Using Bayesian Methods to Rank Cases Based on Response Propensity During Data Collection

James Wagner

July 30, 2017

Acknowledgement

The National Survey of Family Growth (NSFG) is conducted by the Centers for Disease Control and Prevention's (CDC's) National Center for Health Statistics (NCHS), under contract #200-2010-33976 with University of Michigan's Institute for Social Research with funding from several agencies of the U.S. Department of Health and Human Services, including CDC/NCHS, the National Institute of Child Health and Human Development (NICHD), the Office of Population Affairs (OPA), and others listed on the NSFG webpage (see http://www.cdc.gov/nchs/nsfg/). The views expressed here do not represent those of NCHS or the other funding agencies.

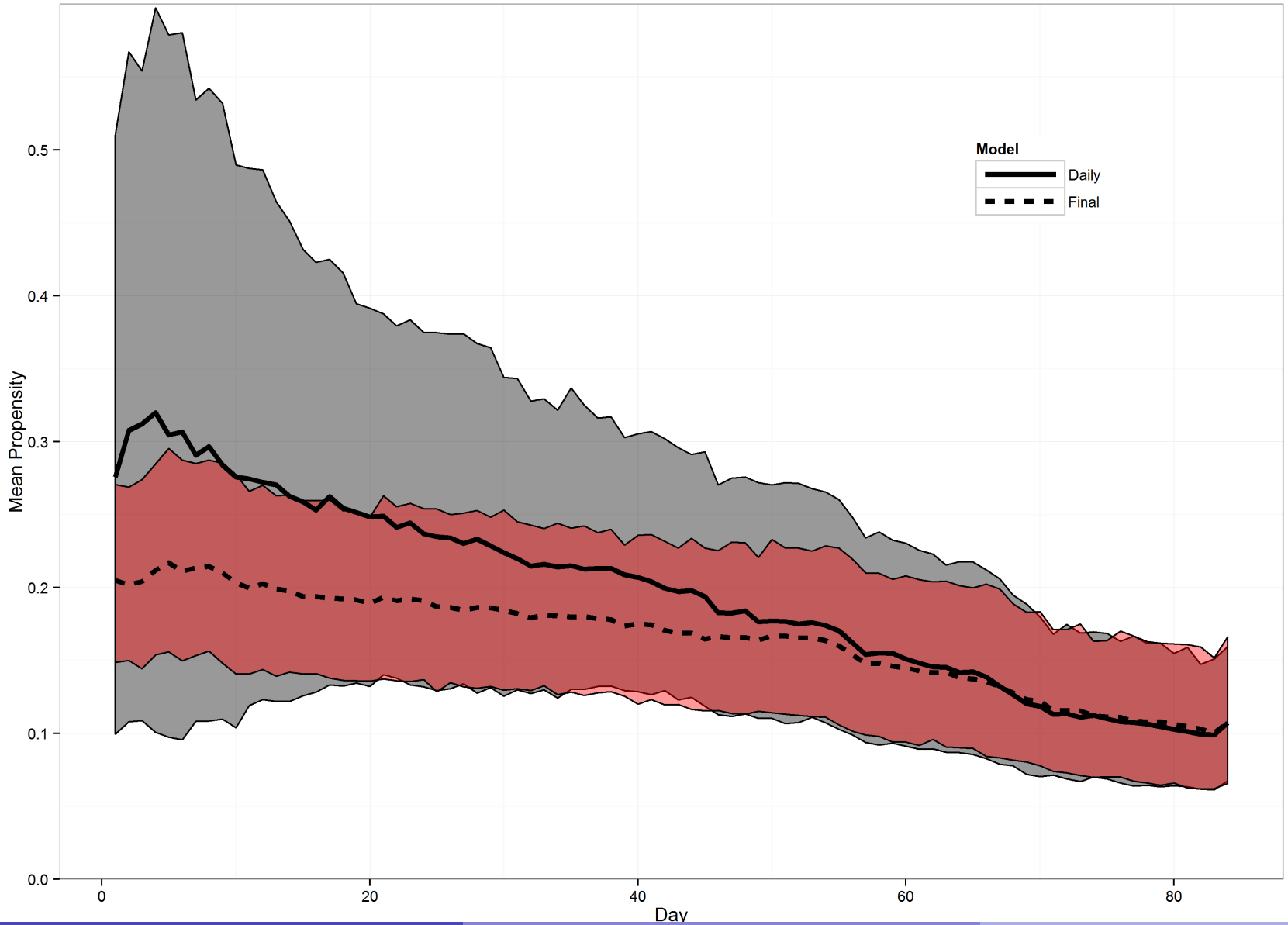
Motivation: Uses for Categories of Estimated Response Propensities

1. Focus effort on low-probability cases
   - *Rosen, et al., 2014*
2. Truncate effort on low-probability cases
   - *Peytchev and Ridenhour, 2013*
3. Match difficult cases to 'best' interviewers
   - *Luiten and Schouten, 2013*

Background

- Response propensity models fit during data collection can be useful
- Model estimates can be biased based on early data
- Bayesian models allow us to add information to the model as a prior
- Difficult to eliminate these biases
- **Can we specify priors such that *rankings* are unbiased?**

Comparison of Two Model Estimates



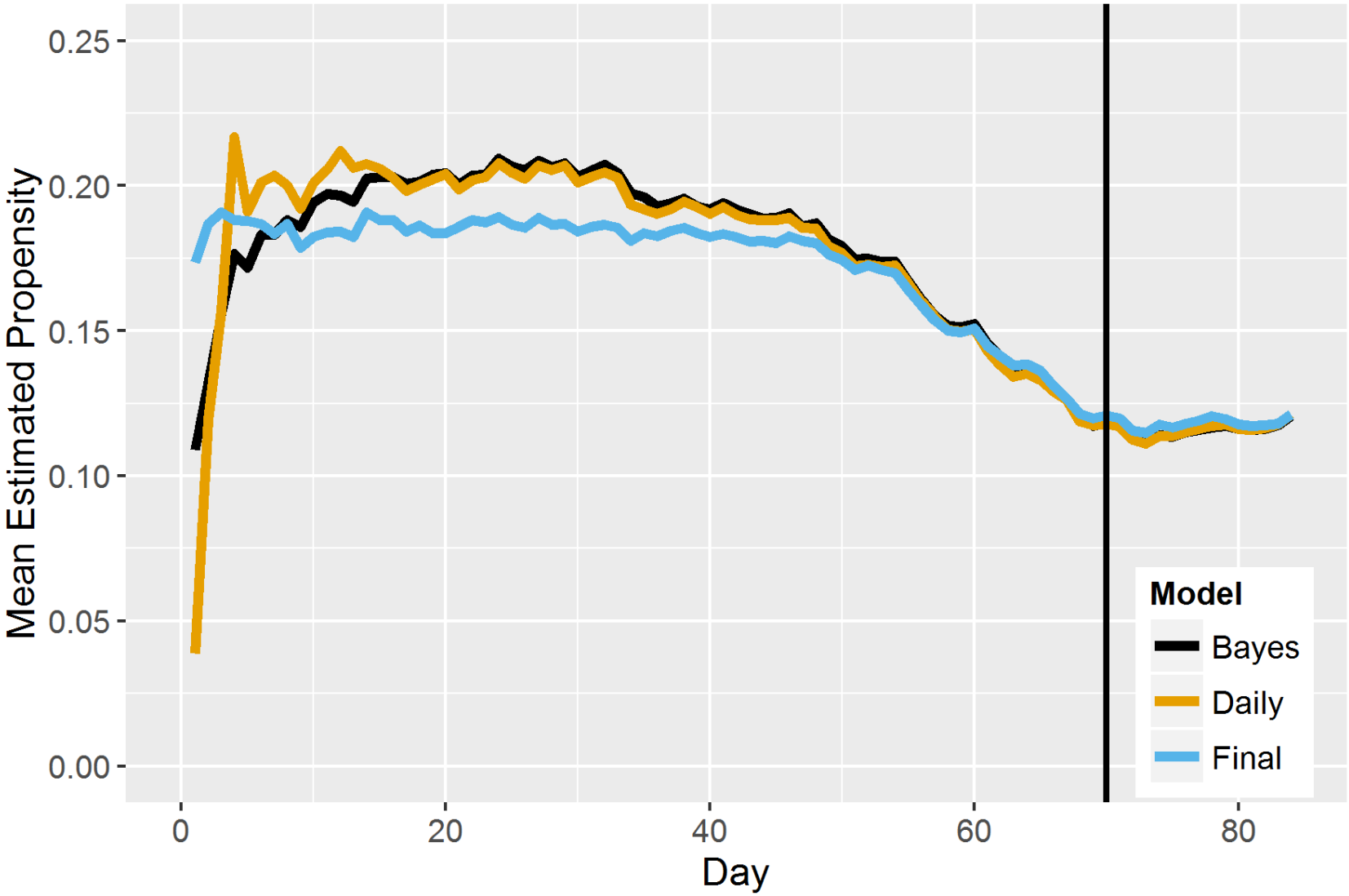
Example: National Survey of Family Growth

- NSFG has quarterly design, 12-week field period
- Survey on feritility and family formation
- Two stages of data collection
  - **Screen** households to identify eligible persons
  - **Interview** eligible persons
- Data include paradata, sampling frame data, commercial data, and (for stage 2) data from screening interview

Can we specify a prior that attenuates bias in rankings?

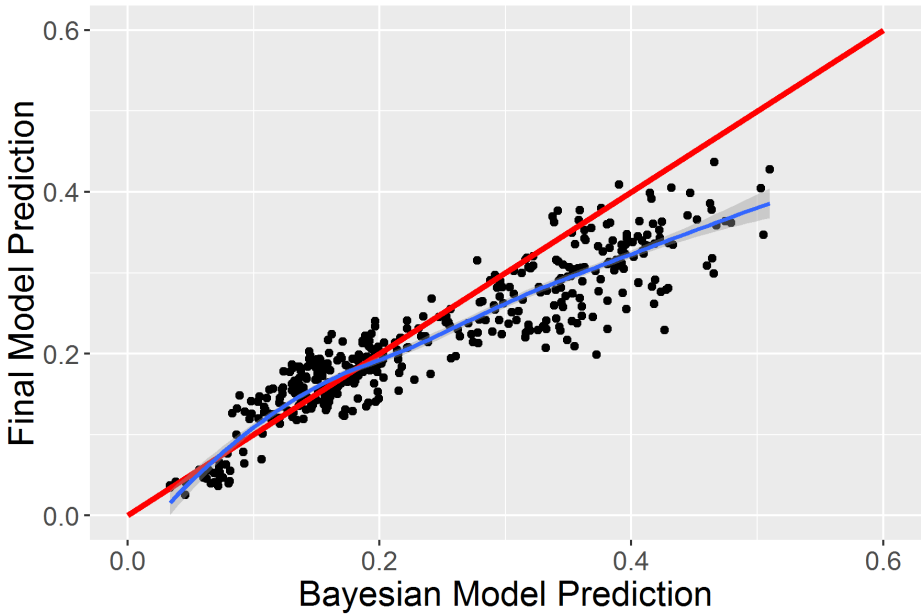
- Set a prior for the intercept using a model
  - $0.209 \times ln(DayNumber) - 2.387$
- Priors for all other coefficients set using data from last 21 days of previous quarter
  - These are the “missing days” from the current quarter

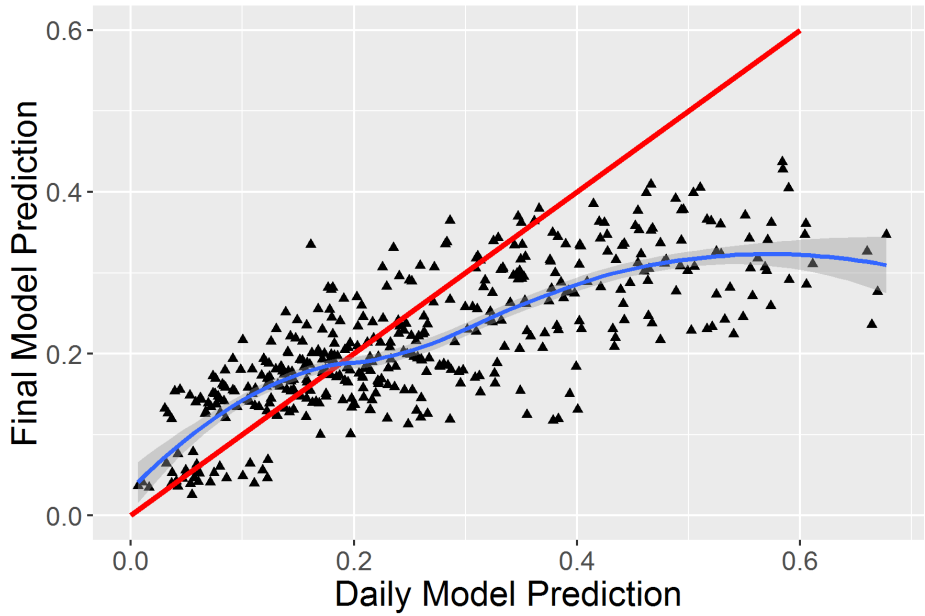
Overall mean: Bayesian model slightly better early



During the first 15 days, the prior information improves the overall estimate

Day 14 Predictions





Quintile Assignment Day 14

Table: Bayes

	Bayes				
Final	0-20	20-40	40-60	60-80	80-100
0-20	76	25	11	0	0
20-40	31	55	22	4	0
40-60	5	31	66	7	2
60-80	0	3	10	75	24
80-100	0	0	0	26	86

	κ	ASE	z	Pr(>|z|)
Unweighted	0.551	0.025	21.7	2.18E-104
Weighted	0.748	0.013	46.13	0.00E+00

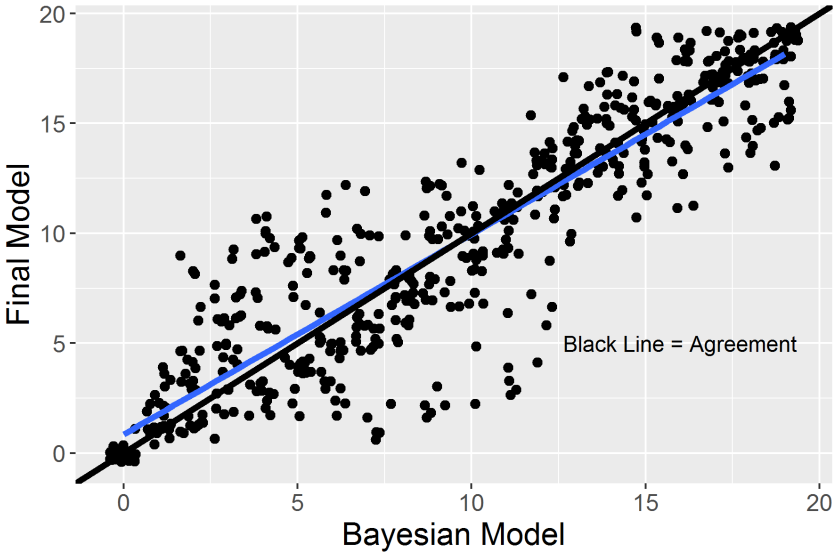
Table: Daily

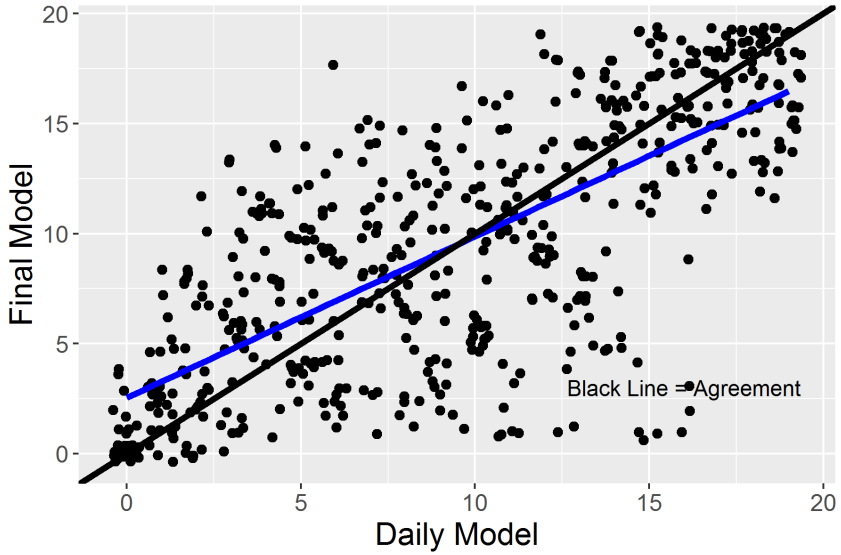
	Daily				
Final	0-20	20-40	40-60	60-80	80-100
0-20	67	18	18	6	3
20-40	29	31	33	19	0
40-60	12	46	30	21	2
60-80	4	16	26	29	37
80-100	0	1	4	38	69

	κ	ASE	z	Pr(>|z|)
Unweighted	0.255	0.026	9.841	7.49E-23
Weighted	0.514	0.023	21.835	1.07E-105

The agreement rate is is 64.0% for the Bayesian model predictions and 40.4% for the Daily model.

Ventile Assignment Day 14





	κ	ASE	z	Pr(>|z|)
Unweighted	0.219	0.019	11.22	3.295e-29
Weighted	0.742	0.013	57.29	0.000e+00

	κ	ASE	z	Pr(>|z|)
Unweighted	0.081	0.015	5.46	4.671e-08
Weighted	0.519	0.021	24.92	4.088e-137

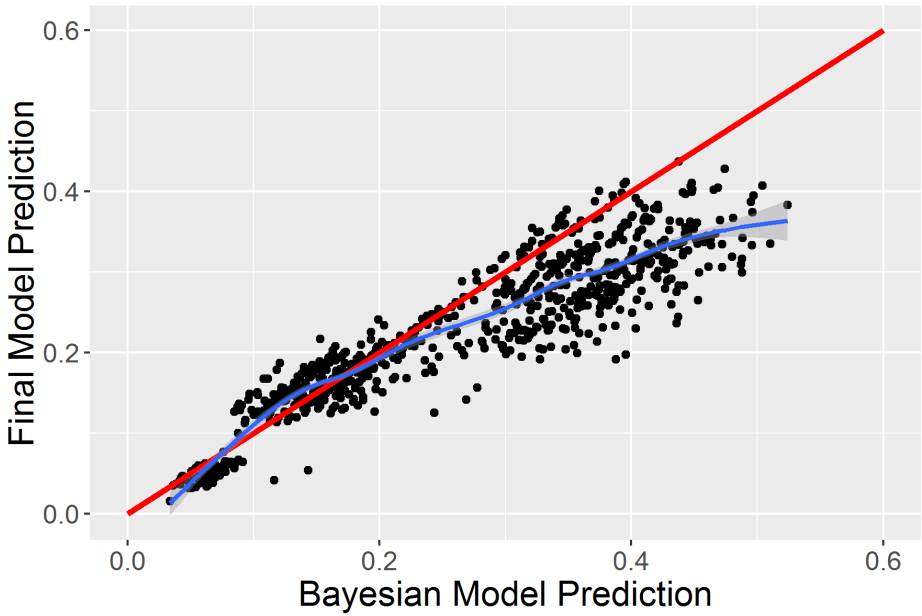
The agreement rate is is 25.8% for the Bayesian model predictions and 12.7% for the Daily model.

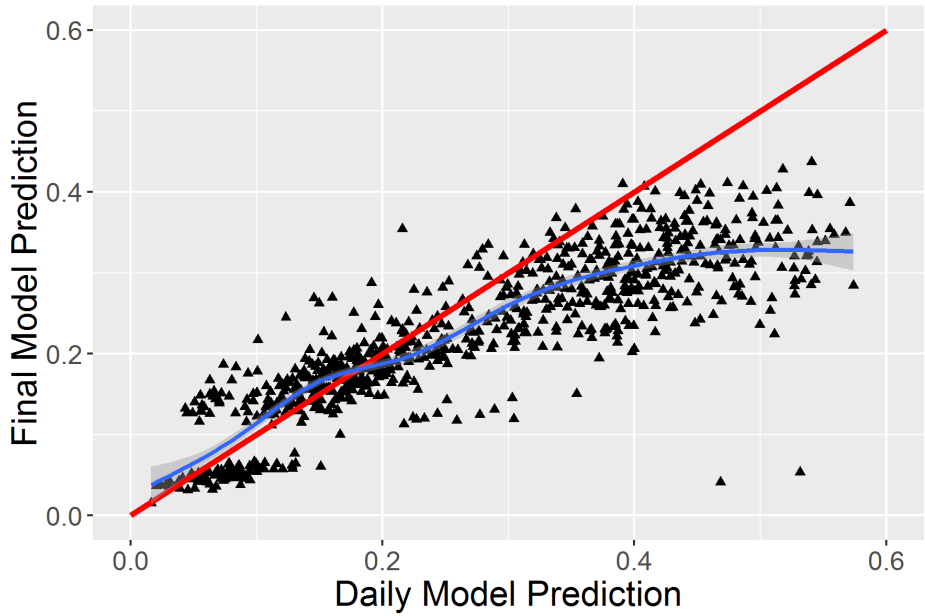
Identifying Fourth Percentile Day 14

	Bayes	
Final	1-4	5-100
1-4	20	3
5-100	3	533

	Daily	
Final	1-4	5-100
1-4	13	10
5-100	10	526

Day 28 Predictions





Quintile Assignment Day 28

Table: Bayes

	Bayes				
Final	0-20	20-40	40-60	60-80	80-100
0-20	176	35	10	0	0
20-40	35	148	36	0	0
40-60	9	38	130	39	4
60-80	0	0	41	103	76
80-100	0	0	2	78	140

	κ	ASE	z	Pr(>|z|)
Unweighted	0.542	0.018	29.85	8.159e-196
Weighted	0.757	0.011	70.47	0.00E+00

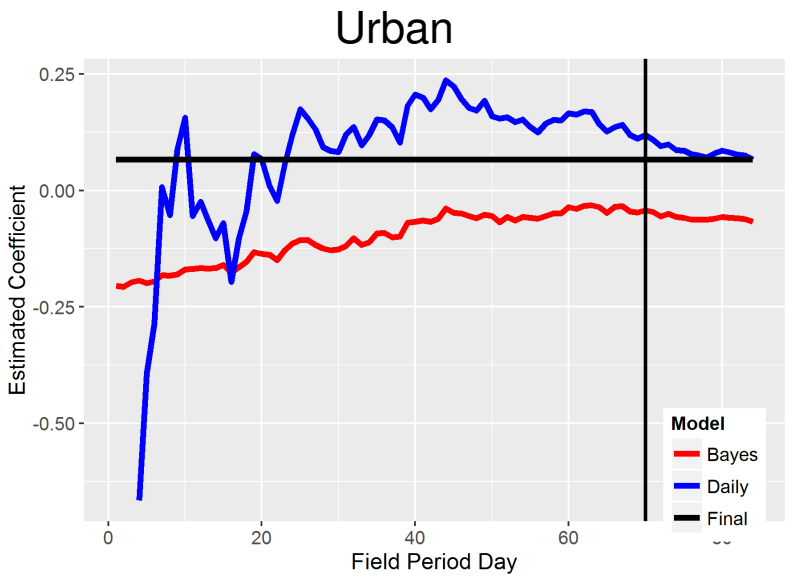
Table: Daily

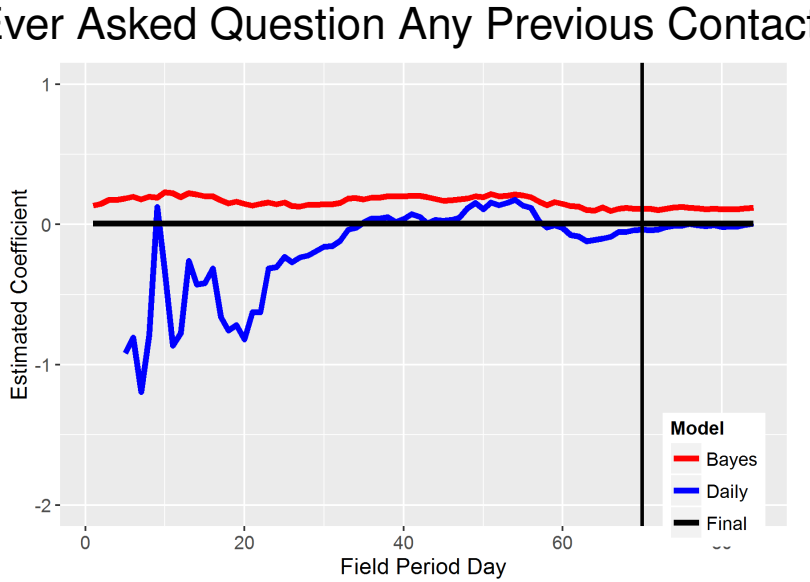
	Daily				
Final	0-20	20-40	40-60	60-80	80-100
0-20	163	32	20	4	2
20-40	57	111	51	0	0
40-60	8	64	114	30	4
60-80	1	4	32	108	75
80-100	0	0	4	78	138

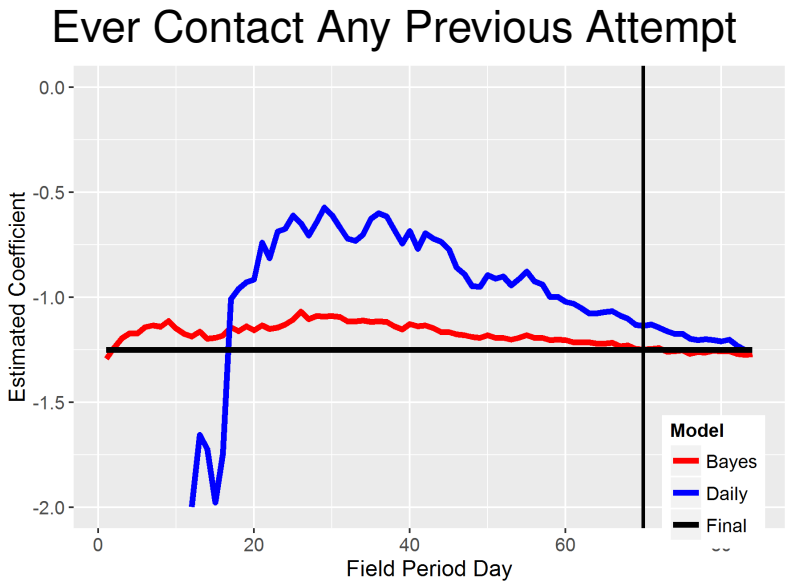
	κ	ASE	z	Pr(>|z|)
Unweighted	0.470	0.019	25.28	5.089e-141
Weighted	0.704	0.012	57.18	0.000e+00

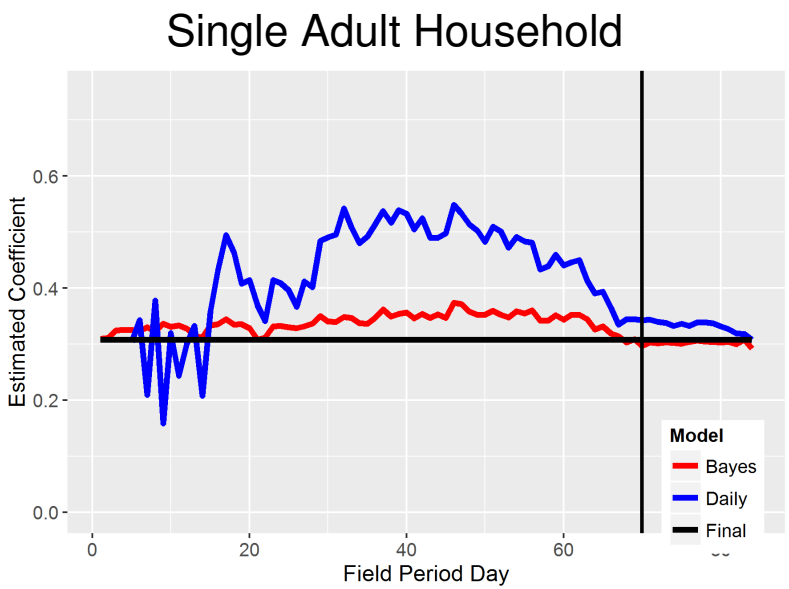
The agreement rate is is 63.4% for the Bayesian model predictions and 57.6% for the Daily model.

Estimated Coefficients Related to Misclassification









Conclusion

Lessons Learned

1. Classification is improved with prior information
2. Prior information more valuable early on
3. Where possible, tuning prior information for the setting may be valuable

Thank you!

email: jameswag@umich.edu
blog: http://jameswagnersurv.blogspot.com/