Widening access to confidential data with the synthpop package

Gillian Raab

Administrative Data Research Centre – Scotland

THE UNIVERSITY of EDINBURGH

Administrative Data Research Network

An ESRC Data Investment

Outline

- Background to our project
  - Context of longitudinal studies
  - Synthpop package
- Features of the synthpop package
  - Methods of synthesising
  - Methods of evaluating how well the synthetic data correspond to the original
- Final thoughts

SYLLS project – from 2013

- To develop tools that can be used by staff with access to the original data to produce synthetic data extracts that can be made available more freely than the original data.
- Researchers can explore the synthetic data and develop analysis code
- Teaching data sets are another use
- Originally we worked for the staff at the Scottish Longitudinal Study – and we still do
- But we now have a wider remit within ADRN-S to work with all staff making administrative data available

UK longitudinal studies

- ONS-LS, SLS, NILS
  - Censuses linked and to other data sets
  - Users apply for an extract
  - Need to analyse it in a safe setting or by sending in code to be analysed by staff
- SYLLS project – from 2013
  - To develop methods and tools that LS-DSU staff can use to produce synthetic data extracts that can be supplied to users to analyse on their own computers
  - Code run on the synthetic data can then be run on the original LS data for publication

Current situation

- Longitudinal studies
  - SLS Permissions obtained to release synthetic extracts and first examples are coming through
  - NILS Almost
  - ONS-LS Unsure
- BUT
  - The synthpop package is available and is being used by others

A software tool for producing synthetic versions of sensitive microdata

R package

synthpop

version 1.3-0

http://cran.r-project.org/package=synthpop

Completely synthetic data

What is it?

Data that resembles the original data

But contains no records that correspond to real individuals or other units

History

Originally proposed for disclosure control over 20 years ago

Many theoretical papers from the early 2000's

Real applications started to appear a few years later

US Bureau of the Census

Others in Canada, New Zealand, Germany

Disclosure risk

Not zero, but evaluations of applications suggest it is low.

The LS data are released only to accredited researchers

Perceived risk may be as important as actual risk

Observed (input)

Sex	Age	Education	Marital status	Income	Life satisfaction
FEMALE	57	VOCATIONAL/GRAMMAR	MARRIED	800	PLEASED
MALE	41	SECONDARY	UNMARRIED	1500	MIXED
FEMALE	18	VOCATIONAL/GRAMMAR	UNMARRIED	NA	PLEASED
FEMALE	78	PRIMARY/NO EDUCATION	WIDOWED	900	MIXED
FEMALE	54	VOCATIONAL/GRAMMAR	MARRIED	1500	MOSTLY SATISFIED
MALE	20	SECONDARY	UNMARRIED	-8	PLEASED
FEMALE	39	SECONDARY	MARRIED	2000	MOSTLY SATISFIED
MALE	39	SECONDARY	MARRIED	1197	MIXED
FEMALE	38	VOCATIONAL/GRAMMAR	MARRIED	NA	MOSTLY DISSATISFIED
FEMALE	73	VOCATIONAL/GRAMMAR			
FEMALE	54	SECONDARY			
MALE	30	VOCATIONAL/GRAMMAR			
MALE	68	SECONDARY			
MALE	61	PRIMARY/NO EDUCATION			

Synthetic (output)

Sex	Age	Education	Marital status	Income	Life satisfaction
MALE	81	PRIMARY/NO EDUCATION	MARRIED	2100	PLEASED
MALE	54	VOCATIONAL/GRAMMAR	MARRIED	1700	PLEASED
FEMALE	32	VOCATIONAL/GRAMMAR	DIVORCED	870	MIXED
FEMALE	98	PRIMARY/NO EDUCATION	MARRIED	800	MOSTLY DISSATISFIED
FEMALE	50	PRIMARY/NO EDUCATION	MARRIED	NA	MOSTLY SATISFIED
FEMALE	37	VOCATIONAL/GRAMMAR	MARRIED	158	PLEASED
MALE	28	VOCATIONAL/GRAMMAR	NA	1500	MOSTLY SATISFIED
FEMALE	62	PRIMARY/NO EDUCATION	MARRIED	830	MOSTLY SATISFIED
MALE	78	PRIMARY/NO EDUCATION	MARRIED	NA	PLEASED
FEMALE	29	SECONDARY	MARRIED	580	MOSTLY SATISFIED
MALE	59	PRIMARY/NO EDUCATION	MARRIED	1300	MOSTLY SATISFIED
MALE	41	SECONDARY	UNMARRIED	1500	MIXED
MALE	18	SECONDARY	UNMARRIED	-8	PLEASED
FEMALE	73	PRIMARY/NO EDUCATION	WIDOWED	1350	MOSTLY SATISFIED

Data that look (structurally) like original data but contain artificial units only

Creating synthetic data

Assumes some sort of model fits the data

- Fit the model to the data
- Generate synthetic data from the fit to the model

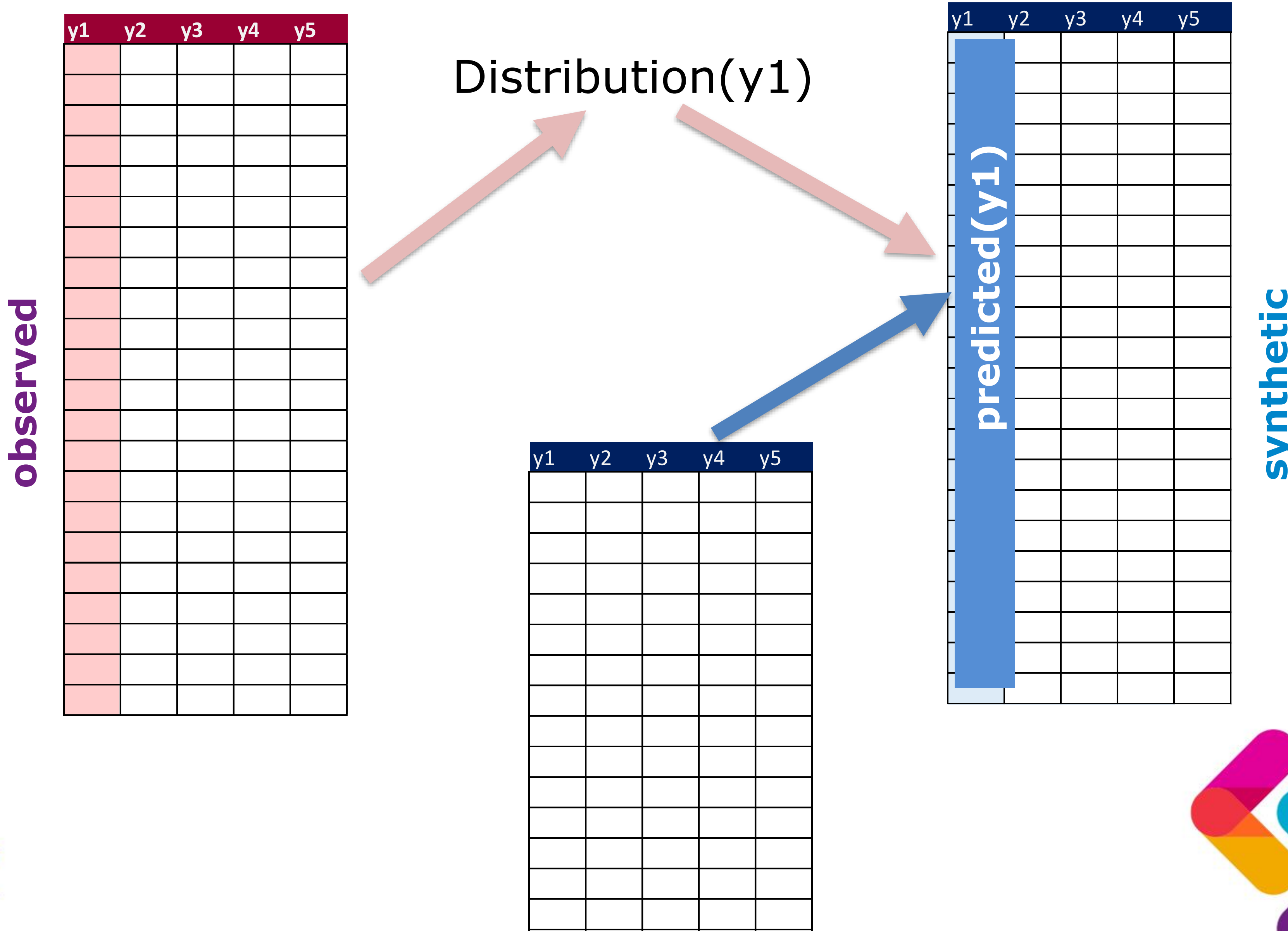
In practice for real data

- Build up from conditional distributions

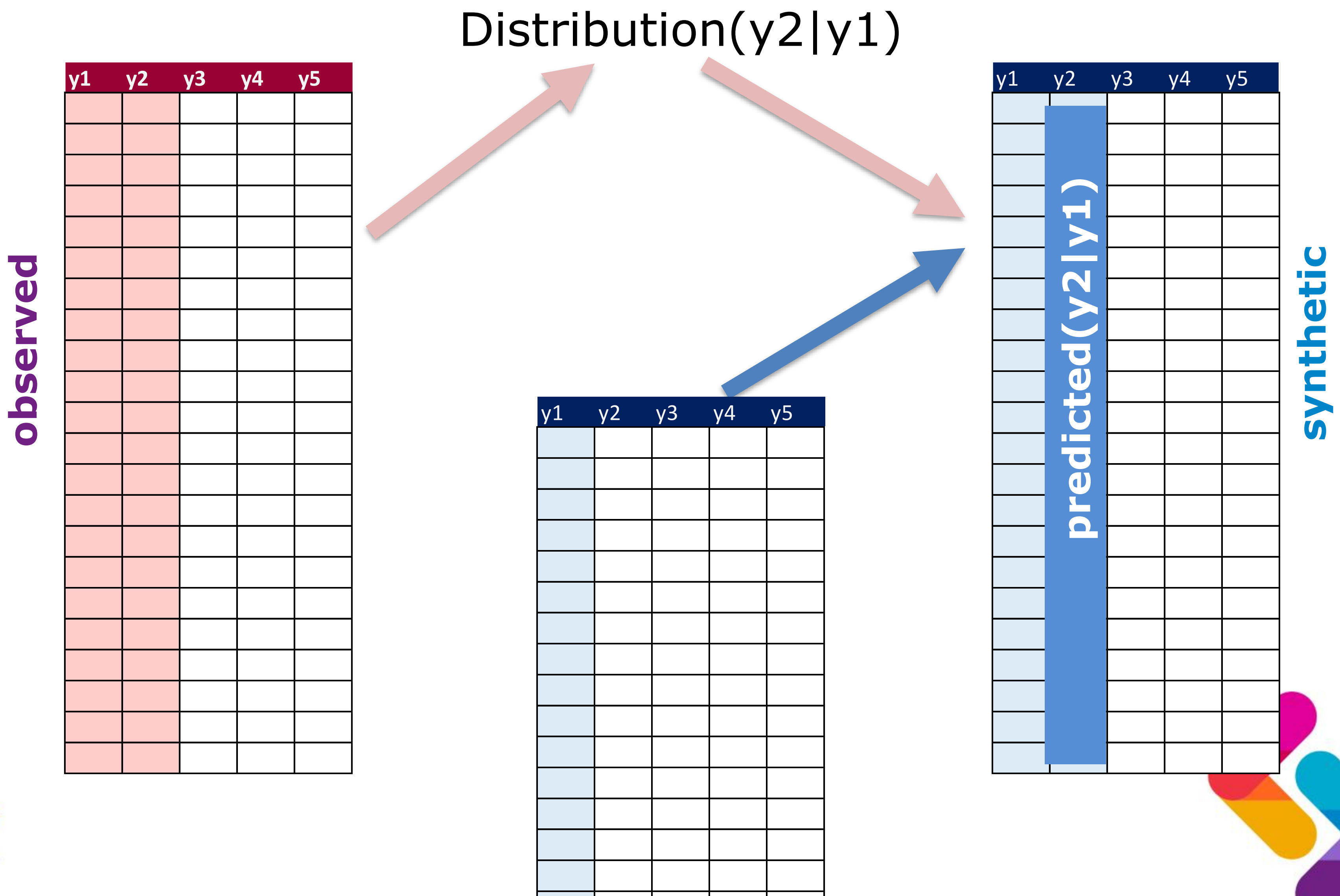
Example

- Start with first variable – fit a distribution– e.g. age
- Generate a sample from this distribution
- Model next variable e.g. sex predicted from age
- Generate simulated data from (sex | age)
- Model education and generate (education| age, sex)
- Generate simulated data from (occupation| sex,age)

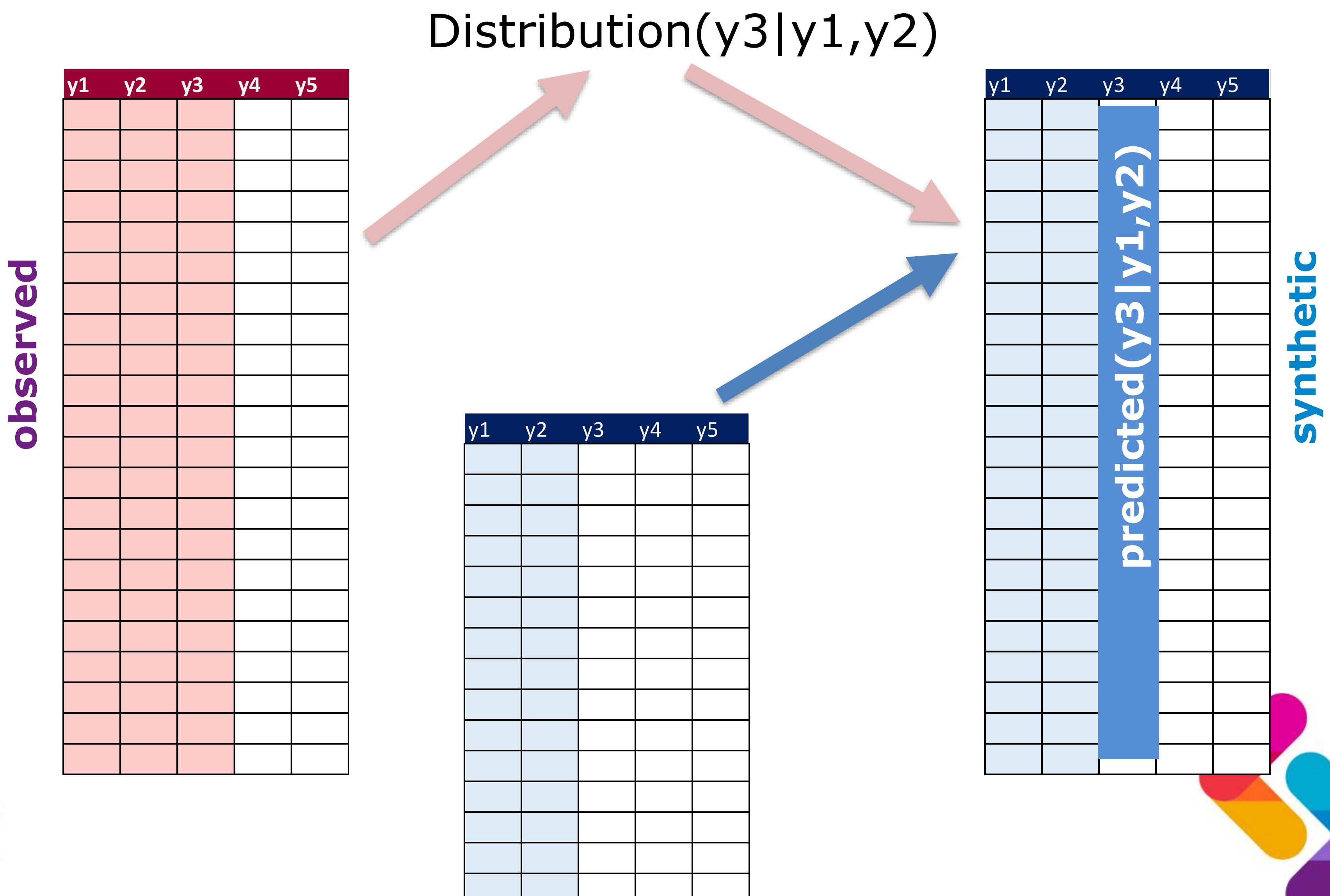
First variable



Second variable

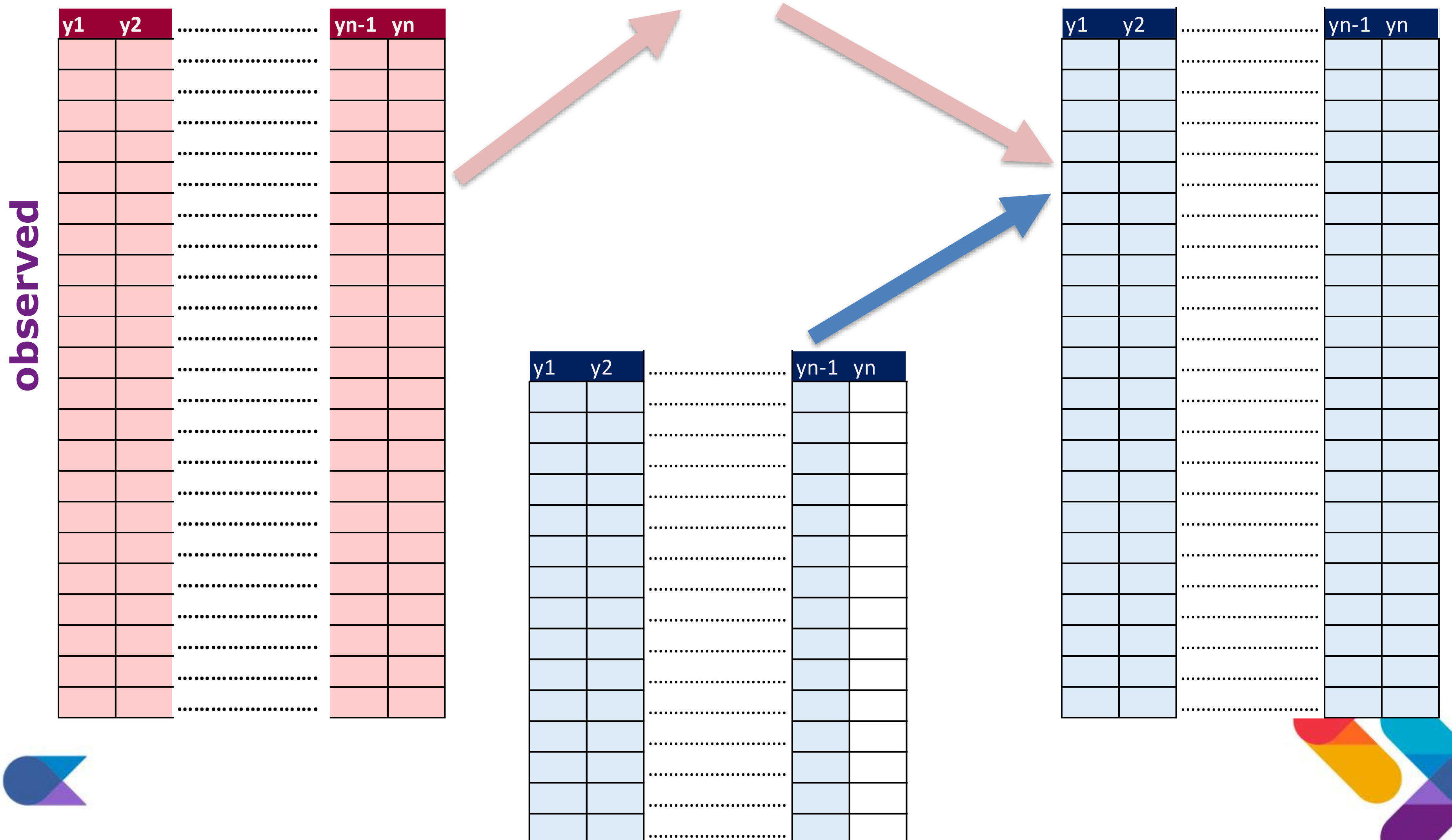


Third variable

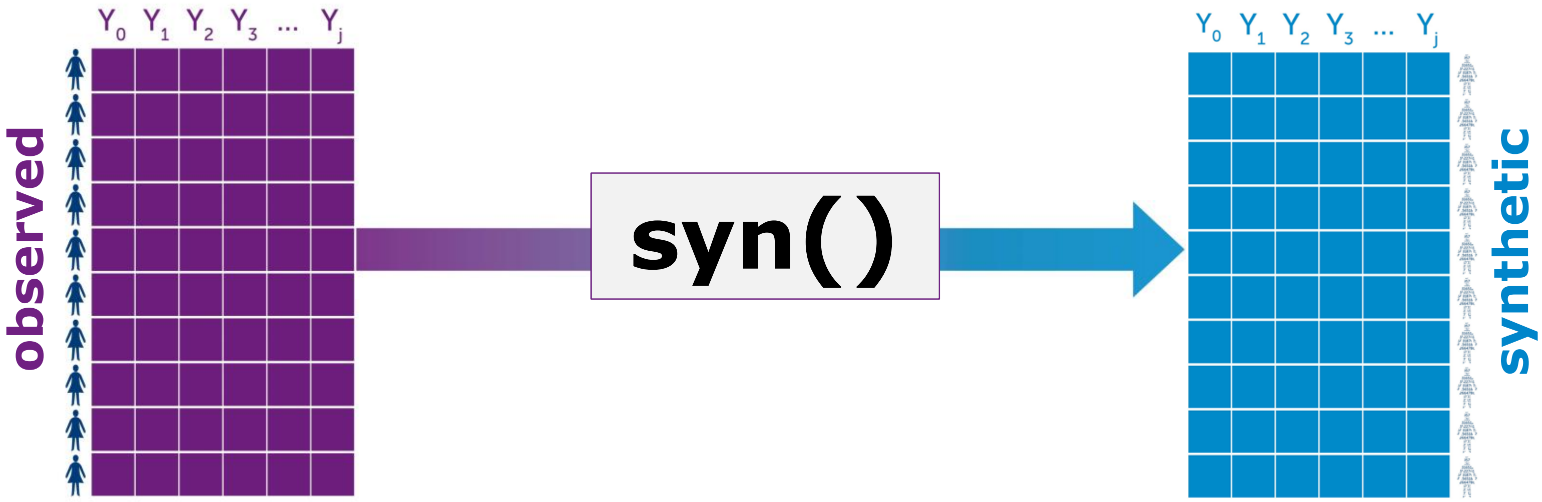


Final step

Distribution(yn|y1,y2,y3,y4,y5,y6,y7,y8,y9,y10,y11,...yn-1)



Generating synthetic data: **synthpop**



Synthesis default parameters

Mysynth1<-syn(data)

Or control how the data are synthesised

	1	2	3	4	5	6	7	8
Variable	age	sex	hh_occ1	mar	agegroup	pperroom	Hh_occ2	disability
Method	sample	logreg	cart	polyreg	~I(floor(age/10))	normrank	mymeth	ctree

Mysynth2<-syn(data, method=***meth*** ,predictor.matrix=***prmat*,**

visit.sequence=c(1,3:6,2,7:8),cont.na=list(age=-8,pperroom=-1),

rules=list(mar = "age < 16"),values=list(mar="Single"),

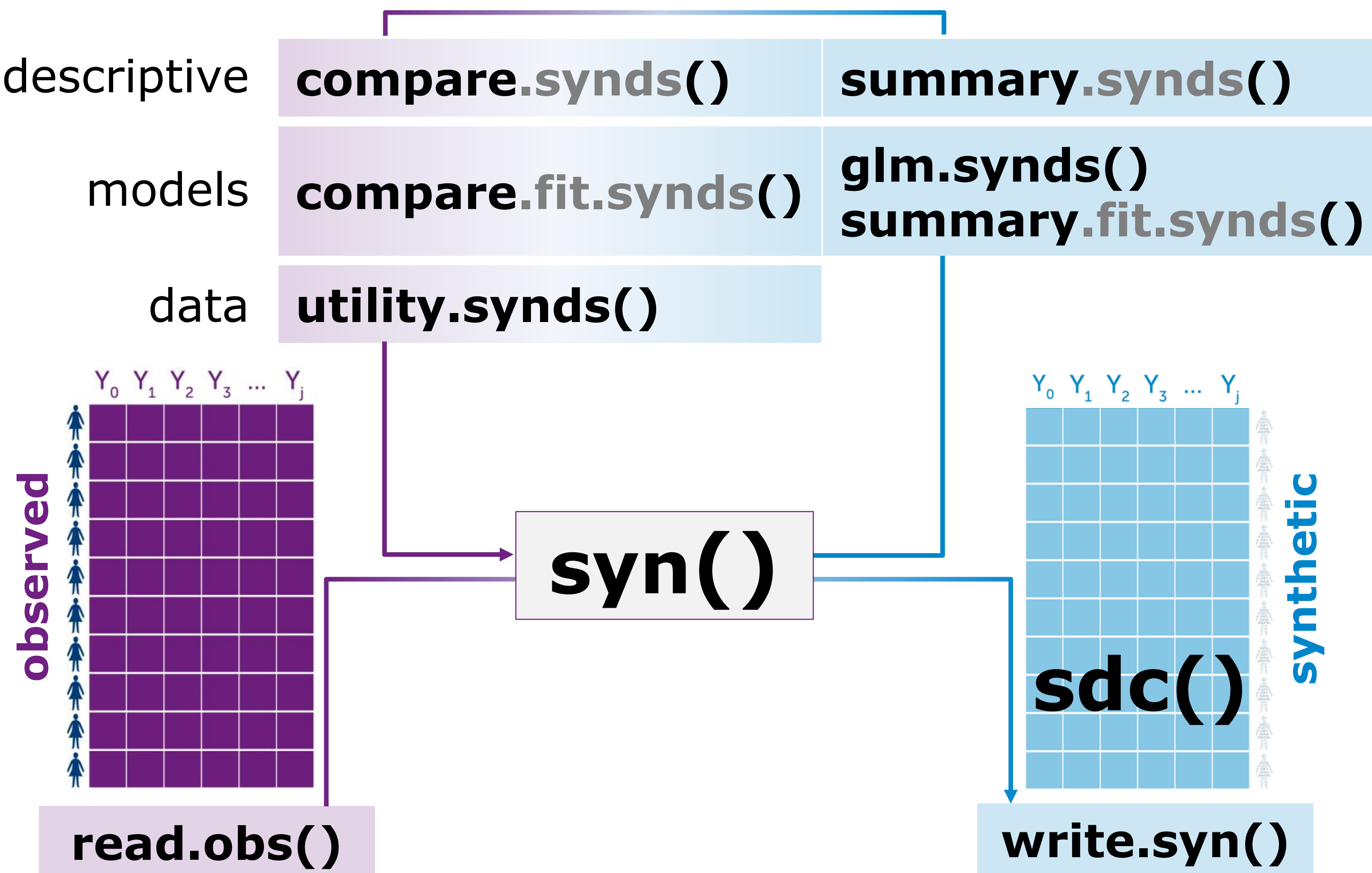
smoothing=list(pperroom="density"), polyreg.maxit=500,

m =5, k=1000, proper=T, models=T, diagnostics=T, *and more……*

Practicalities

- Default parameters
  - Use CART methods √
  - Produce just a single synthetic data set √
  - In the order of variables in the data set ×
  - Use all previously synthesised as predictors ×
  - No specification of rules or checks ×
  - No smoothing continuous variables ×
  - No coding missing value indicators x
  - No stratification into subgroups x

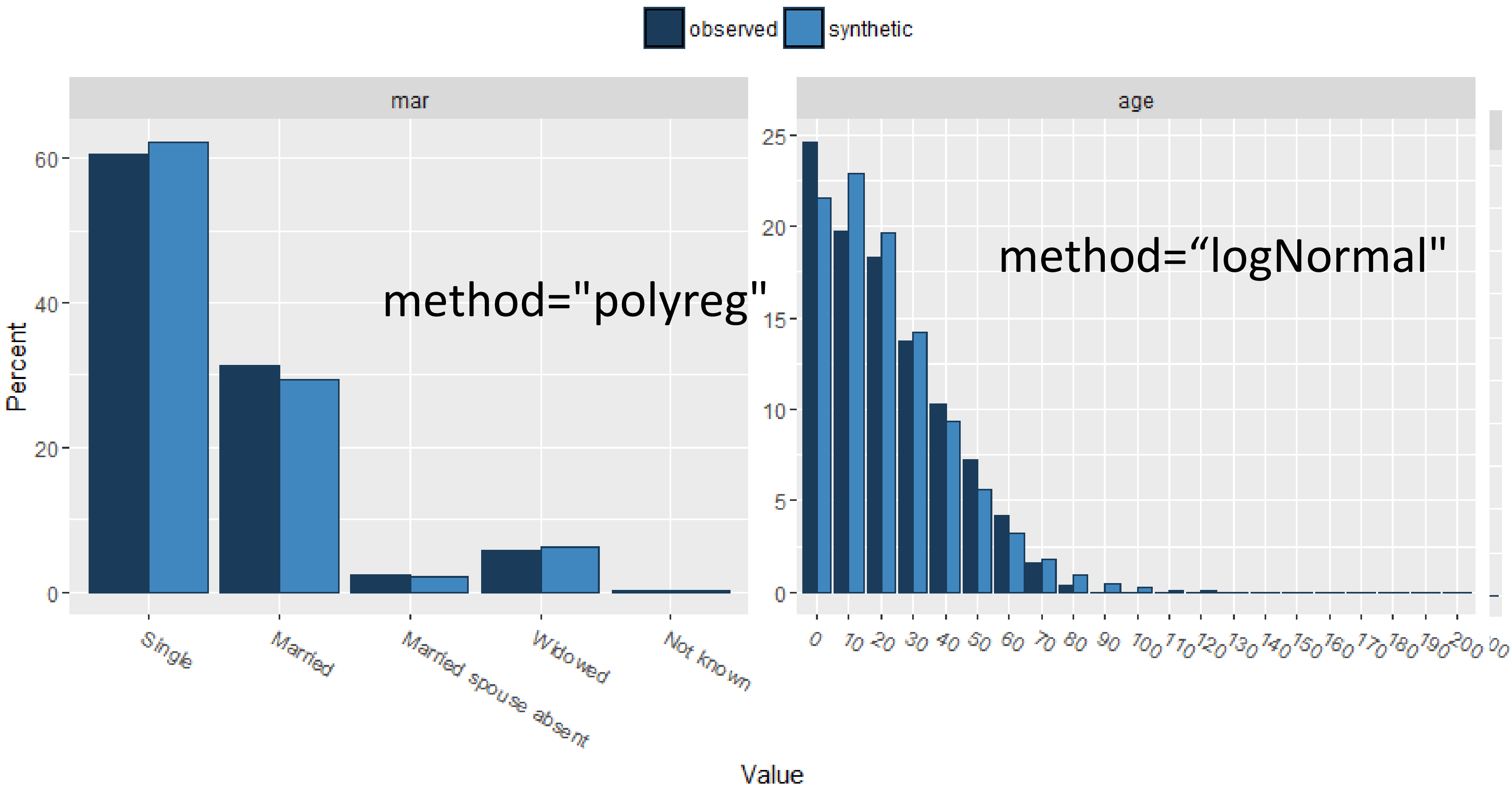
Overview of **synthpop** functions



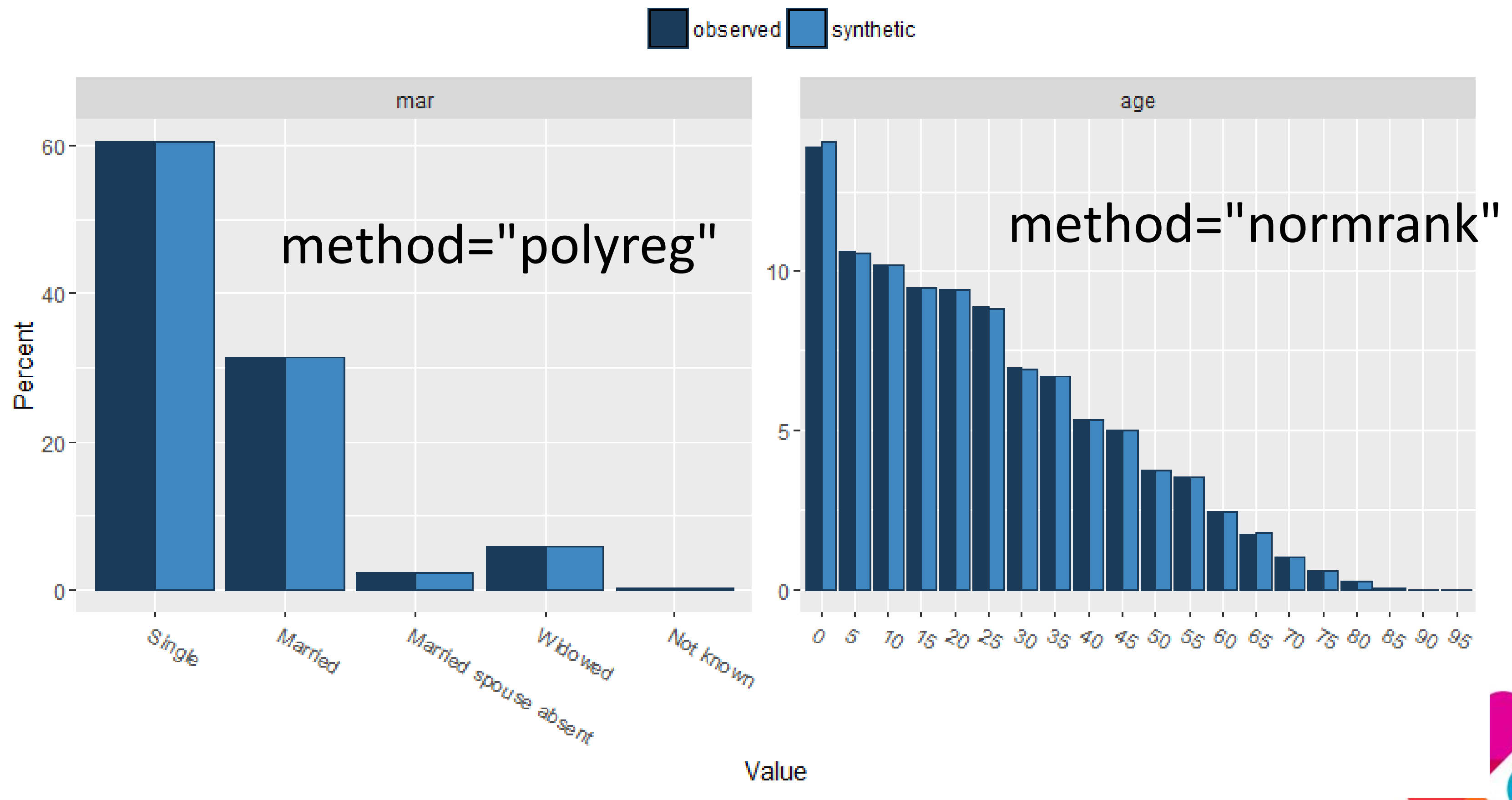
Utility measures – can only be used by the synthesiser

- **Specific** measures: `compare()`
  - Individual variables or two way comparisons
  - Comparing model fits
- **General** measures: `utility.synds()`
  - Based on propensity scores
  - Based on possibly multiway cross-tabulations

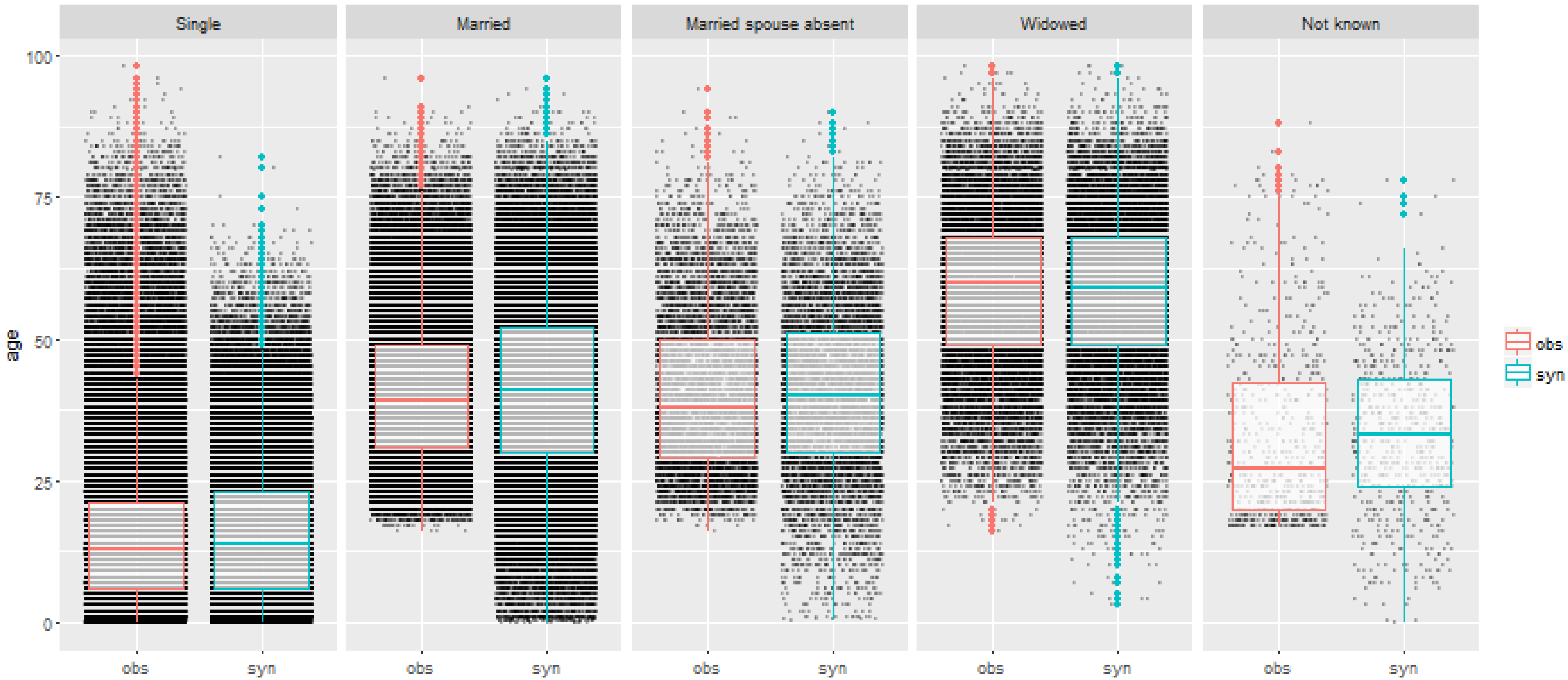
Comparing distributions (1)



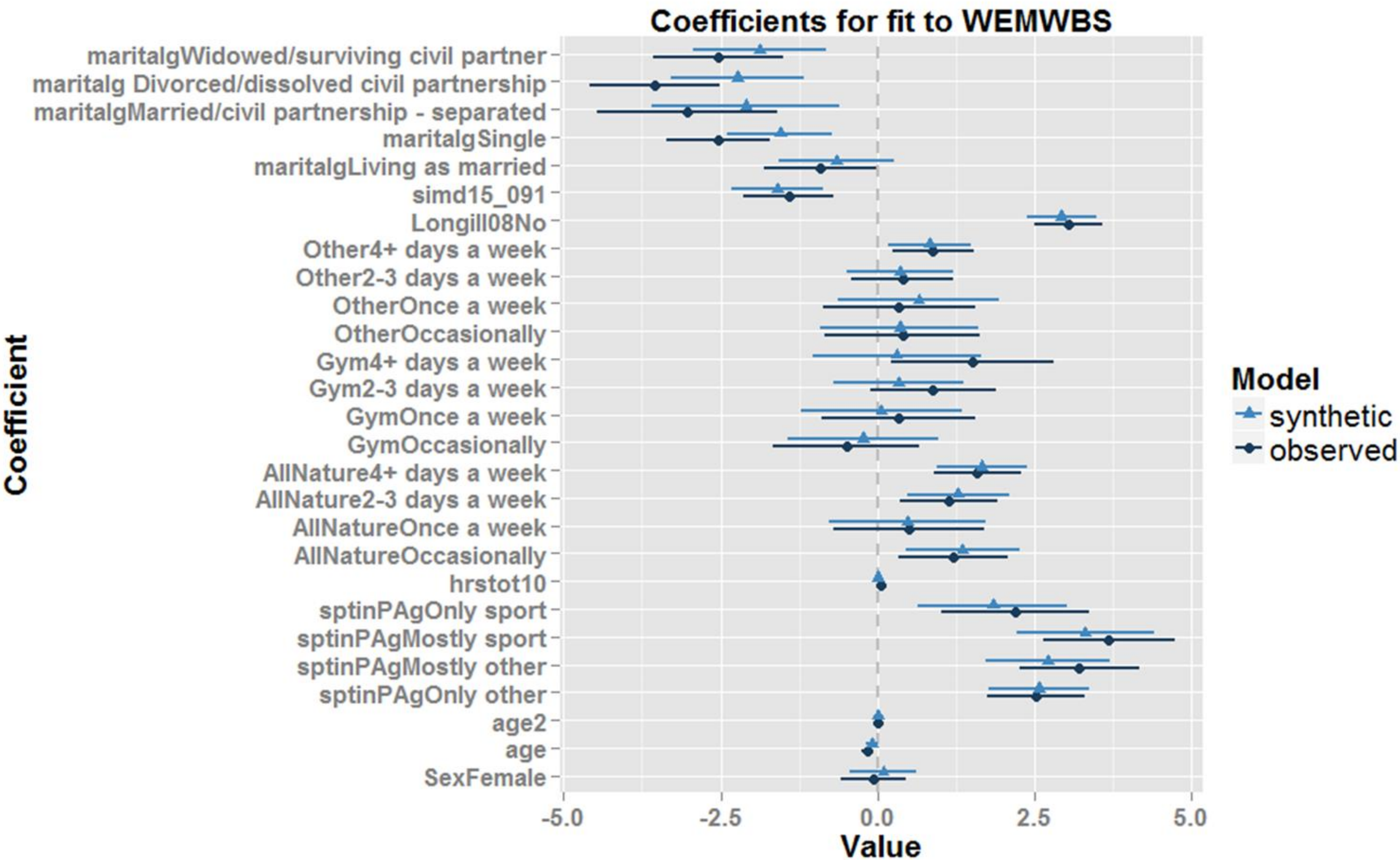
Comparing distributions (2)



Bi-variate visualisation



Comparing results of fitted models



General utility measure

U_{gen} derived from a propensity score method to distinguish original and synthesised data.

Results depend on choice of model for propensity score

Has a known χ^2 distribution if the synthesising model is correct.

Model	**U_{gen}**	**$U_{gen}/83$**	**Influential variables**
1 Parametric with square root normal for age	1,062	12.8	Age, ppreroom, mar, relat, parish
2 Parametric with Normal scores for age	293	3.54	Mar,relat,age, pperroom, parish

U_{tab} can be used to follow these up.

U_{tab} for cross tabulations

U_{tab} also has a known χ^2 distribution if the synthesising model is correct.

Table	U_{tab}	df	Ratio
Age by marital status	29,560	24	1,232
Relationship to head of household by marital status	1,716	36	47.7

sdc() & statistical disclosure control

- Data labelling: `label "false data"`
- Removing replicated uniques: `rm.replicated.uniques`
- Bottom- and top-coding: `recode.vars`, `bottom.top.coding`, `recode.exclude`

- At synthesis stage: `smoothing`, `minbucket`

```
sdc(syn.obj, real, label="false data",
    rm.replicated.uniques = TRUE,
    recode.vars = c("age","income"),
    bottom.top.coding = list(c(NA,85),c(NA,1500)))
```

Final thoughts

- We have provided some tools to create synthetic data
- Real data sets are complicated and large and there are still plenty problems to be overcome
- Larger problems concern persuading administrative data holders to allow the release of synthetic data
  - Hard to explain the process
  - Does not correspond to the usual methods (e.g. data swapping or top-coding) that are used by most data holders at present.
  - Formal disclosure control methods are not available
- But our public participation panel was very positive

Acknowledgements

- Other members of our team – most especially to Beata Nowok who has done the bulk of the work in creating the package and to Chris Dibben from ADRN S who guides us
- We acknowledge the help of the staff of the SLS-DSU staff in implementing the use of the *synthpop* package to create user extracts.
- Gillian Raab is funded by the UK Economic and Social Research Council's Administrative Data Research Centre – Scotland, Economic and Social Research Centre Grant ES/L007487/1.