Optimizing Data Collection Interventions to Balance Cost and Quality Under a Bayesian Framework

Stephanie Coffey
Center for Optimization and Data Science

Presentation at the Michigan Program in Survey Methodology Seminar

March 31, 2021

Background and Motivation

- Increased interest in alternative data collection designs
  - Responsive, Adaptive, Tailored, Targeted Designs
- Apply different data collection features to sample cases
  - Made in pursuit of some data collection goal
  - Survey data collection parameters (SDCPs)
    - Response propensity
    - Costs
    - Survey item response
- Need high quality predictions of SDCPs to make optimal decisions

Statement of Problem

- Responsive and Adaptive Survey Designs
  - Interventions made during data collection
  - Rely on historical data for a survey?
  - Rely on accumulating data?
- Using only data from current round can lead to biased predictions
  - Wagner and Hubbard (2014)[1]

Need a method that combines
external data and current accumulating data
in order to improve predictions of SDCPs

Bayesian Framework for Prediction

- Bayesian methods are a natural solution
  - Systematic way to combine external data with current accumulating data
  - Obtain posterior distributions of coefficients in predictive models of interest:

$$pos(\theta_1, \dots, \theta_n) \propto p(\theta_1, \dots, \theta_n) \prod_i p(y_i | \theta_1, \dots, \theta_n)$$

  - Select k samples from posterior distribution of each coefficient
  - Generate k case-level predictions of an SDCP and average over k predictions

- Recent research on Bayesian methods to improve prediction of SDCPs
  - Schouten et al. (2018)[2] - contact and cooperation propensities
  - West et al. (under review)[3] - response propensity
  - Wagner et al. (2020)[4] - data collection costs
  - Coffey et al. (2020)[5] - response propensity via expert elicitation

Making Interventions Based on SDCPs

- Different data collection features have different properties
- Ideally, survey managers would know characteristics like…
  - *if* a sample member will respond - response propensity
  - *resources needed* to obtain that response - cost
  - *information* a sample member will provide - survey item response
- Schouten et al. (2018) discusses pre-data collection allocation
- Reallocation during data collection
  - Can leverage historical and current accumulating data – better predictions
- Conduct experiment in the National Survey of College Graduates

National Survey of College Graduates

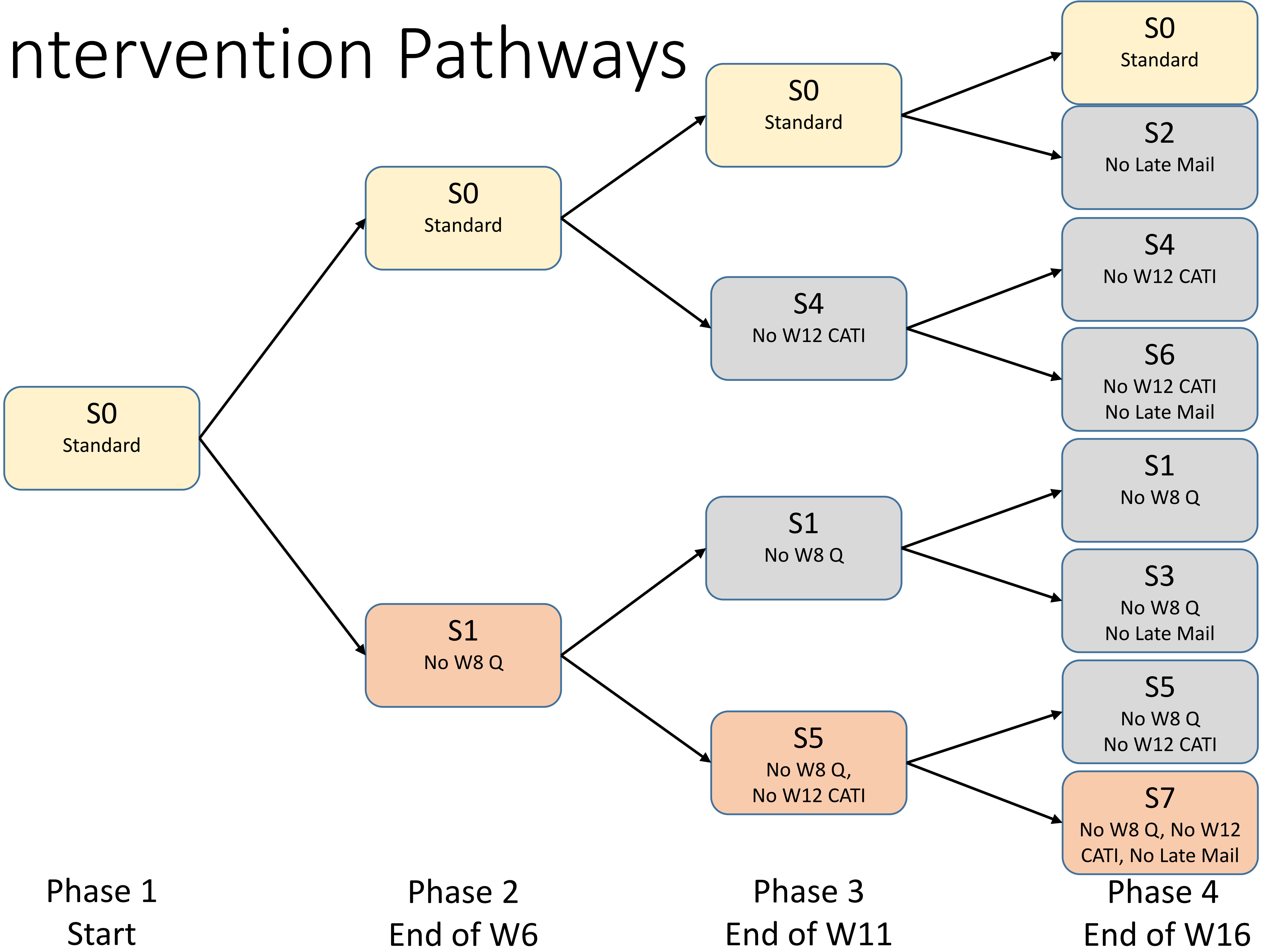
- Sponsored by the National Center for Science and Engineering Statistics within the National Science Foundation
- Conducted by the Census Bureau every 2 years
- Targets college-educated individuals in the US
- Sampled out of the American Community Survey
- Data Collection
  - Six-months
  - Sequential Modes (web, paper, CATI)

National Survey of College Graduates

Phase	Primary Modes	Weeks	Days
1. Web Push Phase	Web	0 – 7	-6 – 49
2. Mail Questionnaire Phase	Web, Mail	8 - 11	50 – 77
3. Telephone Follow-up Phase	Web, Mail, CATI	12 – 17	78 – 119
4. Late Follow-Up Phase	Web, Mail, CATI	18 – 26	120 – 182

- Mix of modes is used to reduce nonresponse error
- Costs of later mode strategies are higher than web self-response
- Costs may not be worth it if sample case
  - Unlikely to respond in more expensive modes
  - Does not contribute information to the survey estimates
- What are the alternate (less costly)?
- How do we identify cases for those strategies?

Intervention Pathways



Phase
1. Web Push Phase
2. Mail Questionnaire Phase
3. Telephone Follow-up Phase
4. Late Follow-Up Phase

i^{th} case
d^{th} day
v covariates
ϵ error
$\hat{C}^R$ cost of response
$\hat{C}^{NR}$ cost of nonresponse

Model Descriptions

- Response propensity (Bayesian Estimates of $\hat{\beta}_v$):

$$\hat{p}_{id} = \hat{p}(y_{id} = 1|X_{id}) = \frac{\exp\left(\sum_{v=0}^{V} \hat{\beta}_v x_{idv}\right)}{1 + \exp\left(\sum_{v=0}^{V} \hat{\beta}_v x_{idv}\right)}$$

- Value of self-reported salary (Bayesian Estimates of $\hat{\beta}_v$):

$$(y_i)^{1/3} = \sum_{v=0}^{V} \hat{\beta}_v x_{idv} + \epsilon_{id}$$

- Cost of response (Estimated from Prior Data Ignoring Error):

$$E(C_i) = \hat{p}_{id}\left(\hat{C}_{id}^R\right) + (1 - \hat{p}_{id})\left(\hat{C}_{id}^N\right)$$

Responsive Design Experiment

- Reduce data collection costs without hurting data quality
- “minimize cost for a small increase in RMSE”
  - Allocate “less impactful” cases to lower cost data collection strategies
  - RMSE of salary – key survey estimate in the NSCG
- Design:
  - Systematic random sample (n=8,000) with cluster size of 2
  - Control group managed with production operational methodology
  - Treatment group managed using responsive design decisions
- Evaluation:
  - Compare actual costs, mean(salary), RMSE(salary), response rates

Optimization Steps

- At each intervention point
- Use priors from historical data + currently accumulating data
- Predict (for nonrespondents)
  - Value of response variable, salary
  - Response propensity under different strategies
  - Cost of different strategies
- Allocate sets of cases to new (cheaper) strategy
- Examine effect on RMSE(salary) and costs
- Determine which cases to allocate to new strategy

Case	Resp Stat	Resp Val	Accrued Costs
390	1	y_i	c_i^p
194	1	y_i	c_i^p
280	1	y_i	c_i^p
227	1	y_i	c_i^p
798	0	--	c_i^p
578	0	--	c_i^p
638	0	--	c_i^p
742	0	--	c_i^p

Case	Resp Stat	Resp Val	Accrued Costs	Impute RVal
390	1	y_i	c_i^p	y_i
194	1	y_i	c_i^p	y_i
280	1	y_i	c_i^p	y_i
227	1	y_i	c_i^p	y_i
798	0	--	c_i^p	$\hat{y}_i$
578	0	--	c_i^p	$\hat{y}_i$
638	0	--	c_i^p	$\hat{y}_i$
742	0	--	c_i^p	$\hat{y}_i$

Case	Resp Stat	Resp Val	Accrued Costs	Impute RVal
390	1	y_i	c_i^p	y_i
194	1	y_i	c_i^p	y_i
280	1	y_i	c_i^p	y_i
227	1	y_i	c_i^p	y_i
798	0	--	c_i^p	$\hat{y}_i$
578	0	--	c_i^p	$\hat{y}_i$
638	0	--	c_i^p	$\hat{y}_i$
742	0	--	c_i^p	$\hat{y}_i$

$\hat{\hat{y}}^T$

Case	Resp Stat	Resp Val	Accrued Costs	Impute RVal	Dist ($\hat{y}_i$-$\hat{\bar{y}}$)	Dist Rank
390	1	y_i	c_i^p	y_i		
194	1	y_i	c_i^p	y_i		
280	1	y_i	c_i^p	y_i		
227	1	y_i	c_i^p	y_i		
798	0	--	c_i^p	$\hat{y}_i$	$\hat{d}_i$	1
578	0	--	c_i^p	$\hat{y}_i$	$\hat{d}_i$	2
638	0	--	c_i^p	$\hat{y}_i$	$\hat{d}_i$	3
742	0	--	c_i^p	$\hat{y}_i$	$\hat{d}_i$	4

$\hat{\bar{y}}^{T}$

<table>
<tr><th>Case</th><th>Resp Stat</th><th>Resp Val</th><th>Accrued Costs</th><th>Impute RVal</th><th>Dist ($\hat{y}_i$-$\hat{\bar{y}}$)</th><th>Dist Rank</th><th>Next Strat</th></tr>
<tr><td>390</td><td>1</td><td>y_i</td><td>c_i^p</td><td>y_i</td><td></td><td></td><td></td></tr>
<tr><td>194</td><td>1</td><td>y_i</td><td>c_i^p</td><td>y_i</td><td></td><td></td><td></td></tr>
<tr><td>280</td><td>1</td><td>y_i</td><td>c_i^p</td><td>y_i</td><td></td><td></td><td></td></tr>
<tr><td>227</td><td>1</td><td>y_i</td><td>c_i^p</td><td>y_i</td><td></td><td></td><td></td></tr>
<tr><td rowspan="2">798</td><td rowspan="2">0</td><td rowspan="2">--</td><td rowspan="2">c_i^p</td><td rowspan="2">$\hat{y}_i$</td><td rowspan="2">$\hat{d}_i$</td><td rowspan="2">1</td><td>0</td></tr>
<tr><td>1</td></tr>
<tr><td rowspan="2">578</td><td rowspan="2">0</td><td rowspan="2">--</td><td rowspan="2">c_i^p</td><td rowspan="2">$\hat{y}_i$</td><td rowspan="2">$\hat{d}_i$</td><td rowspan="2">2</td><td>0</td></tr>
<tr><td>1</td></tr>
<tr><td rowspan="2">638</td><td rowspan="2">0</td><td rowspan="2">--</td><td rowspan="2">c_i^p</td><td rowspan="2">$\hat{y}_i$</td><td rowspan="2">$\hat{d}_i$</td><td rowspan="2">3</td><td>0</td></tr>
<tr><td>1</td></tr>
<tr><td rowspan="2">742</td><td rowspan="2">0</td><td rowspan="2">--</td><td rowspan="2">c_i^p</td><td rowspan="2">$\hat{y}_i$</td><td rowspan="2">$\hat{d}_i$</td><td rowspan="2">4</td><td>0</td></tr>
<tr><td>1</td></tr>
</table>

$$\hat{\bar{y}}^T$$

Case	Resp Stat	Resp Val	Accrued Costs	Impute RVal	Dist ($\hat{y}_i$-$\hat{\bar{y}}$)	Dist Rank	Next Strat	Future Costs	Resp Prop
390	1	y_i	c_i^p	y_i					
194	1	y_i	c_i^p	y_i					
280	1	y_i	c_i^p	y_i					
227	1	y_i	c_i^p	y_i					
798	0	--	c_i^p	$\hat{y}_i$	$\hat{d}_i$	1	0	$\hat{c}_i{}^{S0}$	$\hat{\rho}_i{}^{S0}$
							1	$\hat{c}_i{}^{S1}$	$\hat{\rho}_i{}^{S1}$
578	0	--	c_i^p	$\hat{y}_i$	$\hat{d}_i$	2	0	$\hat{c}_i{}^{S0}$	$\hat{\rho}_i{}^{S0}$
							1	$\hat{c}_i{}^{S1}$	$\hat{\rho}_i{}^{S1}$
638	0	--	c_i^p	$\hat{y}_i$	$\hat{d}_i$	3	0	$\hat{c}_i{}^{S0}$	$\hat{\rho}_i{}^{S0}$
							1	$\hat{c}_i{}^{S1}$	$\hat{\rho}_i{}^{S1}$
742	0	--	c_i^p	$\hat{y}_i$	$\hat{d}_i$	4	0	$\hat{c}_i{}^{S0}$	$\hat{\rho}_i{}^{S0}$
							1	$\hat{c}_i{}^{S1}$	$\hat{\rho}_i{}^{S1}$

$\hat{\bar{y}}^{T}$

Case	Resp Stat	Resp Val	Accrued Costs	Impute RVal	Dist ($\hat{y}_i$-$\hat{\hat{y}}$)	Dist Rank	Next Strat	Future Costs	Resp Prop	Resp Class
390	1	y_i	c_i^p	y_i						
194	1	y_i	c_i^p	y_i						
280	1	y_i	c_i^p	y_i						
227	1	y_i	c_i^p	y_i						
798	0	--	c_i^p	$\hat{y}_i$	$\hat{d}_i$	1	0	$\hat{c}_i{}^{S0}$	$\hat{\rho}_i{}^{S0}$	1
							1	$\hat{c}_i{}^{S1}$	$\hat{\rho}_i{}^{S1}$	1
578	0	--	c_i^p	$\hat{y}_i$	$\hat{d}_i$	2	0	$\hat{c}_i{}^{S0}$	$\hat{\rho}_i{}^{S0}$	0
							1	$\hat{c}_i{}^{S1}$	$\hat{\rho}_i{}^{S1}$	1
638	0	--	c_i^p	$\hat{y}_i$	$\hat{d}_i$	3	0	$\hat{c}_i{}^{S0}$	$\hat{\rho}_i{}^{S0}$	1
							1	$\hat{c}_i{}^{S1}$	$\hat{\rho}_i{}^{S1}$	0
742	0	--	c_i^p	$\hat{y}_i$	$\hat{d}_i$	4	0	$\hat{c}_i{}^{S0}$	$\hat{\rho}_i{}^{S0}$	0
							1	$\hat{c}_i{}^{S1}$	$\hat{\rho}_i{}^{S1}$	0

$\hat{\hat{y}}^T$

Case	Resp Stat	Resp Val	Accrued Costs	Impute RVal	Dist $(\hat{y}_i - \hat{\bar{y}})$	Dist Rank	Next Strat	Future Costs	Resp Prop	Resp Class	Allocation 0% RVal	Allocation 0% Cost
390	1	y_i	c_i^p	y_i							y_i	c_i^p
194	1	y_i	c_i^p	y_i							y_i	c_i^p
280	1	y_i	c_i^p	y_i							y_i	c_i^p
227	1	y_i	c_i^p	y_i							y_i	c_i^p
798	0	--	c_i^p	$\hat{y}_i$	$\hat{d}_i$	1	0	$\hat{c}_i^{S0}$	$\hat{\rho}_i^{S0}$	1	$\hat{y}_i$	$c_i^p + \hat{c}_i^{S0}$
							1	$\hat{c}_i^{S1}$	$\hat{\rho}_i^{S1}$	1		
578	0	--	c_i^p	$\hat{y}_i$	$\hat{d}_i$	2	0	$\hat{c}_i^{S0}$	$\hat{\rho}_i^{S0}$	0	--	$c_i^p + \hat{c}_i^{S0}$
							1	$\hat{c}_i^{S1}$	$\hat{\rho}_i^{S1}$	1		
638	0	--	c_i^p	$\hat{y}_i$	$\hat{d}_i$	3	0	$\hat{c}_i^{S0}$	$\hat{\rho}_i^{S0}$	1	$\hat{y}_i$	$c_i^p + \hat{c}_i^{S0}$
							1	$\hat{c}_i^{S1}$	$\hat{\rho}_i^{S1}$	0		
742	0	--	c_i^p	$\hat{y}_i$	$\hat{d}_i$	4	0	$\hat{c}_i^{S0}$	$\hat{\rho}_i^{S0}$	0	--	$c_i^p + \hat{c}_i^{S0}$
							1	$\hat{c}_i^{S1}$	$\hat{\rho}_i^{S1}$	0		

$\hat{\bar{y}}^T$

Case	Resp Stat	Resp Val	Accrued Costs	Impute RVal	Dist ($\hat{y}_i$-$\hat{\bar{y}}$)	Dist Rank	Next Strat	Future Costs	Resp Prop	Resp Class	**Allocation 0%** RVal	**Allocation 0%** Cost
390	1	y_i	c_i^p	y_i							y_i	c_i^p
194	1	y_i	c_i^p	y_i							y_i	c_i^p
280	1	y_i	c_i^p	y_i							y_i	c_i^p
227	1	y_i	c_i^p	y_i							y_i	c_i^p
798	0	--	c_i^p	$\hat{y}_i$	$\hat{d}_i$	1	0	$\hat{c}_i{}^{S0}$	$\hat{\rho}_i{}^{S0}$	1	$\hat{y}_i$	$c_i^p + \hat{c}_i{}^{S0}$
							1	$\hat{c}_i{}^{S1}$	$\hat{\rho}_i{}^{S1}$	1		
578	0	--	c_i^p	$\hat{y}_i$	$\hat{d}_i$	2	0	$\hat{c}_i{}^{S0}$	$\hat{\rho}_i{}^{S0}$	0	--	$c_i^p + \hat{c}_i{}^{S0}$
							1	$\hat{c}_i{}^{S1}$	$\hat{\rho}_i{}^{S1}$	1		
638	0	--	c_i^p	$\hat{y}_i$	$\hat{d}_i$	3	0	$\hat{c}_i{}^{S0}$	$\hat{\rho}_i{}^{S0}$	1	$\hat{y}_i$	$c_i^p + \hat{c}_i{}^{S0}$
							1	$\hat{c}_i{}^{S1}$	$\hat{\rho}_i{}^{S1}$	0		
742	0	--	c_i^p	$\hat{y}_i$	$\hat{d}_i$	4	0	$\hat{c}_i{}^{S0}$	$\hat{\rho}_i{}^{S0}$	0	--	$c_i^p + \hat{c}_i{}^{S0}$
							1	$\hat{c}_i{}^{S1}$	$\hat{\rho}_i{}^{S1}$	0		

$\hat{\bar{y}}^{T}$ $\hat{\bar{y}}^{A00}$ $\hat{C}^{A00}$

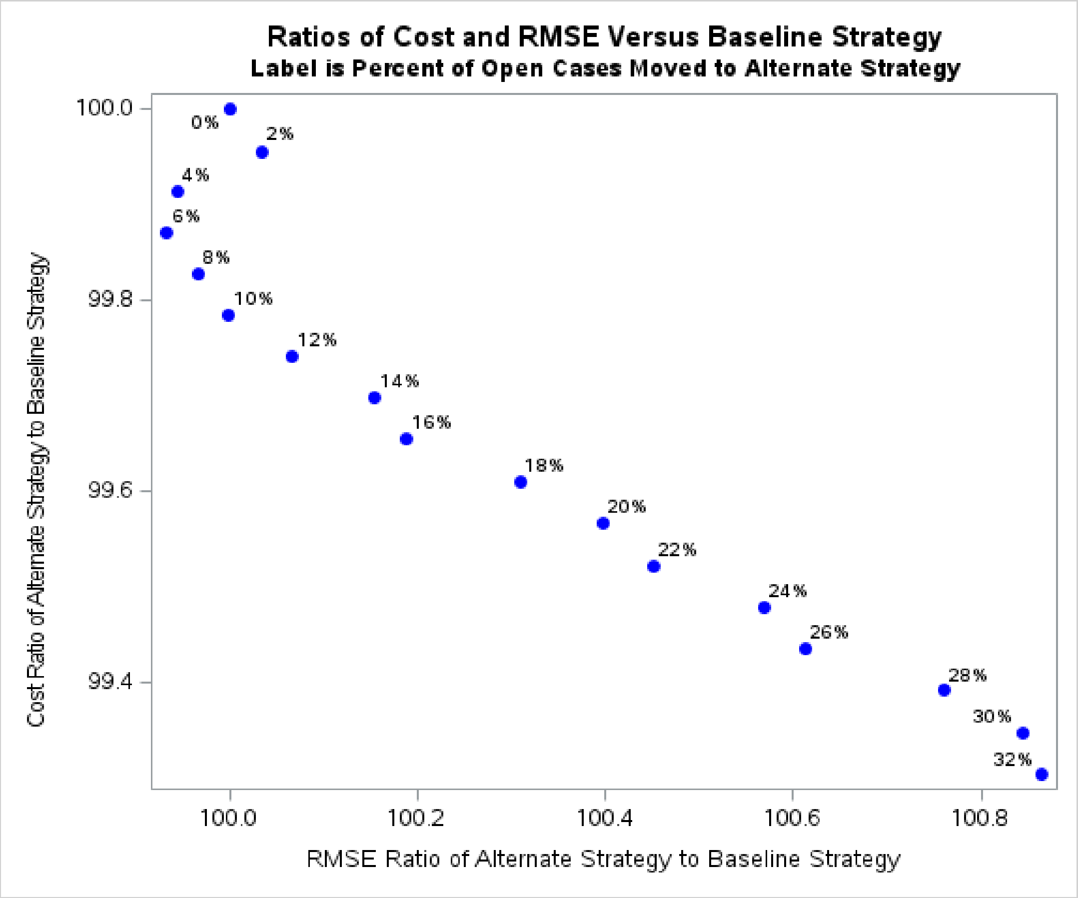
Case	Resp Stat	Resp Val	Accrued Costs	Impute RVal	Dist $(\hat{y}_i - \hat{\bar{y}})$	Dist Rank	Next Strat	Future Costs	Resp Prop	Resp Class	Allocation 0% RVal	Allocation 0% Cost	Allocation 50% RVal	Allocation 50% Cost
390	1	y_i	c_i^p	y_i							y_i	c_i^p	y_i	c_i^p
194	1	y_i	c_i^p	y_i							y_i	c_i^p	y_i	c_i^p
280	1	y_i	c_i^p	y_i							y_i	c_i^p	y_i	c_i^p
227	1	y_i	c_i^p	y_i							y_i	c_i^p	y_i	c_i^p
798	0	--	c_i^p	$\hat{y}_i$	$\hat{d}_i$	1	0	$\hat{c}_i^{S0}$	$\hat{\rho}_i^{S0}$	1	$\hat{y}_i$	$c_i^p + \hat{c}_i^{S0}$		
							1	$\hat{c}_i^{S1}$	$\hat{\rho}_i^{S1}$	1			$\hat{y}_i$	$c_i^p + \hat{c}_i^{S1}$
578	0	--	c_i^p	$\hat{y}_i$	$\hat{d}_i$	2	0	$\hat{c}_i^{S0}$	$\hat{\rho}_i^{S0}$	0	--	$c_i^p + \hat{c}_i^{S0}$		
							1	$\hat{c}_i^{S1}$	$\hat{\rho}_i^{S1}$	1			$\hat{y}_i$	$c_i^p + \hat{c}_i^{S1}$
638	0	--	c_i^p	$\hat{y}_i$	$\hat{d}_i$	3	0	$\hat{c}_i^{S0}$	$\hat{\rho}_i^{S0}$	1	$\hat{y}_i$	$c_i^p + \hat{c}_i^{S0}$	$\hat{y}_i$	$c_i^p + \hat{c}_i^{S0}$
							1	$\hat{c}_i^{S1}$	$\hat{\rho}_i^{S1}$	0				
742	0	--	c_i^p	$\hat{y}_i$	$\hat{d}_i$	4	0	$\hat{c}_i^{S0}$	$\hat{\rho}_i^{S0}$	0	--	$c_i^p + \hat{c}_i^{S0}$	--	$c_i^p + \hat{c}_i^{S0}$
							1	$\hat{c}_i^{S1}$	$\hat{\rho}_i^{S1}$	0				

$\hat{\bar{y}}^{T}$ $\quad$ $\hat{\bar{y}}^{A00}$ $\quad$ $\hat{C}^{A00}$ $\quad$ $\hat{\bar{y}}^{A50}$ $\quad$ $\hat{C}^{A50}$

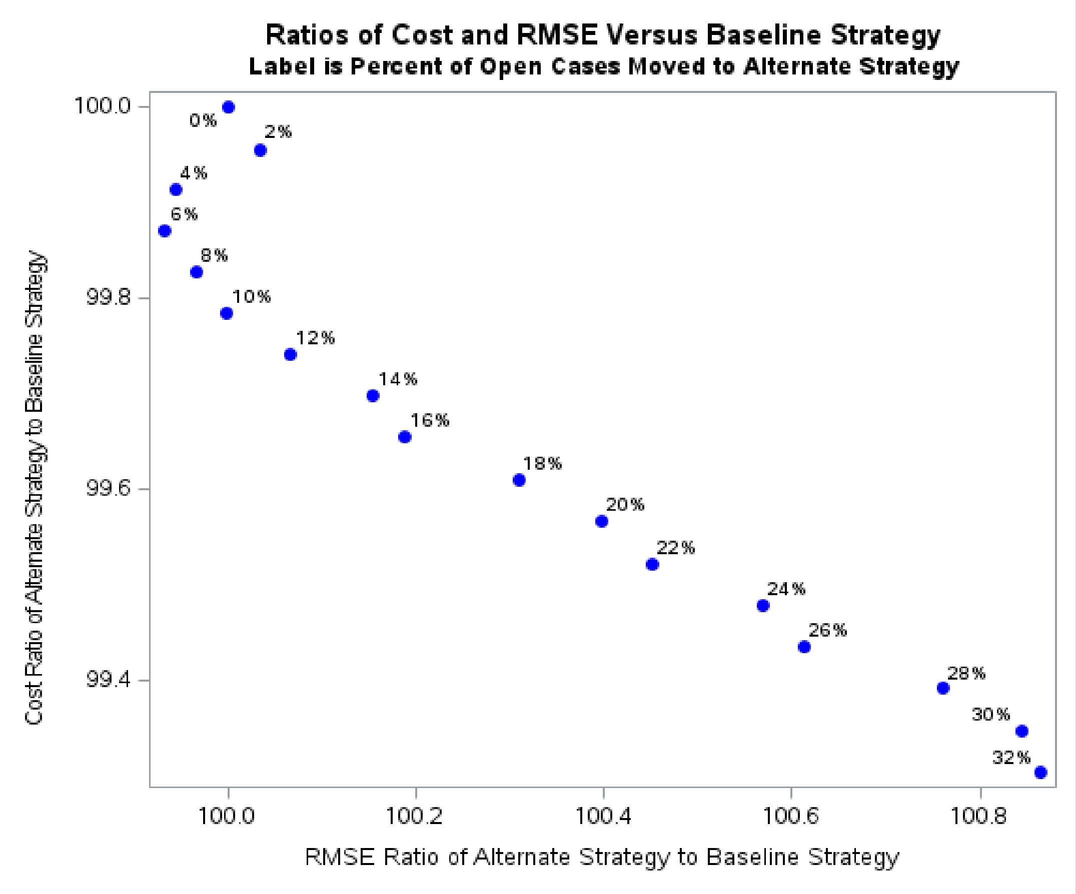
Optimization Output

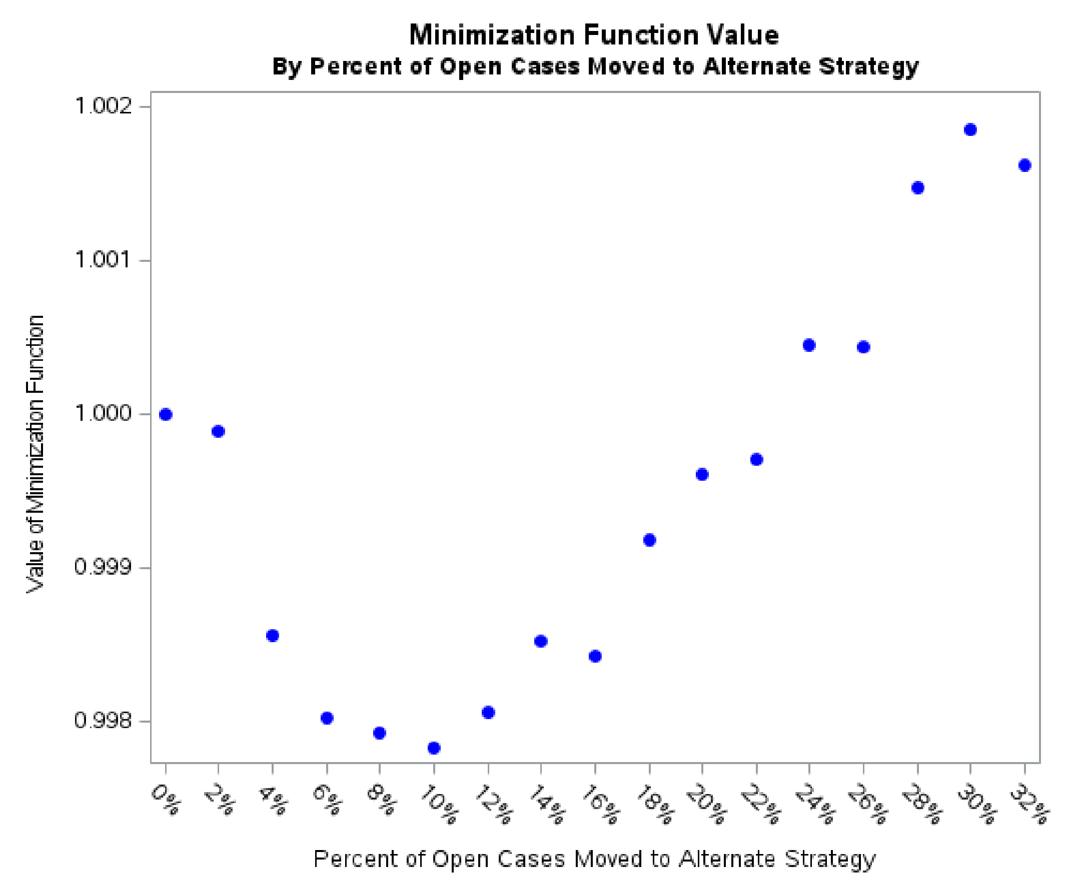
- Predicted responses:
  - Assuming full response – target mean: $\hat{\bar{y}}^{\boldsymbol{T}}$
  - Different strategies: Baseline strategy: $\hat{\bar{y}}^{\boldsymbol{A00}}$; Alternate strategy: $\hat{\bar{y}}^{\boldsymbol{A50}}$
- RMSE for each strategy:
  - $RMSE(S^{A}) = \left(\hat{\bar{y}}^{\boldsymbol{A}} - \hat{\bar{y}}^{\boldsymbol{T}}\right)^2 + \mathrm{Var}\left(\hat{\bar{y}}^{\boldsymbol{A}}\right)$
- Total costs for baseline and alternate strategy
  - $\hat{C}^{A00} = \sum_{i\in R} c_i^p + \sum_{i\in S}(c_i^p + \hat{c}_i{}^{A00})$
  - $\hat{C}^{A50} = \sum_{i\in R} c_i^p + \sum_{i\in S^{A00}}(c_i^p + \hat{c}_i{}^{A00}) + \sum_{i\in S^{A50}}(c_i^p + \hat{c}_i{}^{A50})$
- Ratios of alternate vs baseline: $\left(\frac{RMSE\left(S^{A50}\right)}{RMSE\left(S^{A00}\right)} \;, \; \frac{\hat{C}^{A50}}{\hat{C}^{A00}}\right)$

Decision Point #1:
Replace Questionnaire with Web Invite



Decision Point #1: Week 6
Replace Questionnaire with Web Invite





Results:
Data Collection Costs

	Treatment	Control	Sig.
Sample Size	8,000	8,000	
	Data Collection Costs		
Mean Cost-per-Case	$26.81	$29.57	*
Median Cost-per-Case	$20.22	$26.81	

$*sig\ (\alpha = 0.05)$

Results:
mean(Salary) & RMSE(Salary)

Salary Cutoff for Estimation	**$1,000,000**	
Treatment Group	Treatment	Control
% Respondents Included	100.00%	99.94%
Mean Salary ($)	84,082.10	84,250.02
RMSE Salary	62,776.47	61,940.82
Bias in Mean Salary ($)	-167.92 (-)	
% Difference RMSE	1.35% (-)	

$*sig\ (\alpha = 0.05)$

Results: Response Rate

	Treatment	Control	Sig.
Sample Size	8,000	8,000	
	Response Rate		
Unweighted Response Rate	57.08%	58.23%	-
Percent of Response from Web	85.92%	83.50%	*
Percent of Response from Mail	8.59%	10.32%	-
Percent of Response from CATI	5.50%	6.18%	-

$*sig\ (\alpha = 0.05)$

Conclusions

- In our pre-experiment research, Bayesian methods led to reduced prediction error (RP, salary)
- Possible to implement:
  - Bayesian prediction models in a production setting
  - Decision framework that balances data collection costs and quality
- Positive experimental results:
  - Saved approximately 9% of data collection costs ($p < 0.05$)
  - Mean value of self-reported salary decreased 0.20% (ns)
  - RMSE of mean(salary) increased 1.3% (ns)
  - Unweighted response rate decreased 1.15% (ns)
  - In-line with the predicted expectations
- These methods show promise for improving data collection outcomes

Limitations and Future Work

- Consider multiple survey items
  - Experiment only focused on one survey item, salary
- Improve predictive models and utilize a fully Bayesian approach
  - Experiment was not fully Bayesian because of cost models
- Incorporate survey weights
  - Weighted mean maybe significantly different from unweighted mean
  - Weight variability can increase variance of key survey estimates

References

[1] Wagner, J. and Hubbard, F. (2014), Producing unbiased estimates of propensity models during data collection. *Journal of Survey Statistics and Methodology*, 2, 323-342.

[2] Schouten, B., Mushkudiani, N., Shlomo, N., Durrant, G., Lundquist, P., Wagner, J. (2018), A Bayesian Analysis of Design Parameters in Survey Data Collection. *Journal of Survey Statistics and Methodology*, 6, 431-464.

[3] West, B.T., Wagner, J., Coffey, S., and Elliott, M.R. (revise and resubmit), The Elicitation of Prior Distributions for Bayesian Responsive Survey Design: Historical Data Analysis vs. Literature Review. Submitted to the *Journal of the Royal Statistical Society (Series A)*, May 2019. Retrieved from https://arxiv.org/ftp/arxiv/papers/1907/1907.06560.pdf

[4] Wagner, J., Coffey, S., Elliott, M.R., and West, B.T. (in press). Comparing the Ability of Regression Modeling and Bayesian Additive Regression Trees to Predict Costs in a Responsive Survey Design Context. *Journal of Official Statistics*.

[5] Coffey, S., West, B.T., Wagner, J., Elliott, M.R. (2020). What Do You Think? Using Expert Opinion to Improve Predictions of Response Propensity Under a Bayesian Framework. *methods, data, analyses,* 14(2), 159-194.

[6] Spiegelhalter, D. J., K. R. Abrams and J. P. Myles (2004), *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Chichester ; Hoboken, NJ, John Wiley & Sons.

[7] O’Hagan, A. (2019). Expert knowledge elicitation: subjective but scientific. *The American Statistician*, 73, 69-81.

Questions?