Statistical Analysis of List Experiments

Kosuke Imai

Princeton University

Joint work with Graeme Blair

July 4, 2012

Motivation

- Validity of much empirical social science research relies upon accuracy of *self-reported* behavior and beliefs
- Challenge: eliciting truthful answers to sensitive survey questions
  e.g., racial prejudice, corruptions, fraud, support for militant groups
- Social desirability bias, privacy and safety concerns
- Lies and non-responses
- Solution: Indirect rather than direct questioning
  1. Randomization: Randomized response technique
  2. Aggregation: List experiment (item count technique)

List Experiment: An Example

- The 1991 National Race and Politics Survey (Sniderman et al.)
- Randomize the sample into the treatment and control groups
- The script for the control group

```
Now I'm going to read you three things that sometimes
make people angry or upset. After I read all three,
just tell me HOW MANY of them upset you. (I don't
want to know which ones, just how many.)

(1) the federal government increasing the tax on
gasoline;
(2) professional athletes getting million-dollar-plus
salaries;
(3) large corporations polluting the environment.
```

List Experiment: An Example

- The 1991 National Race and Politics Survey (Sniderman et al.)
- Randomize the sample into the treatment and control groups
- The script for the treatment group

```
Now I'm going to read you four things that sometimes
make people angry or upset. After I read all four,
just tell me HOW MANY of them upset you. (I don't
want to know which ones, just how many.)

(1) the federal government increasing the tax on
gasoline;
(2) professional athletes getting million-dollar-plus
salaries;
(3) large corporations polluting the environment;
(4) a black family moving next door to you.
```

Methodological Challenges

- List experiment is becoming popular:

  Kuklinski et al., 1997a,b; Sniderman and Carmines, 1997; Gilens et al., 1998; Kane et al., 2004; Tsuchiya et al., 2007; Streb et al., 2008; Corstange, 2009; Flavin and Keane, 2010; Glynn, 2010; Gonzalez-Ocantos et al., 2010; Holbrook and Krosnick, 2010; Janus, 2010; Redlawsk et al., 2010; Coutts and Jann, 2011

- Standard practice: Use difference-in-means to estimate the proportion of those who answer yes to sensitive item

- Getting more out of list experiments:
  1. Who are more likely to answer yes?
  2. Who are answering differently to direct and indirect questioning?
  3. Can we study multiple sensitive items in one survey?
  4. Can we detect failures of list experiments?
  5. Can we correct violations of key assumptions?

- Recoup the efficiency loss due to indirect questioning

Overview of the Project

- Goals:
  1. Develop *multivariate regression analysis* methodology
  2. Develop statistical tests to detect *failures of list experiments*
  3. Develop methods to correct deviations from key assumption
  4. Develop open-source software to implement the proposed methods
  5. Applications in Afghanistan (joint work with J. Lyall) and Nigeria

- References:
  1. Imai, K. "Multivariate Regression Analysis for the Item Count Technique." *Journal of the American Statistical Association*
  2. Blair, G. and K. Imai. "Statistical Analysis of List Experiments." *Political Analysis*
  3. Blair, G. and K. Imai. `list: Statistical Methods for the Item Count Technique and List Experiments` availalble at http://cran.r-project.org/package=list

Identification Assumptions

1. Randomization of the Treatment
2. No Design Effect: The inclusion of the sensitive item does not affect answers to control items
3. No Liars: Answers about the sensitive item are truthful

Under these assumptions, difference-in-means estimator is unbiased

New Multivariate Regression Estimators

- Notation:
  - J: number of control items
  - N: number of respondents
  - T_i: binary treatment indicator ($1 =$ treatment, $0 =$ control)
  - X_i: pre-treatment covariates
  - Y_i: outcome variable

- The nonlinear least squares regression model:

$$Y_i \;=\; \underbrace{f(X_i, \gamma)}_{control\ \ items} \;+\; \underbrace{T_i \cdot g(X_i, \delta)}_{sensitive\ \ item} + \epsilon_i$$

- Differeince-in-means: no covariate
- Linear model: $f(x, \gamma) = x^\top \gamma$ and $g(x, \delta) = x^\top \delta$
- Logit model: $f(x, \gamma) = J \cdot \mathrm{logit}^{-1}(x^\top \gamma)$ and $g(x, \delta) = \mathrm{logit}^{-1}(x^\top \delta)$
- Two-step estimation with appropriate standard error

- Define a type of each respondent by
  - total number of yes for control items $Y_i(0)$
  - truthful answer to the sensitive item Z_i^*
- A total of $(2 \times (J + 1))$ types
- Example: three control items ($J = 3$)

Y_i	Treatment group	Control group
4	(3,1)	
3	(2,1) (3,0)	(3,1) (3,0)
2	(1,1) (2,0)	(2,1) (2,0)
1	(0,1) (1,0)	(1,1) (1,0)
0	(0,0)	(0,1) (0,0)

Extracting More Information from List Experiments

- Define a type of each respondent by
  - total number of yes for control items $Y_i(0)$
  - truthful answer to the sensitive item Z_i^*
- A total of $(2 \times (J+1))$ types
- Example: three control items ($J = 3$)

Y_i	Treatment group	Control group
4	(3,1)	
3	(2,1) (3,0)	(3,1) (3,0)
2	(1,1) (2,0)	(2,1) (2,0)
1	~~(0,1)~~ ~~(1,0)~~	(1,1) ~~(1,0)~~
0	~~(0,0)~~	~~(0,1)~~ ~~(0,0)~~

- *Joint distribution* of $(Y_i(0), Z_i^*)$ is identified

Maximum Likelihood and Bayesian Estimation

- Model for sensitive item as before: e.g., logistic regression

$$\Pr(Z^*_{i,J+1} = 1 \mid X_i = x) \;=\; \mathrm{logit}^{-1}(x^\top \delta)$$

- Model for control items given the response to sensitive item: e.g., binomial or beta-binomial logistic regression

$$\Pr(Y_i(0) = y \mid X_i = x, Z^*_{i,J+1} = z) \;=\; J \times \mathrm{logit}^{-1}(x^\top \psi_z)$$

The Likelihood Function

- Mixture structure:

$$\prod_{i \in \mathcal{J}(1,0)} (1 - g(X_i, \delta)) h_0(0; X_i, \psi_0) \prod_{i \in \mathcal{J}(1,J+1)} g(X_i, \delta) h_1(J; X_i, \psi_1)$$
$$\times \prod_{y=1}^{J} \prod_{i \in \mathcal{J}(1,y)} \{g(X_i, \delta) h_1(y-1; X_i, \psi_1) + (1 - g(X_i, \delta)) h_0(y; X_i, \psi_0)\}$$
$$\times \prod_{y=0}^{J} \prod_{i \in \mathcal{J}(0,y)} \{g(X_i, \delta) h_1(y; X_i, \psi_1) + (1 - g(X_i, \delta)) h_0(y; X_i, \psi_0)\}$$

  where $\mathcal{J}(t, y)$ represents a set of respondents with $(T_i, Y_i) = (t, y)$

- Maximizing this function is difficult

Missing Data Framework

- Consider $Z^*_{i,J+1}$ as partially missing data
- The complete-data likelihood has a much simpler form:

$$\prod_{i=1}^{N} \left\{ g(X_i,\delta) h_1(Y_i - 1; X_i, \psi_1)^{T_i} h_1(Y_i; X_i, \psi_1)^{1-T_i} \right\}^{Z^*_{i,J+1}} \times \{(1 - g(X_i,\delta)) h_0(Y_i; X_i, \psi_0)\}^{1-Z^*_{i,J+1}}$$

- The EM algorithm: only separate optimization of $g(x,\delta)$ and $h_z(y; x, \psi_z)$ is required
  - weighted logistic regression
  - weighted binomial logistic regression
- Easy to develop the Gibbs sampling algorithm

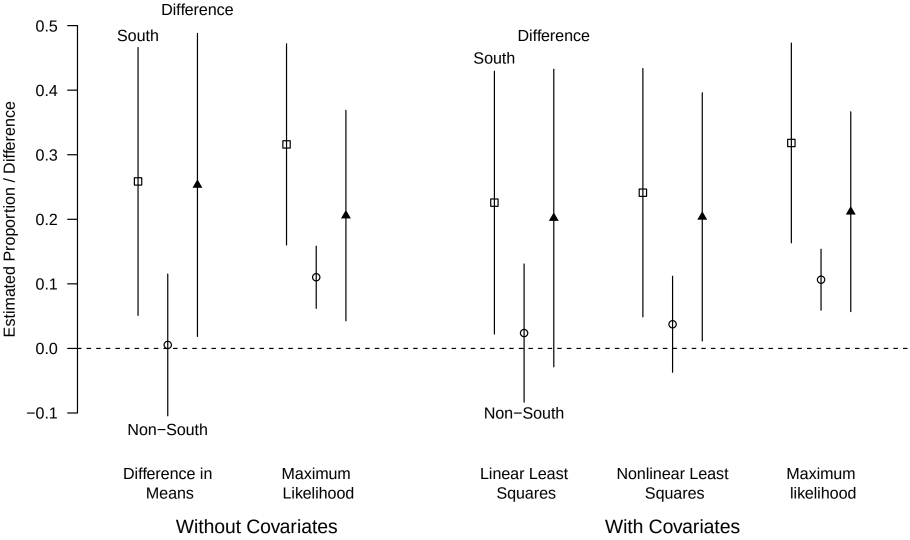
Empirical Application: Racial Prejudice in the US

- Kuklinski *et al*. (1997 JOP): Southern whites are more prejudiced against blacks than non-southern whites – no "New South"
- The limitation of the original analysis:

  So far our discussion has implicitly assumed that the higher level of prejudice among white southerners results from something uniquely "southern," what many would call southern culture. This assumption could be wrong. If white southerners were older, less educated, and the like – characteristics normally associated with greater prejudice – then demographics would explain the regional difference in racial attitudes

- Need for a multivariate regression analysis

Estimated Proportion of Prejudiced Whites



- MLE yields more efficient estimates

Studying Multiple Sensitive Items

- The 1991 National Race and Politics Survey includes another treatment group with the following sensitive item

  (4) "black leaders asking the government for affirmative action"

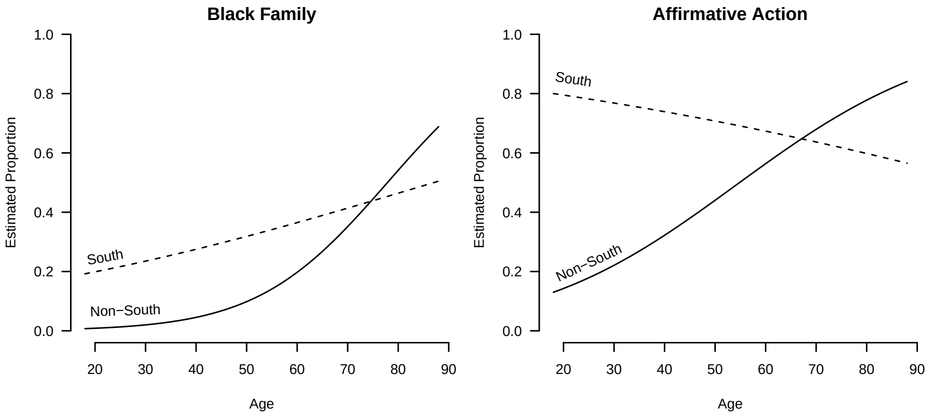
- Use of the same control items permits joint-modeling
- Same assumptions: No Design Effect and No Liars
- Extension to the design with K sensitive items

Multivariate Regression Results

- How do the patterns of generational changes differ between South and Non-South?
- Original analysis dichotomized the age variable without controlling for other factors

	Sensitive Items				Control Items	
	Black Family		Affirmative Action			
Variables	est.	s.e.	est.	s.e.	est.	s.e.
`intercept`	−7.575	1.539	−5.270	1.268	1.389	0.143
`male`	1.200	0.569	0.538	0.435	−0.325	0.076
`college`	−0.259	0.496	−0.552	0.399	−0.533	0.074
`age`	0.852	0.220	0.579	0.147	0.006	0.028
`South`	4.751	1.850	5.660	2.429	−0.685	0.297
`South` × `age`	−0.643	0.347	−0.833	0.418	0.093	0.061
control items $Y_i(0)$	0.267	0.252	0.991	0.264		

Generational Changes in South and Non-South



- Age is important even after controling for gender and education
- Gender is not, contradicting with the original analysis

Measuing Social Desirability Bias

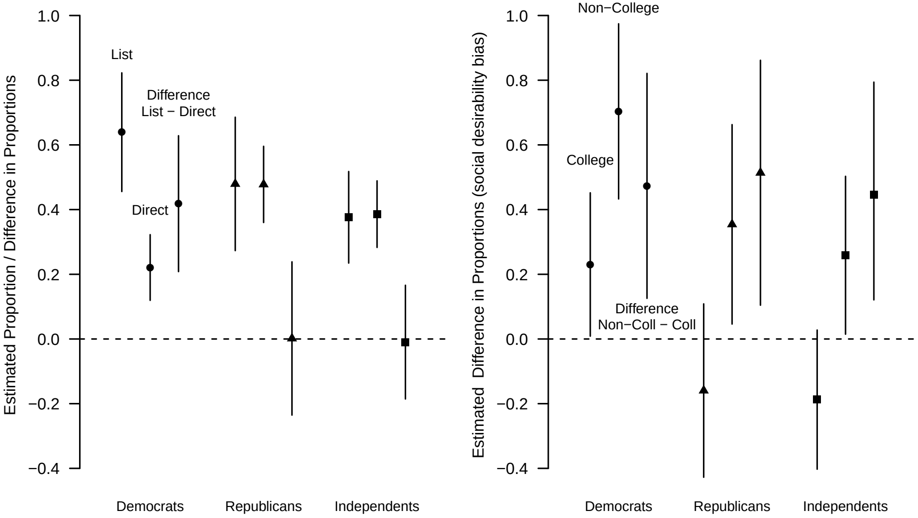
- The 1994 Multi-Investigator Survey (Sniderman et al.) asks list experiment question and later a direct sensitive question:

  `Now I'm going to ask you about another thing that sometimes makes people angry or upset.`

  `Do you get angry or upset when black leaders ask the government for affirmative action?`

- Difference between direct and indirect responses
  $\Longrightarrow$ measure of social desirability bias

Differences for the Affirmative Action Item



When Can List Experiments Fail?

- Recall the two assumptions:
  1. No Design Effect: The inclusion of the sensitive item does not affect answers to non-sensitive items
  2. No Liars: Answers about the sensitive item are truthful
- Design Effect:
  - Respondents evaluate non-sensitive items relative to sensitive item
- Lies:
  - Ceiling effect: too many yeses for non-sensitive items
  - Floor effect: too many noes for non-sensitive items
- Both types of failures are difficult to detect
- Importance of choosing non-sensitive items
- Question: Can these failures be addressed statistically?

Hypothesis Test for List Experiments Failures

- Under the null hypothesis of no design effect and no liars, we expect all types $(y,1) > 0$ and $(y,0) > 0$

$$\begin{aligned}\pi_1 &= \Pr(\text{type} = (y,1)) = \Pr(Y_i \leq y \mid T_i = 0) - \Pr(Y_i \leq y \mid T_i = 1) \geq 0 \\ \pi_0 &= \Pr(\text{type} = (y,0)) = \Pr(Y_i \leq y \mid T_i = 1) - \Pr(Y_i < y \mid T_i = 0) \geq 0\end{aligned}$$

- Alternative hypothesis: *At least* one is negative
- A multivariate one-sided LR test for each $t = 0, 1$

$$\hat{\lambda}_t = \min_{\pi_t} (\hat{\pi}_t - \pi_t)^\top \widehat{\Sigma}_t^{-1} (\hat{\pi}_t - \pi_t), \quad \text{subject to } \pi_t \geq 0,$$

- $\hat{\lambda}_t$ follows a mixture of χ^2
- Difficult to characterize least favorable values under the joint null
- Bonferroni correction: Reject the joint null if $\min(\hat{p}_0, \hat{p}_1) \leq \alpha/2$
- GMS selection algorithm to increase statistical power

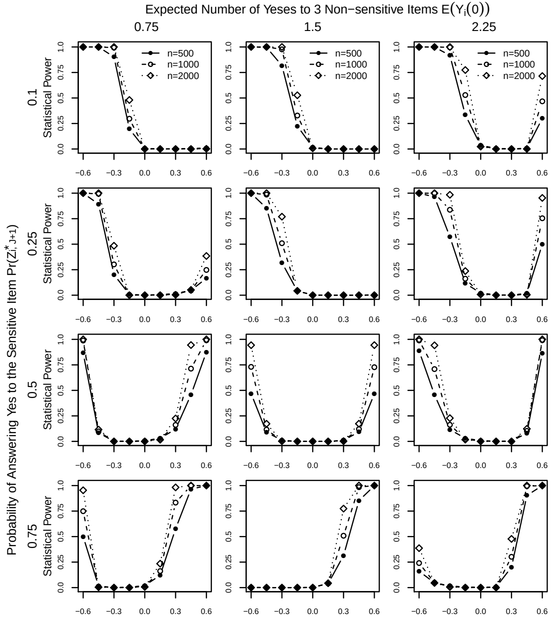
The Racial Prejudice Data Revisited

- Did the negative proportion arise by chance?

	Observed Data				Estimated Proportion of Respondent Types			
	Control		Treatment					
Response	counts	prop.	counts	prop.	$\hat{\pi}_{y0}$	s.e.	$\hat{\pi}_{y1}$	s.e.
0	8	1.4%	19	3.0%	3.0%	0.7	-1.7%	0.8
1	132	22.4	123	19.7	21.4	1.7	1.0	2.4
2	222	37.7	229	36.7	35.7	2.6	2.0	2.8
3	227	38.5	219	35.1	33.1	2.2	5.4	0.9
4			34	5.4				
Total	589		624		93.2		6.8	

- p-value $= 0.022$

Statistical Power of the Proposed Test



- Power is highly assymetric
- Depends on sensitive and control items
- Implications
  1. interpretation
  2. design

Modeling Ceiling and Floor Effects

- Potential liars:

Y_i	Treatment group	Control group
4	(3,1)	
3	(2,1) (3,0) (3,1)*	(3,1) (3,0)
2	(1,1) (2,0)	(2,1) (2,0)
1	(0,1) (1,0)	(1,1) (1,0)
0	(0,0) (0,1)*	(0,1) (0,0)

- Proposed strategy: model ceiling and/or floor effects under an additional assumption
- Identification assumption: conditional independence between items given covariates
- ML regression estimator can be extended to this situation
- A similar strategy applicable to design effects

Multivariate Regression Results

Variables	Ceiling Effects Alone est.	Ceiling Effects Alone s.e.	Floor Effects Alone est.	Floor Effects Alone s.e.	Both Ceiling and Floor Effects est.	Both Ceiling and Floor Effects s.e.
Intercept	−1.291	0.558	−1.251	0.501	−1.245	0.502
Age	0.294	0.101	0.314	0.092	0.313	0.092
College	−0.345	0.336	−0.605	0.298	−0.606	0.298
Male	0.038	0.346	−0.088	0.300	−0.088	0.300
South	1.175	0.480	0.682	0.335	0.681	0.335
Prop. of liars						
Ceiling	0.0002	0.0017			0.0002	0.0016
Floor			0.0115	0.0000	0.0115	0.0000

- Essentially no ceiling and floor effects
- Main conclusion for the affirmative action item seems robust

Concluding Remarks and Practical Suggestions

- List experiments: alternative to the randomized response method
- Advantages: easy to use, easy to understand
- Challenges: loss of information, violation of assumptions
- We develop a set of methods for list experiments

- Suggestions for analysis:
  1. Estimate proportions of types and test design effects
  2. Conduct multivariate regression analyses
  3. Investigate the robustness of findings to ceiling and floor effects

- Suggestions for design:
  1. Select control items to avoid skewed response distribution
  2. Avoid control items that are ambiguous and generate weak opinion
  3. Conduct a pilot study and maximize statistical power