A Comparison of Imputation Methods for the ARMS Data

Presentation by: Joshua D. Habiger (U. S. Carolina, NISS, Ok. St. U.)

Joint work with:
Sujit Ghosh (NCSU), Barry Goodwin(NCSU), Darcy Miller(NASS),
Michael Robbins(NISS), Kirk White(ERS), ... and many more

Aug. 2, 2010

Outline

- **Simulation Setup**
- Results
- Conclusion

Why Simulate?

- Difficult to assess imputation method using real data since “true value” is unknown
- Solution: simulation study
  1. **Use synthetic data** with no missing values
  2. **Poke holes in synthetic data**
  3. **Impute for missing values**
  4. **Compare imputed data to original data**
  5. Repeat 1 - 4

Generate Synthetic data?

- Standard simulation approach: **generate synthetic data**
  - Synthetic data should mimic ARMS data
- Problem: Difficult to ensure generated synthetic data mimics ARMS data

Other synthetic data

- Solution:
  - ~~Generate synthetic data~~
  - **Use real data** from nonrefusable items as synthetic data
- Advantage
  - Nonrefusable ARMS data may more closely mimic refusable ARMS data
- “Disadvantage”
  - Results may not apply to more standard non-ARMS like data

For this study...

- Use 24 fully observed variables and poked holes in 6.

Group of Variables
GROSS VALUE OF SALES
REGION
FARM TYPE
TOT. WHEAT HARVESTED
CORN FOR SILAGE
⋮
CORN GRAIN ACRE HARV.
CORN TOT. PRODUCTION

Making Data Missing

- X_{qn} is value of q'th variable for n'th individual (standardized)
- Let

$$\mathit{logit}(\Pr(x_{qn} \text{ is observed})) = \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \ldots \beta_Q x_{Qn}$$

Missingness Mechanism

$$logit(\Pr(x_{qn} \text{ is observed})) = \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \ldots \beta_Q x_{Qn}$$

- Choice of β_q's allows for MCAR, MAR, NMAR
- Example: For x_{1n}
  - MCAR: $\beta_1 = \beta_2 = \ldots = \beta_Q = 0$
  - MAR: $\beta_1 = 0$, but $\beta_2 \neq 0$ or $\beta_3 \neq 0$ or ... or $\beta_Q \neq 0$
  - NMAR: $\beta_1 \neq 0$
- We will look at MCAR, MAR, NMAR with response rate = .5

Impute for Missing Values

- NASS - nearest neighbor type method
- ABB - Approximate Bayesian Bootstrap
- SR2 - Sequential regression w/ Normal model
- SR3 - Sequential regression w/ Skew Normal model
- ISR2 - **Iterative** sequential regression w/ Normal model
- ISR3 - **Iterative** sequential regression w/ Skew Normal model

Goal

- Goal: Impute in a manner s.t. joint distribution structure preserved
- Joint distribution structure metrics (computed on positive portions)
  - mean
  - variance
  - covariance (log scale)

What do we mean “Preserved”?

- x original data and $\hat{x}_k$ k’th imputed data set
- $\theta(x), \theta(\hat{x}_k)$ represent a metric (marginal mean, marginal variance, covariance) computed on $x, \hat{x}_k$

$$\theta(x) \approx \theta(\hat{x}_k)$$

We will compute

$$\textbf{\% change in } \boldsymbol{\theta} = \mathbf{100} \left(\frac{\boldsymbol{\theta(\hat{x}_k) - \theta(x)}}{\boldsymbol{\theta(x)}} \right)$$

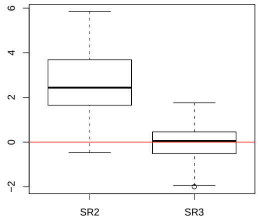
Outline

- Simulation Setup
- **Results**
- Conclusion

MCAR: Skew Normal vs. Normal

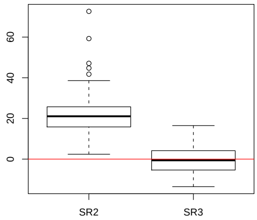
MEAN

Corn Acre Har



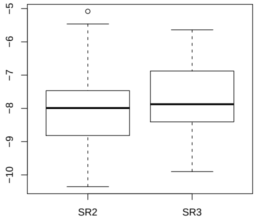
VARIANCE

Corn Acre Har



COVARIANCE

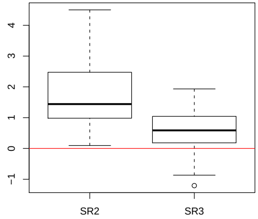
Corn Acre Har vs. Corn Tot Prod



MAR: Skew Normal vs. Normal

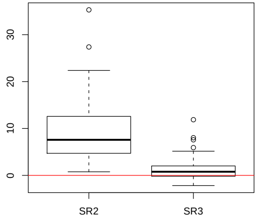
MEAN

Corn Acre Har



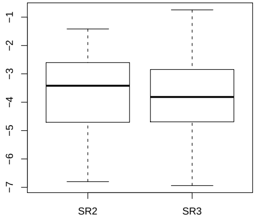
VARIANCE

Corn Acre Har



COVARIANCE

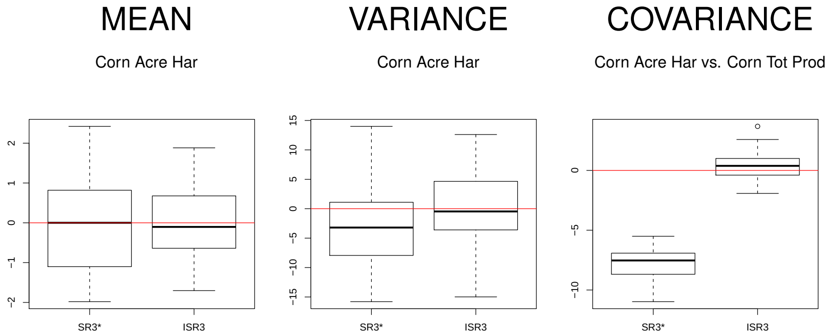
Corn Acre Har vs. Corn Tot Prod



Conclusion 1

- **Skew** Normal model $\gg$ Normal model
  - Difference especially apparent for mean and variance

MCAR: To Iterate or Not To Iterate???



MEAN

Corn Acre Har

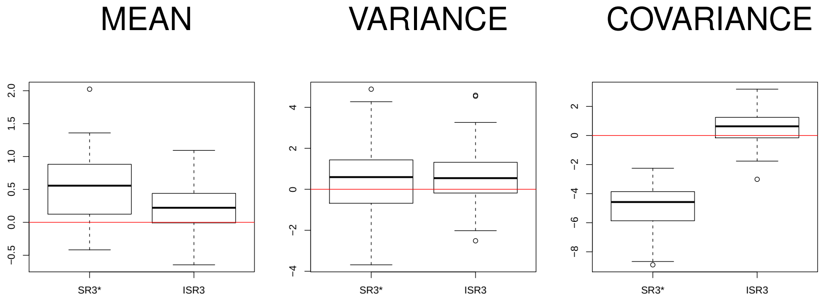
VARIANCE

Corn Acre Har

COVARIANCE

Corn Acre Har vs. Corn Tot Prod

MAR: To Iterate or Not To Iterate???

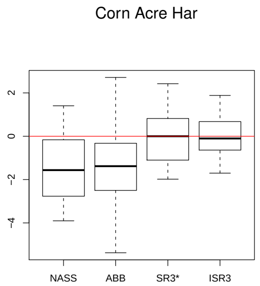


Conclusion 2

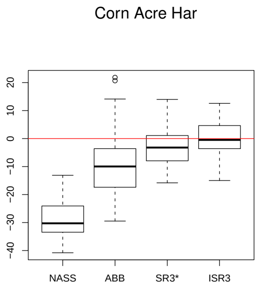
- Iterative SR $\gg$ SR
  - Difference especially apparent for covariance
  - The improvement can be only moderate in less extreme settings

MCAR

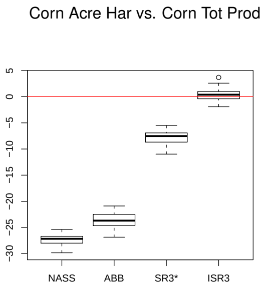
MEAN



VARIANCE



COVARIANCE

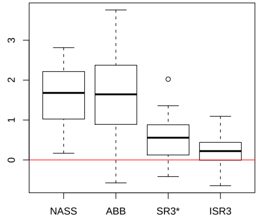


	NASS	ABB	SR	ISR
MEAN	✓	✓	✓	✓
VARIANCE	x	✓	✓	✓
COVARIANCE	x	x	x	✓

MAR

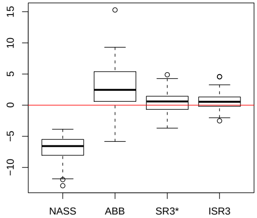
MEAN

Corn Acre Har



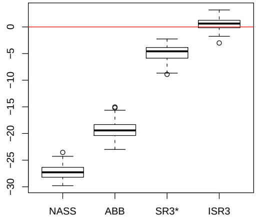
VARIANCE

Corn Acre Har



COVARIANCE

Corn Acre Har vs. Corn Tot Prod

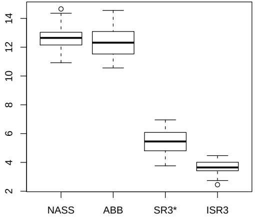


	NASS	ABB	SR	ISR
MEAN	x	x	✓	✓
VARIANCE	x	✓	✓	✓
COVARIANCE	x	x	x	✓

NMAR

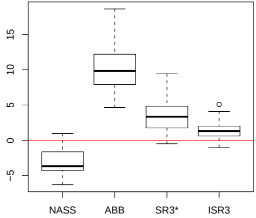
MEAN

Corn Acre Har



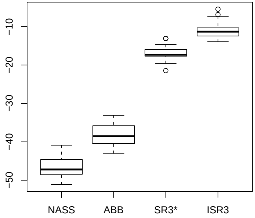
VARIANCE

Corn Acre Har



COVARIANCE

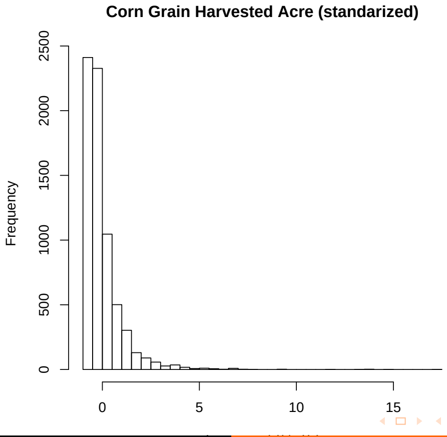
Corn Acre Har vs. Corn Tot Prod



	NASS	ABB	SR	ISR
MEAN	x	x	x	x
VARIANCE	✓	x	✓	✓
COVARIANCE	x	x	x	x

MAR and NMAR Missingness

- Strange behavior?



Outline

- Simulation Setup
- Results
- **Conclusion**

Concluding Remarks

1. Normal $\ll$ Skew Normal
2. NASS$\ll$ ABB $\ll$ SR3 $\ll$ ISR3
3.

	NASS	ABB	SR3	ISR3
mean	0	0	+	+
variance	-	0	+	+
covariance	-	-	0	+

Thank you

Thanks for Listening