Statistics Minus The Math: An Introduction for the Social Sciences

Nathan Favero

2024-02-11

Table of contents

Introduction

This version (1.2) was updated 1/28/2023.

Version 1.1 available at https://nathanfavero.com/teaching-tools/. The only change from 1.1 is that the discussion of transforming variables now appears in Ch. 2 (rather than Ch. 3).

This book was largely adapted from the public domain resource *Online Statistics Education: A Multimedia Course of Study* (http://onlinestatbook.com/ Project Leader: David M. Lane, Rice University).

1 Models and Uncertainty

Before I leave my house each morning, I need to decide whether to take an umbrella. So I check my phone to see whether it's supposed to rain. Instead of giving me a direct yes or no answer, the weather tells me the percentage chance of rain for the day.

Why does the weather app give me a percentage? Because there's uncertainty. Science has done a lot to help us understand the weather. And as our understanding of the weather improves, our predictions get better. But we still can't predict rain perfectly.

Facing uncertainty is a common problem when we're looking at data. Whether we're trying to explain the weather, human behavior, or even plant growth, we can't make perfect predictions because there are things we can't fully explain with our current scientific knowledge.

In statistics, we have several tools that allow us to acknowledge uncertainty. This enables us to build models like the ones powering my weather app—models that give us a prediction that includes a description of how uncertain we are. Some days we are 100% sure it will rain, other days only 60%.

In order to build these models that acknowledge uncertainty, we need a way to talk about what we do know and what we don't know. Let me give a very simple example of a model that accounts for uncertainty:

$$happiness = 3.0 + 2.3 \times income + \varepsilon \tag{1.1}$$

This model attempts to explain one's level of happiness based on their income. You might notice that it looks very similar to the regression equations we saw in Chapter 3. That's because regression is one of the main tools used to estimate a model that includes uncertainty.

What does this model mean in practical terms? Well, there are no obvious units we can use to quantify the amount of happiness someone experiences, so the exact values of the numbers we see are not particularly meaningful. But the fact that there's a positive number (2.3) that is being multiplied by income implies that as income gets bigger, happiness gets larger.

The key part of this equation that I want to focus on is the little Greek letter at the end of the equation: . This letter is called "epsilon," and it is often used to represent what we call an **error term** (also sometimes called a **disturbance term**). The error term ( ) represents everything else besides income that affects happiness. By including an error term, we are acknowledging that we can't perfectly predict one's level of happiness based on their income.

We think that knowing one's income will help us predict their happiness, but we know there are other factors we won't be able to measure or identify that will also affect happiness. Thus, if all we know about someone is their income, we will have uncertainty about their exact level of happiness. By including an error term ( ) in the model, we make clear that we only claim to have a partial understanding of happiness, not a complete one.

Think for a moment about how few topics we could study if we didn't have the freedom to build models that include uncertainty. We'd only be able to build a model of a dependent variable after we had identified (and measured) *all* of the factors that affect that variable! We wouldn't be able to build a model of rain since we don't know all of the factors that affect the rain. We couldn't build a model of voting behavior since we don't know everything that affects how someone will vote. By including an error term in our model, we can build models even when our understanding of something is incomplete.

The first part of our model that appears on the right side of the equation (3.0 + 2.3 x *income*) is sometimes described as the *systematic* part of our model. It's what we would use to build a prediction of happiness if all we know about some is their income level. Suppose, for example, that someone has an income of 4 units (perhaps income is measured in tens of thousands of dollars of annual income, so a salary of $40,000 is coded as a 4). According to our model, that person's happiness would be:

$$happiness = 3.0 + 2.3 \times (4) + \varepsilon$$

$$happiness = 12.2 + \varepsilon$$

We, therefore, predict that someone with an income of 4 will have a happiness of 12.2, but we also acknowledge that their actual happiness will likely be a bit different from our prediction since our model indicates that their actual happiness will equal 12.2 plus the value of the error term ( ).

The error term describes something unknown, so we can't measure it or directly observe it. But what we can do is talk about its characteristics using concepts from probability theory. Specifically, we're going to describe the value of the error term as being randomly selected. You may have dealt with randomness in math classes before using examples such as coin flips, die rolls, or drawing cards from a 52-card deck. Just as the likelihood of different outcomes from parlor games can be described using probability, we're going to use probability to describe different possible values for the error term of a statistical model.

1.1 Assumptions about error terms

It's easy to write out an equation that includes an error term, but we are not going to be able to do much with our model unless we make some assumptions about the error term. One of the most important (and challenging) parts of doing statistical analysis is making assumptions

about the possible values of the error term. Different assumptions about the error term can result in very different conclusions.

Let's again consider the simple model of happiness that was introduced above:

$$happiness = 3.0 + 2.3 \times income + \varepsilon$$

We might assume the following things about the error term ( ):

1. The values of the error term ( ) can be described by a normal distribution with a mean of 0
2. Knowing someone's income doesn't help us predict the values of the error term ( )

What do these two assumptions mean?

First, if the error term ( ) follows a normal distribution with a mean of zero, that means that (according to our model), people are just as likely to have a positive value of the error term as they are to have a negative value of the error term. In other words, all those factors we haven't accounted for in our model are equally likely to push people in the direction of being happier or in the direction of being less happy. Our model and assumptions tell us that if we predict happiness purely based on income, we'll *overestimate* some people's happiness, and we'll *underestimate* an equal number of people's happiness.[1]

Second, these assumptions allow us to describe how much individual observations will tend to deviate from our income-based predictions. We haven't specified in our assumptions what the standard deviation is for the normal distribution for the error term ( ), but statistical analysis will let us estimate the standard deviation of an error term. And we know that there is a 95% chance of drawing a value within two standard deviations of the mean for any normal distribution. So whatever the standard deviation of the error term ( ) is, we would expect that 95% of the time, the error term will take on a value that is within two standard deviations of zero. Conversely, 5% of the time, the error term will take on a value that is more than two standard deviations from zero. Suppose that the standard deviation of the error term ( ) happens to be three. If we have a dataset containing the income and happiness of 1,000 randomly selected people, we would expect that about 950 of these people will have a level of happiness that falls within six units of our income-based prediction. But for about 50 of these people, our prediction of their happiness will be off by more than six units.

Third, our assumptions imply that income is not tied in any consistent way to (the total sum of) factors other than income that also affect peoples' happiness. Remember, the error term ( ) represents all factors other than income that affect satisfaction. If income is related to these other factors, then the value of income should help us predict the value of the error term. For example, if having a stable environment in childhood tends to cause both higher incomes

[1]Note that these deviations from our prediction don't imply that our model is wrong; our model explicitly acknowledges that we'll get only imperfect estimates if we predict happiness based on income, since the unobserved error term ( ) also contributes to happiness.

and greater happiness in adulthood, the error term will partially reflect the effect of childhood stability on happiness, so high incomes (which are partially caused by childhood stability) will be probably be predictive of a more positive error term. This would constitute a violation of our assumptions since we specifically indicated that income wasn't predictive of the error term. As this example illustrates, our assumptions about error terms are often quite strict, making it rather difficult in practice to build good models that account for uncertainty.

1.2 Models and probabilistic thinking

Despite the difficulty inherent in building models that accommodate uncertainty, we have little alternative unless we wish to only build models of things we think we can predict with 100% accuracy. And fortunately, our models do not always have to be perfectly correct in order to generate useful predictions or explanations. As the statistician George Box famously said, "all models are wrong, but some are useful."

An important part of learning to do good statistical analysis is learning to think clearly about models so that you can pick out a model that is useful for whatever it is you want to accomplish. And the first step toward understanding many statistical models is learning to think about the world in probabilistic terms, as we've done here in this reading. Probabilistic thinking asks questions like:

- Based on what I do know and what I don't know, what can I predict?
- How does adding or removing different pieces of information change my prediction?
- How much uncertainty is there in my prediction?
- How often will my prediction differ greatly from what actually happens (even if my model is correct)?