How to Use the Internet for Election Surveys

Simon Jackman and Douglas Rivers

Stanford University and Polimetrix, Inc.

May 9, 2008

Theory and Practice

		Theory	
		Works	**Doesn't work**
Practice	**Works**	Great!	Black magic
	Doesn't work	Too bad	Don't!

What Works in Theory and Practice

- Current Population Survey
- Biennial Registration and Voting Supplements in November of election years
- High response rate (92% in 2004)
- Overstates registration and voting slightly (about 2-3%)
- Aside from misreporting, it can't have large errors
  - By method of Cochran, Mosteller and Tukey, error bound is about 7%

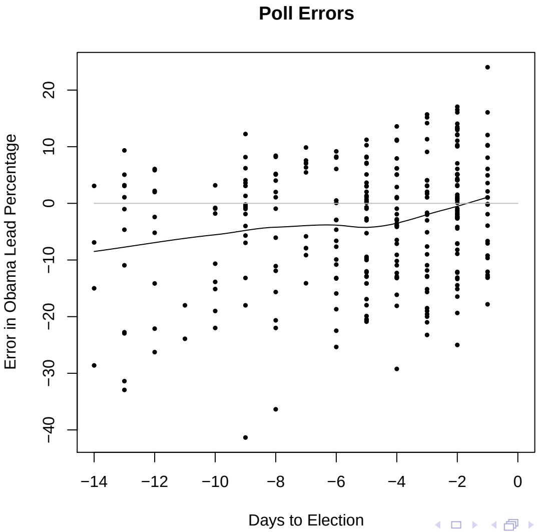
2004 National Election Study

- 66% response rate in pre-election wave with 88% reinterview rate for an overall response rate of 58%
- 89% registration rate and 76% turnout rate are too high.
  - Versus CPS estimates of 72% registration and 64% turnout
  - Actual was about 70% registration and 61% VEP turnout
- CMT bounds for NES are typically around $\pm 30\%$
- But, in practice, it works reasonably well.

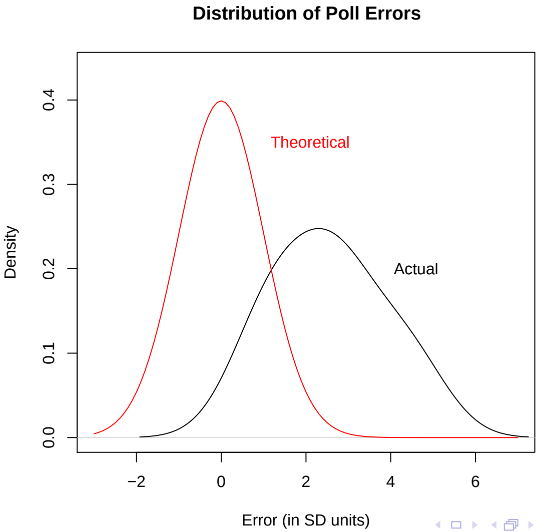
RDD Telephone Surveys

- Actual response rates for short field periods are very low (15-20%)
- Higher contact (and response rates) possible for longer field periods
- Within-household selection usually non-random
  - Gallup uses quotas
  - ABC/Washington Post uses unequal probabilities of selection, but ignores in weighting
  - Oldest male/youngest female—something that fails in theory, but seems to work in practice

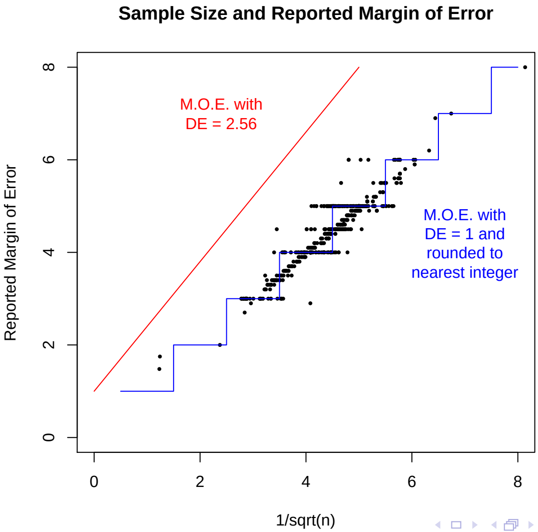
Errors in Democratic Primary Polls in 2008



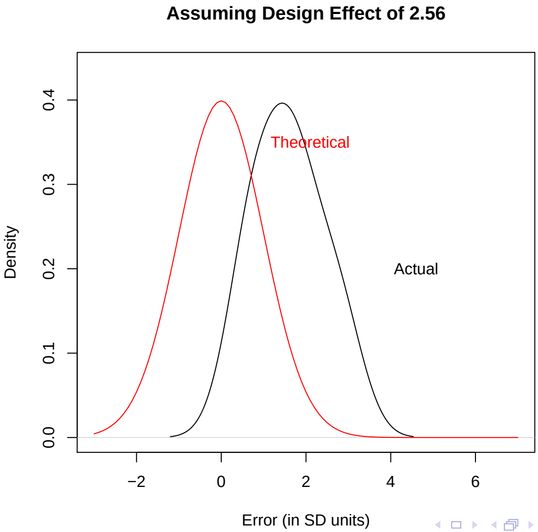
Actual and Theoretical Error Distribution in NH



Don’t believe the MOE!



Incorporating a Design Effect for Weighting



2008 Web Election Surveys

- 2008 is the first year with large Internet-based election surveys
  - NES (dedicated panel, recruited by KN)
  - NAES and AP/Yahoo/Harvard (KN access panel)
  - CCAP, CCES, and campaigns (YG/Polimetrix access panel)

Benefits of RDD Recruitment

- Coverage of non-Internet households
- RDD widely accepted as “probability-based”
- Can calculate a meaningful response rate
- Avoids “volunteers”

Problems with RDD Recruitment

- Expensive
- Small
- Overuse
- Response rates are low
- Attrition is high
- Practical fixes require abandoning theoretical purity

An Alternative Approach

- Sample selection from a panel with unknown selection probabilities
  - Use large opt-in panel
- Availability of large amounts of auxiliary data on *both* population and panel from voter and consumer databases, high quality probability samples
  - We know a lot about both the target population and the panelists and should use it.
- Sample matching used for selection of subsamples from the panel
  - Works in both theory and practice

Comparison of RDD with Web Opt-ins

Group	Unweighted RDD	Web Opt-ins	2004 ACS	2003 CPS Internet
Blacks	7.9%	4.3%	11.8%	9.3%
Hispanics	4.8%	3.3%	13.5%	7.2%
Postgrad	17.2%	23.3%	9.4%	14.7%
Age 18-24	6.4%	8.7%	10.3%	16.0%
Male	41.9%	58.8%	48.9%	48.7%
Married	57.7%	60.4%	54.3%	55.3%

Bias and SEs

- Standard errors measure sampling variability, not bias.
- Possible to calculate SEs without knowing data generating process
- Observations are, at a minimum, exchangeable and usually independent.
- No logical difference between nonresponse and self-selection.
  - In both cases, the selection probabilities are unknown
  - Validity of estimates depends upon untestable modeling assumptions
- Standard approach for both RDD and Web panels leave substantial amounts of bias

Auxilliary Information

- Voter and consumer databases provide a sampling frame for social science research.
- Frame contains a large amount of auxiliary information that can be used for bias reduction.
  - Age, gender, vote history, address, name
  - Home value, children, interests, magazines
- Data available for everyone—both panelists and population.

Idea: *Draw a sample from the frame and find the closest matching respondents from a panel.*

Sample Matching

- Recruit a large *reservoir* of respondents who are accessible for interviewing (the “panel”).
- Obtain a population *frame* containing auxiliary information for matching.
- Select a *target sample* from the frame.
- For each unit in the target sample, find the closest matching unit in the reservoir. This is the *matched sample*.

Variants:

- Use a high quality sample from another source as the target sample.
- Dynamic matching: match to multiple studies simultaneously using a flow of invitations.

Notation

$$\begin{aligned} N &= \text{size of panel} \\ n &= \text{size of sample} \\ X &= \text{covariates } (k \text{ vector}) \\ Y &= \text{survey measurements} \\ Z &= \text{panel membership indicator} \end{aligned}$$

Panel

$$(\tilde{X}_1, \tilde{Y}_1), \ldots, (\tilde{X}_N, \tilde{Y}_N)$$

Distributions and Parameters

$$
\begin{aligned}
f_X(x) &= \text{density of } X \text{ in population} \\
\tilde{f}_X(x) &= \text{density of } X \text{ conditional on } Z_i = 1 \\
f_{Y|X}(y|x) &= \text{conditional distribution of } Y \text{ given } X \\
\tilde{f}_{Y|X}(y|x) &= \text{conditional distribution in the panel} \\
\mu(x) &= E(Y|X = x) = \int y f_{Y|X}(y|x)\, dx \\
\theta_0 &= E(Y) = \int \mu(x)\, f_X(x)\, dx \\
\sigma_1^2(x) &= V(Y|X = x, Z = 1)
\end{aligned}
$$

Assumptions

- **IID Data Generating Process** (X_i, Y_i, Z_i) are i.i.d.
- **Ignorable Selection** $f_{Y|X}(y|x) = \tilde{f}_{Y|X}(y|x)$
- **Continuous Covariates** X has a continuous distribution with bounded convex support
- **Overlap** The support of X is the same in the panel as in the population with density bounded away from zero
- **Continuity** $\mu(x)$ is Lipschitz continuous
- **Regularity** $V(Y|X, Z = 1)$ is uniformly bounded

Matching Process

- **Target Sample:** Choose a (stratified) random sample of size n from the frame $(X_1, \ldots, X_n)$.
- For each element of the target sample, find the closest matching element $M(i)$ in the panel:

$$M(i) = j \text{ iff } |X_i - \tilde{X}_j| \leq |X_i - \tilde{X}_\ell| \text{ for all } \ell \text{ in the panel}$$

Let $X_i^* = \tilde{M}(i)$ and $Y_i^* = \tilde{Y}_{M(i)}$.

Matched Sample

$$(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$$

Matching Estimator

$$\tilde{\theta} = n^{-1} \sum_{i=1}^{n} \tilde{Y}_i$$

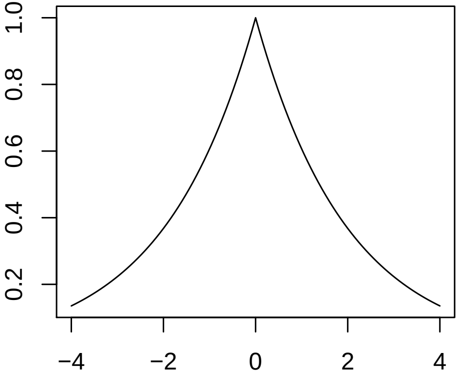
Scalar Matching

- The conditional density of X_i^* given $X_i = x$ is

$$N\tilde{f}(x)[1 - \tilde{F}_X(x + |x^* - x|) + \tilde{F}_X(x - |x^* - x|)]^{N-1}$$

  where $\tilde{F}_X$ is the distribution function of X in the panel.
- Conditional on $X_i = x$, the limiting distribution of $N\,(X_i^* - x)$ is Laplace with mean zero and variance $1/2\tilde{f}_X(x)^2$.
- The approximate distribution of $X_i^* - X_i$ is, thus, Laplace with mean zero and variance $O(1/N^2)$

Asymptotic Distribution of $X^* - X$



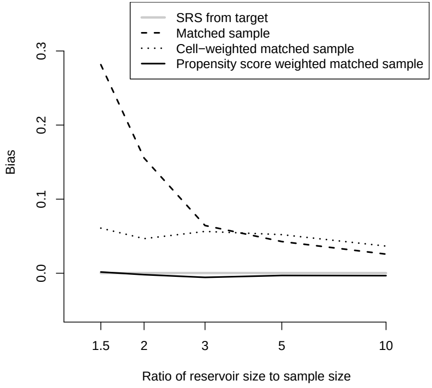
Theoretical Results

- When matching on a $\sqrt{n}$-consistent estimator of the propensity score, the matched sample is consistent and asymptotically normal.
- When using nearest-neighbors matching and the number of matching variables is greater than two, the estimate is consistent, but involves a bias of order $O(n^{k/2}/N)$ where k is the number of matching variables.
- In general, propensity score matching is to be preferred.

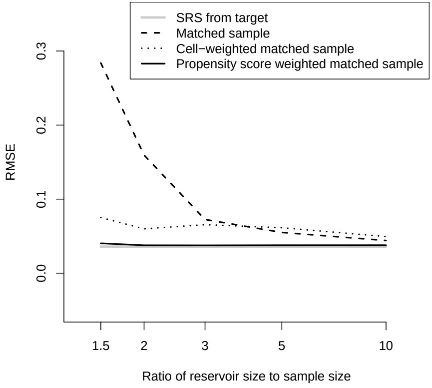
Monte Carlo Simulations

- Covariates have different means and covariance structure
  - Support of X is $[-1, 1] \times [1, 1]$
  - Population distribution is truncated bivariate normal with mean zero, SD 1, and correlation 0.3
  - Panel distribution is truncated bivariate normal with mean $(0.2, -0.3)$, SD 1, and correlation -0.5
- $Y|X \sim N(X_1 + X_2/2, 1)$ (ignorable selection)
- Sample size of $n = 1000$
- Panel size of $N = 1500, 2000, 5000$ or 10000.
- 1000 Monte Carlo Simulations

Simulated Bias of Estimators



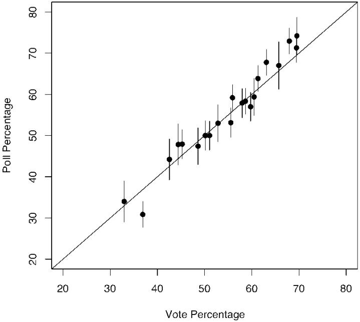
Simulated RMSE of Estimators



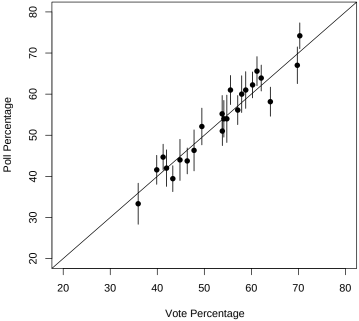
Empirical Application: 2006 CCES

- Consortium of 37 universities
- Pre- and post election interviews
- $n = 37,000$
- $N = 129,000$
- Frame was 2005 ACS matched to 2004 NEP Exit Poll
- 7-point party ID and 5-point ideology imputed from 2004 NAES
- Matched on age, race, gender, education, income, martial status, party ID, and ideology.

Senate Estimates and 95% Confidence Intervals



Governor Estimates and 95% Confidence Intervals



Comparison of Accuracy

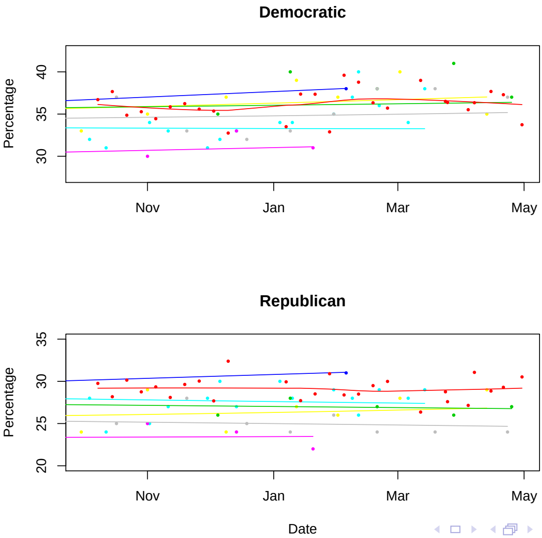
Source	n	Mean Error	RMSE
Phone	255	2.76	8.34
Rasmussen (IVR)	83	3.82	8.47
SurveyUSA (IVR)	63	3.4	7.25
Zogby (Internet)	72	4.86	9.36
Polimetrix (Internet)	40	-0.47	5.21

Source: Blumenthal and Franklin (2007)

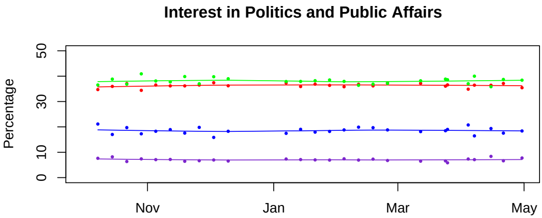
Some Lessons from 2006

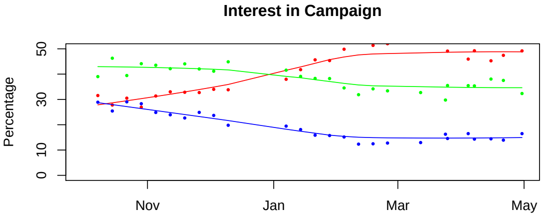
- Some categories are in short supply, even in large panels:
  - Older, low education minorities
- Panelists recruited through public opinion surveys have high levels of political interest (and over-report registration and turnout)
  - The much-maligned “paid survey takers” are helpful for matching unregistered voters.
- Matching is imperfect, so weighting is important to remove remaining biases.
  - However, much smaller weights are needed after matching.
- Panel attrition requires modeling regardless of the method of sampling.
  - The question is not whether you do weighting, but whether you do it well.

Party ID Trends 2007-08



Political and Campaign Interest 2007-08





Trial Heats 2008