Analysing larger data in R

The first rule of management is delegation

Thomas Lumley, University of Auckland, New Zealand

Do you even have big data?

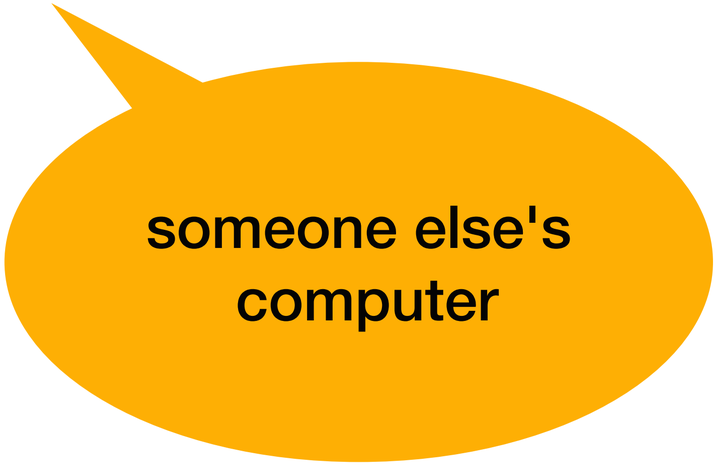
- Buy a bigger computer
- Or even just use your laptop

Using databases

- loading
- aggregation
- duckreg
- discretising first
- subsampling plus polishing

Basic idea

- Many statistical computations are sums of subsets of large vectors
- Relational databases (OLAP systems) are good at this
- When the large vectors live in the cloud, the sums are done there



Parquet

file format with efficient lookup

- `yellow_tripdata_202*.parquet` are a year of monthly taxi trip files

```
> dbGetQuery(con, "select count(*) from read_parquet('*.parquet') ")
  count_star()
1     40259776
```

← **Post**

Keith Ng
@keithng.bsky.social

R-related PSA: Parquet has really improved my quality of life, mostly just by things loading instantly thus giving me less time to switch over to social media while I'm waiting for things to load

January 28, 2025 at 11:53 AM

Everybody can reply

dbplyr/duckplyr

Relational algebra but in R

NZ census tables:

```
Year,Age,Ethnic,Sex,Area,count
2018,000,1,1,01,795
2018,000,1,1,02,5067
2018,000,1,1,03,2229
2018,000,1,1,04,1356
2018,000,1,1,05,180
2018,000,1,1,06,738
2018,000,1,1,07,630
2018,000,1,1,08,1188
2018,000,1,1,09,2157
```

plus file of labels for each field

nzcensus/censusduck.R

survey

complex survey analysis

`svydesign`, `svrepdesign` will accept a database table instead of a data frame

Each function call will automatically load just the variables it needs.

Data transformations happen as data is read: can work with read-only access.

Bioconductor has much more sophisticated version in DelayedArray (Peter Hickey)

Regression

frequency weights

Weighted linear regression was invented to work with averages of replicates

If X are covariate patterns, Y is mean for each pattern, and W is count, then

```
lm(Y~X, weights=W)
```

Same approach works with glms.

(python package duckreg: Akhil Rao; R package Grant McDermott)

Regression

other approaches

Compute $X^T X$ and $X^T Y$ in the database.

Read in all the data, but in chunks that R can handle (`biglm`), and compute QR decomposition of X incrementally

Subsampling and polishing

Read in a random sample of *n* observations from *N* and compute $\tilde{\beta}$

In the database, compute the sum of score function at $\beta = \tilde{\beta}$
(for logistic regression, just $x(y - \hat{\mu})$)

Do one-step Newton update.

Result is asymptotically efficient if $n = N^{1/2+\delta}$

`dbglm` package (Shanqing Cao: needs updating)

Federated learning?

Algorithms are similar to those for collaborative model fitting across institutions

- one round of communication <-> one database query

- database polishing <-> one-step updates
- ??? <-> surrogate likelihood (Jordan et al, Duan et al)

There are one-round or low-round federated algorithms for other problems, eg linear mixed model. Can they be adapted?

nature > nature communications > articles > article

Article | Open access | Published: 30 March 2022

DLMM as a lossless one-shot algorithm for collaborative multi-site distributed linear mixed models

$$y_{ij} = X_{ij}\beta + z_{ij}u_i + \epsilon_{ij}$$

Sufficient statistics are X^TX, X^Ty, y^Ty, n for each site, easily computed by database queries if i is suitable.

Any questions?