Westat

Improving Lives Through Research®

Model Selection and Its Important Roles in Surveys – Jiming Jiang

Discussion

Andreea L. Erciulescu, Westat

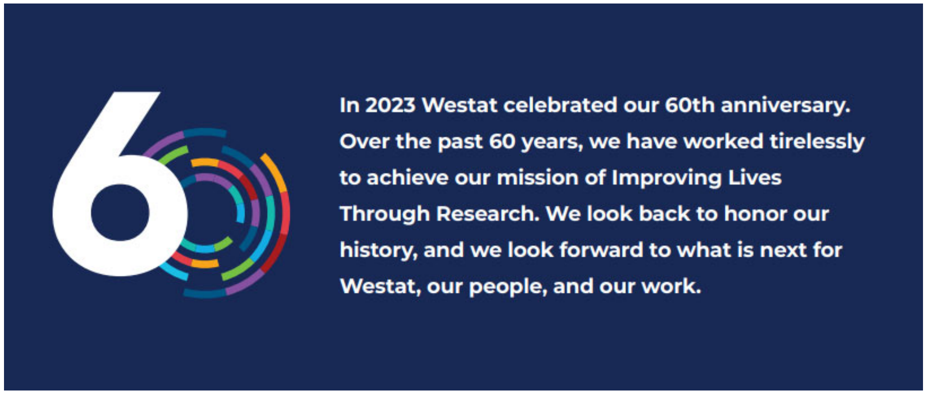
The 2023 Morris Hansen Lecture

The views presented are those of the author(s) and do not represent the views of any Government Agency/Department or Westat.

Morris Hansen and Westat

‘The affiliation of Morris Hansen with Westat on November 1, 1968, upon his retirement from the Bureau of the Census, requires special mention… when he joined the Company it was an announcement to the world that we were to be taken seriously by the contract research community.’

Edward C. Bryant, July 1981



In 2023 Westat celebrated our 60th anniversary. Over the past 60 years, we have worked tirelessly to achieve our mission of Improving Lives Through Research. We look back to honor our history, and we look forward to what is next for Westat, our people, and our work.

Westat. (personal communication, 2023)

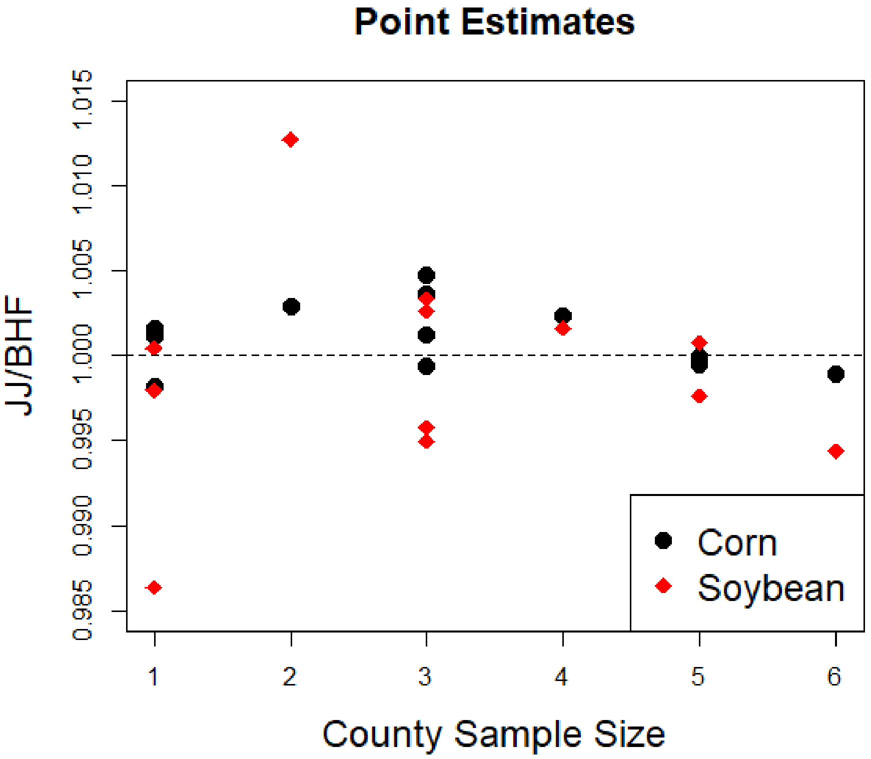
Outline

- General lecture remarks
- A small area estimation (SAE) application of fence methods to Census Bureau data
- A framework for variable selection used in recent Westat SAE projects

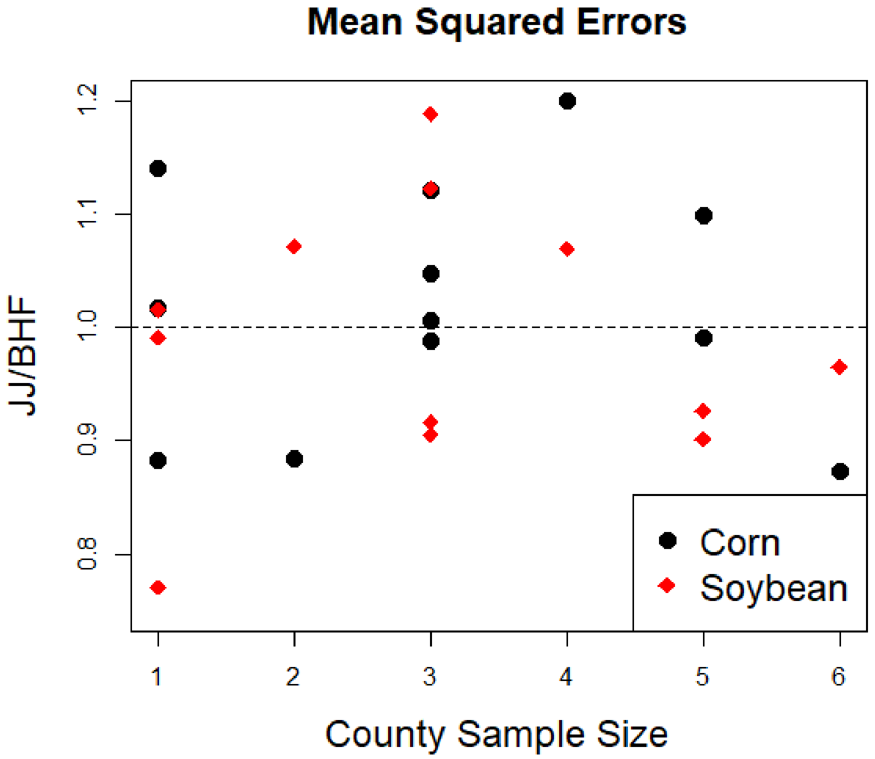
Jiming Jiang's Lecture

- Information in the 21st Century
  - Borrowing strength
- Statistical modeling in surveys
  - SAE
- Model/variable selection
  - Traditional: information criteria, shrinkage selection/estimation
  - Proposed: fence methods
  - Example: unit-level modeling (Battese-Harter-Fuller)

Jiming Jiang’s Lecture Example: Additional Results



Point Estimates



Mean Squared Errors

Gaining more parsimonious models without noticeable losses…

Area-level Modeling Application: Data

- U.S. Census Bureau’s Small Area Income and Poverty Estimates (SAIPE) Program
  - School district, county, state, age group
  - Funds allocation by the U.S. Department of Education
  - Current Population Survey (CPS) $\xrightarrow{time}$ American Community Survey (ACS)
  - Internal Revenue Services (IRS), Supplemental Nutritional Assistance Program (SNAP), Census
- Publicly-available research datasets: https://www.census.gov/srd/csrmreports/byyear.html
  - *Bell and Franco (2017); Erciulescu, Franco, Lahiri (2021)*

Area-level Modeling Application: Modeling

- State-level Fay-Herriot model

$$y_i \sim N(\theta_i\,, D_i)$$
$$\theta_i \sim N(x_i'\beta\,, A)$$

  - Sample data $(\, y_i\,,\, D_i\,)$: CPS93 survey direct estimates of poverty rates for 5–17-year-old children, with associated smoothed variance estimates
  - Covariates x_i': IRSPR93 pseudo-poverty rates, IRSNF93 non-filer rates, SNAP93 participation proportions, CEN89RSD residuals from a Fay-Herriot model fitted to census estimates of children in poverty

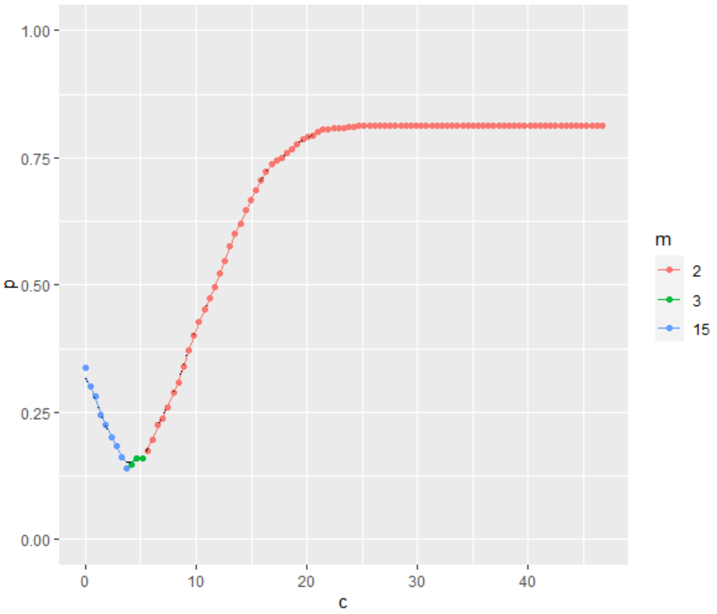
- Estimation: restricted maximum likelihood
- Covariates selection: Akaike Information Criteria (AIC) and Adaptive Fence

Area-level Modeling Application: Results

- AIC results: M1234

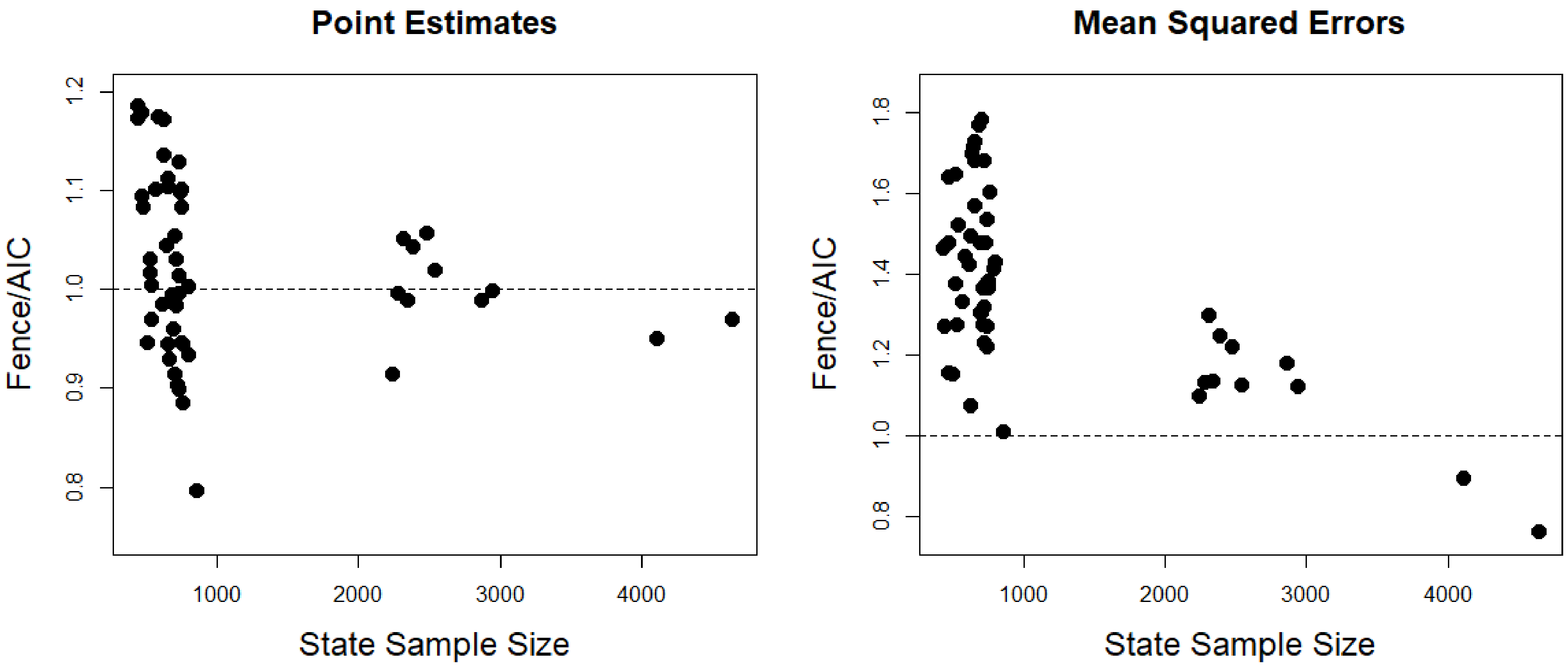
Model	M1	M2	M3	M4	M12	M13	M14	M23	M24	M34	M123	M124	M134	M234	M1234
AIC	333.54	354.88	320.46	355.53	328.20	321.00	329.21	320.90	353.32	317.49	319.75	322.06	316.92	317.76	**314.30**

- Fence results: M3
  - c = 26.36



Is the parsimony criterion too strong?

Area-level Modeling Application: Results cont.



Point Estimates

Mean Squared Errors

Should minimization of a criterion function be considered?

A Framework for SAE Variable Selection

- Build a pool of variables related to the quantity of interest
  - Check definitions, reference years, completeness, possible error sources
  - Rely on subject-matter expertise
- Conduct an initial selection
  - Check for redundancy, outliers, transformations, associations
- Conduct a second selection
  - Use information criteria, shrinkage selection, decision trees, cross-validation

Perhaps fence methods could be included in the second selection…

Framework Application 1: Multi-fold Models

- Prevalence to having a personal doctor for the U.S. Centers for Disease Control and Prevention
- The Behavioral Risk Factor Surveillance System; 2018
- Area-level univariate linear (on arsine-square-root scale) **three-fold** models
- 3,142 small areas defined as counties; 3,114 with sample data

Erciulescu et al. (2022)

Framework Application 1: Multi-fold Models cont.

- Dozens county-level potential covariates
- Least absolute shrinkage selection operator (LASSO) and cross-validation
- 9 selected variables:
  - Education (1), race/ethnicity (3), home ownership (1), changes in location (2), health insurance (1), tax (1)

Framework Application 2: Multivariate Models and Prediction Needs

- Proficiency measures of adult competency for the U.S. National Center for Education Statistics
- The Program for the International Assessment of Adult Competencies, U.S.; 2013-2017
- Area-level univariate and **bivariate** linear three-fold models
- 3,142 small areas defined as counties; **185** with sample data

Ren et al. (2022)

Framework Application 2: Multivariate Models and Prediction Needs cont.

- 70 county-level potential covariates, 24 state-level potential covariates
- LASSO and cross-validation
- 7 selected variables
    - Education (2), poverty (1), race/ethnicity (2), health insurance (1), occupation (1)

Framework Application 3: High-dimensional Model Matrix

- Employee compensation components for the U.S. Bureau of Labor Statistics
- The National Compensation Survey; 2017
- Area-level bivariate linear (on log scale) model
- 668,938 small areas defined as crosstabulations of census divisions, six-digits Standard Occupational Classification (SOC) system codes, work levels, and job characteristics (time/incentive, full-time/part-time, union/nonunion); 16,107 with sample data

Erciulescu and Opsomer (2022)

Framework Application 3: High-dimensional Model Matrix cont.

- **21,454** columns in the model matrix defined as crosstabulations of census divisions, six-digits SOC system codes, work levels, and job characteristics (time/incentive, full-time/part-time, union/nonunion), and their two-way interactions
- Second selection only
  - LASSO, decision trees, stepwise selection using AIC
- 17 selected variables
  - Work level (7), SOC code (2), full-time/part-time x time/incentive (1), full-time/part-time x work level (1), SOC code x full-time/part-time (1), SOC code x work level (4), union/nonunion x work level (1)

Summary

- Model selection versus variable selection
- Parsimony
- Presence of minimization criteria
- Goodness of fit versus predictive power
- Complex survey design
- Complex modeling

References

- Bryant E.C. (1981) “20 Years and Counting. A Personal History of Westat’s Early Years.” Unpublished manuscript.

- Battese G.E., Harter, R.M., and Fuller, W.A. (1988) “An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data.’’ Journal of the American Statistical Association, 83(401). https://doi.org/10.1080/01621459.1988.10478561

- Bell W. R. and Franco C. (2017) “Small Area Estimation – State Poverty Rate Model Research Data Files.’’ Available at https://www.census.gov/srd/csrmreports/byyear.html [accessed October 19, 2023]

- Erciulescu A.L., Franco C., Lahiri P. (2021) “The Use of Administrative Data in Small Area Estimation.” Administrative Records for Survey Methodology. Ed. A. Y. Chun, Ed. M. Larsen, Co-Ed. J. Reiter, Co-Ed. G. Durrant. Wiley.

- Erciulescu A.L., Li J., Krenzke T., Town M. (2022) “Hierarchical Bayes small area estimation for county-level prevalence to having a personal doctor.’’ Statistical Methods & Applications. https://doi.org/10.1007/s10260-022-00678-7

- Erciulescu A.L. and Opsomer J.D. (2022) “A model-based approach to predict employee compensation components.’’ Journal of the Royal Statistical Society, Series C, 71, 5, 1503-1520. https://doi.org/10.1111/rssc.12587

- Fay R.E. and Herriot R.A. (1979) “Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data.” Journal of the American Statistical Association, 74(366a), 269-277. https://doi.org/10.1080/01621459.1979.10482505

- Jiang J. (2023) “Model selection and its important roles in surveys.” 31st Annual Morris Hansen Lecture. Washington DC, November 2023

- Ren W., Li J., Erciulescu A.L., Krenzke T., Mohadjer L. (2022) “A variable selection method for small area estimation modeling of the proficiency of adult competency.’’ Stats 5, no. 3: 689-713. https://doi.org/10.3390/stats5030041

Thank you!

AndreeaErciulescu@westat.com