United States Department of Agriculture

Estimating Farm Household Statistics Under Decreasing Response Rates

23rd Pacioli Conference
September 27-30
Belgrade, Serbia

Daniel Prager, Ph.D.
Economic Research Service
United States Department of Agriculture

The views expressed are those of the author and should not be attributed to the Economic Research Service or USDA

What is the issue?

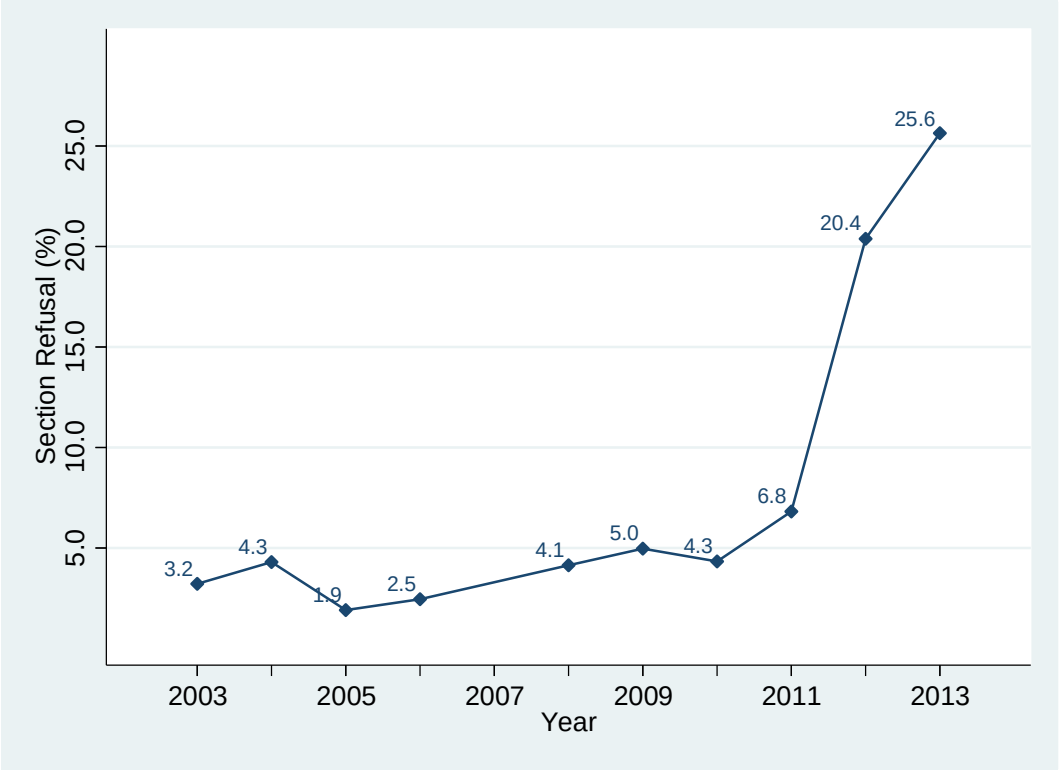
- Current ERS imputation methodology uses conditional mean
  - Outdated statistical method with well-known issues
  - National Research Council (2008) review of ARMS recommended exploring multivariate methods for imputation

- Research Questions:
  - Survey methods: How can we improve the response rate within the household section
  - Statistical methods: Can we improve on the existing imputation methodology?
  - Applied: Do measures of household well-being change significantly with new imputation method? How much?

What is the issue? (cont'd)

- The Agricultural Resource Management Survey (ARMS) is a complex survey administered by the United States Department of Agriculture (USDA)
  - Jointly undertaken by:
    - National Agricultural Statistical Service – NASS (statistics)
    - Economic Research Service – ERS (economics)
  - USDA's primary source of information on financial condition and production practices of nation's farm households
    - Sample size of ~17,000 to ~29,000 usable surveys, depending on year
  - Full datasets available to academic and other researchers
  - Survey suffers from non-response:
    - Similar non-response to other federal surveys
    - Unit, Section, Item refusals
- ERS imputes for missing data in Household Section (HH) of ARMS

Urgency: HH section refusal rate increased in 2012 and 2013



ERS Is Working to Improve HH Non-Response

- Respondents are reluctant to answer the household section
  - Household section is last section of 1-2 hour survey
  - Privacy concerns may dissuade some respondents
    - ARMS has a very good track record on maintaining confidentiality
  - Importance of household data may be unclear in a farm costs and returns survey
  - ARMS became an all-mail survey in 2012, with enumerated follow-up for non-response*

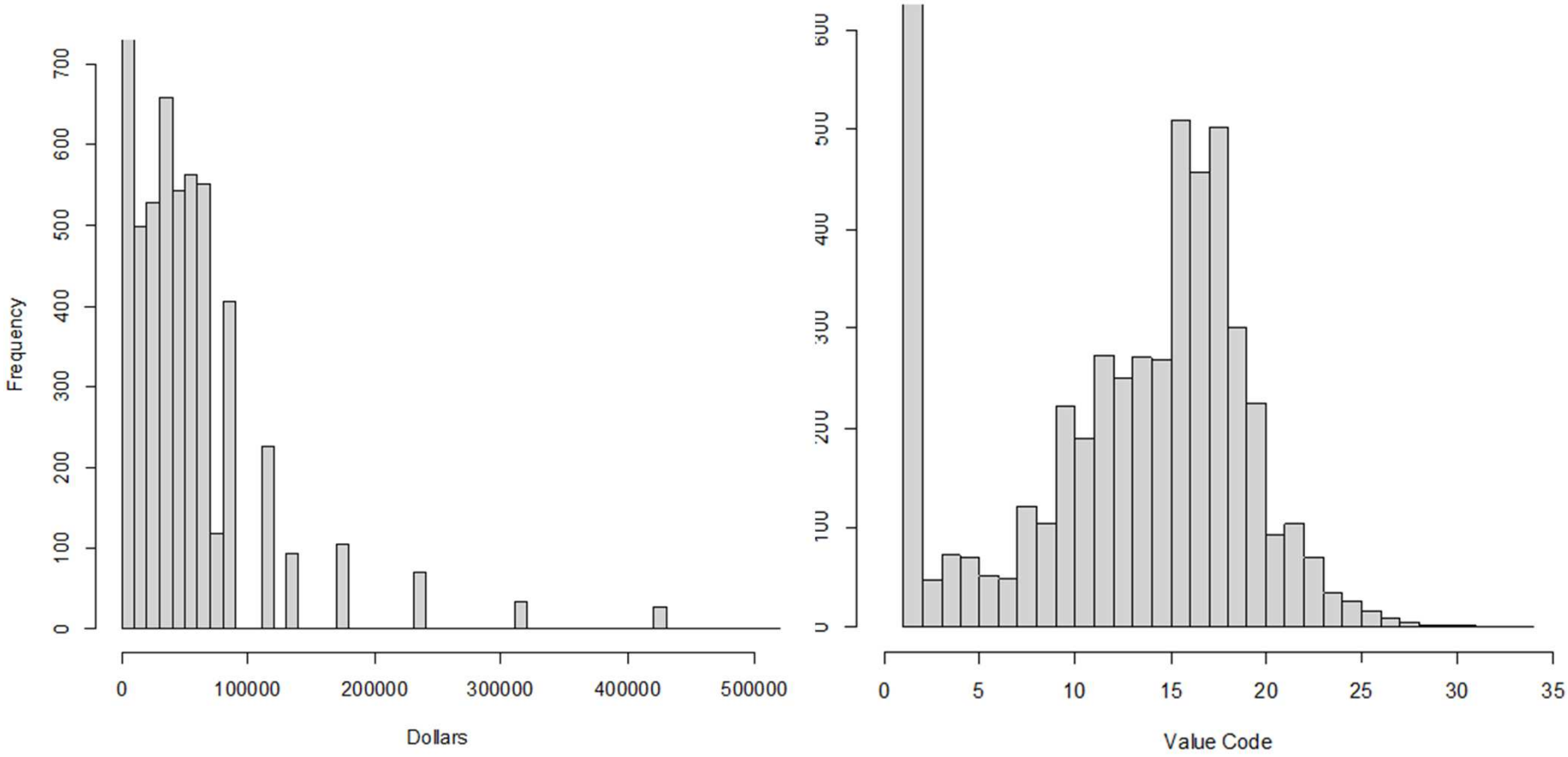
- But, many ways to improve both response rate and inference available from completed surveys
  - Educate enumerators and impress upon them the importance of this HH data
  - Assure respondents of the security of their responses
    - Data safeguarded and never shared with tax authorities
  - Mail surveys may both help and hinder non-response
  - Focus of this talk: new method to improve imputed data for researchers

ARMS Household section

- Asks respondents about
  - Off-farm income, assets, and debt (e.g. wages, dividends, mortgage)
  - Household expenditures (e.g. food, rent, healthcare)
- Responses are value coded
  - Value codes correspond to range of dollar amounts
  - Can be negative for some items (e.g. off-farm business income)

Dollar Range	Value Code
None	1
$1-499	2
$10,000-$14,999	10
$100,000-$124,999	20
$7,500,000-$9,999,999	33
$10,000,000 and over	34

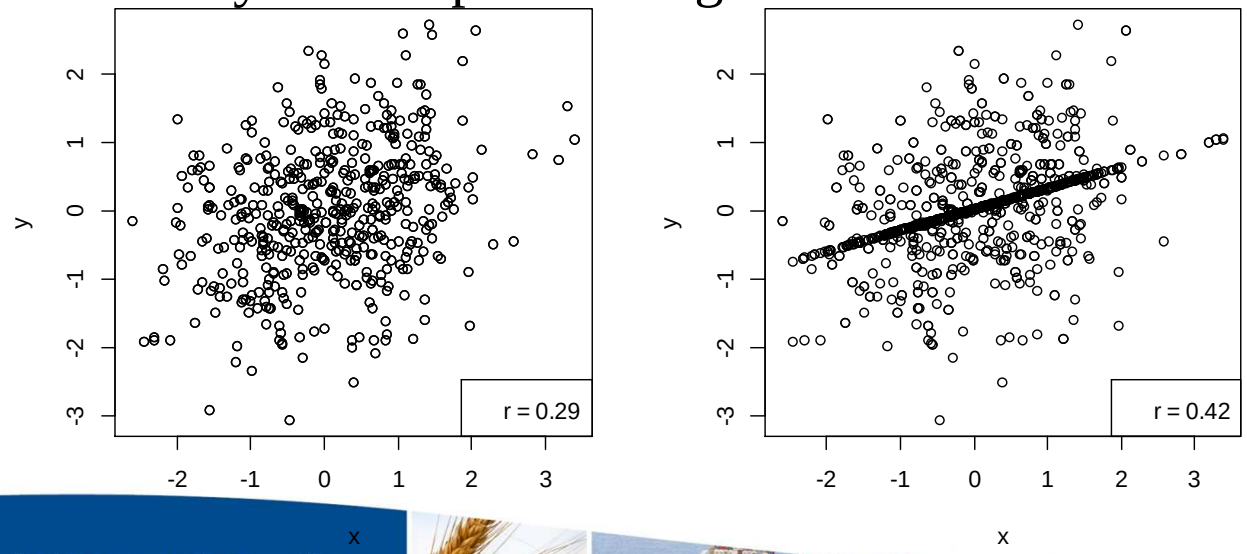
Off-farm wage data is right-skewed in dollar terms and more evenly distributed in value codes



Source: 2013 *Agricultural Resource Management Survey* (ARMS)

Current HH Imputation Methodology

- Imputes using conditional mean from donor group
  - Stratified on key variables: operator age, education, region, occupation
- Issues with conditional mean imputation
  - Acceptable for computing totals and averages
  - Artificially lowers the variance (Little and Rubin, 2002)
  - Sensitive to what the mean is conditioned on
  - Distorts multivariate relationships in the data – below graphs have 50% y data imputed using $\bar{x}$



Iterative Sequential Regression (ISR)

- Earlier project developed ISR specifically to impute for missing data in ARMS in 2011
  - Bayesian approach to imputing missing values using all available data
  - See JASA paper by Robbins, Ghosh and Habiger (2013) for details
- USDA currently uses ISR to impute for farm-level variables
  - E.g. sales, expenses
- Transformations used to achieve approximate normality
  - Skew normal, log normal, empirical CDF
  - Joint multivariate model achieved with Gaussian copula
  - Sequence of regression models built using expert knowledge and economic theory
- Markov Chain Monte Carlo (MCMC) methods used to sample from posterior distributions for parameters and imputations
  - Gibbs sampling
  - Iterative process (I step and P step) until convergence
- Resulting imputations then transformed back to original scale

ISR pros and cons; why we need to transform HH variables

- ISR Advantages
  - Preserves covariance structure between variables
  - Also preserves marginal characteristics
  - Can handle zero values
- ISR Disadvantages
  - ISR can only impute for continuous variables as currently designed
  - ISR could introduce spurious correlation into the data
  - Would ideally have multiple datasets available for analysis
- One issue: HH variables must be transformed to be approximately normal so ISR methodology can be used

→ Our goal: Develop robust transformation that captures information provided by ordering of data

- Value codes in HH section represent ordinal data
  - » Value codes get a increase response rates (as compared with actual dollars)

Maximum Likelihood Estimation (MLE) Method

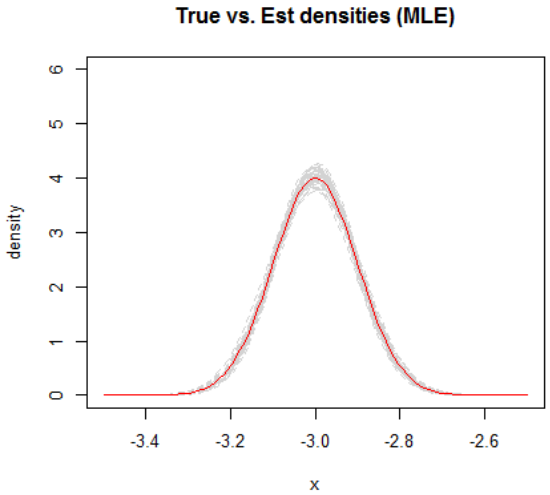
- Suppose Y takes on value codes k=1,2,…m
  - Identify latent variable X using observed Y's and known cut points, $c_{k-1} < X \leq c_k$
  - Let Pr(Y= k) = Pr($c_{k-1} < X \leq c_k$) = F(c_k) – F(c_{k-1})
  - Where F(c_0) = 0 and F(c_m) = 1
- Assume suitable class of parametric family for F
  - Likelihood function of θ is given by
    - $L(\theta) = \prod_{k=1}^{m}[F_\theta(c_k) - F_\theta(c_{k-1})]^{n_k}$
    - Where $n_k = \sum_{i=1}^{n} I(Y_i = k)/n$
- Using suitable numerical optimization method we can obtain MLE estimate, $\hat{\theta}$
- Can use transformation T(y) = $\Phi^{-1}[F_{\hat{\theta}}(c_y + c_{y-1})/2)]$ to obtain normally distributed data

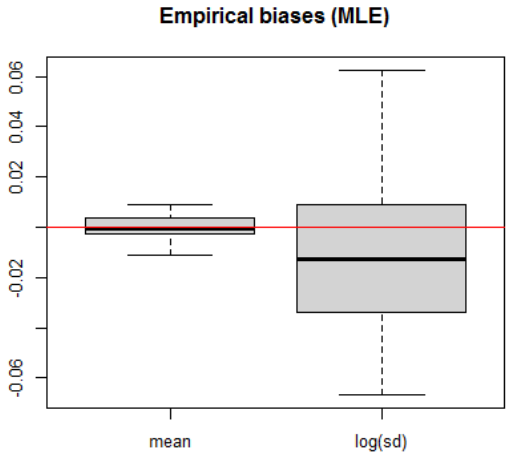
Anderson-Darling Method

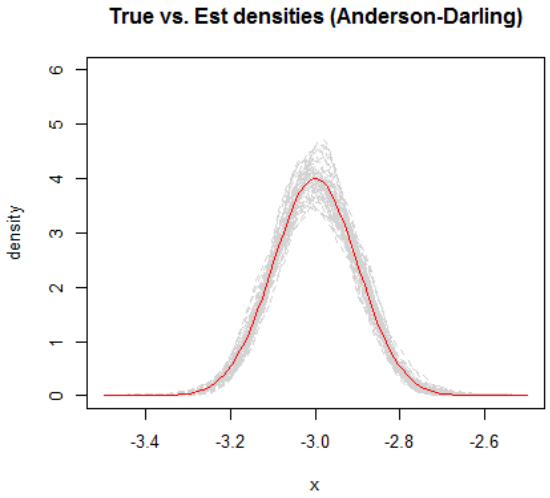
- Non-parametric method for transformation
- Objective function based on Anderson-Darling statistic
  - Choose weights that minimize difference between the empirical CDF and a CDF based on smooth class of distributions (F_θ) at each value code
  - Use log of empirical data because data are highly skewed
  - F_θ is a mixture of weighted polynomials (e.g. Beta's or B-splines)
  - Constrained optimization problem: weights must be non-negative and sum to 1
  - Quadratic programing methods used to choose weights

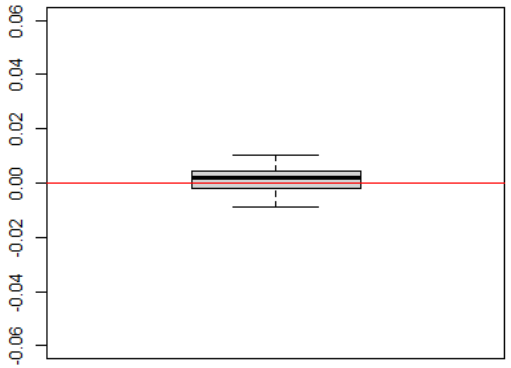
- Advantages
  - Makes no distributional assumptions on underlying data
  - QP methods ensure convergence

MLE and AD method both unbiased with data simulated from normal distribution, MLE has less variability

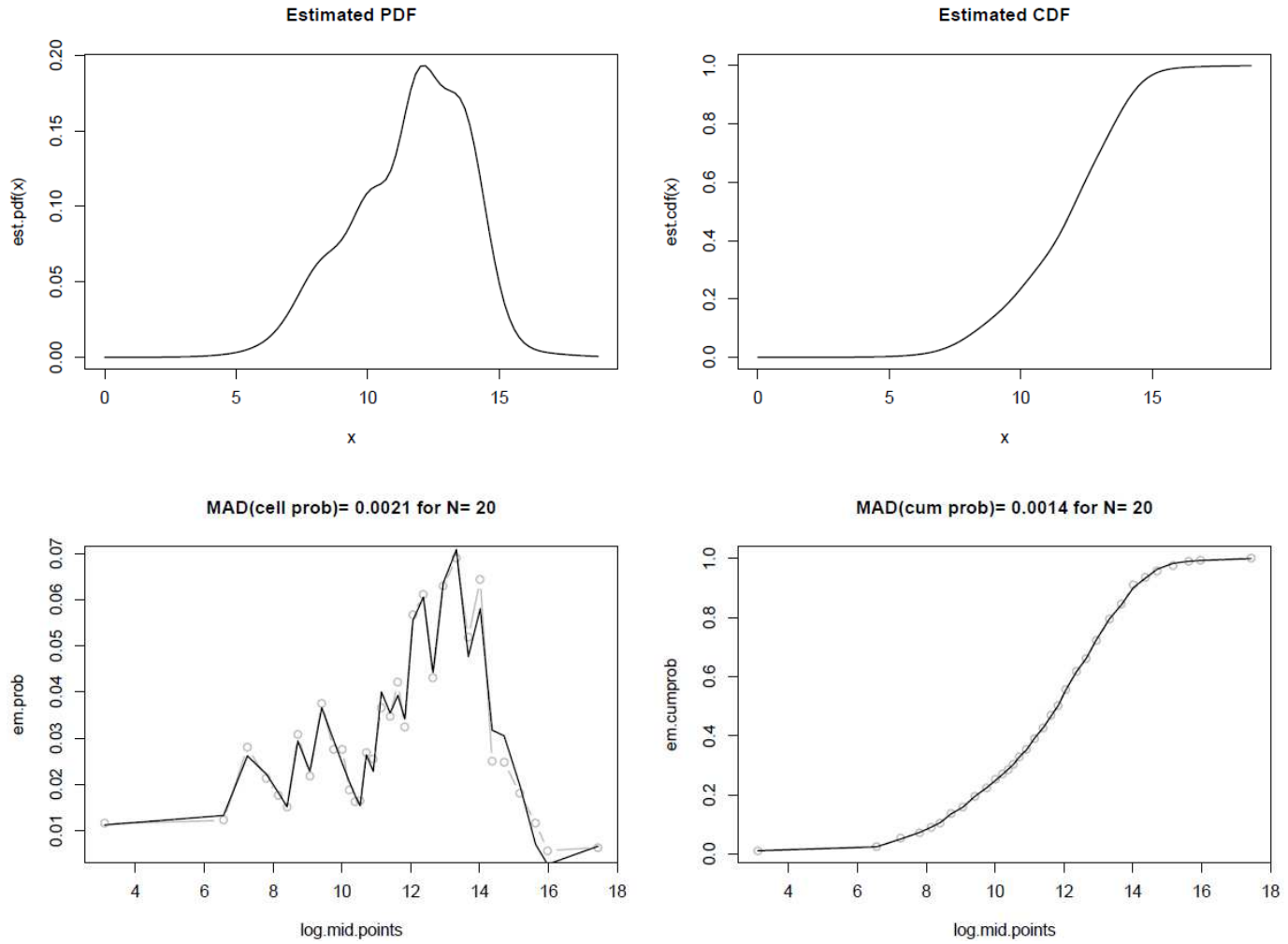








AD method more flexible, better fit to HH data with unknown distribution



Next Steps

- Test both transformation methods with Simulation Study
  - Examine bias and variance
  - Computational efficiency
- Build imputation models for HH variables
  - Use economic theory and expert knowledge (possibly use data mining)
- Run simulation study using adapted version of ISR
  - “Poke holes” in observed ARMS data
  - Impute for created missing data using ISR and conditional mean
  - Compare bias and mean-square error of both imputation methods
- Future research will explore how measures of farm household well-being change under new imputation methodology

Thank you

Questions, comments or suggestions,
please contact:

Daniel Prager
Economic Research Service, USDA
daniel.prager@ers.usda.gov
+1 202 694-5528