Journal of Data Science xx (xx), 1–10 DOI: xx.xxxx/xxxx_xxxx
July 2023

My wonderful paper

Abstract

This is the abstract

1 Introduction

[background - In a data analysis workflow]

In a data analysis workflow -> going back to diagnose the descripency between the expectation and the result is often time-consuming and challenging.

For consumers of data analysis, it is often difficult to evaluate the analysis outcome. Not always have the script to run, even difficult when different expectations are held.

We propose here a concept called analysis plan 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. (think more closer here how our approach help to solve the problem)

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 4 provides examples of analysis plan [more details]. (need another section here or before examples?) Section 6 concludes the paper.

2 Analysis plan

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

TODO: more about analysis plan itself here

TODO: what if the expectation is "wrong"

The whole point being how can be build trust in other's analysis. Some provides scripts to run the analysis but that's not enough. We need to know what to expect from the analysis and that is often not explicitly specified.

An analysis plan is a set of analysis steps combined with expectations. Expectations represent our belief about certain aspects of the analysis, independent of the analysis itself. It can be divided into two types: *outcome expectation* and *plan expectation*. Outcome expectation refers to what we anticipate from the main result of the analysis based on prior knowledge.

They shape how we interpret the results and assess whether they are consistent with existing knowledge or indicate the need for updates (Grolemund and Wickham, 2014). For example, in public health, prior research shows the average increase in mortality rate per unit increase in PM10 is about 0.50% (Liu et al., 2019). This serves as an expectation for similar future studies. Plan expectations concern the intermediate steps within the analysis rather than the final outcome. They serve as checkpoints to detect deflection in the analysis process For example, we may expect temporal data to be ordered with no gaps and duplicates, or expect that temperature will be a significant covariate in the linear regression model of PM10 on mortality.

(might be useful) Analysis plans can be constructed at various granularities, at the highest level, one may only has a plan of the specific method used for analysing data and the expected outcome. This provides little guidance when a deviation from expectation occurs. At the lowest level, one may have a plan for each data entry and every data handling steps. This provides too much detail and may not be practical in practice.

Experienced analysts often have implicit expectation about the outcome and rely on a few "directional signs" to check when the outcome deviate from those expectation. However, these expectations are rarely made explicit within the analysis workflow. This makes it challenging for consumers of the analysis to evaluate the results, since it becomes difficult to disentangle whether discrepancies arise from differing expectations or from the use of statistical technique, without running the analysis themselves. Non-expert analysts, lacking prior knowledge or instinct, may not have clear expectations of the results. This can lead to reduced confidence of the analysis and makes it more difficult and time-consuming to diagnose the cause of the deviation when the results don't align with expectations. By explicitly formulating these expectations, an analysis plan can guide the analysis process, facilitate the communication and evaluate the validity of the results.

The expectations can be thought of as a set of unit tests used to validate the results of data

analysis. By specifying a range of values for these tests, multiple versions of the dataset can be generated to satisfy different sets of plan expectations. This allows us to present what we called the "result universe" – the complete set of possible results that can be obtained from one data analysis process. By visualizing the result universe, data analysis consumers can observe how changes in expectations affect the results and the range of alternative outcomes that could arise under different conditions. This enables them to evaluate the outcomes based on their own plan expectations and gain a broader perspective on how the actual results produced by analysts fit within this spectrum of possibilities, promoting transparency and trust in analysis.

Furthermore, by generating multiple versions of the data, we can emulate various scenarios within the same context for students to exercise judgement when conducting data analysis in a classroom setting.

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 Method

- multiple tests can be generated to diagnose different aspects of an analysis
- good tests

  1) are responsive to an unexpected result - accuracy,

2) motivate actions of analysts to investigate the results. This points towards a smaller set of independent tests

- Since both the plan/ outcome expectations can be phrased as unit tests, which output binary outcomes, consider using the plan expectations as predictors of the outcome expectations. This allows to generate the confusion matrix of predicting the outcome with the plan expectation.
- For an analysis, ideally, good plan expectations should maximize the detection of unexpected outcomes while minimize the false positive discovery, which suggest the use of precision and recall (REF) for evaluating the performance of the plan expectations.
  - precision: the proportion of unexpected results (TP) out of all the predicted unexpected results (TP + FP)
  - recall: the proportion of unexpected results (TP) out of all the actual unexpected results (TP + FN)
- A logic regression (ref) is used to model the relationship between the plan and outcome expectations. (justify the use of Logic regression)
- on independence of the tests

3.1 Toy example revisited

- provide interpretation at different scenarios:
  1) one test is flagged, the prediction is as expected:
  2) multiple tests are flagged, the prediction is unexpected,
  3) no test is flagged, the prediction is unexpected,
  4) no test is flagged, the prediction is as expected

4 Applications

Three examples are presented to illustrate how the concept of analysis plan can be applied to data analysis. [toy example]. Section 4.1 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]

4.1 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

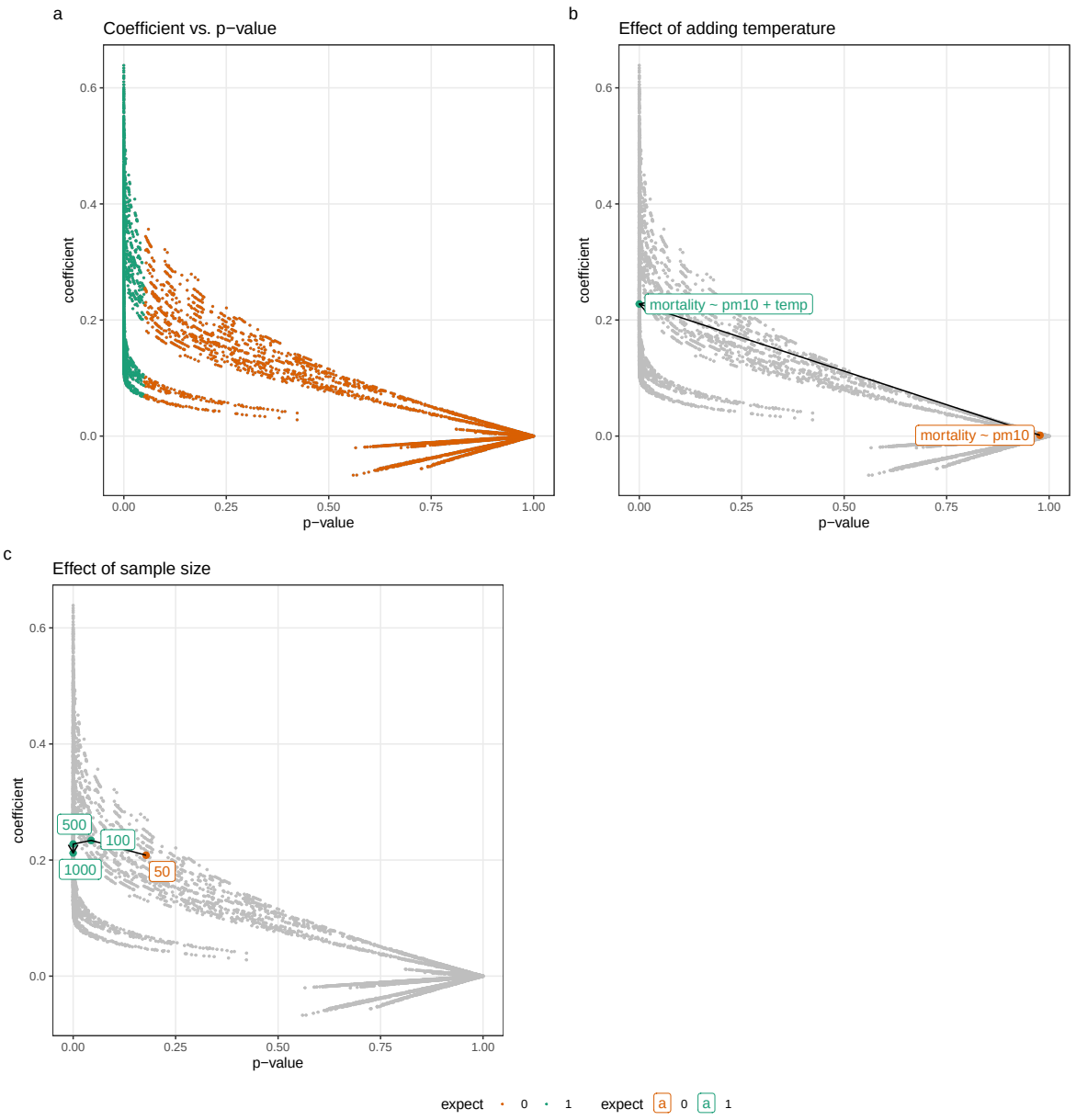

Figure 1: The result universe of linear regression model to study the effect of PM10 on mortality: a) colored by whether the p-value of PM10is significant (less than 0.05), b) the effect of adding temperature to the model for a sample size of 500, c) the effect of increasing sample size for a fixed correlation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

Source: Article Notebook

5 Discussion

- how to systematically simulate data is still unknown, sensitivity of the simulation to the resutls
- currently no automated way to generate unit tests

6 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.

Cong Liu, Renjie Chen, Francesco Sera, Ana M Vicedo-Cabrera, Yuming Guo, Shilu Tong, Micheline SZS Coelho, Paulo HN Saldiva, Eric Lavigne, Patricia Matus, et al. Ambient particulate air pollution and daily mortality in 652 cities. *New England Journal of Medicine*, 381(8):705–715, 2019.