

# Scavenging for covariates to use in selectivity assessment and correction strategies

Trudy Ann Cameron  
University of Oregon

Sketch of projects undertaken in collaboration with J.R. DeShazo (UCLA),  
Mike Dennis (KN), Rick Li (KN), and Jason Lee (ICFI)

prepared for workshop on

## Sample Representativeness: Implications for Administering and Testing Stated Preference Surveys

Resources for the Future  
October 2, 2006

## SUMMARY: Concerns and prospects

1. Exogenous weights based on relative frequencies in estimating sample and the population—“look busy”;
2. There are several phases of selection between the initial random contact with potential panelists and final membership in estimation sample—evidence of different processes at each stage
  - need proprietary variables to expand set of regressors at each stage;
3. “Marginal” versus “conditional” selection models for each phase (can be explored even with identical regressors);
4. Comprehensive selectivity (overall selection from RDD contacts to estimating sample) – for “govt” variable;
5. Propensities and probabilities as ad hoc shifters for preferences;

6. Climate study – final-step selection (invited participants to estimating sample): capturing “salience” of survey topic

7. Advocate:

- make maximum use of opportunities to geocode respondents' locations and take advantage of indicators of neighborhood attitudes (related to salience) that can also be geocoded;
- use salience variables to model selection that occurs AFTER potential respondents become aware of the survey topic;
- appropriate variables will be idiosyncratic to each study;
- requires considerable ingenuity and persistence.

8. Near-term research agenda?

## About those weights....

### Common strategy:

1. Define “bins”
2. Calculate relative freq.. in each bin for population
3. Calculate relative freq. in each bin for estimating sample
4. Calculate “exogenous weights”:  
$$\text{rel. freq. (pop)} / \text{rel. freq. (sample)}$$

### Problem:

1. With sample, can create any cross-tabulation that seems to be relevant or important; e.g. age by gender by income by ... by political ideology by avidity concerning resource....?
2. Some cross-tabs are provided by the Census, some are not
3. Cross-tabs involving specialized variables are not available in the Census at all? Harris Interactive (HI)-type benchmarking RDD attitude surveys? Still not enough to capture salience for all possible survey topics.

## **Work-around?**

1. Assume distributions of different population variables are independent
2. Construct “*joint*” distribution by using *products of marginals*
3. Skip desired variables that are not available in Census

## **A few misgivings...**

Usual weights are not going to be “perfect”

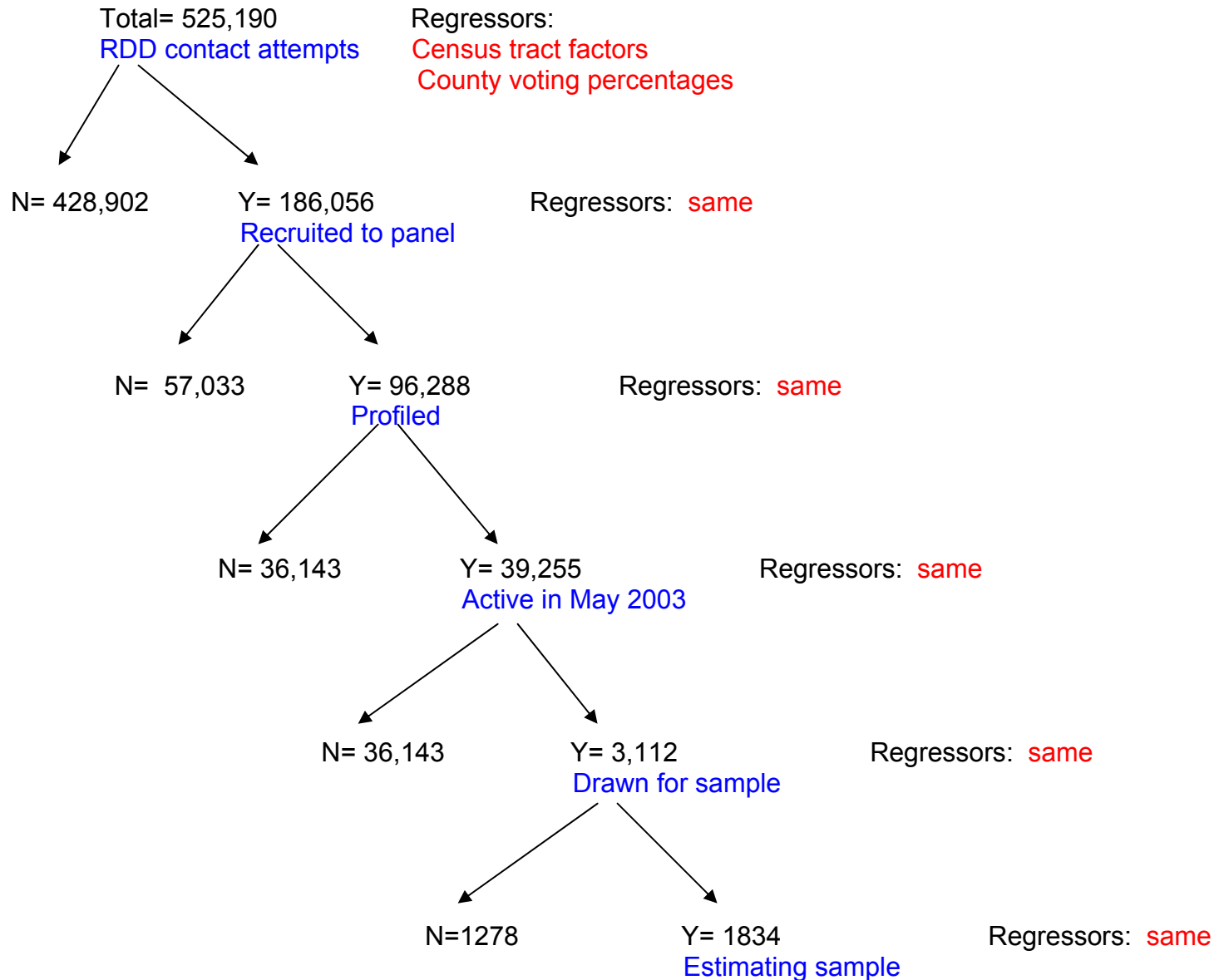
1. not sufficient resolution; some variables unavailable
2. needed Census cross-tabs not necessarily available (may require independence assumption and products of marginals)

## **Well known problem:**

Exogenous weights based on *observables* cannot capture systematic differences in distributions of *unobservables* between population and sample.

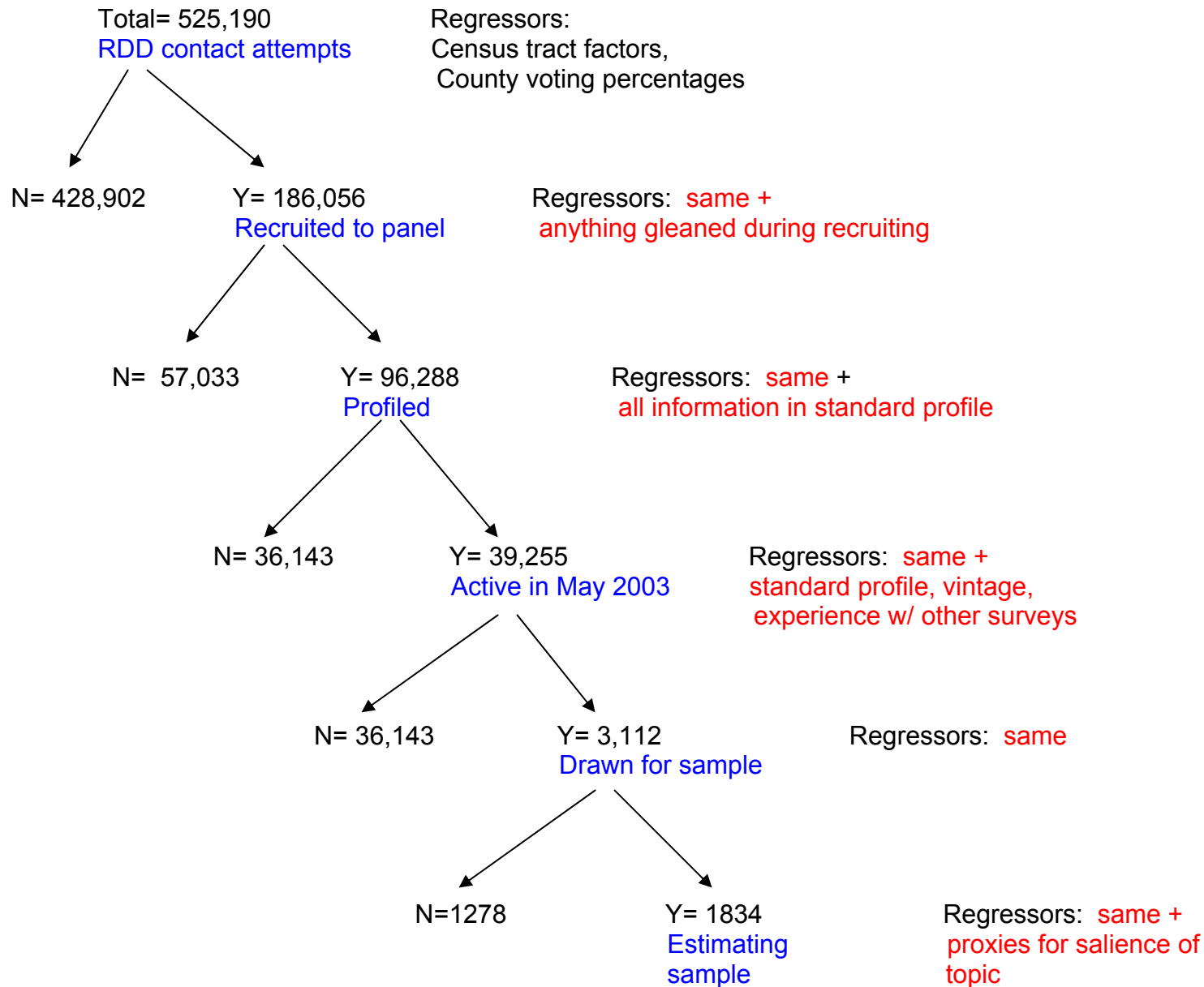
# Exploring the “phases” of selection (identical regressors)

Statistical Lives Project (Knowledge Networks); Cameron and DeShazo (2006)  
“Attrition Patterns in a Major Internet Consumer Panel”



# Exploring the phases of selection – expanding the set of regressors

Statistical Lives Project (Knowledge Networks) – other variables exist



## Marginal versus Conditional Selection Models

“Marginal” models:

1. RDD contacts → Recruited to panel
2. RDD contacts → Profiled
3. RDD contacts → Still active in panel in May 2003
4. RDD contacts → Drawn for survey
5. RDD contacts → Present in estimating sample for model

“Conditional” models:

6. RDD contacts → Recruited to panel
7. Recruited to panel → Profiled
8. Profiled → Still active in panel in May 2003
9. Still active → Drawn for survey
10. Drawn for survey → Present in estimating sample for model

Problem: Most response rate calculations, and most existing selection-correction models address only the transition in 10. We have estimated all ten models in Cameron and DeShazo “Attrition Patterns...”

Table 2 – **Marginal** participation models (n=525,190) from initial RDD contacts  
down to “private preferences” estimating sample)

| Variable                    | recruited             | profiled              | active                | drawn                 | estimating           |
|-----------------------------|-----------------------|-----------------------|-----------------------|-----------------------|----------------------|
| <i>census factor avail.</i> | 0.8850<br>(21.11)***  | 0.6580<br>(13.11)***  | 0.5685<br>(8.17)***   | -0.5344<br>(8.07)***  | -0.6734<br>(9.65)*** |
| well-to-do prime            | -0.0543<br>(29.15)*** | -0.0521<br>(24.13)*** | -0.0953<br>(32.93)*** | -0.1165<br>(15.02)*** | -0.0888<br>(9.42)*** |
| well-to-do seniors          | -0.0538<br>(27.88)*** | -0.0179<br>(8.12)***  | 0.0263<br>(9.34)***   | 0.0106<br>(1.54)      | 0.0352<br>(4.02)***  |
| single renter twenties      | -0.0662<br>(35.91)*** | -0.0474<br>(22.50)*** | -0.0545<br>(19.43)*** | -0.0253<br>(3.56)***  | -0.0318<br>(3.61)*** |
| unemployed                  | -0.0082<br>(4.25)***  | -0.0398<br>(17.84)*** | -0.0246<br>(8.54)***  | -0.0050<br>(0.70)     | -0.0200<br>(2.21)**  |
| minority single moms        | 0.0124<br>(5.63)***   | -0.0521<br>(20.41)*** | -0.0446<br>(13.73)*** | -0.0101<br>(1.32)     | -0.0273<br>(2.74)*** |
| thirty-somethings           | -0.0502<br>(22.74)*** | -0.0368<br>(14.79)*** | -0.0195<br>(5.78)***  | -0.0035<br>(0.39)     | -0.0169<br>(1.57)    |
| working-age disabled        | 0.0193<br>(8.79)***   | 0.0018<br>(0.73)      | 0.0099<br>(3.14)***   | 0.0211<br>(2.81)***   | 0.0069<br>(0.73)     |
| some college, no graduation | 0.0439<br>(22.16)***  | 0.0520<br>(23.06)***  | 0.0376<br>(12.89)***  | -0.0108<br>(1.49)     | -0.0067<br>(0.74)    |
| elderly disabled            | -0.0407<br>(20.96)*** | -0.0337<br>(14.96)*** | -0.0102<br>(3.53)***  | 0.0180<br>(2.62)***   | 0.0241<br>(2.84)***  |
| rural farm self-employed    | 0.0177<br>(5.64)***   | 0.0050<br>(1.38)      | 0.0153<br>(3.38)***   | 0.0439<br>(4.38)***   | 0.0419<br>(3.31)***  |
| low mob. stable neighborhd  | 0.0352<br>(16.32)***  | 0.0171<br>(6.95)***   | 0.0158<br>(5.01)***   | -0.0029<br>(0.37)     | -0.0016<br>(0.16)    |
| Native American             | 0.0409<br>(16.92)***  | 0.0546<br>(19.48)***  | 0.0528<br>(14.59)***  | 0.0343<br>(4.04)***   | 0.0285<br>(2.62)***  |

|                               |                       |                       |                       |                       |                       |
|-------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| Female                        | 0.0038<br>(1.64)      | 0.0015<br>(0.57)      | 0.0099<br>(2.88)***   | 0.0066<br>(0.79)      | 0.0009<br>(0.09)      |
| health-care workers           | 0.0168<br>(7.78)***   | 0.0185<br>(7.50)***   | 0.0202<br>(6.29)***   | 0.0372<br>(4.78)***   | 0.0374<br>(3.82)***   |
| Asian-Hisp-lang isolated      | -0.0310<br>(15.07)*** | -0.0562<br>(23.41)*** | -0.0651<br>(20.49)*** | -0.0454<br>(5.75)***  | -0.0604<br>(5.78)***  |
| <i>Vote percentage avail.</i> | 0.1142<br>(2.47)**    | 0.2026<br>(3.81)***   | 0.2399<br>(3.31)***   | -0.9356<br>(12.12)*** | -1.0114<br>(11.94)*** |
| Gore percent in 2000          | -0.0840<br>(4.34)***  | -0.0430<br>(1.95)*    | -0.0355<br>(1.25)     | -0.1049<br>(1.51)     | -0.2096<br>(2.41)**   |
| Nader percent in 2000         | 0.0132<br>(0.11)      | 0.0914<br>(0.70)      | -0.0277<br>(0.16)     | 0.6544<br>(1.60)      | 0.9018<br>(1.81)*     |
| Constant                      | -1.3075<br>(22.27)*** | -1.7341<br>(24.97)*** | -2.2139<br>(23.02)*** | -0.9995<br>(13.23)*** | -0.9354<br>(12.08)*** |
| Observations                  | 525,190               | 525,190               | 525,190               | 525,190               | 525,190               |

Table 3 - **Conditional** participation models (declining sample sizes);  
from initial RDD contacts down to “private preferences” estimating sample

| Variables                   | recruited             | profiled              | active                | drawn                | estimating          |
|-----------------------------|-----------------------|-----------------------|-----------------------|----------------------|---------------------|
| <i>census factor avail.</i> | 0.8850<br>(21.11)***  | -0.0834<br>(0.87)     | 0.0998<br>(0.76)      | -0.1366<br>(0.49)    | -1.1354<br>(2.51)** |
| well-to-do prime            | -0.0543<br>(29.15)*** | -0.0229<br>(7.42)***  | -0.1053<br>(23.33)*** | -0.0885<br>(7.72)*** | 0.1019<br>(3.42)*** |
| well-to-do seniors          | -0.0538<br>(27.88)*** | 0.0377<br>(11.85)***  | 0.0787<br>(17.44)***  | -0.0273<br>(2.63)*** | 0.0898<br>(3.62)*** |
| single renter twenties      | -0.0662<br>(35.91)*** | 0.0032<br>(1.03)      | -0.0356<br>(8.01)***  | 0.0140<br>(1.28)     | -0.0354<br>(1.33)   |
| unemployed                  | -0.0082<br>(4.25)***  | -0.0587<br>(18.40)*** | 0.0124<br>(2.71)***   | 0.0195<br>(1.83)*    | -0.0476<br>(1.89)*  |
| minority single moms        | 0.0124<br>(5.63)***   | -0.1057<br>(30.02)*** | -0.0069<br>(1.32)     | 0.0292<br>(2.49)**   | -0.0694<br>(2.55)** |
| thirty-somethings           | -0.0502<br>(22.74)*** | 0.0001<br>(0.04)      | 0.0149<br>(2.90)***   | 0.0103<br>(0.75)     | -0.0555<br>(1.58)   |
| working-age disabled        | 0.0193<br>(8.79)***   | -0.0218<br>(6.05)***  | 0.0150<br>(2.97)***   | 0.0253<br>(2.19)**   | -0.0550<br>(2.03)** |
| some coll, no graduation    | 0.0439<br>(22.16)***  | 0.0372<br>(11.45)***  | -0.0076<br>(1.63)     | -0.0516<br>(4.69)*** | 0.0141<br>(0.53)    |
| elderly disabled            | -0.0407<br>(20.96)*** | -0.0069<br>(2.07)**   | 0.0339<br>(7.15)***   | 0.0324<br>(3.02)***  | 0.0230<br>(0.92)    |
| rural farm self-employed    | 0.0177<br>(5.64)***   | -0.0165<br>(3.22)***  | 0.0243<br>(3.30)***   | 0.0680<br>(4.35)***  | -0.0031<br>(0.09)   |
| low mob stable neighborhd   | 0.0352<br>(16.32)***  | -0.0148<br>(4.13)***  | 0.0027<br>(0.53)      | -0.0172<br>(1.46)    | 0.0023<br>(0.08)    |
| Native American             | 0.0409<br>(16.92)***  | 0.0444<br>(11.19)***  | 0.0196<br>(3.34)***   | 0.0096<br>(0.70)     | 0.0219<br>(0.67)    |

|                               |                       |                       |                      |                     |                     |
|-------------------------------|-----------------------|-----------------------|----------------------|---------------------|---------------------|
| female                        | 0.0038<br>(1.64)      | -0.0017<br>(0.44)     | 0.0188<br>(3.42)***  | 0.0104<br>(0.80)    | -0.0177<br>(0.59)   |
| health-care workers           | 0.0168<br>(7.78)***   | 0.0138<br>(3.88)***   | 0.0117<br>(2.30)**   | 0.0408<br>(3.42)*** | -0.0017<br>(0.06)   |
| Asian-Hisp-lang isolated      | -0.0310<br>(15.07)*** | -0.0637<br>(18.78)*** | -0.0439<br>(8.70)*** | 0.0001<br>(0.01)    | -0.0695<br>(2.46)** |
| <i>Vote percentage avail.</i> | 0.1142<br>(2.47)**    | 0.2172<br>(2.93)***   | 0.1336<br>(1.20)     | -0.5793<br>(2.51)** | -0.5950<br>(1.39)   |
| Gore percent in 2000          | -0.0840<br>(4.34)***  | 0.0318<br>(1.01)      | -0.0093<br>(0.21)    | -0.0189<br>(0.18)   | -0.4453<br>(1.75)*  |
| Nader percent in 2000         | 0.0132<br>(0.11)      | 0.2183<br>(1.16)      | -0.1174<br>(0.45)    | 0.4108<br>(0.66)    | 0.7355<br>(0.50)    |
| Constant                      | -1.3075<br>(22.27)*** | -0.1171<br>(0.97)     | -0.4341<br>(2.54)**  | -0.6385<br>(1.79)*  | 2.1218<br>(5.95)*** |
| Observations                  | 525,190               | 186,056               | 96,288               | 35,463              | 3,112               |

Comprehensive selection: “RDD contacts” to “Estimating sample”

From Cameron and DeShazo (2006) “A Comprehensive Assessment of Selection in a Major Internet Panel for the Case of Attitudes toward Government Regulation”

“Outcome” variable is “govt”:

“People have different ideas about what their government should be doing. How involved do you feel the government should be in regulating environmental, health and safety hazards?”

1=minimally involved ..... 7=heavily involved.

The “govt” question was asked in both the “prevention” and the “treatment” version of the “public health policy” surveys

---

### Selectivity equation from jointly estimated model

---

|   |                      |            |
|---|----------------------|------------|
| Census tract factors available <sup>a</sup> | -0.7427              | (12.50)*** |
| “well-to-do prime aged”                     | -0.1185              | (15.00)*** |
| “well-to-do seniors”                        | 0.0324               | (4.62)***  |
| “single renter twenties”                    | -0.0324              | (4.34)***  |
| “unemployed”                                | 0.0062               | (0.85)     |
| “minority single moms”                      | -0.0194              | (2.37)**   |
| “thirty-somethings”                         | -0.0181              | (2.13)**   |
| “working-age disabled”                      | -0.0013              | (0.17)     |
| “some college, no graduation”               | -0.0183              | (2.44)**   |
| “elderly disabled”                          | 0.0055               | (0.76)     |
| “rural farm self-employed”                  | 0.0425               | (4.09)***  |
| “low mobility stable neighborhood”          | -0.0094              | (1.18)     |
| “Native American”                           | 0.0407               | (4.59)***  |
| “female”                                    | 0.0112               | (1.29)     |
| “health-care workers”                       | 0.0192               | (2.33)**   |
| “Asian-Hisp language isolated”              | -0.0666              | (7.75)***  |
| 2000 vote percentage available              | -1.1264              | (15.32)*** |
| Gore percent (county)                       | -0.1769              | (2.45)**   |
| Nader percent (county)                      | 1.6150               | (3.93)***  |
| Constant                                    | -0.6224              | (9.02)***  |
| $\rho$ (implied Heckman error correlation)  | 0.08462 <sup>b</sup> | (1.52)     |

---

## Why not use Heckman correction models for every study?

Cameron and DeShazo (2006) uses “govt” variable—allows us to use packaged selective correction algorithms (in Limdep: ordered probit with selection correction; or Stata, if we treat the govt variable as continuous)

More typical stated preference estimation context:

Conjoint choice experiments—multiple-alternative conditional logit (or analogous)

Problem:

No easy FIML selectivity correction models for these models

Potential solutions:

Jointly estimated mixed logit model and selection equation. At a minimum, allow “intercept” in selection equation sub-model to be correlated with parameter that gives “intercept” in valuation sub-model.

My one attempt at such a model was “uncooperative.”

Dan Hellerstein’s work?

## What might we do instead? (...or, meanwhile?)

Estimated selection equation provides:

1. Fitted participation “propensities”:

$$P_i = z_i' \gamma$$

2. Fitted participation “probabilities”:

$$\Pi_i = \Phi(z_i \gamma)$$

- Point estimates for  $P_i$  and  $\Pi_i$  can be produced for every initial RDD recruiting contact (e.g.  $N = 525,000+$  for KN).
- Calculate central tendency across set of RDD contacts.
- Calculate *deviations* from this “average” across all RDD recruiting contacts; call these  $p_i$  and  $\pi_i$ .
- If every RDD recruiting contact was equally likely to show up in the estimating sample,  $P_i$  and  $\Pi_i$  would be same for all, and  $p_i = 0$  and  $\pi_i = 0$  for all  $i$ .
- Want to be able to simulate the estimation model under these more-desirable conditions

## One strategy for assessing the potential effects of selection

Q: Do the estimated preference parameters—within the estimating sample—vary systematically with the participation propensities (participation probabilities)?

Generalize each preference parameter  $\beta_j = \beta_{j0} + \beta_{j1}p_i, \forall i$  or  $\beta_j = \beta_{j0} + \beta_{j1}\pi_i, \forall i$

Model predicts that “true” preference parameters—i.e. if every observation in the estimating sample was equally likely to be present—will be give by the baseline parameter  $\beta_{j0}$ .

- If shift parameters  $\beta_{j1}$  *are not* significantly different from zero, this may suggest that the lack of any relationship may extend to the non-participants among the RDD contact (but this is not proof).
- If shift parameters  $\beta_{j1}$  *are* statistically significant, researchers should be concerned that preferences also differ with response propensities between the participant and non-participant subsets.

## Selection models for “Value of a Statistical Illness Profile” study

Classes of explanatory variables:

- Some are general response-rate predictors
  - Some are specific to the salience of health issues
1. 15 Census tract factors (including “health care workers” factor)
  2. County voting percentages in 2000 Presidential election
  3. Number of hospitals in county
  4. Actual mortality from selected causes (in same county over previous 12 years)

Fitted comprehensive participation probability ( $\pi_i$ ) significantly affects only the marginal utility of the log of the present value of future sick-years. Shift is about 3 on a base parameter of about -50. Mean value of probability difference is also *far* less than one.

Model for Conditional Selection Process #10: transition from the “drawn” set of target households to the “estimating” sample

Dredging up proxies for the salience of the survey topic

Example: **climate policy survey** (<http://globalpolicysurvey.ucla.edu>)

Q: What variables might capture the *salience* of climate change impacts AND can be *linked* to each potential respondent in some way

### **Current versus historical temperatures**

Avg temp, mailday+5 in zipcode

Positive temp dev. from normal

Negative temp dev. from normal

### **Perception of vulnerability to climate change**

Insurance payments in state, 78-03

Natural disaster (same state)

Natural disaster (neighboring state)

=1 if county flooding during 1998-2000

County with flooding in 2000

County with flooding in 1999

Geographic risk data available (indicator)  
Tornados recorded >6/1000 sq. mi.  
Tornados recorded >11/1000 sq. mi.  
Tornados recorded >15/1000 sq. mi.  
<100 mi from coast, South hurricane zone  
<100 mi from coast, North hurricane zone  
1/(Distance from ocean)

### **Season when survey was completed**

January  
February  
March  
April  
May  
June

July  
August  
September  
October  
November  
December

### **Major events**

Survey mailed 9/11/01-10/9/01 (World Trade Center)  
Mailed to anthrax localities  
Major sporting event

### **Features of survey instrument**

Design value for program cost

Replacement packet sent  
Included a post-it note  
=1 if shortened survey version

### **County political mix**

County Nader vote percentage  
County Bush vote percentage  
County voter turnout percentage

### **Attributes of address/addressee**

Midpoint of addressee age bracket  
Midpoint addressee family inc. bracket/1000  
Addressee household size  
Census tract median house value  
Addressee length of residence  
1=rural, 0=urban  
Addressee female  
Address is an apartment  
Address is a condominium  
Address is a house

## **Neighborhood characteristics Census factor scores**

Well-to-do prime  
Well-to-do seniors  
Single renter twenties  
Unemployed  
Minority single moms  
Thirty-somethings  
Working-age disabled  
Some college, no grad  
Elderly disabled  
Rural, farm, self-empl  
Low mobility older neighb  
Native American  
Asian-Hisp. language isol.

---

**Selection portion of climate model (persistent variables)**

---

| <b>Features of survey instrument</b>           | Coef.   | t-stat    |
|--|---------|-----------|
| =1 if shortened survey version                 | 0.2084  | (2.74)*** |
| <b>Vulnerability</b>                           |         |           |
| Natural disaster (same state)                  | -0.1006 | (1.06)    |
| Geographic risk data available                 | 0.1622  | (1.82)*   |
| Tornados recorded >11/1000 sq. mi.             | -0.1586 | (1.89)*   |
| <100 mi from coast, North hurricane zone       | -0.1675 | (3.50)*** |
| <b>Seasonality</b>                             |         |           |
| September                                      | -0.1264 | (2.19)**  |
| October  | -0.0333 | (0.65)    |
| <b>County political mix</b>                    |         |           |
| County Nader vote                              | 1.7355  | (1.71)*   |
| <b>Attributes of address/addressee</b>         |         |           |
| Midpoint of addressee age bracket              | 0.0106  | (2.07)**  |
| (Midpoint of addressee age bracket)/100        | -0.0100 | (2.03)**  |
| Midpoint of addressee family inc. bracket/1000 | 0.9575  | (1.96)**  |
| Continued....                                  |         |           |

Continued....

**Census tract factor analysis scores**

|  |          |            |
|--|----------|------------|
| Well-to-do prime                                     | 0.0716   | (3.87)***  |
| Well-to-do seniors                                   | 0.0690   | (3.56)***  |
| Unemployed   | -0.0750  | (3.92)***  |
| Minority single moms                                 | -0.1113  | (4.82)***  |
| Working-age disabled                                 | -0.0513  | (2.48)**   |
| Some college, no grad                                | 0.0599   | (3.15)***  |
| Native American                                      | 0.0610   | (2.38)**   |
| Asian-Hisp. language isol.                           | -0.0664  | (3.16)***  |
| Constant   | -1.3036  | (11.69)*** |
| <b>Estimated error correlation (rho)<sup>a</sup></b> | -0.089   | -0.36      |
| <hr/>  |          |            |
| Observations   | 7527     |            |
| Log L  | -4806.24 |            |
| <hr/>  |          |            |

## Research Agenda

1. Figure out some easily implemented models that allow us to mimic Heckman selectivity correction models in the context of multiple conditional logit estimation (e.g. Stata clogit).
2. Find everything you can that might account for differences in the salience of your specific survey topic across individuals
3. There is “deeper” systematic selection in going from the original RDD contact attempts to the set of active panelists in the KN sample at any point in time.
  - Whether these common phases of selection matter for a specific survey is an empirical question to be asked for every study.
  - A finding of minimal evidence of selection bias for one study using KN is encouraging, but does not yet guarantee that no other study will be affected by selectivity biases.
  - Need enough studies to produce a “weight of the evidence.”