Many variables (part 1)

2022-09-08

Contents

Reading: 1

Reading:

Required Reading (everyone):

- Statistical Rethinking, Ch. 5 ("The Many Variables and The Superfluous Waffles"), section 5.1, pp. 123–144.

Reading Notes:

This chapter gets at some very important issues with linear regression modeling (these same issues will apply to more complicated regression models we will study later). The chapter begins by introducing an illustrative example of the difference between correlation and causation. States with greater numbers of Waffle House restaurants have higher divorce rates, but the restaurants aren't the cause of divorce. Rather, due to historical accidents, there is a spurious correlation between Waffle House restaurants and high divorce rates. In this chapter we examine different ways to try to explore correlations between multiple variables and try to figure out which correlations might be causal.

There are several possibilities:

- Variable A has a direct causal influence on variable B
- Variable A has a spurious relationship with variable B: they may be correlated, but this is because a third variable C has a causal effect on both A and B, but there is no direct link between A and B.
- Variable A has an indirect causal relationship with variable B: A has a causal effect on a third variable C, which has a causal effect on B.

The book introduces a new kind of notation, the *Directed Acyclic Graph* (*DAG*), which represents causal relationships as arrows (*directed*), and it's *acyclic*, meaning that there are no loops or cycles, where two variables have direct or indirect causal relationships on each other. The R library `dagitty` can draw *directed acyclic graphs* for you and it can also analyze causal relationships in DAGs. DAGs help us answer the question, *Is there any additional value in knowing a variable, once I already know all of the other predictor variables?*

The chapter shows us how to set up and analyze *multiple regression models* where the dependent variable we are trying to predict depends on many predictor variables. We also learn a number of plots, such as the residual plots in Fig. 5.4, which we can use to diagnose how well a model works for your data and to help us figure out the set of causal relationships that would best describe the data (that is, they help us figure out the DAG that best describes the relationships among a bunch of variables).