Good Statistical Practice (GSP): guidence for using R for scientific publication

Caroline X. Gao
Centre for Youth Mental Health (Orygen), University of Melbourne; School of Public Health and Preventive Medicine, Monash University
Email: caroline.gao@orygen.org.au

3 December 2020

Contents

Introduction

This short practice guidance is designed based on a few guidelines and practice principles, books and jounal articles, including:

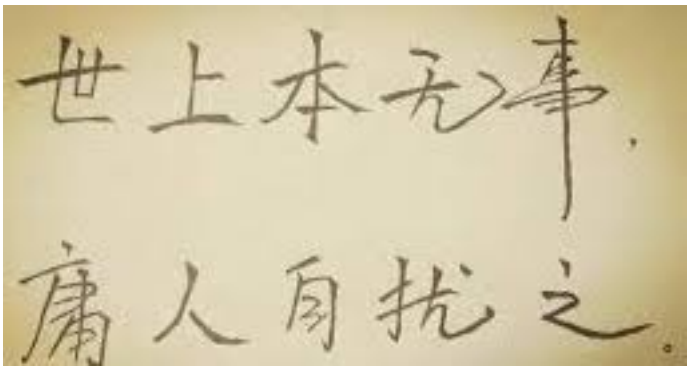
- ASA "Ethical Guidelines for Statistical Practice"
- Reproducible Research with R and R Studio
- Efficient R programming

The aim of this guidance is to promote accountability, reproducibility and integrity of statistical practice in Orygen health service and outcome research team. The guidance is divided into four parts: 1) planning for analysis, 2) data cleaning, 3) analysis, and 4) reporting.

Planning for anlaysis

Rule number 1: The analysis will need to be planed before you touch the data. Your analysis may deviate from the analysis plan to some degrees, which may relate to data integrity of the variable selected, model fitting factors (i.e. multicollinearity, heteroscedasticity etc) and change of research questions. However, unless your aim is purely exploratory (i.e. identify latent clusters) or predictive, your analysis should be guided by the analysis plan to prevent "fishing" results. Remember "If you torture the data long enough, it will confess to anything - Ronald H. Coase", which is really against the scientific principle. The scientific evidence is based on a collection of studies and findings rather than a single paper. If you have a negative finding, you are obligated to publish it !!!

Rule number 2: Be aware of the "Complexity Bias". "Life is really simple, but we insist on making it complicated".



The famous quote is not by Confucius (it is from the New Book of Tang published about 1500 years after his death), and it is not well translated, but you get the idea. We often find it easier

to face a complex problem than a simple one, and often feel that a complex solution must be better than an easier one. However this is often not true in applied statistics.

More often than not, ingenious statistical designs are surprisingly simple. An famous example is the Cox model for survival analysis, which simplifies the needs to describe the baseline hazard function with proportional hazard assumption. Another example, Dijkstra's algorithm (algorithm for finding the shortest paths between nodes in a graph), the author, Edsger Dijkstra (the author), once said "One of the reasons that it is so nice was that I designed it without pencil and paper. I learned later that one of the advantages of designing without pencil and paper is that you are almost forced to avoid all avoidable complexities."

This is the same with your analysis, always force yourself to avoid complexities if you could achieve what you need with a simpler model. There are numerous benefits for simpler models, i.e. less likely to over-fitting with your model, easier to communicate with your audience, less likely to make mistakes etc. If a logistic regression works, there is no need to use a structure equation model.

Rule number 3: Get your analysis plan approved or reviewed. I think we all do this to some degrees. Some studies have strict protocols on who and when the analysis plan will need to be approved together with other ethical requirements. Other studies will only require you to discuss with your supervisors. Regardless, it will be better to have a written analysis plan with review and/or approval and avoids confusions down the track. Sometimes you might want or need to preregister your study, which is considered as a part of the open science practice.

Last rule: Choose a script language (by no means SPSS) for your analysis. If you have already started your analysis in SPSS, please abandoned it ASAP. My metaphor: If R is in its early adolescent years, SPSS is a toddler and it suffers from the Peter Pan Syndrome (it will never grow up).

Sorry for being a bit offensive, but the reality is that most of the skilled SPSS users that I know have already or have been considering to learn R or Stata. The rest of them are no longer doing much analysis. So to avoid long term pain, change to R (Stata is also good, since this is a practice guide for R, I won't touch too much on Stata). There are lots of good online training materials for R:

- R for Reproducible Scientific Analysis from Software Carpentry
- R Programming on Coursera by Roger Peng,
- An Introduction to R
- R for Data Science.

If you are a skilled R user but do not use Rmarkdown or Bookdown, I would also recommend you to read Yihui Xie's two books:

- R Markdown: The Definitive Guide
- bookdown: Authoring Books and Technical Documents with R Markdown (for advanced users)

Well, all books from bookdown.org are worth reading.

One more remark: you can use causal diagrams to assist the design your analysis.

R setup, project managment and Rmarkdown

Setup R on your computer

One thing that you have to remember: R is a fast evolving language. It has a new version every few months (sometimes two versions in one months) with funny names :)

```
library(rversions)
tail(r_versions())
```

```
##     version                date                 nickname
## 115   3.6.2 2019-12-12 08:05:03   Dark and Stormy Night
## 116   3.6.3 2020-02-29 08:05:16    Holding the Windsock
## 117   4.0.0 2020-04-24 07:05:34              Arbor Day
## 118   4.0.1 2020-06-06 07:05:16          See Things Now
## 119   4.0.2 2020-06-22 07:05:19        Taking Off Again
## 120   4.0.3 2020-10-10 07:05:24 Bunny-Wunnies Freak Out
```

Although R kept on updating, you do not need to re-install R all the time. But I tend to update my R half-yearly. It doesn't take a long time to update everything now a days. The first thing that you need to do after installing R is to install your commonly used packages. A good way to install + load packages is to use the *pacman* package, which is much faster than typing install.package() and library(). I normally store the names of my commonly used packages somewhere. So when I need to re-install R, I will call *pacman* to install all of those packages for me at once (only takes about 10 minutes).

```
#load libraries
library(pacman)
p_load("dplyr", "tidyr", "ggplot2")
```

Project managment with R

Use Rmarkdown/Rnotebook for everything

Data cleaning

Inporting data

Tidy-version data cleaning routine

Notes for yourself and others

validity checking

Common pitfalls

Documentation for the never ending data cleaning process

Analysis

Exploratory phase

Statistical modeling with R

Extract results

Advanced topics

Loops

Functions

Render anlaysis results from different dataset with the same rmarkdown file

Reporting

One-stop shop

A good graph takes forever

Write up of the analysis results

- Report the nature and source of the data, validity of instrument and data collection process (i.e. response rate and any possible bias).
- Report any data editing procedures, including any imputation and missing data mechanisms
- When reporting analyses of volunteer data or other data that may not be representative of a defined population, includes appropriate disclaimers and, if used, appropriate weighting.

- Include the complete picture of the analysis results, which may require presenting tables and figures in appendix tables. For example, when reporting a series of multivariate regression models between an exposure and different outcomes, you can choose to include a summary table of adjust coef between exposure and different outcomes in the main text and include all the individual regression model results in the Appendix. The reader can use the appendix tables to understand the impact of confounding variables in the model.
- Report prevalence of outcomes or weighted prevalence of outcomes for representative samples.
- Report point estimate, 95% confidence interval and p-value in results
- Use graphical representations for reporting interaction effects (marginal plot)
- Acknowledges statistical and substantive assumptions made in the execution and interpretation of any analysis.
- Reports the limitations of statistical inference and possible sources of error.
- Where appropriate, addresses potential confounding variables not included in the study.
- Conveys the findings in ways that are meaningful and visually apparent and to the user/reader. This includes properly formatted tables and meaningful graphics (use guidelines by Gordon and Finch (2015)).
- To aid peer review and replication, shares the data (or synthetically generated data) used in the analyses whenever possible/allowable
- Provide all analysis code either as an Appendix or in open repositories such as Github

Advanced topics

Write a paper using R

Advanced Latex

Version control

Version control framework

Github

Publication

Reference

Gordon, Ian, and Sue Finch. 2015. "Statistician Heal Thyself: Have We Lost the Plot?" *Journal of Computational and Graphical Statistics* 24 (4): 1210–29. https://minerva-access.unimelb.edu.au/bitstream/handle/11343/55491/Gordon%20and%20Finch%202014%20DOI%20version.pdf;jsessionid=557C01A51F9EC550925E94F564ED970A?sequence=1.