# Do-it-yourself multiple imputation: Mode-effect correction in a public opinion survey

Stas Kolenikov

Abt SRBI

Stata Conference 2014

# Motivation

- Are you more likely to admit illicit drug use to a stranger in a personal interview, or over the Internet anonymously?
- When an interviewer reads response options to you over the phone, do you still remember the first one when they are done with a long list?
- Are you more likely to provide an open-ended response on the phone, for the interviewer to enter it, or type it in the web survey?
- Do you always scroll down for the long list of response options when doing a survey on your smartphone?

These are all examples of mode effects present in human population surveys collected over several modes.

Methodology reference: Kolenikov and Kennedy (2014)

# Study: Rice University PALS

Portraits of American Life Study (PALS):

- Second wave of data collection (2012)
- 1,879 items in the instrument, 363 analytic variables, 1,418 observations
- Survey modes:
  - Web mode as the primary mode of data collection
  - Phone mode for non-response follow-up (e.g., no Internet access)
  - Built-in methodological experiment: 13% of cases randomized into phone, no web mode offered

`http://www.palsresearch.org/`

# Project objectives

1. Identify variables that suffer from mode effects
2. Adjust for mode effects, if possible
3. Provide methodologically correct inference for the adjusted data, if possible

# How can we adjust for mode effects?

# Existing methods

- Ostrich method: ignore mode effects, pool data across modes
- Report only, do not adjust: cross-tabulate response by mode, eye-ball the extent of differences
- Regression adjustment (Elliott et al. 2009): run a regression with explanatory variables including `i.mode`, report `margin mode` for the reference mode
- Missing data problem:
  - Unobservable counterfactuals (as in causal inference literature, Morgan and Winship (2007))
  - Measurement error, multiple imputation (Powers et al. 2005)

# Proposed adjustment

- Implied utility of the response of person $i$ to item $j$:

$$y^*_{ij} = \beta' x_i + \gamma m_i + \epsilon_{ij}$$
$$\epsilon_{ij} \sim \Lambda(\epsilon)$$
$$y_{ij} = \mathbb{I}[y^*_{ij} \geq 0]$$

$x_i$ = demographic variables, $m_i$ = mode (0=web, 1=phone)

- Estimate on the survey data
- Simulate for $m_i = 1$ without the mode effect $\hat{\gamma}$:

$$\tilde{\epsilon}_{ij} \sim \Lambda(\epsilon | \epsilon > -\hat{\beta}' x_i - \hat{\gamma} m_i), \quad y_{ij} = 1$$
$$\tilde{\epsilon}_{ij} \sim \Lambda(\epsilon | \epsilon < -\hat{\beta}' x_i - \hat{\gamma} m_i), \quad y_{ij} = 0$$
$$\tilde{y}^*_{ij} = \hat{\beta}' x_i + \tilde{\epsilon}_{ij}$$
$$\tilde{y}_{ij} = \mathbb{I}[\tilde{y}^*_{ij} \geq 0]$$

# Proposed adjustment

- Implied utility of the response of person $i$ to item $j$:

$$y^*_{ij} = \beta' x_i + \gamma m_i + \epsilon_{ij}$$
$$\epsilon_{ij} \sim \Lambda(\epsilon)$$
$$y_{ij} = \mathbb{I}[y^*_{ij} \ge 0]$$

$x_i$ = demographic variables, $m_i$ = mode (0=web, 1=phone)

- Estimate on the survey data
- Simulate for $m_i = 1$ without the mode effect $\hat{\gamma}$:

$$\tilde{\epsilon}_{ij} \sim \Lambda(\epsilon | \epsilon > -\hat{\beta}' x_i - \hat{\gamma} m_i), \quad y_{ij} = 1$$
$$\tilde{\epsilon}_{ij} \sim \Lambda(\epsilon | \epsilon < -\hat{\beta}' x_i - \hat{\gamma} m_i), \quad y_{ij} = 0$$
$$\tilde{y}^*_{ij} = \hat{\beta}' x_i + \tilde{\epsilon}_{ij} + 0 m_i$$
$$\tilde{y}_{ij} = \mathbb{I}[\tilde{y}^*_{ij} \ge 0]$$

# Mode effect adjustment as `mi` problem

Single imputation suffers from random noise, hence...

1. Add estimation noise ( `_se[`$\hat\gamma$`]` )
2. Impute conditional residual $\tilde\epsilon$
3. Repeat 1–2 for $m = 1, \dots, M$
4. Analyze the data accounting for complex survey structure (weights, clusters, ... )
5. Combine analyses with the imputed responses using Rubin's multiple imputation rules

# What do we need to adjust?

# Significant mode effects: pt 1

Survey data analysis part:

```
foreach x of varlist outcomes {
 svy : tab `x' mode
 post summary1 ("`x'") (p-value)
}
```

Detecting signal with FDR (Benjamini and Hochberg 1995):

```
use summary1, clear
sort p-value
levelsof outcome if p-value < 0.10*_n/_N
push r(levels) back to the caller
```

# Significant mode effects: pt 2

Survey data analysis part:

```
foreach x of varlist outcomes {
 svy : logit `x' demographics mode
 post summary2 ("`x'") (p-value)
}
```

Detecting signal with FDR (Benjamini and Hochberg 1995):

```
use summary2, clear
sort p-value
levelsof outcome if p-value < 0.10*_n/_N
push r(levels) back to the caller
```

# Mode effect adjustment

```
svy : logit outcome demographics mode
```

```
predict utility, xb
```

`gen epsilon = invlogit(U) if outcome == 1`,

$U \sim U(\Lambda^{-1}($ `-utility` $), 1)$

`replace epsilon = invlogit(U) if outcome == 0`,

$U \sim U(0, \Lambda^{-1}($ `-utility` $))$

```
gen adj_utility = utility - (_b[mode] + rnormal()*_se[mode])
```

```
gen adj_outcome = ( adj_utility + epsilon > 0 )
```

# How do I trick Stata to `mi`?

# Introduction to multiple imputation

Little and Rubin (2002):

1. Come up with a solid univariate or joint distribution of the missing values
2. Impute independently $m = 1, \dots, M$ times
3. Estimate the model of interest, obtain estimates $\hat{\theta}^{(m)}$ and their variances $v^{(m)}$
4. Post point and variance estimates:

$$\hat{\theta}_{\mathrm{MI}} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}^{(m)}$$

$$v_{\mathrm{MI}}[\hat{\theta}_{\mathrm{MI}}] = \frac{1}{M} \sum_{m=1}^{M} v^{(m)} + \left(1 + \frac{1}{M}\right) \frac{1}{M-1} \sum_{m=1}^{M} (\hat{\theta}^{(m)} - \hat{\theta}_{\mathrm{MI}})^2$$

# Introduction to `mi`

Implemented in Stata via official `mi`, user-written `ice+mim` (Royston 2005)

1. Declare data to contain multiple imputations: `mi set` *`style`*
2. Declare the variables to be imputed or retained as is: `mi register`
3. Impute the missing values: `mi impute` *`method`*
4. Combine the results: `mi estimate:` *`command`*

I am trying to hack Step 3.

# mi set wide

My favorite style is `mi set wide`:

- Single data file (vs. multiple files in `mi set flongsep`)
- Imputations for variable `x` are stored as `_1_x, _2_x, ...` in the same observation (vs. additional observations in `mi set flong` or `mi set mlong`)
- Observations with missing values are tagged with the `mi` system variable `_mi_miss`

```
local M = 20
generate mi_outcome = outcome if mode=="web"
mi set wide
mi set M = `M'
mi register imputed mi_outcome
forvalues m=1/`M' {
 do Slide12.do
 replace _`m'_mi_outcome = adj_outcome if mode=="phone"
}
* verify internal consistency:
mi update
```

# What did we get?

# Substantive findings

297 variables ⇒ 19 with significant Rao and Scott (1981) cross-tabs ⇒ 16 with sufficient sample size ⇒ 4 with significant regression effects

- In the past 12 months have you helped directly by giving some of your time to close family?
- In the past 12 months have you helped directly by giving some of your time to neighbors?
- In the past five years, have you had a major financial crisis?
- Number of persons outside your home that you feel closest to (continuous)

# Magnitudes of mode effects

| Variable | Unadjusted | | With corrections | |
|---|---|---|---|---|
| | Estimate | Std. err. | Estimate | Std. err |
| Helped family | 77.1% | (1.6%) | 74.4% | (2.0%) |
| Helped neighbors | 38.8% | (2.0%) | 35.5% | (2.3%) |
| Financial crisis | 32.9% | (2.3%) | 35.1% | (2.7%) |

Relative bias: $\sim 6.3\%$

Relative increase in the standard error of the estimate: $\sim 20.3\%$

# Are we there yet?

# New stuff

- New method for mode effect adjustment
  - utility concept from microeconomics
  - extensions to ordinal models
  - adjusts point estimates as much or more than other methods (considered good)
  - adjusts standard errors in a believable way
- Workflow: 8 do-files, 2 ado-files, $\sim$ 36kbytes / $\sim$ 1000 lines of code
  - cycles over variables to be tested for mode effects
  - multiple testing corrections are incorporated
  - creating and passing to and fro the lists of variables with detected mode effects
- A complete implementation of custom multiple imputations

# If I were doing this today...

I would have:

- ...used Robert Picard's `project`
- ...used `char _dta[]` to exchange variable lists instead of `c_local`

# Questions?

**Thank you!**

Contact: kolenikovs@srbi.com

# References I


Benjamini, Y. & Hochberg, Y. (1995), 'Controlling the false discovery rate: A practical and powerful approach to multiple testing', *Journal of the Royal Statistical Society. Series B (Methodological)* **57**(1), 289–300.

Elliott, M. N., Zaslavsky, A. M., Goldstein, E., Lehrman, W., Hambarsoomians, K., Beckett, M. K. & Giordano, L. (2009), 'Effects of survey mode, patient mix, and nonresponse on CAHPS hospital survey scores.', *Health services research* **44**(2 Pt 1), 501–518.

Kolenikov, S. & Kennedy, C. (2014), 'Evaluating three approaches to statistically adjust for mode effects', *The Journal of Survey Statistics and Methodology* **2**(2), 126–158.

Little, R. J. A. & Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, Wiley Series in Probability and Statistics, 2nd edn, Wiley, New York.

# References II


Morgan, S. L. & Winship, C. (2007), *Counterfactuals and Causal Inference: Methods and Principles for Social Research (Analytical Methods for Social Research)*, Cambridge University Press.

Powers, J. R., Mishra, G. & Young, A. F. (2005), 'Differences in mail and telephone responses to self-rated health: Use of multiple imputation in correcting for response bias', *Australian and New Zealand Journal of Public Health* **29**, 149154.

Rao, J. & Scott, A. (1981), 'The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables', *The Journal of the American Statistical Association* **76**, 221–230.

Royston, P. (2005), 'Multiple imputation of missing values: update', *Stata Journal* **5**(2), 188–201.