Symposium on Data Science and Statistics (SDSS) 2019, Bellevue, WA

An R Package for Linear Mediation Analysis with Complex Survey Data

Yujiao Mai, Jiahui Xu, Deo Kumar Srivastava, Hui Zhang
Department of Biostatistics

St. Jude Children's Research Hospital
Finding cures. Saving children.

5/31/2019

Outline

1. Classic Mediation Analysis
2. Complex Surveys
3. Problem & Solution
4. Software & Application

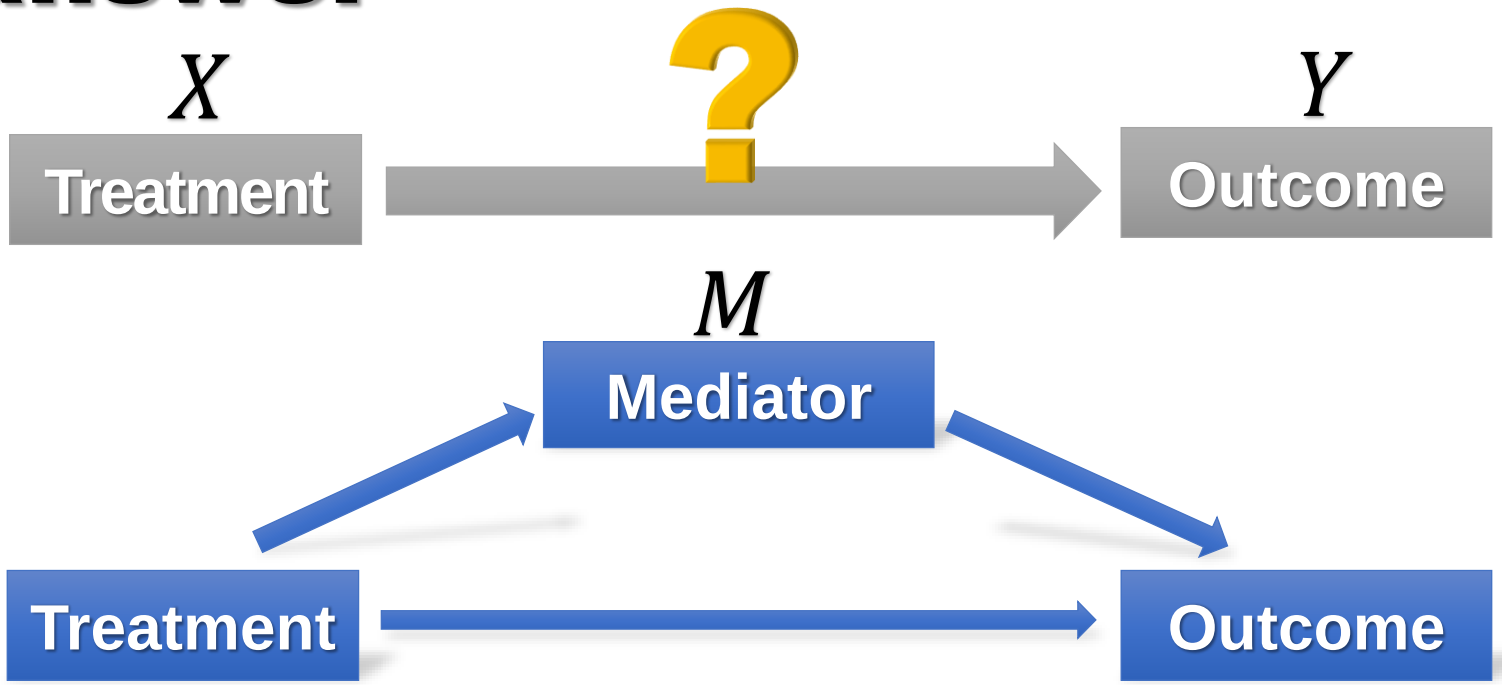
Section 1

Classic Mediation Analysis

Questions to answer

Regression analysis:
"Does it work?"

Mediation analysis:
"How does it work?"



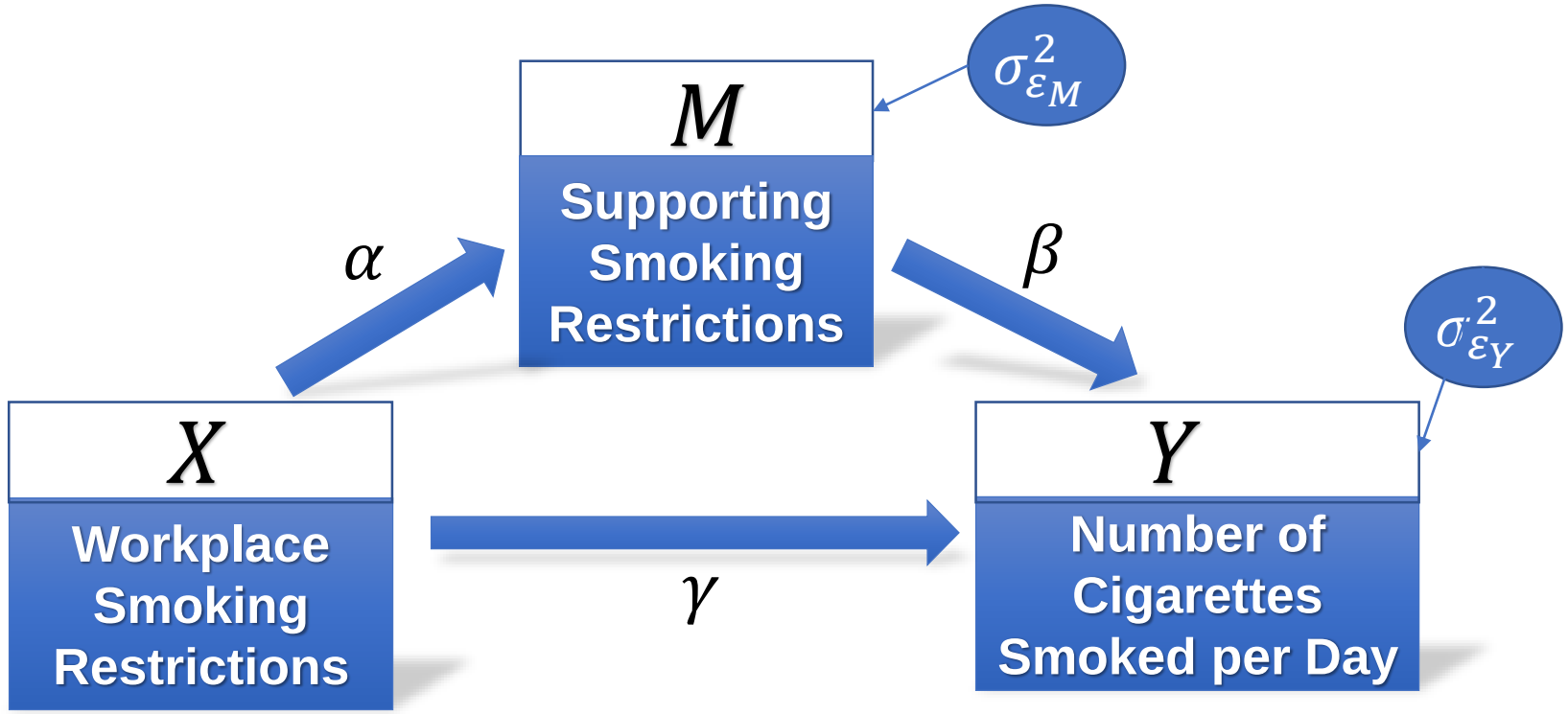
Example 1

Model

(Continuous Variables M, Y)

$$M|X = \mu_M + \alpha X + \varepsilon_M$$
$$Y|X, M = \mu_Y + \gamma X + \beta M + \varepsilon_Y$$

- Direct effect: γ
- Mediation effect: $\alpha\beta$



- Estimator: $F_{ML}(\theta; \boldsymbol{S}_n) = log|\boldsymbol{\Sigma}(\theta)| + trace[\boldsymbol{S}_n \boldsymbol{\Sigma}(\theta)^{-1}] - log|\boldsymbol{S}_n| - p$

$\boldsymbol{S}_n$: *sample covariance matrix*

$\boldsymbol{\Sigma}(\theta)$: *estimated covariance matrix,* $\theta = (\mu_M, \mu_Y, \alpha, \gamma, \beta, \sigma^2_{\varepsilon_M}, \sigma^2_{\varepsilon_Y})$, p: *number of parameters*

$$\text{VAR}(M; \theta) = VAR(\mu_M + \alpha X) + \sigma^2_{\varepsilon_M}$$

Model (Continue)

- Significance Test: $\mathrm{H}_0: \alpha\beta = 0; \mathrm{H}_1: \alpha\beta \neq 0$

  1) *Sobel's Test (Sobel, 1982)*

$$T_{Sobel} = \hat{\alpha}\hat{\beta} \Big/ \sqrt{\hat{\beta}^2 \widehat{Var}(\hat{\alpha}) + \hat{\alpha}^2 \widehat{Var}(\hat{\beta})} \overset{H_0}{\sim} N(0,1) \text{ as } n \to \infty$$

  2) *Resampling/bootstrap*

Assumptions

True Model
Independent Residuals
Normality
Independent-identically-distributed sample data

Section 2

Complex Surveys

Multi-stage Sampling *(Wolter, 2007)*

Suppose the population (40,000 students) locate in two states.

Each state has 20 districts.

Each district has 10 schools.

Each school has 100 students.

Sample weights (ω_0) = the number of students represented

State	District	School	Student	Y	ω_0
1	1	1	1	9.8	2500
1	1	1	2	7.5	2500
1	1	2	3	8.3	2500
1	1	2	4	4.5	2500
1	2	3	5	4.5	2500
1	2	3	6	5.1	2500
1	2	4	7	2.3	2500
1	2	4	8	6.5	2500
2	3	5	9	8.1	2500
2	3	5	10	3.2	2500
2	3	6	11	4.5	2500
2	3	6	12	5.8	2500
2	4	7	13	6.6	2500
2	4	7	14	8.1	2500
2	4	8	15	1.2	2500
2	4	8	16	6.5	2500

$$\omega_0 = \frac{1}{\pi_{PSU} \times \pi_{sch|PSU} \times \pi_{stud|sch}} = \frac{1}{\frac{2}{20} \times \frac{2}{10} \times \frac{2}{100}} = 2{,}500$$

Problem: Estimate the Mean of *Y*

- Disaggregated estimates vs. aggregated estimates

  (Group-specific effects) *(Generalized effects)*

For aggregated estimates:

- Point estimate is consistent when including the sample weights

- Standard errors is underestimated even when taking into account the sample weights

  Adjustments:

  Taylor series linearization (TSL)

  Bootstrap

  Jackknife repeated replications (JRR)

  Balanced repeated replications (BRR)

Balanced repeated replications

(Wolter, 2007)
(Fay & Train, 1995)

State	District	School	Student	Y	ω_0	ω'_1	ω'_2	ω'_3	ω'_4
1	1	1	1	9.8	2500	0	2×2500	0	2×2500
1	1	1	2	7.5	2500	0	2×2500	0	2×2500
1	1	2	3	8.3	2500	0	2×2500	0	2×2500
1	1	2	4	4.5	2500	0	2×2500	0	2×2500
1	2	3	5	4.5	2500	2×2500	0	2×2500	0
1	2	3	6	5.1	2500	2×2500	0	2×2500	0
1	2	4	7	2.3	2500	2×2500	0	2×2500	0
1	2	4	8	6.5	2500	2×2500	0	2×2500	0
2	3	5	9	8.1	2500	0	0	2×2500	2×2500
2	3	5	10	3.2	2500	0	0	2×2500	2×2500
2	3	6	11	4.5	2500	0	0	2×2500	2×2500
2	3	6	12	5.8	2500	0	0	2×2500	2×2500
2	4	7	13	6.6	2500	2×2500	2×2500	0	0
2	4	7	14	8.1	2500	2×2500	2×2500	0	0
2	4	8	15	1.2	2500	2×2500	2×2500	0	0
2	4	8	16	6.5	2500	2×2500	2×2500	0	0

Note. The Primary sampling unit (PSU) is District.

Note. The Primary sampling unit (PSU) is District.

Replicate sampling weights $\boldsymbol{\omega}'_r, r = 1,2,\ldots,R$

R is the number of replications, $R = 4$ in the case.

$$SE_{BRR}(\hat{\mu}) = \sqrt{\frac{\sum_{r=1}^{R}(\hat{\mu}_r - \hat{\mu})^2}{R}}$$

$\hat{\mu}_r$ is the estimate using replicate weight $\boldsymbol{w}'_r$.

$\hat{\mu}$ is the estimate using original(main) sample weights

Section 3

Problem & Solution

Problem to Solve

How can mediation analysis work with complex surveys?

Classic Mediation Analysis	Complex Surveys
i-i-d sample: Independent	Within cluster: Dependent
i-i-d sample: Identically distributed	Between cluster: Heterogenous
i-i-d sample: Equal possibility	Unequal-sized strata: Unequal possibility

Disaggregated Estimates

Aggregated Estimates

Potential Solutions

- Aggregated Estimates

  Design-based method:

  - Taylor series linearization (TSL)
  - Bootstrap
  - Jackknife repeated replications (JRR)
  - Balanced repeated replications (BRR)

Most national/international surveys do not provide the cluster indicator.

Apply *Balanced Repeated Replications* to Analysis of $\alpha\beta$

(*Mai et al., 2019*)

Estimator

Maximum Likelihood

$$F_{ML}(\theta; \boldsymbol{S}_n) = log|\boldsymbol{\Sigma}(\theta)| + trace[\boldsymbol{S}_n\boldsymbol{\Sigma}(\theta)^{-1}] - log|\boldsymbol{S}_n| - p$$

\+ Weighted covariance matrix

Replace $\boldsymbol{S}_n$ with weighted sample covariance matrix $\boldsymbol{S}_{wn}$

Test Statistic

Asymptotical Normality (*Bishop, 1975; Rao, 1973*)

$$\hat{\alpha}\hat{\beta} \sim N(\mu, \sigma^2) \text{ as } n \to \infty$$

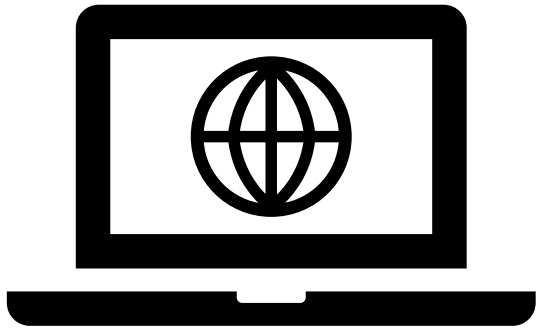
Use BRR Standard Errors

$$T_{\text{BRR}} = \frac{\hat{\alpha}\hat{\beta}}{\text{SE}_{\text{BRR}}(\hat{\alpha}\hat{\beta})} \overset{H_0}{\sim} N(0,1) \text{ as } n \to \infty$$

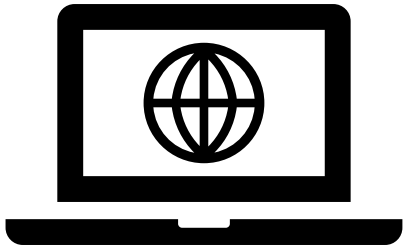
$$\text{SE}_{\text{BRR}}(\hat{\alpha}\hat{\beta}) = \sqrt{\frac{1}{R(1-f)^2}\sum_{r=1}^{R}(\hat{\alpha}_r\hat{\beta}_r - \hat{\alpha}_0\hat{\beta}_0)^2}$$

Section 4

Software & Application



Software Packages



- R package ‘MedSurvey’ (Feb. 2019)

  Flexible and complex models

  https://CRAN.R-project.org/package=MedSurvey

Application

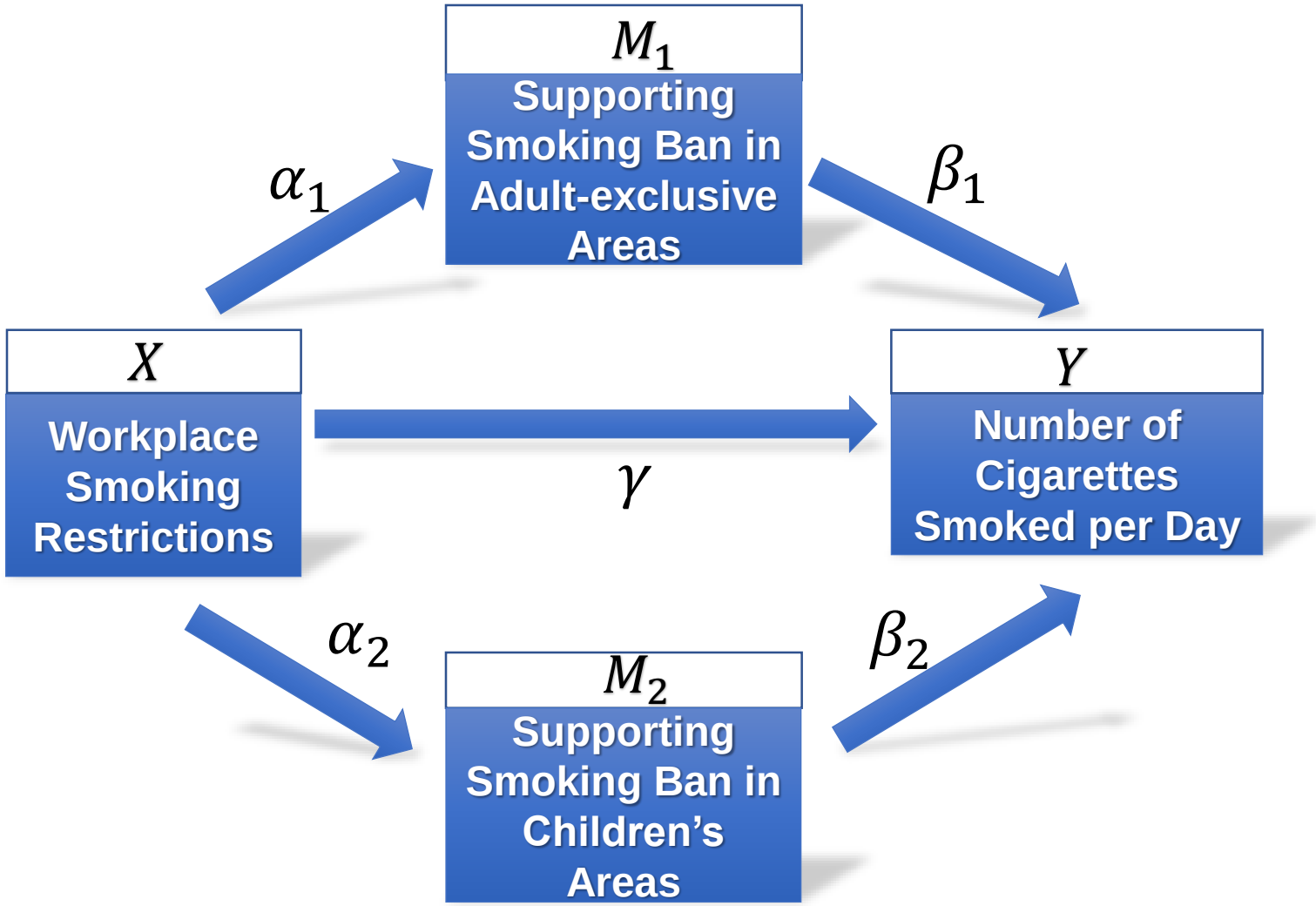
Data:

2014-15 CPS Tobacco Use Supplement (TUS; U.S. Department of Commerce and U.S. Census Bureau 2016), employed adult daily smokers (Non-Hispanic White males only).

Survey Design:

Balanced Repeated Replications $R = 160$.

$f = 0.5$ is suggested.



```r
## Package and options
library("MedSurvey")
## Data and related information
MedData
R <- 160
wgtnames <- paste("repwgt", seq(0,R,by=1), sep="")
mwgtname=wgtnames[1]
repwgtnames=wgtnames[2:(R+1)]

## Sepcify the model 1
model1 <- '
  numcg ~ u0*1 + gamma0*workban + b1*sp_adltban + b2*sp_kidsban
  sp_adltban ~ u1*1 + a1*workban
  sp_kidsban ~ u2*1 + a2*workban
  sp_adltban ~~ sp_kidsban
  a1b1 := a1*b1
  a2b2 := a2*b2
  total := gamma0 + (a1*b1) + (a2*b2)
 '

## Fit the model 1
fit.BRR <- med.fit.BRR(model=model1, data=MedData, mwgtname=mwgtname,
                       repwgtnames=repwgtnames, fayfactor=0.5, parallel = 'parallel')

## View the summary results of the mediation analysis
med.summary(fit=fit.BRR, med.eff=c('a1b1' , 'a2b2'))
```

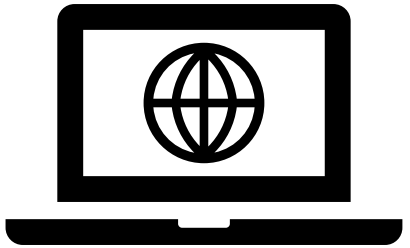
Results

Table 1. Tests for Mediation Effects between Smoking Ban at Work and Number of Cigarettes Smoked per Day among Male NH White Daily Smokers

Mediator	Not Weighted				BRR Weighted			
	$\hat{\alpha}\hat{\beta}$	$\widehat{SD}$	p-value	Adjusted p-value	$\hat{\alpha}\hat{\beta}$	$\widehat{SD}_{\mathrm{BRR}}$	p-value	Adjusted p-value
M_1	−0.0170	0.0066	0.010	0.020	−0.0156	0.0073	0.032	0.063
M_2	0.0018	0.0033	0.588	0.612	0.0004	0.0045	0.937	1.000

Note. *M*1 = Supporting Smoking Ban in Adult-exclusive Areas.
*M*2 = Supporting Smoking Ban in Children's Areas.
n =2,260 (Estimated population size is 3,294,568).
Adjusted p values adjusted for the multiple tests using Holm's method (Holm,1979).

Software Packages



- R package ‘MedSurvey’ (Feb. 2019)

  Flexible and complex models

  https://CRAN.R-project.org/package=MedSurvey

- R shiny (March 2019)

  User-friendly interface

  https://sjbiostat.shinyapps.io/MedSurvey/

R shiny

MedSurvey

https://sjbiostat.shinyapps.io/MedSurvey/

MedSurvey | Model Fitting | Example | Manual | User Map | Contact

A data file with header

Browse... MedData.csv

Upload complete

Y (Outcome)

numcg

X (Predictor)

workban

M (Mediators)

sp_adltban sp_kidsban

Zm (Covariates for mediator)

Zy (Covariates for outcome)

PRTAGE

Main weight

repwgt0

Replicate weights

repwgt

Replicate weights #

160

Fay factor

0 0.5 1

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Submit

Summary | Download

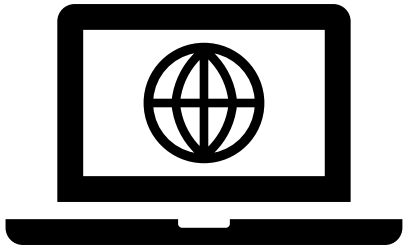
Multimediation with Complex Survey Data:

	Effect	Estimate	BRR SE.	P-Value	Adjusted P-Value
sp_adltban	a1b1	-0.015456	0.004574	0.000728	0.001456
sp_kidsban	a2b2	-0.005366	0.003852	0.1636	0.1636

Note:
p Value adjustment method is holm.
Standard errors type is BRR SE.

Look, this's our product.

Software Packages



- R package ‘MedSurvey’ (Feb. 2019)

  Flexible and complex models

  https://CRAN.R-project.org/package=MedSurvey

- R shiny (March 2019)

  User-friendly interface

  https://sjbiostat.shinyapps.io/MedSurvey/

- SAS macros ‘MedBRR’ (Dec. 2018)

  Super large-scale datasets

  https://github.com/YujiaoMai/MedSurvey

References

- Bishop, Y. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: theory and practice*. Cambridge, Massachusetts: Cambridge, Mass., MIT Press.
- Fay, R. E., & Train, G. F. (1995). Aspects of survey and model-based postcensal estimation of income and poverty characteristics for states and counties. In *Proceedings of the Section on Government Statistics, American Statistical Association, Alexandria, VA* (pp. 154-159).
- Judkins, D. R. (1990). Fay's method for variance estimation. *Journal of Official Statistics*, *6*(3), 223-239.
- Mai, Y., Ha, T., & Soulakova, J. N. (2019). Multimediation Method With Balanced Repeated Replications For Analysis Of Complex Surveys. *Structural Equation Modeling: A Multidisciplinary Journal*.
- Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed., Vol. 2). New York, NY: John Wiley & Sons, Inc.
- Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, *13*, 290–312.
- U.S. Department of Commerce, & U.S. Census Bureau. (2016). National Cancer Institute and Food and Drug Administration co-sponsored Tobacco Use Supplement to the Current Population Survey. 2014-15.
- Wolter, K. (2007). *Introduction to variance estimation*. New York, NY: Springer.

Acknowledgements

- The research is sponsored by American Lebanese Syrian Associated Charities (ALSAC).

St. Jude Children's Research Hospital
Finding cures. Saving children.

Questions?

Thank you !