More Information is Better! Where Can We Get It and How Can We Use It?

Stephanie Coffey

Joint Statistical Meetings

Baltimore, 7/31/2017

Overview

- Motivate using survey progress monitoring
- Explain the survey progress metric of interest
- Discuss why we need information/priors
- Mention some potential sources of information
- Illustrate different ways one type of info might be used
  - Non-Bayesian and Bayesian methods for using external information
  - Compare to the actual survey progress metric
  - Pros and cons of different ways to use information
- Discussion

Motivation

- Surveys are expensive
  - Fieldwork is a high proportion of costs of surveys
- Surveys are complex - can encompass many data collection and processing operations
  - Some ops can be managed/monitored fairly easily
  - Fieldwork is decentralized and relies on interviewers
- Understanding and monitoring progress in the field is important for meeting overall goals

Motivation

- Managing fieldwork requires expectations of what “should” or “needs to” happen
- Variety of metrics developed and monitored
  - # of hours per complete
  - # of attempts before complete
  - # of days between attempts

Motivation

- Expectations are used in multiple ways
  - Overall data collection progress
    - *"Does it look like we will meet overall survey goals?"*
    - *"How much more/less money/time/features/staff should we have used to better meet goals?"*
  - Monitor specific interviewer behavior
    - *"Does this interviewer look like they are on track to met performance metrics?"*
    - *"Did this interviewer meet final performance metrics?"*

Progress Metric Example

- Time Lag Between 1st Attempt and 1st Contact
  - How many days it takes from the time we first attempt a case to when we first make contact
- Why could lags be longer (or shorter)?
  - Interviewers
  - Geographic Location of Cases
  - Weather
- Why is this interesting?
  - Contact and response are really two different processes
  - Response cannot occur without contact
  - Metric might not be consistent over time

External Information (Priors)

- Need to base expectations on “something”
  - Early Part of Current Survey Implementation
  - Past Implementations of the Same Survey
  - Past Data for Similar Surveys
  - Heuristics from Field Management
  - Historical Aggregate Measures from Field
  - Information in the Literature
  - Past Data for Different Surveys

External Information (Priors)

- Need to base expectations on “something”

Early Part of Current Survey Implementation

Past Implementations of the Same Survey

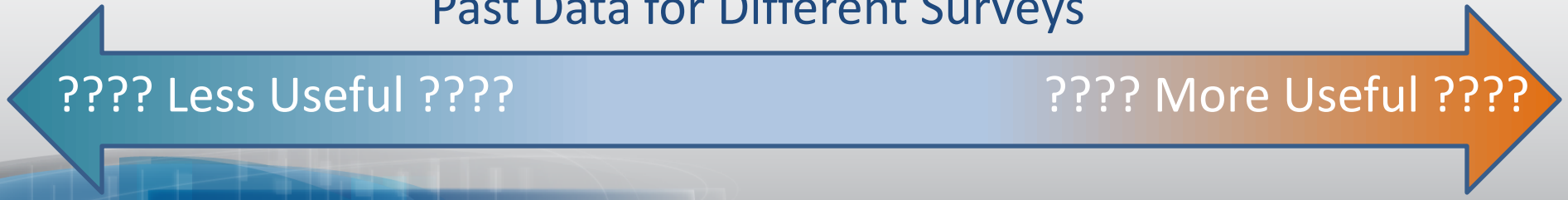
Past Data for Similar Surveys

Heuristics from Field Management

Historical Aggregate Measures from Field

Information in the Literature

Past Data for Different Surveys



Illustration

- Simulated data to represent a national monthly survey at Census
  - 3 months of “historical” data
  - 1 month of “current” data
- Outcome variable: Lag in Days
- Auxiliary covariates:
  - Census region
  - Block-group level sociodemographic characteristics
  - Experience (in years) of FR
  - Reassignment indicator

Illustration

- Display what the “actual” average lag-by-day is for “current” month (benchmark)

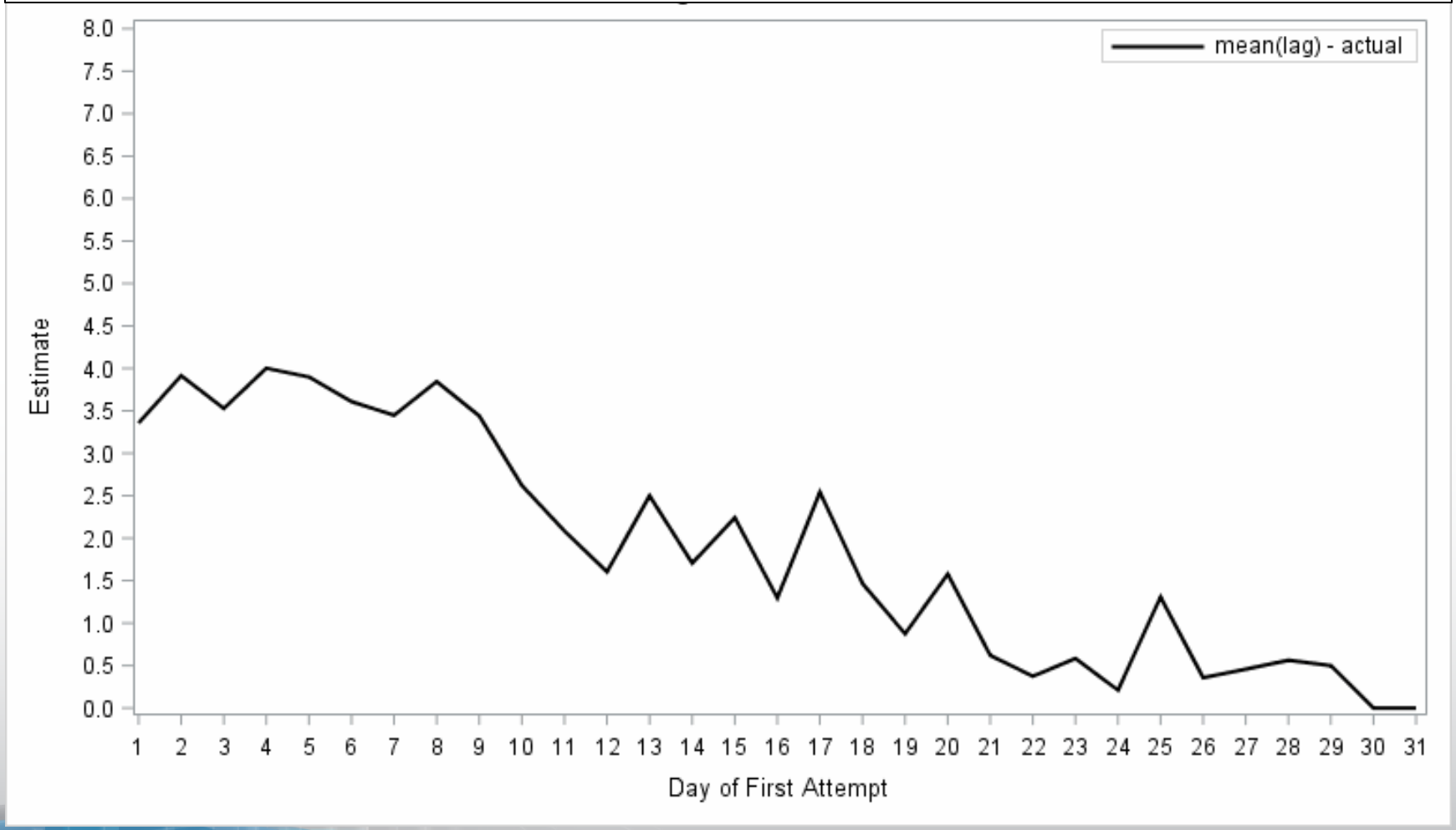
Illustration

- Display what the “actual” average lag-by-day is for “current” month (benchmark)
- Predict estimated lag-by-day for “current” month using:
  - End of month average (3 months historical data)
  - Regression-based prediction (3 months historical data)
  - Regression-based prediction (first week of current data)
  - Regression-based prediction (current data & priors from 3 months historical data)

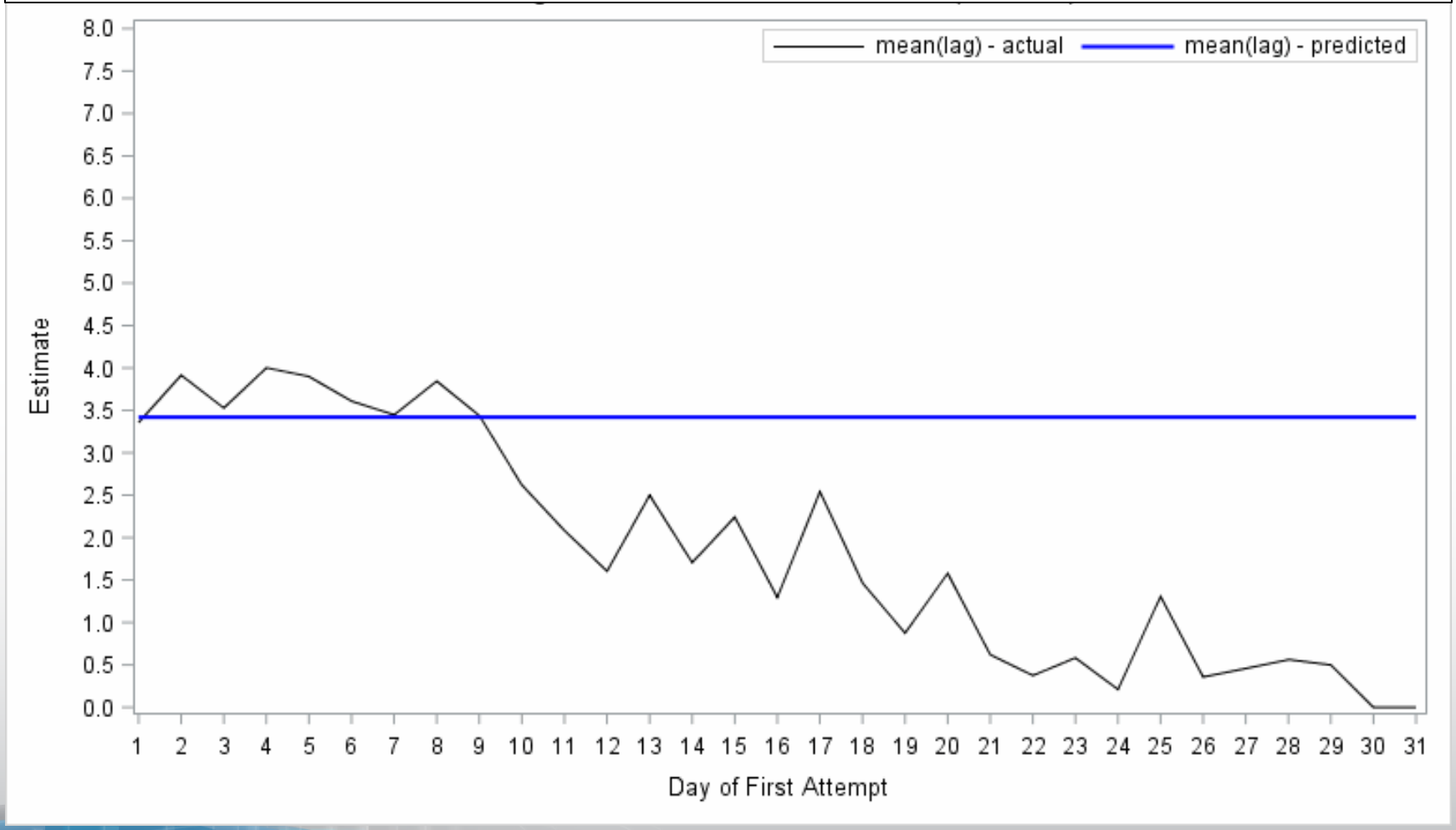
Illustration

- Display what the “actual” average lag-by-day is for “current” month (benchmark)
- Predict estimated lag-by-day for “current” month using:
  - End of month average (3 months historical data)
  - Regression-based prediction (3 months historical data)
  - Regression-based prediction (first week of current data)
  - Regression-based prediction (current data & priors from 3 months historical data)
- Used zero-inflated negative binomial model
  - 60% of cases have contact at first attempt (lag = 0)
  - Variance of the lag is much larger than the mean

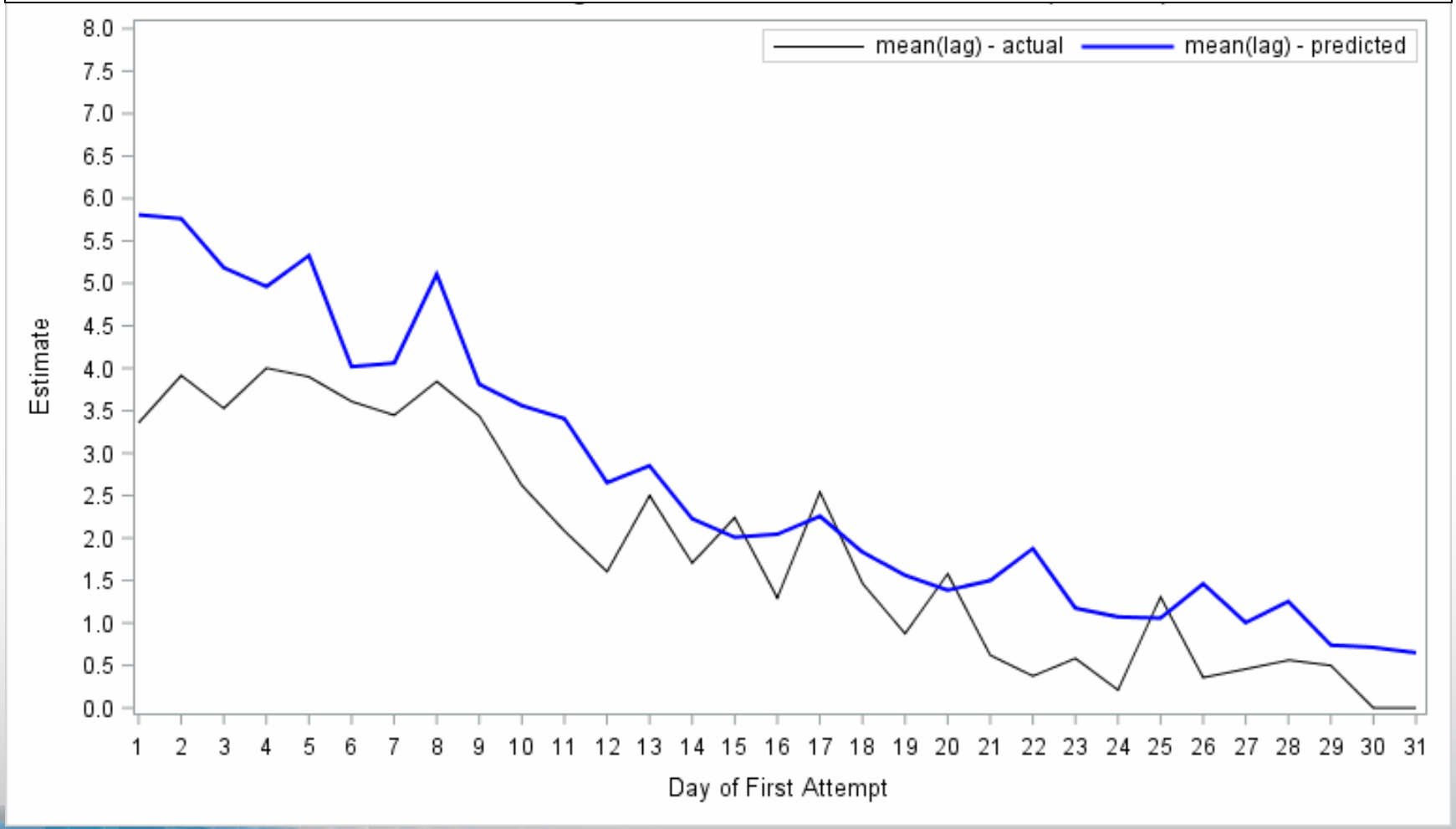
Actual Mean Lag by Day of First Attempt Using Current Round Data



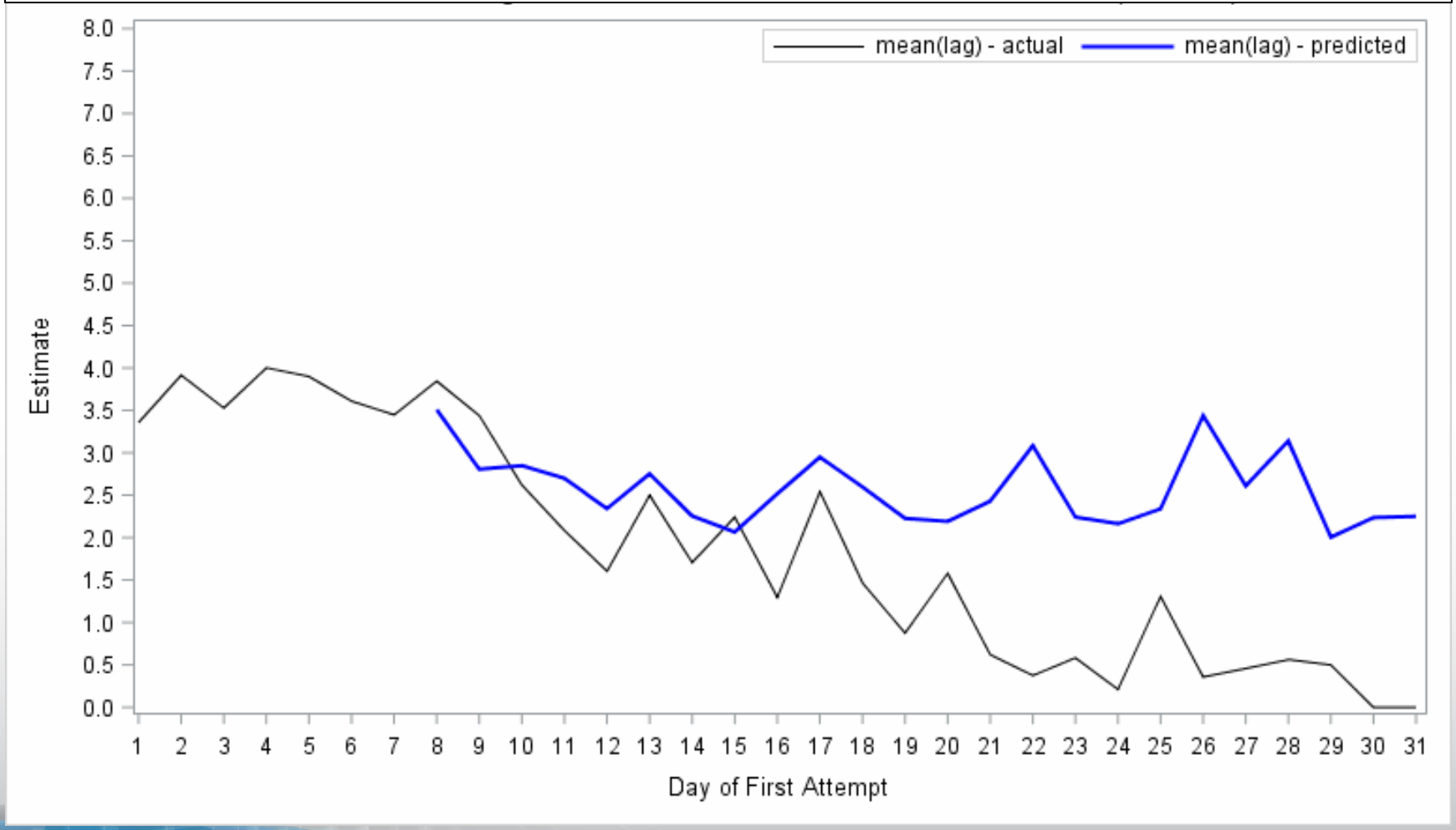
Predicted vs Actual Mean Lag by Day of First Attempt Using Mean of 3 Months Historical Data (Method 1)



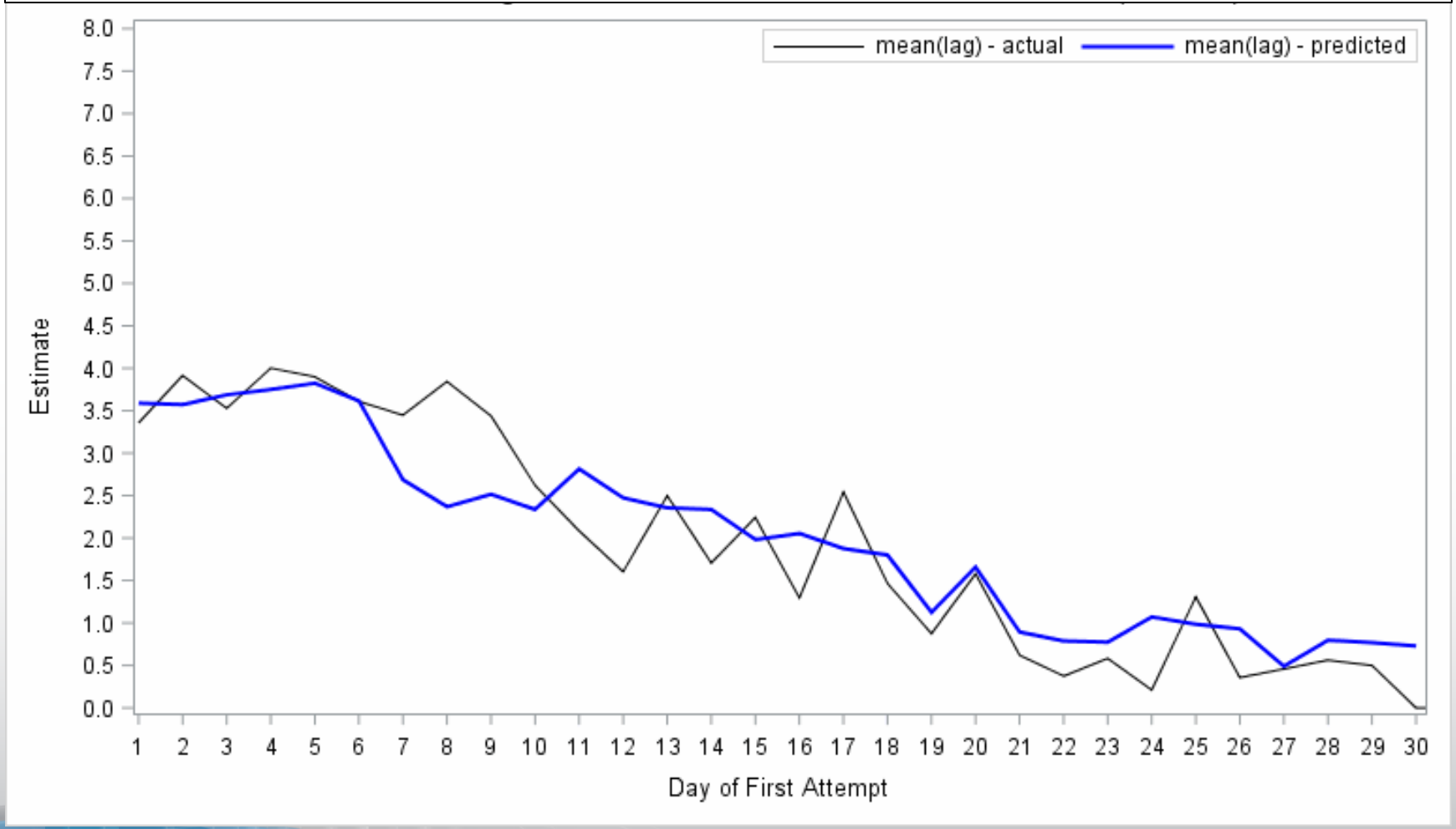
Predicted vs Actual Mean Lag by Day of First Attempt Using ZINB Parameters from Historical Data (Method 2)



Predicted vs Actual Mean Lag by Day of First Attempt Using ZINB Parameters 1st Week Current (Method 3)



Predicted vs Actual Mean Lag by Day of First Attempt Using ZINB Priors and Current Data (Method 4)



Discussion

- Method 4 provided improved estimation
  - Improvements early and late
  - Still diverged where the prior and current data were “different” (or there wasn’t enough data)
- Why did this work well?
  - Behavior of the lag was similar across months
  - Borrow strength from prior months about how the lag changes within a month
- What if our external information isn’t as useful?
  - Increase the variance of priors to reduce influence

Discussion

The Bayesian method performed the best in this example…

Is the Bayesian method always the best way?

Are other ways “wrong”?

Not really…it’s important to consider the goal of the outcome being estimated

Contact Information

stephanie.coffey@census.gov