An Application of Multiple Imputation and Sampling Based Estimation

Haluk Gedikoglu

Cooperative Research Programs
Lincoln University of Missouri

July 26, 2012

Outline

- Background
- Objectives
- MI Imputation Step
- MI Completed Data Analysis
- MI Sampling Based Estimation
- Application
- Conclusions

Outline

- Background
- Objectives
- MI Imputation Step
- MI Completed Data Analysis
- MI Sampling Based Estimation
- Application
- Conclusions

Outline

- Background
- Objectives
- MI Imputation Step
- MI Completed Data Analysis
- MI Sampling Based Estimation
- Application
- Conclusions

Outline

- Background
- Objectives
- MI Imputation Step
- MI Completed Data Analysis
- MI Sampling Based Estimation
- Application
- Conclusions

Outline

- Background
- Objectives
- MI Imputation Step
- MI Completed Data Analysis
- MI Sampling Based Estimation
- Application
- Conclusions

Outline

- Background
- Objectives
- MI Imputation Step
- MI Completed Data Analysis
- MI Sampling Based Estimation
- Application
- Conclusions

Outline

- Background
- Objectives
- MI Imputation Step
- MI Completed Data Analysis
- MI Sampling Based Estimation
- Application
- Conclusions

Outline

- Background
- Objectives
- MI Imputation Step
- MI Completed Data Analysis
- MI Sampling Based Estimation
- Application
- Conclusions

Background

- Missing data is a problem that occurs frequently in survey data.
- Missing data can cause biased estimates and reduced efficiency for the regression estimates (Rubin, 1987).
- The standard procedure on Stata is to use only complete observations, which is called list-wise deletion.

Background, Cont'd

- List-wise deletion can lead to a loss of significant number of observations. For example in the current study list-wise deletion leads to a loss of 43% of the data.
- Overtime, different methods have been used to handle missing data, including single imputation and multiple imputation.
- Simple imputation treats imputed values as known in the analysis, which understates the variance of the estimates and overstates the precision.

Background, Cont'd

- Multiple imputation addresses this problem by creating multiple sets of imputed data and take into account the sampling variability due to missing data, which is called between-imputation variability.
- Although statistical literature has been developed for missing data imputation, the use of these methods have been relatively low in applied fields, such as agricultural household survey analysis.
- There are many practical problems that have not been answered in applying missing data imputation methods, such as how to analyze the data when all the variables have missing observations.
- Implications of sampling based estimation for missing data imputation, when all the variables have missing observations, should be analyzed.

Objectives

- Analyze the implications of multiple imputation when all the variables have missing observations.
- Analyze the implications of multiple imputation when sampling based estimation is used for stratified random sampling.

Data Augmentation

- ***Multiple Imputation*** is based on simulation from a Bayesian posterior distribution of missing data.
  - Data Augmentation (an iterative Markov Chain Monte Carlo method). Data augmentation consists of two steps, an I step (imputation step) and a P step (posterior step), which are preformed at each iteration $t = 0, 1, \ldots, T$(Schafer, 1997).
    - I-Step: At iteration t of the I step, the missing values in are replaced with draws from the conditional posterior distribution of given observed data and the current values of model parameters and independently for each observation (Little and Rubin, 2002).

$$\mathbf{x}_{i(m)}^{(t+1)} \sim P\left(x_{i(m)}|z_i, x_{i(0)}, \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Sigma}^{(t)}\right), i = 1, \ldots, N$$

Data Augmentation, cont'd

- P-Step: During the P step new values of model parameters and are drawn from their conditional posterior distribution given observed data and data imputed in the previous I step $\mathbf{x}_{i(m)}^{(t+1)}$:

$$\mathbf{\Sigma}^{(t+1)} \sim P\left(\mathbf{\Sigma}|z_i, x_{i(0)}, \mathbf{x}_{i(m)}^{(t+1)}\right)$$

$$\mathbf{\Theta}^{(t+1)} \sim P\left(\mathbf{\Theta}|z_i, x_{i(0)}, \mathbf{x}_{i(m)}^{(t+1)}\right)$$

- I and P steps are repeated until the MCMC sequence $\left(\mathbf{X}_m^{(t)}, \mathbf{\Theta}^{(t)}, \mathbf{\Sigma}^{(t)}\right)$ converges to the stationary distribution $P\left(\mathbf{X}_m, \mathbf{\Theta}, \mathbf{\Sigma}|\mathbf{Z}, \mathbf{X}_0\right)$.

Expectation-Maximization Algorithm

- The EM algorithm iterates the expectation step (E step) and maximization step (M step) to maximize the log-likelihood function:

$$l_l(\boldsymbol{\Theta}, \boldsymbol{\Sigma}|\mathbf{X_0}) = \sum_{s=1}^{s} \sum_{i \in I(s)} \{-0.5 \ln(|\boldsymbol{\Sigma}_\mathbf{s}|) - 0.5(\mathbf{x}_{i(o)} - \boldsymbol{\Theta}'_{(s)} z_i)' \Sigma_s^{-1} (x_{i(o)} - \boldsymbol{\Theta}_{(s)} z_i)\}$$

  - E- Step:Following Little and Rubin (2002) the expectations $E\left(\sum_{s=1}^{N} x_i x_i'\right)$ and $E\left(\sum_{s=1}^{N} z_i x_i'\right)$ are computed with respect to the conditional distribution $P\left(\mathbf{X}_m | \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Sigma}^{(t)}, \mathbf{X}_0\right)$.
  - M- Step: During the M step, the model parameters are updated using the computed expectations of the sufficient statistics:

$$\boldsymbol{\Theta}^{(t+1)} = (\mathbf{Z}'\mathbf{Z})^{-1} E\left(\sum_{i=1}^{N} z_i x_i'\right)$$

$$\boldsymbol{\Sigma}^{(\mathbf{t+1})} = \frac{1}{N + \lambda + p + 1} \left\{ E(\sum_{i=1}^{N} x_i x_i') - E(\sum_{i=1}^{N} z_i x_i')(\mathbf{Z}'\mathbf{Z})^{-1} E(\sum_{i=1}^{N} z_i x_i') + \boldsymbol{\Lambda}^{-\mathbf{1}} \right\}$$

MI Estimation Stage

- The results obtained from M completed-data analyses are combined into a single multiple-imputation based estimation results.
- Let$\{(\widehat{\mathbf{q}}_{\mathbf{i}}, \widehat{\mathbf{U}}_{\mathbf{i}}) : i = 1, 2, \ldots, M\}$ be the completed-data estimates of $\mathbf{q}$ and the respective variance covariance estimates $\mathbf{U}$ from M imputed datasets. The multiple imputation estimate of $\mathbf{q}$ is $\bar{q}_M = \frac{1}{M} \sum_{i=1}^{M} \widehat{q}_i$.
- The var-cov estimate of $\bar{q}_M$ (total) is $\mathbf{T} = \bar{\mathbf{U}} + (\mathbf{1} + \frac{\mathbf{1}}{\mathbf{M}})\,\mathbf{B}$, where $\bar{\mathbf{U}} = \frac{1}{M} \sum_{i=1}^{M} \widehat{\mathbf{U}_{\mathbf{i}}}/\mathbf{M}$ is the within-imputation var-cov matrix and $\mathbf{B} = \frac{\mathbf{1}}{\mathbf{M}} \sum_{\mathbf{1}=\mathbf{1}}^{\mathbf{M}} (\mathbf{q}_{\mathbf{i}} - \bar{\mathbf{q}}_{\mathbf{M}})(\mathbf{q}_{\mathbf{i}} - \bar{\mathbf{q}}_{\mathbf{M}})'/(\mathbf{M} - \mathbf{1})$ is the between-imputation variance-covariance matrix.

MI Sampling Based Estimation

- For each strata h, the sampling weights are calculated as $W_h = N_h/n_n$, where N_h is the number of observations in population in strata h and n_n is the number of observations sampled in strata h.
- Sampling weights W_h are used in the estimation stage for each imputation $m = 1 \ldots M$.
- Within variance-covariance estimate $\mathbf{\bar{U}} = \frac{1}{M}\sum_{i=1}^{M} \widehat{\mathbf{U_i}}/\mathbf{M}$ includes $\widehat{\mathbf{U_i}}$ is computed using Taylor series linearization.
- Degrees of freedom is now the small-sample method, which is $\tilde{v}_{mi} = \left(\frac{1}{(M-1)\widehat{\gamma}^{-2}} + \frac{1}{\widehat{v}_{obs}}\right)^{-1}$.

Data

- A mail survey of 2,995 livestock farmers was conducted in Iowa and Missouri in Spring 2011.
- Farmers were stratified by farm sales and by type of livestock.
- The effective response rate for the survey was 21 percent.

Missing Data Table

```
. misstable summarize, all
```

Variable	Obs=.	Obs>.	Obs<.	Unique values	Min	Max
rrsoybean	22		450	2	0	1
age	12		460	60	24	89
lowned	6		466	194	0	1832
lrentout	11		461	46	0	1200
lrentin	13		459	108	0	4000
watqual	9		463	5	1	5
airqual	29		443	5	1	5
globalwarm	9		463	5	1	5
Othfarm	20		452	5	1	5
neighbors	22		450	5	1	5
bank	23		449	5	1	5
contractor	23		449	5	1	5
university	19		453	6	0	5
USDA	25		447	5	1	5
Other	26		446	5	1	5
state	5		467	2	0	1
educop	14		458	5	1	5
educsp	99		373	5	1	5
offfarmop	36		436	6	1	6
offfarmsp	120		352	6	1	6
dairycaau	5		467	73	0	1571.429
beefcaau	4		468	56	0	2200
beefcoau	6		466	75	0	750
swinele55au	5		467	37	0	300
swinebi55au	6		466	56	0	4000
broilerau	7		465	22	0	35
turkeyau	6		466	34	0	1000
sheepau	5		467	38	0	60
otherau	6		466	49	0	12000
hirelabor	9		463	2	0	1
fs1_9	16		456	2	0	1
fs10_49	16		456	2	0	1
fs50_99	16		456	2	0	1
fs100_249	16		456	2	0	1
fs250_499	16		456	2	0	1
fs500	16		456	2	0	1
strata			472	30	1	49
weight			472	29	10	2169
fsc			472	30	10	32805

Setting Data as Multiple Imputation (MI)

```
. mi set mlong

 . mi set M = 10
 (10 imputations added; M = 10)

. mi register imputed rrsoybean age lowned lrentout lrentin watqual airqual glob
> alwarm Othfarm neighbors bank contractor university USDA Other state educop ed
> ucsp offfarmop offfarmsp dairycaau beefcaau beefcoau swinele55au swinebi55au b
> roilerau turkeyau sheepau otherau hirelabor fs1_9 fs10_49  fs50_99 fs100_249 f
> s250_499 fs500
(199 m=0 obs. now marked as incomplete)
```

When missing data is not imputed, only 273 out of 472 observations are used, which leads to a loss of 43% of observations.

```
Probit regression                                 Number of obs   =        273
                                                  LR chi2(33)     =     155.88
                                                  Prob > chi2     =     0.0000
Log likelihood = -110.87757                       Pseudo R2       =     0.4128
```

Using Multivariate Normal Distribution(MVN)

```
. mi impute mvn rrsoybean age lowned lrentout lrentin watqual airqual globalwarm
>  Othfarm neighbors bank contractor university USDA Other state educop educsp o
> fffarmop offfarmsp dairycaau beefcaau beefcoau swinele55au swinebi55au broiler
> au turkeyau sheepau otherau hirelabor  fs50_99 fs100_249 fs250_499 fs500, add(
> 10) rseed(2232) noisily emlog emoutput force


Expectation-maximization estimation      Number obs            =        471
                                         Number missing        =        632
                                         Number patterns       =         88
Prior: uniform                           Obs per pattern: min  =          1
                                                          avg  =   5.352273
                                                          max  =        273

Observed log likelihood =  -29374.08 at iteration 26



Performing MCMC data augmentation ...

Multivariate imputation                 Imputations =       10
Multivariate normal regression                added =       10
Imputed: m=1 through m=10                   updated =        0

Prior: uniform                           Iterations =     1000
                                            burn-in =      100
                                            between =      100
```

Convergence of Data-Augmentation

- We use the worst linear function (WLF), developed by Schafer (1997) is used to detect the convergence and autocorrelation for the Data-Augmentation.
- WLF corresponds to the linear combination of parameter estimates where the coefficients are chosen such that this function has the highest asymptotic rate of missing information.
- WLF can be calculated as (Schafer, 1997): $w(\theta) = \widehat{v}'(\theta - \widehat{\theta})$

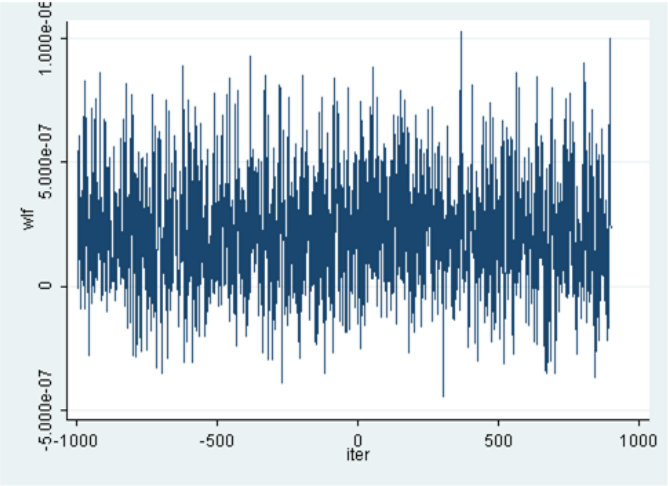
```
. mi impute mvn rrsoybean age lowned lrentout lrentin watqual airqual globalwarm
>  Othfarm neighbors bank contractor university USDA Other state educop educsp o
> fffarmop offfarmsp dairycaau beefcaau beefcoau swinele55au swinebi55au broiler
> au turkeyau sheepau otherau hirelabor fs10_49  fs50_99 fs100_249 fs250_499 fs5
> 00, mcmconly burnin(2000) rseed(2232) savewlf(wlf)

. use wlf, clear

. tsset iter
```

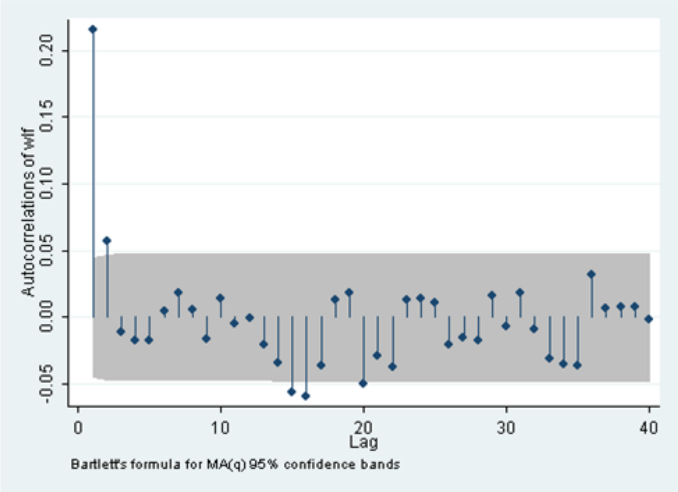
Convergence of Data-Augmentation

```
. tsline wlf, ytitle(Worst linear function) xtitle(Burn-in period)
```



Convergence of Data-Augmentation

```
. ac wlf, title(Worst linear function) ytitle(Autocorrelations)
```



Imputed Data

	Observations per *m*			
Variable	complete	incomplete	imputed	total
rrsoybean	450	22	22	472
age	460	12	12	472
lowned	466	6	6	472
lrentout	461	11	11	472
lrentin	459	13	13	472
watqual	463	9	9	472
airqual	443	29	29	472
globalwarm	463	9	9	472
Othfarm	452	20	20	472
neighbors	450	22	22	472
bank	449	23	23	472
contractor	449	23	23	472
university	453	19	19	472
USDA	447	25	25	472
Other	446	26	26	472
state	467	5	5	472
educop	458	14	14	472
educsp	373	99	99	472
offfarmop	436	36	36	472
offfarmsp	352	120	120	472
dairycaau	467	5	5	472
beefcaau	468	4	4	472
beefcoau	466	6	6	472
swinele55au	467	5	5	472
swinebi55au	466	6	6	472
broilerau	465	7	7	472
turkeyau	466	6	6	472
sheepau	467	5	5	472
otherau	466	6	6	472
hirelabor	463	9	9	472
fs50_99	456	16	16	472
fs100_249	456	16	16	472
fs250_499	456	16	16	472
fs500	456	16	16	472

Imputed Data vs. Non-Imputed Data

. mi xeq 0 5 10: summarize

Comparison of the Mean Between No Imputation and Multiple Imputations

Variables	**No Imputation**	**MVN Multiple Imputation**	
	m=0	**m=5**	**m=10**
Roundup Ready Corn	0.466	0.464	0.466
Age	53	53	53
Owned Land	235	234	234
Land Rented Out	20	20	20
Land Rented In	170	167	166
Missouri(Base=Iowa)	0.490	0.489	0.489
Non-Family Labor	0.283	0.282	0.284
Environmental Perceptions			
Water Quality	3.994	3.994	3.994
Managing Manure	4.115	4.104	4.113
Global Warming	2.544	2.541	2.541
Farm Sales			
$1-$9,999	2.573	2.571	2.574
$10,000-$49,999	1.718	1.716	1.716
$50,000-$99,999	1.866	1.864	1.864
$100,000-$249,999	1.490	1.490	1.490
$250,000-$499,999	2.210	2.210	2.210
$500,000 or more	2.145	2.145	2.145
Off-Farm Income			
Farm Operator	2.614	2.614	2.614
Spouse	2.842	2.842	2.842
Education			
Farm Operator	2.744	2.758	2.742
Spouse	2.736	2.735	2.801

Imputed Data vs. Non-Imputed Data, cont'd

. mi xeq 0 5 10: summarize

Comparison of the Mean Between No Imputation and Multiple Imputations

Variables	**No Imputation**	**MVN Multiple Imputation**	
	m=0	**m=5**	**m=10**
Roundup Ready Corn	0.496	0.496	0.496
Age	12	12	12
Owned Land	256	257	256
Land Rented Out	103	104	103
Land Rented In	337	338	337
Missouri(Base=Iowa)	0.500	0.500	0.500
Non-Family Labor	0.453	0.452	0.453
Environmental Perceptions			
Water Quality	1.193	1.193	1.197
Managing Manure	1.094	1.094	1.072
Global Warming	1.355	1.355	1.361
Farm Sales			
$1-$9,999	0.343	0.343	0.343
$10,000-$49,999	0.448	0.448	0.448
$50,000-$99,999	0.370	0.370	0.368
$100,000-$249,999	0.412	0.412	0.418
$250,000-$499,999	0.341	0.341	0.341
$500,000 or more	0.269	0.269	0.272
Off-Farm Income			
Farm Operator	1.625	1.625	1.644
Spouse	1.474	1.474	1.498
Education			
Farm Operator	1.151	1.151	1.181
Spouse	1.258	1.258	1.297

MI Regression

```
. mi estimate,  dftable vartable ufmitest  : logistic rrsoybean age lowned lrent
> out lrentin state hirelabor watqual airqual globalwarm fs50_99 fs100_249 fs250
> _499 fs500 offfarmop offfarmsp educop educsp dairycaau beefcaau beefcoau swine
> le55au swinebi55au broilerau turkeyau sheepau otherau
```

```
Logistic regression                              Number of obs     =        472
                                                 Average RVI       =     0.0957
DF adjustment:   Large sample                    DF:     min       =      45.11
                                                         avg       =    3349.41
                                                         max       =   16588.59
Model F test:    Unrestr. FMI                    F(  26, 1109.5)   =       3.56
Within VCE type:          OIM                    Prob > F          =     0.0000
```

MI Regression

Regression Results for Roundup Ready Soybean

Variables	No Imputation			Multivariate Normal Imputation				
	Coeff.	Std.Err.	p-Value	Coeff.	Std.Err.	p-Value	DOF	Inc.S.E.(%)
Age	1.001	0.015	0.948	0.024	0.010	0.021	12358	1.38
Owned Land	1.001	0.001	0.174	0.001	0.001	0.056	639	6.52
Land Rented Out	0.999	0.002	0.819	-0.001	0.001	0.302	4247	2.38
Land Rented In	1.003	0.001	0.005	0.002	0.001	0.032	243	11.29
Missouri (Base=Iowa)	0.319	0.118	0.002	-1.037	0.261	0.000	4210	2.4
Non-Family Labor	1.301	0.494	0.488	-0.260	0.294	0.378	1749	3.79
Environmental Perceptions								
Water Quality	0.746	0.133	0.100	-0.250	0.129	0.053	429	8.13
Managing Manure	1.130	0.209	0.510	0.226	0.151	0.135	140	15.77
Global Warming	0.868	0.111	0.271	-0.135	0.091	0.138	24002	0.98
Farm Sales								
$50,000-$99,999	3.586	1.630	0.005	0.955	0.329	0.004	4274	2.38
$100,000-$249,999	7.554	4.030	0.000	1.368	0.374	0.000	3419	2.67
$250,000-$499,999	16.078	12.982	0.001	2.169	0.569	0.000	653	6.44
$500,000 or more	9.137	11.341	0.075	2.263	0.964	0.019	576	6.91

Setting Data as Survey

```
. mi svyset _n [pweight=weight], strata(strata) fpc(fsc) vce(linearized) singleu
> nit(certainty)

      pweight: weight
          VCE: linearized
  Single unit: certainty
     Strata 1: strata
         SU 1: <observations>
        FPC 1: fsc

. mi estimate, ufmitest  : svy linearized : logistic rrsoybean age lowned lrento
> ut lrentin state hirelabor watqual airqual globalwarm  fs1_9 fs10_49  fs50_99
> fs100_249 fs250_499 fs500 offfarmop offfarmsp educop educsp dairycaau beefcaau
>  beefcoau swinele55au swinebi55au broilerau turkeyau sheepau otherau

Multiple-imputation estimates                    Imputations       =         10
Survey: Logistic regression                      Number of obs     =        456

Number of strata   =        30                   Population size   = 97933.342
Number of PSUs     =       456
                                                 Average RVI       =    0.1559
                                                 Complete DF       =       426
DF adjustment:   Small sample                    DF:     min       =     33.14
                                                         avg       =    265.40
                                                         max       =    411.08
Model F test:    Unrestr. FMI                    F(  27,   211.9)  =      2.65
Within VCE type:   Linearized                    Prob > F          =    0.0001
```

MI Sampling Based Regression

Regression Results for Roundup Ready Soybean

Variables	Multiple Imputation					Multiple Imputation (Sampling Based)				
	Coeff.	Std.Err.	p-Value	DOF	Inc.S.E.(%)	Coeff.	Std.Err.	p-Value	DOF	Inc.S.E.(%)
Age	0.024	0.010	0.021	12358	1.38	0.014	0.013	0.276	374	2.63
Owned Land	0.001	0.001	0.056	639	6.52	0.002	0.001	0.094	88	17.32
Land Rented Out	-0.001	0.001	0.302	4247	2.38	-0.004	0.002	0.050	155	10.54
Land Rented In	0.002	0.001	0.032	243	11.29	0.002	0.001	0.073	56	24.95
Missouri (Base=Iowa)	-1.037	0.261	0.000	4210	2.4	-1.192	0.351	0.001	410	1.51
Non-Family Labor	-0.260	0.294	0.378	1749	3.79	-0.019	0.400	0.962	325	3.95
Environmental Perceptions										
Water Quality	-0.250	0.129	0.053	429	8.13	-0.219	0.152	0.149	365	2.86
Managing Manure	0.226	0.151	0.135	140	15.77	0.159	0.183	0.385	95	16.21
Global Warming	-0.135	0.091	0.138	24002	0.98	-0.102	0.128	0.428	399	1.89
Farm Sales										
$50,000-$99,999	0.955	0.329	0.004	4274	2.38	1.156	0.485	0.018	431	0.64
$100,000-$249,999	1.368	0.374	0.000	3419	2.67	1.720	0.547	0.002	401	1.81
$250,000-$499,999	2.169	0.569	0.000	653	6.44	2.541	0.929	0.007	291	4.93
$500,000 or more	2.263	0.964	0.019	576	6.91	1.921	1.468	0.192	336	3.67

Impact of Missingness on Estimates

Impact of Missing Observations on Variable Estimates

Variables	Multiple Imputation			Multiple Imputation(S.B.)		
	RVI	FMI	Rel.Eff.	RVI	FMI	Rel.Eff.
Age	0.028	0.027	0.997	0.053	0.051	0.995
Owned Land	0.135	0.121	0.988	0.376	0.285	0.972
Land Rented Out	0.048	0.046	0.995	0.222	0.188	0.982
Land Rented In	0.239	0.199	0.980	0.561	0.377	0.964
Missouri (Base=Iowa)	0.048	0.047	0.995	0.030	0.030	0.997
Non-Family Labor	0.077	0.073	0.993	0.081	0.076	0.992
Environmental Perceptions						
Water Quality	0.169	0.149	0.985	0.058	0.055	0.994
Managing Manure	0.340	0.264	0.974	0.351	0.270	0.974
Global Warming	0.020	0.019	0.998	0.038	0.037	0.996
Farm Sales						
$50,000-$99,999	0.048	0.046	0.995	0.013	0.013	0.999
$100,000-$249,999	0.054	0.052	0.995	0.036	0.035	0.996
$250,000-$499,999	0.133	0.120	0.988	0.101	0.094	0.991
$500,000 or more	0.143	0.128	0.987	0.075	0.071	0.993

Conclusions

- Although multiple imputation is a very robust method, care should be given when addressing practical questions.
- When complex survey design is used for data collection, sampling based estimation should be used for more realist standard errors.