Strengthening Analytical Thinking for Observational Studies (STRATOS): COLLABORATIVE PROJECT ON INITIAL DATA ANALYSIS FOR REGRESSION MODELS

Lara Lusa 1,2 , Marianne Huebner 3 , Willi Sauerbrei 4 , Georg Heinze 5 on behalf of STRATOS TG3 and TG2

- Department of Mathematics, Faculty of Mathematics, Natural Sciences and Information Technologies,University of Primorska, Koper/Capodistria, Slovenia; lara.lusa@famnit.upr.si
- Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana,Ljubljana, Slovenia
- Department of Statistics and Probability, Michigan State University, East Lansing, USA
- Faculty of Medicine and Medical Center, University of Freiburg, Germany
- Medical University of Vienna, Center for Medical Data Science, Institute of Clinical Biometrics,Vienna, Austria

The complexity of observational studies requires expertise in multiple areas of statistics. Previous short reports in the Biometric Bulletin focused on specific STRATOS topic groups (TG) and panels, but collaborative projects have been ongoing. The TG3 developed a body of work on systematic approaches of initial data analysis (https://stratosida.github.io/papers.html) while TG2 publications aim to provide guidance for variable and function selection in multivariable analysis. Our collaboration focused on providing practical solutions for analysts to implement initial data analysis steps as a prerequisite to regression analyses. This provides knowledge about underlying features of the data and is needed to ensure the appropriateness of the statistical models and correct interpretation of results.

The initial data analysis topic group proposed an IDA framework that can be incorporated into the research flow [1], and includes data cleaning, data screening, possible updates of pre-planned statistical analyses, and reporting of IDA findings. We described ten simple rules to explain IDA and the benefits of adopting IDA in practice [2].

Data screening examines the properties of data before addressing the research question with the main statistical analysis. The aim is to obtain information relevant for subsequent intended statistical analyses. Such explorations investigate missing data and univariate and multivariable descriptions. Importantly, these explorations do not anticipate analysis directly related to the research question, namely associations between outcome and predictors are not explored. IDA findings support the original analysis strategy, can guide revisions and the correct presentation and interpretation of the results of the analyses. IDA strategies require careful consideration of the research question. IDA steps can then be incorporated in analysis plans to manage the scope and for reproducibility. Some decisions about the statistical analysis may be conditional on the IDA results but are pre-planned before the main data analyses are undertaken.

The recently published paper “Regression without regrets –initial data analysis is a prerequisite for multivariable regression” [3] addressed this problem in the context of cross-sectional studies that intend to use regression models with numerical, binary or count outcomes. This resulted in a checklist for an IDA plan focused on three IDA domains:missing values, univariate distributions, and the multivariable system of predictors. The resulting IDA plan balances exhaustive investigation of the dataset with utility. The approach was illustrated in a study with the aim to fit a diagnostic prediction model for the presence or absence of bacteria in the blood, in a study where demographic variables and about 50 routinely collected laboratory variables were available for about 15,000 patients. While the paper focused on a predictive research question, the recommendations can also be used for descriptive or explanatory research questions, to pre-plan data screening efforts for clinical trial data or for research aims that include a plan for using modern algorithmic approaches.

Building on this project, TG3 developed an IDA checklist for longitudinal data with an example of complex surveys [4]. The goal of this study was to extend the check list to studies where participants are measured repeatedly over time, which poses additional challenges for IDA and requires additional considerations. In the longitudinal setting, IDA explorations can quickly become overwhelming even with a small number of variables. To manage this we provide guidance explaining which aspects of the data may be explored prior to undertaking the statistical analyses that address the main research question. Different time metrics, the description of how much data was collected through the study (how many observations and at which times), missing values across time points, including drop-out, and longitudinal trends of variables were considered, resulting in two additional domains: participation profile and longitudinal trends. This study was based on collaborations with STRATOS members from the missing data (TG1) and measurement error (TG4) topic groups.

To facilitate the implementation of the checklists in other projects and encourage the use of a systematic approach and of reproducible strategies, extensive and well documented R code was provided for the examples discussed in the papers (https://stratosida.github.io/resources.html). This includes many examples of data visualization.

Both papers discussed the need to integrate the IDA plan in a statistical analysis plan to increase transparency, reproducibility and avoid ad-hoc decisions. These considerations motivated a STRATOS wide project to develop guidelines for statistical analysis plans for observational studies that includes IDA steps and prepares for documenting deviations based on IDA results. This project is registered with the EQUATOR network as a reporting guideline under development for observational studies. We would like to point out the importance of incorporating IDA discussion and tools in curricula to build researchers' capacity and empower them to employ a systematic use of IDA in their studies.

We acknowledge the contributions from the co-authors of the four papers cited: Mark Baillie, Frank E. Harrell, Frank Lawrence, Katherine J. Lee, Saskia le Cessie, Cecile Proust-Lima, Carsten O. Schmidt and Werner Vach.

REFERENCES

[1] Huebner M, le Cessie S, Schmidt CO, Vach W. on behalf of the Topic Group 'Initial data Analysis' of the STRATOS initiative. A contemporary conceptual framework for initial data analysis. Observational Studies 2018; 4: 171-192. https://doi.org/10.1353/obs.2018.0014

[2] Baillie M, le Cessie S, Schmidt CO, Lusa L, Huebner M. for the Topic Group "Initial Data Analysis" of the STRATOS Initiative. Ten simple rules for initial data analysis. PLoS Comput Biol 2022; 18(2): e1009819. https://doi.org/10.1371/journal.pcbi.1009819

[3] Heinze G, Baillie M, Lusa L, Sauerbrei W, Schmidt CO, Harrell FE, Huebner M on behalf of TG2 and TG3 of the STRATOS initiative. Regression without regrets -initial data analysis is a prerequisite for multivariable regression. BMC Med Res Methodol. 2024 Aug 8;24(1):178. https://doi.org/10.1186/s12874-024-02294-3

[4] Lusa L, Proust-Lima C, Schmidt CO, Lee KJ, le Cessie S, Baillie M, Lawrence F, Huebner M., on behalf of TG3 of the STRATOS initiative. Initial data analysis for longitudinal studies to build a solid foundation for reproducible analysis. PLoS One. 2024 May 29;19(5):e0295726. https://doi.org/10.1371/journal.pone.0295726