Alternative Modeling Approaches to Predicting Costs in a Responsive Design Framework

James Wagner[1,2], Michael Elliott[1,2,3], Brady T. West[1,2], Stephanie Coffey[2,3]

[1] Survey Research Center, Institute for Social Research, Univ. of MI-Ann Arbor
[2] Joint Program in Survey Methodology, Univ. of MD-College Park
[3] U.S. Census Bureau

Acknowledgements / Disclaimers

2

Overview

- Background
  - Responsive Survey Design
  - Monitoring incoming data from the field
- Problem
  - Need predictions of costs under alternative designs
- NSFG
  - Current design
  - Data
- Methods
  - Multilevel Regression Models
  - Bayesian Additive Regression Trees (BART)
- Results
- Conclusions

Responsive Survey Design (RSD)

- *Uncertainty* has become an issue in survey design
- **RSD makes use of incoming data** from the field to address this uncertainty

  – *Groves and Heeringa (2006)*

- Develop **indicators of cost and error**
- **Planned interventions** when costs increase or errors stabilize/increase

4

Problem

- Costs and errors vary over time
- RSD has used proxy indicators
  - Costs: call attempts
  - Nonresponse error: stabilized estimates, "phase capacity"
- Inaccurate indicators may lead to inefficient designs
- ***Can we improve the accuracy of cost predictions?***

NSFG

- National Survey of Family Growth 2011-2019 (NSFG)
- Continuous data collection
  - A new sample is released every **quarter**
- Two-stage data collection:
  - <u>Screener</u> interview to identify eligible persons
  - <u>Main</u> interview of selected person
- Two phases of data collection:
  - **Phase 1:** 10 week data collection
  - **Phase 2:** Subsample remaining cases
    - Oversample higher likelihood of interview cases and eligible/likely eligible cases
    - Reduce interviewer workload by 2/3
    - Change data collection model: increased interview token of appreciation, interviewer behavior change
  - Combine data and response rates from two phases using weights

NSFG: Data

- We use the data available at the time the predictions would be made
  - Paradata: highly correlated with interviewer hours *(Wagner, 2019)*
  - Paradata and other characteristics of the sample for the **future time periods** we are predicting are **not available** at time of prediction

- Use **lagged values**: the values from two weeks prior to the time period being predicted

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
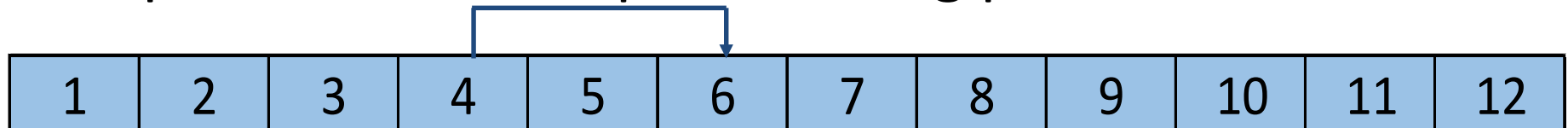
NSFG: Data

- Data include:
  - **Interviewer ID**
  - **Phase** (i.e. the design change)
  - **Lagged values of the following:**
    - **Area characteristics**: Census Division, Population eligibility rate, urbanicity, etc.
    - **Interviewer observations**: access problems, safety concerns, etc.
    - **Commercial data**: Age of first person, etc.
    - **Paradata**: Number of screened interviews, number of trips, number of active lines, etc.

8

NSFG: Data

- Variables are **summarized to the interviewer-week level**
  - Call attempts have sample characteristics, interviewer observations, etc.
  - Categorical variables: Mode
    - EXAMPLE: **Modal urbanicity from cases that were attempted two weeks prior to the week being predicted**
  - Continuous variables: Mean
    - EXAMPLE: Mean population eligibility rate from cases that were attempted two weeks prior
  - Paradata: Sums
    - EXAMPLE: **Number of completed screening interviews two weeks prior to the week being predicted**
    - EXAMPLE: Number of active sampled units ("lines") two weeks prior

9

NSFG: Data

- Use data from previous quarters and current quarter phase one to predict phase two costs
  - Predictions for Q22-Q27
- Outcome variable: **phase two interviewer hours**
  - This is the major cost driver of the phase
- Secondary costs: incentives
  - Not predicted in these models
  - Some are prepaid, i.e. known
  - Postpaid incentive more readily predicted from propensity models

Method

- Three methods of prediction:
  - **1) Multilevel regression models**
    - Random intercept for each interviewer
    - All covariates described earlier included in the model
  - **2) Bayesian Additive Regression Trees** (BART, *Chipman, et al., 2010*)
    - Sum of trees method
    - Priors constrain the inclusion of predictors
    - Possible to examine how frequently each predictor is included
    - Treatment of interviewers as: **a) Random Intercept vs b) Fixed Effect**

Method

- Compare:
  1. **Total** **Predicted vs observed** hours in phase 2
  2. **Interviewer-level** **predicted vs observed** hours in phase 2
- Measures of accuracy

  1. Mean squared error: $MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$

  2. Mean absolute error: $MAE = \frac{\sum_{i=1}^{n}|Y_i - \hat{Y}_i|}{n}$

  - *i* indexes interviewer-week, *Y* is the observed hours worked, $\hat{Y}$ is the predicted hours, *n* is the number of interviewer-weeks

Results: Interviewers are consistent

- <u>MLM models:</u>
  ICC: 0.21-0.25

- <u>Fixed Effect</u>
  <u>BART models</u>: 20
  most frequently
  included
  variables
  - 8 are
    interviewer IDs
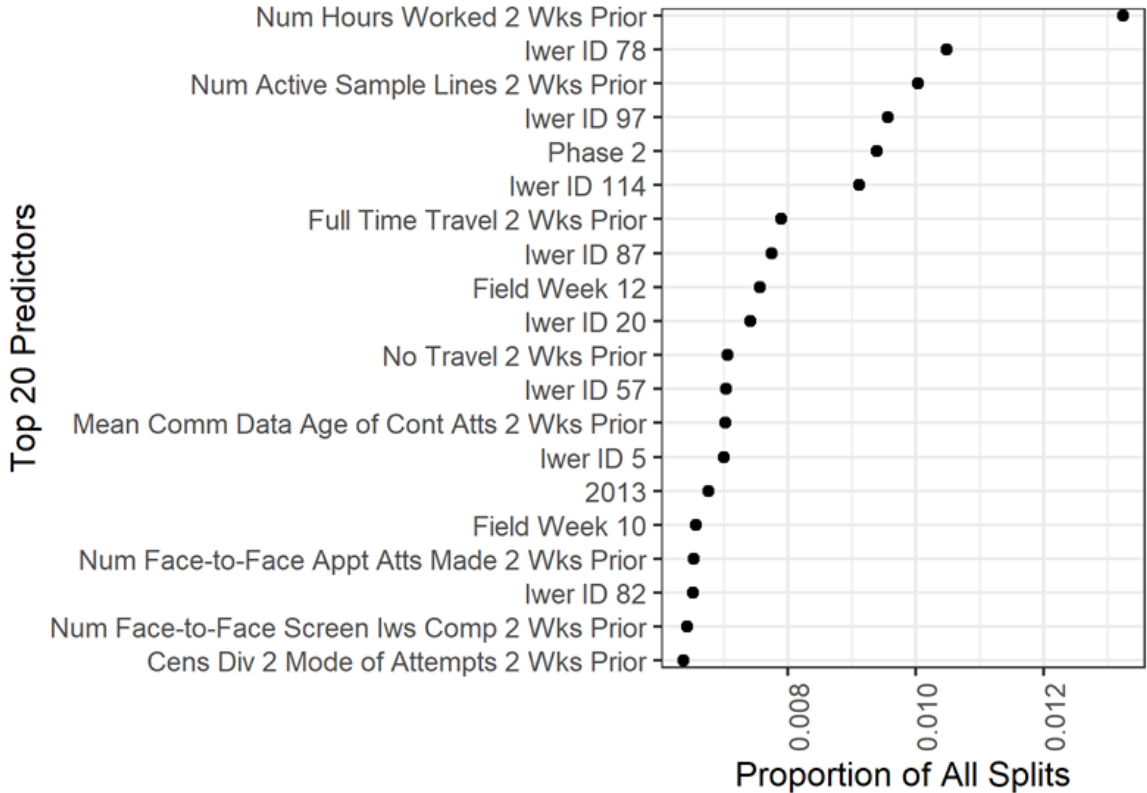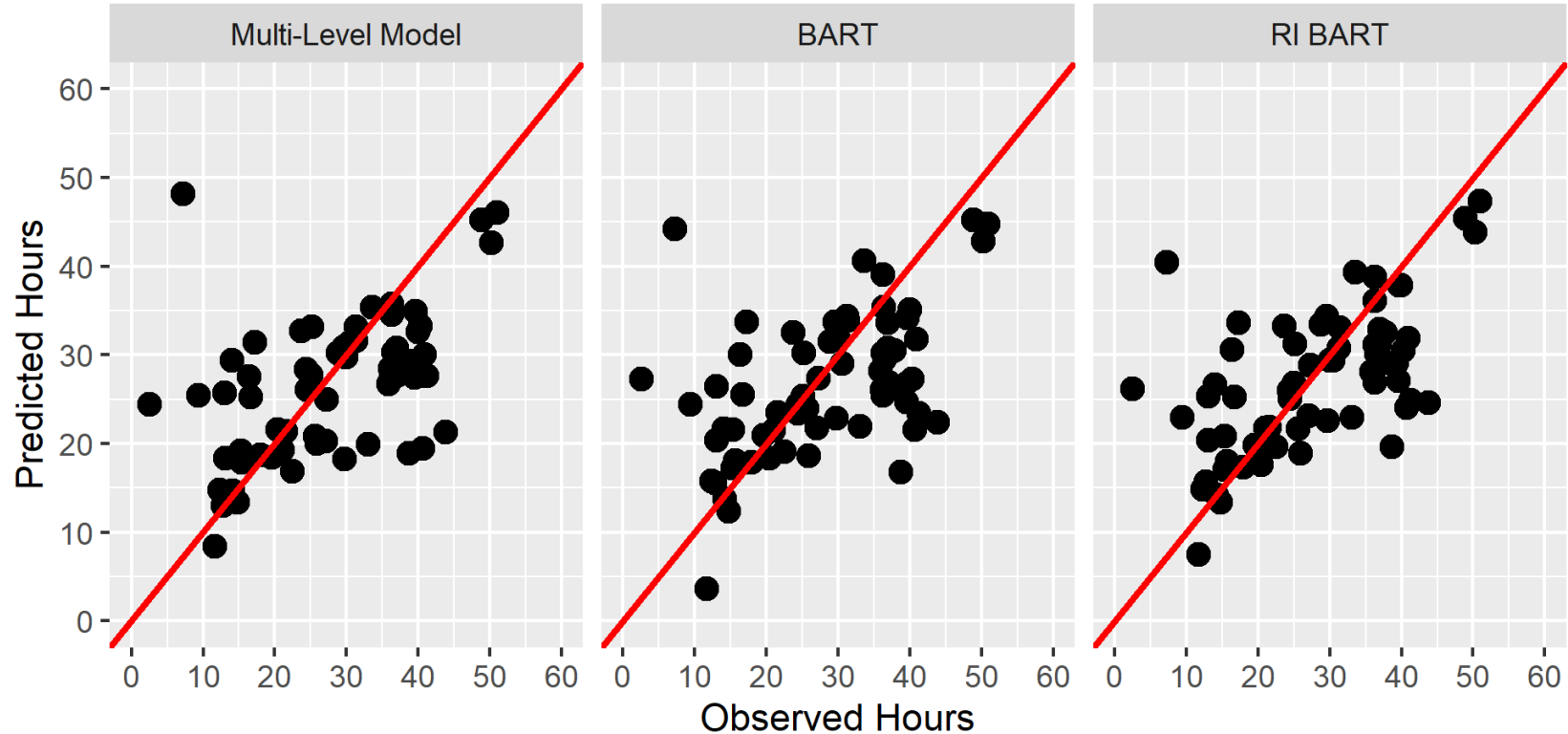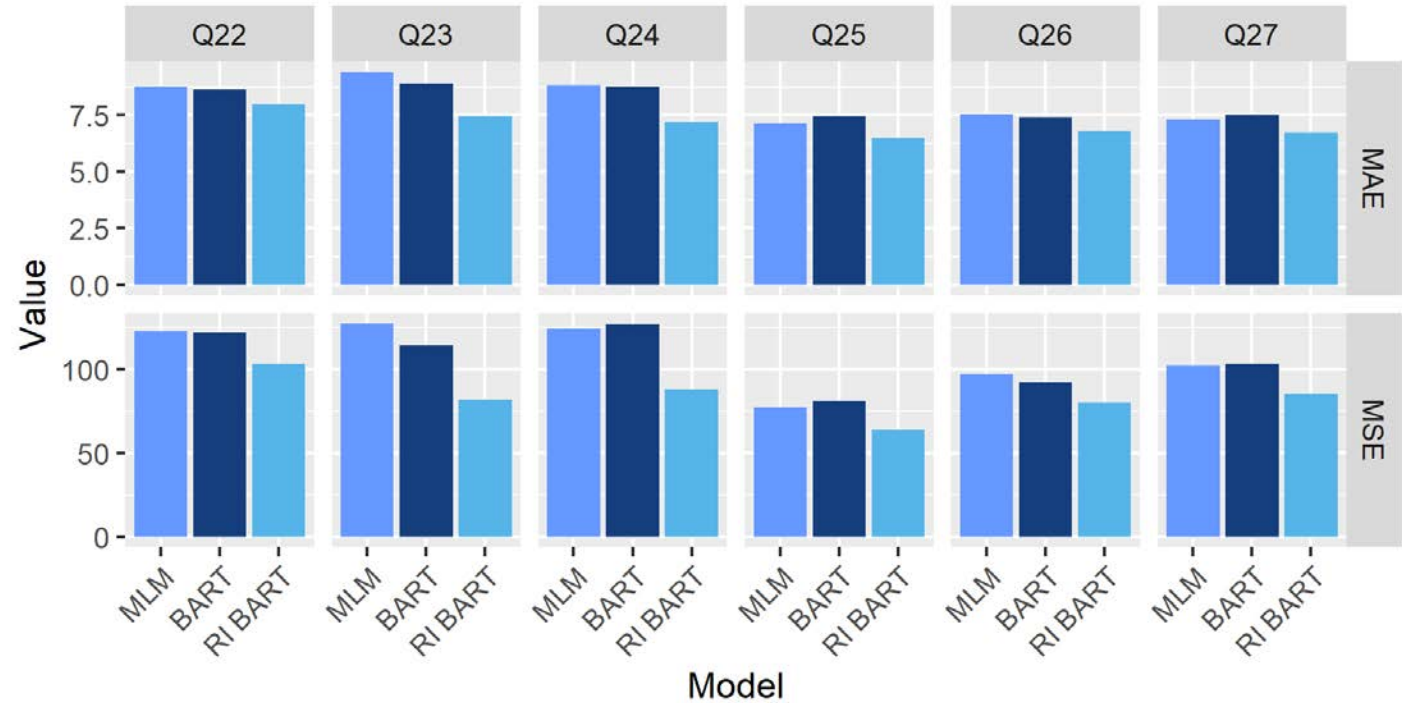
Conclusions
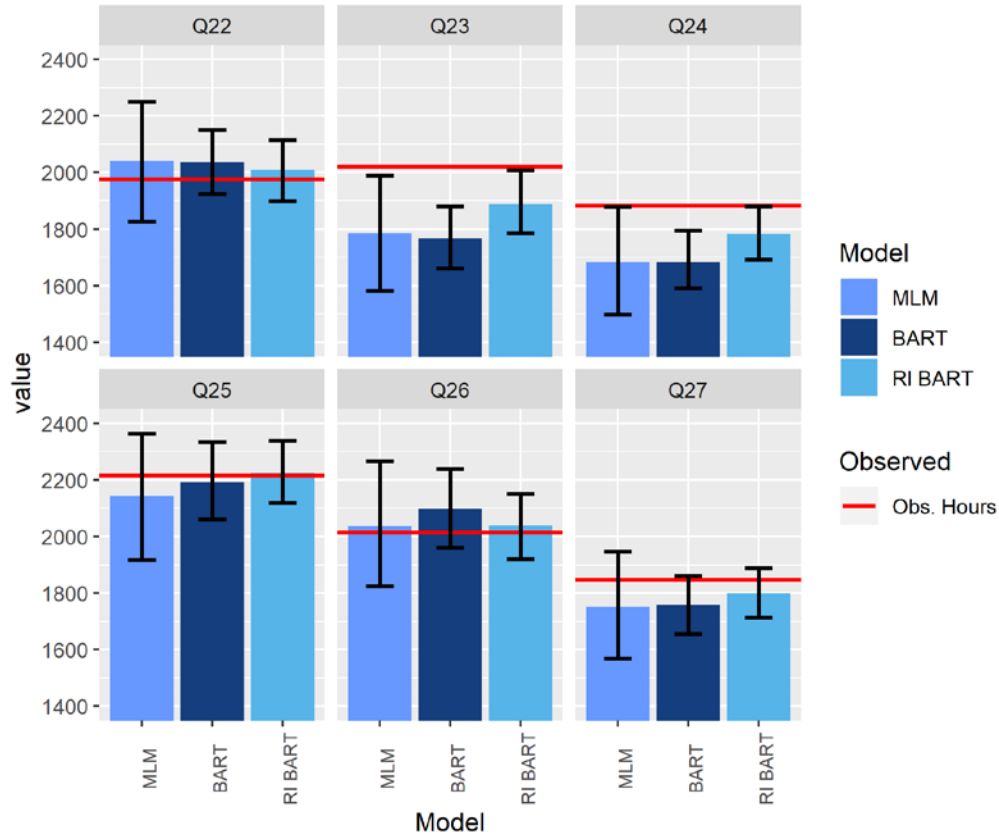
- Existing methods can be used to predict survey costs

- In this problem, **knowing the interviewer is very useful since their behavior is consistent**

- For this problem, **BART models improve predictive accuracy**
  - Including interviewers as a **random intercept** in the BART model produces the most accurate predictions

Thank You!

- See these results in a forthcoming *Journal of Official Statistics* paper

- Email: jameswag@umich.edu

References

- Chipman, H. A., E. I. George and R. E. McCulloch (2010). "BART: Bayesian additive regression trees." <u>The Annals of Applied Statistics</u> **4(1): 266-298.**

- Groves, R. M. and S. G. Heeringa (2006). "Responsive design for household surveys: tools for actively controlling survey errors and costs." <u>Journal of the Royal Statistical Society: Series A (Statistics in Society)</u> **169(3): 439-457.**

- Wagner, J. (2019). "Estimation of Survey Cost Parameters Using Paradata." <u>Survey Practice</u> **12(1): 1-10.**

19

Data

Source	Predictor	Description
	NEWIWERID	Interviewer ID Number
TIMESHEETS	NHOURS_LAG2	Number of hours worked by the interviewer in the week two weeks prior to the current week
	TRAVEL_LAG2	How much did the interviewer participate in overnight travel in the week two weeks prior to the current week: NONE, SOME, or ALL week.
	DAYS_WORKED_LAG2	The number of days worked (i.e. days with an entry in the timesheet) in the week two weeks prior to the current week
SAMPLING FRAME	QTR	The quarter of production (Q1-Q27)
	YEAR	The calendar year of production (2011-2018)
	CENSUS_DIV_MODE_LAG2	The modal Census Division of the lines attempted by an interviewer in the week two weeks prior to the current week
	CENS_REG_MODE _LAG2	The modal Census Region of the lines attempted by an interviewer in the week two weeks prior to the current week
	EST_ELIG_RATE_MEAN_LAG2	This is the mean of the Census Block Group level data about the estimated eligibility rate. The data are at the Block Group level, but the value here is the average over all contact attempts for the week that is two weeks prior to the current week.
	EST_ELIG_15_49_ACS_MEAN_ LAG2	The mean of the estimated eligibility rate for the Census Block Group reported in the American Community Survey. The data are at the Block Group level, but the value here is the average over all contact attempts for the week that is two weeks prior to the current week.
	ELIG_NEVER_PCT_MEAN_LAG 2	This is the percentage of eligible persons living in the Census Tract who have never been married. The data are at the Tract level, but the value here is the average over all contact attempts for the week that is two weeks prior to the current week.
	OCC_RATE_MEAN_LAG2	This is the Census Block Group level occupancy rate. The data are at the Block Group level, but the value here is the average over all contact attempts for the week that is two weeks prior to the current week.
	DOMAIN_MODE_LAG2	The domain is set at the Census Block Group (BG) level and assigned to housing units within each BG. All BGs are assigned to a domain based upon the following definitions: 1) <10% of Block Group African-American and <10% Hispanic, 2) >=10% of Block Group African-American and <10% Hispanic, 3) <10% of Block Group African-American and >=10% Hispanic, and 4) >=10% of Block Group African-American and >=10% Hispanic. The mode is for the domain of the lines that are attempted in the week two weeks prior to the current week
	URBAN_MODE_LAG2	The mode of the urbanicity (assigned at the case level) of the attempts made during the week that is two weeks prior to the current week, where 1=Major Metropolitan Area, 2=Minor Metropolitan Area, 3=Non-Metropolitan Area, 4=Remote Area.

20

Source	Predictor	Description
INTERVIEWER OBSERVATIONS	STRUCTURE_TYPE_MODE _LAG2	The mode of the structure type variable of the cases that were attempted in the week that is two weeks prior to the current week . 1=Single family home, 2=Structure with 2 to 9 units, 3=Structure with 10+ units,4= Mobile home, 5=Other.
	BLACCESS_GATED_MEAN_LAG2	The mean of an area segment-level observation about whether there is a gated community in the area segment. This is observed at the segment level but the value here is average over all contact attempts for the week that is two weeks prior to the current week.
	BLACCESS_SEASONAL_HAZARD_ MEAN_LAG2	The mean of an area segment-level observation about whether there is a potential seasonal hazard preventing access to the area segment (e.g. unplowed roads). This is observed at the segment level but the value here is average over all contact attempts for the week that is two weeks prior to the current week.
	BLACCESS_UNIMPROVED_ROADS _MEAN_LAG2	The mean of an area segment-level observation about whether there are unimproved roads limiting access to the area segment. This is observed at the segment level but the value here is average over all contact attempts for the week that is two weeks prior to the current week.
	BLACCESS_OTHER_MEAN	The mean of an area segment-level observation about whether there other (i.e. not gated, seasonal hazards, or unimproved roads) factors limiting access to the area segment. This is observed at the segment level but the value here is average over all contact attempts for the week that is two weeks prior to the current week.
	LRESIDENTIAL_MEAN	The mean of an area segment-level observation about whether the area is completely residential or also includes some commercial structures. This is observed at the segment level but the value here is average over all contact attempts for the week that is two weeks prior to the current week.
	INON_ENGLISH_SPEAKERS_MEA N_LAG2	The mean of an area segment-level observation about whether the area has evidence of non-English speakers. This is observed at the segment level but the value here is average over all contact attempts for the week that is two weeks prior to the current week
	BLNON_ENGLISH_LANG_SPANIS_ MEAN_LAG2	The mean of an area segment-level observation about whether the area has evidence of Spanish speakers. This is observed at the segment level but the value here is average over all contact attempts for the week that is two weeks prior to the current week.
	ISAFETY_CONCERNS_MEAN_LAG 2	The mean of an area segment-level observation about whether the interviewer had concerns about their safety on the first visit. This is observed at the segment level but the value here is average over all contact attempts for the week that is two weeks prior to the current week.
	MANYUNITS_MEAN_LAG2	The mean of an observation at the housing unit level indicating whether the sampled housing unit has 1=more than one unit, or 0=1 unit. This is observed at the housing unit level but the value here is average over all contact attempts for the week that is two weeks prior to the current week.
	CHILDRENUNDER15_MEAN_LAG 2	The mean of an observation at the housing unit level indicating whether the interviewer believes that there are children under the age of 15 living in the housing unit (1=Yes, 0=No). This is observed at the housing unit level but the value here is average over all contact attempts for the week that is two weeks prior to the current week.
	ALLAGEOVER45_MEAN_LAG2	The mean of an observation at the housing unit level indicating whether the interviewer believes that persons living in the housing unit are all over the age of 45 (1=Yes, 0=No). This is observed at the housing unit level but the value here is average over all contact attempts for the week that is two weeks prior to the current week.

Source	Predictor	Description
COMMERCIAL DATA	MSG_MATCHQUALITY_MEAN_LAG2	A variable indicating the estimated quality of the match of commercially-available data to the address (1-5). The data are at the case level, but the value here is the average over all contact attempts for the week that is two weeks prior to the current week.
	MSG_AGE_MEAN_LAG2	The mean age of the first person from the commercially-available data where those data are available. The data are at the case level, but the value here is the average over all contact attempts for the week that is two weeks prior to the current week
	MSG_INCOME_MEAN_LAG2	The mean of the estimated household income for cases with a match to commercially-available data. The data are at the case level, but the value here is the average over all contact attempts for the week that is two weeks prior to the current week
LEVEL OF EFFORT PARADATA	PHA**SE**	The phase of the NSFG design (first phase occurs in weeks 1-10, phase 2 during weeks 11-12).
	LAG2.ACTIVE_LINES	The number of active lines from 2 weeks prior to the current week for each interviewer.
	TRIPS_LAG2	The total number of unique visits to an area segment (derived from call record data) from 2 weeks prior to the current week for each interviewer.
	FTFNOCONTACT_LAG2	The total number of Face-to-face contact attempts that resulted in no contact from 2 weeks prior to the current week for each interviewer.
	FTFCONTACT_LAG2	The total number of Face-to-face contact attempts that resulted in a contact with only agreement for a general callback from 2 weeks prior to the current week for each interviewer.
	FTFAPPT_LAG2	The total number of Face-to-face contact attempts that resulted in setting an appointment from 2 weeks prior to the current week for each interviewer.
	MAINIW_LAG2	The total number of main interviews (all main interviews are completed face-to-face) from 2 weeks prior to the current week for each interviewer.
	FTFMAINRESIST_LAG2	The total number of Face-to-face contact attempts that resulted in the sampled person expressing concerns from 2 weeks prior to the current week for each interviewer.
	FTFMAINNI_LAG2	The total number of Face-to-face contact attempts that resulted in a final noninterview from 2 weeks prior to the current week for each interviewer.
	FTFMAINNS_LAG2	The total number of Face-to-face contact attempts that resulted in a final nonsample from 2 weeks prior to the current week for each interviewer.
	FTFSCRNIW_LAG2	The total number of Face-to-face contact attempts that resulted in a screening interview from 2 weeks prior to the current week for each interviewer.
	FTFSCRNRESIST_LAG2	The total number of Face-to-face contact attempts that resulted in the sampled housing unit expressing concerns prior to completing a screening interview from 2 weeks prior to the current week for each interviewer.
	FTFSCRNNI_LAG2	The total number of Face-to-face contact attempts that resulted in the sampled housing unit being finalized as a noninterview prior to completing a screening interview from 2 weeks prior to the current week for each interviewer.
	FTFSCRNNS_LAG2	The total number of Face-to-face contact attempts that resulted in the sampled housing unit being finalized as nonsample prior to completing a screening interview from 2 weeks prior to the current week for each interviewer.
	FTF_MAINNS_INEL_LAG2	The total number of Face-to-face contact attempts that resulted in the sampled person being finalized as ineligible prior to completing a screening interview from 2 weeks prior to the current week for each interviewer.
	ACTIVE_LINES_LAG2	The number of active sampled units two weeks prior to the current week for each interviewer.
	TEL_ALL_LAG2	The total number of telephone attempts made by each interviewer two weeks prior to the current week.

Predictive Accuracy – Interviewers (2)

		MSE	MAE
Q22	BART Model	121.75	8.64
	Multi-Level Model	122.71	8.75
	RI BART Model	**102.98**	**7.97**
Q23	BART Model	114.26	8.87
	Multi-Level Model	127.14	9.40
	RI BART Model	**82.04**	**7.45**
Q24	BART Model	126.89	8.74
	Multi-Level Model	123.83	8.82
	RI BART Model	**87.91**	**7.18**
Q25	BART Model	80.94	7.42
	Multi-Level Model	77.03	7.11
	RI BART Model	**63.93**	**6.48**
Q26	BART Model	91.90	7.41
	Multi-Level Model	97.36	7.52
	RI BART Model	**80.09**	**6.77**
Q27	BART Model	103.16	7.49
	Multi-Level Model	102.44	7.31
	RI BART Model	**84.97**	**6.73**