Research Seminar Data Science Task 3 — Workshop "An introduction to the Multiple Imputation DOCtor R package"

Daan Brugmans (S1080742)

For Task 3 of the Research Seminar Data Science, I attended a workshop hosted by prof. Kate Tilling and dr. Ellie Curnow from the University of Bristol. Since this was a workshop (and not a seminar), my experience may deviate from what is usual for this task.

I. Summary

The workshop was about performing Multiple Imputation (MI) responsibly using the *Multiple Imputation DOCtor (midoc)*, an R package under development by the hosts of the workshop. In practice, most datasets any data scientist will work with will have some data missing, which must be accounted for. An increasingly popular method with which to account for missing data is MI, which is generally more robust and statistically sound than other, simpler imputation methods. As MI has seen more use, it has also become more accessible, with libraries such as *mice* providing the option to apply MI on a dataset in just a single function call. However, as MI has become easier to apply, it has also become easier to apply it erroneously. For example, it is fairly common practice to supply an MI model with as much auxiliary variables as possible, which it can use to impute data to the existing distribution, without considering which variables should and should not be passed onto the MI model. This is because the inclusion of some auxiliary variables into the MI model may cause bias in the imputed data, skewing the final distribution away from the original distribution.

To guide data scientists through the process of choosing an imputation method that is considered valid to the data, and to mitigate new instances of improper imputation, prof. Tilling and dr. Curnow are developing the R package *midoc*. midoc offers two main tools to the data scientist's toolkit:

1) midoc implements varying functions that help a developer perform imputation that is valid to the data, taking causality amongst variables into account. For example, midoc can transparently show in what patterns data are missing amongst variables, and, given a directed acyclic graph (DAG) of the dataset, advice the developer on which imputation methods are suitable for the data at hand, which ones are likely to introduce bias, and show the developer the evidence it has gathered in support of its advice.
2) midoc includes an interactive webpage that will guide the user through all the steps necessary in order to determine the right imputation method and settings in a low-code environment. Written and presented in natural language, the user learns how and why certain steps must be taken during this process, while performing them themselves.

II. Experience about Asking Questions

The hosts were very easy to approach for questions. During the hands-on parts of the workshop, prof. Tilling and dr. Curnow would approach attendees themselves, providing answers to questions and help with the instructions from midoc. I found the webpage guiding users through the process of determining a valid imputation method easy to follow and transparent. Because of this, I found myself not needing to raise a question. Those who did raise questions received answers that were clear and well grounded.

III. Reflection

I enjoyed attending the workshop and feel that I have learned a valuable new insight into a topic I was already somewhat familiar with. In past projects, I have also used Multiple Imputation without considering that the inclusion of certain variables into the imputation model would lead to bias in my data. Having attended this workshop, I have now become aware that such careless use of MI can have negative effects on the data I work with.

I also appreciate that this workshop builds on top of the knowledge that I had learned during the Bayesian Networks and Causal Inference course. I think this was a fun and interesting way of applying DAGs and causal inference on a real-world problem.

In the future, I would like to try to apply midoc to a real-world dataset. For the workshop, I only used midoc on a toy dataset provided by the hosts. I would like to more thoroughly test midoc's capabilities on a dataset with more variables. For example, I am considering to apply midoc on a project I will be partaking in for the Machine Learning in Practice course, should missing data be a relevant topic for that project.