmx.psyc.virginia.edu)
- EQS (www.mvsoft.com)
- PROC CALIS (support.sas.com)
- Stata **sem** (www.stata.com)
- lavaan (R package) (lavaan.ugent.be)

Direct ML

With no missing data, the multivariate normal likelihood is

$$L(\theta) = \prod_i f(\mathbf{y}_i \mid \boldsymbol{\mu}(\theta), \Sigma(\theta))$$

where

$$f(\mathbf{y}) = \frac{\exp[-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})]}{(2\pi)^{k/2} \mid \Sigma \mid^{1/2}}$$