Journal of Data Science xx (xx), 1–3 DOI: xx.xxxx/xxxx_xxxx
July 2023

My wonderful paper

Abstract

This is the abstract

1 Introduction

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 and 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.

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

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). From the literature, analysts may expect a significant PM10 coefficient in the linear model from the literature. This is the outcome expectation: p-value of PM10 coefficient is less than 0.05. There are multiple factors that can affect the outcome expectation of this analysis, for example, 1) sample size, 2) model specification, and 3) correlation structure between variables. Temperature is often an important confounder to consider in such study (add reference).

To construct the result universe, we can simulate datasets with different configerations of the above factors. 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

$$\text{cor}(\text{mortality}, \text{PM10}) \in [-0.01, 0]$$

$$\text{cor}(\text{mortality}, \text{temperature}) \in [-0.6, -0.2]$$

$$\text{cor}(\text{PM10}, \text{temperature}) \in [0.2, 0.6]$$

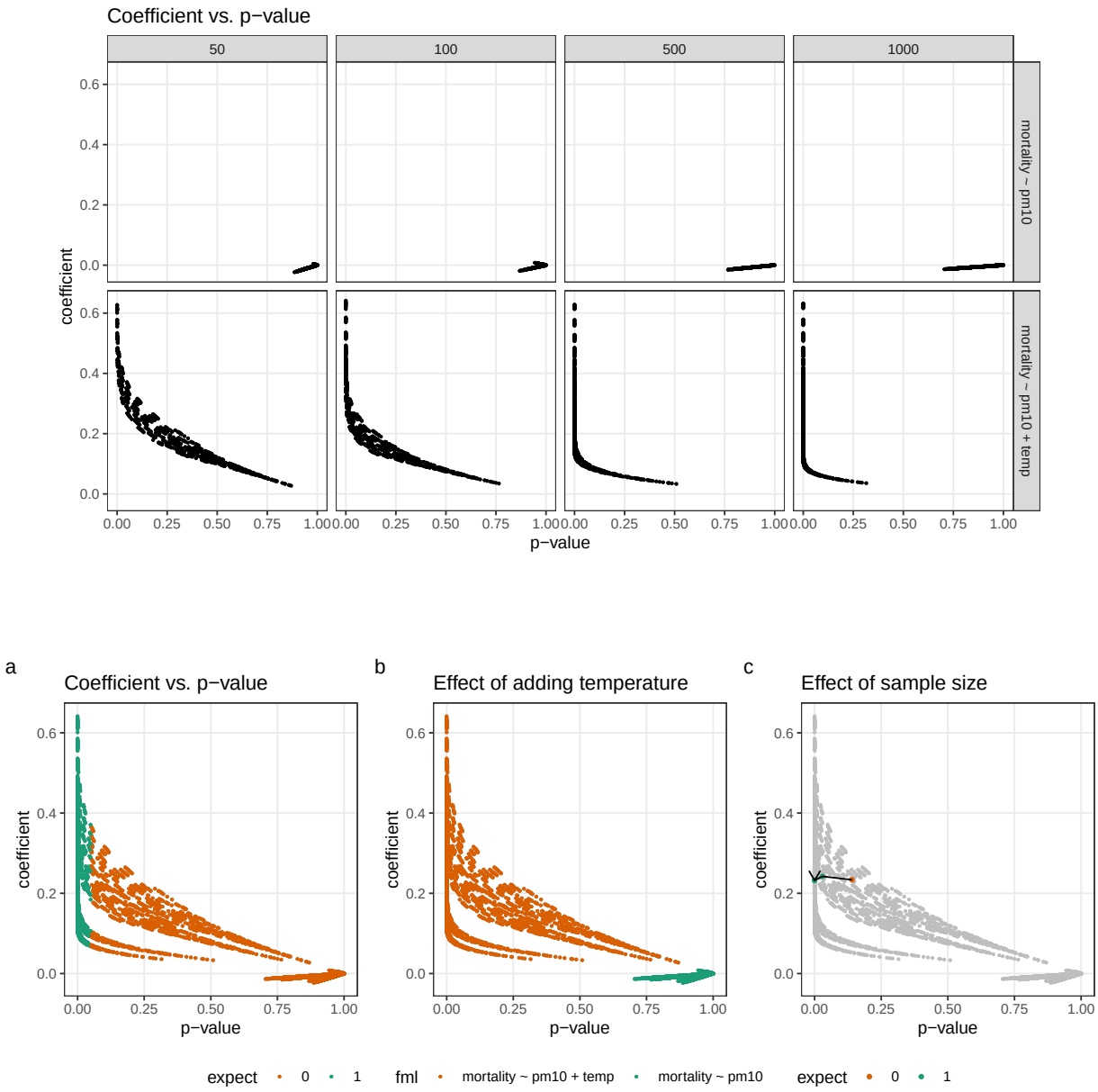

Figure 1

Figure 1 shows that result universe of the linear regression model with a) colored by whether the p-value of PM10 is significant (less than 0.05), b) highlighting the effect of adding temperature to the model for a fixed correlation structure, and c) highlighting the effect of increasing sample size.