A Partially Successful Attempt to Integrate a Web-Recruited Cohort into an Address-Based Sample

Presenter: Phillip S. Kott

Collaborators: Matthew Farrelly and Kian Kamyab
with an assist from Joe McMichael

$$\sum_R d_k \, [1 + \exp(\mathbf{m}_k{}^T \mathbf{g})] \mathbf{c}_k = \mathbf{T}_\mathbf{c}$$

⇑	⇑	⇑
Model variables	**Calibration variables**	**Calibration targets**

Overview

- **The Oregon Marijuana Study: an address-based sample (ABS) supplemented by Facebook recruits** (preliminary results)
- **Adjusting the ABS respondent sample for selection bias …**
- **While at the same time, calibrating the Facebook recruits to the Internet respondents of the ABS sample**
- **Testing whether the previous step was appropriate**
- **Creating analysis weights and delete-a-group jackknife replicate weights to do estimation**
- **Some concluding remarks**

The Oregon Marijuana Study

An ABS of one adult per Oregon household in 2015 was given a 20-minute questionnaire on marijuana use and attitudes.

Roughly half responded via mail, half Internet

More responses were recruited via Facebook.

Poor response on race and household size questions.

How can we weight the result to draw inferences? (Question was not asked until after the data was collected)

Potential Calibration Variables

Sample size – 1,989
(mail response – 722; mail-to-web – 640; recruit – 627;
a respondent needed to give age, sex, or education level)

Missing number of adults in household – over 800

Missing race = black – over 1,300

Used to calibrate the ABS sample to the population

Missing Age group (six levels) – 3
Missing Sex – 76
Missing Education (three levels) – 173

Added to calibrate recruit cohort to mail-to-web cohort

In politics TODAY, do you consider yourself ….
Republican, Democrat, Independent,
No preference,
No or invalid answer (*treated as a separate level*)

The Selection Model

The probability that an Oregon adult was sampled and then responded to the ABS survey is assumed to be a logistic function of three categorical variables: age group, sex, and education level. (Better would be to assume only a probability of response, if the probabilities of selection were known)

The probability that an Oregon adult was recruited into the sample via Facebook is assumed to be a logistic function of the above three categorical variables and party affiliation.

The population that would respond by Internet when given the chance (represented by the mail-to-web cohort) is assumed to be the same as the population that could be recruited via Facebook. *An assumption that will be tested.*

SUDAAN : WTADJX

WTADJX implements calibration weighting *allowing the model (*MODEL) *and calibration* (CALVARS) *variables to differ*.

In our case, response for the ABS sample is a function of the categorical (CLASS) calibration variables with Oregon population targets (POSTWGT).

Response to Facebook recruitment is a function of categorical model variables having the same target totals as internet respondents to the ABS survey.

If these variables are multiplied by 1 for Facebook recruits and by -1 for ABS internet respondents, then they form calibration variables with target totals equal to 0.

SAS/SUDAAN Code

Recruit cohort: TYPE = 1; X = 1; Z = 1; ABS = 0

Mail-to-web cohort: TYPE = 2; X = 0; Z = -1; ABS = 1

Mail cohort: TYPE = 3; X = 0; Z = 0; ABS = 1

```
PROC WTADJX DATA = D ADJUST = POST  DESIGN = WR;
WEIGHT _ONE_;  NEST _ONE_;  LOWERBD 1;  VAR [ ….];
CLASS SEX AGE EDU PARTY;   * after imputing missing values;
MODEL _ONE_ = SEX*ABS  AGE*ABS  EDU*ABS SEX*X  AGE*X
EDU*X PARTY*X/NOINT;   (NOINT = no intercept)
CALVARS SEX*ABS  AGE*ABS  EDU*ABS SEX*Z  AGE*Z  EDU*Z
PARTY*Z/NOINT;
POSTWGT [population totals for the categories,  16 zeroes];
VDIFFVAR TYPE (1,2);
```

SAS/SUDAAN Code

DESIGN = WR (with replacement);

ADJUST = POST (outside targets);

WEIGHT _ONE_ (starts with weights = 1);

NEST _ONE_ (no clusters or strata);

LOWERBD 1 (adjustment factor never less than 1);

Find the **g** such that: $\sum_R d_k \; [1 + \exp(\mathbf{m}_k{}^T \mathbf{g})] \; \mathbf{c}_k = \mathbf{T}_\mathbf{c}$

⇑	⇑	⇑	⇑	⇑
Starting weight (=1)	**LOWERBD**	**MODEL variables**	**CALVAR variables**	**POSTWGT targets**

VDIFFVAR TYPE (1,2) (difference between estimated means for TYPEs)

WTFINAL is the calibrated weight

Inverse of bracketed term is the estimated probability of selection.

Variables for VAR Statement

Ordered response when item response;
Whether there was an item response

Do you now smoke **cigarettes**

Do you now smoke **electronic cigarettes**

Do you now drink **alcohol**

- ☐ 1 Every day
- ☐ 2 Some days
- ☐ 3 Rarely
- ☐ 4 Not at all

When you drink, **how many drinks** do you usually have?

- ☐ 1 One
- ☐ 2 Two
- ☐ 3 Three
- ☐ 4 Four or more

Variables for VAR Statement

What is **your opinion** about **legalizing** the use of marijuana by **adults**? (*used for testing*)

What do **most people in your state** think about legalizing the use of marijuana use by **adults**?

- ☐ 1 It should not be legal for any purpose
- ☐ 2 It should be legal only for medical use
- ☐ 3 It should also be legal for recreational use

What is **your opinion** about the use of marijuana by **adults**

What is **your opinion** about the use of marijuana by **teenagers**?

- ☐ 1 It is okay to use every day
- ☐ 2 It is okay to use some days
- ☐ 3 It is not okay to use at all

Variables for VAR Statement

Would it bother you if people were smoking marijuana **in public**?

In your opinion …

should people be allowed to use **edible marijuana** in places they are not allowed to smoke it?

is **edible marijuana**, such as food or candy, **safer** to use than marijuana that is smoked?

is **vaping marijuana**, such as through an e-cig or e-vaporizer device, **safer** than smoking marijuana in a joint or pipe?

does legalization of **medical marijuana** lead to more teenagers trying marijuana?

does the legalization of **recreational marijuana** lead to more teenagers trying marijuana?

- ☐ 1 Definitely yes
- ☐ 2 Probably yes
- ☐ 3 Probably not
- ☐ 4 Definitely not

Variables for VAR Statement

Have you ever tried **marijuana**, even one time?

In your opinion, does the legalization of **recreational marijuana** lead to more people driving under the influence of marijuana?

Do you think people convicted of possessing more than an allowable amount of **marijuana** should serve **time in jail**?

Are you aware of any stores or shops in or near your community that sell **marijuana**?

- ☐ 1 Yes
- ☐ 2 No

Now that **recreational marijuana** is legal in Oregon, will your usage…

- ☐ 1 Increase
- ☐ 2 Stay the same
- ☐ 3 Decrease

Why Party Affiliation?

Before Calibration Weighting:

	Type			
Party Affiliation	**Facebook Recruit**	**Mail-to-Web**	**Mail**	**Total**
No answer	14.83	4.06	6.09	
Republican	13.88	17.97	22.02	
Democrat	25.52	33.75	29.78	
Independent	18.82	22.81	21.47	
No preference	26.95	21.41	20.64	
Total	627	640	722	1989

Party Affiliation

After Calibration Weighting:

	Type			
Party Affiliation	**Facebook Recruit**	**Mail-to-Web**	**Mail**	**Total**
No answer	2.88	2.88	5.25	
Republican	17.62	17.62	21.59	
Democrat	27.98	27.98	26.42	
Independent	24.05	24.05	20.81	
No preference	27.47	27.47	25.94	
Total	1531798	1531798	1579221	4642817

WTADJX's linearization variance estimator

Ignoring finite population correction

$$\text{var}\left(\frac{\sum_R w_k y_k}{\sum_R w_k}\right) = \quad \text{, where } w_k = d_k(1+\exp(\mathbf{m}_k{}^T\mathbf{g})) \text{ is the calibrated weight}$$

$$\text{var}\left(\frac{\sum_R w_k e_k}{\sum_R w_k}\right) \quad \text{, where } e_k = y_k - \mathbf{c}_k{}^T\mathbf{b},$$

$$\text{and } \mathbf{b} = \left[\sum_R d_j \exp(\mathbf{m}_j{}^T\mathbf{g})\mathbf{m}_j\mathbf{c}_j{}^T\right]^{-1}\sum_R d_j \exp(\mathbf{m}_j{}^T\mathbf{g})\mathbf{m}_j y_j$$

Treat the $w_k \approx d_k(1+\exp(\mathbf{m}_k{}^T\boldsymbol{\gamma}))$ like design weights in a linearization variance estimator ($\boldsymbol{\gamma}$ is the selection-model parameter estimated by $\mathbf{g}$)

Holm-Bonferroni Procedure

The conservative HB procedure is not only a overall multiple comparison test but also assesses each individual comparison.

Sort the 20 (or 40) differences by their *p*-values.

For HB20_.1 (*as an example*):

Difference with lowest *p*-value out of 20 is significant at .1 level if *p*-value is less than HB20_.1 critical value (.1/20).

Difference with second lowest *p*-value is significant at .1 level if *p*-value is less than HB20.1 critical value (.1/19).

Continue until first not-significant difference.

Smallest *p* Values vs Critical Holm-Bonferroni Values

VARIABLE	Estimated difference	*p* value	Critical HB40_.1 HB20_.05	Critical HB20_.1	Critical HB40_.05
More DUI?	0.11	0.00247	0.00250	0.00500	0.001000
Edible MJ in public?	-0.23	0.00371	0.00256	0.00526	0.001026
How legal?	0.11	0.00658	0.00263	0.00556	0.001053
Adult frequency?	-0.13	0.01619	0.00270	0.00588	0.001081
Is edible MJ safer?	-0.17	0.02260	0.00278	0.00625	0.001111
Guest use in home?	-0.18	0.04079	0.00286	0.00667	0.001143
Is vaping safer?	0.10	0.05260	0.00294	0.00714	0.001176
More teenage use?	0.12	0.08722	0.00303	0.00769	0.001212
Response to vaping Q	0.05	0.09704	0.00313	0.00833	0.001250

Jackknife Weights (from Kott 2006)

Randomly sort ABS and recruit respondent samples.

Systematically assign respondents to one of 30 jackknife groups.

Create the r^{th} set of jackknife replicate weights by setting the replicate weights of respondents in the r^{th} group to zero and multiply the calibrated weight for respondents outside the group by 30/29.

Recalibrate each replicate without a *lowerbd*.

Scale the calibrated and jackknife weights assigned to mail-to-web (by .65) and recruit (by .35) cohorts to eliminate double counting.

Standard-Error Results (ignoring fpc)

Computing the standard errors of the 40 differences with jackknife weights (and DIFFVAR) rather than through WTADJX increased SE measures by 4.8% on average ($\log(SE_{JK}/SE_{WTADJX})$);
6.0% median, interquartile range from 1.0% to 12.1%.

This is consistent with theory (linearization tends to underestimate calibrated estimates' SEs; replication to overestimate)

Incorporating the recruit cohort into the ABS sample decreased SEs by 8.6% on average (comparing jackknife SE to jackknife SE);
7.5% median, interquartile range from 4.2% to 12.1%.

Using a more traditional jackknife (which is more likely to fail to calibrate) returns nearly the same results.

Some Concluding Remarks

Think about analysis before data are collected.

Using nonprobability samples relies on assumptions, which need to be clearly stated and tested when possible.

Selection modeling is analogous to nonresponse modeling.

An estimated difference not being statistically significant does not mean the actual difference is 0.

When appropriately calibrated (using WTADJX or an equivalent program in R) the decrease in SE from adding nonprobability samples is less than the sample-size increase implies.

Useful References

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 65–70.

Kott, P. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 133–142.

RTI International (2012). *SUDAAN Language Manual, Release 11.0.* Research Triangle Park, NC: RTI International.

Singh, A., Dever, J., and Iannacchione, V. (2004). Efficient estimation for surveys with nonresponse follow-up using dual-frame calibration. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 3919–3930.

Tille, Y. and Matei, A., (2013). *Package 'Sampling.'* A software routine available online at http://cran.r-project.org/web/packages/sampling/sampling.pdf.