What is a Statistical Model?

Greg Kochanski

November 3, 2010

1 Reading a Statistical Model

Statistical models in R (and in much of the rest of the statistical world) use a specialized notation. They look something like this:

$$Z \sim W + X + Y. \tag{1}$$

That string of symbols[1] means Z is the dependent variable[2] (the thing that we are trying to predict), and W, X, and Y are the independent variables[3] (these are the things we will predict with). You could read this as "Z is modeled as a function of W, X, and Y." Note that the + signs don't quite mean addition, they are closer to "and". If this is a linear model, then we can easily write this in standard mathematical notation as

$$Z = a \cdot X + b \cdot Y + c \cdot Z + \epsilon. \tag{2}$$

Where a, b, and c are constants that multiply the independent variables. Also, ϵ is a uncorrelated Gaussian random variable: it is the unexplained part of the data, the noise, the place to hide prediction errors.

A simple linear regression or ANOVA[4] with one independent variable would look like this: $Z \sim X$, which means that Z depends on X. If we are talking about an ANOVA, it means that the model predicts a different value of Z for each different category of X.

An even simpler regression with no independent variable is $Z \sim 1$, which means that Z does not depend on anything. Specifically, we are modelling Z as a constant, a. Mathematically, that means $Z = a + \epsilon$. (Note that one can combine the 1 with independent variables to add a constant to any linear model. Note also that the constant term is sometimes not written down.)

[1] "Equation" isn't really the right term, because no one really expects the two sides to be equal. This describes a statistical way to predict data, and no one really expects the prediction to be perfect. So, the left and right sides should have similar values, not equal values. Perhaps it should be called a "similarity"?

[2] This is typically your data.

[3] These are typically the conditions you are investigating. They are also often called "factors".

[4] Analysis of Variance. ANOVA is a statistical technique that can be very useful on categorical data. (Words are a good example of categorical data.)

Finally, statistical models can include interaction terms, like this:

$$Z \sim X \star Y, \tag{3}$$

which which is most useful for categorical or ordinal[5] data. It means that our model for Z can be different for each combination of levels of X and Y. In practice, one looks up a value for each X, Y combination, as from a table.[6] This contrasts with $Z \sim X + Y$ where you look up the effect for X, then the effect for Y, and then add them together.[7]

Writing a statistical model is helpful because it forces you to think about what depends on what.

2 Exclusive and Nonexclusive descriptions

Any real object has a myriad of properties. When we say "The table is brown." we do not mean that its only property is brownness. That would be hard to imagine: it would be something close to a Platonic Ideal, perhaps. Rather, saying that "The table is brown." simply means that brownness is one of its many properties. A table can be brown and flat and wood and antique all at once. Most properties can be combined, fairly arbitrarily, with the word "and".

But, when we talk about statistical models, we do not talk that way. When we say *grammaticality* $\sim$ *frequency*, we are saying something more specific than the casual English translation that "grammaticality depends on usage frequency". In analogy with brown tables, people often incorrectly assume that "depends on usage frequency" is just one of many properties that "grammaticality" may have. But, that is not so, because it is not consistent with the mathematics behind the statistical model.

Rather, *grammaticality* $\sim$ *frequency* is the entire description of what correlates with "grammaticality". Think of describing the colour of the table: if it is brown, it is not green. Or, think of marriage instead of friendship: if John is married to Alice, he is not married to anyone else. It means that the grammaticality score for a phrase depends on the frequency with which the phrase occurs, and nothing else. It means precisely what Equation 3 says. It means that after you subtract off the *frequency* effect from *grammaticality*, whatever remains is uncorrelated, random, Gaussian noise. Anything that violates this assumption invalidates the statistical analysis.

So (with some exceptions) it makes no sense to make two different tests on the same data set. The first test asserts that the true situation is

$$\text{grammaticality} \sim \text{frequency}, \tag{4}$$

[5]Ordinal data are things like "small", "medium", "big". You know the order in which they occur, but they are not quite numbers. Specifically, you do not really know how big the gaps in between them are.

[6]For instance, if X has 4 levels and Y has 4 also, then the model can take on $4 \times 4 = 16$ different values, and it takes 16 numbers to fully define the model.

[7]The model still predicts 16 different possible outcomes, but it only requires $4 + 4 = 8$ numbers to be fully defined.

and if the other test asserts that

$$\text{grammaticality} \sim \text{length}, \tag{5}$$

then they are (generally) inconsistent. The first says it depends only on the frequency with which the phrase occurs, the second says it depends only on the length of the phrase. Generally, if we were to find that one effect exists, the other test would then become invalid.[8] Generally, if both tests turn out significant, then neither is valid.

So, what are these exceptions? When can you run two tests on the same data set without one invalidating the other? The answer is in the mathematics. The assumption built into these statistical models is that "... whatever remains is uncorrelated, random, Gaussian noise." If you can meet that assumption with two simultaneous tests, you are in good shape. To understand when we might meet this requirement think about a model that includes both effects:

$$\text{grammaticality} \sim \text{frequency} + \text{length}. \tag{6}$$

Suppose Equation 6 were to fit the data perfectly. Then, the error term in a simpler model (e.g. Equation 4) would just be the *length* term in the more complex model (Eq. 6). So, we would meet the mathematical requirement that the error term be uncorrelated Gaussian noise if (and only if) *length* has a Gaussian distribution and it is uncorrelated with *frequency*.

This leads to a good rule of thumb: if your independent variables are correlated (e.g. Table 1), you must do a large multi-parameter regression. Only if they are uncorrelated[9] can you do a sequence of small, single parameter regressions and get meaningful results.[10]

Why does Table 2 indicate trouble? Simply because if your data were distributed that way and you saw an effect, you could not easily tell whether it was caused by the phrase length or by the frequency. They go together, so it is hard to disentangle which one is important. Wait! It's worse than that. They could both be important, in any combination: $1/3 \times \text{length} + 2/3 \times \text{frequency}$, for example. In fact, it can be even worse than *that*: the two factors could be working against each other. It might be that the effect is caused entirely by *frequency*, but is partially cancelled out by *length* working in the other direction. As a result, when there is strong correlation between independent variables, you might know there is a large effect but be entirely unable to describe its origin.

In such a case, the problem with doing tests on one variable at a time (e.g. Equations 4 and 5) is that they might falsely tell you that *both* variables show

[8] We would invalidate the other test's results of "not significant", opening up the possibility that there might be a correlation from both *length* and *frequency* to *grammaticality*.

[9] Equivalently, you can say the independent variables are independent of one another. (This terminology is a bit confusing because the first "independent" is a statistical term, and the second "independent" is a term from probability theory.) What the second "independent" means is that knowing the value of one of your independent variables does not help you to predict the value of any others.

[10] We normally ignore the requirement for Gaussian distributions, because Gaussians are pretty common. But, strictly speaking, if the distributions of your independent variables were very non-Gaussian, there would be difficulties.

	frequency=high	frequency=medium	frequency=low
length=short	24	25	26
length=medium	23	22	24
length=long	25	25	24

Table 1: *The distribution of two independent variables that are nearly uncorrelated. Each box counts the number of phrases in a corpus with that frequency and that length. You can see very little relationship between the length and the frequency of a phrase: all combinations are possible. So, if you know the frequency of a phrase, you cannot make any useful prediction about its length.*

	frequency=high	frequency=medium	frequency=low
length=short	50	10	2
length=medium	20	60	4
length=long	3	3	40

Table 2: *The distribution of two independent variables that are strongly correlated. Each box counts the number of phrases in a corpus with that frequency and that length. You can see that high frequency phrases and short phrases go together: there are many such examples. But there are few examples of long, high-frequency phrases. So, if you know the frequency of a phrase, you can make a pretty good guess about its length.*

a significant positive effect, when (in reality) one variable might have no effect, or might even have a negative effect.

3 Many Small Tests and Fishing Expeditions

The other problem with using data more than once is that you need to use Bonferroni corrections, to avoid fooling yourself and your readers.[11]

An essential feature of a statistical hypothesis test is that there is a chance of getting a significant result, a chance of rejecting H_0, even if the null hypothesis is actually perfectly correct. This chance is the test's *significance level*. A false rejection might happen (strictly by chance) if all the female subjects of your experiment happened to have been raised by nannies who spoke a Bantu language. Then, you may be surprised to find that a group that appear to be perfectly normal Englishwomen are all far better at clicking and discriminating between clicks than their male counterparts. You would (honestly, but incorrectly) conclude that women are different, even though the true explanation is just the luck of the draw. You would later be surprised again, unpleasantly, when someone eventually proves that you got the wrong answer. There is no perfect defense against this kind of surprise, but a well-planned experiment may help. And, ultimately the best strategy is to do the statistics right and work at a relatively strict significance level (like 1% or 0.1% instead of 5%).

Getting false significances is much more likely if you do repeated tests with the same data and you don't make Bonferroni corrections. For instance, if you do two tests at a 1% significance level, there is a 2% chance that one will be falsely positive. Often, though, many tests are done, not just two. In linguistics, it is common for many factors to be plausibly connected to a result. Let's look at an example:

> Research Question: Do people prefer the sentence "I thought X was a deluxe Y" versus "I thought X was an opulent Y"?

The answer could reasonably depend on all kinds of details about X and Y.

- Is X a pronoun or other kind of noun phrase? (E.g. "I thought he was an opulent bastard.")
- Is X a animate or non-animate object? (E.g. "I thought an iPhone was a deluxe iPod.")
- Is Y a simple noun or a noun phrase? (E.g. "I thought his pie was a deluxe pile of goo.")
- ...

[11] The point of doing statistics is to avoid fooling yourself when you find an interesting pattern in your data. You do statistics when you want to know if the pattern is likely to appear again if you repeated your experiment.

With a bit of linguistic imagination, you could probably think up a few other syntactic or semantic properties that might reasonably make one form preferred relative to the other. So, let's assume that there are six binary features that might affect the preference via X and six via Y.

- If we test these 12 properties one at a time at a 5% confidence level, we end up with a 46% chance of a false positive from one of the 12 tests.
- Or, if we split the properties into one group of six for X and another group for Y, we would get 36 two-way combinations of properties. If we test all those at a 5% significance level we would expect that at least one would turn up significant in the statistical tests, even if there are really no effects at all.
- Or, we could realize that there are $2^{12} = 4096$ twelve-way combinations of factors and test each combination separately. Then, even at a 1% significance level, we would expect 41 false positives (give or take a few).

There is actually nothing wrong with this as long as it is reported clearly so that your readers understand that there is a larger-than-usual chance of false positives. But, that means in any of the three examples above, there is no reason to get excited about a single statistically significant result. If you do lots of tests, no single result is important any more, because you expect some significances to happen accidentally.

The Bonferroni correction is there to prevent this dilution of significance. It is a very simple procedure: you just divide your target significance level by a count of how many tests that you are going to do; then you do the individual tests at this smaller significance level. Then, the overall chance of getting a false significance stays at 1% (or 5% or 0.1%, whatever you choose) no matter how many tests you run.