A Comparison of Imputation Methods under Large Samples and Different Censoring Levels

Jose A. Lopez, Ph.D.

Assistant Professor of Agribusiness

Agricultural & Applied Economics Association's

2011 AAEA Annual Meeting, Pittsburgh, Pennsylvania, July 24-26, 2011

Outline

1. Introduction
2. Imputer’s Models
3. Analyst’s Models
4. Data and Procedures
5. Results and Discussion
6. Concluding Remarks

Introduction

- Censored observations
  - Survey design, implementation, and institutional constrains
  - Common problem
  - Usually takes place in high proportions
  - The value of an observation is partially known (also called item nonresponse)

Item Nonresponse

- Only on the dependent variable
  - Use of parametric models
  - The probit and tobit models, or their multinomial versions
- Only on an independent variable
  - Several methods and approaches
  - Excluding censored observations, deductive imputation, cell mean imputation, hot-deck imputation, cold-deck imputation, complete case analysis, regression imputation, EM algorithm, MCMC algorithm

Outline

1. Introduction
2. Imputer’s Models
3. Analyst’s Models
4. Data and Procedures
5. Results and Discussion
6. Concluding Remarks

Excluding Censored Observations

- Easy to implement
- It discards incompletely recorded units and focuses only on the completely recorded observations (Little and Rubin 2002)
  – Complete-case analysis
- “It can lead to serious bias, however, and it is not usually very efficient, especially when drawing inferences for subpopulations.” (Little and Rubin 2002, p. 19).

Deductive Imputation

- The researcher deduces the missing value by using logic and the relationships among the variables.
- If the geographical location of a household is missing, it can be recovered by using other variables such as the consecutive order of household interviews and the time period when the household was interviewed.

Cell Mean Imputation

- Zero-order missing price procedure (Cox and Wolgenant 1986)
- Fill-in with means analysis (Little and Rubin 2002)
- It consists of grouping the observations (e.g., households) into classes (e.g., strata and state) and using the non-missing values of the variable of interest (e.g., non-missing prices) to impute the missing values of the variable of interest (e.g., missing prices).
- The more specific the classes are (e.g., strata and county), the more likely the researcher is to obtain an estimate that is closer to the true value.
- The variance in the imputed variable decreases.
- To avoid losing variability in the variable of interest, the researcher may alternatively use the mean and standard deviation from the non-missing values of the variable of interest and generate values for imputation from a normal distribution with this mean and this standard deviation.

Hot Deck Imputation

- The term *hot deck* dates back to the time computer programs and datasets were punched on cards (Lohr1999, p. 275) .
- The card reader used to warm the data cards, so the term *hot deck* was used to refer to the data cards being analyzed.
- Similar to cell mean imputation.

Cold Deck Imputation

- It uses a dataset other than the dataset being analyzed to impute the missing value.
- These datasets may be from a previous survey or from another source.
- Cold deck imputation is common in time series datasets.

Regression Imputation

- Cox and Wohlgenant (1986)
  - First-order missing price procedure
  - It combines cell mean imputation with regression imputation
- Simple regression imputation

Cox and Wohlgenant's (1986)

- First, compute the regional mean prices (mp_i) using the non-missing prices
- Second, calculates the corresponding deviations from the regional mean prices (dmp_i)

$$dpm_i = p_i - mp_i$$

- Third, regresses dmp_i as a function household characteristics

$$dpm_i = \mathbf{z_i}'\boldsymbol{\beta}_\mathbf{i} + e_i$$

- Fourth, the missing prices are imputed

$$\tilde{p}_i = \widehat{dmp}_i + mp_i$$

EM Algorithm

- The EM algorithm finds the MLE of the vector of parameters by iterating two steps until the iterations converge.
- The expectation step (E-step) computes the conditional expectation of the complete-data log likelihood given the observed data and the parameter estimates.

EM Algorithm (Cont.)

- The maximization step (M-step) estimates the parameters that maximize the complete-data log likelihood from the E-step
- The observed-data log likelihood being maximized can be expressed as follows
- $$\log L(\boldsymbol{\theta} \mid X_{obs}) = \sum_{g=1}^{G} \log L_g(\boldsymbol{\theta} \mid X_{obs})$$
- $$\log L_g(\boldsymbol{\theta} \mid X_{obs}) = -\frac{n_g}{2} \log |\Sigma_g| - \frac{1}{2} \sum_{hg} (\mathbf{x}_{hg} - \boldsymbol{\mu}_g)' \Sigma_g^{-1} (\mathbf{x}_{hg} - \boldsymbol{\mu}_g)$$
- G = number of groups with distinct missing patterns
- $\log L(\boldsymbol{\theta}|X_{obs})$ = the observed-data log likelihood from the g^{th} group
- n_g = the number of observations in the g^{th} group
- The summation is over the household observations in the g^{th} group
- $\mathbf{x}_{hg}$ = a vector of observed values corresponding to observed variables
- $\boldsymbol{\mu}_g$ = the mean vector
- $\sum_g$ = the associated covariance matrix.

MCMC Algorithm

- The Markov Chain Monte Carlo (MCMC) has applications in Bayesian inference.
- This approach consists of a data augmentation procedure that is implemented in two steps.
- The imputation step (I-step) draws values for X_{mis} from a conditional predictive distribution of X_{mis} given X_{obs}.
- That is, with a current estimate of $\boldsymbol{\theta}^{(t)}$ at the t^{th} iteration,
- $\boldsymbol{\theta}^{(t+1)} \sim \Pr(X_{mis} \mid X_{obs}, \boldsymbol{\theta}^{(t)})$

MCMC Algorithm (Cont.)

- The posterior step (P-step) draws values for **θ** from a conditional distribution of **θ** given X_{obs}
- $\boldsymbol{\theta}^{(t+1)} \sim \Pr\left(\boldsymbol{\theta} \mid X_{obs}, X_{mis}^{(t+1)}\right)$
- The two steps are iterated creating a Markov chain
- $\left(X_{mis}^{(1)}, \boldsymbol{\theta}^{(1)}\right), \left(X_{mis}^{(2)}, \boldsymbol{\theta}^{(2)}\right), \ldots$
- which converges in distribution to $\Pr(X_{mis}, \boldsymbol{\theta} \mid X_{obs})$

Outline

1. Introduction
2. Imputer's Models
3. Analyst's Models
4. Data and Procedures
5. Results and Discussion
6. Concluding Remarks

Almost Ideal Demand System (AIDS)

- The Marshallian demand function for commodity *i* in share form is specified as
- $$w_{ih} = \alpha_i + \sum_j \gamma_{ij} \log(p_{jh}) + \beta_i \log\left(\frac{m_h}{P_h}\right) + \varepsilon_{ih}$$
- w_{ih} = the budget share for commodity *i* and household *h*
- p_{jh} = the price of commodity *j* and household *h*
- m_h = total household expenditure on the commodities being analyzed
- α_i, β_i and γ_{ij} = parameters
- ε_i = a random term of disturbances
- P_h = a price index

AIDS (Cont.)

- In a nonlinear approximation, the price index P_h is defined as
- $$\log(P_h) = \alpha_0 + \sum_k \alpha_k \log(p_{kh}) + \frac{1}{2}\sum_k \sum_j \gamma_{kj} \log(p_{kh}) \log(p_{jh})$$
- The demand theory properties of adding-up, homogeneity and symmetry can be imposed on the system of equations by restricting parameters in the model as follows
- Adding-up: $\sum_i \alpha_i = 1, \quad \sum_j \gamma_{ij} = 0, \quad \sum_i \beta_i = 0$
- Homogeneity: $\sum_i \gamma_{ij} = 0$
- Symmetry: $\gamma_{ij} = \gamma_{ji}$

AIDS (Cont.)

- The Marshallian (uncompensated) and the Hicksian (compensated) price elasticities as well as the expenditure elasticities can be computed from the estimated coefficients

- Marshallian Price Elasticity

$$e_{ij} = -\delta_{ij} + \frac{\gamma_{ij}}{w_i} - \frac{\beta_i}{w_i}\left(\alpha_j + \sum_k \gamma_{kj} \ln p_k\right)$$

- Hicksian Price Elasticity

$$e_{ij}^c = e_{ij} + w_j e_i$$

- Expenditure Elasticity

$$e_i = 1 + \frac{\beta_i}{w_i}$$

Outline

1. Introduction
2. Imputer's Models
3. Analyst's Models
4. Data and Procedures
5. Results and Discussion
6. Concluding Remarks

ENIGH

- Mexican data on household income and weekly expenditures
  - *Encuesta Nacional de Ingresos y Gastos de los Hogares* (2008)
- Seven food sources of protein were analyzed in this study
  - i = 1, 2, …7, where 1 = meat, 2 = dairy, 3 = eggs, 4 = tubers, 5 = vegetables, 6 = legumes, 7 = fruits
- A subsample of 3,572 households
- p_i , i = 1, …, 7, were randomly censored at two levels
  - 30% censoring level
    - 2,500 non-missing price observations
    - 1,072 censored price observations
  - 70% censoring level
    - 1,072 non-missing price observations
    - 2,500 censored price observations
- Only one missing data pattern is considered (i.e., all prices were censored for the same instance).
- q_i , i = 1, …, 7, are NOT censored

Outline

1. Introduction
2. Imputer's Models
3. Analyst's Models
4. Data and Procedures
5. Results and Discussion
6. Concluding Remarks

Imputation Methods and Approaches

- Excluding censored observations (ECO)
- Cell mean imputation (CM)
- Cox and Wohlgenant’s first-order missing price procedure (CW)
- Simple regression imputation (SR)
- The EM algorithm
- The MCMC algorithm

Variable	Description
p00_11	Household members who are less than 12 years old.
p12_64	Household members who are or are between 12 and 64 years old.
p65_more	Household members who are or are older than 65 years old.
inc	Household income.
rural	“1” for household locations with a population of 14,999 people or less and “0” if otherwise.
urban	“1” for household locations with a population of 15,000 people or more and “0” if otherwise.
element	“1” if the household decision maker has elementary school education or less and “0” if otherwise.
highsch	“1” if the household decision maker has high school education or if he/she is a high school graduate and “0” if otherwise.
college	“1” if the household decision maker has some college, college or incomplete university education and “0” if otherwise.
university	“1” if the household decision maker has completed university or has some graduate school education and “0” if otherwise.
NE	“1” if the household is located in the Northeast region of Mexico and “0” if otherwise.
NW	“1” if the household is located in the Northwest region of Mexico and “0” if otherwise.
CW	“1” if the household is located in the Central-West region of Mexico and “0” if otherwise.
C	“1” if the household is located in the Central region of Mexico and “0” if otherwise.
SE	“1” if the household is located in the Southeast region of Mexico and “0” if otherwise.
d_car	“1” if the household has a 4-wheel vehicle and “0” if otherwise.
d_refri	“1” if the household has a refrigerator at home and “0” if otherwise.
supermkt	“1” if the household purchased the protein product or commodity from a supermarket and “0” if somewhere else.

Observed and Imputed Prices
VARIABILITY

p_i	No Censoring		30 % Censoring Level												
	Observed Prices		Excluding Cen. Obs.		Cell Mean		Cox & Wohlgenant		Simple Regression		EM Algorithm		MCMC Algorithm	
	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean
p_1	46.4608	0.3650	47.0064	0.4462	47.0064	0.3071	47.0651	0.3141	46.9953	0.3124	46.9953	0.3124	46.9948	0.3123
p_2	23.7807	0.4708	23.9239	0.5504	23.9239	0.3785	23.7270	0.3893	23.8325	0.3874	23.8325	0.3874	23.8344	0.3874
p_3	18.7620	0.1311	18.8758	0.1769	18.8758	0.1216	18.8716	0.1252	18.8804	0.1242	18.8804	0.1242	18.8810	0.1242
p_4	15.5820	0.5964	16.0031	0.7511	16.0031	0.5165	16.0858	0.5219	16.0884	0.5180	16.0884	0.5180	16.0860	0.5180
p_5	13.3280	0.1362	13.1985	0.1662	13.1985	0.1143	13.2242	0.1189	13.2155	0.1173	13.2155	0.1173	13.2162	0.1173
p_6	18.6618	0.2500	18.4720	0.2282	18.4720	0.1571	18.4876	0.1615	18.5022	0.1591	18.5022	0.1591	18.5021	0.1591
p_7	10.3969	0.1455	10.4638	0.1685	10.4638	0.1159	10.4885	0.1184	10.4776	0.1177	10.4776	0.1177	10.4770	0.1177

p_i	No Censoring		70 % Censoring Level												
	Observed Prices		Exculdign Cen. Obs.		Cell Mean		Cox & Wohlgenant		Simple Regression		EM Algorithm		MCMC Algorithm	
	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean
p_1	46.4608	0.3650	45.2598	0.6193	45.2598	0.1938	45.3959	0.2255	45.3696	0.2156	45.3696	0.2156	45.3730	0.2156
p_2	23.7807	0.4708	23.4655	0.8953	23.4655	0.2794	23.9321	0.3333	23.6935	0.3108	23.6935	0.3108	23.6877	0.3107
p_3	18.7620	0.1311	18.5115	0.1558	18.5115	0.0487	18.5492	0.0568	18.4960	0.0547	18.4960	0.0547	18.4973	0.0547
p_4	15.5820	0.5964	14.6550	0.9537	14.6550	0.2977	14.5172	0.3249	14.5298	0.3079	14.5298	0.3079	14.5285	0.3079
p_5	13.3280	0.1362	13.6131	0.2372	13.6131	0.0740	13.6248	0.0844	13.6234	0.0834	13.6234	0.0834	13.6229	0.0834
p_6	18.6618	0.2500	19.0796	0.6189	19.0796	0.1937	18.8062	0.2198	19.0082	0.2119	19.0082	0.2119	19.0097	0.2119
p_7	10.3969	0.1455	10.2498	0.2817	10.2498	0.0879	10.1045	0.0972	10.2020	0.0926	10.2020	0.0926	10.2015	0.0926

Note: p_i, i = 1, 2, …,7, where 1 = meat, 2 = dairy, 3 = eggs, 4 = tubers, 5 = vegetables, 6 = legumes, and 7 = fruits.

Observed and Imputed Prices

BEST ESTIMATES FROM SIMPLE COMPARISON (NOT RECOMENDED)

	No Censoring		30 % Censoring Level											
	Observed Prices		Excluding Cen. Obs.		Cell Mean		Cox & Wohlgenant		Simple Regression		EM Algorithm		MCMC Algorithm	
p_i	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean
p_1	46.4608	0.3650	47.0064	0.4462	47.0064	0.3071	47.0651	0.3141	46.9953	0.3124	46.9953	0.3124	46.9948	0.3123
p_2	23.7807	0.4708	23.9239	0.5504	23.9239	0.3785	23.7270	0.3893	23.8325	0.3874	23.8325	0.3874	23.8344	0.3874
p_3	18.7620	0.1311	18.8758	0.1769	18.8758	0.1216	18.8716	0.1252	18.8804	0.1242	18.8804	0.1242	18.8810	0.1242
p_4	15.5820	0.5964	16.0031	0.7511	16.0031	0.5165	16.0858	0.5219	16.0884	0.5180	16.0884	0.5180	16.0860	0.5180
p_5	13.3280	0.1362	13.1985	0.1662	13.1985	0.1143	13.2242	0.1189	13.2155	0.1173	13.2155	0.1173	13.2162	0.1173
p_6	18.6618	0.2500	18.4720	0.2282	18.4720	0.1571	18.4876	0.1615	18.5022	0.1591	18.5022	0.1591	18.5021	0.1591
p_7	10.3969	0.1455	10.4638	0.1685	10.4638	0.1159	10.4885	0.1184	10.4776	0.1177	10.4776	0.1177	10.4770	0.1177

	No Censoring		70 % Censoring Level											
	Observed Prices		Exculdign Cen. Obs.		Cell Mean		Cox & Wohlgenant		Simple Regression		EM Algorithm		MCMC Algorithm	
p_i	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean	Mean (Pesos/Kg)	Std. Err. of Mean
p_1	46.4608	0.3650	45.2598	0.6193	45.2598	0.1938	45.3959	0.2255	45.3696	0.2156	45.3696	0.2156	45.3730	0.2156
p_2	23.7807	0.4708	23.4655	0.8953	23.4655	0.2794	23.9321	0.3333	23.6935	0.3108	23.6935	0.3108	23.6877	0.3107
p_3	18.7620	0.1311	18.5115	0.1558	18.5115	0.0487	18.5492	0.0568	18.4960	0.0547	18.4960	0.0547	18.4973	0.0547
p_4	15.5820	0.5964	14.6550	0.9537	14.6550	0.2977	14.5172	0.3249	14.5298	0.3079	14.5298	0.3079	14.5285	0.3079
p_5	13.3280	0.1362	13.6131	0.2372	13.6131	0.0740	13.6248	0.0844	13.6234	0.0834	13.6234	0.0834	13.6229	0.0834
p_6	18.6618	0.2500	19.0796	0.6189	19.0796	0.1937	18.8062	0.2198	19.0082	0.2119	19.0082	0.2119	19.0097	0.2119
p_7	10.3969	0.1455	10.2498	0.2817	10.2498	0.0879	10.1045	0.0972	10.2020	0.0926	10.2020	0.0926	10.2015	0.0926

Note: p_i, i = 1, 2, …,7, where 1 = meat, 2 = dairy, 3 = eggs, 4 = tubers, 5 = vegetables, 6 = legumes, and 7 = fruits.

Root Mean Square Error (RMSE) and Root Mean Square Percent Error (RMSPE)

- A simple comparison of the mean prices obtained from the dataset with no censored prices with the mean prices obtained from the various imputation approaches is inappropriate because positive errors would cancel out with negative errors.
- To appropriately evaluate which method generated the best imputations, the RMSE and the RMSPE for price p_i are defined as
- $$RMSE = \sqrt{\frac{1}{(H*l)}\sum_{h=1}^{H*l}\left(p_{ih}^{imputed} - p_{ih}^{actual}\right)^2}$$
- $$RMSPE = \sqrt{\frac{1}{(H*l)}\sum_{h=1}^{H*l}\left(\frac{p_{ih}^{imputed} - p_{ih}^{actual}}{p_{ih}^{actual}}\right)^2}$$
- Similar definitions are used for the Marshallian and Hicksian price elasticities as well as the expenditure elasticities.

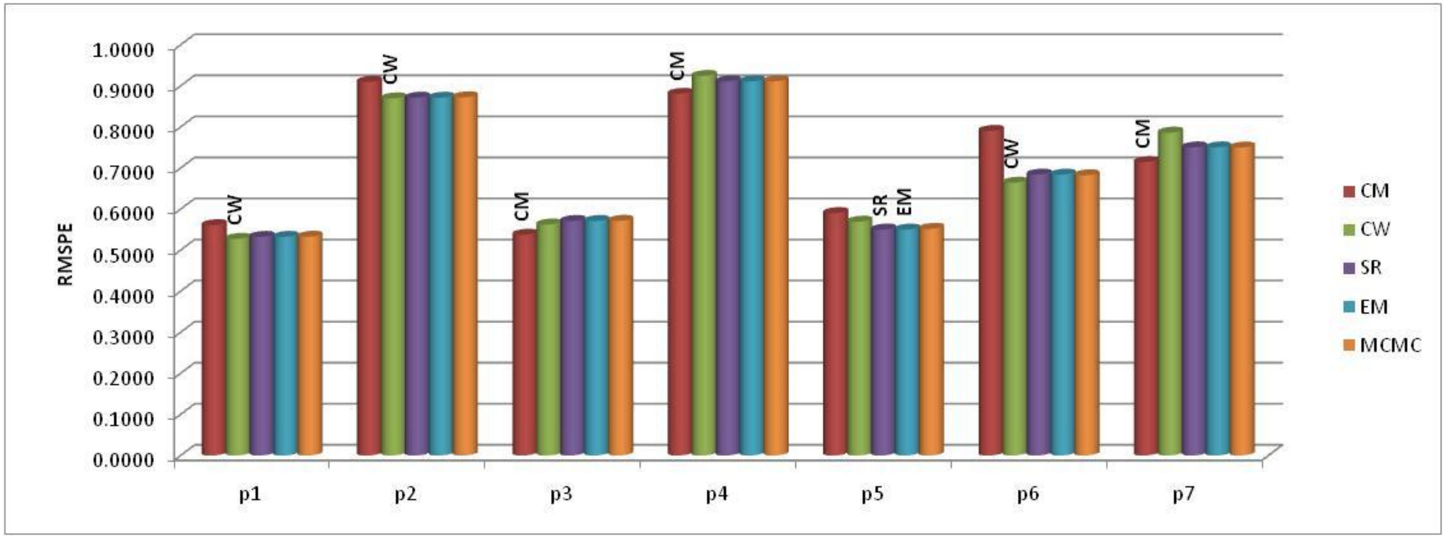
RMSE and RMSPE for Imputed Prices **(RECOMMENDED)**

	30% Censoring									
	CM		CW		SR		EM		MCMC	
	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE
p_1	15.9498	0.5609	15.0249	0.5277	15.1083	0.5325	15.1083	0.5325	15.1139	0.5328
p_2	23.6157	0.9100	22.4628	0.8696	22.4946	0.8713	22.4946	0.8713	22.5092	0.8724
p_3	4.6705	0.5376	4.4238	0.5624	4.4348	0.5711	4.4348	0.5711	4.4406	0.5716
p_4	22.1532	0.8809	21.8287	0.9245	22.0666	0.9111	22.0666	0.9111	22.0679	0.9113
p_5	6.0702	0.5903	5.7229	0.5693	5.8029	0.5502	5.8029	0.5502	5.8044	0.5520
p_6	9.4277	0.7907	9.2105	0.6643	9.2567	0.6841	9.2567	0.6841	9.2574	0.6825
p_7	6.2683	0.7147	6.2678	0.7862	6.2593	0.7504	6.2593	0.7504	6.2635	0.7500
Overall	38.5966	1.9215	37.1921	1.8945	37.4087	1.8796	37.4087	1.8796	37.4223	1.8802
	70% Censoring									
	CM		CW		SR		EM		MCMC	
	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE
p_1	15.4196	0.5142	15.2015	0.5040	15.1526	0.5082	15.1525	0.5082	15.1572	0.5083
p_2	22.6790	0.9412	21.8412	0.9802	21.7891	0.9366	21.7891	0.9366	21.7817	0.9365
p_3	9.1615	0.6595	8.9020	0.6733	8.9764	0.6827	8.9764	0.6827	8.9763	0.6818
p_4	29.3571	0.9543	29.4960	1.0555	29.5222	1.0570	29.5222	1.0570	29.5311	1.0588
p_5	6.5642	0.5079	6.3488	0.5125	6.4298	0.5079	6.4298	0.5079	6.4293	0.5077
p_6	9.8939	0.8132	10.2613	0.6832	10.3302	0.7581	10.3301	0.7581	10.3298	0.7570
p_7	9.2151	0.7564	9.1513	0.7763	9.1047	0.7508	9.1047	0.7508	9.1035	0.7508
Overall	43.8608	1.9968	43.4365	2.0284	43.4447	2.0285	43.4447	2.0285	43.4483	2.0287

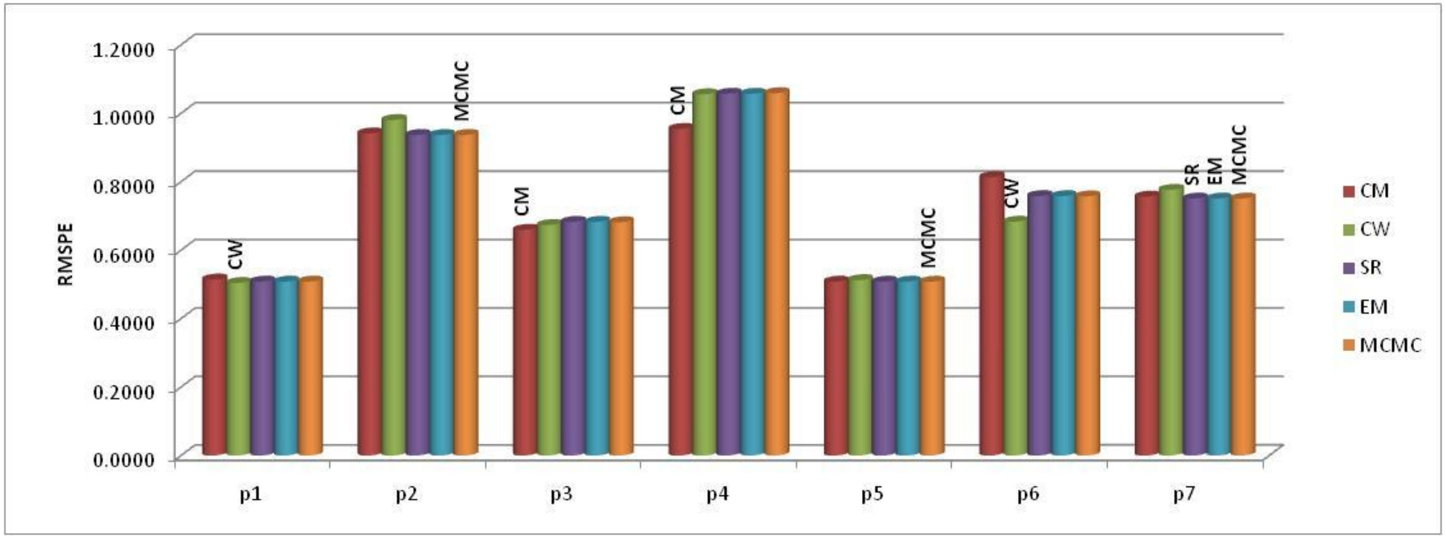
Note: p_i, i **=** 1, 2, …,7, where 1 = meat, 2 = dairy, 3 = eggs, 4 = tubers, 5 = vegetables, 6 = legumes, and 7 = fruits.

Best Estimates from RMSPE Comparison

(a) 30% Censoring Level



(b) 70% Censoring Level



Marshallian Own-Price Elasticity Estimates Under 0%, 30%, and 70% Censoring Levels.

	No Censoring	30% Censoring					70% Censoring				
		ECO	CM	CW	EM	MCMC	ECO	CM	CW	EM	MCMC
e_{11}	-0.9300	-0.9267	-0.9120	-0.9288	-0.9189	-0.9184	-0.9412	-0.9035	-0.9472	-0.8995	-0.8999
e_{22}	-1.1009	-1.1216	-1.0772	-1.1050	-1.1102	-1.1097	-1.0532	-0.9946	-1.0067	-1.0437	-1.0444
e_{33}	-0.6560	-0.6783	-0.6487	-0.6292	-0.6360	-0.6353	-0.5835	-0.4990	-0.4441	-0.4615	-0.4624
e_{44}	-0.8196	-0.8312	-0.7960	-0.7527	-0.7661	-0.7648	-0.7816	-0.7479	-0.4537	-0.4244	-0.4240
e_{55}	-0.8924	-0.9227	-0.9138	-0.9256	-0.9211	-0.9211	-0.8313	-0.8230	-0.8574	-0.8379	-0.8377
e_{66}	-0.6477	-0.6289	-0.6098	-0.6043	-0.6172	-0.6170	-0.7132	-0.6976	-0.5750	-0.5815	-0.5810
e_{77}	-0.7998	-0.8063	-0.8089	-0.7862	-0.7917	-0.7920	-0.7797	-0.7851	-0.6921	-0.7873	-0.7873

Note: e_{ij}, $i = j = 1, 2, \ldots, 7$, where 1 = meat, 2 = dairy, 3 = eggs, 4 = tubers, 5 = vegetables, 6 = legumes, and 7 = fruits.

Marshallian Own-Price Elasticity Estimates
Under 0%, 30%, and 70% Censoring Levels.
BEST ESTIMATES FROM SIMPLE COMPARISON **(NOT RECOMMENDED)**

	No Censoring	30% Censoring					70% Censoring				
		ECO	CM	CW	EM	MCMC	ECO	CM	CW	EM	MCMC
e_{11}	-0.9300	-0.9267	-0.9120	$\boxed{-0.9288}$	-0.9189	-0.9184	$\boxed{-0.9412}$	-0.9035	-0.9472	-0.8995	-0.8999
e_{22}	-1.1009	-1.1216	-1.0772	$\boxed{-1.1050}$	-1.1102	-1.1097	$\boxed{-1.0532}$	-0.9946	-1.0067	-1.0437	-1.0444
e_{33}	-0.6560	-0.6783	-0.6487	$\boxed{-0.6292}$	-0.6360	-0.6353	$\boxed{-0.5835}$	-0.4990	-0.4441	-0.4615	-0.4624
e_{44}	-0.8196	$\boxed{-0.8312}$	-0.7960	-0.7527	-0.7661	-0.7648	$\boxed{-0.7816}$	-0.7479	-0.4537	-0.4244	-0.4240
e_{55}	-0.8924	-0.9227	$\boxed{-0.9138}$	-0.9256	-0.9211	-0.9211	-0.8313	-0.8230	$\boxed{-0.8574}$	-0.8379	-0.8377
e_{66}	-0.6477	$\boxed{-0.6289}$	-0.6098	-0.6043	-0.6172	-0.6170	-0.7132	$\boxed{-0.6976}$	-0.5750	-0.5815	-0.5810
e_{77}	-0.7998	$\boxed{-0.8063}$	-0.8089	-0.7862	-0.7917	-0.7920	-0.7797	-0.7851	-0.6921	$\boxed{-0.7873}$	$\boxed{-0.7873}$

Note: e_{ij}, $i = j = 1, 2, \ldots, 7$, where 1 = meat, 2 = dairy, 3 = eggs, 4 = tubers, 5 = vegetables, 6 = legumes, and 7 = fruits.

RMSE and RMSPE for After-Imputation Marshallian Own-Price Elasticity Estimates **(RECOMMENDED)**

TEXAS A&M UNIVERSITY COMMERCE

	30% Censoring							
	CM		CW		EM		MCMC	
	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE
e_{11}	0.1113	0.3157	0.0736	0.1418	0.0911	0.2216	0.0928	0.2293
e_{22}	0.2086	0.1450	0.2392	0.1678	0.2663	0.1866	0.2679	0.1865
e_{33}	0.2644	8.3271	0.2972	13.0268	0.2854	11.6893	0.2869	11.8010
e_{44}	0.6481	5.5709	0.7454	4.8520	0.8006	5.6926	0.8078	5.7448
e_{55}	0.2019	3.1883	0.1551	2.3391	0.1671	2.6428	0.1664	2.6433
e_{66}	0.8954	20.7972	1.0144	25.9407	0.9566	20.6203	0.9566	20.6394
e_{77}	0.3860	8.4851	0.4097	8.9000	0.4028	7.6417	0.4018	7.2458
All e_{ij}, i = j	1.2399	24.8028	1.3884	30.8365	1.3809	25.6848	1.3854	25.6482
All e_{ij}, i, j = 1, ..., 7	267.0056	396.0614	267.0065	971.9238	267.0070	694.2498	267.0070	683.6587

	70% Censoring							
	CM		CW		EM		MCMC	
	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE
e_{11}	0.1070	0.3619	0.1070	0.3619	0.1138	0.4675	0.1124	0.4585
e_{22}	0.3186	0.2096	0.3186	0.2096	0.2513	0.1339	0.2447	0.1328
e_{33}	0.7800	21.5138	0.7800	21.5138	0.8629	24.6745	0.8663	25.9810
e_{44}	0.9381	9.8106	0.9381	9.8106	2.5436	31.5370	2.5802	32.1871
e_{55}	0.5910	12.7837	0.5910	12.7837	0.4668	10.1846	0.4746	10.4610
e_{66}	0.4019	38.0428	0.4019	38.0428	0.5798	26.6884	0.5920	29.7259
e_{77}	1.2067	42.6318	1.2067	42.6318	1.1940	38.0270	1.2246	37.1436
All e_{ij}, i = j	1.8890	63.1460	1.8890	63.1460	3.0447	62.1747	3.0913	63.9058
All e_{ij}, i, j = 1, ..., 7	267.0114	994.2068	267.0114	994.2068	267.0317	2167.6406	267.0326	2212.9051

Note: e_{ij}, $i = j = 1, 2, \ldots, 7$, where 1 = meat, 2 = dairy, 3 = eggs, 4 = tubers, 5 = vegetables, 6 = legumes, and 7 = fruits.

RMSE and RMSPE for After-Imputation Elasticity Estimates
SUMMARY

	30% Censoring							
	CM		CW		EM		MCMC	
	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE
AIDS Parameters	0.0044	0.8789	0.0029	1.4332	0.0032	1.1755	0.0032	1.1617
Expenditure Elasticities	1.8777	39.3499	1.6597	35.2016	1.7649	31.6429	1.7683	32.3191
Marshallian Elasticities	267.0056	396.0614	267.0065	971.9238	267.0070	694.2498	267.0070	683.6587
Hicksian Elasticities	267.0100	1767.3065	267.0103	1884.7531	267.0117	2204.6346	267.0118	2236.7423
Overall Elasticities	377.6107	1811.5699	377.6106	2120.8888	377.6124	2311.5791	377.6125	2339.1131
	70% Censoring							
	CM		CW		EM		MCMC	
	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE	RMSE	RMSPE
AIDS Parameters	0.009	2.290	0.008	5.778	0.008	4.938	0.008	4.950
Expenditure Elasticities	3.329	68.857	3.329	68.857	2.330	75.951	2.352	76.221
Marshallian Elasticities	267.011	994.207	267.011	994.207	267.032	2,167.641	267.033	2,212.905
Hicksian Elasticities	267.024	1,561.863	267.024	1,561.863	267.044	4,890.050	267.045	4,795.167
Overall Elasticities	377.6349	1852.7291	377.6349	1852.7291	377.6556	5349.4884	377.6572	5281.7028

Outline

1. Introduction
2. Imputer's Models
3. Analyst's Models
4. Data and Procedures
5. Results and Discussion
6. Concluding Remarks

Concluding Remarks

- Even when there was small variability among the imputer's models, relatively larger variability was found from the analyst's model.
- A "simple comparison" of the mean prices or elasticities is inappropriate because positive errors would cancel out with negative errors; therefore, it is recommended to compute the RMSE & RMSPE.
  - ECO approach excluded
  - ECO approach may be unfeasible when a 30% censoring occurs in each price at different times (i.e., the complete-case data may have few observations).
- The imputation method or approach that provides the best estimates varies across the imputed variables (i.e., p_i, $i = 1, 2, \ldots, 7$) and across the ultimately desired measures (i.e., e_{ij}, e_i, e^c_{ij}, $i, j = 1, 2, \ldots, 7$).

Concluding Remarks

- Results are sensitive to censoring levels
  - At high levels of censoring (e.g., 70%), a simple method (e.g., CM) may provide satisfactory or even better estimates than sophisticated methods.
- It is recommended that a portion of the dataset is set aside for validation purposes and the imputation method that would be chosen is selected from an analysis from the ultimately desired measures.

Thank You!

Protein Sources and ENIGH (2008) Codes

- MEAT
  - BEEF = A025-A037
  - PORK = A038-052
  - PROCESSED MEAT = A053-A056
  - CHICKEN = A057-A061
  - PROCESSED POULTY MEAT = A062
  - OTHER MEAT = A063-A065
  - FRESH FISH = A066-A067
  - SHELLFISH = A072-A074
- DAIRY
  - MILK = A075-A081
  - CHEESE = A082-A088
  - OTHER MILK DERIVED PRODUCTS = A089-A092
- EGSS
  - EGGS = A093-A094
- TUBERS
  - RAW OR FRESH TUBERS = A101-A104
  - PROCESSED TUBERS = A105-A106
- VEGETABLES
  - FRESH AND POD VEGETABLES = A107-A132
  - PROCESSED AND POD VEGETABLES = A133-A136
- LEGUMES
  - LEGUMES = A137-A141
  - PROCESSED LEGUMES = A142-A143
- FRUITS
  - FRESH FRUITS = A147-A170
  - PROCESSED FRUITS = A171-A172

ENIGH (2008) Codes (Cont.)

2. DAIRY

A075	Leche pasteurizada de vaca
A076	Leche condensada
A077	Leche evaporada
A078	Leche en polvo entera o descremada
A079	Leche modificada o maternizada
A080	Leche no pasteurizada (leche bronca)
A081	Otras leches: de burra, de cabra, de soya
	QUESOS
A082	Queso amarillo en rebanadas o para untar
A083	Queso añejo y cotija
A084	Queso chihuahua
A085	Queso fresco
A086	Queso manchego
A087	Queso oaxaca o asadero
A088	Otros quesos
	OTROS DERIVADOS DE LA LECHE
A089	Crema
A090	Mantequilla
A091	Bebidas fermentadas de leche
A092	Otros derivados de la leche

3. EGGS

A093	Huevo de gallina blanco y rojo
A094	Otros huevos: codorniz, pata, pava etcétera

4. TUBERS

A101	Betabel y camote
A102	Papa
A103	Rábano
A104	Otros tubérculos
A105	Harina para puré de papa
A106	Papas fritas en bolsa o a granel

6. LEGUMES

A137	Frijol en grano
A138	Garbanzo en grano
A139	Haba amarilla o verde en grano
A140	Lenteja en grano
A141	Otras leguminosas en grano
A142	Frijol procesado
A143	Otras leguminosas procesadas

ENIGH (2008) Codes (Cont.)

5. VEGETABLES

A107	Acelgas, espinacas y verdolagas
A108	Aguacate
A109	Ajo
A110	Brócoli
A111	Calabacita y calabaza
A112	Cebolla
A113	Chayote
A114	Chícharo
A115	Chile jalapeño
A116	Chile poblano
A117	Chile serrano
A118	Otros chiles
A119	Cilantro
A120	Col y repollo
A121	Ejote
A122	Elote
A123	Epazote
A124	Jitomate
A125	Lechuga
A126	Nopal
A127	Pepino
A128	Perejil y yerbabuena
A129	Tomate verde
A130	Zanahoria
A131	Otras verduras
A132	Germinados de maíz, de soya, de trigo
A133	Chiles envasados
A134	Chile secos o en polvo
A135	Verduras y legumbres envasadas
A136	Verduras y legumbres congeladas

ENIGH (2008) Codes (Cont.)

6. FRUITS

A147	Anona, chirimoya, guanábana
A148	Cereza, frambuesa, fresa, zarzamora
A149	Chabacano, durazno, melocotón
A150	Chicozapote y mamey
A151	Ciruela y jobo
A152	Guayaba
A153	Lima
A154	Limón
A155	Mandarina, nectarina, tangerina
A156	Toronja
A157	Mango
A158	Manzana y perón
A159	Melón
A160	Naranja
A161	Papaya
A162	Pera
A163	Piña
A164	Pitahaya y tuna
A165	Plátano macho y de castilla
A166	Plátano verde y tabasco
A167	Otros plátanos (chiapas, dominico, guineo, manzano, dorado
A168	Sandía
A169	Uva
A170	Otras frutas: garambullo, granada, higo, jícama, kiwi, etcéter
A171	Frutas en almíbar y conserva
A172	Frutas cristalizadas, enchiladas y secas