

My wonderful paper

Abstract

Unexpected results in data analysis can prompt researchers to examine data quality and analysis steps. While some researchers can typically diagnose issues effectively with a few checks, others may struggle to identify appropriate checks to diagnose the problem. [an example of check] These checks are often informal and difficult to trace and discuss in publications, resulting in others questioning the trustworthiness of the analysis. To address this, we formalize the informal checks into an *analysis plan* that encompasses the analysis steps and (the unit tests): one test for whether the result meets expectations and multiple tests for checking the analysis. We then present a procedure to assess the quality of these unit tests based on their accuracy and redundancy on simulated versions of the original data. The accuracy is assessed using binary classification metrics, *i.e.*, precision and recall, while redundancy is calculated using mutual information. This procedure can be used to conduct a sensitivity analysis, compare different analysis plans, and to identify the optimal cutoff point for the unit tests.

Table of contents

1	Introduction	3
2	Literature review	4
2.1	Diagnosing unexpected outcomes in data analysis	4
2.2	Data analysis checks in statistical software	4
2.3	Unit tests	5
3	Analysis plan	6

3.1	Framing checks into unit tests	6
3.2	A toy example	7
4	Method	9
4.1	A workflow to assess the quality of unit tests	9
4.2	Toy example revisited	10
5	Applications	10
5.1	Linear regression	11
6	Discussion	14
7	Conclusion	14

1 Introduction

In a data analysis, an experienced researcher can often quickly assess whether a result meets their expectation, while others may be unsure of what to anticipate from the analysis or how to recognize when results diverge from expected norms. (These expectations are often based on prior knowledge, domain expertise, or common sense and they can often be framed into a Boolean question that can be answered by a unit test. For example, in a linear regression model, one may expect the p-value of the coefficient to be less than 0.05. On top of the expectation on the final results, researchers may also have expectations on the intermediate steps of the analysis process. For example, one may expect the data doesn't contain outliers, etc. These expectations are often checked early on to prevent an unexpected propagate through the analysis process.)

This expectation-setting process is often implicit and rarely documented or discussed in publications. (There are most publication requires open source code for publication, however, it is no instruction or guidance on checking ...). Yet, making them explicit is crucial for 1) helping junior researchers interpret results and diagnose the unexpected, 2) providing checkpoints that support independent replication or application of methods to new data, 3) aligning assumption across researchers from different backgrounds.

In this paper, we formalize these informal checks into an *analysis plan* comprising the analysis steps along with the associated unit tests. This formalization allows us to examine assumptions made about the data during analysis and to compare different unit tests in diagnosing unexpected outcomes. We then introduce a procedure to assess the quality of these unit tests based on their accuracy and redundancy across simulated version of the original data. The accuracy is assessed using binary classification metrics – precision and recall – while redundancy is measured using mutual information. We illustrate these concepts with a simple step count example and a scenario of diagnosing unexpected regression results when studying the effect of PM10 on mortality.

The rest of the paper is organized as follows: Section 3 describes the concept of analysis plan in detail. Section 5 provides examples of analysis plan [more details]. (need another section here or before examples?) Section 7 concludes the paper.

2 Literature review

2.1 Diagnosing unexpected outcomes in data analysis

(Peng et al., 2021)

TODO: what if the expectation is “wrong”

2.2 Data analysis checks in statistical software

In data analysis, checks are often performed to ensure the quality of the data and the validity of the model. For data checks,

- The `assertr` package (Fischetti, 2023) provides a set of functions to check the quality of data, including the presence of missing values, the distribution of variables, and the relationship between variables.
- Similarly, the `pointblank` package (Iannone et al., 2024) is a data validation tool that allows users to define data quality checks.

For model checks,

- the `performance` package (Lüdtke et al., 2021) provides a set of functions to assess the performance of a model, including residual analysis, goodness-of-fit tests.
- The `yardstick` package (Kuhn et al., 2024) provides the calculation of model performance metrics in the `tidymodels` (Kuhn and Wickham, 2020) ecosystem.

These checks test on specific aspect of the data or the model but doesn’t reflect the underlying

assumption of the data analysts about the data or the problem context. It may be obvious to see the assumption if someone is testing for data quality on whether a column in the data frame is between two values. In a much complex analysis, it is difficult to trace the assumptions made by the data analysts. Currently, most of these assumptions are not documented or discussed in the analysis process, or at least in a machine readable format.

2.3 Unit tests

Why develop unit tests in software engineering? Because it is difficult to predict the outcome of a program, especially when the program is complex. Because the loss associated with a program failure is costly.

Software testing relies on “oracles” to provide the expected output necessary for verifying test results. For example, to test whether the program correctly calculates $1 + 1$, one need to supply the correct answer, 2. However, establishing this “correct” output can sometimes be challenging, where obtaining a solution may be difficult without the program itself. This situation leads to the oracle problem (Barr et al., 2014). In data analysis, the similar oracle problem can happen, as the “truth” of an outcome, the expectation, depends on the underlying theory or interpretation. For example, in a linear regression model, the significance of a coefficient may be expected or unexpected based on the theory, making it challenging for researcher with a different theory to assess the results and the analysis.

Existing tools for checks in data analysis including those check for data quality and model diagnostics. In software development,

- the `testthat` package (Wickham, 2011) provides the infrastructure for testing R code. It allows users to write unit tests for functions and packages.
- The `assertthat` package (Wickham, 2019) helps to write assertion statements that are supposed to be true at a certain point in the code for defensive programming.

3 Analysis plan

3.1 Framing checks into unit tests

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

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.

3.2 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.

4 Method

4.1 A workflow to assess the quality of unit tests

In real-world applications, it is rare to create a set of unit tests can fully guarantee expected results. On one hand, there is the cost of effort involved in manually developing all these tests; on the other, there is the inherent complexity of the problem. (can we - if we're thriving for detecting 95% of the cause?). However, the quality of unit tests can be evaluated by simulated data. One set of tests is considered better than another if a small set of tests can reliably detect unexpected outcomes, which brings two criteria in the evaluation metric: accuracy and parsimony.

Accuracy refers to a set of tests' ability to accurately detect unexpected outcomes while minimizing false positives and false negatives. A false positive can indicate (caution or skepticism on checking the data), whereas a false negative suggests the tests may lack sensitivity to unexpected outcomes. Since both plan and outcome expectations can be framed as unit tests with binary outcomes, one approach is to predict the outcome expectation based on the plan expectation. Performance of the tests can then be evaluated using precision and recall metrics from the confusion matrix:

- precision: the proportion of correctly identified unexpected results (true positives) out of all the predicted unexpected results (true positives + false positives)
- recall: the proportion of correctly identified unexpected results (true positives) out of all the actual unexpected results (true positives + false negatives)

The second criteria is parsimony in the tests. While tests may score high on accuracy, they may be less effective at explaining the reasons behind unexpected results. This could happen if a set of tests are all tangentially related to the cause of the unexpected results, but none addressing the root cause. It may also occur if the tests are correlated with one another, leading to redundancy.

- explain mutual information
- explain logic regression
- explain combining precision, recall, and independence together through “means”
- A logic regression (ref) is used to model the relationship between the plan and outcome expectations. (justify the use of Logic regression)

4.2 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

5 Applications

Three examples are presented to illustrate how the concept of analysis plan can be applied to data analysis. [toy example]. Section 5.1 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]

5.1 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 ($p\text{-value} \leq 0.05$) 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 ($\text{mortality} \sim \text{PM10}$), 2) linear model with PM10 and temperature as covariates ($\text{mortality} \sim \text{PM10} + \text{temp}$). 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](#)

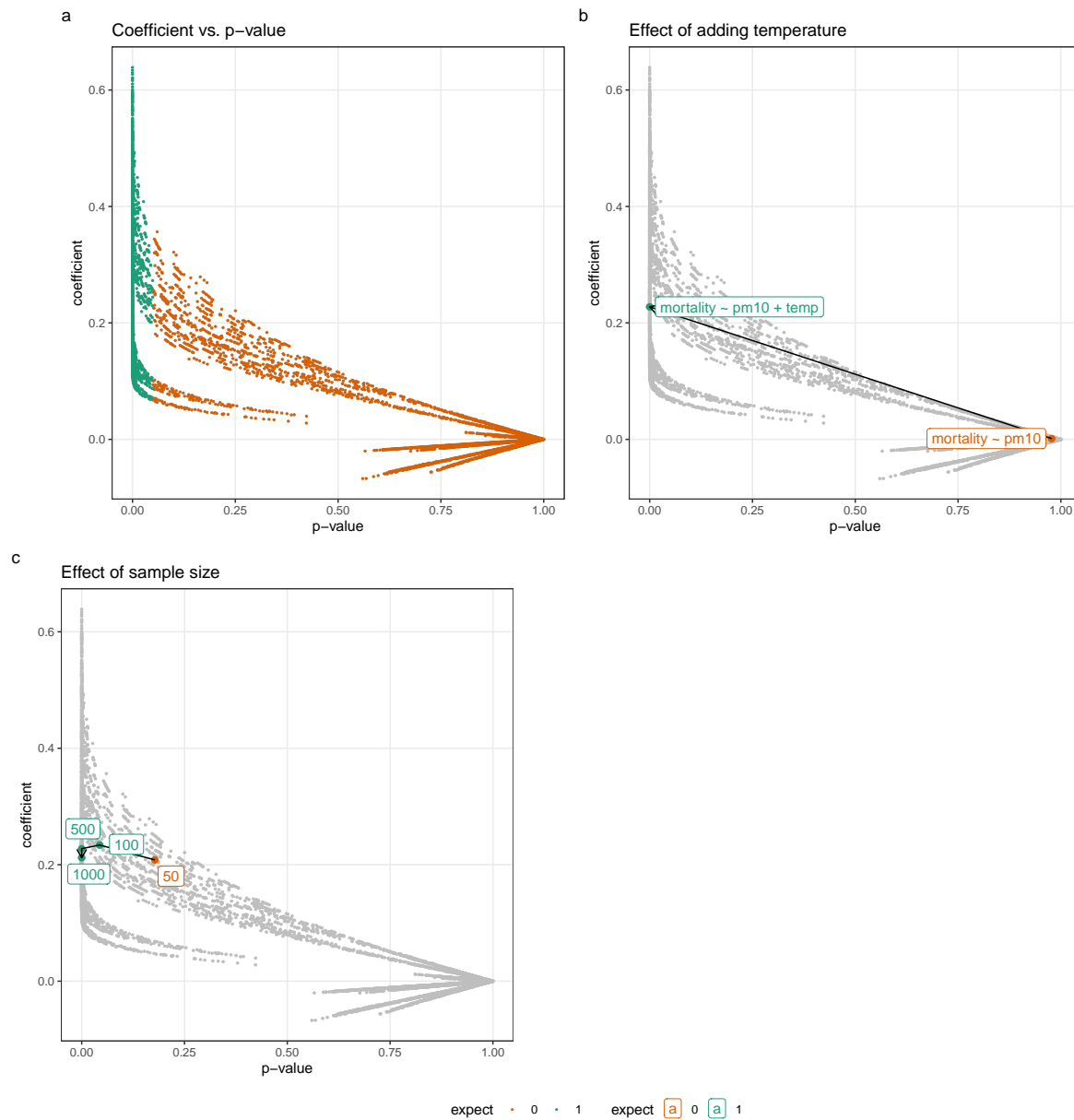


Figure 1: The result universe of linear regression model to study the effect of PM10 on mortality: a) colored by whether the p-value of PM10 is significant (less than 0.05), b) the effect of adding temperature to the model for a sample size of 500, c) the effect of increasing sample size for a fixed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

Source: [Article Notebook](#)

6 Discussion

- how to systematically simulate data is still unknown, sensitivity of the simulation to the results
- plotting is a critical way to check data and they can still be frame into a unit test. it is a open problem to how to encode the visualization into the unit tests. Maybe a procedure like confirm plot (this looks alright to you) and then press the button to continue
- currently no automated way to generate unit tests. It is interesting to see the automation of generating unit tests, although it requires the inputs from experts across a wide array of common scenarios.

7 Conclusion

References

- Earl T Barr, Mark Harman, Phil McMinn, Muzammil Shahbaz, and Shin Yoo. The oracle problem in software testing: A survey. *IEEE transactions on software engineering*, 41(5): 507–525, 2014.
- Tony Fischetti. *assertr: Assertive Programming for R Analysis Pipelines*, 2023. URL <https://CRAN.R-project.org/package=assertr>. R package version 3.0.1.
- 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>.
- Richard Iannone, Mauricio Vargas, and June Choe. *pointblank: Data Validation and Organization of Metadata for Local and Remote Tables*, 2024. URL <https://CRAN.R-project.org/package=pointblank>. R package version 0.12.2.
- Max Kuhn and Hadley Wickham. *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.*, 2020. URL <https://www.tidymodels.org>.

- Max Kuhn, Davis Vaughan, and Emil Hvitfeldt. *yardstick: Tidy Characterizations of Model Performance*, 2024. URL <https://CRAN.R-project.org/package=yardstick>. R package version 1.3.1.
- 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.
- Daniel Lüdecke, Mattan S. Ben-Shachar, Indrajeet Patil, Philip Waggoner, and Dominique Makowski. performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, 6(60):3139, 2021. doi: 10.21105/joss.03139.
- Roger D. Peng, Athena Chen, Eric Bridgeford, Jeffrey T. Leek, and Stephanie C. Hicks. Diagnosing Data Analytic Problems in the Classroom. *Journal of Statistics and Data Science Education*, 29(3):267–276, September 2021. doi: 10.1080/26939169.2021.1971586. URL <https://doi.org/10.1080/26939169.2021.1971586>.
- Hadley Wickham. testthat: Get started with testing. *The R Journal*, 3:5–10, 2011. URL https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.
- Hadley Wickham. *assertthat: Easy Pre and Post Assertions*, 2019. URL <https://CRAN.R-project.org/package=assertthat>. R package version 0.2.1.