Missing data

The best thing to do about missing data is not to have any. —Gertrude Cox

The only thing we know for sure about a missing data point is that it is not there, and there is nothing that the magic of statistics can do do change that. The best that can be managed is to estimate the extent to which missing data have influenced the inferences we wish to draw. —Wainer, 2009

- Dealing with missing data
  - General strategies
  - Simple Imputation
  - Estimation with missing data (EM algorithms)
  - Multiple imputation
  - Plots for missing data

Missing data mechanisms

Recent progress in dealing with missing data has come from an understanding of the *reasons* why data may be missing. These can be described under three categories:

The probability that a data value is missing . . .

MCAR (missing completely at random) . . . is *unrelated to the observed or missing* values. Like tossing a coin.

MAR (missing at random) . . . may depend on *other* observed variables, but not on the variable which is missing. e.g., divorced people less likely to their report income.

MNAR (missing NOT at random) . . . depends on the missing value itself. e.g., people with high or low income less likely to report it.

- When data is MCAR or MAR, missing data is said to be *ignorable* — valid methods of analysis exist that don't require modeling the process that generates missing data.
- Otherwise (MNAR), missing data is *non-ignorable* — unbiased results require a secondary model for missingness.
- Cold comfort: Except in special circumstances (substantitive knowledge of data collection process), one cannot determine whether data is MCAR, MAR or MNAR.

Dealing with missing data

- Standard software uses one of two procedures for missing data:
  - Complete case analysis (listwise deletion) — Discard data with *any* missing variables
  - Available case analysis (pairwise deletion)— Discard data with missing values on the analysis variables
- E.g., in linear models (regression, ANOVA, etc.):
  - univariate statistics based on available data
  - missing on *any* predictor $\rightarrow$ case deleted (listwise)
  - missing on response $\rightarrow$ case not used, but a fitted value is generated.
  - multiple responses (MANOVA, Multivar regression): software differs
- Caveats:
  - Must assume at least *missing-at-random* (MAR): "missingness" on X is unrelated to value of X
  - Failure of MAR $\rightarrow$ results (predicted values, coefficients) are biased
  - Factor analysis, PCA, etc: available case analysis can $\rightarrow$ improper correlation matrices (not PD)

Missing data: General strategies

- ***Deletion***: Decreases power, but unbiased if missing is MCAR.
- ***Single imputation***: replace missing by 'suitable estimates', use complete-case analysis.
- ***Weighting***: discard if any missing, but weight complete cases to compensate for incomplete cases.
- ***Direct analysis*** of incomplete data. Two forms:
  - Available case analysis
  - Maximum likelihood estimation over available data (e.g., `PROC MIXED`, E-M algorithm)
- ***Multiple imputation***:
  - Impute $m > 1$ from appropriate distribution for each missing observation.
  - Combine $\rightarrow$ estimates, std. errors that incorporate missing-data uncertainty.
- See: General FAQ 25: Handling missing or incomplete data `http://ssc.utexas.edu/consulting/answers/general/gen25.html`

Missing data: Bias and Power

- Various missing data techniques have different consequences, depending on the missing data mechanism [a]

TABLE 1.1 Parameter Bias and Statistical Power Problems of Common Missing Data Techniques

Missing Data Technique	Missingness Mechanism: MCAR	MAR	MNAR
Listwise deletion	Unbiased, low power	Biased, low power	Biased, low power
Pairwise deletion	Unbiased, inaccurate power	Biased, inaccurate power	Biased, inaccurate power
Maximum likelihood	**Unbiased, accurate power**	**Unbiased, accurate power**	Biased, accurate power
Multiple imputation	**Unbiased, accurate power**	**Unbiased, accurate power**	Biased, accurate power

Note. Recommended techniques are in boldface.

[a] Table from D. A. Newman (2009), "Missing data techniques and low response rates"

Missing data: Imputation

- Trade-offs
  - $+$ Get complete data $\rightarrow$ use software not handling missings
  - $+$ Makes good use of info on incomplete cases
  - $-$ Analyses overstate precision: nominal 95% CI may have only 80%–90% coverage; p-values $< .05$ may be really 0.10–0.20!
- Some imputation methods:
  - Unconditional imputation: fill in the grand mean
    - Only gives illusion of progress!
    - Estimates of variance are understated!
  - Conditional imputation
    - Regression-based imputation: Fill-in $\hat{x}$ using other Xs as predictors, i.e., $\mathcal{E}(X \mid \text{others})$.
    - Sub-group means, i.e., $\mathcal{E}(X \mid \text{group})$.
    - Cluster-based imputation: group into clusters; fill-in cluster mean
  - Stochastic conditional imputation
    - Regression: Fill-in $\hat{x}_i + r_i$, where $r_i \sim \mathcal{N}(0, \sigma^2)$.
    - "Hot-deck" imputation: cluster, then fill-in randomly chosen observation in same cluster.
  - Multiple imputation: Impute $m \geq 2$ values for each missing obs.; use variability of these imputations to correct std. errors, p-values

Single Imputation: Examples

Auto data: some missing values for repair record in 77 & 78

- Unconditional means (Not A Good Idea)

```
%include data(auto);
data auto;
   set auto;
   r77 = rep77;  *-- copy original vars;
   r78 = rep78;
proc standard REPLACE;
   var rep77 rep78;
proc print;
   where (r77=. or r78=.);
```

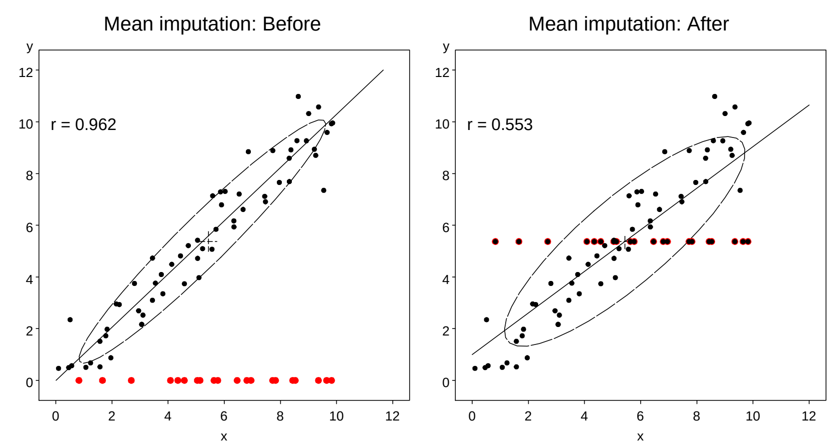
Output $\rightarrow$

```
                          -- imputed --     -- orig --
    MODEL          ORIGIN   REP77    REP78    R77  R78

AMC SPIRIT           A       3.20     3.41     .    .
BUICK OPEL           A       3.20     3.41     .    .
FORD FIESTA          A       3.20     4.00     .    4
MERC. MONARCH        A       3.20     3.00     .    3
PEUGEOT 604 SL       E       3.20     3.41     .    .
PLYM. HORIZON        A       3.20     3.00     .    3
PLYM. SAPPORO        A       3.20     3.41     .    .
PONT. PHOENIX        A       3.20     3.41     .    .
```


What's wrong with mean substitution?

- Problems:
  - Corrupts marginal distribution of each imputed variable: s^2 too small ($\bar{x}$ OK)
  - Corrupts covariances and correlations with other variables: $|r|$ too small



Single Imputation: Conditional means

- Conditional means (by region of origin)

```
proc sort data=auto;
   by origin;
proc standard REPLACE;
   by origin;
   var rep77 rep78;
```

Output →

MODEL	ORIGIN	REP77 (-- imputed --)	REP78	R77 (-- orig --)	R78
AMC SPIRIT	A	2.98	3.02	.	.
BUICK OPEL	A	2.98	3.02	.	.
FORD FIESTA	A	2.98	4.00	.	4
MERC. MONARCH	A	2.98	3.00	.	3
PLYM. HORIZON	A	2.98	3.00	.	3
PLYM. SAPPORO	A	2.98	3.02	.	.
PONT. PHOENIX	A	2.98	3.02	.	.
PEUGEOT 604 SL	E	2.90	4.00	.	.

- Better than filling in grand means
- In general, use demographic (age, gender, ...) or other variables to form homogeneous subgroups

Single Imputation: Regression estimates

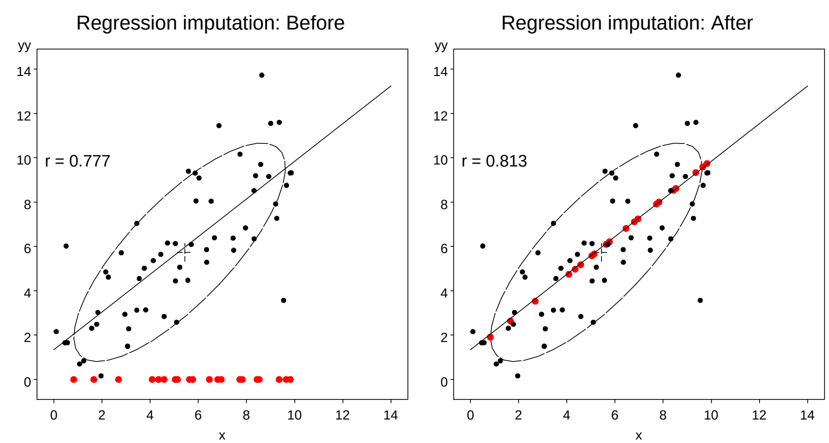
- Regression estimates
  - Include vars with missing as dependents
  - Replace missing values with predicted values from regression
  - PROC REG or lm(): Output data set with predicted values

```
proc reg data=auto;
   model rep77 rep78 = price mpg hroom rseat trunk
                       weight length turn displa gratio;
   output out=newauto p=p77 p78;
```

→ (not quite right; non-missings are replaced too!)

MODEL	ORIGIN	P77	P78	REP77	REP78
AMC SPIRIT	A	3.85	4.38	.	.
BUICK OPEL	A	3.76	3.58	.	.
FORD FIESTA	A	2.76	3.70	.	4
MERC. MONARCH	A	3.00	3.06	.	3
PLYM. HORIZON	A	3.42	4.34	.	3
PLYM. SAPPORO	A	4.07	3.87	.	.
PONT. PHOENIX	A	3.08	2.86	.	.
PEUGEOT 604 SL	E	3.31	4.20	.	.

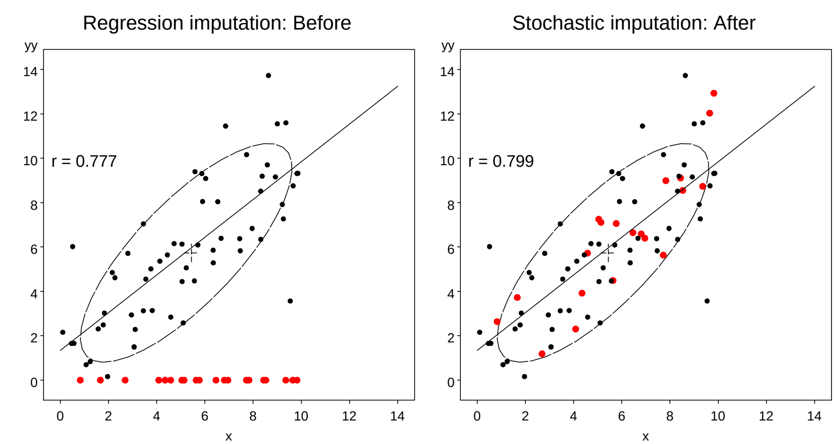
- Problems:
  - Inflates covariances and correlations with other variables



- Inflates covariances and correlations with other variables
- Makes results seem more precise than they are

Single Imputation: Stochastic imputation

- Stochastic imputation
  - Use regression estimate, plus a random quantity, $N(0, \sigma^2)$
  - Counteracts upward bias on correlations



Single Imputation: Cluster-based

- Cluster-based imputation: `PROC FASTCLUS`
  - Assign observations to clusters
  - Fill in cluster means for missing observations

$\rightarrow$

```
proc fastclus IMPUTE data=auto out=auto2
     maxclusters=10 delete=1 summary;
   var price -- gratio;
   id model;
```

```
    MODEL     ORIGIN  REP77  REP78  CLUSTER  DIST  _IMPUTE_

AMC SPIRIT       A     2.5    2.7      1    702.05     2
BUICK OPEL       A     3.6    4.0      8    347.71     2
FORD FIESTA      A     3.6    4.0      8    291.73     1
MERC. MONARCH    A     2.5    3.0      1    408.48     1
PEUGEOT 604 SL   E     3.0    2.7      2    941.74     2
PLYM. HORIZON    A     3.6    3.0      8    274.28     1
PLYM. SAPPORO    A     3.9    4.4      4    411.99     2
PONT. PHOENIX    A     2.5    2.7      1    410.23     2
```


Problems with single imputation

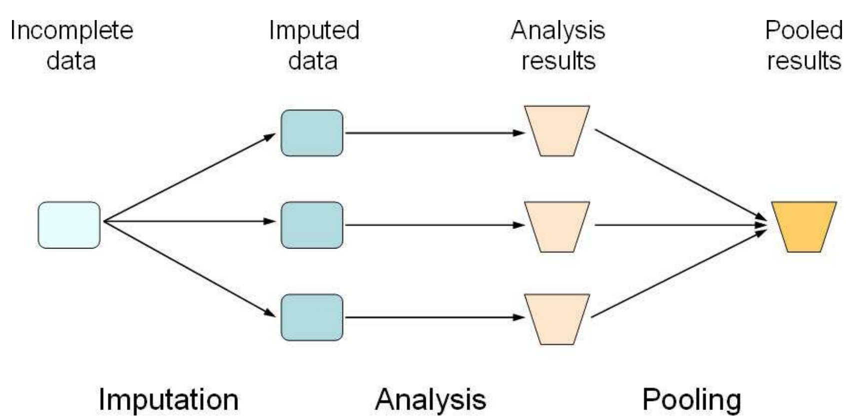
- Subsequent analyses do not reflect *missing data uncertainty*
  - sample size, N and degrees of freedom are overstated
  - confidence intervals too narrow
  - Type I error rates too high
- Problem gets worse as rate of missing and model complexity (number of parameters) increase
- Example:
  - 30% missing
  - One confidence interval (regression coeff., odds ratio, ...)

Nominal coverage (%)	90	95	99
Actual coverage (%)	77	85	94

  - Testing a 10-parameter H_0 (e.g., regression F-test)

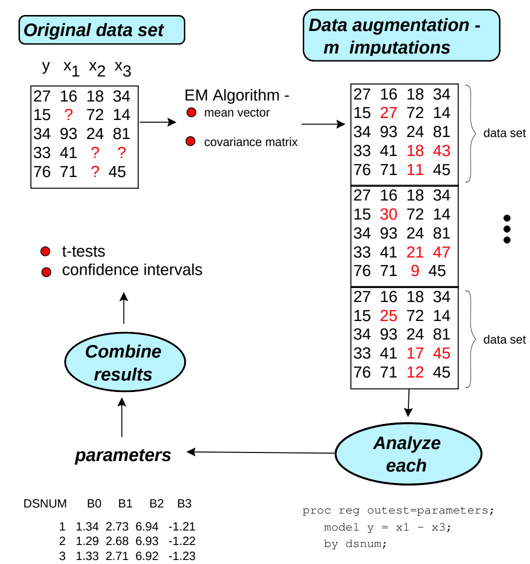
Nominal α	0.10	0.05	0.01
Actual α	0.57	0.45	0.25

Multiple imputation: Overview



- Imputation: Fill in missing entries m times, drawing from a distribution
- Analysis: Analyze each of the m completed data sets using *standard* complete-data techniques
- Pooling: Combine the m estimates into a final result, accounting for within- and between-imputation variance

Multiple imputation: Steps



Multiple imputation

- Obtaining valid inferences from imputed data: Little and Rubin (1987), Rubin (1987), Schafer (1997)
- Missing values replaced by $m > 1$ simulated versions, $(3 \leq m \leq 10)$.
- Each imputed complete dataset is analyzed by standard methods,
- Results combined to produce estimates and confidence intervals that incorporate missing-data uncertainty
- High efficiency, even for small m.

$$\text{Rel. Efficiency} = \left(1 + \frac{\gamma}{m}\right)^{-1}$$

where γ = rate of missing info (about a parameter)

	γ				
m	.1	.3	.5	.7	.9
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	95	93	96

See The Multiple Imputation FAQ page,
`http://www.stat.psu.edu/~jls/mifaq.html`

Multiple imputation: Combining estimates

Rubin (1987) method for MI inference, ***scalar quantities*** (θ)—

- m imputations $\rightarrow$ m estimates, $\hat{\theta}_i$, each with an estimated sampling variance $\widehat{var}(\theta_i)$.
- MI point estimates: average the m values of $\hat{\theta}_i$

$$\bar{\theta} = \frac{1}{m} \sum_i^m \hat{\theta}_i$$

- Proper tests and CI for imputed data must take into account:
  - ***Within-imputation variance***: average sampling variance of the m estimates.

$$\bar{W} = \sum \widehat{var}(\theta_i)/m$$

  - ***Between-imputation variance***: variability of the estimates across m imputations.

$$B = \sum (\hat{\theta}_i - \bar{\theta})^2/(m-1)$$

- These are combined to give the ***Total-imputation variance*** of $\bar{\theta}$,

$$T \equiv var(\bar{\theta}) = \bar{W} + (1 + \frac{1}{m})B$$

Multiple imputation: Significance tests and CI

- MI hypothesis tests: $t_{obs} = \bar{\theta}/\sqrt{T} \sim t_{df}$
- MI adjusted confidence interval: $\bar{\theta} \pm t_{df}\sqrt{T}$
- Degrees of freedom:

$$df = (m-1)\left(1 + \frac{m\bar{W}}{(m+1)B}\right)^2$$

- Fraction of missing info (γ), relative increase in variance due to nonresponse (r):

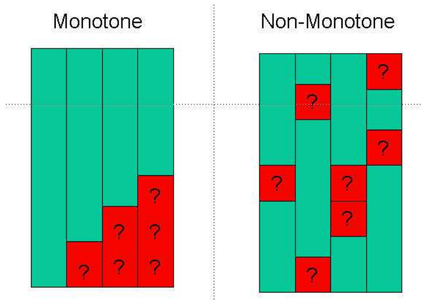
$$\gamma = \frac{r + 2/(df+3)}{r+1} \qquad r = \frac{T - \bar{W}}{\bar{W}} = \frac{(1+m^{-1})B}{\bar{W}}$$

Multiple imputation: Software

- SAS:
  - SAS V8.1+: `PROC MI`, `PROC MIANALYZE` (`www.sas.com/rnd/app/papers/multipleimputation.pdf`)
  - The `miplot` macro for visualizing missing data and multiple imputations
- SPSS Statistics 17+
  - Missing Value Analysis (MVA) - listwise, pairwise, EM, Regression
  - Multiple Imputation $\rightarrow$ { Analyze Patterns, Impute Missing Data Values}
  - See: Help $\rightarrow$ Case studies $\rightarrow$ Missing values option
- R packages
  - mice - Multiple Imputation by Chained Equations (comprehensive!)
  - VIM - Visualizing and Imputation of Missing Values
  - mi - state-of-art bleeding edge methods
- Other software listed at `http://ssc.utexas.edu/consulting/answers/general/gen25.html`

Multiple imputation: `PROC MI` and `PROC MIANALYZE`

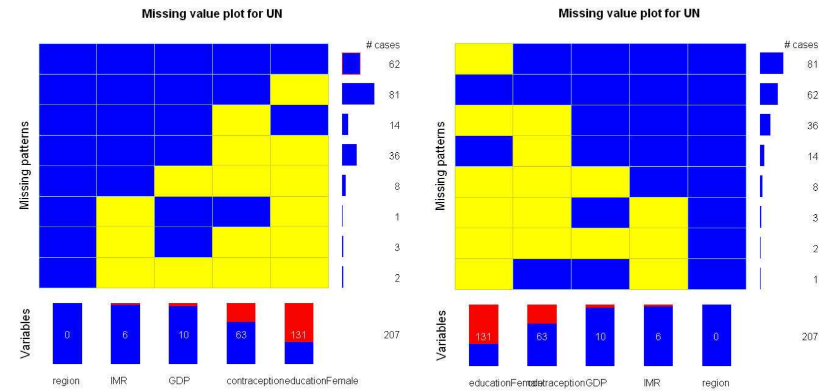
- `PROC MI` — Different methods for different missing patterns:



- Monotone missing data pattern: $Y_j = . \Rightarrow Y_k = ., \forall k > j$.
  - Parametric regression (assumes multivariate normality)
  - Non-parametric, propensity scores method
- Non-monotone missing data pattern
  - Markov chain monte carlo (MCMC) for all (comp. intensive) [default!]
  - MCMC $\rightarrow$ monotone pattern. Then, use monotone methods.
  - Predictive mean matching (PMM) now more widely used
- Generates m imputed data sets, with index variable `_Imputation_`.

Missing data patterns: Example

Data from United Nations (Fox, 2008) on Infant mortality in relation to GDP, contraception and female education, visualized by missing data patterns



- For the variables of interest, there is no permutation that gives a monotone pattern
- Therefore, the more general MCMC methods must be used.

Multiple imputation: `PROC MI` and `PROC MIANALYZE`

- *Any* analysis step BY `_Imputation_`, producing estimates. (`PROC REG`, `PROC GLM`, `PROC MIXED`, `PROC GENMOD`, etc.)

```
proc reg data=outmi outest=outreg covout noprint;
   model Oxygen = RunTime RunPulse;
    by _Imputation_;
```

- `PROC MIANALYZE`
  - Combines estimates, á la Rubin (1987), Schafer (1997)
  - Provides both univariate (t) and multivariate (F) tests.
- R: `mice` package is analogous; `mi` package is state-of-art, but more complex
  - `mice()`: generates multiple imputation data set; many imputation methods.
  - `pool()`: pools multiple imputation estimates

Baseball data: `PROC MI` and `PROC MIANALYZE`

- Salary and performance data for $n = 322$ players
- Salary missing for 18% of players (monotone pattern)
- Model: $\log(\text{salary}) \sim \min(\text{years}, 7)$ + trpc + batavgc
- MI assumes:
  - Variables are approx. *multivariate normal* (transform first if not)
  - Model used for imputation is consistent with the analysis model
  - i.e., preserves essential features of the data (e.g., ordinal variables, integer variables).

0. Screen and transform variables

Data screening and preliminary analysis (ignoring missing data) led us to transform `salary` $\rightarrow$ `log(salary)` and re-express `years` as linear up to 7, but flat thereafter `min(years, 7)`.

The primary consideration is that variables are modeled in the correct form (non-linearity corrected or taken into account).

basemi.sas ···

```
%include data(baseball);
*-- Screen/Transform variables;
data baseball;
    set baseball;
    if salary ^=.
        then logsal = log(salary);
    years7 = min(years,7);
    trpc = (runsc + rbic + homerc) / years;
    label logsal = 'log Salary'
        trpc='Total career runs/year'
        years7='Years, up to 7';
```


1. Generate m imputed data sets

··· basemi.sas ···

```
title 'Proc MI: Regression method (monotone)';
proc mi data=baseball seed=42424241 out=basemi;
    monotone method=regression;
    var years7 trpc batavgc logsal;
    run;
```

Notes:

- Variables *must* be listed so that missing pattern is monotone (`logsal` last)
- $\rightarrow$ generates $m = 5$ copies with imputed values for `logsal`.

Printed output:

```
                          The MI Procedure
                         Model Information

     Data Set                               WORK.BASEBALL
     Method                                 Regression
     Number of Imputations                  5
     Seed for random number generator       42424241

                        Missing Data Patterns

 Group     years7     trpc     batavgc     logsal        Freq      Percent

     1     X          X        X           X              263        81.68
     2     X          X        X           .               59        18.32

                       Missing Data Patterns
         -------------------------Group Means-----------------------
Group           years7            trpc           batavgc             logsal

    1         5.224335       94.246679        262.794677           5.927417
    2         5.237288       73.385852        255.474576                  .
```

Notes:

- Are there differences in means among different missing patterns?
- NB: large difference for `trp` — why?
- Is there evidence that data is not MAR?

2. Analyze m complete data sets

- Use output dataset from `PROC MI` as input to `PROC REG`
- Obtain output dataset containing parameter estimates (and covariance matrices)
- Use `by _Imputation_;` to repeat analysis m times

··· basemi.sas ···

```
proc reg data=basemi noprint outest=outreg covout;
   model logsal = years7 trpc batavgc;
   by _Imputation_;
   run;
```

Print parameter estimates:

··· basemi.sas ···

```
proc print data=outreg;
   id _Imputation_;
   by _Imputation_;
   where (_Type_ = 'PARMS');
   var Intercept years7 trpc batavgc;
   title2 'Parameter estimates from imputed data sets';
   run;
```

Output:

Parameter estimates from imputed data sets

Imputation	Intercept	years7	trpc	batavgc
1	2.62497	0.25947	.007388496	.004761202
2	2.56478	0.25190	.007207141	.005198336
3	2.48735	0.25515	.006764982	.005615983
4	2.76897	0.25300	.008008711	.004010438
5	3.26130	0.25139	.008340637	.002135631

3. Combine results with PROC MIANALYZE

- Use output dataset from PROC REG as input to PROC MIANALYZE

··· basemi.sas ···

```
title 'Proc MIANALYZE to combine and test';
proc mianalyze data=outreg mult edf=318;
    var Intercept years7 trpc batavgc;
    run;
```

Printed output:

The MIANALYZE Procedure

Multiple Imputation Variance Information

Parameter	Variance: Between	Variance: Within	Variance: Total	DF
Intercept	0.095088	0.095850	0.209955	12.38
years7	0.000010826	0.000219	0.000232	241.56
trpc	0.000000399	0.000000537	0.000001015	16.254
batavgc	0.000001878	0.000001779	0.000004032	11.73

Parameter estimates, standard errors and CI:

Multiple Imputation Parameter Estimates

Parameter	Estimate	Std Error	95% Confidence Limits		DF
Intercept	2.741474	0.458209	1.74651	3.736435	12.38
years7	0.254181	0.015215	0.22421	0.284153	241.56
trpc	0.007542	0.001008	0.00541	0.009675	16.254
batavgc	0.004344	0.002008	-0.00004	0.008730	11.73

In contrast to complete case analysis (throws away any missing data) or single imputation:

- estimated coefficients are unbiased, if missing salary is ignorable (MCAR or MAR),
- standard errors & CIs typically smaller than complete case analysis, and reflect uncertainty due to imputation.

$$SE_{\text{single impute}} \leq SE_{\text{MI}} \leq SE_{\text{complete case}}$$

- like Goldilocks, this is just about right!

3. Combine results with PROC MIANALYZE

Individual hypothesis tests ($H_0 : \theta_i = 0$):

Multiple Imputation Parameter Estimates

Parameter	Theta0	t for H0: Parameter=Theta0	Pr > |t|
Intercept	0	5.98	<.0001
years7	0	16.71	<.0001
trpc	0	7.49	<.0001
batavgc	0	2.16	0.0519

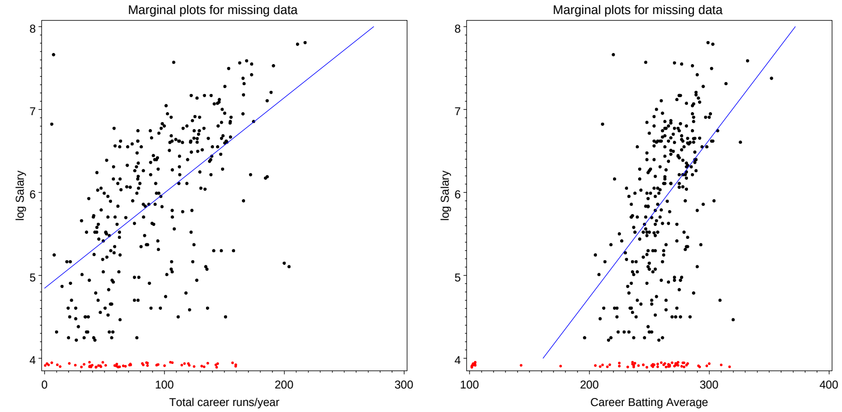
Multivariate hypothesis test ($H_0 : \theta_1 = \theta_2 = \cdots = 0$):

Multiple Imputation Multivariate Inference
Assuming Proportionality of Between/Within Covariance Matrices

Avg Relative Increase in Variance	Num DF	Den DF	F for H0: Parameter=Theta0	Pr > F
0.502386	4	94.202	7499.13	<.0001

Plots for missing data

- Marginal plots
  - Ordinary bivariate plots ignore all missing observations
  - Instead, show missing observations as marginal points
  - Are the missing observations consistent with the marginal distributions? (weak test of MAR)
- Example: Baseball data, marginal plots for Career runs/year (`trpc`) and Career batting average (`batavgc`)
  - Only `logsal` is missing here.
  - Missing observations shown at margins (red).
  - For illustration, $\sim 10\%$ of observations with missing salary also had `batavgc` set to missing.



What we're looking for here is any indication that the marginal distributions of the missing cases seems to differ systematically from those of the complete cases.

Plots for missing data

- Imputation plots
  - For missing observations, calculate typical value (mean, median) and variability (std., stderr, IQR) over the m imputed data sets.
  - Show fully observed data as points, missings as typical value (mean, median), with error bars for variability, over the m imputed data sets.
  - `miplot` macro: takes input data, imputed data $\rightarrow$ plot for (x, y); std. error bars for missing on one, diamonds for missing on both.

··· baseimplt.sas

```
proc mi data=baseball out=basemi nimpute=10;
    monotone method=regression;
    var years7 trpc batavgc logsal;
    run;

%miplot(data=baseball, imputed=basemi,
    x=trpc,    y=logsal, id=name);
%miplot(data=baseball, imputed=basemi,
    x=batavgc, y=logsal, id=name);
```


Plots:

These plots simply show the observed, non-missing data, together with the average for imputed data and variability over imputations.

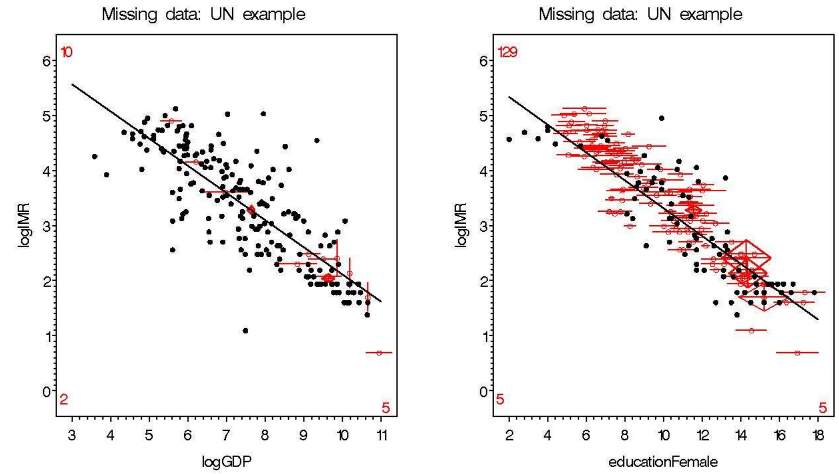
UN data: `PROC MI`, and miplot macro

- Fit a model predicting logIMR from logGDP, contraception and educationFemale
- Use multiple imputation to take account of missings (non-monotone: use MCMC)
- Plot the data, showing the uncertainty around each missing observation

··· UNmissing.sas ···

```
*-- Complete case analysis;
proc reg data=UN;
  model logIMR = logGDP contraception educationFemale;
  run;
*-- Run PROC MI, using MCMC method;
proc mi data=UN seed=21355417 nimpute=6 out=outmi;
  mcmc chain=multiple initial=em;
  var logIMR logGDP contraception educationFemale;
run;
%miplot(data=UN, imputed=outmi, id=country, y=logIMR, x=logGDP, interp=rl);
%miplot(data=UN, imputed=outmi, id=country, y=logIMR, x=educationFemale,
        interp=rl);
```


UN data: `PROC MI`, and `miplot` macro



- For logGDP, there is very little missing data
- For educationFemale, there is a lot

Missing data: Summary

- Make every effort to avoid missing data, or failing that, to understand why data is missing.
- Understand missing data mechanisms (MCAR, MAR, MNAR) and their implications
  - MCAR: unbiased with all methods
  - MAR: only unbiased with ML methods and multiple imputation
  - NMAR: possibly biased for all methods
- Avoid default methods (listwise deletion, pairwise deletion), unless fraction missing is quite small.
  - Avoid default fixups (mean imputation, etc.) where possible
- Use multiple imputation to take proper account of missings
- Plot the data, showing the uncertainty around each missing observation
- Sensitivity analysis: how do your *substantive* results depen on how you handled missing data?
  - Do complete case analysis
  - Do a better missing-data analysis
  - Compare substantive conclusions, decide how to report

References

Fox, J. *Applied Regression Analysis and Generalized Linear Models*. Sage Publications, Thousand Oaks, CA, 2 edition, 2008.

Little, R. J. A. and Rubin, D. B. *Statistical Analysis with Missing Data*. John Wiley and Sons, New York, 1987.

Rubin, D. B. *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons, New York, 1987.

Schafer, J. L. *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London, 1997.