Deloitte.

Missing In Action

Strategies for Handling Missing Data

CAS RPM Seminar
Las Vegas
March, 2009

Mike Greene
Jim Guszcza
Deloitte Consulting

Antitrust Notice

The Casualty Actuarial Society is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.

Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding – expressed or implied – that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.

It is the responsibility of all seminar participants to be aware of antitrust regulations, to prevent any written or verbal discussions that appear to violate these laws, and to adhere in every respect to the CAS antitrust compliance policy.

What is the Problem?

- Insurance data is often incomplete, inconsistently coded, and generally "dirty".

- In particular: when building a predictive model – or performing any type of statistical analysis – one must decide how to handle observations containing missing information.

- Key considerations
  - Potential losses of predictive power
  - Potentially biased parameter estimates
  - "Overly optimistic" measures of confidence in parameter estimates
  - Implementation

Perspective

- Many commonly used techniques for handling missing data are considered "unacceptable" in textbook presentations.

- Twofold goal:
  - Better understand why common techniques are considered "unacceptable".
  - Better understand the practical advantages and disadvantages of various methods.

- Our goal here is not to advocate one approach above others.
  - Practical decisions are likely to be context-dependent.
  - We will give a survey - no attempt at completeness!

- "The only really good solution to the missing data problem is not to have any." --Paul Allison

Topics

- Background Concepts
- Traditional Methods
- Maximum Likelihood
- Expectation Maximization
- Multiple Imputation
- Case Studies

Deloitte.

Background Concepts

Missing Completely At Random
Missing At Random
Not Missing At Random

Missing Completely at Random [MCAR]

- Prob(*Y* is missing | *X*,*Y*) = Prob(*Y* is missing)
- Missingness of *Y* depends neither on the value of *Y* nor on the values of any other variables in one's data.
  - There is rhyme or reason to why *Y* is missing.
  - The subpopulation for which *Y* is not missing is a random sample of the full population.
- This is the least bad type of "missingness".
- Unfortunately MCAR is also a fairly strong assumption.

Missing at Random [MAR]

- Prob(*Y* is missing | *X*,*Y*) = Prob(*Y* is missing | *X*)
- Example:
  - Prob(credit_m | age,credit) = Prob(credit_m | age)
- Probability of missing credit is dependent on age…
- …but within a given age, the probability of missing credit is not dependent on the value of credit.
- More realistic than MCAR
- Note: we can test whether (Y is missing) depends on X
- But we can't test whether (Y is missing) depends on Y

Ignorability

- Ignorable: the parameters to be estimated are independent of the parameters governing the missing data mechanism.
- Ignorable example: a tape containing the credit score for a random sample of historical policies was destroyed.
- Non-ignorable example: a tape containing the credit score for premier-tier policies was destroyed.
- If ignorability is violated than textbook methods for handing missing data don't apply.

➔ For the purpose of this presentation we will ignore ignorability.
(i.e. assume ignorability is satisfied)

Not Missing At Random (aka Selection Bias)

- Prob(*Y* is missing | *X*,*Y*) ≠ Prob(*Y* is missing | *X*)
- Example: people with poor credit are more likely to have "thin files" or missing credit reports altogether.
- i.e. the probability of credit being missing depends on the value of credit itself.
- In other words, the sub-sample of policies for which credit is not missing is a "biased sample" of the overall population.
    - The probability of missing credit is not independent of the quality of a policyholder's credit.

In Summary

- Missing Completely at Random [MCAR]
  - **Prob(*Y* is missing | *X*,*Y*) = Prob(*Y* is missing)**
  - The missingness of *Y* is independent of everything in your data.
  - I.e. no particular reason why a given observation is missing.
  - I.e. the observations with non-missing *Y* are a random sample of the total population.

- Missing at Random [MAR]
  - **Prob(*Y* is missing | *X*,*Y*) = Prob(*Y* is missing | *X*)**
  - The probability that *Y* is missing is independent of the value of *Y*, conditioned on the other variables *X* in your data.

- Not Missing at Random [NMAR]
  - **Prob(*Y* is missing | *X*,*Y*) ≠ Prob(*Y* is missing | *X*)**
  - The probability that *Y* is missing is dependent on the value of *Y* itself.
  - i.e. we face *selection bias*: the observations for which *Y* is observed are a biased sample of the total population.

What to Do?

- MCAR and MAR: broadly speaking, 2 types of options
  - Traditional methods
    - Delete observations with missing data
    - Add dummy variables to flag missing observations
    - Impute missing values
  - Newer methods
    - Maximum likelihood, facilitated by the EM algorithm
    - Multiple Imputation [MI]

- NMAR
  - "Off the shelf" textbook methods no longer apply
  - We need to *model* the process that generated missing data
  - i.e. we need to account for selection bias in our model
  - Classic example: the Heckman adjustment

Deloitte.

Traditional Methods

Listwise deletion
Missing value dummy variables
Simple imputation

Traditional Method #1: Complete Case Analysis

- The most obvious strategy for handling missing values is to simply disregard all observations for which any variable is missing.
  - This is the default option in SAS and some R functions

- Advantages
  - Simple
  - Works with any statistical procedure
  - MCAR ➔ unbiased estimates

- Disadvantages
  - Inefficient use of data
  - Will lead to biased results if data isn't MAR
  - In an extreme case, you could be left with (virtually) no data
  - Imagine a scenario in which X_1 is missing on the odd observations and X_2 is missing on the even observations.
    ➔ you'd throw out all of your data

Traditional Method #2: Missing Value Dummies

- For each variable with missing observations, create a separate {0,1} dummy variable indicating the presence of missing data.
  - Notation: *X*_m = 1 iff *X* is missing

- Example: Suppose we are regressing claim frequency on AGE, CREDIT, and TENURE
  - Suppose CREDIT is MCAR 50% of the time
  - (e.g. we wanted to save money by ordering credit on only half the population)
  - If we include CREDIT_m as well as CREDIT in the model, we can recode missing values of CREDIT to *any* number.
  - intuition: including CREDIT_m means that the model contains a full degree of freedom to accommodate missing values of CREDIT.
  - β_{CREDIT_m} adjusts to the value used to impute missing values of CREDIT.
    - ➔ β_{CREDIT} is unaffected.

- Big problem: in general this leads to biased β estimates!
  - We will come back to this

Traditional Method #3: Mean Imputation

- Marginal Mean Imputation
  - Impute missing values of *X* with the mean or median value of the non-missing values of *X*.
  - The most commonly used technique
- Pro: simple, intuitive
- Con: produces potentially biased parameter estimates
- Con: standard error estimates are often biased downward
  - Why: a regression model can't tell the difference between a dataset that has a large number of values bunched up at the mean and a dataset for which missing values have been replaced with the mean.
  - Mean-imputation gives the appearance of having more information about β_X than we really have.
- Textbook advice: avoid this method.

Simulated Example

- Use of mean imputation and dummy variables are fairly widespread, so let's explore the potential pitfalls.

- Simulation: simulate 10,000 draws from a bivariate normal distribution.

$$\begin{pmatrix} X \\ Z \end{pmatrix} \sim N(0, \Sigma) \quad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

- We can vary ρ to understand the effect of multicollinearity

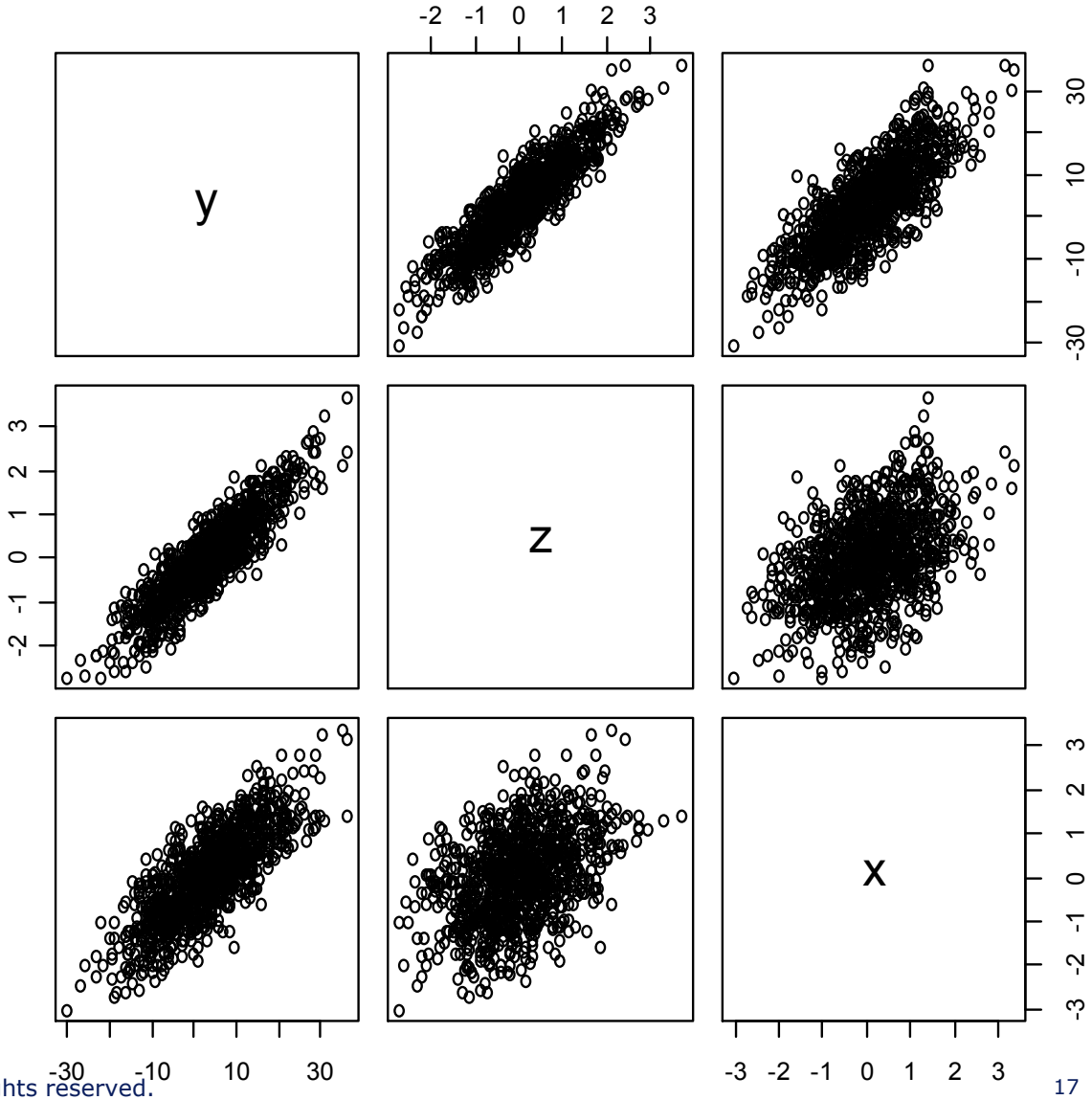
- Next let: $Y = 3 + 5*X + 7*Z + \varepsilon$ where $\varepsilon \sim N(0,1)$

- Finally assign missing values to a randomly selected half of the values of *Z*.
    - Values of *Z* are Missing Completely at random [MCAR]

Simulate 1000 Data Points

Y = 3 + 5*X + 7*Z + e cor(X,Z)=0.4

- Let's assume that
  $\rho_{X,Z} \equiv \text{cor}(X,Z) = 0.4$
- Illustrate 3 methods
  - Listwise deletion
  - Dummy variable
  - Mean imputation




```
> round(cor(dat),2)
     y    z    x
y 1.00 0.89 0.78
z 0.89 1.00 0.43
x 0.78 0.43 1.00
```

Live Demo: Common Ways of Handling Missing Data

- We will select a random 50% of the *Z* values in our data and blank them out.
  - These values of *Z* are "missing completely at random" [MCAR]
  - The most innocuous form of missingness

- Next illustrate the effects of traditional methods #1,2,3
  - Method1: Y ~ X + Z (on complete cases)
  - Method2: Y ~ X + Z + Z.m (on all cases)
  - Method3: Y ~ X + Z* (on all cases)

  Where Z* = mean(Z) if missing; Z*=Z otherwise

Only Method #1 yields unbiased β estimates

```
> round(coef(t1),3)
(Intercept)           x           z
      2.983       4.982       7.018
> round(coef(t2),3)
(Intercept)           x           z         z.m
      2.907       6.582       6.346   63449.427
> round(coef(t3),3)
(Intercept)           x           z
      2.989       6.582       6.346
```

From Whence the Bias?

- Here is a clue: after we recode the missing values of Z to the mean value of Z, the correlation structure of the data is changed.

- Original data:

  (Before missing data was introduced)

```
> doSim(); corr(dat)
     y    x    z
y 1.00 0.78 0.89
x 0.78 1.00 0.43
z 0.89 0.43 1.00
```

- After deletion of missing cases:

```
> dat$z[miss==1] <- NA
> corr(dat)
     y    x    z
y 1.00 0.78 0.90
x 0.78 1.00 0.43
z 0.90 0.43 1.00
```

- After mean imputation:

```
> dat$z[miss==1] <- 0
> corr(dat)
     y    x    z
y 1.00 0.78 0.64
x 0.78 1.00 0.30
z 0.64 0.30 1.00
```

- Mean imputation is an inherently "univariate" activity.
- It does not preserve the correlation structure of the data.
- Regression parameters are derived from these correlations.

Where Do We Stand?

- The traditional methods of dummy variable indicators and mean imputation lead to biased parameters even in the "least bad" case of data that is missing completely at random [MCAR].
    - But – consider how serious the bias is likely to be and how damaging the bias would be to your business outcome.

- Listwise deletion is better behaved but still no panacea
    - Leads to a loss of "power" (higher standard errors).
    - In many practical situations would cause you to throw away a majority of your data.
    - Assumes that the missing cases are an unbiased sample of the total population.

➔ It makes sense to explore other strategies for handling missing data.

No Mean Feat

- Mean imputation and the dummy variable techniques are "myopic": they impute missing values without regard to how this affects the correlation structure of the data.
- More advanced methods take the opposite approach.
- They attempt to impute missing values in a way that preserves the correlation structure of the data.
    - Maximum Likelihood
    - Expectation-Maximization Algorithm
    - Multiple Imputation

Deloitte.

Maximum Likelihood Methods

Monotonicity
EM Algorithm

Maximum Likelihood Refresher

- Remember what maximum likelihood means.
- Suppose our data $\{y,\ldots,y_n\}$ is iid distributed with pdf $f(y_i|\theta)$.
- Likelihood function: $$L(\theta) = \prod_{i=1}^{n} f(y_i \mid \theta)$$
- We find the value of θ that maximizes $L(\theta)$: θ_{MLE}
- Properties of θ_{MLE}:
  - **Consistent:** θ_{MLE} is approximately unbiased in large samples
  - **Asymptotically efficient:** In the limit as $n\rightarrow\infty$ the standard error of θ_{MLE} is at least as small as s.e. for any other consistent estimator.
  - **Asymptotically normal:** In the limit as $n\rightarrow\infty$ θ_{MLE} has an approximately normal distribution.

Maximum Likelihood With Missing Data

- Illustration: suppose that we have n complete observations of y but observations $m+1$, $m+2$, …, n are missing for x.

y_1	$y_2 \cdots$	y_m	$y_{m+1} \cdots$	y_n
x_1	$x_2 \cdots$	x_m	NA…	NA

- Basic idea: when setting up the likelihood function, integrate across the missing values of x:

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

where

$$g(y \mid \theta) = \sum_{x} f(x, y \mid \theta)$$

Monotnicity

- The likelihood "factorization" described on the previous page only works if the data's "missingness" follows a monotonic pattern.
- Example:

Monotonic: data are missing for X_i ➔ they are also missing for Xj with $j>i$.

monotonic					non-monotonic				
obs	**X1**	**X2**	**X3**	**X4**	**obs**	**X1**	**X2**	**X3**	**X4**
1	-56	122	-107	43	1	-56	122	.	.
2	-23	36	-22	-30	2	.	.	-22	-30
3	156	40	-103	.	3	156	40	-103	90
4	7	11	-73	.	4	.	.	.	.
5	13	-56	.	.	5	13	-56	-63	.
6	172	179	.	.	6	172	179	.	69
7	46	50	.	.	7	.	.	.	.
8	-127	.	.	.	8	.	-197	.	.
9	-69	.	.	.	9	.	70	-114	.
10	-45	.	.	.	10	-45	.	125	-38

- This limits the practical applicability of the maximum likelihood approach.

The Expectation-Maximization Algorithm

- The EM algorithm: an approach to maximum likelihood estimation.
- As with maximum likelihood we need to make a distributional assumption about our data.
- e.g.: $(X_1, X_2, X_3, X_4) \sim \text{MVN}(\mu, \Sigma)$
- Failure of monotonicity ➔ hard to set up and maximize the likelihood function.
- EM: iterative 2-step process for approximating the MLE.
    - Expectation step
    - Maximization step

How EM Works

- Step 0: make an initial guess at the parameters θ of your model for the data.
- **E**xpectation: fill in the missing values given the current estimate of the unknown parameters θ.
- **M**aximization: re-estimate θ by maximizing the likelihood of the data in its currently imputed form.
- Repeat the E-step and M-step until you get convergence.

MEMEMEMEMEMEMEMEME

- Example: suppose our data contains only the single variable X.
  - Data: X={100, 80, 100, 110, 120, 90, NA , NA , NA , NA}
- Assume X is normally distributed ➔ μ_{MLE} = mean(X)
- Here the EM algorithm is simple, and quickly converges to what we know to be true.
  - Step 0: guess: μ = -9999
  - E-step: X={100, 80, 100, 110, 120, 90, -9999, -9999, -9999, -9999}
  - M-step: $\mu_{MLE} \approx$ -3940
  - E-step: X={100, 80, 100, 110, 120, 90, -3940, -3940, -3940 , -3940}
  - M-step: $\mu_{MLE} \approx$ -1516
  - E-step: X={100, 80, 100, 110, 120, 90, -1516, -1516, -1516, -1516}
  - …
  - E-step: X={100, 80, 100, 110, 120, 90, 100, 100, 100, 100}

More Realistic

- With multiple variables $\{X_1, X_2, X_3, X_4\}$ a common assumption is that the data is MVN(μ, Σ).

- Here the EM algorithm is simple, and quickly converges to what we know to be true.
  - Step 0: make initial guess: (μ_0, Σ_0)
  - E-step$_1$: using (μ_0, Σ_0) calculate regression coefficients for linear model that predicts
    - X_1 in terms of X_2, X_3, X_4
    - X_2 in terms of X_1, X_3, X_4
    - Etc
  - M-step$_1$: calculate MLE estimate (μ_1, Σ_1) using the data from E-step$_1$
  - Rinse, repeat
  - … (μ_2, Σ_2) , (μ_3, Σ_3) , (μ_4, Σ_4) , …
  - …continue until your estimate of (μ, Σ) converges.

EM Observations

- EM is a clever way of calculating maximum likelihood estimates
  - Does not require monotonicity in pattern of missing data
  - Does require a statistical model of one's data (often MVN)
- In a regression setting, the EM step would not distinguish between the predictive and target variables
  - All variables are used in EM routine
  - You should throw the target variable in along with the predictive variables.
- EM uses the correlation structure of the data to impute missing values.
  - Your current estimate of $\sum$ allows you to compute the regression parameters that relate X_i to all of the other variables in the data.

Deloitte.

Multiple Imputation

Motivation
Case Studies

Multiple Imputation

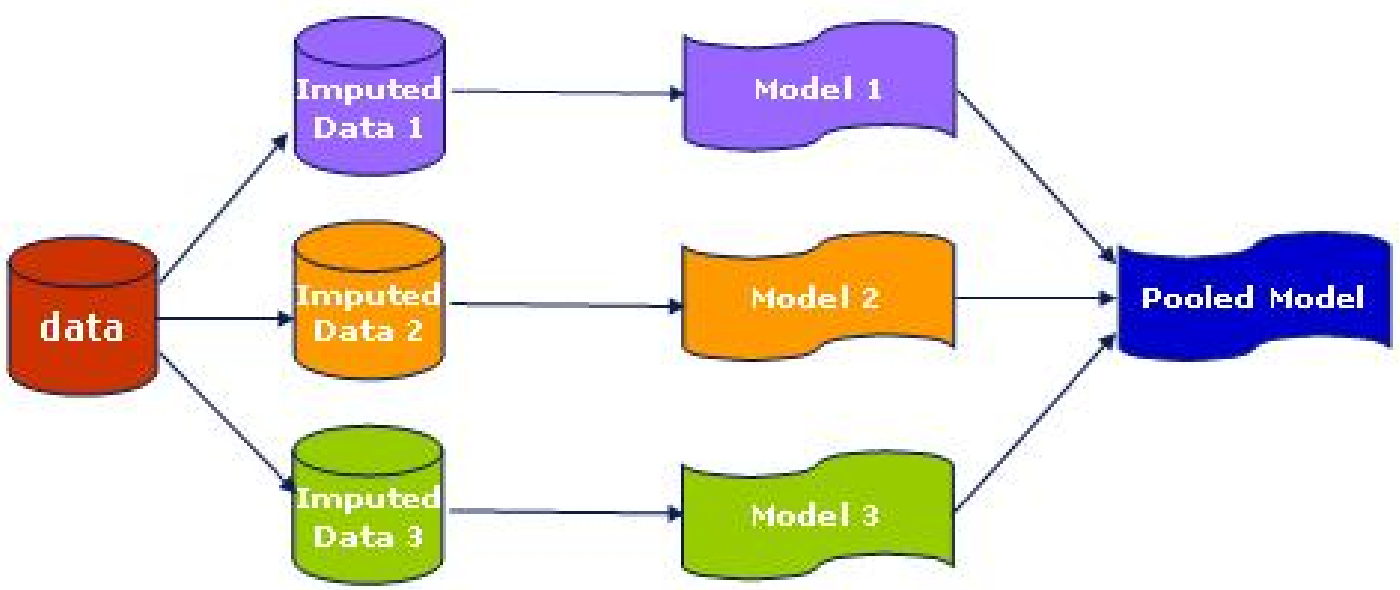
- EM is great, but in practice is hard to implement for anything more complex than traditional (log)linear models.

- Multiple Imputation has same advantages as EM
    - Consistency
    - Asymptotic efficiency
    - Asymptotic normality

- Further advantage of MI: can be used for any type of model (not just linear regression)

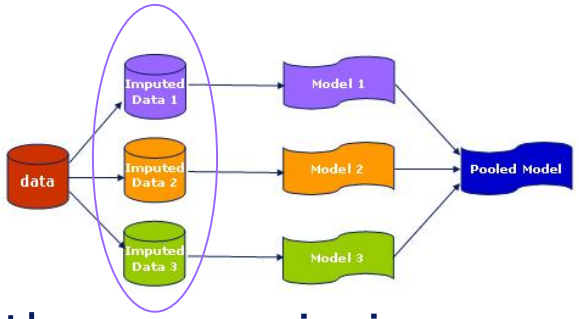
Multiple Imputation

- 3 major steps: imputation ➔ modeling ➔ pooling



- Start with incomplete data ➔ impute missing values multiple times ➔ perform your analysis on each imputed dataset ➔ pool results into final model.

Performing the Multiple Imputations

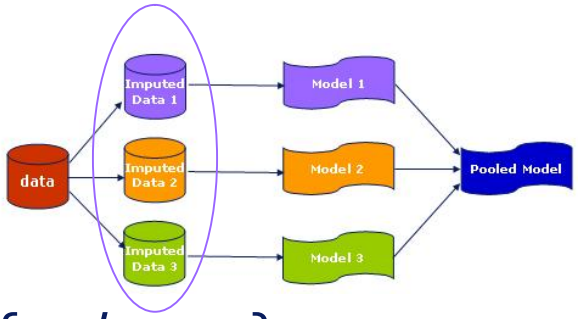


- Suppose we have two variables X and Y, with many missing values of X.
- We can estimate the missing values of X by using a regression equation: $x_i \approx a + by_i$
- Roughly speaking: each imputed value is computed using the formula

$$\tilde{x}_i = a + by_i + u_i \quad where \quad u_i \sim n(0, s_{x|y})$$

- $s_{x|y}$ is the estimate of the standard deviation of the regression of X on Y.
- Therefore each imputed value of X involved a random draw from the regression of X on Y.

Known Unknowns



- But this formula isn't quite right: it treats $\{a, b, s_{x|y}\}$ as known quantities from which we can randomly draw multiple imputed values of *X*.

$$\tilde{x}_i = a + by_i + u_i \quad where \quad u_i \sim n(0, s_{x|y})$$

- In actuality we don't know the values of these parameters.
  - We only have estimates.

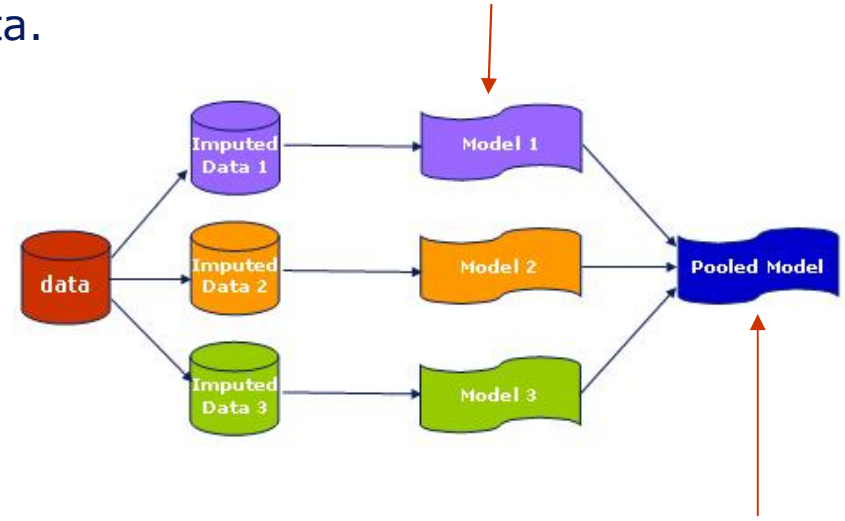
➔ it is best to draw a different set of parameters $\{a, b, s_{x|y}\}$ for each imputed dataset.

- These would be drawn from the Bayesian posterior distribution of $\{\alpha, \beta, \sigma_{x|y}\}$
  - The process of doing this is known as "data augmentation"

Pooling the Multiple Models

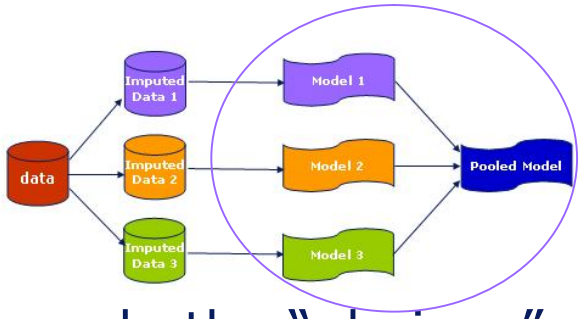
- Once we've created the imputed datasets (typically 5-10) we build our model or perform our analysis separately on each of the datasets.

- Note that we can perform any statistical analysis on the multiply imputed datasets.
  - Regression, GLM, Quantile Regression, Principal Components, …
  - On each of the imputed datasets we proceed as if there were no missing data.



- Next we pool the results into a single model.

Pooling the Multiple Models



- To calculate the pooled model parameters, we do the "obvious" thing:

$$\hat{\theta} = \frac{1}{M}\sum_{k=1}^{M}\hat{\theta}_k$$

- To calculate standard errors we must take into account both within variation and between variation:

$$s.e.(\hat{\theta}) = \sqrt{\hat{\sigma}^2_{within} + \left(1+\frac{1}{M}\right)\hat{\sigma}^2_{between}}$$

$$s.e.(\hat{\theta}) = \sqrt{\frac{1}{M}\sum_{k=1}^{M}s_k^2 + \left(1+\frac{1}{M}\right)\left(\frac{1}{M-1}\right)\sum_{k=1}^{M}\left(\hat{\theta}_k - \bar{\hat{\theta}}\right)^2}$$

- These two formulas are generic, and apply to the parameters of any models fit on multiple imputed datasets.

Multiple Imputation in Practice

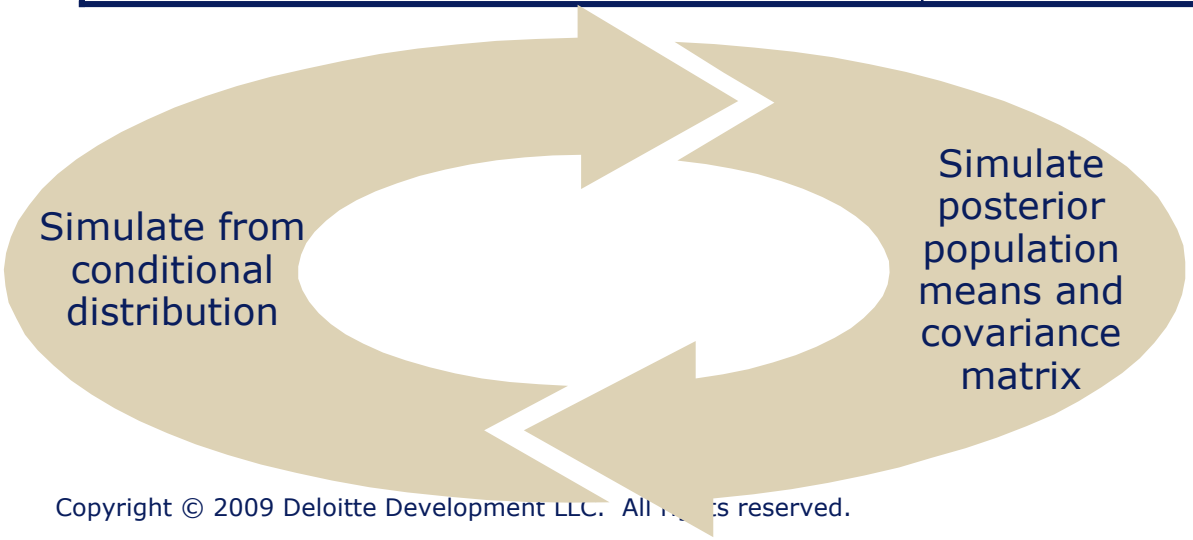
- SAS has many MI algorithms integrated into **PROC MI**

Method	Description
Regression Method	Parameters for regression models are simulated using Bayesian inference to introduce appropriate variation
Markov Chain Monte Carlo (MCMC)	Utilize a Markov Chain and Bayesian inference to impute missing values
Propensity Score Method	Propensity scores of missing values combined with Bayesian bootstrapping to make random draws from similar observations
Predictive Means Matching (PMM)	Similar to the Regression Method above, but randomly select from 'similar' cases for random draws

Markov Chain Monte Carlo (MCMC)

- MCMC utilizes a long chain of simulated variables, drawing from the conditional covariance matrix to impute missing values

Pros	Cons
Does not require monotonicity	Computationally intensive
Can utilize informative priors to get quicker convergence	Convergence may not be guaranteed
Can be used to produce monotonicity	"Reasonable" values may not be obtained – MVN assumption



$$p(\boldsymbol{\theta}|y) = \frac{p(y|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(y|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

Propensity Scoring

- Calculate propensity of missingness for an observation
- Group observations by propensity
- Bayesian bootstrapping to generate space of missing values by group
- Impute via random sampling from missing value space

Pros	Cons
Generates values only from within the known distribution	Does not utilize correlations among covariates

Propensity Scoring is not suitable for imputing values on predictive variables for use in regression analyses!

Predictive Means Matching

- Build a regression model for missing values on a variable (requires monotonicity)
- Draw from the MVN distribution of parameters
- Generate predictive values on all cases
- To impute, find cases with *similar* predicted values and randomly draw one, assigning the observed values

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} \rightarrow \vec{\sigma}_{MVN} \begin{bmatrix} \beta'_0 \\ \beta'_1 \\ \beta'_2 \\ \beta'_3 \\ \beta'_4 \end{bmatrix}$$

Obs.	Var1	Var1_HAT
1	9	10.1
2	.	10.2
3	12	10.3
4	.	10.4
5	11	10.5
…	…	…

Case Study #2 – MI In Practice

References

We followed the presentation of:

- Allison, P. D. (2002) Missing Data, Sage Publications.

The EM Algorithm came from:

- Dempster, A. P., Laird, N. M., Rubin, D. B. (1977) "Maximum Likelihood Estimation from Incomplete Data Using the EM Algorithm," *Journal of the Royal Statistical Society Series B.*

And Multiple Imputation came from:

- Rubin, D. B. (1987) Multiple Imputation for Nonresponse in Surveys, Wiley.