Geocentric Models

2022-09-06

Contents

Reading:

Required Reading (everyone):

- Statistical Rethinking, Ch. 4 ("Geocentric Models").

Reading Notes:

In chapter 3, we focused on sampling, a *nonparametric* method for statistical analysis. Here, we focus on *parametric* methods that use functional forms for probability distributions. The most famous one is the Gaussian, or Normal, distribution. Parametric functions rarely describe actual data exactly. They are simplified approximations. So why do we use them? Because in science it's often better to be simple and approximately right than to be absurdly complicated in the quest to represent every detail of the data.

There is a time and place for very complicated and detailed mathematical modeling, but elegance and simplicity are important, and valuable, in science. Albert Einstein famously said that scientific theories and mathematical models "should be as simple as possible, but no simpler." The statistician George Box said "all models are wrong, but some are useful." Simplicity is a virtue, but it's not an absolute one. How simple or complex a model should be depends on what you're using it for. If you are designing a rocket engine that astronaut's lives will depend on, your engineering calculations need to use a very complex model that takes account of all the details of the rocket engine, but if you're designing an engine for a toy model rocket that weighs a pound or two and that you'll be launching for fun to go a few hundred feet in the air, a much simpler model will suffice. McElreath uses the term *geocentric model* to describe mathematical and statistical models that deliberately omit lots of detail about the real system, but which are useful for our purposes and much easier to understand than more complicated models.

A big piece of this chapter concerns the Normal distribution. Even when data is not well described by Normal distributions, if we take a bunch of measurements and add them together or multiply them together, the result of the addition or multiplication will usually approximate a Normal distribution, especially when we are adding or multiplying a large number of measurements. This result comes from the *central limit theorem* of statistics.

This chapter walks us through how to use both grid-sampling and quadratic approximations to analyze data using simple models: First, estimating normal distributions to describe the probability distribution of data. Then using linear regression models to measure the relationships between different measured variables, adding complexity to regression models by extending them from linear to quadratic or higher polynomial forme. And finally, using nonparametric regression methods, such as B-spline and cubic spline regression. All of these methods are valuable golems for you to have in your toolbox. I expect that fitting Gaussian models to statistical distributions and performing linear regression analysis is familiar to most of you, but I expect that few of you have seen nonparametric spline regression before, so I will spend a lot of time in class discussing that part of the chapter.

Some important things to pay attention to in the chapter:

- The difference between *probability mass* and *probability density* on p. 26.
- The language for describing models on p. 27. Understand the different parts: Which variables are *data*, and which are *parameters*? Note the notation of the "~" symbol for *stochastic relationships* (see p. 78.
- The importance of plotting both your data and your models, to understand what's going on and to check whether the model makes sense, as well as inspecting tables of marginal distributions (section 4.4.3 on pp. 98–110).
- The importance of *standardizing* variables for modeling, especially with nonlinear models (p. 111). Standardizing a variable means subtracting the mean and dividing by the standard deviation. $z = (x - \bar{x})/\sigma_x$, where z is the standardized variable, x is the unstandardized variable, $\bar{x}$ is the mean of x, and σ_x is the standard deviation.
  - Standardizing variables has two advantages:
    1. It helps statistical analysis algorithms perform better. Some algorithms don't work well if x is very much larger than 1 or very much smaller than 1.
    2. It helps you understand the results of your regression analyses. There are convenient rules of thumb for interpreting regression coeffiients when the variables have been standardized.