You are in the editable version of the document, please return to the Preprint in case you do not want to make any modifications or additions.

The good, the bad and the improved practices for statistical modelling - A crowdsourcing project

Author: Ilse Coolen

Original contributors: Jessica Schaaf, Michael Aristodemou, Jordy van Langen, Nick Judd, Liza Rozman, Nebbe Al-Moula, Ben Kretzler, Rogier Kievit, Eleni Zimianiti, Léa Michel, Sophie Hofman.

Additional contributors:

Practices in statistical modelling

Kicking off with the famous George Box words that "all models are bad, but some can be useful" (Box & Draper, 1987); but how to avoid pitfalls and ensure that a model does indeed fall within the useful category?

Statistical modelling is a powerful tool to answer a wide range of research questions and provides valuable insights in understanding complex systems or phenomena. However, in fields such as psychology, most researchers are not statisticians or do not have an expert background in statistical modelling. Whether you are stepping into the realm of modelling for the first time or have been navigating it for a while, chances are that you are self-taught out of curiosity or necessity for your research. The lack of widespread knowledge about this statistical technique results in some common pitfalls that can reduce the quality of the models developed.

Therefore, I organised a lab meeting with the Lifespan Cognitive Dynamics Lab, which has a strong focus on statistical approaches such as SEM, linear mixed modelling, mixture modelling and related approaches, with the goal to gather some bad and good practices in modelling. The aim of this post is to raise awareness and give you the handles to avoid bad practices in statistical modelling and turn them into good ones.

Though this list will not be exhaustive, I hope it will aid fellow researchers in navigating the complexities of statistical modelling with greater confidence by becoming aware of potential pitfalls.

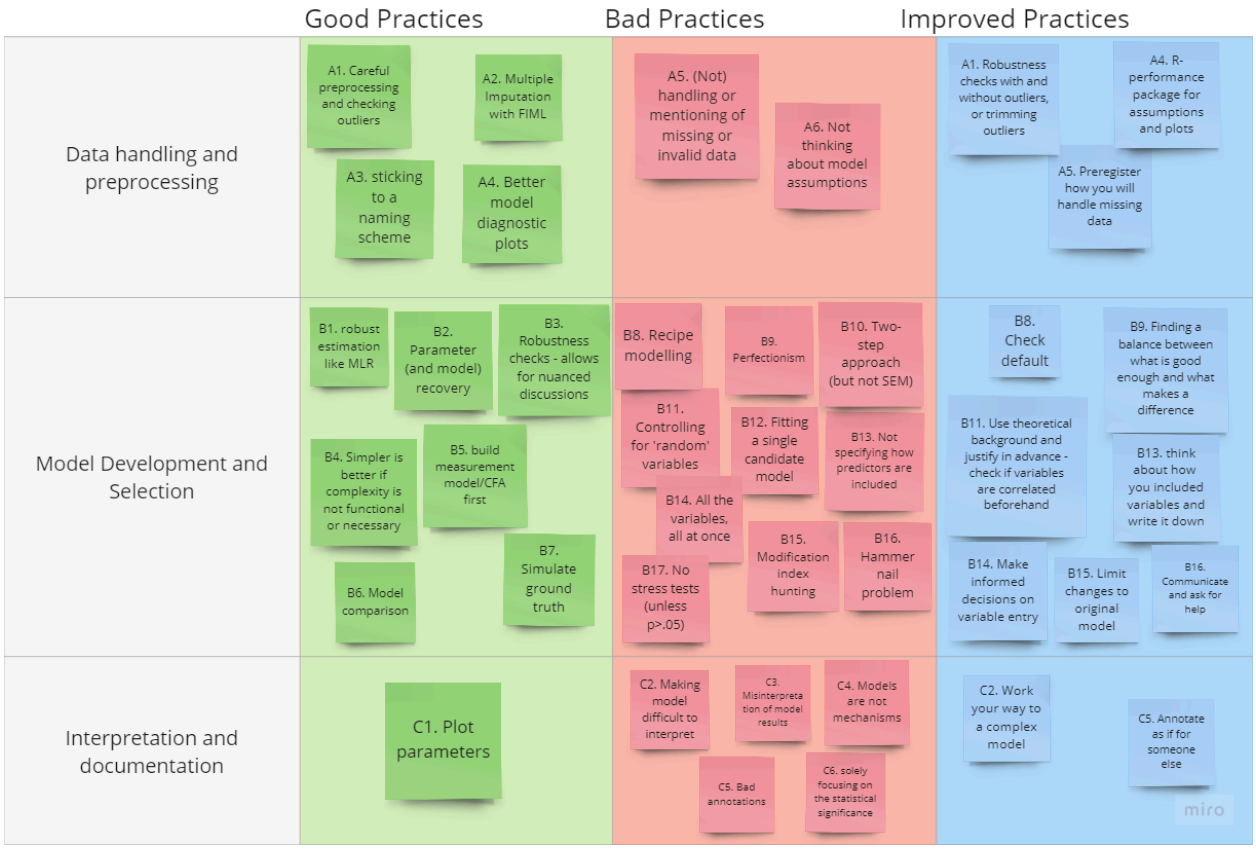
Insights from our lab meeting

I asked each lab member to prepare at least one bad and one good practice that they had previously encountered or experienced in modelling, which we then categorised in one of 3 categories during the meeting: (A) data handling and pre-processing, (B) model development and selections, and (C) documentation and interpretation. The categories weren't

communicated beforehand to allow out-of-the-box thinking. Each practice was discussed and nuances were added where needed. For each bad practice we tried to come up with a way around it or turn them into a good practice. Figure 1 shows the final board with all the originally suggested practices that were discussed during the lab meeting for an overview. Each of the practices will be explained in more details.

Figure 1.

Overview of practices discussed within the LCD lab meeting



The good, the bad and the improved

A. Data handling and pre-processing

Pre-processing of your data is the first crucial step in any modelling project. Before diving into building a model, it's important to thoroughly understand the data. This initial phase shapes the quality and reliability of the models that follow.

Good practice A1 – Careful pre-processing and checking for outliers

It is important to assure that your data is free from errors and inconsistencies prior to starting data analyses. One way of assuring this is to identify outliers as they can have a disproportionate impact on model performance. Although not all lab members handle

outliers the same way it is crucial to understand the nature of the outliers and judge whether they align with a realistic response.

Improved Practice A1 – Robustness checks without outliers and trimming

We opted to nuance Good practice A1 to discuss some good practices in how to handle outliers. Deleting outliers is not always recommended as this results in missing data not missing at random. Additionally, in research in children, variability can be high and techniques to identify outliers might be too conservative. Therefore, many researchers in developmental psychology often identify outliers on a trial level, but decide to keep all data on subject level that can be judged as a realistic performance. One good practice on decided whether to keep outliers or not, is by checking for robustness of your findings with and without outlying values. This allows to test the effect of outliers on your findings. Another way to assure a maximum of data retention and taking into account that performance is high for some participants is often referred to as "winsorising" or "trimming." This technique involves bringing extreme values 'to the fence' (e.g., at a certain percentile) to mitigate their influence on the analysis while retaining the data points.

Good practice A2 – FIML for missing data

Chances are that you will have some missing data and dealing with these missing data is a large part of modelling. While you can have a very large list of good and bad practices on handling missing data alone, we agreed that using Full Information Maximum Likelihood (FIML) estimations is a good options to deal with data missing at random in CFA and SEM models, retaining a maximum of the sample size without resulting in a large bias of the model fit (see for example Köse, 2014).

Good practice A3 - Sticking to a naming scheme

A good habit to already implement in pre-processing is sticking to understandable and easy variable names. This will make it easier to develop and interpret your model. It also enhances the transparency and reproducibility of your research findings. Some recommendations for naming schemes are to pick names that reflect their meaning and remaining consistent in naming and formatting between variables. A codebook to accompany your data is never a bad idea.

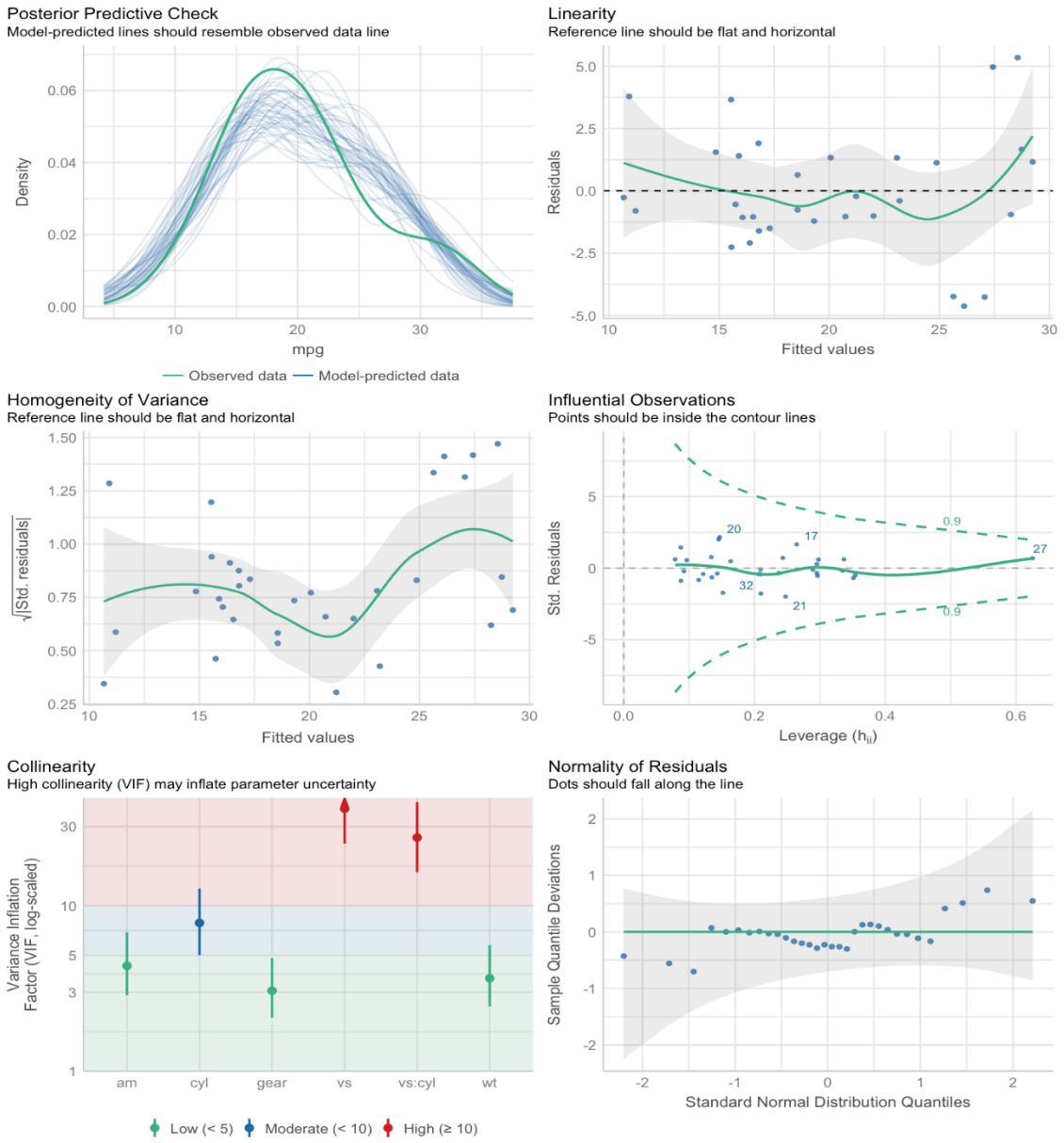
Good practice A4 - Better model diagnostic plots

Plotting your data as part of the pro-processing data is a great way to understand your data. This allows for the identification and understanding of assumption violations, outliers, missing data in a visual way. Visualising your data contributes to understanding violations and exploring certain relations between variables, which are crucial to understand in order to develop your model. The 'performance' package on R (Lüdecke et

al., 2021) was demonstrated and recommended for diagnostic plots on your data, see Figure 2.

Figure 2.

Diagnostic plots with the R package 'Performance' by Lüdecke et al., 2021.



Bad practice A5 - (Not) handling or mentioning of missing or invalid data

The worst way to handle missing data is probably to not handle them. Although this is probably not a common bad practice, the practice of not mentioning what has been

done to handle data in manuscript is regrettably not so uncommon. The way one handles missing data should be well-thought through and justified. While there is not one solution for all with missing data, there are recommendations available depending on the amount and type of missing data in the literature and in the good practice A2 mentioned above. Regardless of what massing data technique you opt for, this should be clearly reported as well as the amount of missing data.

Improved practice A5 – Preregistrating missing handling technique

To avoid the pitfall (bad practice 1a) of not handling missing data and/or not reporting them, researchers could think about the techniques they will use in a preregistration. This does not mean that one cannot change the way they handle the missing data in the case where another technique might seem more appropriate upon seeing the data, but this will allow them to actively think about it and justify the method or change in method. Authors can refer back to the preregistration and be reminded to ensure that missing data techniques are reported in manuscripts.

Bad practice A6 - Not thinking about model assumptions

Neglecting to check for violations of key assumptions like normality, independence, linearity, or homoscedasticity may lead to a lack of understanding of the data and applying (default) settings without critically evaluating their appropriateness for the data. By disregarding model assumptions, researchers risk producing biased estimates, misleading interpretations, and unreliable predictions.

I refer back to good practice A4 - Better model diagnostic plots, as a means to avoid this pitfall and check for model assumptions in a visual way.

B. Model development and selection

From the selection of appropriate modelling techniques to the fine-tuning of model parameters, every decision made in this phase has a profound impact on the accuracy, interpretability, and generalisability of the resulting models. Different approaches and methods need to be carefully considered in developing and selecting a model to identify the final (set of) model(s) that best capture the underlying patterns in the data while avoiding common pitfalls.

Good practice B1 - Robust estimation like MLR

Although we would not recommend always using MLR as an estimator without thinking about what would work best for your data and variables, it is likely that your data will not follow a multivariate normality or have other assumption violations, making a stringent estimator like MLR a better solution. Before deciding to use the MLR estimator, remember that the estimator should fit the variables used (e.g., categorical or continuous).

Good practice B2 - Parameter (and model) recovery

Parameter (and model) recovery involves checking if a statistical model accurately estimates the true parameters and replicates the underlying structure of the data it's meant to represent. This practice includes testing the model's performance by creating simulated data with known parameters and seeing how well the model can identify these parameters. It ensures that the model reliably captures relations between variables, making it trustworthy for real-world use.

Good practice B3 - Robustness checks

We have already talked about robustness checks with and without outliers, but beyond this, we believe this is a good practice in many cases. If you are unsure whether the results will hold up if you add another parameter or condition, robustness checks provide a means to assess the validity of the model findings. This could also hold for robustness checks with the same model but different datasets. If the results are robust, this allows for generalisation of your findings. If they do not, this allows for a nuanced discussion and understanding of the results.

Good practice B4 - Simpler is better if complexity is not functional or necessary

A simple model is often easier to interpret than a complex one, making it easier to communicate your findings. The simplest model also relates to the principle of parsimony, which advocates for using the fewest assumptions or entities necessary to explain a phenomenon. By favouring simplicity without sacrificing explanatory power, researchers can adhere to the principle of parsimony and avoid unnecessary complexity. Adding meaningless variables to a model makes it more likely to overfit your data and capture noise rather than the underlying patterns you want to test. This can make it appear as though the model fits the data better, but these added variables might not be meaningful. Note that this is the case when you are assessing underlying patterns in your data and not noise. Purposely capturing noise or other fluctuations in your model might require a more complex model (e.g., DSEM). Indeed, there is a risk when always sticking to the simplest model as you may overlook important patterns in your data when not adding parameters, variables or conditions that are functional or necessary in your model (e.g., time-series data). In doubt, robustness checks can be advised.

Good practice B5 - Building your measurement model/CFA first in latent SEM models

This good practice involves initially constructing the measurement model or Confirmatory Factor Analysis (CFA) before proceeding with the structural relations in the model. This practice allows researchers to establish the validity and reliability of the latent variables measured before adding extra parameters or conditions. This will add to the understanding of the basic measurement model.

Good practice B6 - Model comparison

Models should have a strong theoretical background to accurately test a specific hypothesis. However, developing one model under the assumption that this is the true model could result in suboptimal models potentially leading to inaccurate representations of the data. Comparing models allows us to select the most suitable model that accurately describes the data and to test multiple hypotheses. It helps in assessing model complexity, avoiding overfitting, and identifying potential model misspecifications. By comparing alternative models, researchers can make informed decisions about which model best fits the data, leading to more robust and reliable findings. Note that comparisons between models should be between nested models, so a different version of the same model (e.g., Y ~ A + B + C vs. Y ~ A + B), not different models (e.g., Y ~ A + B + C vs. Y ~ A + B + D).

Good practice B7 - Simulate ground truth

Simulating ground truth refers to the act of generating simulated data based on the relations you expect to find. This practice allows you to validate a model on data for which you know the true relations between variables before applying the model to your collected data. This helps assess whether the model is able to accurately capture the underlying patterns that you are looking for.

Bad practice B8 – Recipe modelling

Recipe modelling refers to a bad practice where researchers apply a predefined set of modelling techniques or procedures without considering the specific characteristics of their data or the underlying assumptions of the models. This approach treats statistical modelling as a one-size-fits-all recipe, rather than a tailored and thoughtful process that takes into account the unique aspects of the research question, data, and context. Common 'recipes' are sticking to the default or copy-pasting.

Sticking to the default settings as a bad practice needs a little more nuance as default settings are often default for a good reason. Most of the existing R packages for modelling such as *sem* (Fox, 2006) and *Lavaan* (Rosseel, 2012) are easy to use and have a whole variety of default settings that can be tailored to specific research questions or data characteristics. These default settings are most likely to be the correct option for your data, hence why these are set as default, however, it is a bad practice to always stick to the same default without asking questions and verifying whether these are suitable for your model.

Similarly, copy-pasting procedures from previous studies or published papers without adapting them to the specific context of the current study can lead to the application of methods that may not be well-suited to the data or research question, resulting in biased or misleading results.

Improved practice B8 – Verify default settings

Verifying the default settings or the settings of a previously used model, exploring other options and what would be most suitable to fit your new data and research questions is advisable. Some of the common default settings that are worth verifying are the estimator, constraints, scaling, missing data handling, fit indices and of course the model itself.

Bad practice B9 – Perfectionism

While it is good to want to thrive to the perfect model, it should not stand in the way of progress. Moreover, perfectionism may lead to a relentless pursuit of an overly complex model that fits the data extremely well but fails to generalise to other data. Another pitfall is that perfectionism may eventually lead to a model that has gone through excessive changes and additions so that the findings go in the direction of your hypothesis, while all the previous models did not.

Improved practice B9- Finding a balance

Finding a balance between what model is good enough and what makes a difference is key. When you find that your model findings are stable after robustness checks, for example on other data, you can be more confident that your model does not need further improvement. When a change to the model does not make theoretical sense or does not add meaning to the model, it probably does not belong in your model.

Bad practice B10 - Two-step approach

This was a bit of a controversial bad practice, as we previously recommended a two-step approach in good practice B5 by assessing the measurement model in SEM in a first step and only adding structural relations in a second step. However, it is still worth noting the dangers of a two-step approach. This two-step approach refers to the splitting of the analyses in two steps rather than adding the second step to the first as was recommended in B5. One could for example extract estimates from a measurement model and use these estimates in a second step to uncover relations between them without including the original measurement model in the final model. Changes that are made in a certain step can affect the outcome of previous steps. Instead we recommend a two-step approach where the second model includes the original model, allowing us to verify whether the measurement model changes with the additional parameters.

Bad practice B11 - Controlling for 'random' variables

Adding variables without a theoretical or empirical basis can lead to biased estimates and reduced validity of the model. With this practice you may obscure real relations and effects in your model leading to incorrect conclusions. A recent example that has received more attention is adding the control variable of age in research exploring development. When the aim of a research question is to understand development over time, we would essentially remove the part of development that is related to age. This is

often forgotten during interpretations, as development without age does not make much sense and is this difficult to comprehend or interpret.

Improved practice B11 - Theoretical background, justifying in advance, and verify correlations.

To avoid the pitfall of adding 'random' variables, researchers need to carefully think of the theoretical background and justifications of all variables before starting to build the model. This could be done in a preregistration phase as this provided time to sit down and think about each variable. Before building the model, one should also explore correlations between the variables. Each added variable should relate to the outcome measure in the model. If there is no correlation, the variable is likely not a good addition to the model as this could result in finding a relation in the model that is due to chance.

Bad practice B12 - Fitting a single candidate model

When you fit a single candidate model that provides adequate fit indices, you might miss models that are better suited. This could lead to confirmation bias and premature conclusions about a model, while there might be a better-fitted one.

This bad practice can be contrasted with good practice B6 – model comparison

Bad practice B13 - Not specifying how predictors are included

Failing to clearly specify the treatment of variables in statistical modelling can lead to ambiguity, misinterpretation, and reproducibility issues. For example, the way categorical variables are coded and specified in the model (e.g., as factor or as numerical) and continuous variables are transformed (e.g., centred) need to be thought through and reported.

Improved practice B13- Think about how to included variables

Thinking about how to include variables in the model before you start building the model (e.g., preregistration) can be helpful. When you decide on a specific way, make sure that this is reported somewhere (e.g., annotation in analyses script, preregistration, methods) so that it is easy to track down again.

Bad practice B14 - All the variables, all at once

When you have multiple variables to build your model, entering them all at once can lead to overfitting, unnecessary complexity and a loss of information. When you overfit your model by including all variables simultaneously, you are more likely to capture noise and less likely to end up with a model that is generalisable to other data. You are also adding complexity to the model that might not be necessary and is likely to complicate interpretations of the findings. Indeed, by entering all variables at once, you might miss important information about certain interactions between variables, making it difficult to fully understand your data.

Improved practice B14 - Make informed decisions

Link back to the literature and your research question to make informed decisions on which variables to enter when to avoid entering them blindly all together. Think about expected relations and interactions and correlations between your variables. This way you can assess whether all variables are necessary to be in your final model and allows you to understand the relations between all variables better.

Bad practice B15 - Modification index hunting

While modification indexes can provide valuable insight to suggest areas of improvement for a model and help to understand bad fit, excessively looking for modification indexes to improve your model might lead to index hunting until the model fits. You may end up with a model that is far from the hypothesised theory-based model, but might lead you to draw the conclusion that the hypothesised model fits the data well.

Improved practice B15 – Limit the changes to your original model

Every change to the original model should be reported, including the changes tested but not retained for the final model. Every change should also be meaningful and in line with the theoretical model. Another option is to report the difference between the original theoretical model and the final model including their fit indices.

Bad practice B16 – The hammer-nail problem

Related to the bad practice B1 – recipe modelling, is the hammer nail problem. If you only have a hammer, you will treat every problem as a nail. In modelling, this translates to having experience with one specific statistical modelling technique or script and treating every research question with this skill/script. Rather than letting the research question lead to the statistical analyses necessary, you transform the research question until it fits or you fail to accurately address your research question.

Improved practice 16 – Communicate and ask questions

Make sure that your research question is primary and that the analysis accurately fits this question. Collaborations or communicating your research plan with colleagues are a great way to learn new skills and to make sure that your analysis fits your question. Preregistered reports also provide the option to receive feedback from fellow researchers on the fit between the analytical technique and the question.

Bad practice B17 - No stress tests (unless p>.05)

This practice refers to the testing, or rather not testing how a model might change under different conditions, such as done in robustness checks described previously. We have already established that robustness checks are good practices, but not having them when your hypothesised and original model fits is also not recommended. In doing this you ignore the uncertainty of models and it could give you a false sense of security about your original model. We recommend the good practice B3 – robustness checks and good-practice B6 – model comparison to avoid this pitfall.

C. *Interpretation and documentation*

How we interpret and document our models is crucial. Every step impacts how reliable, understandable, and reproducible a model is. It is essential to ensure a thorough understanding of the data's patterns, while also documenting our process. Good documentation and annotations boosts reproducibility, allowing others to check our work and build upon it.

Good practice C1 – Plot parameters

Plotting your model parameters provides a visual way of interpreting your results. These plots can be simplified to make interpretations easier as they can focus on the research questions and specific patterns of interest.

Bad practice C2 - Making models difficult to interpret

Complex models including multiple variables and parameters can be difficult to understand. The purpose of a model is not to be complex or include as many parameters at the same time as are possible, but it is to provide a framework to understand relations and patterns in your data that could generalise to other similar data. These relations and patterns should remain interpretable.

Improved practice C2 - Work your way to a complex model

Sometimes a model requires complexity if this fits with the data and the research question. This might make it hard to understand the underlying patterns found. By gradually working your way to a more complex model, each step of the way will provide you with necessary information to understand specific relations. These steps can be reported to help readers gain the same understanding that you have reached.

Bad practice C3- Misinterpretation of Model Results

There are many ways of misinterpreting model results, such as assuming causation from correlation, cherry-picking (emphasising only those results that support your hypothesis), generalising findings for a non-representative sample, not accounting for assumption violations, etc. Improved practice B16 – communicating and asking questions provides a way to avoid these kinds of pitfalls. Communicate about your results and your interpretation to verify whether these are in line with the interpretations by colleagues.

Bad practice C4 - Models are not mechanisms

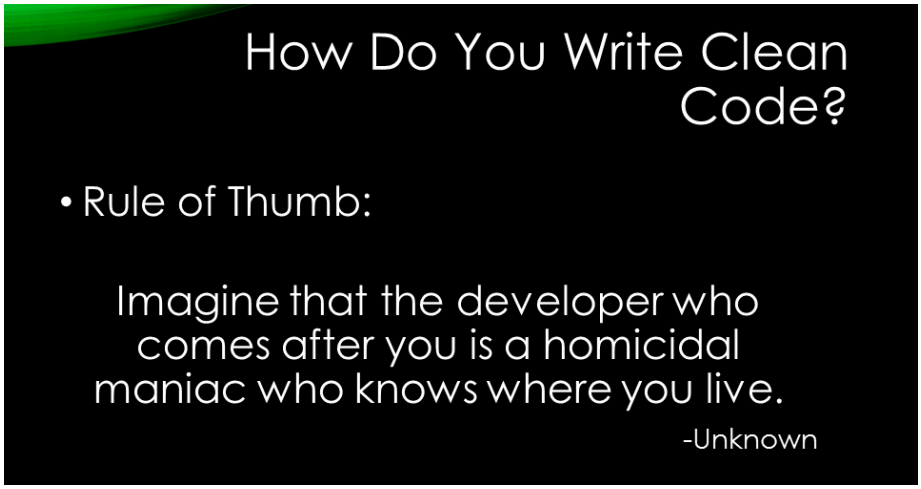
Statistical models describe and predict underlying relations and patterns, not direct causal mechanisms, as they do not explain the how and why. A model can help to suggest potential mechanisms or give you insight in underlying relations of potential mechanistic processes, but most models should not be interpreted as mechanisms.

Bad practice C5 - Bad annotations

The assumption that you will remember what you have done in a previous analytical model and why you have done this, is probably wrong. Pre-processing your data, understanding your data and building your models takes time. It would be a pity to have to take the same time just to remember what you have done and why. Additionally, understanding somebody else's script is even more difficult, limiting the reproducibility of your analyses.

Improved practice C5 – Annotate your script as if for someone else

Clear annotations provide a means for you and other researchers to understand what was done and for what reasons. In doubt, write as many comments as you think might be necessary for someone else to understand each step and replicate your script.



Bad practice C6 - solely focusing on the statistical significance

The statistical significance of a model can be important in certain contexts, but it is not the sole determinant of the model's utility or validity. We previously already talked about the pitfall of no stress testing when a model is significant, but immediately discarding a model based on a non-significant result without looking at other fit indices and why the model was not significant, will not provide you with a clear understanding of the data.

Conclusion

Three good practices seemed to return on frequent occasions in our discussion: (1) visualise data, (2) perform robustness checks, and (3) only add meaningful complexity to a model.

(1) Visualising your data can help you understand assumption violations at the pre-processing stage and simplify interpretations of your model.
(2) Robustness checks whether it be on other datasets, with or without outliers, different parameters or when comparing different nested models, are a great way to provide

more strength to your model or nuance the findings and enhance understanding of the data. If in doubt, trust in robust!

(3) Only add variables, interactions and parameters when they have theoretical value to answer your research question.

Do you have additional good or bad practices to contribute or believe that certain aspects should be nuanced, please don't hesitate to reach out to coolenilse@gmail.com or directly add your contributions to the "Edit_Good and bad modelling practices" and together we can continue to refine modelling practices! For more information with regards to adding your contribution, see "Contribution_guidelines"

References

Box, G. E., & Draper, N. R. (1987). *Empirical model-building and response surfaces*. John Wiley & Sons.

Fox (2006), "Structural Equation Modeling With the sem Package in R." *Structural Equation Modeling,* 13:465-486

Köse, A. (2014). The effect of misssing data handling methods on goodness of fit indices in confirmatory factor analysis. *Educational Research and Reviews, 9*(8), 208.

Lüdecke et al., (2021). performance: An R Package for Assessment, Comparison and Testing of Statistical Models. *Journal of Open Source Software, 6*(60), 3139. https://doi.org/10.21105/joss.03139

Rosseel (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2), 1-36. URL http://www.jstatsoft.org/v48/i02/