Publication-Ready Data Analyis with R

E. F. Haghish

February 28, 2021

In this document I will demonstrate how to create and organize a computational project and how to produce a dynamic document as you are developing the data analysis.

Producing such a document within R might appear to be tedious at the first glance. But it is not, if you you make yourself familiar with the Rmarkdown workflow. On the bright side, a data analysis developed in these manners is transparent and easy to comprehend and inspect. The project is also easy to share, improving the practice of collaboration on data analysis.

The data analysis is organized in several Rmd files. The **MAIN.Rmd** file execute them all within a single document. For the following example, I use a data set from the lavaan package: <https://lavaan.ugent.be/tutorial/cfa.html>.

* The analysis is carried out using R version 3.5.1 on Mac OSX version 10.14.6
* The following R packages are also required:
  + pander v. 0.6.3
  + lavaan v. 0.6.7.
  + psych v. 2.0.12
  + semPlot v. 1.1.2

# Results

## Preparation

Dynamic text is text that includes scalars obtained from R. Such a text paragraph obtains and interprets values from R on the fly. Here is an example:

301 students from two school participated in the study, of which, 1 subject (female, age 13) was excluded from the study because of her unknown grade. Both schools roughly contributed the same number of seven- and eight-grader subjects as you in the table below.

*Table 1*: *Example of a* ***dynamic table***

|  |  |  |
| --- | --- | --- |
|  | seven-grade | eight-grade |
| **Grant-White** | 79 | 65 |
| **Pasteur** | 78 | 78 |

The table above is not necessary, but shows how easy it is to construct a table in R. The actual work is done by the **pander** package. I will construct a couple of more tables in the following sections

## Descriptive analysis

Whether you should write the entire document here and produce a publication-ready Microsoft Word or LaTeX document is up to you. The point is documenting what is carried out on the data and making all decisions trackable. Note that **interpretation** of statistical analysis includes a number of decisions that a researcher makes. This is why combining interpretations with statistical code provide a clearer picture of what is going on in the data analysis.

## Analysis

There is no need for the entire analysis to be written in Rmd files. You can write R script files and source them within the dynamic document. Below, I source a function that I have written to extract the values of a CFA output from the lavaan package. By doing so, you can keep your dynamic documents clean and down to point.

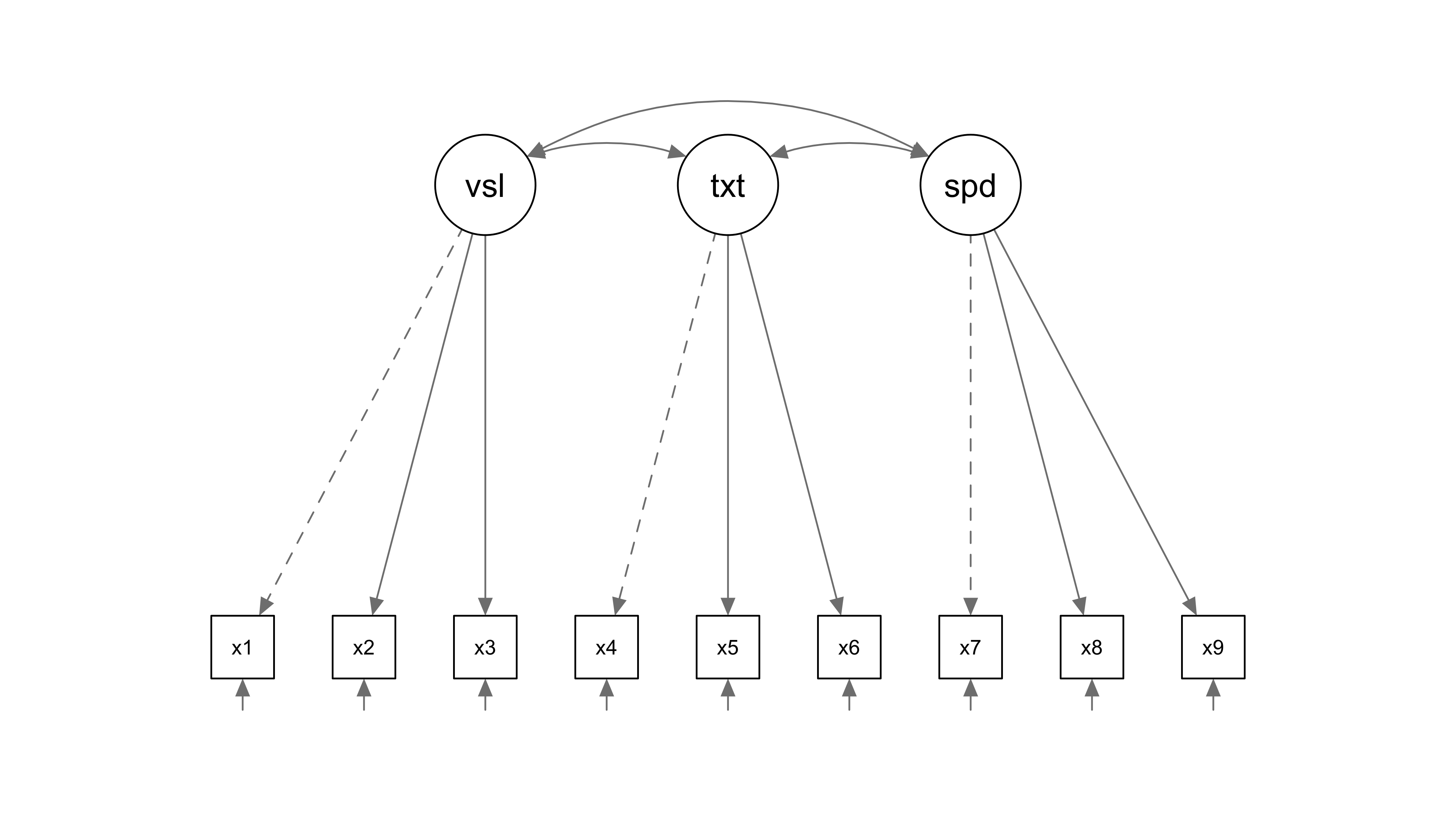
source("../FUNCTIONS/getLavaanCFA.R")

The R code above is printed in the document, but as we are trying to keep the MS Word document clean, we can exclude the R code and output from the document using variety of options:

|  |  |
| --- | --- |
| **Option** | **Description** |
| echo=FALSE | exclude the R code but keep its output |
| results=hide | keep R code but hide the results |
| include=FALSE | hide both R code and outputs |

### CFA analysis

Including a figure in the document is automatic. Although, you can also export it as a PNG file and import it in the document. At top of the **MAIN** file, I have increased the DPI resolution of the figures generated by R to make them ready for print. Here is an example of a path diagram for the CFA model.



More importantly, once we run the factor analysis, we can extract the values and use them to construct a dynamic table or point them out in the dynamic text. Below, I extract the values and generate a dynamic table with them. Such a table is usually useful when comparing multiple models together:

*Table X*: Estimates of the CFA model

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| (df) | RMSEA | RMSEA 95% CI | CFI | TLI | SRMR | P-value |
| 85.94 (24) | 0.093 | 0.072 - 0.114 | 0.930 | 0.895 | 0.066 | 0.000 |

Alternatively, you can use the extracted values in a dynamic text and refer to the one by one. For example, the RMSEA value was estimated to be 0.093… Note that I also included a mathematical notation in the table for .

### Last words

Preparing this project took only a few hours of work. I didn’t do much of an analysis here, but the template is ready to be used for a real project. Therefore, practicing data analysis in this manner is neither unrealistic nor time-consuming. It’s just a matter of habit.

Once you practice your data analysis with such a discipline, it becomes easy to follow and review your logic for planning and conducting the data analysis. And that is a step forward to reproducible research. As a final word, I want to point out that the majority of the R code and its output was excluded in the MS Word. This is also what happens in the actual scientific papers, i.e. only the absolute-necessary outputs are included in the manuscript. Therefore, a dynamic document like this indeed profides the full picture of the process of data analysis.