Journal of Data Science xx (xx), 1–5 DOI: xx.xxxx/xxxx_xxxx
July 2023

My wonderful paper

Abstract

This is the abstract

1 Introduction

(background)

In this paper, a concept called analysis plan is proposed to describe the logical structure of a data analysis. An analysis plan is a set of analysis steps plus their expected outcomes. It is a formal representation of the analysis process and can be used to guide the analysis process, to communicate and compare the analysis process to others, and to evaluate the analysis process. The concept of analysis plan is illustrated with examples. The implications of the concept for data analysis practice is discussed.

The analysis plan described in this paper should be differentiated from the pre-specifies analysis plan document often used in biostatistics to specifies the hypothesis, data collection mechanism, statistical procedures etc of randomized experiments.

The rest of the paper is organized as follows: Section 2 describes the concept of analysis plan in detail. Section 3 provides examples of analysis plan [more details]. (need another section here or before examples?) Section 4 concludes the paper.

2 Analysis plan

describe/ define what analysis plan is

analysis plan and outcome plan

analysis plan as unit tests to divide the "result universe", which allows us to answer questions like:

- how would the results change if the value of a unit test change
- whether our outcome expectation aligns with the plan expectation, meaning whether "it is possible for the combination of plan expectations to produce the outcome expectation"

3 Examples

Three examples are presented to illustrate how the concept of analysis plan can be applied to data analysis. [toy example]. Section 3.2 illustrates how constructing the result universe in a linear regression model of PM10 on mortality can help understand the impact of sample size, model specification, and variable correlation structure on data analysis. [example three]

3.1 A toy example

3.2 Linear regression

Consider a linear regression model to study the effect of PM10 on mortality (provide context of using PM10 to study mortality). Analysts may expect a significant (p-value ≤ 0.05) PM10 coefficient in the linear model from the literature. This is the *outcome expectation*. There are multiple factors that can affect the outcome expectation of linear regression, which here is called *plan expectation*, for example, 1) sample size, 2) model specification, and 3) correlation structure between variables. Adequate sample size is required to achieve the desired power to detect the significance of PM10 on mortality. Temperature is often an important confounder to consider in such study (add reference). From some domain knowledge, an analyst may expect that the significance of PM10 coefficient can be attained by adding temperature to the model. Analysts may also expect certain correlation structure between PM10, temperature, and mortality, and the distribution of each variable.

To build the result universe, datasets can be simulated to either meet and fail these plan expectations, allowing the analysts to observe the significance of PM10 coefficient. Here, sample sizes of 50, 100, 500, and 1000 are considered. Two model specifications are included: 1) linear model with PM10 as the only covariate (mortality $\sim$ PM10), 2) linear model with PM10 and temperature as covariates (mortality $\sim$ PM10 + temp). A grid-based approach is used to simulate correlation structure. Reasonable ranges of correlation between the three variables are $\text{cor}(\text{mortality}, \text{PM10}) \in [-0.01, 0]$, $\text{cor}(\text{mortality}, \text{temperature}) \in [-0.6, -0.2]$, and $\text{cor}(\text{PM10}, \text{temperature}) \in [0.2, 0.6]$.

Figure 1 shows that result universe of the linear regression model and how a change of decision in one of the plan expectations above affect the outcome expectation. Panel a) is colored by the outcome expectation – whether a significant p-value is found in the PM10 coefficient. Panel b) shows the effect of adding temperature to the model and the results show that the significance of PM10 coefficient can be achieved by adding temperature to the model for a sample size of 500. Panel c) shows that increasing sample size from 50 to 100 enhances the significance of p-value for PM10 and the significance remains with further increases in sample size.

A result universe constructed in this example can be presented to analysts to answer the what-if questions raised in the data analysis. What if the sample size is increased? What if temperature is added to the model? What would the results expect to be changed when the correlation structure is different? For analysts, expectations can be used as unit tests to divide the result universe, which allows them to understand the impact of each factor on the outcome expectation.

The result universe also provides a holistic view of how the results obtained by the analysts are situated in all possible results. This can be seen as a direct towards trustworthy data analysis, where the audience of the analysis to exercise their own cognitive model [@grolemund_cognitive_2014] to evaluate the results reported.

Source: Article Notebook

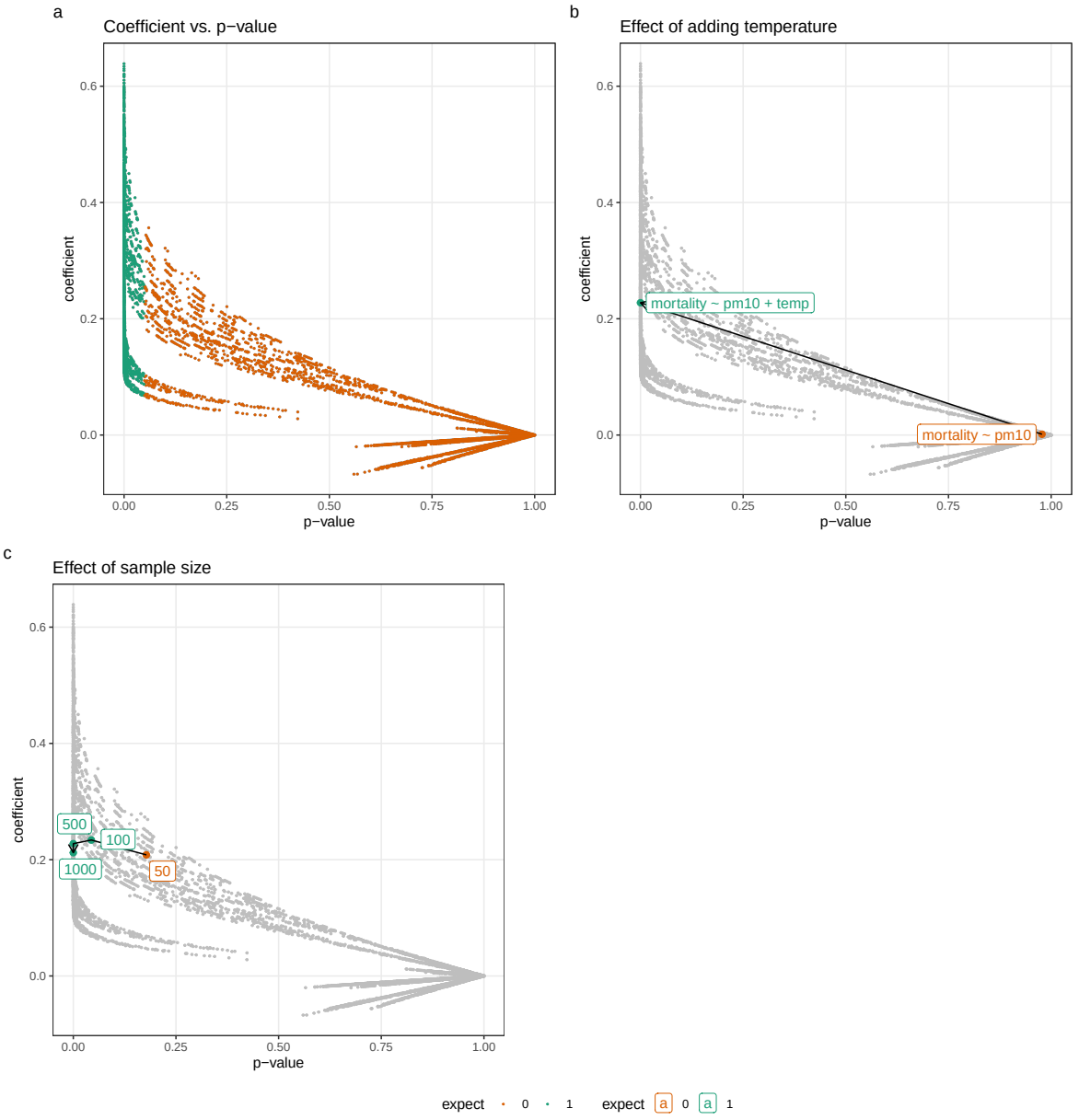

Figure 1: The result universe of linear regression model to study the effect of PM10 on mortality: a) colored by whether the p-value of PM10is significant (less than 0.05), b) the effect of adding temperature to the model for a sample size of 500, c) the effect of increasing sample size for a fixed correlation structure.

Source: Article Notebook

4 Conclusion

5 References