The importance of computing in statistical analysis

Examples from a package in R

Antonio Gasparrini

Medical Statistics Department
London School of Hygiene and Tropical Medicine

Centre for Statistical Methodology
LSHTM, 25 October 2013

Two questions

Take a sample of articles published 10 years ago. Two questions:

1. How many of the methods proposed in methodological papers have been actually **applied** in substantive analyses?
2. How many of the results published in the substantive papers could be exactly **reproduced** (if you had the data?)

Two questions

Take a sample of articles published 10 years ago. Two questions:

1. How many of the methods proposed in methodological papers have been actually **applied** in substantive analyses?

2. How many of the results published in the substantive papers could be exactly **reproduced** (if you had the data?)

Two questions

Take a sample of articles published 10 years ago. Two questions:

1. How many of the methods proposed in methodological papers have been actually **applied** in substantive analyses?
2. How many of the results published in the substantive papers could be exactly **reproduced** (if you had the data?)

Some answers

In both questions, the **software implementation** of the statistical methods is likely to play an important part

The application of statistical methodologies also depends on how well/much the software is **documented** and **user-friendly**

The reproducibility of results depends instead on availability of **details** on the statistical methods and (possibly) **scripts** of the analysis

Software requirements

- Reliable
- Well documented
- User-friendly
- General
- Flexible
- Fast

Software requirements

- Reliable
- Well documented
- User-friendly
- General
- Flexible
- Fast

Software requirements

- Reliable
- Well documented
- User-friendly
- General
- Flexible
- Fast

Software requirements

- Reliable
- Well documented
- User-friendly
- General
- Flexible
- Fast

Software requirements

- Reliable
- Well documented
- User-friendly
- General
- Flexible
- Fast

Software requirements

- Reliable
- Well documented
- User-friendly
- General
- Flexible
- Fast

Layers of complexity

Different **layers** of complexity:

1. Knowledge on the **theory of the statistical methods**, required for planning the analysis and interpret the results
2. Knowledge on the **use of the software**, required for correctly performing the analysis
3. Knowledge on the **computational aspects** of the statistical methods, required for developing software (and maybe more)

Layers of complexity

Different **layers** of complexity:

1. Knowledge on the **theory of the statistical methods**, required for planning the analysis and interpret the results
2. Knowledge on the **use of the software**, required for correctly performing the analysis
3. Knowledge on the **computational aspects** of the statistical methods, required for developing software (and maybe more)

Layers of complexity

Different **layers** of complexity:

1. Knowledge on the **theory of the statistical methods**, required for planning the analysis and interpret the results
2. Knowledge on the **use of the software**, required for correctly performing the analysis
3. Knowledge on the **computational aspects** of the statistical methods, required for developing software (and maybe more)

Reproducible research

An approach to research which ensures that all the steps of a study, results included, can be exactly repeated

A concept related to how much some research is **accessible**

In the context of statistical analysis, this translates in the **software**, **code** and **original data** to be available to the research community

Perspectives

- **Author** *vs.* **reader** of the article
- **Applied** *vs.* **methodological** research
- **Developer** *vs.* **user** of the software

Multivariate meta-analysis

Extension of traditional meta-analytical models to pool estimates of **multiple outcomes** from several studies

Applied in different contexts: clinical trials, network meta-analysis, multi-parameter functions

Statistically, just a **special case** of linear mixed-effects models (LME)

The model

For k outcomes $\hat{\mathbf{y}}_i$ estimated in in each study i:

$$\hat{\mathbf{y}}_i \sim \mathrm{N}\left(\mathbf{X}_i\beta\,,\,\mathbf{S}_i + \mathbf{\Psi}\right)$$

with $\mathbf{S}_i$ and $\mathbf{\Psi}$ as **within** and **between** (co)variance matrices

Alternatevely, as a linear mixed-effects (LME) model:

$$\hat{\mathbf{y}}_i = \mathbf{X}_i\beta + \mathbf{b}_i + \epsilon_i$$

with between-study random effects $\mathbf{b}_i \sim \mathrm{N}\left(\mathbf{0}, \mathbf{\Psi}\right)$ and within-study errors $\epsilon_i \sim \mathrm{N}\left(\mathbf{0}, \mathbf{S}_i\right)$

The model

For k outcomes $\hat{\mathbf{y}}_i$ estimated in in each study i:

$$\hat{\mathbf{y}}_i \sim \mathrm{N}\left(\mathbf{X}_i\boldsymbol{\beta}\,,\, \mathbf{S}_i + \boldsymbol{\Psi}\right)$$

with $\mathbf{S}_i$ and $\boldsymbol{\Psi}$ as **within** and **between** (co)variance matrices

Alternatevely, as a linear mixed-effects (LME) model:

$$\hat{\mathbf{y}}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{b}_i + \boldsymbol{\epsilon}_i$$

with between-study random effects $\mathbf{b}_i \sim \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Psi}\right)$ and within-study errors $\boldsymbol{\epsilon}_i \sim \mathrm{N}\left(\mathbf{0}, \mathbf{S}_i\right)$

Software implementation

Several estimators available: fixed, ML and REML, method of moments

Likelihood-based models can be fitted in **SAS** using the procedures for LME models

Ian White has developed the command mvmeta in **Stata**

The package mvmeta

Multivariate meta-analytical models implemented in the **R package mvmeta**, available in the R CRAN currently with version 0.3.5

Main functions:

- `mvmeta()`
- `predict()` and `blup()`
- `qtest()`
- `mvmetaSim()` and `simulate()`

See `example0.R`

Documentation

The documentation of a package consists of **help pages** for each function and optionally for the package, plus optionally package **vignettes**

The documentation, as mentioned earlier, plays a key role in usability of the package

In particular, the different sources of documentation should be **structured** accordingly with the different level of complexity

See `example1.R`

Usage and syntax

Usage of the functions mvmeta() and lm():

```
mvmeta(formula, S, data, subset, method="reml", model=TRUE,
  contrasts=NULL, offset, na.action, control=list())
```

```
lm(formula, data, subset, weights, na.action, method="qr",
  model=TRUE, x=FALSE, y=FALSE, qr=TRUE, singular.ok=TRUE,
  contrasts=NULL, offset, ...)
```

See example2.R

Reproducible analysis

Article:

"*Multivariate meta-analysis for non-linear and other multi-parameter associations*"

- **Paper** is open access
- **Web appendix** with info on the computational methods
- **R code**
- Access to the **data**

See **the paper** and **this web page**

Full or profiled likelihood

Defining $\boldsymbol{\Sigma}_i = \mathbf{S}_i + \boldsymbol{\Psi}$, the restricted log-likelihood ℓ_R is:

$$\ell_R = -\tfrac{1}{2}(n-q)\log\pi - \tfrac{1}{2}\sum_{i=1}^{m}\log|\boldsymbol{\Sigma}_i| - \tfrac{1}{2}\log\left|\sum_{i=1}^{m}\mathbf{X}_i^{\mathsf{T}}\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i\right| +$$
$$-\tfrac{1}{2}\sum_{i=1}^{m}\left[\left(\hat{\boldsymbol{\theta}}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}\right)^{\mathsf{T}}\boldsymbol{\Sigma}_i^{-1}\left(\hat{\boldsymbol{\theta}}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}\right)\right]$$

Given $\hat{\boldsymbol{\Sigma}}_i$:

$$\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\Sigma}}_i) = \left(\sum_{i=1}^{m}\mathbf{X}_i^{\mathsf{T}}\hat{\boldsymbol{\Sigma}}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{m}\mathbf{X}_i^{\mathsf{T}}\hat{\boldsymbol{\Sigma}}_i^{-1}\hat{\boldsymbol{\theta}}_i\ ,$$

$$V\left[\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\Sigma}}_i)\right] = \left(\sum_{i=1}^{m}\mathbf{X}_i^{\mathsf{T}}\hat{\boldsymbol{\Sigma}}_i^{-1}\mathbf{X}_i\right)^{-1}.$$

Full or profiled likelihood

Defining $\boldsymbol{\Sigma}_i = \mathbf{S}_i + \boldsymbol{\Psi}$, the restricted log-likelihood ℓ_R is:

$$\ell_R = -\tfrac{1}{2}(n-q)\log\pi - \tfrac{1}{2}\sum_{i=1}^{m}\log|\boldsymbol{\Sigma}_i| - \tfrac{1}{2}\log\left|\sum_{i=1}^{m}\mathbf{X}_i^{\mathsf{T}}\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i\right| + \\ -\tfrac{1}{2}\sum_{i=1}^{m}\left[\left(\hat{\boldsymbol{\theta}}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}\right)^{\mathsf{T}}\boldsymbol{\Sigma}_i^{-1}\left(\hat{\boldsymbol{\theta}}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}\right)\right]$$

Given $\hat{\boldsymbol{\Sigma}}_i$:

$$\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\Sigma}}_i) = \left(\sum_{i=1}^{m}\mathbf{X}_i^{\mathsf{T}}\hat{\boldsymbol{\Sigma}}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{m}\mathbf{X}_i^{\mathsf{T}}\hat{\boldsymbol{\Sigma}}_i^{-1}\hat{\boldsymbol{\theta}}_i\ ,$$

$$V\left[\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\Sigma}}_i)\right] = \left(\sum_{i=1}^{m}\mathbf{X}_i^{\mathsf{T}}\hat{\boldsymbol{\Sigma}}_i^{-1}\mathbf{X}_i\right)^{-1}\ .$$

Fitting procedure

$$\ell_R = -\tfrac{1}{2}(n-q)\log\pi - \tfrac{1}{2}\sum_{i=1}^{m}\log|\boldsymbol{\Sigma}_i| - \tfrac{1}{2}\log\left|\sum_{i=1}^{m}\mathbf{X}_i^{\mathsf{T}}\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i\right| + \\ -\tfrac{1}{2}\sum_{i=1}^{m}\left[\left(\hat{\boldsymbol{\theta}}_i - \mathbf{X}_i\hat{\beta}\right)^{\mathsf{T}}\boldsymbol{\Sigma}_i^{-1}\left(\hat{\boldsymbol{\theta}}_i - \mathbf{X}_i\hat{\beta}\right)\right]$$

Computational issues with:

- Full-matrix or study-component approach
- Derivatives of ℓ_R Ψ
- Positive-definiteness of Ψ
- Structuring Ψ

Fitting procedure

$$\ell_R = -\tfrac{1}{2}(n-q)\log\pi - \tfrac{1}{2}\sum_{i=1}^{m}\log|\boldsymbol{\Sigma}_i| - \tfrac{1}{2}\log\left|\sum_{i=1}^{m}\mathbf{X}_i^{\mathsf{T}}\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i\right| + \\ -\tfrac{1}{2}\sum_{i=1}^{m}\left[\left(\hat{\boldsymbol{\theta}}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}\right)^{\mathsf{T}}\boldsymbol{\Sigma}_i^{-1}\left(\hat{\boldsymbol{\theta}}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}\right)\right]$$

Computational issues with:

- Full-matrix or study-component approach
- Derivatives of ℓ_R $\boldsymbol{\Psi}$
- Positive-definiteness of $\boldsymbol{\Psi}$
- Structuring $\boldsymbol{\Psi}$

Fitting procedure

$$\ell_R = -\tfrac{1}{2}(n-q)\log\pi - \tfrac{1}{2}\sum_{i=1}^{m}\log|\boldsymbol{\Sigma}_i| - \tfrac{1}{2}\log\left|\sum_{i=1}^{m}\mathbf{X}_i^{\mathsf{T}}\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i\right| + \\ -\tfrac{1}{2}\sum_{i=1}^{m}\left[\left(\hat{\boldsymbol{\theta}}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}\right)^{\mathsf{T}}\boldsymbol{\Sigma}_i^{-1}\left(\hat{\boldsymbol{\theta}}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}\right)\right]$$

Computational issues with:

- Full-matrix or study-component approach
- Derivatives of ℓ_R $\boldsymbol{\Psi}$
- Positive-definiteness of $\boldsymbol{\Psi}$
- Structuring $\boldsymbol{\Psi}$

Fitting procedure

$$\ell_R = -\tfrac{1}{2}(n-q)\log\pi - \tfrac{1}{2}\sum_{i=1}^{m}\log|\boldsymbol{\Sigma}_i| - \tfrac{1}{2}\log\left|\sum_{i=1}^{m}\mathbf{X}_i^{\mathsf{T}}\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i\right| + \\ -\tfrac{1}{2}\sum_{i=1}^{m}\left[\left(\hat{\boldsymbol{\theta}}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}\right)^{\mathsf{T}}\boldsymbol{\Sigma}_i^{-1}\left(\hat{\boldsymbol{\theta}}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}\right)\right]$$

Computational issues with:

- Full-matrix or study-component approach
- Derivatives of ℓ_R $\boldsymbol{\Psi}$
- Positive-definiteness of $\boldsymbol{\Psi}$
- Structuring $\boldsymbol{\Psi}$

The importance of computing

Nowadays, computing plays an essential role in statistical analysis

Beside obvious advantages, this also presents some problems about reproducibility

Also, the use of more complex computational techniques requires an expertise beyond the usual statistical skills

Software implementation

The implementation of statistical techniques in a software provides several benefits

However, software development takes time and efforts, and should be carefully planned

Documentation is essential to improve the usability of the software

Software and computing in general should be considered an inherent part of any statistical research