Designing statistical models

2022-09-20

Contents

Reading:

Required Reading (everyone):

- Statistical Rethinking, Ch. 6 ("The Haunted DAG & The Causal Terror").

Reading Notes:

There are several layers to this chapter:

1. At the simplest level, this chapter builds upon what we learned about *confounding* and *masking* effects among predictor variables in a regression analysis to introduce three new potential problems:

    a. Multicollinearity
    b. Post-treatment bias
    c. Collider bias

    It then generalizes from these, in the "Confronting Confounding" section, to give us a general way to think about potential problems that arise from diffrent kinds of causal relationships among predictor variables and the dependent variable.

2. Simultaneously to developing this content about different kinds of problems with regression, the book also models a very important method: Before analyzing experimental or observational data, it's useful to generate *synthetic data* (also called "artificial data" or "fake data") from a specified statistical process model (e.g., a DAG) and testing whether your regression analysis will give useful and sensible results when you know beforehand what the answers should be.

    For instance, if you generate a bunch of synthetic data by drawing it from a normal distribution with mean μ and standard deviation σ (`y <- rnorm(1000, mu, sigma)`), then when you analyze that data, the posterior distributions for μ and σ should be compatible with the actual μ and σ you used to generate your data.

3. There is also a layer of philosophical commentary about what you can and can't expect statistical analysis to tell you. As McElreath writes in the "Rethinking" box at the bottom of p. 173:

    > Model `m6.7` misleads because it asks the wrong question, not because it would make poor predictions. ... [P]rediction and causal inference are just not the same task. No statistical procedure can substitute for scientific knowledge and attention to it.

As you read this chapter, pay attention to all three layers of the discussion.

A few details I'd like you to pay particular attention to are:

- The concept of *non-identifiability* that comes up in the section on multicollinearity. McElreath says that multicollinearity is only an issue for interpreting models, not for statistical analysis. This is true in principle, but multicollinearity can be a big problem for practical regression analysis when we get to Monte Carlo sampling in Chapter 9, because it can prevent a Monte Carlo analysis from converging. We'll discuss this when we get to that point in the semester.

- The contrasting errors of omitting important variables and including problematic variables, in the section on post-treatment bias. If there's a problem with omitting variables, but also with including too many variables, how do we come up with a sensible model?

  McElreath goes into the philosophical mode at several places, saying at the beginning of the chapter,

  > The previous chapter demonstrated some amazing powers of multiple regression. ... This may encourage the view that, when in doubt, just add everything to the model and let the oracle of regression sort it out.
  >
  > Regression will not sort it out. Regression is indeed an oracle, but a cruel one. It speaks in riddles and delights in punishing us for asking bad questions.

  And in the summary at the end of the chapter, he writes

  > Multiple regression is no oracle, but only a golem. It is logical, but the relationships it describes are conditional associations, not causal influences. Therefore, additional information, from outside the model, is needed to make sense of it.

- In the section on post-treatment effects, there is good discussion on the similarities and differences between *experimental* and *observational* studies. This is worth reading carefully.

- The collider bias effect is subtle and easy to misunderstand. Read carefully and make sure you understand it. Following along with the code can be helpful. Play with code and data to understand how a spurious correlation can arise when you select data to analyze from a larger population, based on some selection criterion.

  Can you see how collider bias relates to the problem of *non-identifiability* in the section on *multicollinearity*?

- Section 6.4 synthesizes all the different problems we studied in chapters 5 and 6, and presents a more general theory of statistical confounding, based on different patterns that causal DAGs can take. McElreath then discusses a general approach, called "shutting the backdoor" to dealing with these confounding effects.

McElreath concludes that DAGs are useful, but they won't solve all the problems you'll run into in statistical analysis. Having a solid theory of the thing you're studying is an important complement to DAG analysis of the structure of your model.