Project Protocol

1. Title

Evaluating Statistical Quality in Health and Biomedical Research

2. Project team roles & responsibilities

Lee Jones (PhD Candidate)
Associate Professor Dimitrios Vagenas (Primary Supervisor)
Professor Adrian Barnett (Associate Supervisor)

The PhD candidate will be responsible for all aspects of the project, including planning, analysis and dissemination of results. Associate Professor Vagenas and Professor Barnett will provide supervision and be involved in planning and critical review of all results and publications.

3. Background information

3.1. Project outline

This study will examine statistical quality in health and biomedical research by focusing on linear regression models which provide foundational knowledge used throughout health and medical research. We are hoping to recruit 40 statisticians to review 100 randomly selected published papers from PLOS ONE, a large medical and science journal at the forefront of open data access. Each selected paper will be rated twice to ensure reliability of conclusions. Our results will highlight the most common issues for regression analyses and demonstrate where training and reporting guidelines need to be strengthened.

3.2. Introduction/background information

To improve the proportion of research that is translated from the lab into clinical practice, it is vital that all decisions are based on robust evidence. Statistical models provide us with tools to understand the variability of data allowing us to estimate the effectiveness of new treatments or understand underlying relationships in health systems. Unfortunately, when statistical methods are used poorly, they can provide misleading results leading to wasted resources and patients receiving ineffective or even harmful treatments.

There are many things to consider when analysing data including, the type of data, and whether it is continuous or categorical data. Statistical tests require certain assumptions to be met in order for the results to be valid. If assumptions of tests are not met the results may be misleading. At best this may cause estimates to be inaccurate without changing the conclusion of the study. At worst assumption violation can cause results to be completely invalidated and the original conclusions are overturned.

Discussion of statistical assumptions are frequently absent from publications (Thiese, Arnold, and Walker, 2015), with one study in the biomedical area showing assumptions being mentioned in only 20% of papers (Fernandez-Nino, Hernandez-Montes, and Rodriguez-Villamizar, 2018), while another study in psychology found 92% of articles were unclear about their assumption checks (Ernst and Albers, 2017). It is estimated that over half million papers

per year are published in journals (based on data from *PubMed*) with most research papers including statistical methods and interpretation.

3.3. Rationale/justification

This study aims to contribute to this growing area of research by exploring current statistical practice and identify common statistical misconceptions and errors made by researchers. We aim to make recommendations for improving statistical training for health researchers. Currently there are few studies examining assumptions for linear regression models, with no previous studies in a sample where the original data are available. This project will also be unique, as to this author's knowledge, no studies have used a sample of statisticians to review statistical assumptions with standard practice using two raters.

4. Study objectives

4.1. Hypotheses

The primary aim is to estimate the prevalence of statistical behaviours and will not be based on hypothesis testing.

Hypotheses that will be used are:

H1: There is an association between how p-values are displayed (categorical/continuous) and discussing the scientific importance of parameter estimates
H1: There is an association between having a strategy to check assumptions and performing residual checks (correctly checking assumptions)
H1: Statistician ratings of the same paper are reliable (reliability >= 0.7)

4.2. Research questions/ aims

- What is the prevalence of publication author teams who have demonstrated in their manuscript that they have checked linear regression assumptions?
- Are author teams checking assumptions correctly?
- What is the prevalence of statistical reporting behaviours?
- Are author teams who report P-values using categories more likely to not have discussed the scientific importance of the parameter estimates?
- Are author teams who have a strategy for assumptions more likely to perform residual checks?
- What is the agreement in the statisticians' rating of statistical quality?

4.3. Outcome measures

- Proportion of papers which check linear regression assumptions
- Of the papers which mentioned assumptions, the proportion which were correct
- Proportion of papers using recommended statistical reporting behaviours (e.g. 95% confidence intervals reported)
- Agreement between statistician raters of the statistical assumptions

5. Study design and survey development

5.1. Study type/design

Cross-sectional observational

5.2. Survey development

The survey was developed to gain an understanding of current reporting practices for linear regression analysis and adapted from the Statistical Analyses and Methods in the Published Literature (SAMPL) regression guidelines (Lang and Altman, 2013). A literature review was undertaken to identify common errors made by researchers when reporting linear regression, a comprehensive list of 55 questions was developed to assess statistical quality. It was decided by the research team, which consists of three accredited biostatisticians, to reduce survey burden by substantially reducing the number of items. The survey was shortened by focusing on only questions important to assessing assumptions and interpretation of linear regression, wording was improved by the research team, through an iterative process of reviewing papers. Five independent experts (four biostatisticians and an epidemiologist) assessed the survey for readability and length, by reviewing two randomly selected papers. Feedback was used to further reduce the questions to the current checklist of 27 items, which can be seen in the Appendix.

6. Randomisation

The randomisation process of selected papers will occur in two steps, as described below.

6.1. Article selection and randomisation

Articles will be selected from PLOS ONE which have the term 'linear regression' in the methods and results section from 2019 using the *rplos* package in R (Chamberlain, Boettiger, and Ram, 2018). Papers that match the inclusion criteria (see below) will then be randomly ordered and the first 100 papers meeting the exclusion criteria will be selected. A full list of DOIs of included and excluded papers will be made available for transparency.

Inclusion and exclusion criteria

Inclusion criteria:

- PLOS ONE articles published between January 1st 2019 to December 31st 2019
- "Linear regression" in the article's methods and results section
- Subject area of health
- Original research articles, not editorial, commentaries, etc

Exclusion criteria:

- Linear regression models that have accounted for clustering or random effects.
- Non-parametric linear regression, Bayesian or other alternative methods to linear regression
- Linear regression which was not a part of the primary analyses of the article and was related to pre-processing the data or verify an instrument or method of data collection. An example of this is linear regression used to calibrate an instrument to a reference sample.

Exclusion criteria were selected to make models comparable by excluding analysis which do not have the same assumptions or are more complex in nature. Using automation via a computer algorithm to exclude these papers may also exclude papers that incorrectly use linear regression, for example, ignoring clustered data. Therefore, the primary researcher will read papers starting with the first in the random series until a total of 100 papers is reached that meet inclusion and exclusion criteria. The number of papers excluded and the reasons will be reported.

Our approach for selecting papers will miss papers that used linear regression but did not mention this in their material and methods section. Finding these papers would require the researchers to read a wide selection of papers that would be time consuming and may not yield many additional papers.

6.2. Random allocation of papers to statisticians

The second step is to allocate the papers to statisticians. This will be achieved using a one-way random design for inter-rater reliability of the statistician. Fleiss (1981) recommends if there is no interest in comparing mean of several raters, then a simple random sample of raters from the overall pool of raters can be chosen. Hence, we will therefore randomly allocate papers using the following approach:

- 5 papers will be randomly allocated per statistician
- Papers will then be randomly reallocated to different statisticians, ensuring that no statistician is given the same paper twice

The final number of reviews per statistician may not be exactly five because some statisticians may not complete all five reviews. In this case we will consider: i) asking reviewers who have already completed five reviews to complete one more review, or ii) recruiting additional statistical reviewers.

7. Study population

7.1. Participants

40 statisticians will be recruited. Profession accreditation in the statistical field is voluntary, this means only a very small proportion of statisticians are accredited. Statistician often come from diverse backgrounds such as formal statistics degree but commonly other fields such as ecology, psychology, economics and may identify as statisticians, data scientists and data analysts.

7.2. Statistician Inclusions

Statisticians will have to either be employed or previously employed as a statistician, data scientist or data analyst.

7.3. Conflict of interest

Statisticians will be asked if they have a perceived conflict of interest in reviewing papers, e.g. one of the paper authors is a colleague. If a conflict of interest is identified, a different paper or set of papers will be sent to the statistician.

7.4. Recruitment strategies, timeframe

Given the diversity roles that may be identified as a statistician, participants will be target via email and Twitter through professional societies, universities and organisations such as CSIRO.

The recruitment period will be six months, statisticians will be given a total of 8 weeks to complete the survey, with reminders sent out every two weeks. If there is no response after this time, papers will be reallocated. Therefore, it is expected for data collection to take up to 12 months, with a further 12 months required for data analysis and publication. Starting July 2020 with completion in July 2022.

7.5. Consent

A returned consent form from participants will be taken as consent.

7.6. Participant withdrawal

Participants may withdraw from the study at any time. If this occurs, providing they have consented their partial response may be used. If participants do not wish for their partial/complete responses to be used in the study, they can contact the research team to withdraw their response. Should this occur, their response will be archived but not used in the research. Participants can also request for any information already obtained from them to be destroyed upon withdrawal.

8. Procedures

8.1. Screening of participants

Emails and advertising for the study will clearly state inclusion criteria. Agreement of consent in the study will be evidence of employment with the following statement included in the consent form.

- Meet the selection criteria of having employment (current/past) as a statistician

8.2. Data collection

1. A survey with basic demographic information from statisticians, and 2. reviews of five published papers using linear regression (see appendix)

8.3. Data collection/gathering techniques: How will you collect/gather the information?

Data will be collected using Key Survey.

8.4. Impact of and response to missing data e.g. extrapolation; participant withdrawal

If papers are not reviewed by the statistician within 8 weeks, these papers will then be reallocated to another statistician to avoid missing data.

8.5. Safety and adverse effects/events

No adverse events are expected.

9. Statistical plan

9.1. Sample size

The primary outcome of this study is to understand current reporting practices of authors in published manuscripts in regard to linear regression with a focus on assumptions. Previous studies show that the prevalence of reporting assumptions ranges from about 0 to 13% with most assumptions being reported under 10% of the time. Prevalence of reporting behaviours will be estimated by a random sample of papers meeting the search criteria of 'linear regression'. In 2018 there were 1,152 articles which satisfied these criteria.

It was deemed feasible to recruit 40 statisticians. From our own experience and through feedback we got during the development stage, it was feasible and not too onerous to have each statistician review five papers. This will achieve a sample size of 100 papers with each

paper rated twice to ensure robustness of results. This approach is perfectly acceptable as an alternative to conventional sample size calculation as explained by (Bacchetti, 2010). This would result in being able to detect a sample proportion for any statistical assumption of 0.050 using a two-sided 95% confidence interval with a width equal to 0.100 and was calculated using a test for one proportion with exact Clopper-Pearson confidence intervals, using PASS (2019) version 12. This is a relatively narrow confidence interval that will be able to differentiate between statistical practices that were completed well or badly.

9.2. Data analysis

This is an exploratory study to examine the statistical quality of published papers in the health and biomedical field. Reporting behaviours are categorical and will be described using frequency and percentages and be presented with 95% confidence intervals. Chi-square will be used to examine if there is an association between reporting behaviours. Fisher's exact will be used if expected counts are small. Reliability of raters will be examined using Gwet-kappa (Gwet, 2008). In the case of disagreement between raters, a third statistician will be used to clarify disagreements. The level of agreement will be reported and all raw data including the third rating will be available upon publication of results. A quality score will be created with questions that are common to all papers. Generalised mixed models will be used to assess if years of rater experience or the quality of the paper effect reliability, and to account for missing data if required.

An initial analysis will be created in R using scrambled data. This will be sent around the group for review. Once all investigators are happy with the approach, the data will be unscrambled and the real results shown. This approach helps iron-out errors in the code and/or data before seeing the real results, and hence reduces the chances of bias in the results.

10. Data management and record keeping

10.1. Confidentially and privacy

Statisticians reviewing papers will be deidentified. Participants (Statisticians) will be provided with a participant information sheet that detail privacy and confidentiality issues. The data from reviewed papers is publicly available will be identifiable through doi's and any code reproduced will be stored on GitHub.

Data from reviewed papers will include frequencies of statistical assumptions and other statistical behaviours within papers. Review data will be made available for each paper with the reviewer (statistician) information de-identified.

10.2. Data security

Raw and processed data will be stored on the QUT password accessed network drive. All data will be backed up in two secure password protected locations regularly.

All data will be retained for a minimum of 5 years after publication. Data will be archived using the Research Data Storage Service.

10.3. Record retention

A hard copy of consent forms from statisticians will be filed in a locked cabinet at for

15 years in a secure location at Q block QUT Kelvin Grove building. Statistician reviews of papers will be retained for a minimum of 5 years. Data will be archived using the Research Data Storage Service.

10.4. Secondary use

Results will be published in an open access journal to allow results to be examined for reproducibility and accuracy. The raw data on ratings will be made available.

11. Resources

This project does not require any resources other than the PhD candidates and supervisors time and resources such as Keysurvey readily and freely available for QUT researchers.

12. Results and outcomes

12.1. Plans for return of results to participants

A brief report will be returned to statisticians highlighting the main results of the study, with more detail provided upon publication of results.

12.2. Plans for return of results to PLOS ONE

The results of this study will be provided to PLOS ONE upon publication, along with raw data and code required to reproduce the results in RMarkdown and shared using GitHub (version control and repository for code).

12.3. Publication plan

It is expected that this study will result in one publication in a Q1 journal and be presented at a conference.

12.4. Plans for sharing data and/or future use of data and or follow-up research

This study will be published in an open science journal requiring the original raw data to be shared.

12.5. Project closure processes

- Finalise publications
- Ensure all data is stored and archived adequately including meta data
- Update contact details with data storage and publications

13. References

Bacchetti, P (2010). Current sample size conventions flaws, harms, and alternatives. BMC medicine 8(1), 17.

Chamberlain, S, C Boettiger, and K Ram (2018). rplos: Interface to the Search API for 'PLoS' Journals. R package version 0.8.4. https://CRAN.R-project.org/package=rplos.

Ernst, AF and CJ Albers (2017). Regression assumptions in clinical psychology research practice a systematic review of common misconceptions. PeerJ 5, e3323.

Federer, LM, CW Belter, DJ Joubert, A Livinski, YL Lu, LN Snyders, and H Thompson

(2018). Data sharing in PLOS ONE: An analysis of Data Availability Statements. PloS one 13(5), e0194768.

Fernandez-Nino, JA, RI Hernandez-Montes, and LA Rodriguez-Villamizar (2018). Reporting of statistical regression analyses in Biomedica: A critical assessment review. Biomedica 38, 173–9.

Gwet, KL (2008). Computing inter-rater reliability and its variance in the presence of high agreement. British Journal of Mathematical and Statistical Psychology 61(1), 29–48.

Fleiss, JL (1981). Balanced incomplete block designs for inter-rater reliability studies. Applied Psychological Measurement 5(1), 105–112.

Lang, T and D Altman (2013). Basic statistical reporting for articles published in clinical medical journals: the SAMPL Guidelines. Handbook, European Association of Science Editor

PASS (2019). Sample size and power. https://www.ncss.com/software/pass/ (visited on 09/22/2019).

Thiese, MS, ZC Arnold, and SD Walker (2015). The misuse and abuse of statistics in biomedical research. Biochemia medica: Biochemia medica 25(1), 5–11.

14. Appendices

Survey Questions

Demographic questions

1. What role do you identify as?
2. What is your highest statistical/mathematical education?
3. How many years of experience do you have working as a Statistician/data analyst?
4. How did you find out about this study?

Paper Identification

- Please type the full name of the first author
- Do you have a conflict of interest in reviewing this paper? *

Statistical questions

1. How many linear regressions can you identify in this paper?
2. Does the paper mention a strategy for assessing linear regression assumptions?
3. What was checked with regards to normality?
4. How was normality assessed?
5. Did the authors check the linearity assumption?
6. How was linearity assessed?
7. Did the authors check the homoscedasticity assumption?
8. How was homoscedasticity assessed?
9. Did the authors check the independence of observations?
10. How was independence of observations assessed?
11. Was collinearity of X variables in models evaluated?

12. Have authors checked for outliers in their data?
13. What did they do with respect to outliers?
14. Were any continuous variables transformed, not including categorisation?
15. Were continuous variables on a very large/small scales in the model scaled appropriately?
16. Is there any process for selecting the variables included in the final model?
17. Which variable selection strategy was used?
18. Does the paper mention any statistical significance criteria for including variables?
19. What has been reported? B, CI, SE, R2, F/t, DF
20. Has the direction of the parameter estimates been interpreted?
21. Has the size of the parameter estimates been interpreted?
22. Have authors discussed the scientific importance of parameters estimates?
23. Have p-values been reported?
24. Was linear regression the main analysis in the paper?
25. Was most of the detail for the assumption checks in the supporting information (appendix)?
26. Rate the overall statistical quality of the paper
27. Do you have any other comments you would like to provide with respect to the statistics and their quality in the paper?

***Declaring conflict of interest**

A conflict of interest is anything that interferes with or could be perceived as potentially interfering with, a thorough and objective assessment of a manuscript. Common examples of conflict of interest may include:

- Recent/ current collaborations with any of the authors
- Direct competition or a history of scientific conflict with any of the authors
- An opportunity to profit financially from the work

Do not accept a review of a paper if you have a conflict of interest, or don't feel able to give an objective assessment. If you have a conflict of interest, please contact our research team at lee.jones@qut.edu.au and we will arrange for an alternative paper to be reviewed.