Small Area Modeling of County Estimates for Corn and Soybean Yields in the US

Matt Williams

National Agricultural Statistics Service
United States Department of Agriculture
Matt.Williams@nass.usda.gov

Federal Committee on Statistical Methodology
Research Conference
Nov. 4, 2013

Background

County Agricultural Production Survey (CAPS) Program

- In 2009-2010 the County Estimates Pilot Study was initiated in 5 States.
- New approach for selecting and processing county estimates, consistent with procedures used for the Agricultural Survey programs
- County estimates data are merged with year-end survey data.
- From 2011 onward, the probability design was expanded to cover all States.

Motivation

- CAPS results are used by field offices and HQ to establish county estimates for acreage and yield (production per acre).
  - Other administrative data are examined (Farm Service Agency, remote sensing, etc.)
  - Centralized Database for examining inputs and setting estimates
- Our GOAL is to add value to this process.
  - Efficiency: increased automation?
  - Added functionality: SE's and CV's for estimates?
- We examine *Small Area Models* as a possible solution.

Methods

- The methods described in this presentation are considered *model-based* (Rao, 2003).
  - Nested error (mixed model) regression
  - Combine survey and covariate information.
  - Induce a correlation between small areas, pooling information across the areas.
  - Referred to as "borrowing strength", a phrase ubiquitous in the literature.
- Methods are related to *shrinkage estimates*.
- Models can focus on the County (area) or Record (unit).
- Inference can be conducted with "Frequentist" and "Bayesian" approaches.

Shrinkage Estimates

- Mean Square Error ($E[Y] = \theta$)

$$\begin{aligned} E\left[(\theta - \hat{\theta})^2\right] &= \left(\theta - E[\hat{\theta}]\right)^2 + E\left[\left(E[\hat{\theta}] - \hat{\theta}\right)^2\right] \\ MSE &= Bias^2 + Variance \end{aligned}$$

- Stein's "Paradox" (Efron and Morris, 1977)
  - For $d > 2$, total MSE $\sum_{i=1}^{d} E\left[(\theta_i - \hat{\theta}_i)^2\right]$ is *not* minimzed by the Least Squares estimator $\hat{\theta}_{LS}$.
  - A class of shrinkage estimates $\hat{\theta}_{Sh}$ can be created which dominate $\hat{\theta}_{LS}$

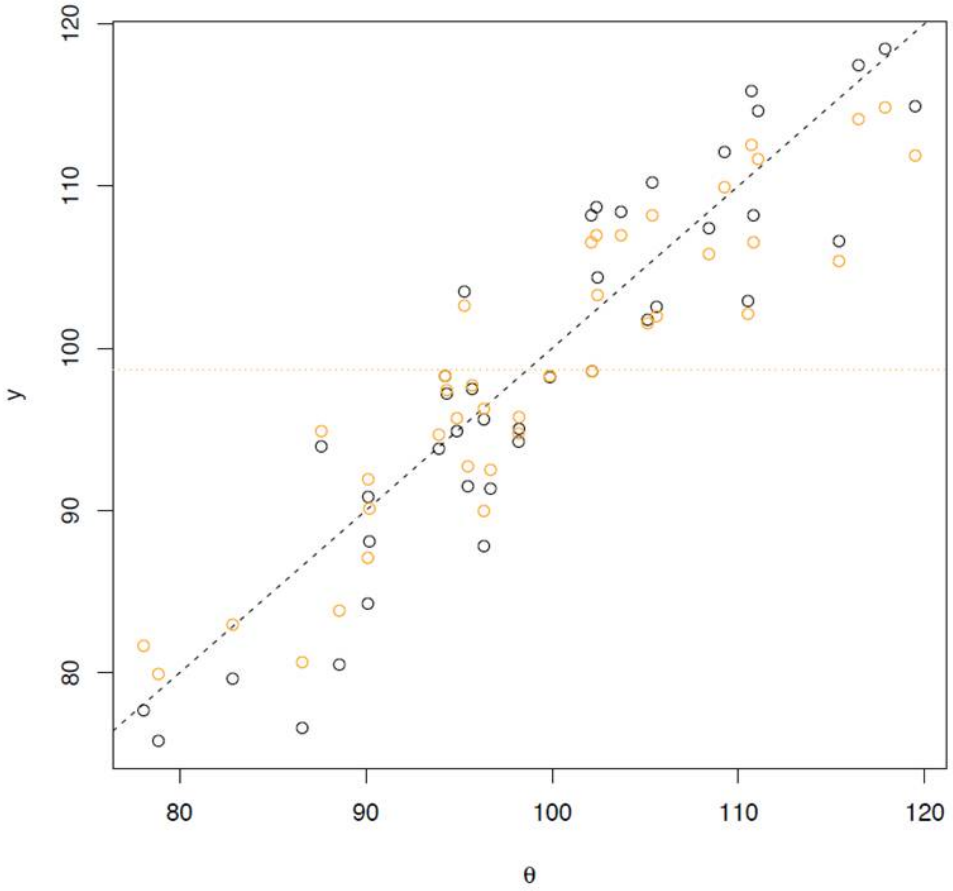
Shrinkage Estimates

- Shrinkage Estimate
  - $\hat{\theta}_{Sh} = \bar{y} + c(\hat{\theta}_{LS} - \bar{y})$
  - c can be chosen several ways.
- Composite Estimate
  - $\hat{\theta}_C = \phi\hat{\theta}_A + (1 - \phi)\hat{\theta}_B$
  - ϕ can be chosen several ways.
- Small Area Models
  - Combine two estimators: Survey Indications and Regression Model
  - Use variance components from the Survey and the Model to estimate ϕ.

Shrinkage Estimates: Example

$$\begin{aligned} \theta &\sim N(100, 10) \\ Y &\sim N(\theta, 5) \end{aligned}$$

$\hat{\theta}$	Bias	RMSE	MAD
Raw Y	33.3	29.6	155
Model	32.4	26.0	134



County vs. Record Level Models

- County (Area) Level
  - Covariates are available at the county level
  - Survey estimates and standard errors (SE's) used directly
  - Composite parameter ϕ determined by survey SE's and model estimate of between-county variance
- Record (Unit) Level
  - Covariates may be used for the record level
  - Survey estimates are modeled via record responses and weights
  - Composite parameter ϕ determined by model estimate of between-county variance and residual variance of the model (with weights)

Parameter Estimation

- Empirical Bayes
  - Frequentist approach
  - Maximum Likelihood and Restricted ML Estimation for Parameters
  - Inference uses "plug-in" estimates of parameters.
- Hierarchical Bayes
  - "Full" Bayesian approach
  - Constructs a posterior using a likelihood and a prior distribution for parameters.
  - Inference uses distributions (or samples from them) of the parameters.
  - Considerations for computation and prior selection

Results: 2010 Pilot States

- Models (14 results)
  - Area Level (EB & HB)
  - Unit Level (EB & HB)
  - Unit Level with Sampling Weights (10)
    - EB (2), HB (3)
    - With & with/out modifications (x2)

Results: 2010 Pilot States

- Models (14 results)
  - Area Level (EB & HB)
  - Unit Level (EB & HB)
  - Unit Level with Sampling Weights (10)
    - EB (2), HB (3)
    - With & with/out modifications (x2)
- Each model used the same area-level covariates
  - National Commodity Crop Productivity Index (NCCPI)
  - Normalized Difference Vegetation Index (NDVI) based

Results: 2010 Pilot States

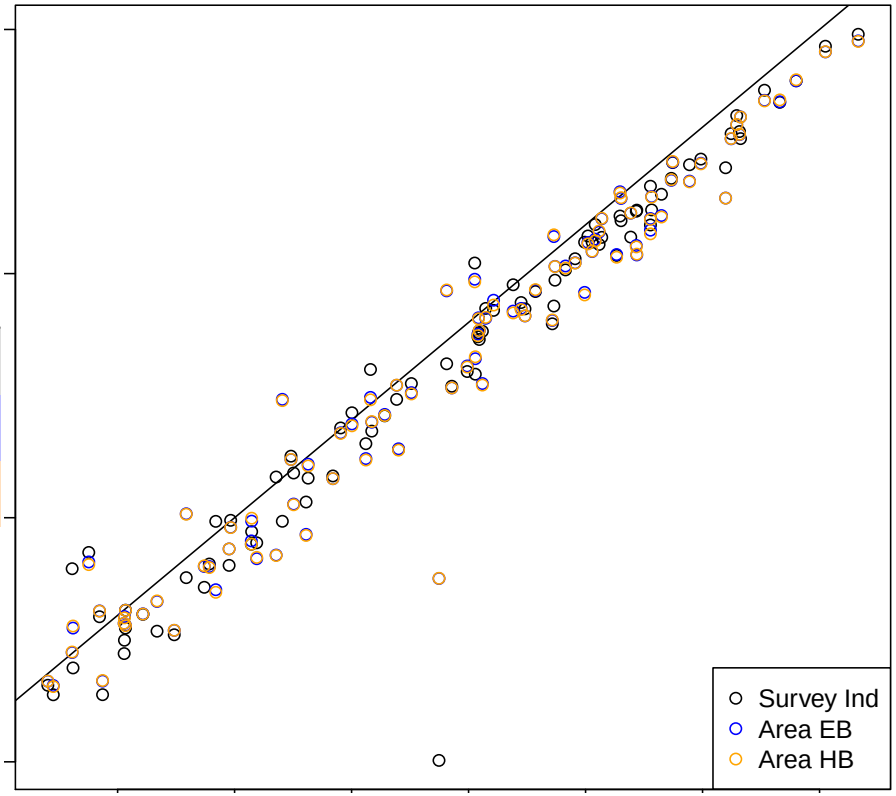
- Models (14 results)
  - Area Level (EB & HB)
  - Unit Level (EB & HB)
  - Unit Level with Sampling Weights (10)
    - EB (2), HB (3)
    - With & with/out modifications (x2)
- Each model used the same area-level covariates
  - National Commodity Crop Productivity Index (NCCPI)
  - Normalized Difference Vegetation Index (NDVI) based
- Survey Indications
  - generally best predictor of final published values
- Area Level models
  - consistently second best
- EB and HB models
  - generally consistent
  - tighter agreement for area-level models

Results: 2011-12 Selected States

- Similar to 2010 Results
- A few states showed modest gains using Models over Survey Indications
- Some states "broke even"
- Many state Models had no improvement over the Survey Indications

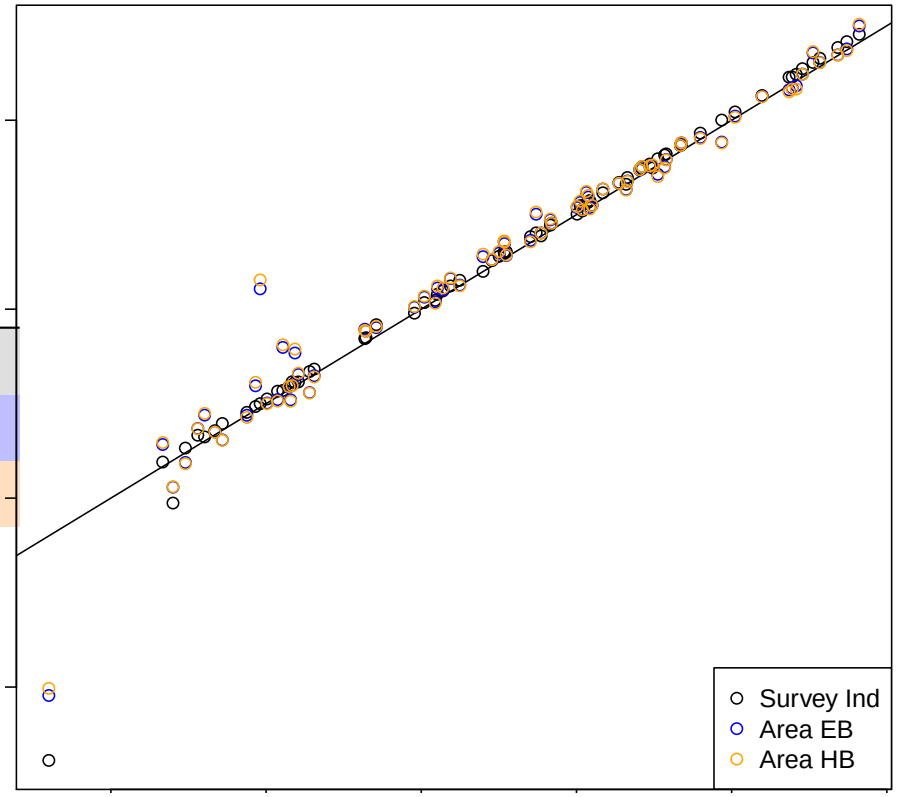
Example: Modest Gain

$\hat{\theta}$	Bias	RMSE	MAD
Direct	1.00	1.00	1.00
EB	0.94	0.81	0.99
HB	0.96	0.82	1.00



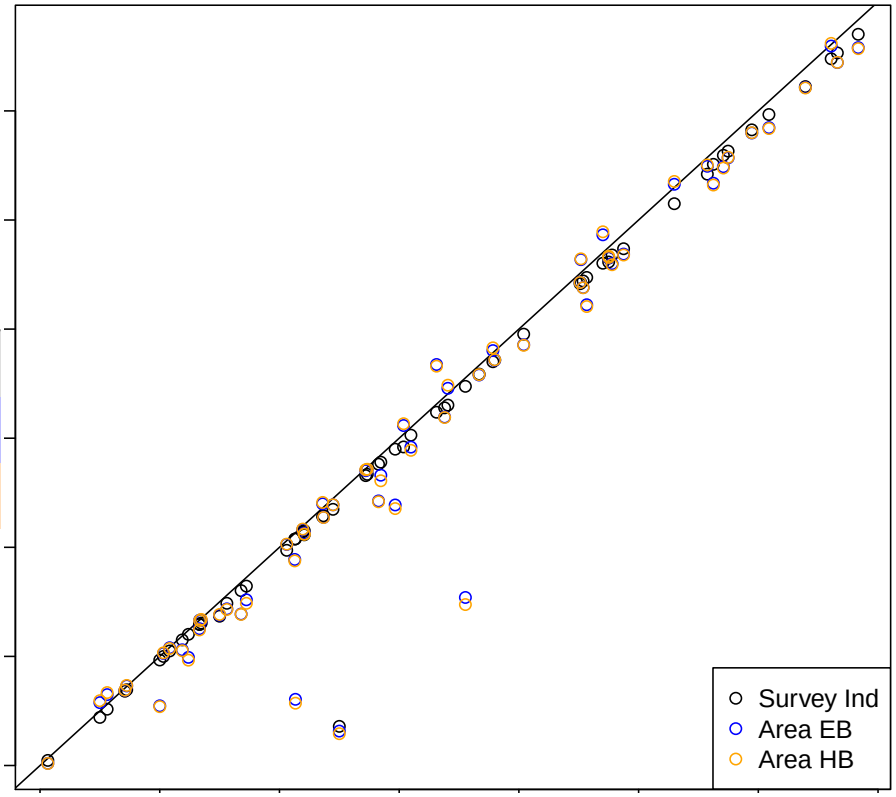
Example: Break Even

$\hat{\theta}$	Bias	RMSE	MAD
Direct	-1.00	1.00	1.00
EB	-3.25	0.93	1.47
HB	-3.38	0.96	1.56



Example: Adding Noise

$\hat{\theta}$	Bias	RMSE	MAD
Direct	1.00	1.00	1.00
EB	1.65	1.61	1.94
HB	1.68	1.65	2.02



Comments

- Final published values considered the true mean.
  - Direction of causality is reversed. Published values did not generate the survey responses.
  - Benchmarks to state values and other considerations (weather, growing conditions, etc).
  - A few large deviations (changes) and some systematic (benchmarking)
- Shrinkage estimates
  - Small amounts of change spread out over all counties
  - No decision to leave some unchanged.
- External Validation
  - How is the current county estimates program being evaluated?
  - Best available given timeliness requirement.
  - What is available later to compare? (Census 2012)

Related Research

- Benchmarking
  - County yield is a ratio. Benchmarking to state yield (ratio) is therefore nonlinear (Williams and Berg, 2013).
  - Multi-stage benchmarks and penalties (districts or variances) are also possible.
- Nonlinear relationships
  - Strong relationships between covariates and responses may be nonlinear.
  - General additive models (GAM's) have been applied to unit level models (Opsomer et al., 2008). Similar methods can be used for area level models.

References

Efron, B. and Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, 236:119–127.

Opsomer, J., Claeskens, G., Ranalli, M., Kauermann, G., and Breidt, F. (2008). Non-parametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society: Series B*, 70:265–286.

Rao, J. (2003). *Small Area Estimation*. John Wiley & Sons, New Jersey.

Williams, M. and Berg, E. (2013). Incorporating user input into optimal constraining procedures for survey estimates. *Journal of Official Statistics*, 29:375–396.