Advances in Comparative Survey Research and Survey Data Harmonization

Marta Kołczyńska

RECSM Summer Methods School 2024

Course outline

Day 1

1. Cross-national surveys: overview of available data sources

2. Survey data quality and comparability: Total Survey Error framework, cross-survey differences in measurement and representation, survey quality and how to measure it

Day 2

3. Framework for survey data harmonization: representativeness and measurement

4. Representation comparability

Day 3

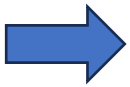
5. Measurement comparability

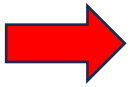
6. Wrap-up

Problems of scale and size

Problem of scale and data subsets. Difference between 1 strongly biased survey among 200 versus 1 strongly biased survey in a subset of 5 surveys from Venezuela.

Combining data





PRECISE NUMBER + PRECISE NUMBER = SLIGHTLY LESS PRECISE NUMBER

PRECISE NUMBER × PRECISE NUMBER = SLIGHTLY LESS PRECISE NUMBER

PRECISE NUMBER + GARBAGE = GARBAGE

PRECISE NUMBER × GARBAGE = GARBAGE

$\sqrt{\text{GARBAGE}}$ = LESS BAD GARBAGE

$(\text{GARBAGE})^2$ = WORSE GARBAGE

$\frac{1}{N}\sum(\text{N PIECES OF STATISTICALLY INDEPENDENT GARBAGE})$ = BETTER GARBAGE

$(\text{PRECISE NUMBER})^{\text{GARBAGE}}$ = MUCH WORSE GARBAGE

GARBAGE − GARBAGE = MUCH WORSE GARBAGE

$\frac{\text{PRECISE NUMBER}}{\text{GARBAGE} - \text{GARBAGE}}$ = MUCH WORSE GARBAGE, POSSIBLE DIVISION BY ZERO

GARBAGE × 0 = PRECISE NUMBER

https://www.explainxkcd.com/wiki/index.php/2295:_Garbage_Math

III. Framework for survey data harmonization

1. Harmonization of representation
2. Harmonization of measurement
3. Latent trend models (if time allows)

Kołczyńska and Bürkner, 10.1093/jssam/smad024

Harmonization of representation

- Survey weights
- Multilevel regression and poststratification (MRP)

Important when the research goal is to estimate levels (means, proportions) overall or by group; less so if the goal is to estimate correlations.

Survey weights: design

Design weights: correct for the unequal probabilities of selection in surveys that use a different sampling design than simple random sampling (SRS)

- Inverse of selection probabilities

$$dweight = 1/\prod_{i=1}^{n} p_i$$

- Calculated based on the sample design (clustered, stratified, multi-stage)
- Also used when some groups or regions are over-sampled by design
- Unrelated to actual item non-response

Example: two-stage sample:

(1) simple random sample of households

(2) within households: Kish grid

Survey weights: design

- Design weights should always be used… when available
- Design weights are typically not provided in cross-national survey datasets
- Nor is information about the design – clusters, strata, etc.
- At the same time, we know that most surveys in cross-national projects use non-SRS

Survey weights

- Poststratification weights: correct sample representativeness by weighting the data to population proportions.

- Population weights: adjust sample sizes to population proportions with a larger entity e.g. for inferences about the EU as a whole.

ESS Weighting guidelines:
https://www.europeansocialsurvey.org/methodology/ess-methodology/data-processing-and-archiving/weighting

Poststratification weights

- Traditional way for representativeness adjustments in survey research.

- In cross-national surveys, poststratification weights typically adjust for age and gender; less often also for the region, urban/rural residence, education or economic status.

- Poststratification weights are constructed by survey producers after data collection.

Poststratification weights

Poststratification weights in the ESS

*have been constructed using information on **age group, gender, education, and region**. The post-stratification weights are obtained by adjusting the design weights in such a way that they will **replicate the distribution of the cross-classification of age group, gender, and education in the population and the marginal distribution for region in the population**. The population distributions for the adjusting variables were obtained from the European Union Labour Force Survey.*

https://www.europeansocialsurvey.org/methodology/ess-methodology/data-processing-and-archiving/weighting

Table2: ESS8 Post-Stratification Weighting Dimensions and source of Control Data by country

Countries ESS8	Post-stratification			
	Weighting Dimensions (G=Gender; A=Age; E=Edu; R=Region)	Source & Year for G, A, E	Definition of Region	Source & Year for Region
Austria	GAE; R	LFS 2016	NUTS2 (9)	Statistics Austria 2016
Belgium	GR; AR; ER	LFS 2016	NUTS1 (3)	LFS 2016
Switzerland	GA; R	LFS 2016	NUTS2 - 6/7 *	LFS 2016
Czechia	GAE; R	LFS 2016	NUTS2 (8)	LFS 2016
Germany	GAE; R	LFS 2016	NUTS1 (16)	LFS 2016
Estonia	GAE	LFS 2016	-	
Spain	GAE; R	LFS 2016	NUTS1 (7)	LFS 2016
Finland	GAE; R	LFS 2016	NUTS2 4/5 *	ESS STATS
France	GAE; R	LFS 2016	NUTS1 (8)*	LFS 2016
Great Britain	GA; R	LFS 2016	NUTS1 (12)	LFS 2016

https://www.europeansocialsurvey.org/sites/default/files/2023-06/ESS8_weighting_strategy.pdf

Poststratification weights

- Poststratification weights have become more complex (accounteing for more factors) over time.

- Old surveys sometimes used case multiplication instead of poststratification weights.
  - Instead of having a weight = 2, the given observation was duplicated and the dataset would include two identical records*.

* Non-unique records is also a quality problem in surveys. Some surveys from cross-national projests contain large proportions of identical records, which is unlikely to happen by chance: https://doi.org/10.18148/srm/2017.v11i1.6557

Analysis of 1721 surveys from 22 cross-national survey projects 1966-2013

Table 47.1 Percentages of weights containing particular weighting variables in time periods.

			Post-stratification weights				
Period	Gender	Age	Economic status	Education	Region	Urbanicity	Total number of surveys
	Percent within surveys using any type of weight						
1966–1980	0	0	0	0	0	0	17
1981–1985	50	93	3	0	33	3	51
1986–1990	69	69	0	3	28	0	80
1991–1995	53	50	0	22	9	25	82
1996–2000	62	59	2	6	43	20	285
2001–2005	39	40	1	17	32	18	344
2006–2009	68	64	1	24	34	22	679
2010–2013	89	90	2	25	80	61	183
Total N	646	637	14	194	407	257	1721

Zieliński, Powałko, and Kołczyńska, 2018. doi.org/10.1002/9781118884997.ch47

Poststratification weights

- There typically is just one poststratification weighting variable per survey (if there is any).

- This weighting variable does not necessary adjust for the factors that are important for the analysis in question.
  - E.g., for political trust education is far more important than age or gender, but not all surveys include education in poststratification weights (e.g. Eurobarometer doesn't)

- It's a good idea to check if the weights have an average of 1, only have positive values, and do not have extremely high values
  - E.g. weight of 90.32 in New Zealand, ISSP 2007

Poststratification weights

- In many surveys weights are poorly documented, and it is not always clear what factors they account.
- Finally, sometimes, especially in older surveys, there are no weights.
- With appropriate population data, we can calculate our own poststratification weights.
- It’s best to have the joint distribution of all poststratification variables, i.e. proportions for all combinations of e.g. age, sex, education.

Poststratification table

- Joint distribution
- Counts / proportions for all combinations of poststratification variables

Example for the US:

4 ethnic groups

2 gender groups

3 age groups

2 education groups

-> 48 combinations

eth	male	age	educ	n	prop_pop
Black	-0.5	18-39	BA+	1275123	0.005581791
Black	-0.5	18-39	no BA	5133125	0.022470013
Black	-0.5	40-59	BA+	1315067	0.005756644
Black	-0.5	40-59	no BA	3943910	0.017264280
Black	-0.5	60+	BA+	663855	0.002905994
Black	-0.5	60+	no BA	3056635	0.013380276
Black	0.5	18-39	BA+	809469	0.003543412
Black	0.5	18-39	no BA	5372170	0.023516421
Black	0.5	40-59	BA+	842037	0.003685977
Black	0.5	40-59	no BA	3762614	0.016470666
Black	0.5	60+	BA+	454763	0.001990704
Black	0.5	60+	no BA	2237694	0.009795400
Hispanic	-0.5	18-39	BA+	881474	0.003858611
Hispanic	-0.5	18-39	no BA	3818142	0.016713737
Hispanic	-0.5	40-59	BA+	696277	0.003047920
Hispanic	-0.5	40-59	no BA	2258622	0.009887012

Poststratification table

Sample

eth	male	age	educ	n	prop_sample
Black	-0.5	18-39	BA+	3	0.0012
Black	-0.5	18-39	no BA	10	0.0040
Black	-0.5	40-59	BA+	10	0.0040
Black	-0.5	40-59	no BA	12	0.0048
Black	-0.5	60+	BA+	18	0.0072
Black	-0.5	60+	no BA	36	0.0144
Black	0.5	18-39	BA+	1	0.0004
Black	0.5	18-39	no BA	2	0.0008
Black	0.5	40-59	BA+	1	0.0004
Black	0.5	40-59	no BA	3	0.0012
Black	0.5	60+	BA+	6	0.0024
Black	0.5	60+	no BA	27	0.0108
Hispanic	-0.5	18-39	BA+	5	0.0020
Hispanic	-0.5	18-39	no BA	12	0.0048
Hispanic	-0.5	40-59	BA+	6	0.0024
Hispanic	-0.5	40-59	no BA	7	0.0028

Population

eth	male	age	educ	n	prop_pop
Black	-0.5	18-39	BA+	1275123	0.005581791
Black	-0.5	18-39	no BA	5133125	0.022470013
Black	-0.5	40-59	BA+	1315067	0.005756644
Black	-0.5	40-59	no BA	3943910	0.017264280
Black	-0.5	60+	BA+	663855	0.002905994
Black	-0.5	60+	no BA	3056635	0.013380276
Black	0.5	18-39	BA+	809469	0.003543412
Black	0.5	18-39	no BA	5372170	0.023516421
Black	0.5	40-59	BA+	842037	0.003685977
Black	0.5	40-59	no BA	3762614	0.016470666
Black	0.5	60+	BA+	454763	0.001990704
Black	0.5	60+	no BA	2237694	0.009795400
Hispanic	-0.5	18-39	BA+	881474	0.003858611
Hispanic	-0.5	18-39	no BA	3818142	0.016713737
Hispanic	-0.5	40-59	BA+	696277	0.003047920
Hispanic	-0.5	40-59	no BA	2258622	0.009887012

Exercise

ex2_weights.R

Properties of poststratification weights

- The smaller the variance of the poststratification weight, the better (more representative) the original sample

- The number of distinct values of the poststratification weight equals the number of categories, by which the weight was calculated

Properties of poststratification weights

- Poststratification weights should not change the sample size, i.e. the mean of weights should be 1
  - If the mean is different than 1, it needs to be rescaled by dividing by the mean
- Weights shouldn’t be set to 0, unless one wants to exclude an observation from the analysis
- Extremely high values should be avoided
  - Trimming

Poststratification weights: iterative

What if we don't have a poststratification table, just marginal distributions?

Joint distributions of population data for more than 3 characteristics are rare, especially for multiple countries.

e.g. the European Labor Force Survey has age-sex-education distributions available through Eurostata (with some gaps)

Sometimes we only have single-variable distributions, separately for sex, age, gender, ethnicity, etc.

Poststratification weights: iterative

Compute Ethnicity weight (weight_eth)

Weight data by weight_eth and generate the weighted frequency table for Male

Compute Male weight (weight_male)

Weight by weight_eth*weight_male and generate the weighted frequency table for Age

Compute Age weight (weight_age)

Weight the data by weight_eth*weight_male*weight_age and generate the weighted frequency table for Education

Compute Education weight (weight_educ)

Combine weights = weight_eth * weight_male * weight_age * weight_educ

Exercise

ex2_weights2.R

MRP = multilevel regression + poststratification

Two main applications:

- Small Area Estimation (SME)
- Analysis of non-probability samples

Used in elections forecasting, especially in non-proportional systems when the winner of the election is not the winner of the popular vote.

Developed and popularized by Andrew Gelman, who calls it *MisterP*.

Gelman and Little. 1997. „Poststratification into many categories using hierarchical logistic regression.” Survey Methodology 23: 127-135.

BTW, Andrew Gelman’s blog: https://statmodeling.stat.columbia.edu/

MRP: the Xbox survey

Xbox survey: 750,148 responses, from 345,858 unique respondents, during the 45 days preceding the 2012 US presidential election.

Xbox survey Respondents:

65% aged 18-29 (electorate: 19%)

93% male (electorate: 47%)

also asked about: education, state, party ID, political ideology, and voting in 2008

„…After adjusting the Xbox responses via multilevel regression and poststratification, we obtain estimates which are in line with the forecasts from leading poll analysts, which were based on aggregating hundreds of traditional polls conducted during the election cycle…"

Wang, Rothschild, Goel, and Gelman, http://dx.doi.org/10.1016/j.ijforecast.2014.06.001

MRP: multilevel model

For MRP, the groups are demographic categories, i.e., age categories, sex, education levels, ethnic groups, etc., not countries as in cross-national analysis or individuals as in panel analysis.

attitude ~ 1 + (1 | age_cat) + sex + (1 | educ_cat) + …

From this model we obtain predictions for each age * sex * education combination.

We weigh these predictions by counts in the poststratification table and average them to obtain the estimated population mean.

MRP: multilevel

Why does it matter that estimates from the multilevel model are weighted, not the data?

Multilevel models use partial pooling of information between groups.

Partial pooling avoids small groups with extreme values of the outcome variable from affecting the overall estimate too much.

- e.g. that one respondent from New Zealand, ISSP 2007, with the weight of 90.32.

MRP: poststratification table

MRP, like constructing weights, requires data about the population.

But, weights (typically) come with the data, and the poststratification table for MRP must be constructed by the researcher herself.

The poststratification table requires population counts for all combination of adjustment factors, e.g. for age categories * sex * education categories * region (not just margins).

The categories of the poststratification table must be the same as in the survey data.

In cross-national projects, the population data should be measured in the same way over time.

MRP: poststratification table

Population data for the poststratification table may come from the census or a high-quality large sample survey.

US: annual American Community Survey.

EU + candidate countries + EFTA: European Labor Force Survey, conducted by national statistics institutes (this is an *ex ante* harmonized survey).

- Aggregated data are available via the Eurostat website.
- Micro-data are available for academic use after registration.

Many countries of the world: 10% decennial census samples in IPUMS International.

Software break

Install `rstan` and `rstanarm`

https://github.com/stan-dev/rstan/wiki/RStan-Getting-Started

1. Configure c++ toolchain

2. install.packages("rstan", repos = c('https://stan-dev.r-universe.dev', getOption("repos")))

```
library(rstan)

example(stan_model, package = "rstan", run.dontrun = TRUE)

Install.packages(rstanarm)
```

Exercise

Stan is a probabilistic programming language for Bayesian inference maintained and developed by a 50+ strong international team.

Rstanarm is a simplified interface to Stan in r.

Stan is named after Stanisław Ulam (1909-1984), the inventor of the Monte Carlo method.

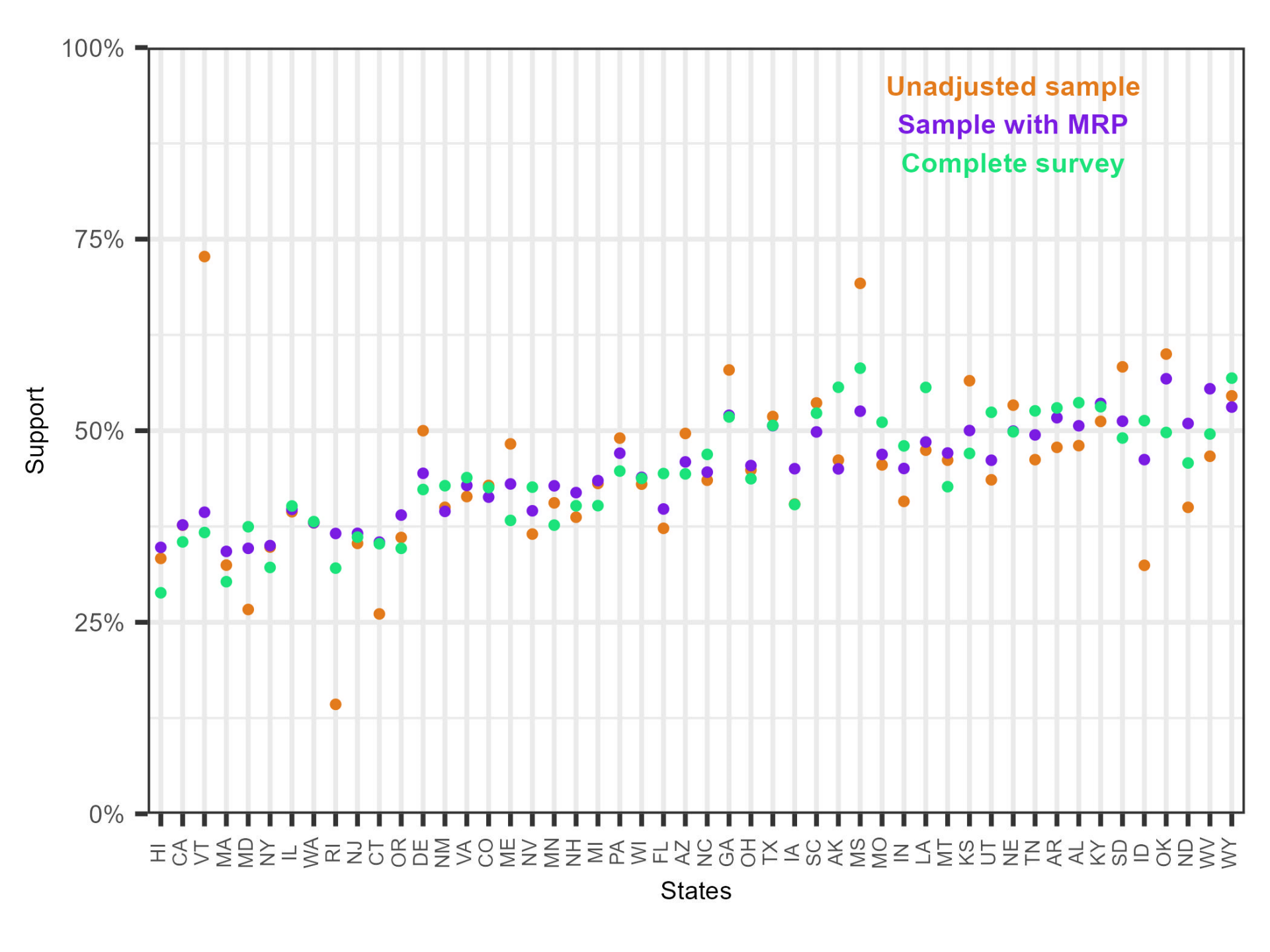
For our purposes, the greatest advantage of the Bayesian approach is that it makes it easy to estimate uncertainty in complex models.

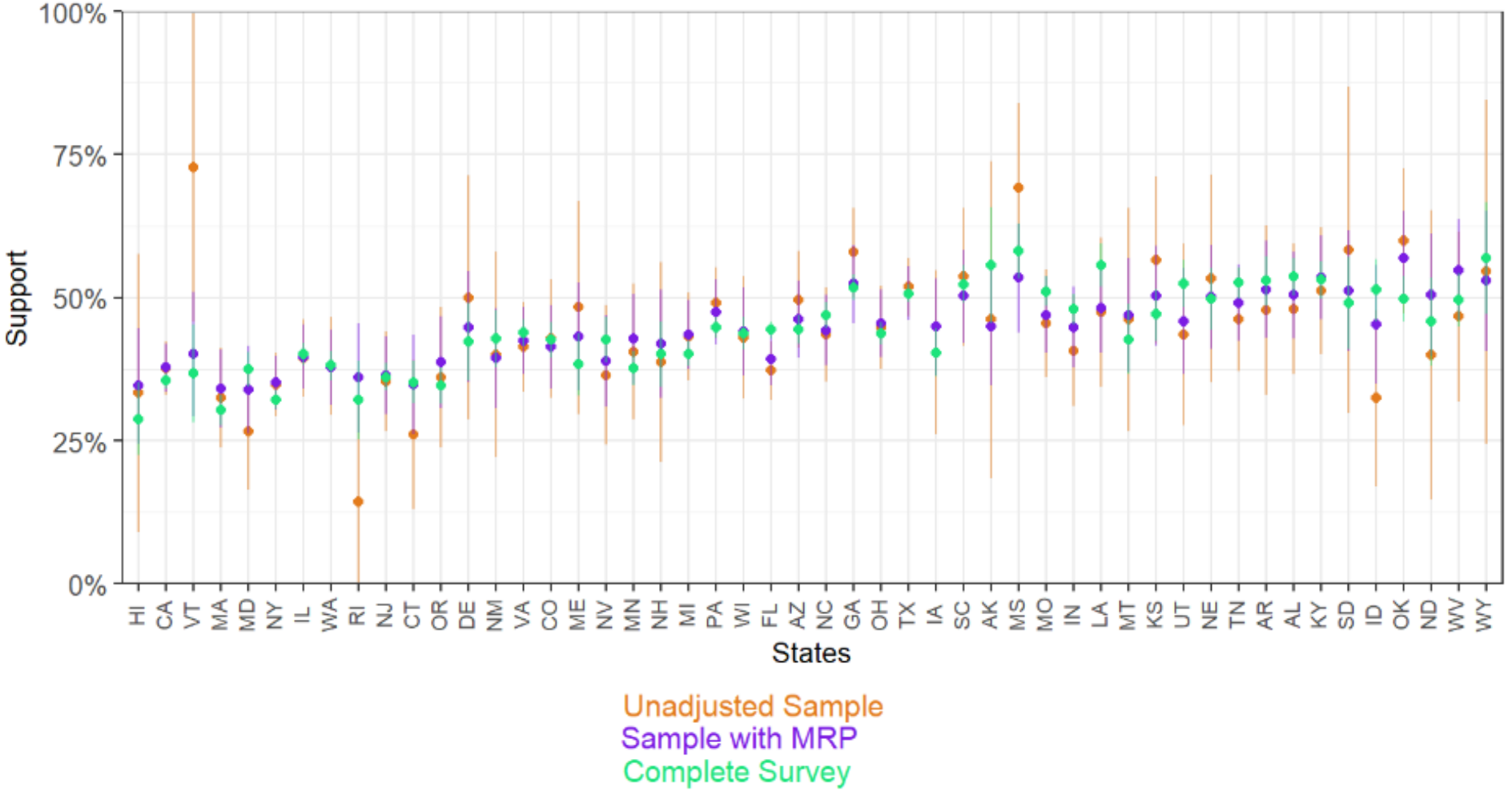
Exercise

ex3_mrp_tutorial.R

Modified version of:

Lopez-Martin, Phillips, and Gelman, 2022, Multilevel Regression and Poststratification Case Studies: https://bookdown.org/jl5522/MRP-case-studies/


Unadjusted sample
Sample with MRP
Complete survey
100%
75%
50%
25%
0%
Support
HI CA VT MA MD NY IL WA RI NJ CT OR DE NM VA CO ME NV MN NH MI PA WI FL AZ NC GA OH TX IA SC AK MS MO IN LA MT KS UT NE TN AR AL KY SD ID OK ND WV WY
States


100%
75%
50%
25%
0%
Support
HI CA VT MA MD NY IL WA RI NJ CT OR DE NM VA CO ME NV MN NH MI PA WI FL AZ NC GA OH TX IA SC AK MS MO IN LA MT KS UT NE TN AR AL KY SD ID OK ND WV WY
States
Unadjusted Sample
Sample with MRP
Complete Survey

MRP: extensions

Similarly, we can poststratify to non-census variables, e.g. voter / non-voter status.

Step 1: estimate the proportion of voters by age and sex based on a high-quality survey dataset (e.g., election studies),

Sex	Age	Estimated turnout (%)	Standard error of the estimate
F	20-34		
F	35-54		
F	55-74		
M	20-34		
M	35-54		
M	55-74		

MRP: extensions

Similarly, we can poststratify to non-census variables, e.g. voter / non-voter status.

Step 1: estimate the proportion of voters by age, sex, and education based on a high-quality survey dataset (e.g., election studies),

Sex	Age	Estimated turnout (%)	Standard error of the estimate
F	20-34	45	+-2 ppoints
F	35-54	50	+-2 ppoints
F	55-74	55	+-2 ppoints
M	20-34	44	+-2 ppoints
M	35-54	51	+-2 ppoints
M	55-74	58	+-2 ppoints

MRP: extensions

Step 2: create 100 plausible poststratification tables based on model predictions from Step 1,

Step 3: perform poststratification 100 times with the 100 poststratification tables,

Step 4: average the 100 poststratified datasets.

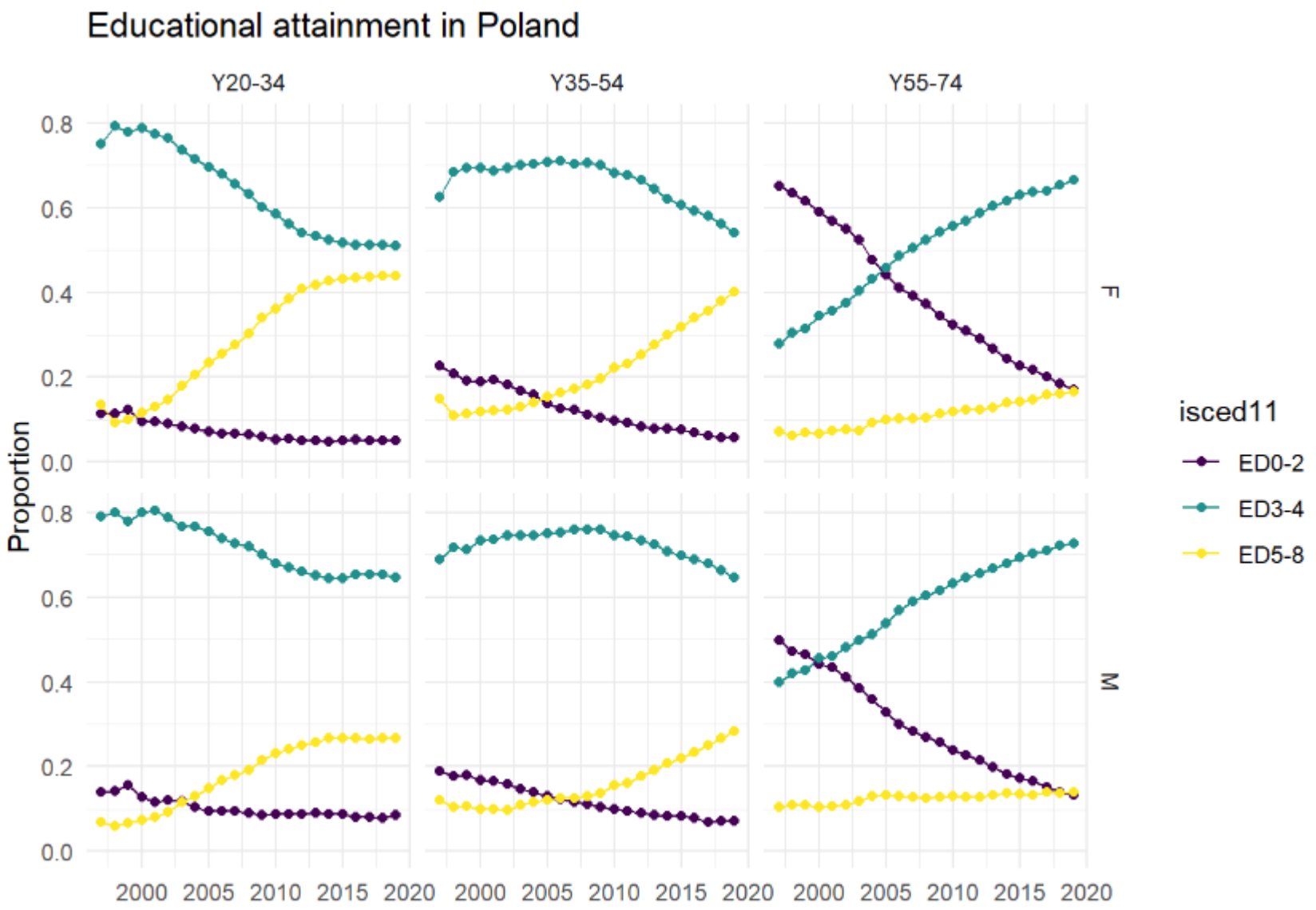
This propagates the uncertainty from education estimates to final results.

MRP: extensions

Imputing parts of the poststratification table.

Example: We have joint distributions by age and sex for all time points, but additionally by education only for some years.

Step 1: impute proportions of education by age and sex for the missing years.


Educational attainment in Poland
Y20-34
Y35-54
Y55-74
F
M
Proportion
0.0
0.2
0.4
0.6
0.8
2000
2005
2010
2015
2020
isced11
ED0-2
ED3-4
ED5-8

MRP: extensions

Step 2: create 100 plausible poststratification tables based on model predictions from Step 1,

Step 3: perform poststratification 100 times with the 100 poststratification tables,

Step 4: average the 100 poststratified datasets.

This propagates the uncertainty from education estimates to final results.

MRP vs weights

- Possible to estiamate quantities also for states that are not in the data; in our example based on the region and Republican vote

$$Pr(y_i = 1) = logit^{-1}(\alpha^{\text{state}}_{\text{s[i]}} + \alpha^{\text{age}}_{\text{a[i]}} + \alpha^{\text{eth}}_{\text{r[i]}} + \alpha^{\text{educ}}_{\text{e[i]}} + \beta^{\text{male}} \cdot \text{Male}_{\text{i}} + \alpha^{\text{male.eth}}_{\text{g[i],r[i]}} + \alpha^{\text{educ.age}}_{\text{e[i],a[i]}} + \alpha^{\text{educ.eth}}_{\text{e[i],r[i]}})$$

where:

$$\begin{aligned}
\alpha^{\text{state}}_{\text{s}} &\sim \text{normal}(\gamma^0 + \gamma^{\text{south}} \cdot \text{South}_{\text{s}} + \gamma^{\text{northcentral}} \cdot \text{NorthCentral}_{\text{s}} + \gamma^{\text{west}} \cdot \text{West}_{\text{s}} \\
&\quad + \gamma^{\text{repvote}} \cdot \text{RepVote}_{\text{s}}, \sigma^{\text{state}}) \text{ for s} = 1,...,50 \\
\alpha^{\text{age}}_{\text{a}} &\sim \text{normal}(0, \sigma^{\text{age}}) \text{ for a} = 1,...,6 \\
\alpha^{\text{eth}}_{\text{r}} &\sim \text{normal}(0, \sigma^{\text{eth}}) \text{ for r} = 1,...,4 \\
\alpha^{\text{educ}}_{\text{e}} &\sim \text{normal}(0, \sigma^{\text{educ}}) \text{ for e} = 1,...,5 \\
\alpha^{\text{male.eth}}_{\text{g,r}} &\sim \text{normal}(0, \sigma^{\text{male.eth}}) \text{ for g} = 1,2 \text{ and r} = 1,...,4 \\
\alpha^{\text{educ.age}}_{\text{e,a}} &\sim \text{normal}(0, \sigma^{\text{educ.age}}) \text{ for e} = 1,...,5 \text{ and a} = 1,...,6 \\
\alpha^{\text{educ.eth}}_{\text{e,r}} &\sim \text{normal}(0, \sigma^{\text{educ.eth}}) \text{ for e} = 1,...,5 \text{ and r} = 1,...,4
\end{aligned}$$

MRP vs weights

- Requires high quality population data
- Makes it easy to include region-level predictors (e.g., Republican vote)
- Best done with Bayesian models, which makes it straightforward to quantify uncertainty
- But, Bayesian models -> computational issues
- Is arguably more complex than using weights