Statistical Analysis of List Experiments

Kosuke Imai

Princeton University

Joint work with Graeme Blair

October 29, 2010

Motivation

- Survey is used widely in social sciences
- Validity of survey depends on the accuracy of self-reports
- Sensitive questions $\Longrightarrow$ social desirability, privacy concerns
  e.g., racial prejudice, corruptions, fraud, support for militant groups
- Lies and non-responses

- How can we elicit truthful answers to sensitive questions?
- Survey methodology: protect privacy through indirect questioning
- Statistical methodology: efficiently recover underlying responses

List Experiments and Project Overview

- **List Experiments** (Raghavarao and Federer, 1979)
  - Also known as total block response, item count technique, and unmatched count technique
  - Use aggregation to protect privacy
  - An alternative to randomized response technique
- Goals of this project:
  1. Develop methods for *multivariate regression analysis*
  2. Develop a statistical test to detect *failures of list experiments*
  3. Adjust for deviations from the standard list experiment assumption

The 1991 National Race and Politics Survey

- 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.
```

The 1991 National Race and Politics Survey

- 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.
```


Notation and Setup

- J: number of non-sensitive items
- N: number of respondents
- T_i: binary treatment indicator

- *Potential* outcomes notation: Neyman, Rubin, Holland
- $Z_{ij}(t)$: potential response to the jth non-sensitive item under treatment status $T_i = t$ where $j = 1, \ldots, J$ and $t = 0, 1$
- $Z_{i,J+1}(t)$: potential response to the sensitive item under treatment status $T_i = t$ where $Z_{i,J+1}(0)$ is truthful

- $Y_i(0) = \sum_{j=1}^{J} Z_{ij}(0)$: potential response under control condition
- $Y_i(1) = \sum_{j=1}^{J+1} Z_{ij}(1)$: potential response under treatment condition
- $Y_i = Y_i(T_i)$: observed response

Standard Analysis

- Assumptions:
  1. No Design Effect: The inclusion of the sensitive item does not affect answers to non-sensitive items, i.e., $\sum_{j=1}^{J} Z_{ij}(0) = \sum_{j=1}^{J} Z_{ij}(1)$
  2. No Liar: Answers about the sensitive item are truthful, i.e., $Z_{i,J+1}(0) = Z_{i,J+1}(1)$
- Difference-in-means estimator:

$$\hat{\tau} = \frac{1}{N_1}\sum_{i=1}^{N} T_i Y_i - \frac{1}{N_0}\sum_{i=1}^{N}(1 - T_i) Y_i$$

- Straightforward and unbiased under the above assumptions
- But, inefficient
- Difficult to explore multivariate relationship
- No existing method allows for multivariate regression analysis

Nonlinear Least Squares (NLS) Estimator

- Generalize the difference-in-means estimator to a multivariate regression estimator
- The Model:

$$Y_i = f(X_i, \gamma) + T_i g(X_i, \delta) + \epsilon_i$$

  - X_i: covariates
  - $f(x, \gamma)$: model for non-sensitive items, e.g., $J \times \mathrm{logit}^{-1}(x^\top \gamma)$
  - $g(x, \delta)$: model for sensitive item, e.g., $\mathrm{logit}^{-1}(x^\top \delta)$
- Two-step estimation procedure:
  1. Fit $f(x, \gamma)$ to the control group via NLS and obtain $\hat{\gamma}$
  2. Fit $g(x, \delta)$ to the treatment group via NLS after subtracting $f(X_i, \hat{\gamma})$ from Y_i and obtain $\hat{\delta}$
- Standard errors via the method of moments
- When no covariate, it reduces to the difference-in-means estimator

Extracting More Information from the Data

- Define a "type" of each respondent by $(Y_i(0), Z_{i,J+1}(0))$
  - $Y_i(0)$: total number of yes for non-sensitive items $\in \{0, 1, \dots, J\}$
  - $Z_{i,J+1}(0)$: truthful answer to the sensitive item $\in \{0, 1\}$
- A total of $(2 \times J)$ types
- Example: two non-sensitive 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* is identified

Extracting More Information from the Data

- Define a "type" of each respondent by $(Y_i(0), Z_{i,J+1}(0))$
  - $Y_i(0)$: total number of yes for non-sensitive items $\{0, 1, \dots, J\}$
  - $Z_{i,J+1}(0)$: truthful answer to the sensitive item $\{0, 1\}$
- A total of $(2 \times J)$ types
- Example: two non-sensitive 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* is identified:

$$\Pr(\text{type} = (y, 1)) = \Pr(Y_i \leq y \mid T_i = 0) - \Pr(Y_i \leq y \mid T_i = 1)$$

Extracting More Information from the Data

- Define a "type" of each respondent by $(Y_i(0), Z_{i,J+1}(0))$
  - $Y_i(0)$: total number of yes for non-sensitive items $\{0, 1, \ldots, J\}$
  - $Z_{i,J+1}(0)$: truthful answer to the sensitive item $\{0, 1\}$
- A total of $(2 \times J)$ types
- Example: two non-sensitive 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* is identified:

$$\Pr(\text{type} = (y, 1)) = \Pr(Y_i \le y \mid T_i = 0) - \Pr(Y_i \le y \mid T_i = 1)$$

$$\Pr(\text{type} = (y, 0)) = \Pr(Y_i \le y \mid T_i = 1) - \Pr(Y_i < y \mid T_i = 0)$$

The Maximum Likelihood (ML) Estimator

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

$$\Pr(Z_{i,J+1}(0) = 1 \mid X_i = x) = \text{logit}^{-1}(x^\top \delta)$$

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

$$\Pr(Y_i(0) = y \mid X_i = x, Z_{i,J+1}(0) = z) = J \times \text{logit}^{-1}(x^\top \psi_z)$$

- Difficult to maximize the resulting likelihood function
- Develop the EM algorithm for reliable estimation

The Likelihood Function

- $g(x, \delta) = \Pr(Z_{i,J+1}(0) = 1 \mid X_i = x)$
- $h_z(y; x, \psi_z) = \Pr(Y_i(0) = y \mid X_i = x, Z_{i,J+1}(0) = z)$:
- The likelihood function:

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

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 and the EM Algorithm

- Consider $Z_{i,J+1}(0)$ as missing data
- For some respondents, $Z_{i,J+1}(0)$ is "observed"
- The complete-data likelihood has a much simpler form:

$$\prod_{i=1}^{N} \left\{ g(X_i, \delta) h_1(Y_i - 1; X_i, \psi_1)^{T_i} h_1(Y_i; X_i, \psi_1)^{1-T_i} \right\}^{Z_{i,J+1}(0)}$$
$$\times \{(1 - g(X_i, \delta)) h_0(Y_i; X_i, \psi_0)\}^{1 - Z_{i,J+1}(0)}$$

- 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
- Both can be implemented in standard statistical software

Empirical Application: Racial Prejudice in the US

- Kuklinski *et al*. (1997) analyzes the 1991 National Race and Politics survey with the standard difference-in-means estimator
- Finding: Southern whites are more prejudiced against blacks than non-southern whites – no evidence for the "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, leaving culture as little more than a small and relatively insignificant residual."
- Need for a multivariate regression analysis

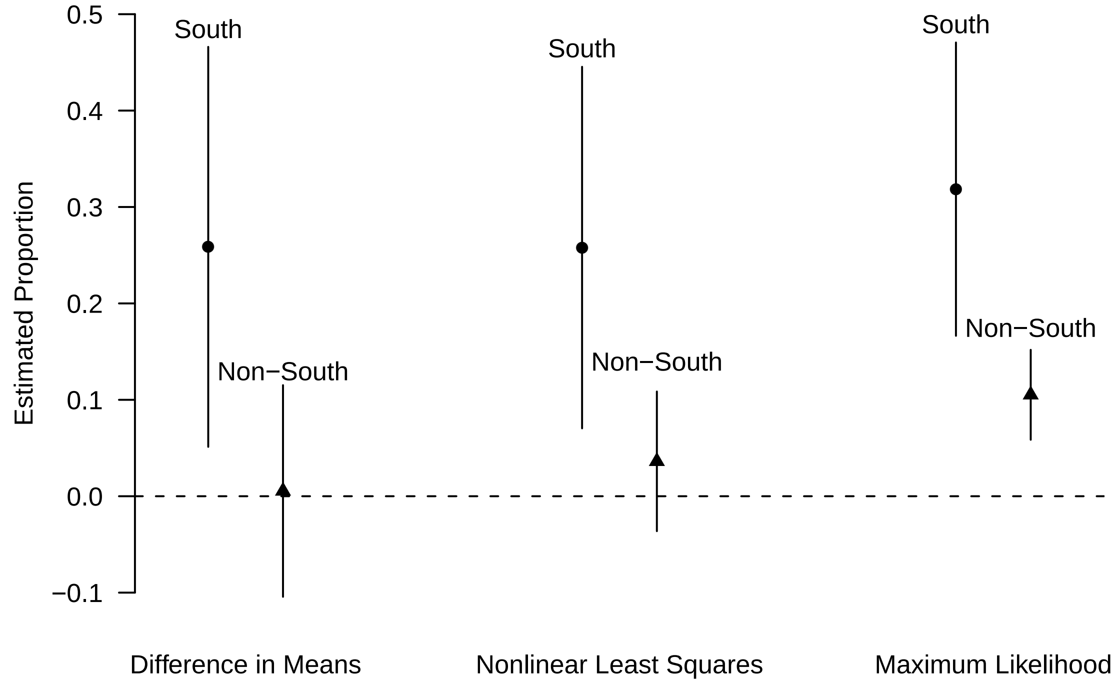
Results of the Multivariate Analysis

- Logistic regression model for sensitive item
- Binomial regression model for non-sensitive item (not shown)
- Little over-dispersion
- Likelihood ratio test supports the constrained model

	Nonlinear Least Squares		**Maximum Likelihood**			
			Constrained		Unconstrained	
Variables	est.	s.e.	est.	s.e.	est.	s.e.
Intercept	−7.084	3.669	−5.508	1.021	−6.226	1.045
South	2.490	1.268	1.675	0.559	1.379	0.820
Age	0.026	0.031	0.064	0.016	0.065	0.021
Male	3.096	2.828	0.846	0.494	1.366	0.612
College	0.612	1.029	−0.315	0.474	−0.182	0.569

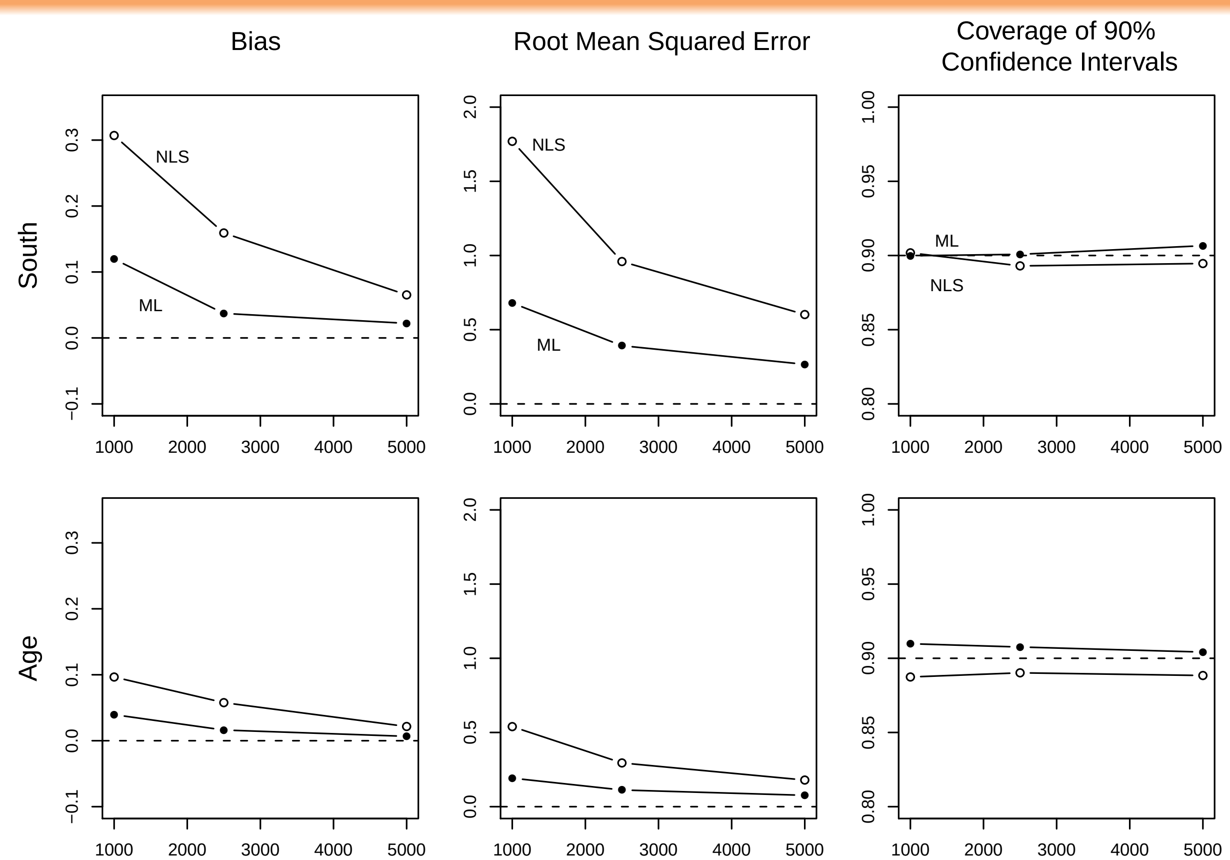
- The original conclusion is supported
- Standard errors are much smaller for ML estimator

Estimated Proportion of Prejudiced Whites



- Regression adjustments and MLE yield more efficient estimates

Simulation Evidence



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 Liar: 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 Detecting List Experiment Failures

- Under the null hypothesis of no design effect and no liar, we

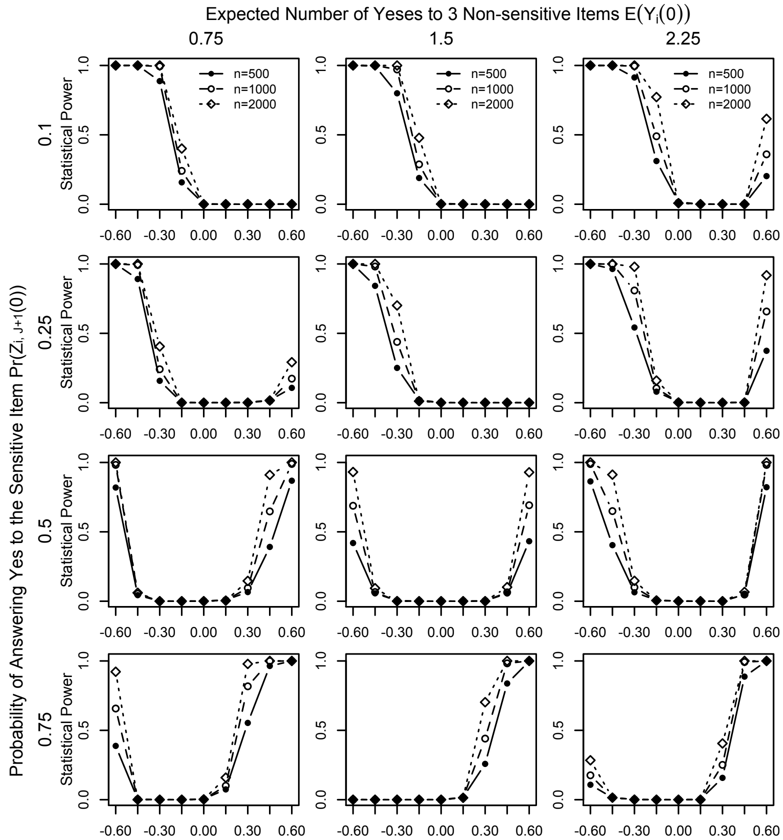
$$\pi_1 = \Pr(\text{type} = (y,1)) = \Pr(Y_i \leq y \mid T_i = 0) - \Pr(Y_i \leq y \mid T_i = 1) \geq 0$$
$$\pi_0 = \Pr(\text{type} = (y,0)) = \Pr(Y_i \leq y \mid T_i = 1) - \Pr(Y_i < y \mid T_i = 0) \geq 0$$

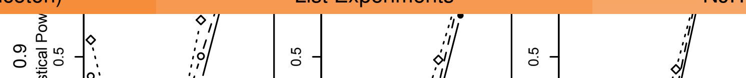
- Alternative hypothesis: At least one is negative
- Test of two stochastic dominance relationships
- A multivariate one-sided LR test for each $t = 0, 1$

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

- $\hat{\lambda}_t$ follows a mixture of χ^2
- Difficult to characterize least favorable values under the joint null
- Bonferroni correction: Reject the joint null if $\min(\hat{p}_0, \hat{p}_1) \leq \alpha/2$
- Failure to reject the null may arise from the lack of power

Statistical Power of the Proposed Test





The Racial Prejudice Data Revisited

- Did the negative proportion arise by chance?

	Observed Data				Estimated Proportion of Respondent Types			
	Control		Treatment					
y value	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	

- Minimum *p*-value: 0.077
- Fail to reject the null with $\alpha = 0.1$

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)

- Previous tests do not detect these liars: proportions would still be positive so long as there is no design effect
- Proposed strategy: model ceiling and/or floor effects under an additional assumption
- Identification assumption: conditional independence between items given covariates
- ML estimation can be extended to this situation

Ceiling and Floor Effects in the Racial Prejudice Data

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`	−5.500	1.023	−5.833	1.129	−5.830	1.125
`male`	0.852	0.496	1.061	0.548	1.069	0.542
`college`	−0.314	0.474	−0.527	0.499	−0.529	0.500
`age`	0.636	0.165	0.719	0.183	0.721	0.179
`south`	1.672	0.559	1.630	0.593	1.626	0.595
Cond. prob. of lying						
Ceiling effects	0.009				0.009	
Floor effects			0.808		0.810	

- Small ceiling effects
- Large floor effects but not many belong to type $(0, 1)$
- Main conclusion appears to be robust

Concluding Remarks

- List experiments: alternative to the randomized response method
- Advantages: easy to use, easy to understand
- Disadvantages:
  1. inefficient
  2. difficult to explore multivariate relationship
  3. the assumptions may be violated
- Our propose methods partially overcome the difficulties
  - multivariate regression analysis for efficient analysis
  - exploration of multivariate relationship
  - statistical tests for detecting list experiment failures
  - modeling ceiling and floor effects
- The importance of design: choice of non-sensitive items

Project Reference

- PAPERS:
  1. Imai. “Statistical Inference for the Item Count Technique.”
  2. Blair and Imai. “Statistical Analysis of List Experiments.”
- SOFTWARE: R package
  Blair, Graeme, and Kosuke Imai. `list: Multivariate Statistical Analysis for the Item Count Technique.`
- PROJECT WEBSITE:

  http://imai.princeton.edu/projects/sensitive.html