Much ado about nothing: methods and implementations to estimate incomplete data regression models

Nicholas J. Horton

Smith College, Northampton, MA, USA and University of Auckland, New Zealand

December 6, 2007, Australasian Biometrics Conference

nhorton@email.smith.edu
http://www.math.smith.edu/~nhorton/muchado.pdf
joint work with Ken P. Kleinman, Department of Ambulatory Care Policy, Harvard Medical School

partial funding support from NIH MH54693
Introduction and motivation

- missing data a common problem
- may be due to design or happenstance
- ignoring missing data may lead to inefficiency
- ignoring missing data may lead to bias
Introduction and motivation

1. many developments in methodology for incomplete data settings
2. software to fit incomplete data regression models is improving (but not yet entirely there!)
3. these methods need to be more widely utilized in practice
What missing data methods are used in practice?

1. Burton and Altman (BJC, 2004), review of missing covariates in 100 cancer prognostic papers

2. Horton and Switzer (NEJM, 2005), missing data methods in the *Journal*
Burton and Altman review of 100 papers (BJC, 2004)

<table>
<thead>
<tr>
<th>APPROACH</th>
<th># PAPERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>no missing or unclear</td>
<td>6</td>
</tr>
<tr>
<td>complete data entry criteria</td>
<td>13</td>
</tr>
<tr>
<td>missing data were reported</td>
<td>81</td>
</tr>
</tbody>
</table>

Nicholas J. Horton
Much ado about nothing
Papers reporting methods
(n=32, subset of 81)

<table>
<thead>
<tr>
<th>APPROACH</th>
<th># PAPERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>available case</td>
<td>12</td>
</tr>
<tr>
<td>complete case</td>
<td>12</td>
</tr>
<tr>
<td>omit predictors</td>
<td>6</td>
</tr>
<tr>
<td>missing indicator</td>
<td>3</td>
</tr>
<tr>
<td>ad-hoc imputation</td>
<td>3</td>
</tr>
<tr>
<td>multiple imputation</td>
<td>1</td>
</tr>
</tbody>
</table>
26 original articles in the *NEJM* (January 2004–June 2005) reported use of missing data methods

<table>
<thead>
<tr>
<th>APPROACH</th>
<th># PAPERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>last value carried forward</td>
<td>12</td>
</tr>
<tr>
<td>mean imputation</td>
<td>13</td>
</tr>
<tr>
<td>sensitivity analysis</td>
<td>2</td>
</tr>
<tr>
<td>multiple imputation</td>
<td>2</td>
</tr>
</tbody>
</table>
We are concerned that very few authors have considered the impact of missing covariate data; it seems that missing data is generally either not recognised as an issue or considered a nuisance that is best hidden. (p.6)
Barriers to use

- methods not well developed (not so true anymore)
- little easy to use software (still somewhat true, more later)
- word count limitations (online methods!)
- not perceived to be critical to a comprehensive analysis (quite common belief)
- no CONSORT equivalent (see Burton and Altman)
Burton and Altman (BJC, 2004) proposed guidelines

1. quantification of completeness of covariate data
   1. if availability of data is an exclusion criterion, specify the number of cases excluded for this reason,
   2. provide the total number of eligible cases and the number with complete data,
   3. report the frequency of missing data for every variable

2. exploration of the missing data
   1. discuss any known reasons for missing covariate data
   2. present the results of any comparisons of characteristics between the cases with or without missing data

3. approaches for handling missing covariate data
   1. provide sufficient details of the methods adopted
   2. give appropriate references for any imputation method used
   3. for each analysis, specify the number of cases included and the associated number of events
Goal

1. Assess the state of the art in general purpose statistical software to fit incomplete data regression models
2. Use a real-world health services dataset with complicated patterns of missingness
Health services motivating example

- Kids’ Inpatient Database (KID)
- developed by Healthcare Cost and Utilization Project (HCUP), Agency for Healthcare Research and Quality (AHRQ)
- Year 2000 dataset contains data from 27 State Inpatient Databases
- Inferential goal: Study predictors of routine discharge (as opposed to leaving AMA, transferring to another facility, or dying) among 10-20 year old subjects with a primary, secondary or tertiary diagnosis of mental health or substance abuse issues, what is predictive of being discharged from a hospitalization in a routine fashion
Predictors with complete data

- AGE (in years)
- LOS (length of stay, in days)
- NDX (number of medical diagnoses)
- WEEKEND (=1 if admitted on a weekend)
- FEMALE (=1 if female)
- OUTCOME (ROUTINE=1) is fully observed
Predictors with missing data

- RACE (1=Caucasian, 2=Black, 3=Hispanic, 4=Other)
- TOTCHG (Total charges, in dollars)
- SEASON (Winter, Spring, Summer, Fall)
- ATYPE (Admission type: 1=emergency, 2=urgent, 3=elective, 4=other)
- reasons for missingness?
- why season and not month?
10 variables, 135344 observations, 12 patterns
4 vars. (40%) have at least one missing value
55770 obs. (41%) have at least one missing value

Breakdown by variable
V 0 name Missing % missing
1 8 TOTCHG 5021 4
2 2 ATYPE 15093 11
3 10 SEASON 15616 12
4 7 RACE 21888 16
Missing data patterns (Splus missing data library)

1234 count
1 .... 79574 <- complete cases
2 ...m 21335 <- missing RACE
3 ..m. 15354 <- missing SEASON
4 .m.. 13601 <- missing ATYPE
5 m... 3665 <- missing TOTCHG
6 ..mm 213 <- missing SEASON + RACE
7 .m.m 234

11 mm.. 1213  (Note: decidedly non-monotone!)
Note: 21,335 subjects have everything observed except RACE
Pointers to the (extensive) literature

- excellent review by Ibrahim, Chen, Lipsitz and Herring (JASA 2005)
- provides a clear and comprehensive review of methods
- example involves only one variable with missing data!
Getting started

This page aims to provide a non-technical introduction to the issues involved in the analysis of datasets with missing observations. The material is extracted from our introductory missing data course (see events [/msu/missingdata/events.html] ). If it raises questions, please go to our frequently asked questions page in the first instance.
Pointers to the (extensive) literature (websites)

UCLA http://stat.ats.ucla.edu

UCLA Academic Technology Services
  Stat Computing  >  Textbook Examples

Missing Data
Paul Allison

This is one of the books available for loan from Academic Technology Services (see Stat Computing > Textbook Examples for other such books, and details about borrowing). We are grateful to Professor Allison for providing us with the data files for the book and for permission to distribute them along with programs showing how to replicate his results in a variety of packages. For more information about Professor Allison's work, see his web site at http://www.ssc.upenn.edu/~allison/. For more information about ordering the Missing Data book please see the Sage publications web site or see Where to buy books for tips on different places you can buy these books.
Introduction

The idea of multiple imputation is to create multiple imputed data sets for a data set for which some observations are missing at random (MAR). The analysis of a statistical model is then done on each of the multiple data sets, then combined to yield a set of results. In general, multiple imputation techniques are considered, since they allow the flexibility of no restriction on the distribution of the distribution of the missing data. However, the assumption of MAR is more restrictive and may not always be appropriate. There are two major approaches in multiple imputations. The first one is based...
Pointers to the (extensive) literature (Books)

- Little and Rubin (2nd edition)
- Schafer (1997)
- Allison (Sage)
- Molenberghs and Kenward (2007)
- Hogan and Daniels (sensitivity analysis, in press)
- Tsiatis (weighting)
- Carpenter monograph (forthcoming)
Pointers to the (extensive) literature (Review papers)

- *Multiple imputation: current perspectives*, Kenward and Carpenter, SMIMR 2007)
- *Multiple imputation review of theory, implementation and software*, Harel and Zhou (2007, SIM)
- *Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models*, Horton and Kleinman (2007, TAS)
Notation

- $Y$ outcome of regression model (univariate for our example)
- $X$ predictor in regression model (typically a vector, $X_1, X_2, \ldots, X_p$, mixed types of variables)
- $f(Y|X, \beta)$ regression model of interest
Introduction and motivation

Missing data models

Implementations

Results

Sensitivity analyses

Summary and discussion

Taxonomy and background

Maximum likelihood

Weighting approaches

Multiple imputation

Specifying the imputation model

Missing data nomenclature: mechanisms

- Introduced by Little and Rubin (text, 1987, 2002)
- Let $R = 1$ denote whether a particular variable (say $Y_2$) is observed in a longitudinal study
- What assumptions are we willing to make regarding the missingness law: $f(R|Y_1, Y_2, X, \gamma)$?
Missing data nomenclature: MCAR (Missing Completely at Random)

- $f(R | Y_1, Y_2, X) = f(R)$
- Missingness does not depend on observed or unobserved quantities
- Example: data fell from the truck
Missing data nomenclature: MAR (Missing at Random)

- \( f(R|Y_1, Y_2, X) = f(R|Y_1, X) \)
- Missingness does not depend on unobserved quantities
- Example: doctor took a subject off a longitudinal trial because they were too sick (based on observed \( Y_1 \))
- misleading name
Missing data nomenclature: NINR (Nonignorable nonresponse)

- \( f(R|Y_1, Y_2, X) = f(R|Y_1, Y_2, X) \) (no simplification)
- Missingness depends on unobserved quantities
- Example: subject missed their observation \( Y_2 \) because they were too sick to get out of bed

Note that \( R \) is a multinomial RV with 11 possible values for the KID dataset
Little and Rubin showed that if MAR missingness, then likelihood based approaches can *ignore* missing data mechanism and still yield the right answer.

MAR impossible to verify without auxiliary information.

NINR models require a lot of work modeling missingness, best used for sensitivity analyses.
Approaches for handling NINR (selection models)

\[ f(Y, R | X) = f(Y | X)f(R | Y, X) \]

(e.g. Diggle and Kenward, JRSS-C, 1994; Fitzmaurice, Laird and Zahner, JASA, 1996)

- fits complete data model for the outcomes \( f(Y|X) \)
- constraints on the non-response model need to be imposed
- identifiability can be problematic
- hard work (remember 11 patterns of missingness for KID study?)
Approaches for handling NINR (pattern-mixture models)

\[ f(Y, R \mid X) = f(R \mid X)f(Y \mid R, X) \]

(e.g. Little, *JASA*, 1993)

- \(f(Y \mid X)\) not modeled directly
- clearer assumptions to ensure identifiability (i.e. structure in conditional mean model includes no interactions between components of \(X\) and \(R\))
- even harder work
we focus on missing predictors (common problem)

- same nomenclature, but different implications in some settings (caveat emptor!)
- assume MAR for most methods
(Partial) taxonomy of missing data methods

- Complete case
- Ad-hoc methods
- Maximum likelihood methods (XMISS)
- Weighting methods
- Multiple imputation
Complete case/available case methods

- **Complete case**
  - Simple
  - Main drawback: inefficient (uses only 59% of the KID dataset!)
  - May yield bias

- **Available case**
  - will use different set of observations based on predictors in a particular model
  - models are not nested
  - difficult to describe
‘ad-hoc’ methods (not recommended)

- last value/observation carried forward (LVCF/LOCF)
- mean imputation
- missing indicator methods
- dropping a predictor from the model
Typically we are interested in \( f(Y|X, \beta) \) where the covariates are assumed fixed.

To gain information from partially observed subjects, posit a distribution for \( f(X|\alpha) \).

Maximize likelihood of \( f(Y, X|\beta, \alpha) \), typically through use of the EM (Expectation-Maximization) algorithm.

unbiased if MAR and model correctly specified

proposed by Ibrahim (1990)
Alternate:

- calculating the Expected value of the missing observations
- Maximizing the complete data log likelihood given those values
- formalized by Dempster, Laird and Rubin (1977)
Ibrahim method of weights

Augmented dataset

<table>
<thead>
<tr>
<th>Original dataset</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>y</td>
<td>x₁</td>
<td>x₂</td>
<td>x₃</td>
<td>wt</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0w₃₁</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0w₃₂</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>0w₅₁</td>
<td>1w₅₂</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0w₇₁</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0w₇₂</td>
<td>1w₇₃</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0w₇₄</td>
<td>1w₇₅</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Nicholas J. Horton  Much ado about nothing
Maximum likelihood

- major task: housekeeping and specification of model for X
- need MCEM for continuous
- Implementations now exist (XMISS)
- some limitations (no continuous RV with missing, only 10 variables with missing values, no control of models for predictors, only 5 levels for categorical variables [MONTH vs. SEASON])
Weighting approaches

- great if only one incomplete predictor (Ibrahim et al JASA 2005)
- plausible to consider if monotone missing
- fiendishly difficult otherwise
Weighting approaches (Rotnitzky, in press)

Not much is available for the analysis of semi-parametric models of longitudinal studies with intermittent non-response. One key difficulty is that realistic models for the missingness mechanism are not obvious. As argued in Robins and Gill (1997) and Vaansteelandt, Rotnitzky and Robins (2007), the [coarsened at random] CAR assumption with non-monotone data patterns is hard to interpret and rarely realistic...More investigation into realistic, easy to interpret models for intermittent non-response is certainly needed.
Multiple imputation

- ‘fill-in’ the missing values with some ‘appropriate’ value to give a completed dataset
- repeat this process multiple times
- combine results from each of these multiple imputations
- originally proposed by Rubin (1978)
- assumes MAR missingness
- requires a model to ‘fill-in’ the values (hardest part!)
Specifying the imputation model

- most complicated task (since running the separate analyses is fast and cheap)
- simple when the predictors and outcome are plausibly multivariate normal
- harder with categorical missing values
- even harder if non-monotone
- Note: the imputation model is of only secondary interest to the analyst!
Specifying the imputation model

1. full specification of joint distribution (Rubin, Schafer)
2. separate chained equations (van Buuren 1999, Raghunathan 1999, Royston 2005)
Full specification of joint distribution

- need joint distribution function for mixture of different types of random variables
- one approach: log-linear model for categorical variables, MVN for remainder conditional on categorical

\[
f(X_1, \ldots, X_9, Y) = f(X_1, \ldots, X_6, Y)f(X_7, X_8, X_9 | X_1, \ldots, X_6, Y)
\]
Full specification of joint distribution

- conditional on categorical variables, are the rest plausibly multivariate normal?
- what about other types of variables?
- proliferation of (nuisance) parameters
- can be computationally challenging
- need to remain “proper” in the sense of Rubin
- potential for bias if mis-specified
- a lot of work!
Chained equations

- impute one value, use that to impute the next with a separate equation, and repeat until convergence
- fit marginal models for each variable with missing values

\[ f(X_1|X_2, \ldots, X_9, Y) \]
\[ f(X_2|X_1, X_3, \ldots, X_9, Y) \]
\[ f(X_3|X_1, X_2, X_4, \ldots, X_9, Y) \]
\[ f(X_4|X_1, X_2, X_3, X_5, \ldots, X_9, Y) \]

then repeat from the top 5 or 10 or 15 times
Chained equations

- run separate chain per imputation (typically 10-25)
- fit main effects only (common default)
- computationally straightforward
- not much theoretical justification
- potential problem: marginal distributions may not correspond to any sensible joint distribution!
Analysis using multiple imputation in SAS/STAT is carried out in three steps

1. imputation is carried out by PROC MI
2. complete data methods are employed using any of the SAS procedures (e.g. PROC GLM, GENMOD, PHREG, or LOGISTIC) with the ‘BY’ statement for each imputed data set
3. results are combined using PROC MIANALYZE
Artificial example (Horton and Lipsitz, TAS 2001)

```sas
proc mi data=allison out=miout nimpute=25 noprint;
   monotone method=reg;
   var y x1 x2;
proc reg data=miout outest=outreg covout noprint;
   model y = x1 x2;
   by _Imputation_;
proc mianalyze data=outreg;
   var Intercept x1 x2;
run;
```
proc mi data=allison out=miout nimpute=25 noprint;
  monotone method=reg;
  var y x1 x2;
proc reg data=miout outest=outreg covout noprint;
  model y = x1 x2;
  by _Imputation_;
proc mianalyze data=outreg;
  var Intercept x1 x2;
run;
Artificial example (Horton and Lipsitz, TAS 2001)

```sas
proc mi data=allison out=miout nimpute=25 noprint;
   monotone method=reg;
   var y x1 x2;
proc reg data=miout outest=outreg covout noprint;
   model y = x1 x2;
   by _Imputation_;
proc mianalyze data=outreg;
   var Intercept x1 x2;
run;
```
Artificial example (Horton and Lipsitz, TAS 2001)

```sas
proc mi data=allison out=miout nimpute=25 noprint;
  monotone method=reg;
  var y x1 x2;
proc reg data=miout outest=outreg covout noprint;
  model y = x1 x2;
  by _Imputation_
proc mianalyze data=outreg;
  var Intercept x1 x2;
run;
```
SAS PROC MI

- SAS PROC MI MCMC statement (appropriate if all variables multivariate normal)
- SAS PROC MI CLASS statement for categorical variables (straightforward if monotone pattern)
- what if not MV normal and non-monotone?
SAS PROC MI for non-monotone (our ‘ad-hoc’ approach)

1. create 20 imputations of the missing values for TOTCHG, using a regression equation based on variables that are complete (simplifying assumption)

2. for each of these imputed datasets, impute missing categorical variables separately for each pattern of missing data

3. code requires some sophistication in SAS (provided in Appendix to our manuscript)
IVEware

- SAS version 9 callable routine built using the SAS macro language
- straightforward to install
- implements chained equation approach
- allows for constraints on imputed values (structural zeroes, bounds on imputations)
datain work.one; mdata impute;
iterations 10; multiples 25;
seed 42; estout mylib.est;
repout mylib.rep; link logistic;
categorical atype nseason race;
dependent routine;
predictor age female los totchg ndx aweekend;
estimates
  race1: race (1) race2: race (0 1) /
  race3: race (0 0 1) /
  atype1: atype (1) atype2: atype (0 1) /
  nseason1: nseason (1) nseason2: nseason (0 1) /
  nseason3: nseason (0 0 1);
print details;
utilizes a bootstrapping-based variant of EM to impute that is fast and robust (black box)
imputation done in a standalone package (or as an add-on library for R)
datasets can be loaded into another package to run analyses and combine results (in SAS using PROC MIANALYZE, in Stata using Royston’s ICE)
Hmisc

```r
f <- aregImpute(~ ROUTINE + AGE + ... + NDX, 
    n.impute=25, defaultLinear=TRUE, data=kidfact)
fmi <- fit.mult.impute(ROUTINE ~ AGE + ... NDX, 
    glm, f, family="binomial",data=kidfact)
impse <- sqrt(diag(Varcov(fmi)))
summary(fmi)
```
imp <- mice(kidfact, im=c("","polyreg","polyreg","","","norm","polyreg","",""), m=25, seed=456)
fitted <- glm.mids(Routine ~ AGE + ... + NDX, family=binomial, data=imp)
result <- pool(fitted)
Other options

- SOLAS (standalone package)
- S-plus missing values library
- Cytel’s XMISS/LogXact
- SPSS
Descriptive statistics

variable percentage
ROUTINE 86%
WEEKEND 20%
FEMALE 54%
WHITE 57%

variable mean (SD)
AGE 16.3 (2.7)
LOS 6.4 (12.7)
TOTCHG $9,230 ($17,371)
NDX 3.5 (2.0)
## Missing data model results (log OR)

<table>
<thead>
<tr>
<th>Package</th>
<th>WEEKEND</th>
<th>FEMALE</th>
<th>BLACK</th>
</tr>
</thead>
<tbody>
<tr>
<td>complete case</td>
<td>-0.058 (0.026)</td>
<td>0.089 (0.021)</td>
<td>-0.018 (0.029)</td>
</tr>
<tr>
<td>Amelia II</td>
<td>-0.027 (0.020)</td>
<td>0.103 (0.016)</td>
<td>-0.066 (0.024)</td>
</tr>
<tr>
<td>ICE</td>
<td>-0.020 (0.020)</td>
<td>0.099 (0.016)</td>
<td>-0.082 (0.024)</td>
</tr>
<tr>
<td>XMISS/LogXact</td>
<td>-0.026 (0.020)</td>
<td>0.105 (0.016)</td>
<td>-0.075 (0.026)</td>
</tr>
<tr>
<td>SAS PROC MI</td>
<td>-0.036 (0.021)</td>
<td>0.119 (0.017)</td>
<td>-0.068 (0.025)</td>
</tr>
<tr>
<td>S-Plus</td>
<td>-0.018 (0.020)</td>
<td>0.098 (0.016)</td>
<td>-0.078 (0.023)</td>
</tr>
</tbody>
</table>
Sensitivity analyses

- MAR may not be tenable
- NINR models require additional specification of joint likelihood
- important way to assess sensitivity to MAR assumption
• assess sensitivity to MAR for logistic regression models using existing imputed datasets
• posit model for missingness (estimable if $\delta = 0$):
• Example: for missing $X_2$:

\[
\text{logit}(P(R = 1| Y, X_1, X_2)) = \gamma_0 + \gamma_1 Y + \gamma_2 X_1 + \delta X_2
\]
Carpenter, Kenward and White (SMIMR, 2007)

- weight results based on fixed sensitivity parameter $\delta$ (only requires imputed values from $X_2$ from each imputed dataset)

$$\tilde{w}_m = \exp \left( \sum_{i=1}^{n_1} -\delta X_{2,i}^m \right)$$

- reweight parameters from imputed datasets (only requires weights and vector of imputation results for parameters of interest)

$$w_m = \frac{\tilde{w}_m}{\sum_{i=1}^{m} \tilde{w}_m}, \quad \hat{\theta}_{NINR} = \sum_{i=1}^{m} w_m \hat{\theta}_m$$
Distribution of $\hat{\theta}$ from 50 imputations (BLACK)
Limitations

- Assumes support is the same under MAR or NINR
- Only allows one non-ignorably missing variable (predictor or outcome)
- Not ideally suited to missingness for KID study
- Undertake four marginal sensitivity analyses (one per missing variable)
### Sensitivity analysis results (log OR)

<table>
<thead>
<tr>
<th>Analysis</th>
<th>BLACK</th>
</tr>
</thead>
<tbody>
<tr>
<td>MI MAR</td>
<td>-0.082 (0.024)</td>
</tr>
<tr>
<td>NINR (ATYPE)</td>
<td>-0.091</td>
</tr>
<tr>
<td>NINR (RACE)</td>
<td>-0.075</td>
</tr>
<tr>
<td>NINR (SEASON)</td>
<td>-0.084</td>
</tr>
<tr>
<td>NINR (TOTCHG)</td>
<td>-0.090</td>
</tr>
</tbody>
</table>
complete case estimator simple, but may be inefficient and biased (particularly when missingness depends on Y or selection biases exist)
‘ad-hoc’ methods not recommended
a variety of models have been proposed in the statistical literature, many of these make simplifying assumptions or have been coded specifically for a given situation.

implementations of missing data methods are available, require imposition of assumptions (MAR) and somewhat considerable effort above and beyond fitting the regression model of interest.

these imputation models yield efficiency gains (of more than 25%)

also may reduce bias (as seen for the WEEKEND and BLACK parameters), assuming MAR.
Summary

- missing data models are not yet commonly utilized in practice, nor is the extent of missingness clearly reported
- sensitivity analyses of the MAR assumption should be carried out routinely
Future work

- “job security for statisticians!”
- assess sensitivity to assumptions
- determine when these methods have greatest potential for benefit
- support for non-monotone models in SAS PROC MI?
- better theoretical justification for chained equations
- use chained equation to get to monotone pattern, then use more principled approaches?
- use of NINR models in this setting (will WinBUGS run with a dataset of this size?), decrease the degree of difficulty of fitting those models
- account for clustering, longitudinal measures and complex survey design
Cautions are needed, however, just as with any statistical methodology. It is clear that if the imputation model is seriously flawed in terms of capturing the missing-data mechanism, then so will be any analysis based on such imputations. ... This is not an additional burden for using Rubin’s method, but rather a fundamental requirement for any general method that attempts to produce statistically and scientifically meaningful results in the presence of incomplete data.

(Barnard and Meng, SMIMR 1999)
The most pressing task, in my opinion, is placing further emphasis on the general recognition and understanding, at a conceptual level, of properly dealing with the missing data mechanism, as part of our ongoing emphasis on the importance of the data collection process in any meaningful analysis.

(Meng, Dial “M” for Missing, JASA 2000)
Much ado about nothing: methods and implementations to estimate incomplete data regression models

Nicholas J. Horton
Smith College, Northampton, MA, USA and University of Auckland, New Zealand

December 6, 2007, Australasian Biometrics Conference

nhorton@email.smith.edu
http://www.math.smith.edu/~nhorton/muchado.pdf