Project Protocol

1. Title

Reproducibility of linear regression in health and biomedical research

2. Project team roles and 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, analyses, and dissemination of results. Associate Professor Vagenas and Professor Barnett will supervise and be involved in the planning and critical review of all results and publications.

3. Background information

3.1. Project outline

This study will examine the reproducibility of results from published papers in the health and biomedical area. We will examine 100 randomly selected papers that have used linear regression from *PLOS ONE*, a large medical and science journal at the forefront of open data access. We will try to replicate the published linear regression(s), checking the robustness of the results. 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

Research has highlighted that up to half of the literature previously reviewed has been found to have statistical errors (Gore, Jones, and Rytter 1977; Nuijten et al., 2016; Wullschleger et al. 2014). One in eight papers (12.5%) has inconsistent results that change the interpretation of conclusions. (Nuijten et al. 2016). The absence of discussion or checking of statistical assumptions in published papers (Thiese, Arnold, and Walker 2015), further exacerbates this issue, underscoring the urgent need for our proposed research.

This project will focus on computational and Inferential reproducibility. Computational reproducibility involves extracting the same results using the same data to verify computational and experimental procedures. Inferential reproducibility, on the other hand, refers to interpreting conclusions from a reanalysis of the original study. (Goodman, Fanelli, and Ioannidis 2016). Our research will gain an understanding of how conclusions may change by analysing the same data and making different analytical choices in cases where the original model has been mis-specified.

3.3. Rationale/justification

This study will make an original contribution by comprehensively assessing the statistical quality of the health literature using a random sample of papers that have shared their raw data. It will be one of the first to include an assessment of the validity of scientific conclusions made by authors, and empirically assess the consequences of assumption violations.

Checks and balances are required to maintain the quality of research and ensure scarce health resources are spent efficiently with maximum benefit to the community. Data sharing enables the reproducibility of data analysis, which likely improves the translation of research by increasing the accountability of researchers and institutions. Although more data are being shared, there has not been a commensurate increase in the verifications of the data shared and the analyses. Hence, this project will make an important and relatively novel contribution to that area.

4. Study objectives

4.1. Aim and hypotheses

The aim is to examine the overall quality of statistical methods in published papers and will not be based on hypothesis testing.

4.2. Research questions

- What proportion of published papers had data available?
- What proportion of the papers stated data was available and had data required to reproduce the analyses?
- What proportion of analyses/ published papers were computationally reproducible?
- What is the proportion of analyses/ published papers that violate statistical assumptions?
- For analysis that did not meet assumptions, did the results change substantially when appropriate analysis was used?
- Have studies with outliers reported them?
- Did studies routinely remove outliers without adequate explanation?
- What proportion of analyses/papers had multicollinearity issues?

4.3. Outcomes

- Proportion of papers where data to reproduce up to three linear regression analyses is available.
- Proportion of papers where authors provided data dictionary (none, partially described, fully described)
- Proportion of papers/analyses for which we could reproduce the results.
- Proportion of papers/analyses that violate statistical assumptions.
- Proportion of papers/analyses that have substantially changed results after reanalysis; this will be measured through change in coefficients (10%), range of confidence intervals (10%), change in model fit (R^2 and AIC), statistical significance, and significant direction change of coefficients. Analyses will be assessed on this range of measures and be categorised into minor change and substantial change.
- Proportion of papers/analyses with outliers not reported in the published paper.
- Proportion of papers/analyses with multicollinearity issues.

All proportions will be reported with a 95% confidence interval with the number of eligible papers as the denominator.

5. Study design

Cross-sectional observational

6. Article selection and randomisation

Articles will be selected from *PLOS ONE*, which has the term 'linear regression' in the methods 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 placed into a randomised order, and the first 100 papers

meeting the exclusion criteria will be selected. A complete list of DOIs of included and excluded papers will be made available for transparency.

6.1. 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 section
- Subject area of health
- Original research articles, not commentaries, editorials, etc

Exclusion criteria:

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

6.2. Linear regression models within papers

Each linear regression model within each paper will be identified and numbered. Up to three linear regression analyses will be selected per paper. When papers have more than three linear regression models, and there is a final model (or primary model of interest), the two other regression models will be chosen at random. If there is no final model, all three regression models will be randomly chosen. This limit of three is to keep the workload manageable and because of an expected strong correlation in practice in models from the same paper.

6.3. Recruitment

All data are publicly available, and no recruitment is required. The data for selected papers will be downloaded from *PLOS ONE* following randomisation. If data from an article has restricted access, this will be classed as a paper we could not verify. Data will not be requested from the authors.

6.4. Consent

All data are publicly available; in order to publish in *PLOS* journals, authors have to provide a data availability statement that states that data may be used for the reproducibility of research as outlined below by PLOS (full policy can be viewed at https://journals.plos.org/plosone/s/data-availability)(accessed May 31 2024.)

> PLOS journals require authors to make all data necessary to replicate their study's findings publicly available without restriction at the time of publication. When specific legal or ethical restrictions prohibit public sharing of a data set, authors must indicate how others may obtain access to the data.
>
> PLOS believes that sharing data fosters scientific progress. Data availability allows and facilitates:
>
> - Validation, replication, reanalysis, new analysis, reinterpretation, or inclusion into meta-analyses
> - Reproducibility of research

7. Procedures

7.1. Data collection

Data from papers will be collected in whatever format is available; this is expected to include xlsx, CSV, statistical packages (SPSS, Stata, etc.), and tables. All data relating to reproducibility will be converted and stored in CSV format. Published papers will be stored in PDF format. We will record the data formats used and whether the authors provided a data dictionary.

7.2. Ethical considerations

Corrections of published literature are common and may cause authors discomfort, as the time required to review and correct errors may be long. Basic statistical errors have been found in up to half of the literature previously reviewed, with one in eight papers having inconsistent results that change the interpretation of conclusions. Therefore, it is likely that this project will identify statistical errors that would ideally be corrected.

There is a small risk that statistical review could lead PLOS ONE to investigate problems in a paper further and result in a retraction. However, retractions are rare, 2 to 4 in 10,000 (Fernandes et al. 2023), retractions may cause reputational damage to authors. This study will not be involved in this process except for providing the initial expert statistical review; it is up to the journal, authors, and institutions to determine appropriate action. However, if a paper has such flaws that a retraction is warranted, it would be unethical for the current team to withhold this information since it could harm patients and waste clinical and research resources. We will share any concerns with PLOS ONE rather than engage in a direct discussion with other authors.

8. Statistical plan

8.1. Sample size determination

The purpose of this study is descriptive and not based on hypothesis testing. Its aim is to understand the prevalence of papers that have data available and are computationally and inferentially reproducible in a random sample of papers. The original sample size of 100 was based on understanding the prevalence of reported linear regression assumptions to be able to detect a sample proportion of 0.05 (5%) using a two-sided 95% confidence interval with a 5% margin of error, which is also adequate for the descriptive purposes of this project.

8.2. Statistical methods

Up to three linear regression analyses will be reproduced from each article with available data. Assumptions of linear regression will be examined using a mix of descriptive statistics, plots and tests. The models will also be checked for multicollinearity, the presence of outliers, and influential observations. Each regression analysis will be assessed using the steps below:

1. Are the residuals of the model independent of each other?

- Describe if the study design is theoretically independent, checking whether there are clusters such as hospitals or the same individual measured multiple times.
- If appropriate, check for serial correlation using the Durbin-Watson test by participant ID and/or time.

- If the study design is not independent (e.g., repeated data from individuals), a marginal/ mixed model will be fit to gain an estimation change in the coefficient and the degree of correlation.

2. Are the residuals of the regression model approximately normally distributed?

- Plot residuals using histogram and QQ-plot
- Describe residuals using descriptive statistics, including mean, SD, median min, max, skewness and kurtosis.
- Normality tests (Shapiro–Wilk test and Kolmogorov-Smirnov test).
- Decision will be based on plots and descriptive statistics.
- Normality tests will be used as a part of the overall picture. However, they will also be used to understand how often tests reject minor deviations from normality.
- Non-normality issues may be resolved using bootstrapping or data transformation.

3. Does the mean of the dependent variable Y change linearly with the model's parameters?

- Plot of residuals vs predicted values to assess the linearity of the overall model.
- Plot of residual vs continuous variables to assess the linearity of individual variables
- If a non-linear relationship is observed, the appropriate model for the data (e.g. quadratic, cubic) will be fit and assessed using residual plots and R^2.

4. Do residuals display homogeneity of variance?

- Plot of residuals vs predicted values to assess patterns such as funnelling.
- Plot of residuals vs key individual variables to assess patterns.
- Test for heteroscedasticity (studentized Breusch–Pagan test).
- Remedies for heteroskedasticity may include using weighted linear regression or robust standard errors to account for bias or transformation of the dependent variable.

5. Are there outliers and influential observations?

- Cook's distance and leverage will be calculated, examined for large values, and plotted to identify problematic data points.
- Sensitivity analysis will be performed for data with identified outliers and influential data points.
- Remedies may include bootstrapping, data transformation or robust regression.

6. Is there collinearity among the X variables?

- Using the Variance Inflation Factor (VIF) with values greater than five is problematic.
- Pairwise correlations between independent variables.
- Changes or problems with standard error in regression models
- A remedy to collinearity issues is to identify variables that are too highly correlated and remove the least important predictor using AIC and R^2.

Overall assumptions and model differences

This study is the second part of a larger statistical quality project; stage one of this project involves the selected papers being statistically reviewed by statisticians. This may help identify statistical problems that may be faced in reproducing results. Linear regression assumptions can be assessed individually; however, any remedies must be viewed in the overall framework of the model. For example, normality problems may be caused by large outliers.

Inconsistencies between original and reproduced regression results will be measured through changes in coefficients (10%), range of confidence intervals (10%), or change in model fit (R^2 and AIC), statistical significance, and significantly changed direction of coefficients. Forest plots will be used to display original and reproduced results. The key results (Z scores) will be standardised to compare estimates and confidence intervals across studies.

9. Data management and record keeping

9.1. Confidentially and privacy

All R code created to reproduce results from published papers will be shared on GitHub, with the identification of the research team and papers (DOIs).

9.2. Data security

Raw and processed data will be stored on the QUT password-accessed network drive.
All data will be backed up regularly in two secure password-protected locations.
All data will be retained for at least five years after publication and archived using the Research Data Storage Service. The research team are statisticians who regularly train on all aspects of data, including analysis, data privacy, and storage.

9.3. Record retention

All publication data will be stored for at least five years after the final publication.

9.4. Secondary use

Results will be published in an open-access journal to allow results to be examined for reproducibility and accuracy.

10. Resources

This project requires no resources other than the PhD candidate and supervisors' time.

11. Results and outcomes

11.1. Plans for return of results

Minor statistical errors will be reported through the PLOS One website, and major errors will be reported directly to PLOS ONE. All data will be reanalysed in a reproducible format using code with R Quarto and shared using GitHub (version control and repository for code). This approach provides complete transparency of results and allows errors to be detected by the broader research community and corrected promptly.

11.2. Publication Plan

This project will be published in a Q1 journal and disseminated at conferences.

11.3. Project Closure Processes

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

12. References

Fernandes, Bianca Barros Parron, Mustafa Reha Dodurgali, Carlos Augusto Rossetti, Kevin Pacheco-Barrios, and Felipe Fregni. 2023. "Editorial - the Secret Life of Retractions in Scientific Publications." *Principles and Practice of Clinical Research (2015)* 9 (1). https://doi.org/10.21801/ppcrj.2023.91.2.

Goodman, Steven N., Daniele Fanelli, and John P. A. Ioannidis. 2016. “What Does Research Reproducibility Mean?” *Science Translational Medicine* 8 (341): 341ps12-341ps12.

Gore, S. M., I. G. Jones, and E. C. Rytter. 1977. “Misuse of Statistical Methods: Critical Assessment of Articles in BMJ from January to March 1976.” *British Medical Journal* 1 (6053): 85–87.

Nuijten, Michèle B., Chris H. J. Hartgerink, Marcel A. L. M. van Assen, Sacha Epskamp, and Jelte M. Wicherts. 2016. “The Prevalence of Statistical Reporting Errors in Psychology (1985–2013).” *Behavior Research Methods* 48 (4): 1205–26.

Thiese, Matthew S., Zachary C. Arnold, and Skyler D. Walker. 2015. “The Misuse and Abuse of Statistics in Biomedical Research.” *Biochemia Medica* 25 (1): 5–11.

Wullschleger, Marcel, Soheila Aghlmandi, Marcel Egger, and Marcel Zwahlen. 2014. “High Incorrect Use of the Standard Error of the Mean (SEM) in Original Articles in Three Cardiovascular Journals Evaluated for 2012.” *PloS One* 9 (10): e110364.

PLOS Data availability statement. Accessed May 31, 2024. https://journals.plos.org/plosone/s/data-availability).