Journal of Data Science xx (xx), 1–8 DOI: xx.xxxx/xxxx_xxxx
July 2023

My wonderful paper

Abstract

This is the abstract

1 Introduction

[background - In a data analysis workflow]

For expert analysts, the expectation of the analysis process is often implicit. They may have a mental model of the analysis process and the expected results. However, for non-expert analysts, the expectation of the analysis process may not be clear. They may not know what to expect from the analysis process, or what to look for when an analysis process fails the expectation. An analysis plan can help to make the expectation of the analysis process explicit. It can help to guide the analysis process, to communicate and compare the analysis process to others, and to evaluate the analysis process.

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, expectation (outcome and plan)
  - expectations can be used to formulate unit tests, they can be used to guide to diagnose in the analysis process: data issue, unreasonable expectation – expectation is "wrong"
  - some expectations can be about data: e.g. check temporal gaps in time series data. Some others are statistical checks: sample size, outliers, model specification, or checks based on domain knowledge: include temperature in the mortality ~ PM10 model.

Q: Whether we should formulate these concept with math notation?? A: only if it helps

An analysis plan is a set of analysis steps plus their expected outcomes. An analysis plan consists of two parts: outcome expectation and plan expectation. Outcome expectation is the expected result of the analysis, for example, a significant p-value or a coefficient estimate within a certain range. Plan expectation is the expected results from the intermediate steps of the analysis, for example, the expected distribution of the data, or the expected correlation structure between variables.

While there is typically one major result expectation in an analysis, multiple plan expectations can be formulated as unit tests to guide the analysis process. These plan expectations serve as road marks to test, at various stages of the analysis, whether the results are within expectation. When the outcome is out of expectation, these plan expectations can serve as unit tests to diagnose the analysis process.

- why it is useful to be explicit about expectation?

- with simulation, the unit tests can be used to construct the "result universe"
- With formulated expectation, one can generate multiple version of the dataset for teaching statistics

When the expectation is formulated, we can also construct values fulfill or not fulfill the expectation. With simulation, this allows to construct the set of all possible results that can be obtained from the analysis process, which we called the result universe. With such a result universe, we can see how the results obtained by the analysts are situated in all possible results. When accompanied by the result universe, it is clearer how different factors in the analysis affects the main 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 and Wickham, 2014) to evaluate the results reported.

2.1 A toy example

Let's think about a 5-day step count. You make a resolution to walk on average 5000 steps a day (your expectation) and using an app to record your step count. After 5 days, the app tells you've walked on average 8000 steps.

It is easy to come up with reasons why an 8000 average step is resulted based on common sense:

1. you may run a 10k on day 1, resulting a high step count on the day (outlier on the right).
2. you left your phone at home on day 3, resulting a zero or minimal step count on the day (outlier on the left).
3. you may realise the step count may increase since you were in a hiking trip in the last five days (average shift).

Based on these reasons, you may devise a set of unit tests to check the step count data, i.e. check the maximum and minimum step count, check the difference between each day.

- If the daily count looks like c(4000, 5000, 5500, 5500, 20000), the maximum check will flag the data for investigate the maximum. The difference between days test will also flag the data
- If the daily count looks like c(20000, 20000, 20000, 20000, 20000), the maximum check will flag the data for investigate the maximum.

Some part of the space is impossible: c(0, 4000, 5000, 5500, 5500) is flagged by the minimal tests but won't cause an average of 8000 average step.

The statistical procedure of averaging 5 numbers "around 5000" to get a mean of 5000 is *consistent* meaning if all the numbers are around 5000, we are guaranteed to get a mean around 5000. We could devise 5 unit tests to check each number. Since you're more familiar with your daily life, you may realise the step count may increase since you were in a hiking trip in the last two days. This may prompt you to check the step count.

In a data analysis, it is not practical to check every entry of the data, a similar strategy of devising tests to check for

- The combination of unit tests are not unique
- The unit tests provide guidance for diagnosing the results, but are not red flags: c(2000, 2000, 5000, 8000, 8000) will likely to fail the max diff test but receive a within expectation mean.

3 Applications

Three examples are presented to illustrate how the concept of analysis plan can be applied to data analysis. [toy example]. Section 3.1 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 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 $\text{cor}(\text{PM10}, \text{temperature}) \in [0.2, 0.6]$.

- add a paragraph to describe the simulation process
- add a fourth panel to describe the comparison of a right/ wrong expectation, i.e. correlation on PM10 and mortality

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. [note: weave the "actual data" into the example linear regression model]

Source: Article Notebook

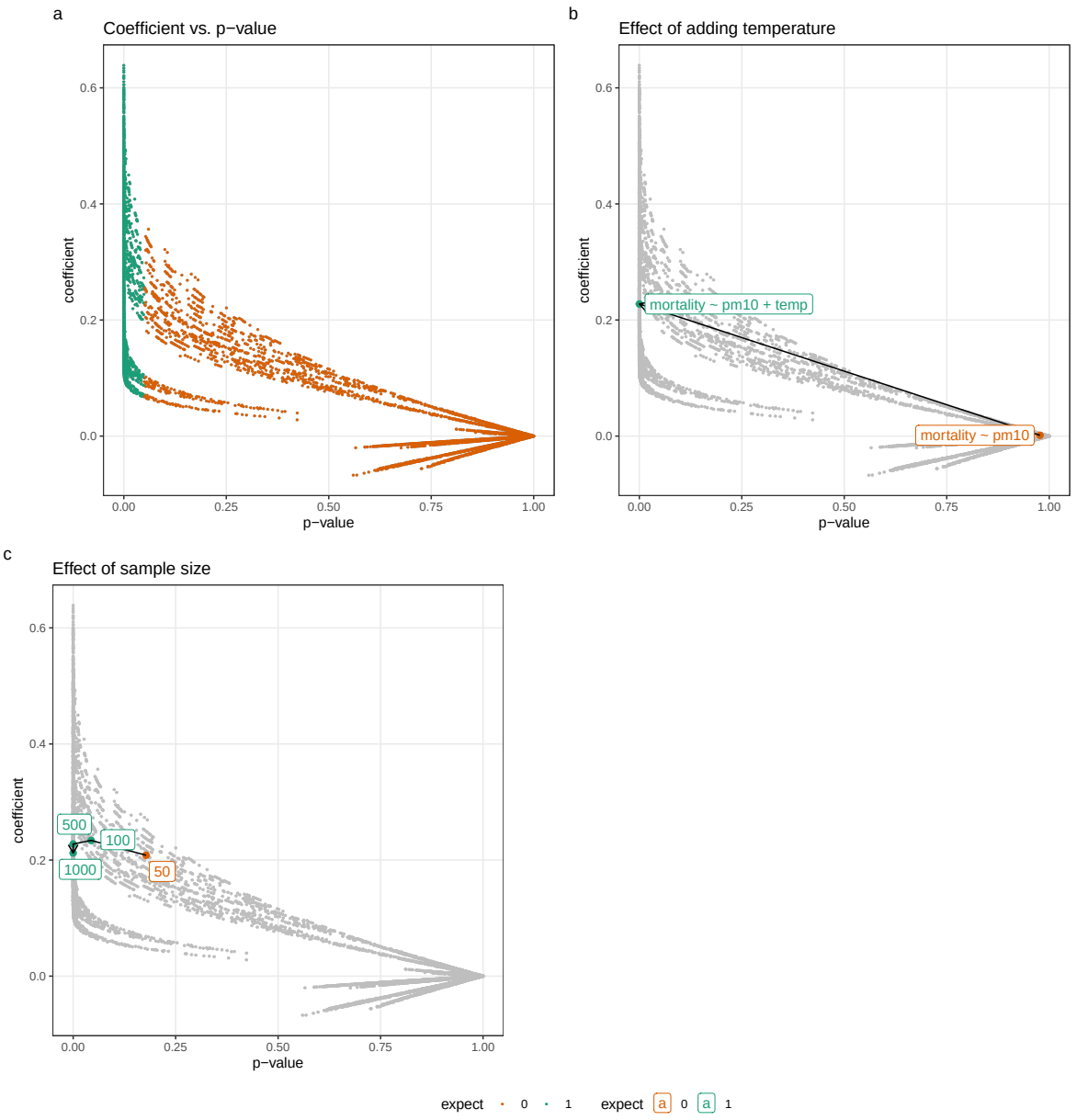

Figure 1: The result universe of linear regression model to study the effect of PM10 on mortality: a) colored by whether the p-value of PM10is significant (less than 0.05), b) the effect of adding temperature to the model for a sample size of 500, c) the effect of increasing sample size for a fixed correlation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

Source: Article Notebook

4 Conclusion

References

Garrett Grolemund and Hadley Wickham. A Cognitive Interpretation of Data Analysis. *International Statistical Review*, 82(2):184–204, 08 2014. doi: 10.1111/insr.12028. URL https://onlinelibrary.wiley.com/doi/10.1111/insr.12028.