Missing Data Imputation in the Bayesian Framework

by

John Draper, David Kadonsky

June 9, 2005

Introduction

- Dr. Paul Robbins of The Ohio State University developed a survey on lawn care beliefs and behaviors in 2001 to challenge commonly held beliefs in the environmental literature.
  – Unfortunately, some variables were missing in the survey data.

Introduction

- Dr. Paul Robbins of The Ohio State University developed a survey on lawn care beliefs and behaviors in 2001 to challenge commonly held beliefs in the environmental literature.
  - Unfortunately, some variables were missing in the survey data.
- A variety of imputation methods (cold deck, hot deck, regression based, and Bayesian) are explored to analyze the advantages and disadvantages of each in the context of these survey data.

Purpose

- The primary purpose of the paper is to focus on developing a comprehensive Bayesian imputation method that would improve on established methods.

Purpose

- The primary purpose of the paper is to focus on developing a comprehensive Bayesian imputation method that would improve on established methods.
- Ideally, the augmented data would more closely resemble the true sampled population and therefore, the logistic regression models derived would lead to more accurate conclusions.

Data

- National telephone survey conducted by the Survey Research Center at The Ohio State University in 2001 on behalf of Dr. Paul Robbins.
- Asked questions on the opinions, behaviors and knowledge of lawn care
- Missing data values come primarily from the household income and housing value variables.

Methods

- Cold Deck Imputation
- Hot Deck Imputation
- Non-random Regression
- Bayesian Imputation

Cold Deck Imputation

- Advantages:
  - simple
  - unbiased for observed sample

Cold Deck Imputation

- Advantages:
  - simple
  - unbiased for observed sample
- Disadvantages:
  - does not use any concomitant information
  - ignores any missing data mechanism
    - if MAR, then cold deck will be biased
  - large variance in imputed missing values

Hot Deck Imputation

- Advantages:
  - fairly simple
  - uses concomitant information
  - can capture crude missing data mechanisms

Hot Deck Imputation

- Advantages:
  - fairly simple
  - uses concomitant information
  - can capture crude missing data mechanisms
- Disadvantages:
  - must discretize continuous concomitant data
  - no distinct values imputed

Non-random Regression

- Advantages:
  – strong use of concomitant data
  – can capture complex missing data mechanisms

Non-random Regression

- Advantages:
  - strong use of concomitant data
  - can capture complex missing data mechanisms
- Disadvantages:
  - deterministic
  - cannot take any exogenous information into account

Bayesian Imputation

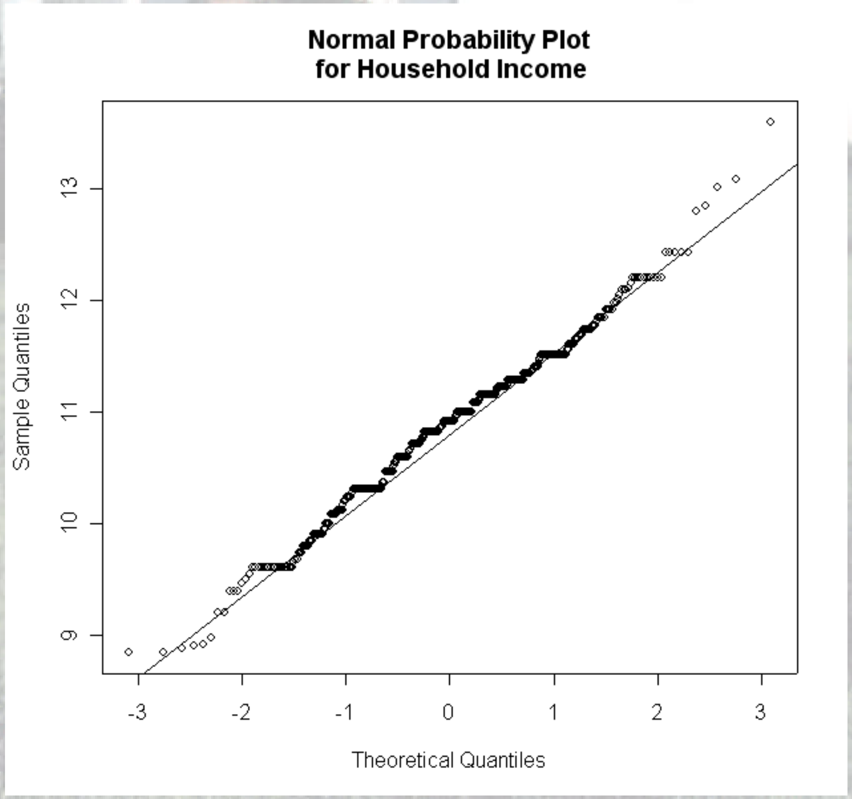
- Advantages:
  - strong use of concomitant data
  - can capture complex missing data mechanisms
  - can take any exogenous information into account

Bayesian Imputation

- Advantages:
  - strong use of concomitant data
  - can capture complex missing data mechanisms
  - can take any exogenous information into account
- Disadvantages:
  - complexity
  - strong distributional assumptions

Bayesian Imputation

- Requires normality for imputation to be reasonable.



Bayesian Imputation

$$[y, ln(income)|x] = [y|ln(income), x1, x2] \times [ln(income)|x1, x2] \qquad (1)$$

where x1 are the variables used to predict y (the final variable of interest), x2 are the variables to predict income and $x=(x_1,x_2)$. This model can be simplified by noting that income does not depend on x1 and once income is given, y does not depend on x2. Therefore, the model reduces to:

$$[y, ln(income)|x] = [y|ln(income), x1] \times [ln(income)|x2] \qquad (2)$$

Bayesian Imputation

- Prior data says: $ln(income) \sim N(\mu_0, \sigma^2)$
- Data model says: $ln(income)|x2 \sim N(\mathbf{X}\beta, \tau^2)$

Bayesian Imputation

- Prior data says: $ln(income) \sim N(\mu_0, \sigma^2)$
- Data model says: $ln(income)|x2 \sim N(\mathbf{X}\beta, \tau^2)$
- Original Model:

$$ln(income) = \mathbf{X}\beta + \epsilon_i \; where \; \epsilon_i \sim N(0, \tau^2)$$

- Taking prior information into account:

$$ln(income) = \mathbf{Y}\beta^* + \epsilon_i \; where \; \epsilon_i \sim N(0, \tau^2)$$

Bayesian Imputation

The prior placed upon β^* is the following:

$$\begin{aligned}
\beta^* &\sim MVN(\mu_0, \Lambda_0) \; where && (3)\\
\mu_0^{\mathbf{T}} &= (\theta, 0, 0, \ldots, 0) \; and && (4)\\
\mathbf{\Lambda}_0 &= \begin{bmatrix} \tau^2 & 0 & 0 & \ldots & 0 \\ 0 & 100 & 0 & \ldots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \ldots & 100 \end{bmatrix} && (5)
\end{aligned}$$

and the prior on $ln(income) \sim N(\theta, \tau^2)$.

Bayesian Imputation

The prior placed upon β^* is the following:

$$\begin{aligned}
\beta^* &\sim MVN(\mu_0, \mathbf{\Lambda_0}) \; where && (3)\\
\mu_0^{\mathbf{T}} &= (\theta, 0, 0, \ldots, 0) \; and && (4)\\
\mathbf{\Lambda_0} &= \begin{bmatrix} \tau^2 & 0 & 0 & \ldots & 0 \\ 0 & 100 & 0 & \ldots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \ldots & 100 \end{bmatrix} && (5)
\end{aligned}$$

and the prior on $ln(income) \sim N(\theta, \tau^2)$.

The likelihood is:

$$ln(income)|\mathbf{Y}, \beta^* \sim N(\mathbf{Y}\beta^*, \sigma^2\mathbf{I}) \qquad (6)$$

Bayesian Imputation

$$\beta^* \mid ln(income), \mathbf{Y} \sim N(\mu_{\mathbf{n}}, \mathbf{\Lambda_n}) \; where \tag{7}$$

$$\mu_{\mathbf{n}} = (\mathbf{\Lambda_0}^{-1} + n(\frac{1}{\sigma^2}\mathbf{I}))^{-1}(\mathbf{\Lambda_0}^{-1} * \mu_0 + n(\frac{1}{\sigma^2}\mathbf{I})\hat{\beta}) \tag{8}$$

$$\mathbf{\Lambda_n}^{-1} = \mathbf{\Lambda_0}^{-1} + n(\frac{1}{\sigma^2}\mathbf{I}) \tag{9}$$

and $\hat{\beta}$ is a k×1 vector of least squares estimates.

Bayesian Imputation

$$\beta^* \mid ln(income), \mathbf{Y} \sim N(\mu_{\mathbf{n}}, \Lambda_{\mathbf{n}}) \; where \tag{7}$$

$$\mu_{\mathbf{n}} = (\Lambda_0{}^{-1} + n(\frac{1}{\sigma^2}\mathbf{I}))^{-1}(\Lambda_0{}^{-1} * \mu_0 + n(\frac{1}{\sigma^2}\mathbf{I})\hat{\beta}) \tag{8}$$

$$\Lambda_{\mathbf{n}}{}^{-1} = \Lambda_0{}^{-1} + n(\frac{1}{\sigma^2}\mathbf{I}) \tag{9}$$

and $\hat{\beta}$ is a k×1 vector of least squares estimates.

- To impute the missing values, one simply computes the posterior predictive mean of the data, Yß*.

Results

- Imputing the housing value
  - To get some idea as to how reliable our imputed estimates are, we chose to use each non-deterministic method to impute the missing data values 5,000 times.
  - The resulting mean values of the full data sets (observed and imputed housing value) were accumulated.

Results

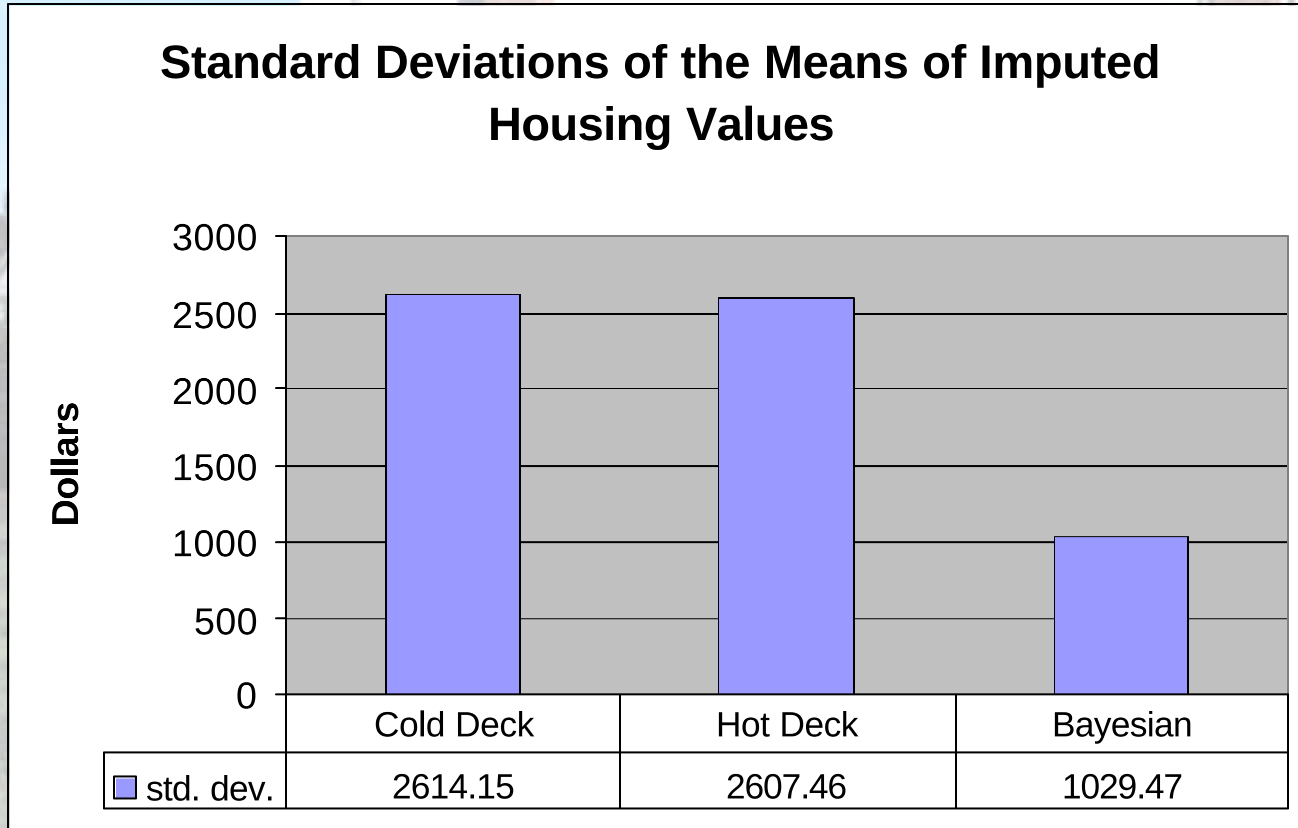
Method	*mean* (in \$)	*std.dev.* (in \$)
Dropped	167900	N/A
Cold Deck	167904.2	2614.15
Hot Deck	168064.3	2607.46
Bayesian	163488	1029.47
Regression	162100	N/A

Results

Method	*mean* (in \$)	*std.dev.* (in \$)
Dropped	167900	N/A
Cold Deck	167904.2	2614.15
Hot Deck	168064.3	2607.46
Bayesian	163488	1029.47
Regression	162100	N/A

- Cold deck and hot deck imputation methods seem to be centered around the sample mean housing value
- Regression suggests that respondents with lower housing values may be underrepresented.
- Bayesian falls between Regression and cold deck

Results



- Bayesian reduces standard deviation of full data means of housing value by 60%

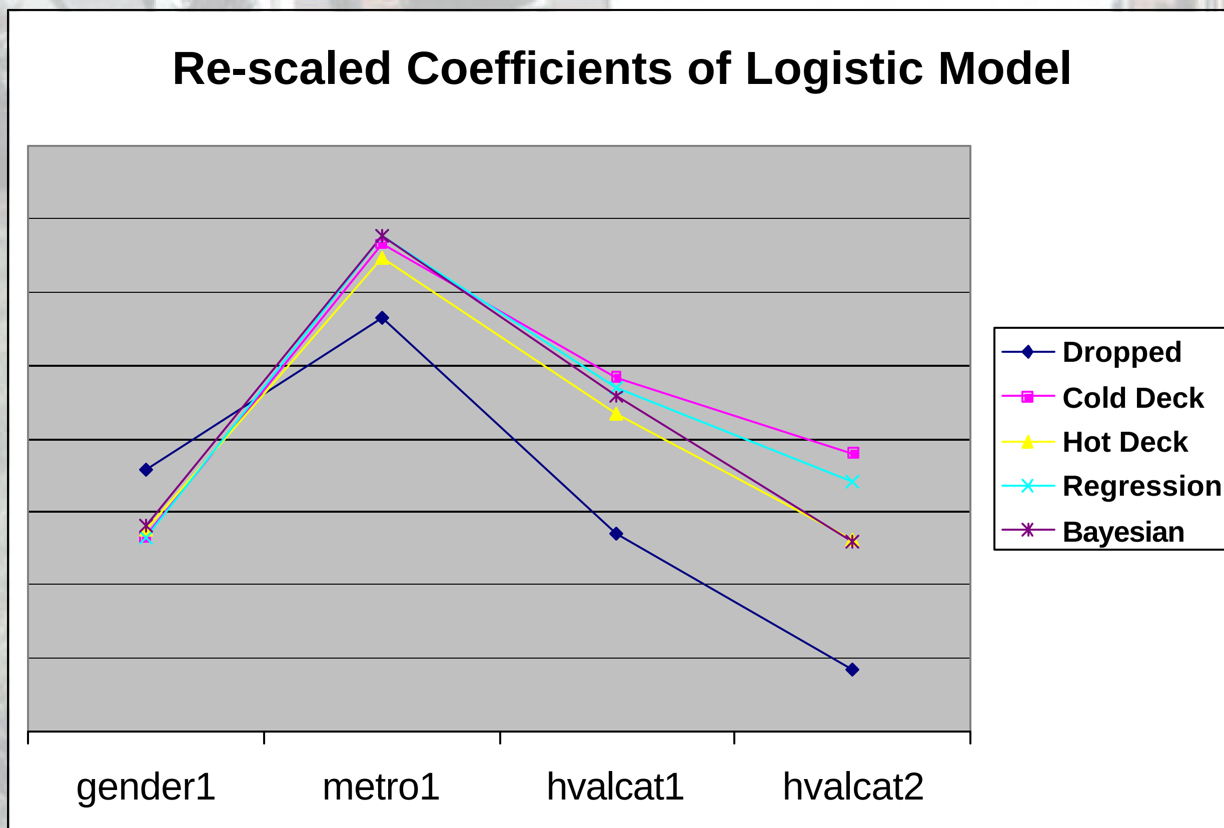
Results – Logistic Model

Method	*Intercept*	*gender1*	*metro1*	*hvalcat1*	*hvalcat2*
Dropped	1.984	0.712	1.132	-0.959	-1.829
Cold Deck	1.679	0.536	1.329	-0.533	-1.24
Hot Deck	1.854	0.547	1.292	-0.634	-1.476
Regression	1.747	0.525	1.353	-0.559	-1.32
Bayesian	1.669	0.557	1.353	-0.589	-1.483

gender1 is coded for Female
metro1 is coded for Urban or Suburban (vs. Rural)
hvalcat1 is coded for Housing Value between $100,000 and $150,000
hvalcat2 is coded for Housing Value less than $100,000

Results – Logistic Model

Method	*Intercept*	*gender1*	*metro1*	*hvalcat1*	*hvalcat2*
Dropped	1.984	0.712	1.132	-0.959	-1.829
Cold Deck	1.679	0.536	1.329	-0.533	-1.24
Hot Deck	1.854	0.547	1.292	-0.634	-1.476
Regression	1.747	0.525	1.353	-0.559	-1.32
Bayesian	1.669	0.557	1.353	-0.589	-1.483



Conclusions

- Very little gained by imputing housing value by cold deck and hot deck.
- Enormous reduction in variance of the imputed means of housing value using Bayesian methods.
- Very little difference in resulting logistic models.

Further Research

- FRITZ imputation
  - (Federal Reserve Imputation Technique Zeta)
- Predictive Computational techniques
- Other Bayesian modeling approaches