Managing and modelling multiple-response data

Thomas Lumley
University of Auckland

@tslumley
notstatschat.rbind.io

Photograph by Melanie Lovell-Smith, from Te Ara

Multiple-response data

- Choose all that apply: what's your ethnicity, what social media do you use, what birds did you see today, what languages do you speak
- An enumerated character type
- factor/category/class is a special case

Photograph by Melanie Lovell-Smith, from Te Ara

Multiple-response data

Managing:

- Read, summarise, graph, tabulate, clean, tidy

Modelling:

- Regression outcome or regression predictor
- Multivariate models for tables

Photograph by Melanie Lovell-Smith, from Te Ara

Two R packages

Managing:

- `rimu`: **r**esponses **i**n **mu**ltiplex

Modelling:

- `rata`: **r**esponses **a**nalysed **t**ogether or **a**part

Photograph by Melanie Lovell-Smith, from Te Ara

Tidy multiple-response data

One observation — one cell

Represent multiple responses as an S3 vector class

- Multiple responses = one observation = one element
- One question = one column in dataset

Import from any format we can manage

Basic verbs mostly based on `forcats` package: `drop`, `union`, `intersection`, `lump`, `flatten`, `recode`, `reorder`, `count`

Tidy multiple-response data

Two approaches to underlying structure

- Base R: **binary presence/absence matrix**
  (cf `survival` package, Genstat, SPSS)

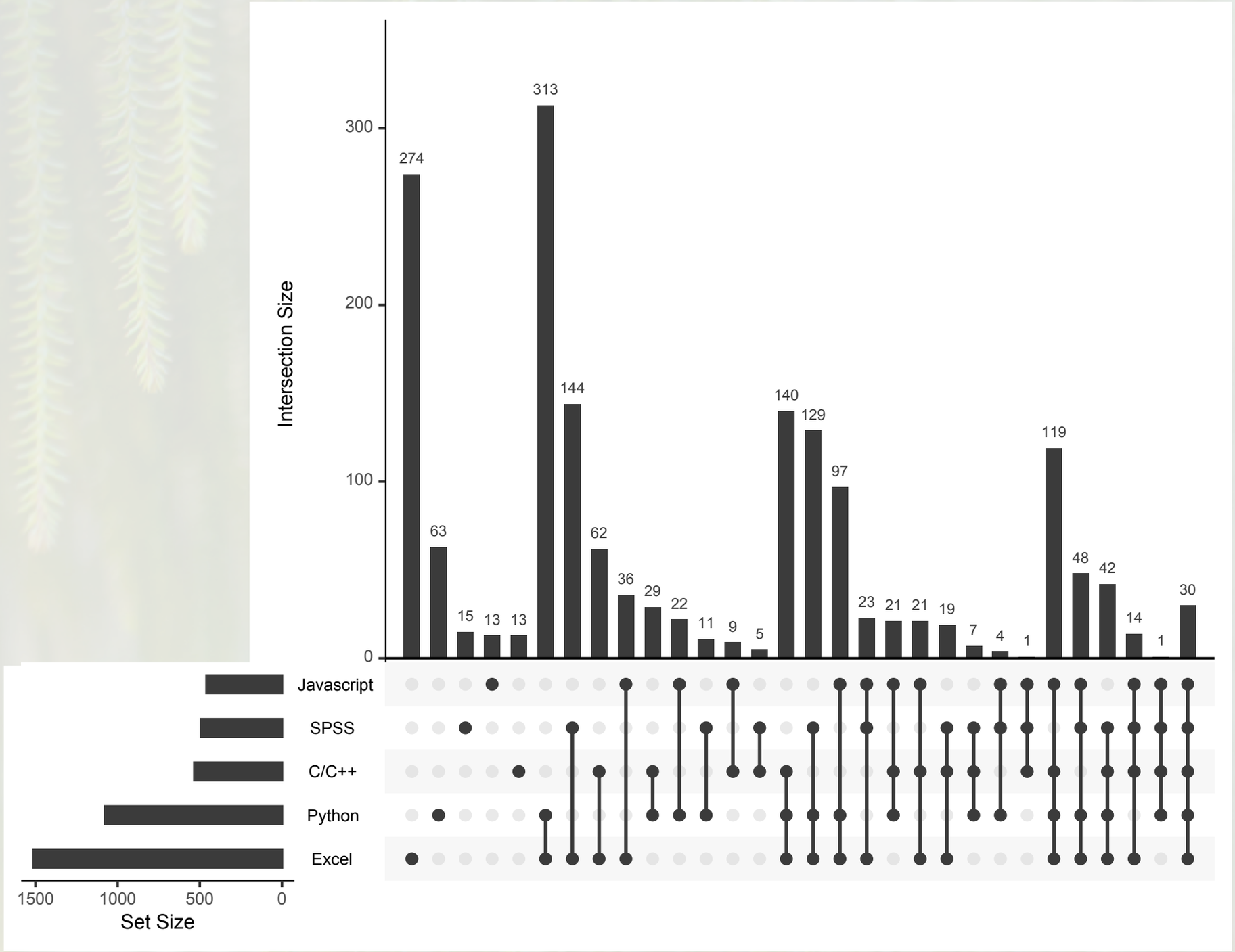
- Tidyverse: **vector of lists** (using `vctrs` package)


```
> head(other_software)
[1] "C/C++Python"
[2] "Excel+Tableau"
[3] "Excel+Tableau"
[4] "C/C++Excel+Go+Java+Javascript+Matlab+PHP+Python+Ruby+Visual
Basic"
[5] "Excel+Mplus+SPSS"
[6] “Excel+Javascript+PHP+Python+scala"

> table(other_software %hasany% c("GENSTAT","Genstat"))
FALSE  TRUE
 1836     2

> common<-mr_lump(other_software, n=15)
> mtable(common)
  C/C++ Excel Go Java Javascript Matlab PHP Python Ruby SAS
   533  1511  47  291        459    436  176   1076  89 393
 SPSS  SQL Stata Tableau Visual Basic Other
  493  59    89     435           360   403
```

> plot(common)


Intersection Size
300
200
100
0
274
63
15
13
13
313
144
62
36
29
22
11
9
5
140
129
97
23
21
21
19
7
4
1
119
48
42
14
1
30
Javascript
SPSS
C/C++
Python
Excel
1500
1000
500
0
Set Size

```
> rowpct(mtable(common, happy))
                1 2  3  4  5
C/C++           0 1  4 23 72
Excel           0 1  3 25 71
Go              0 2  2 19 77
Java            0 2  3 25 70
Javascript      0 1  3 22 74
Matlab          0 1  4 25 70
PHP             0 3  3 22 71
Python          0 1  3 23 73
Ruby            0 2  5 20 73
SAS             0 1  4 24 72
SPSS            0 1  5 21 73
SQL             0 0  0 17 83
Stata           0 1 10 24 65
Tableau         0 1  4 22 73
Visual Basic    0 1  3 24 73
Other           0 2  3 26 69
```

Modelling as outcome

Two main strategies

- Treat it as clustered **binary** data for logistic regression: mrglm (Agresti & Liu)
- Treat it as clustered **multinomial** data: mrloglin (Loughin et al), mrmultinom

Model fitting and inference is done by the VGAM and survey packages, rata does data expansion and rearrangement.

Modelling as outcome

Stacked records for each observed response

Example: where do you get your veterinary information?

```
mrglm(formula, data, family=binomial)
```

$$\text{logit}\, P[Y_{ik} = 1] = \alpha_k + \sum_{h=1}^{H} \gamma_{hk} I(\text{education}_i = h)$$

Photograph by Phil Brown on Wikipedia, public domain

Modelling as outcome

Stacked records for each observed response

Example: where do you get your veterinary information?

```
mrglm(formula, data, family=binomial)
```

```
present(sources)~value(sources)
present(sources)~value(sources)+as.numeric(education)
present(sources)~value(sources)+education
present(sources)~value(sources)*education
```

Photograph by Phil Brown on Wikipedia, public domain

Modelling as outcome

Stacked records for each observed response

Example: where do you get your veterinary information?

```
mrglm(formula, data, family=binomial)
```

0/1

```
present(sources)~value(sources)

present(sources)~value(sources)+as.numeric(education)

present(sources)~value(sources)+education

present(sources)~value(sources)*education
```

Photograph by Phil Brown on Wikipedia, public domain

Modelling as outcome

Stacked records for each observed response

Example: where do you get your veterinary information?

```
mrglm(formula, data, family=binomial)
```

0/1

Factor: consultant, govt, feed companies

```
present(sources)~value(sources)

present(sources)~value(sources)+as.numeric(education)

present(sources)~value(sources)+education

present(sources)~value(sources)*education
```

Photograph by Phil Brown on Wikipedia, public domain

Modelling as outcome

Stacked records for each observed response

Example: where do you get your veterinary information?

```
mrglm(formula, data, family=binomial)
```

0/1

Factor: consultant, govt, feed companies

```
present(sources)~value(sources)

present(sources)~value(sources)+as.numeric(education)

present(sources)~value(sources)+education

present(sources)~value(sources)*education
```

Level of study

Photograph by Phil Brown on Wikipedia, public domain

Modelling as predictor

One record: indicator variable for each level

```
mortality~wide(ethnicity)
```

Factor with stacked record for each observed level

```
mortality~each(ethnicity)
```

One record: single indicator

```
mortality~has(ethnicity, “Maori”)
```

Photograph by Phil Brown on Wikipedia, public domain

github/tslumley/rimu
and on CRAN

github/tslumley/rata

@tslumley
notstatschat.rbind.io

Photograph by Melanie Lovell-Smith, from Te Ara