Small Area Estimation for the Annual Integrated Economic Survey

Stephen J. Kaputa

Composite Methods Staff

Economic Statistical Methods Division

Small Area Estimation Team

- ESMD
  - Stephen Kaputa
  - Yeng Xiong
  - Danny Chang
- EMD
  - Julian Hunt
- CSRM
  - Jerry Maples
  - Serge Aleshin-Guendel
  - Ryan Janicki
  - Gauri Datta

Annual Integrated Economic Survey (AIES)

- Economy-wide survey that replaces multiple independently designed annual surveys
- Designed to produce national level detailed industry estimates and limited industry-by-state estimates
- But what about more detailed estimates?

Small Area Estimation Motivation

- AIES is designed to produce select national and subnational estimates
  - 6-digit NAICS national estimates
  - 3-digit NAICS for select states
  - Annual Payroll, 1st quarter payroll, employment, receipts
- Direct domain estimates
  - Uses only domain-specific information
  - Typically design-based estimates using survey weights given a sample design
- We consider a domain "large" if the direct estimate is of adequate precision
  - For NAICS3 *x* state estimates, target coefficient of variation (CV) used in sample design is 15%

United States® Census Bureau

Motivation

- We consider a domain "small" if the direct estimate is NOT of adequate precision
  - Small state estimates
  - State estimates for detailed industries

- Constraints (e.g., budget and burden) prevent drawing large samples from small domains

- Indirect domain estimates
  - Borrow strength from related areas or time periods to increase "effective" sample size
  - Model-based estimates

United States Census Bureau

Fay – Herriot Model

- An area level model that blends a direct survey estimate with an indirect model-based estimate
  - "Optimal" estimator that reduces the mean squared error (MSE)

- A linking model defines the relationship between the domain estimate and auxiliary information or covariates
  - We need auxiliary information that is a strong predictor of the survey outcome

Notation

- Sampling model (direct survey estimate)

$$\hat{Y}_d^{Dir} = Y_d + e_d^{Dir}$$

$$e_d^{Dir} \sim N\left(0, \sigma_{Dir,d}^2\right)$$

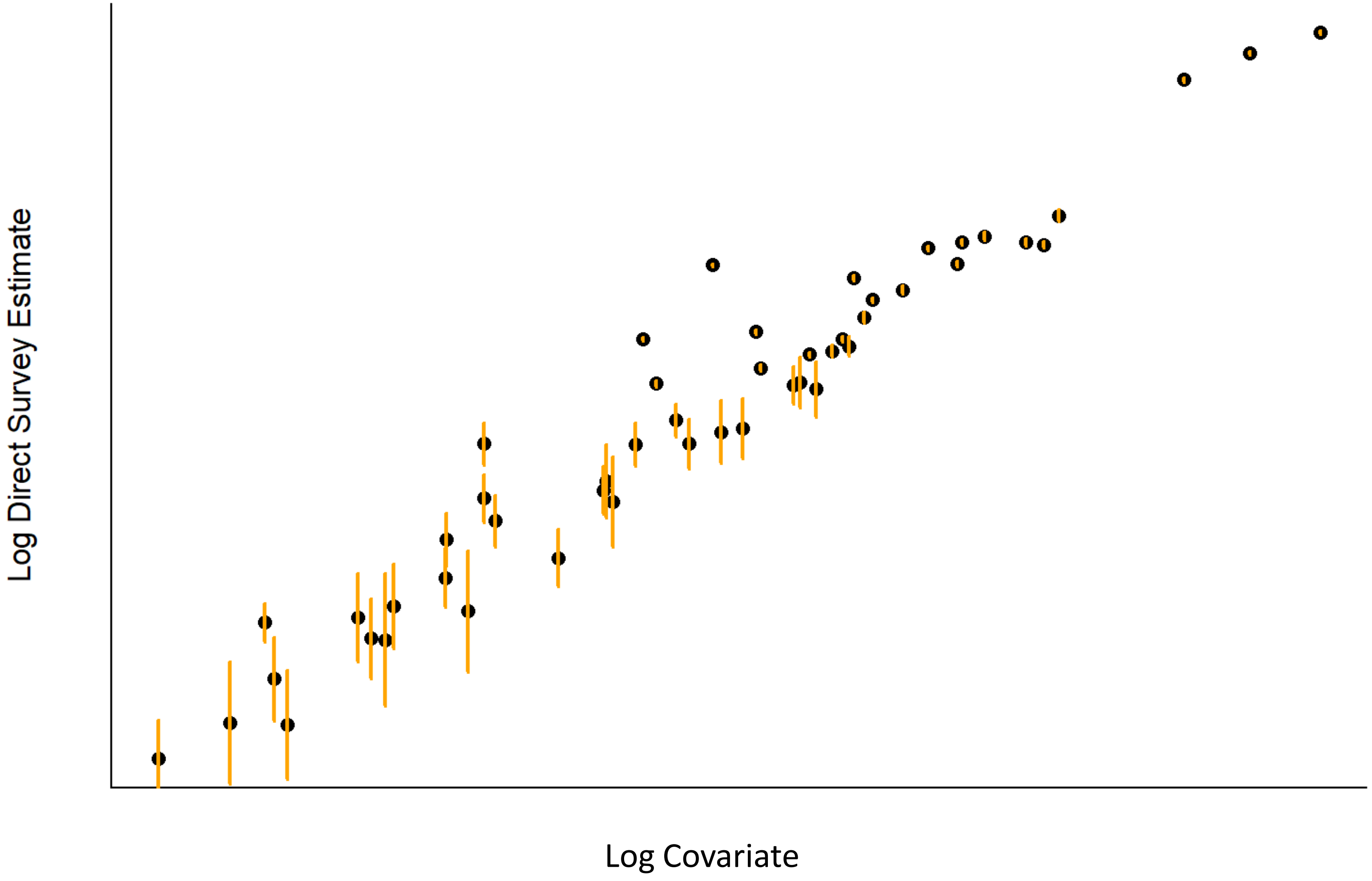
- Linking model (indirect survey estimate)

$$Y_d = X_d' \beta + e_d^{Mod}$$

$$e_d^{Mod} \sim N\left(0, \sigma_{Mod}^2\right)$$

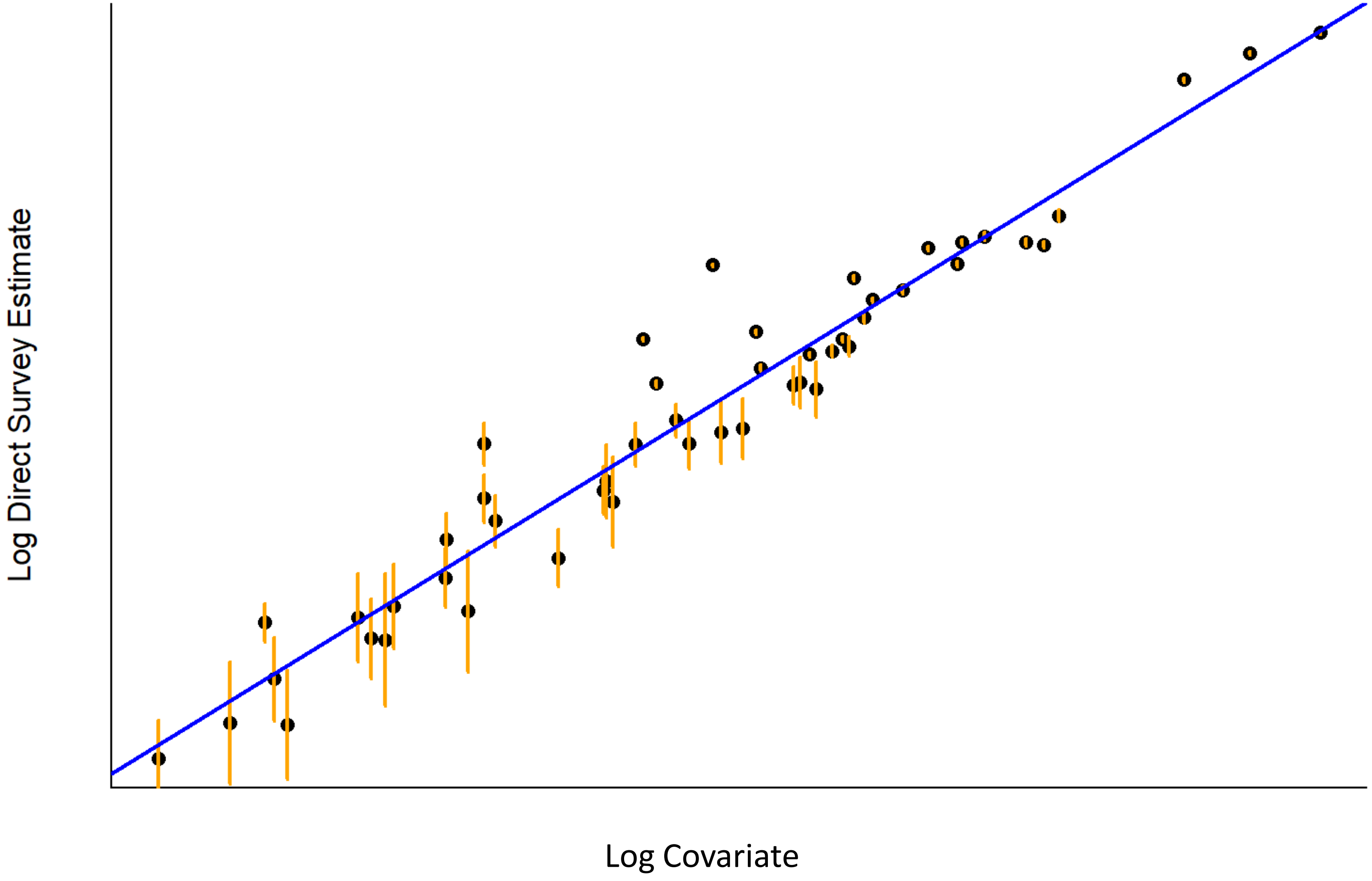
Note: Models are fit independent at the industry estimation level (NAICS 3 or 4), industry subscript has been omitted

Survey Data



Source: U.S. Census Bureau, [Project No. P-7529180 / Approval CBDRB-FY24-ESMD001-012]

Indirect survey estimate (Linking model)



Source: U.S. Census Bureau, [Project No. P-7529180 / Approval CBDRB-FY24-ESMD001-012]

Notation

- Linear mixed model (Fay-Herriot)

$$\hat{Y}_d^{Dir} = X'_d\beta + e_d^{Mod} + e_d^{Dir}$$

- Empirical Best Linear Unbiased Prediction (EBLUP) Estimator

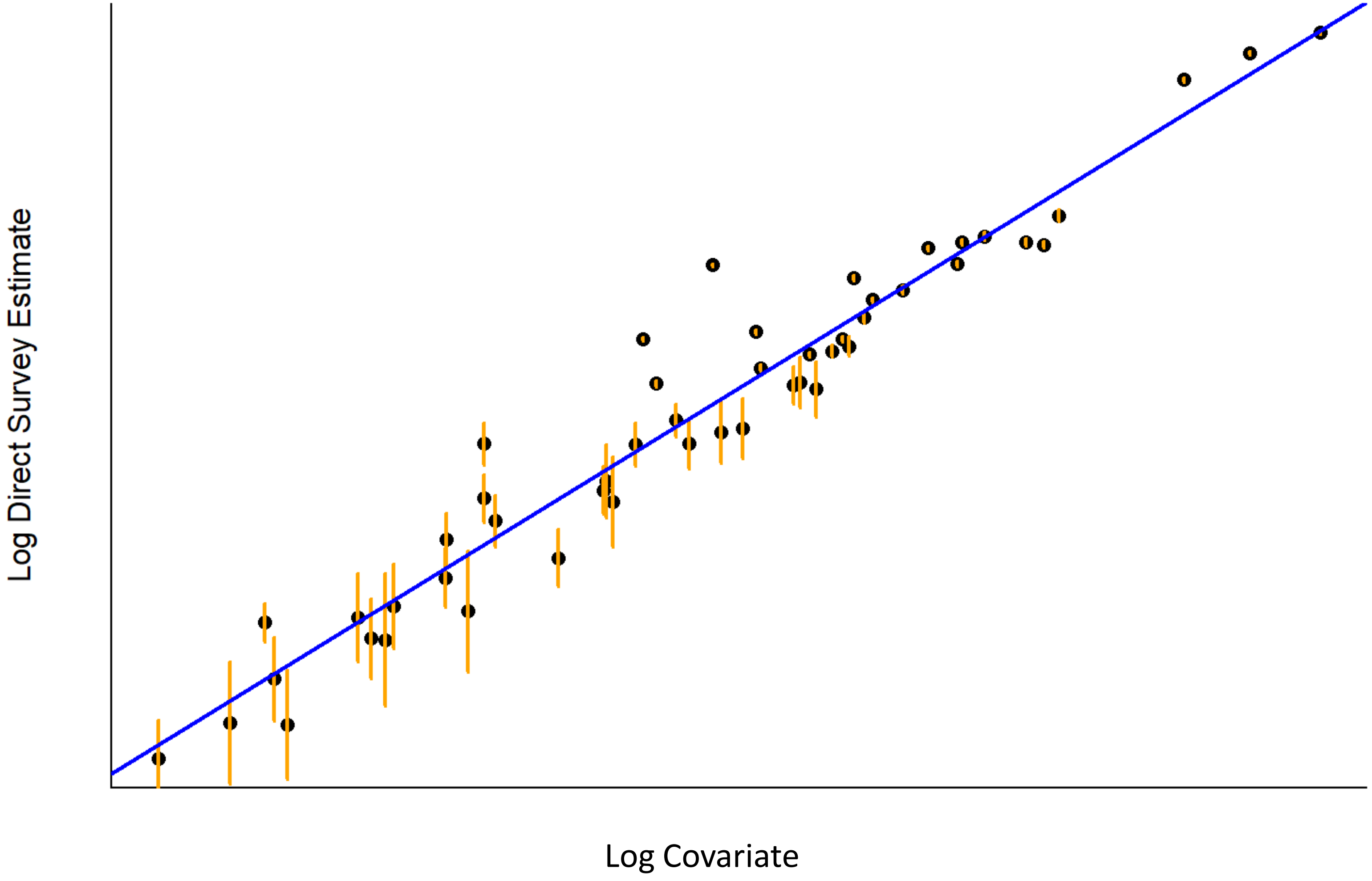
$$\hat{Y}_d^{FH} = \hat{\gamma}_d \hat{Y}_d^{Dir} + (1 - \hat{\gamma}_d) X_d \hat{\beta}$$

- Where $\hat{\gamma}_d$ is the shrinkage factor

$$\hat{\gamma}_d = \frac{\hat{\sigma}_{Mod}^2}{\hat{\sigma}_{Mod}^2 + \sigma_{Dir,d}^2}$$

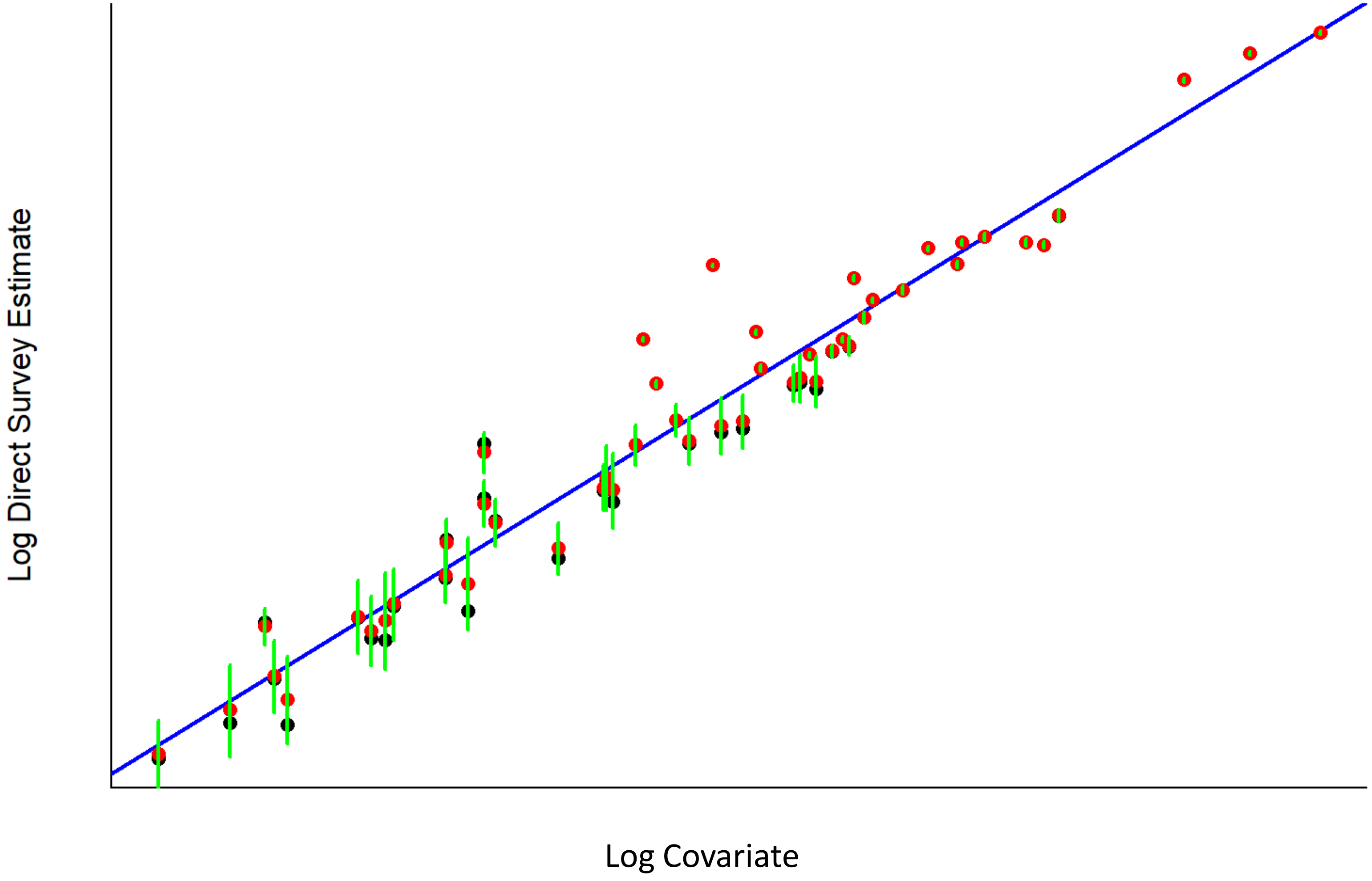
United States® Census Bureau

Indirect survey estimate (Linking model)



Source: U.S. Census Bureau, [Project No. P-7529180 / Approval CBDRB-FY24-ESMD001-012]

Fay-Herriot Estimates



Fay-Herriot model using Hierarchical Bayes

- Transformations are easier
  - Back transforming the log posterior distribution is straightforward
- Additivity requirements or constraints
  - Can include constraints in the model
  - Or posterior distributions can be adjusted to known totals
- Can include informative prior information
- Research is being done using the open-source probabilistic programming language “Stan” in R

United States® Census Bureau

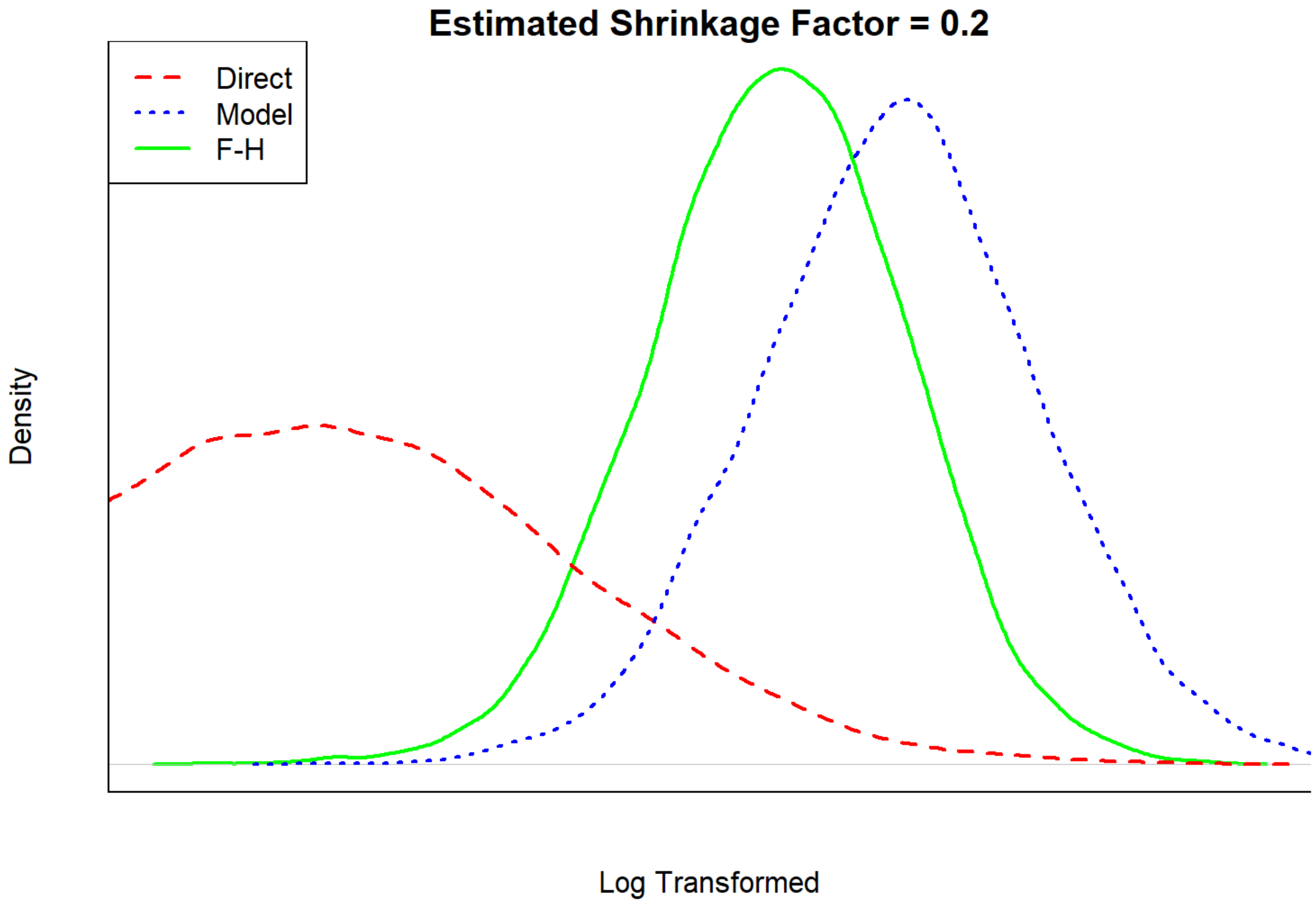
Fay-Herriot model using Hierarchical Bayes

$$\hat{Y}_d^{Dir} \mid Y_d, \sigma^2_{Dir,d} \sim normal(Y_d, \sigma^2_{Dir,d})$$

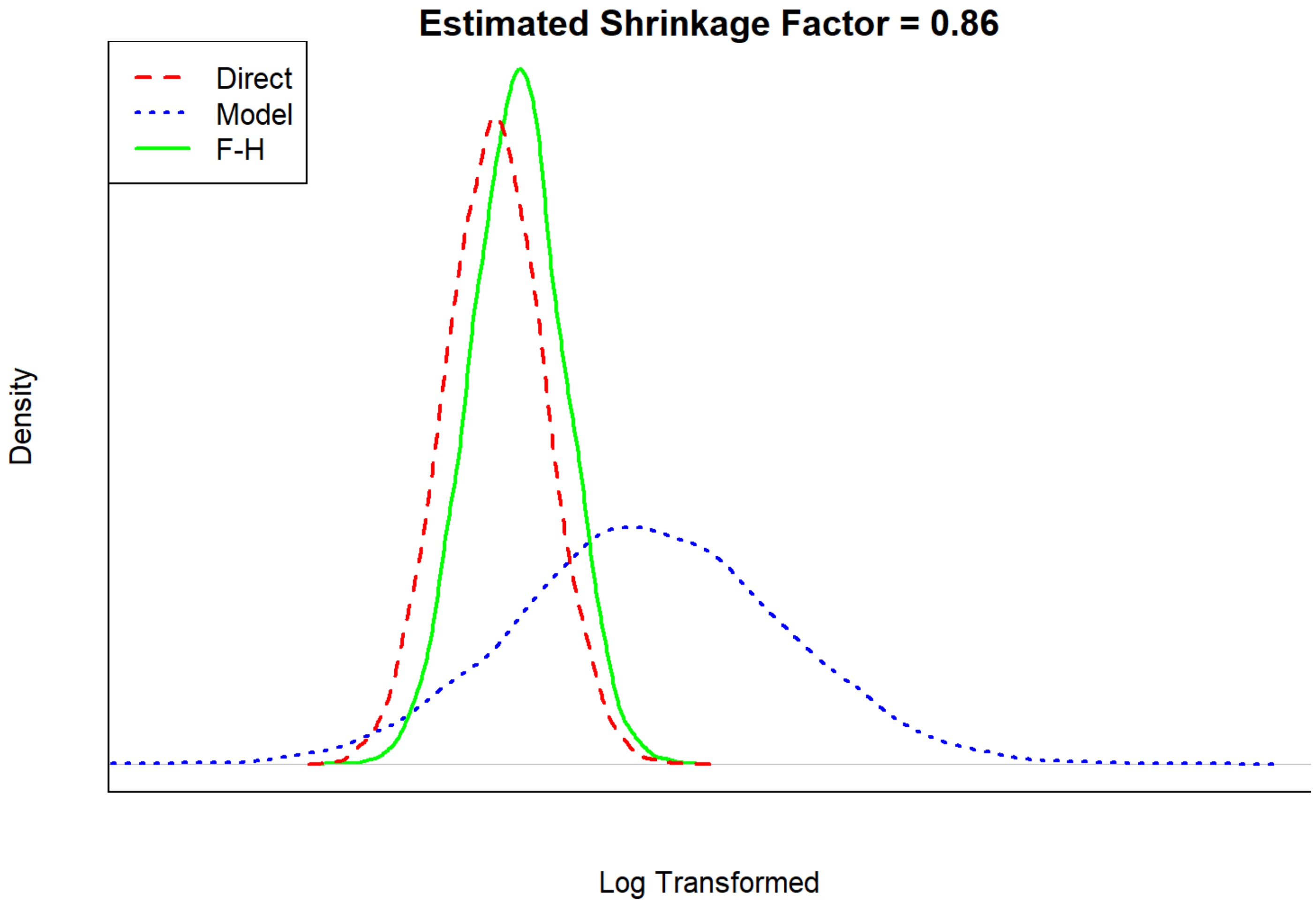
$$Y_d \mid \beta, \sigma^2_{Mod} \sim normal(x'_d\beta, \sigma^2_{Mod})$$

Uninformative flat priors: $\beta, \sigma^2_{Mod} \propto 1$

FH-BH Posterior Distributions



FH-BH Posterior Distributions



AIES Data and Covariates

- Recall that AIES wants to produce state level estimates by NAICS3 or NAICS4 for annual payroll, 1^{st} quarter payroll, employment, and receipts

- County Business Patterns (CBP)
  - Annual series that provides subnational economic data
  - Sources: Business Register, Report of Organization survey, Economic Census, Annual Survey of Manufactures and Current Business Surveys, administrative record sources.
  - Annual payroll, 1^{st} quarter payroll, employment

- Economic Census
  - Conducted by the U.S. Census Bureau and collects data in years ending in 2 and 7
  - Annual payroll, 1^{st} quarter payroll, employment, receipts

- Sourcing good covariates is a large part of the research

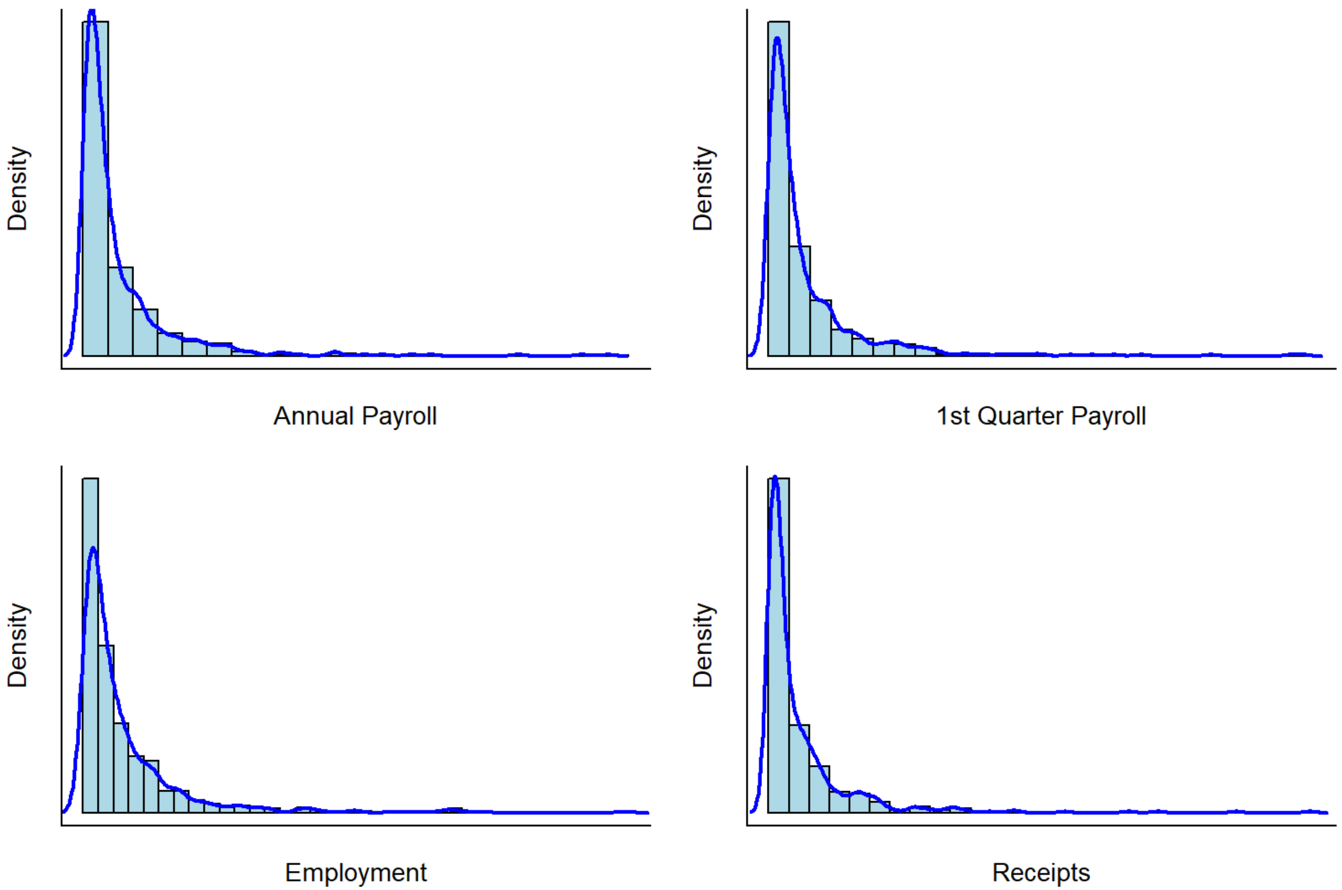
United States® Census Bureau

What does our econ data look like?

- Are they normally distributed?
- What is the correlation structure?
- Which variables have strong predictive power?
- Example using Econ Census data for an undisclosed sector…

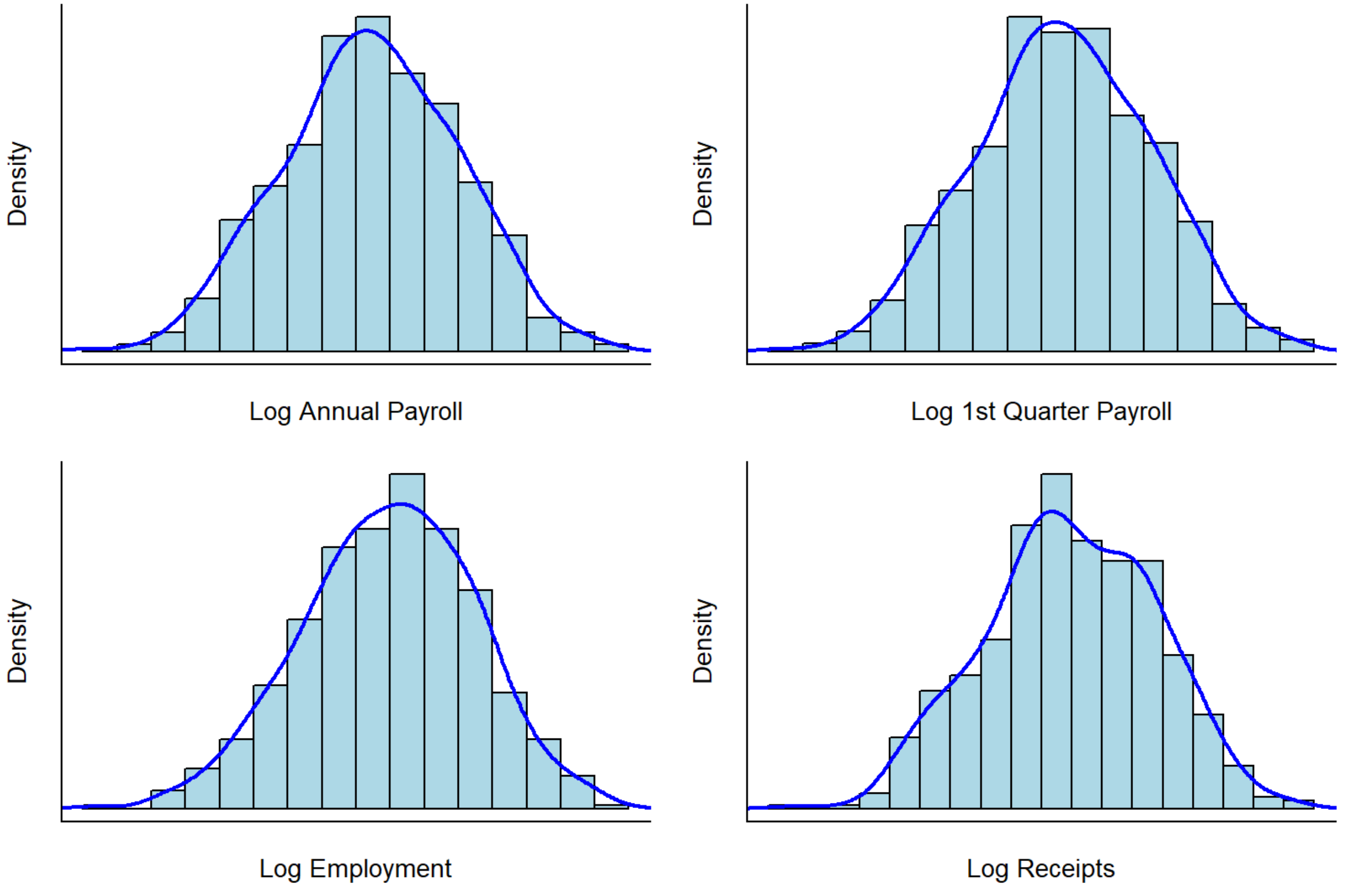
United States® Census Bureau

Econ Census State Estimates



Source: US Census Bureau's 2017 Economic Census

Econ Census State Estimates



Note: Assume log transformation for parameter estimation

Source: US Census Bureau's 2017 Economic Census

United States® Census Bureau

Econ Census State Estimates

Source: US Census Bureau’s 2017 Economic Census

Covariate Evaluations

- Response: Econ Census state estimates
  - Annual payroll, 1st quarter payroll, employment, and receipts

- Covariates
  - Prior year CBP annual payroll (P)
  - Prior year CBP Q1 payroll (Q)
  - Prior year CBP employment (E)
  - Prior Econ Census receipts (R)

- For each response variables we have 15 possible regression models

Covariate Evaluations

- Response: Econ Census state estimates
  - Annual payroll, 1st quarter payroll, employment, and receipts

- Covariates
  - Prior year CBP annual payroll (P)
  - Prior year CBP Q1 payroll (Q)
  - Prior year CBP employment (E)
  - Prior Econ Census receipts (R)

- For each response variables we have 15 possible regression models

United States® Census Bureau

Regression Models

Model	Name
$Y_d = \beta_0 + \beta_1 X_d^{py\_pay} + e_d$	P
$Y_d = \beta_0 + \beta_2 X_d^{py\_qtr1} + e_d$	Q
$Y_d = \beta_0 + \beta_3 X_d^{py\_emp} + e_d$	E
$Y_d = \beta_0 + \beta_4 X_d^{pc\_rcpt} + e_d$	R
$Y_d = \beta_0 + \beta_1 X_d^{py\_pay} + \beta_2 X_d^{py\_qtr1} + e_d$	P+Q
$Y_d = \beta_0 + \beta_1 X_d^{py\_pay} + \beta_3 X_d^{py\_emp} + e_d$	P+E
$Y_d = \beta_0 + \beta_1 X_d^{py\_pay} + \beta_4 X_d^{pc\_rcpt} + e_d$	P+R
$Y_d = \beta_0 + \beta_2 X_d^{py\_qtr1} + \beta_3 X_d^{py\_emp} + e_d$	Q+E
$Y_d = \beta_0 + \beta_2 X_d^{py\_qtr1} + \beta_4 X_d^{pc\_rcpt} + e_d$	Q+R
$Y_d = \beta_0 + \beta_3 X_d^{py\_emp} + \beta_4 X_d^{pc\_rcpt} + e_d$	E+R
$Y_d = \beta_0 + \beta_1 X_d^{py\_pay} + \beta_2 X_d^{py\_qtr1} + \beta_3 X_d^{py\_emp} + e_d$	P+Q+E
$Y_d = \beta_0 + \beta_1 X_d^{py\_pay} + \beta_2 X_d^{py\_qtr1} + \beta_4 X_d^{pc\_rcpt} + e_d$	P+Q+R
$Y_d = \beta_0 + \beta_1 X_d^{py\_pay} + \beta_3 X_d^{py\_emp} + \beta_4 X_d^{pc\_rcpt} + e_d$	P+E+R
$Y_d = \beta_0 + \beta_2 X_d^{py\_qtr1} + \beta_3 X_d^{py\_emp} + \beta_4 X_d^{pc\_rcpt} + e_d$	Q+E+R
$Y_d = \beta_0 + \beta_1 X_d^{py\_pay} + \beta_2 X_d^{py\_qtr1} + \beta_3 X_d^{py\_emp} + \beta_4 X_d^{pc\_rcpt} + e_d$	P+Q+E+R

Note: Response variable and Industry subscript have been removed to simplify notation

Regression Evaluation

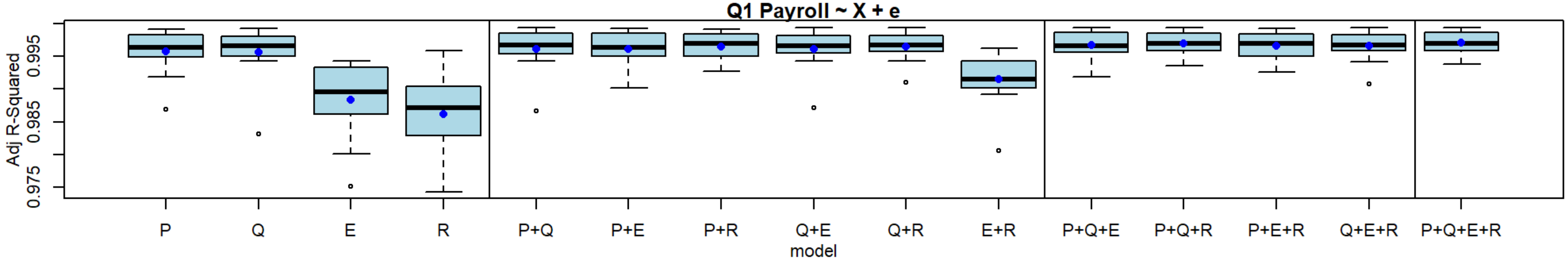
- Fit each regression model (15) at the industry estimation level (≈ 10) for each response variable

- Evaluate model diagnostics (one for each industry)
  - Adjusted R-squared
  - Residual standard error
  - Parameter significance

- Example using data for an undisclosed sector…

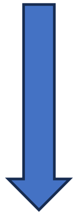
United States® Census Bureau

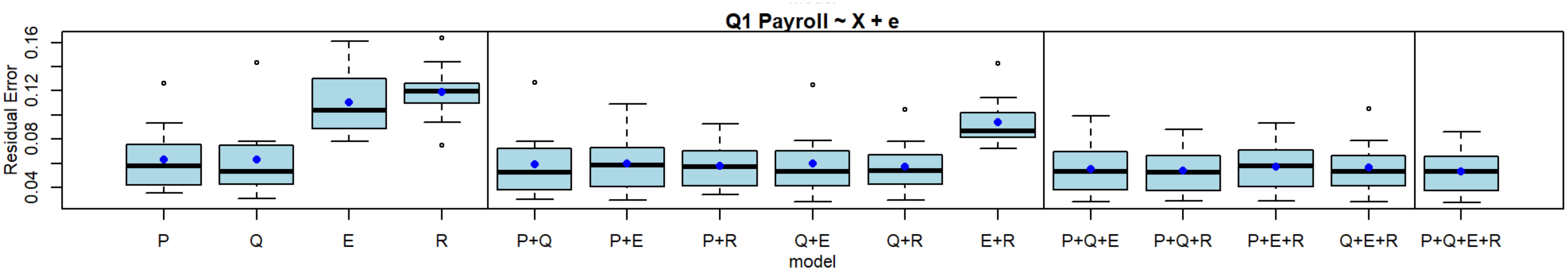
Q1 Payroll Evaluation



Source: U.S. Census Bureau, [Project No. P-7529180 / Approval CBDRB-FY24-ESMD001-012]

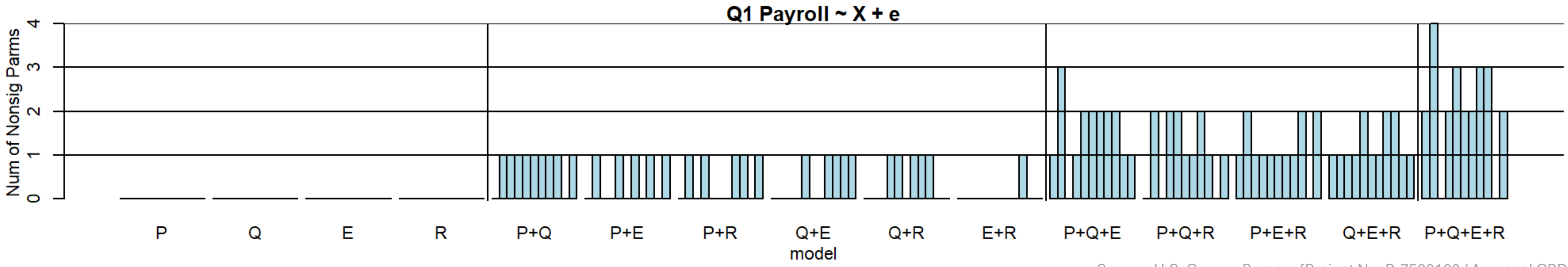
Q1 Payroll Evaluation





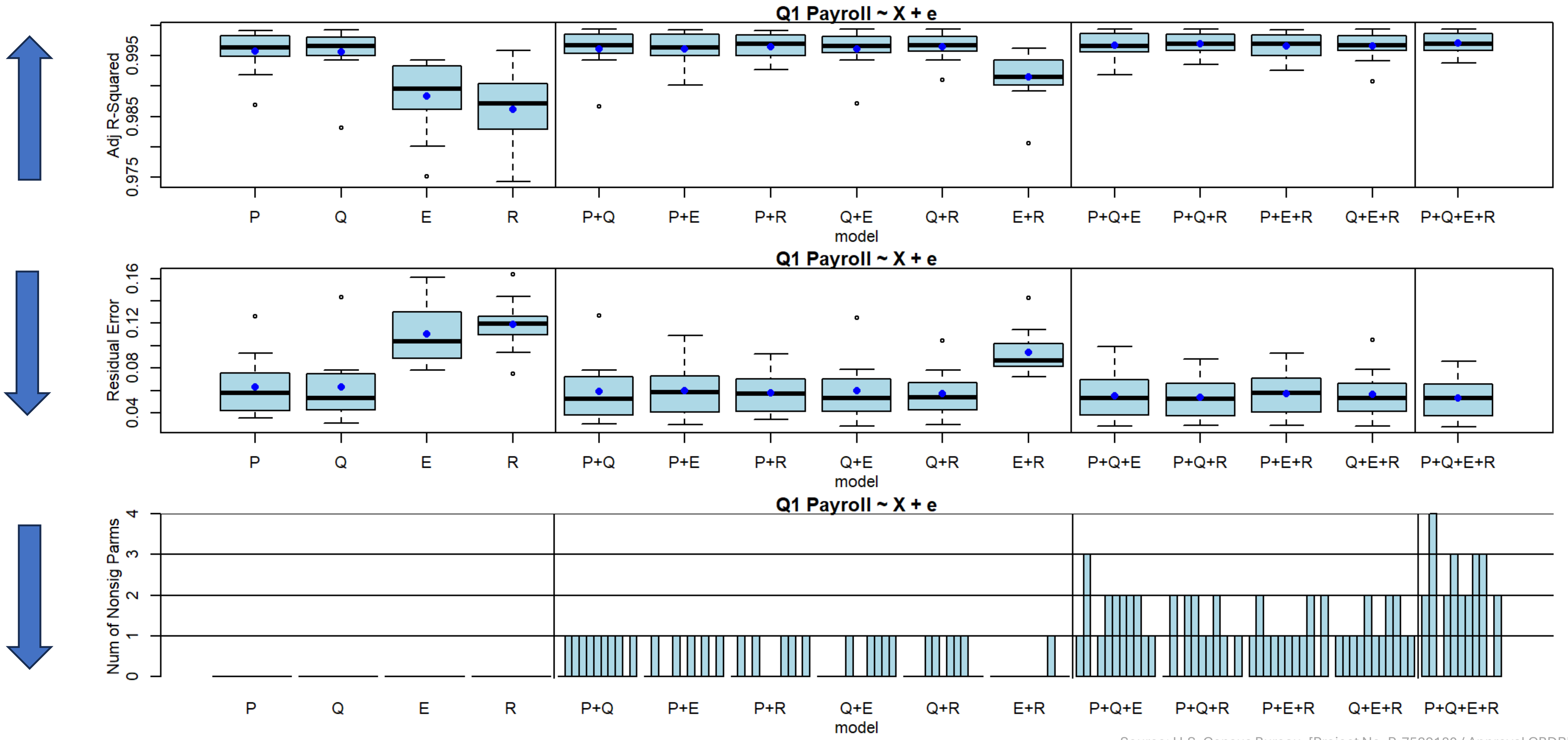
Source: U.S. Census Bureau, [Project No. P-7529180 / Approval CBDRB-FY24-ESMD001-012]

Q1 Payroll Evaluation



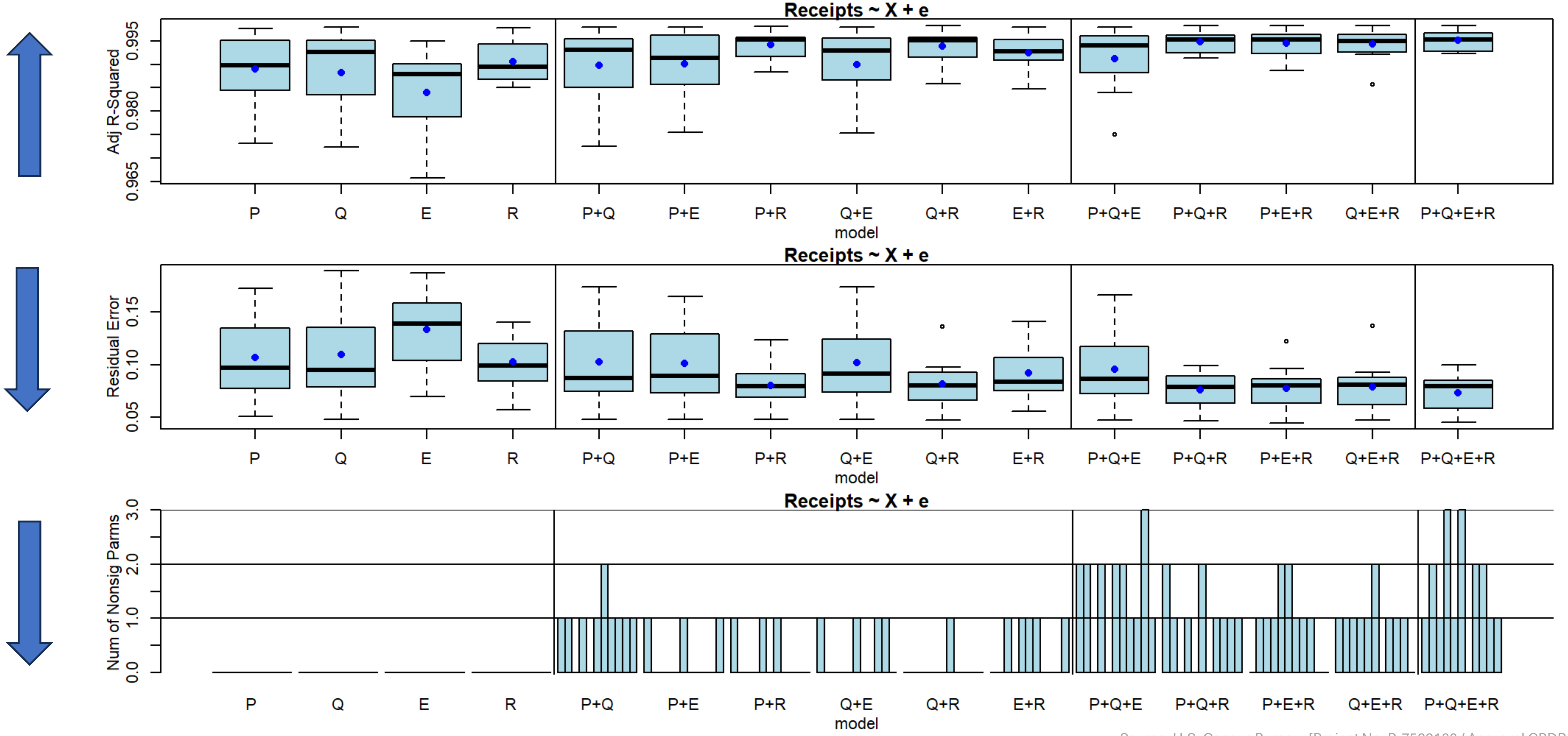
Source: U.S. Census Bureau, [Project No. P-7529180 / Approval CBDRB-FY24-ESMD001-012]

Q1 Payroll Evaluation



Source: U.S. Census Bureau, [Project No. P-7529180 / Approval CBDRB-FY24-ESMD001-012]

Receipts Evaluation



Source: U.S. Census Bureau, [Project No. P-7529180 / Approval CBDRB-FY24-ESMD001-012]

Sector level Suggested Linking Model

- Annual Payroll

$$Y_d^{pay} = \beta_0 + \beta_1 X_d^{\mathrm{py\_}pay} + e_d$$

- Q1 Payroll

$$Y_d^{qtr1} = \beta_0 + \beta_1 X_d^{\mathrm{py\_}qtr1} + e_d$$

- Employment

$$Y_d^{emp} = \beta_0 + \beta_1 X_d^{\mathrm{py\_}emp} + e_d$$

- Receipts

$$Y_d^{rcpt} = \beta_0 + \beta_1 X_d^{\mathrm{py\_}pay} + \beta_2 X_d^{\mathrm{pc\_}rcpt} + e_d$$

Fay - Herriot Example

- Current production sample

- Simulate full response using Census and CBP data
  - Please do not draw inference about the population

- Produce F-H estimates for two variables using “simple” models our sector

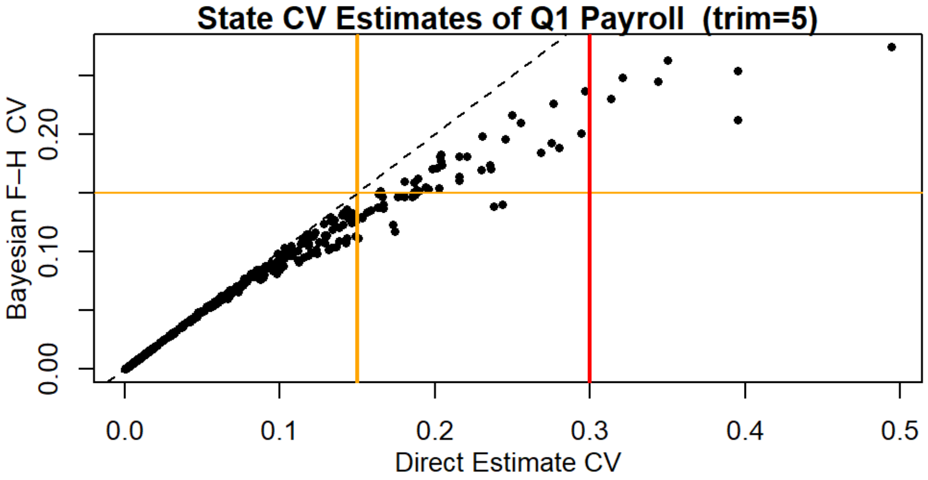
  - Q1 Payroll: $Y_d^{qtr1} = \beta_0 + \beta_1 X_d^{\mathrm{py}\_qtr1} + e_d$
  - Receipts: $Y_d^{rcpt} = \beta_0 + \beta_1 X_d^{\mathrm{py}\_pay} + \beta_2 X_d^{\mathrm{pc}\_rcpt} + e_d$

United States® Census Bureau

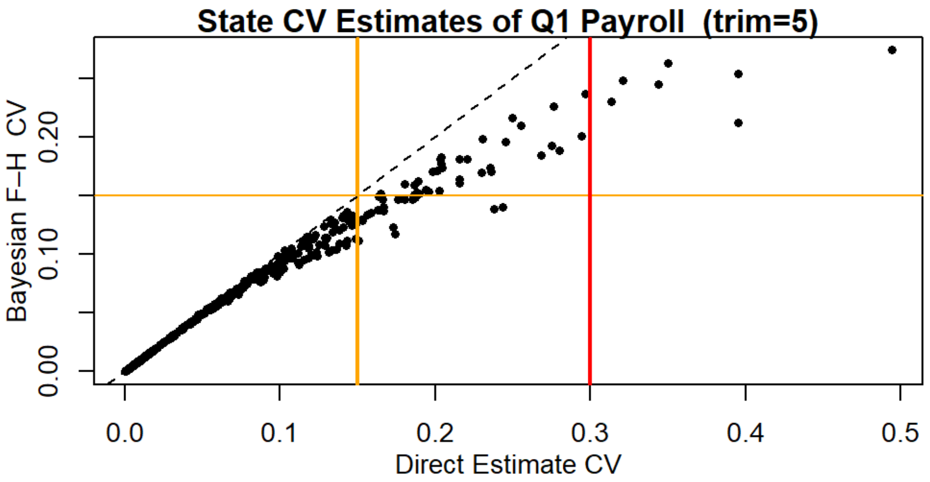
Fay- Herriot Model diagnostics

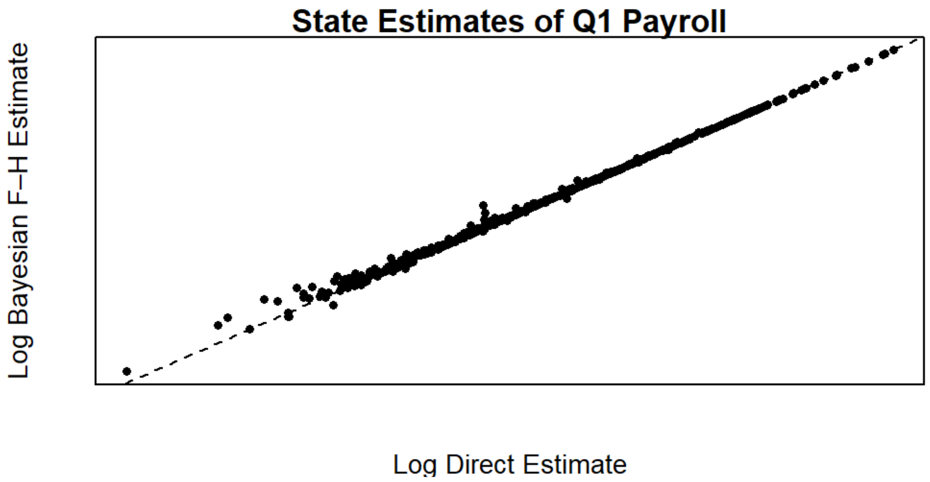
- Coefficient of Variation (CV): $\sigma_d / \hat{Y}_d$
- Change in estimates
- Shrinkage Factor

Q1 Payroll Model Diagnostics



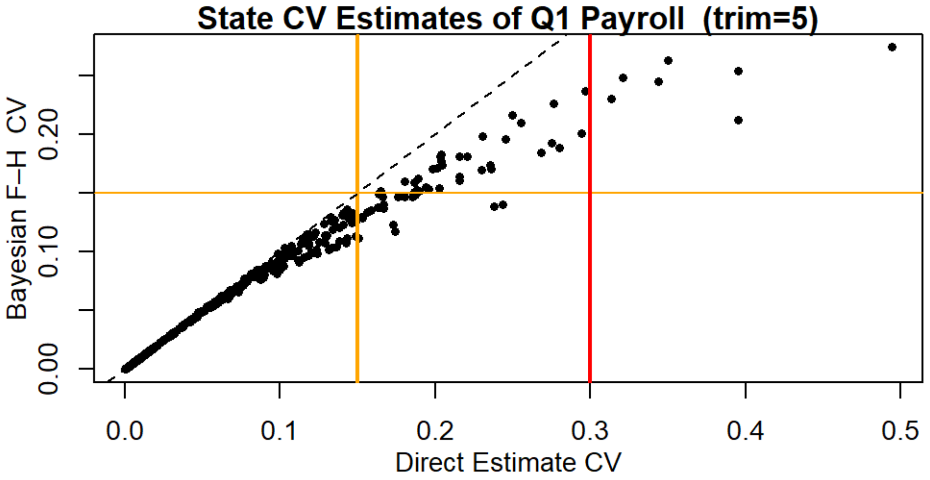
Q1 Payroll Model Diagnostics

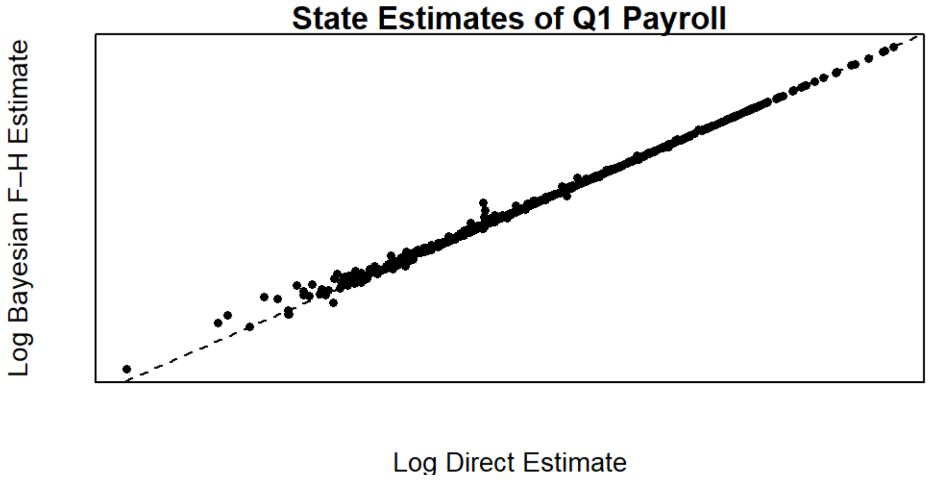


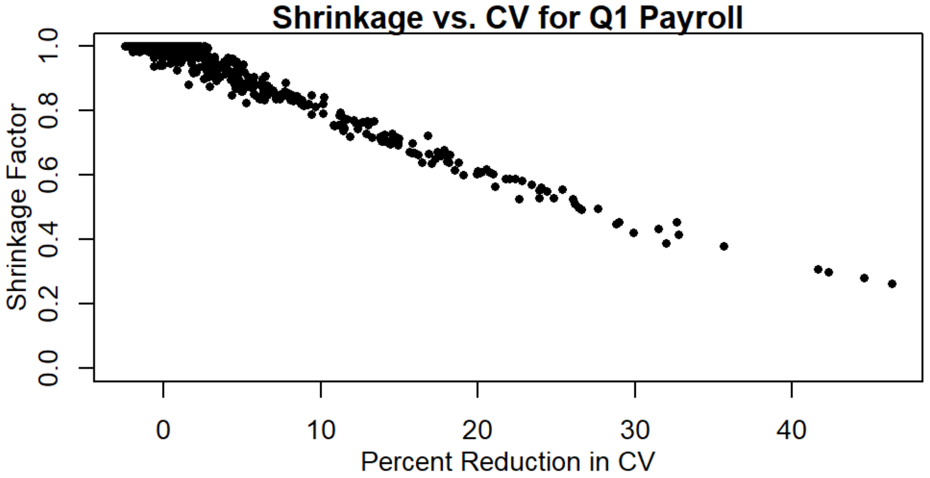


Source: U.S. Census Bureau, [Project No. P-7529180 / Approval CBDRB-FY24-ESMD001-012]

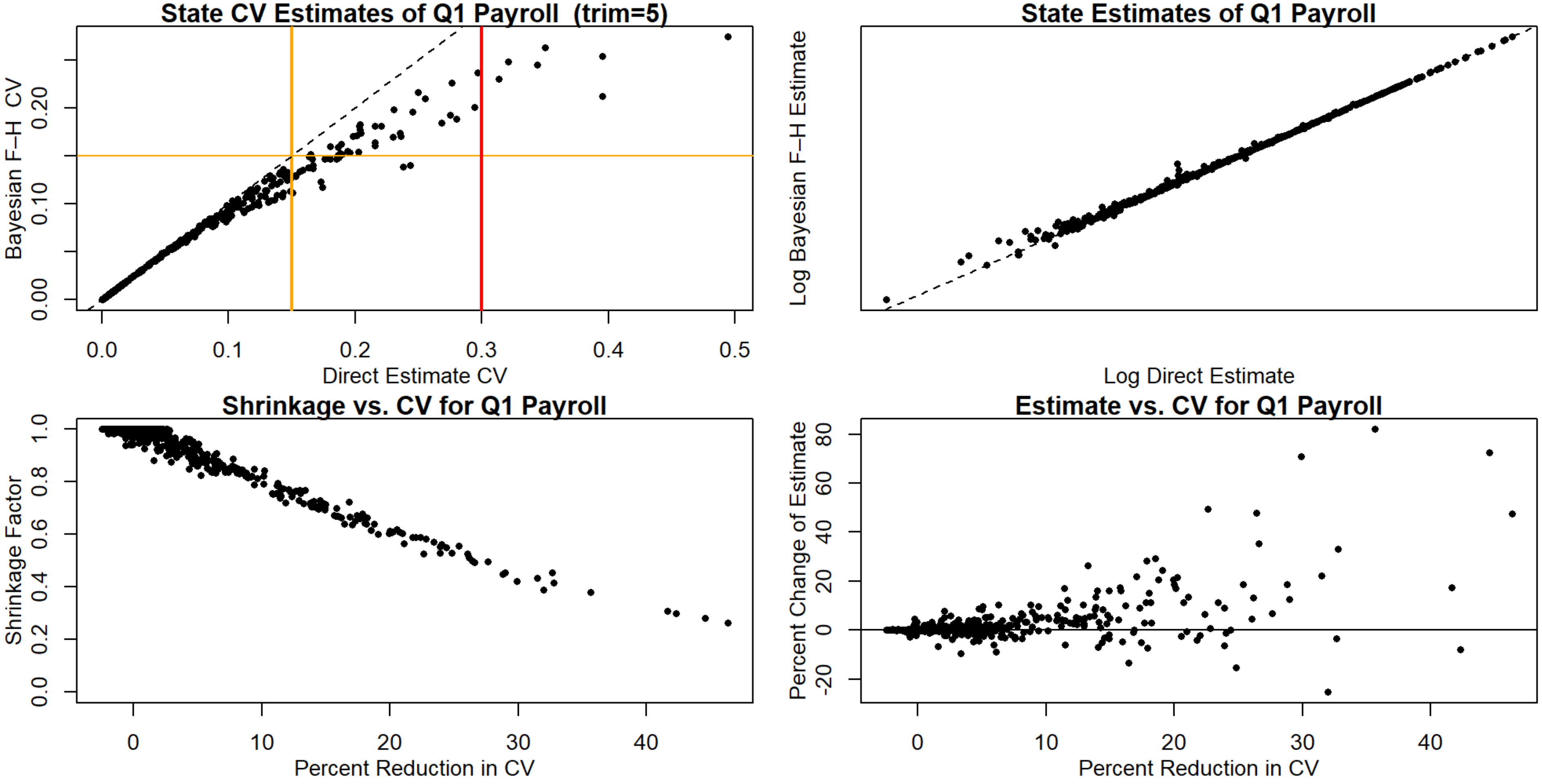
Q1 Payroll Model Diagnostics





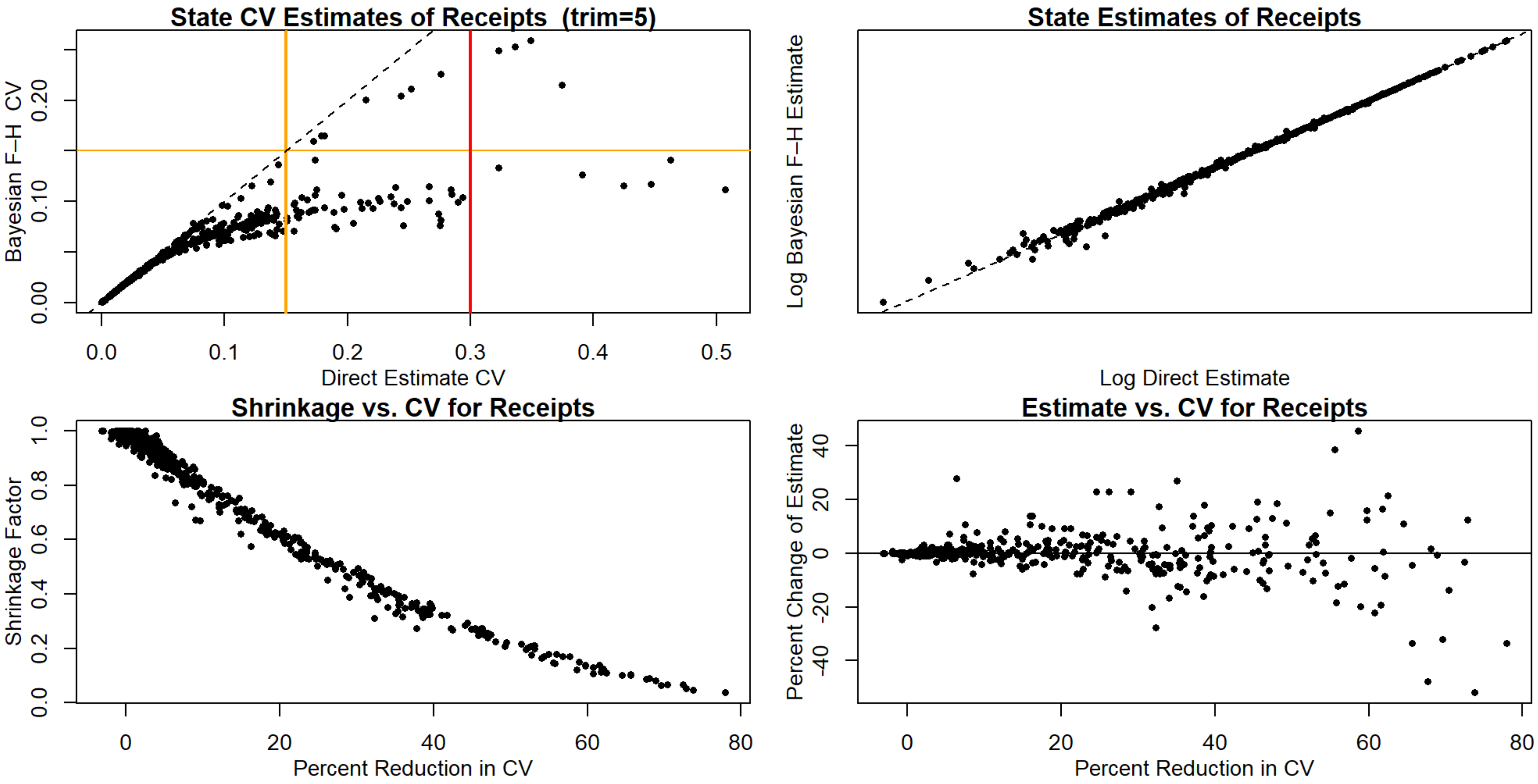


Q1 Payroll Model Diagnostics



Source: U.S. Census Bureau, [Project No. P-7529180 / Approval CBDRB-FY24-ESMD001-012]

Receipts Model Diagnostics



Summary

- F-H model reduces the Coefficient of Variation (CV) when compared to direct survey estimates
- Minimal changes to direct survey estimates of adequate precision
- Can we do better…

Current Research

- Should we only model noncertainty tabulations?
  - Certainty tabulations have no sampling variance
  - Noncertainty covariates are harder to create
    - Link prior year data to the production frame to create cert/noncert tabs
    - Run prior year data through AIES sample design process to create cert/noncert tabs

- Linear mixed model (Fay-Herriot)

$$\hat{Y}_d^{Dir,nc} = X_d^{nc\prime}\beta + e_d^{Mod} + e_d^{Dir}$$

- New Estimator

$$\hat{Y}_d^{FH} = \hat{Y}_d^{Dir,c} + \left[\hat{\gamma}_d \hat{Y}_d^{Dir,nc} + (1 - \hat{\gamma}_d) X_d^{nc\prime}\hat{\beta}\right]$$

United States® Census Bureau

Future Work and Research

- Continue to investigate combined vs noncertainty models
  - Promising results!
- Automatic model selection at the industry within sector level
- Treating sampling variances as estimates
  - Should we be modeling the sampling variance?
  - Can add stability to small domain variances
  - Challenging for the first year of a survey
- Disclosure risk associated with Fay- Herriot estimates

United States® Census Bureau

Thank you!

stephen.kaputa@census.gov

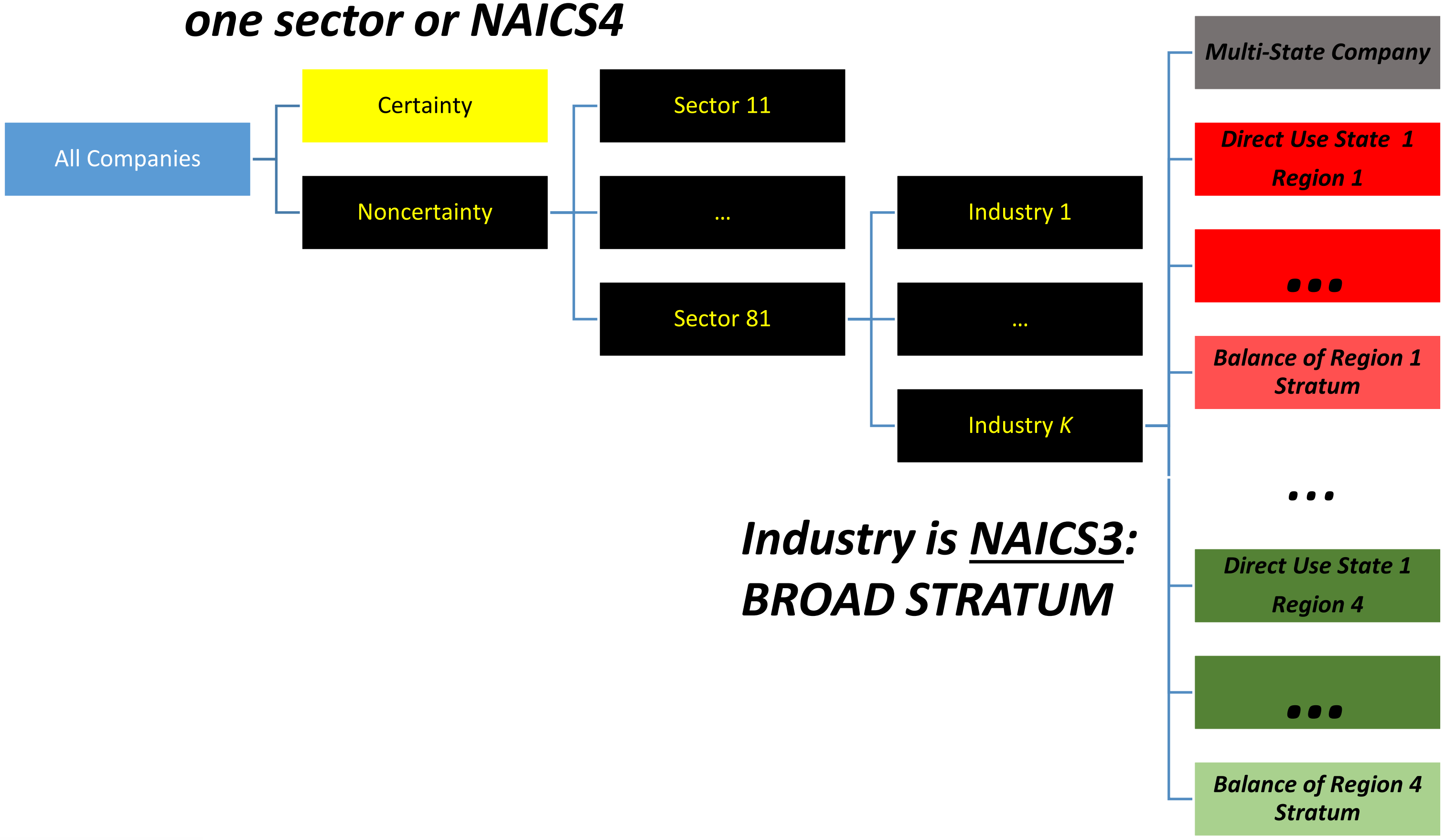
Appendix

AIES Sample Design

- Frame is created from the Business Register
- Sampling unit is company (firm)
- Stratification
  - Certainty – included with probability 1
  - Noncertainty – separated by sector

AIES Stratification of Companies

Operates in more than one sector or NAICS4

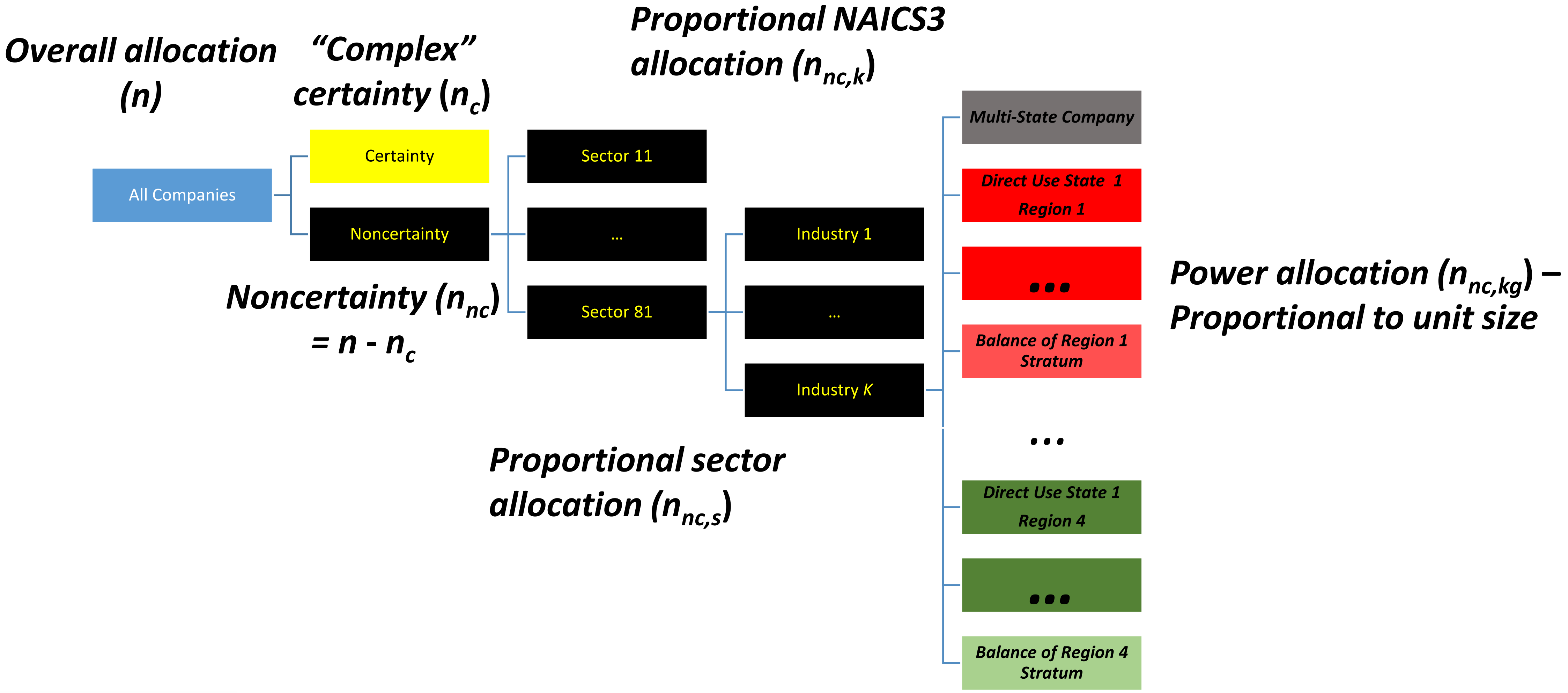


Operates establishments in more than one state

27 Geographic Strata per Industry (NAICS3)

Industry is <u>NAICS3</u>: BROAD STRATUM

AIES Allocation



AIES Sample Design

- Stratified sequential random sampling (Chromy, 1979)
  - Companies sorted within sampling strata
  - Fixed-size unequal probability sample without replacement
- Domain estimates
- Ratio estimation (Post-stratification)
  - Separate adjustments for national and subnational estimates
- Variances are estimated using a bootstrap method (Antal and Tillé, 2011)