18.2: Statistical Models Missing from the Book

Statistics is a huge field. The core tools that I've described in this book (chi-square tests, t-tests, ANOVA and regression) are basic tools that are widely used in everyday data analysis, and they form the core of most introductory stats books. However, there are a ***lot*** of other tools out there. There are so very many data analysis situations that these tools don't cover, and in future versions of this book I want to talk about them. To give you a sense of just how much more there is, and how much more work I want to do to finish this thing, the following is a list of statistical modelling tools that I would have liked to talk about. Some of these will definitely make it into future versions of the book.

- Analysis of covariance In Chapter 16 I spent a bit of time discussing the connection between ANOVA and regression, pointing out that any ANOVA model can be recast as a kind of regression model. More generally, both are examples of *linear models*, and it's quite possible to consider linear models that are more general than either. The classic example of this is "analysis of covariance" (ANCOVA), and it refers to the situation where some of your predictors are continuous (like in a regression model) and others are categorical (like in an ANOVA).
- **Nonlinear regression** When discussing regression in Chapter 15, we saw that regression assume that the relationship between predictors and outcomes is linear. One the other hand, when we talked about the simpler problem of correlation in Chapter 5, we saw that there exist tools (e.g., Spearman correlations) that are able to assess non-linear relationships between variables. There are a number of tools in statistics that can be used to do non-linear regression. For instance, some non-linear regression models assume that the relationship between predictors and outcomes is monotonic (e.g., isotonic regression), while others assume that it is smooth but not necessarily monotonic (e.g., Lowess regression), while others assume that the relationship is of a known form that happens to be nonlinear (e.g., polynomial regression).
- **Logistic regression** Yet another variation on regression occurs when the outcome variable is binary valued, but the predictors are continuous. For instance, suppose you're investigating social media, and you want to know if it's possible to predict whether or not someone is on Twitter as a function of their income, their age, and a range of other variables. This is basically a regression model, but you can't use regular linear regression because the outcome variable is binary (you're either on Twitter or you're not): because the outcome variable is binary, there's no way that the residuals could possibly be normally distributed. There are a number of tools that statisticians can apply to this situation, the most prominent of which is logistic regression.
- **The General Linear Model (GLM)** The GLM is actually a family of models that includes logistic regression, linear regression, (some) nonlinear regression, ANOVA and many others. The basic idea in the GLM is essentially the same idea that underpins linear models, but it allows for the idea that your data might not be normally distributed, and allows for nonlinear relationships between predictors and outcomes. There are a lot of very handy analyses that you can run that fall within the GLM, so it's a very useful thing to know about.
- **Survival analysis** In Chapter 2 I talked about "differential attrition", the tendency for people to leave the study in a non-random fashion. Back then, I was talking about it as a potential methodological concern, but there are a lot of situations in which differential attrition is actually the thing you're interested in. Suppose, for instance, you're interested in finding out how long people play different kinds of computer games in a single session. Do people tend to play RTS (real time strategy) games for longer stretches than FPS (first person shooter) games? You might design your study like this. People come into the lab, and they can play for as long or as little as they like. Once they're finished, you record the time they spent playing. However, due to ethical restrictions, let's suppose that you cannot allow them to keep playing longer than two hours. A lot of people will stop playing before the two hour limit, so you know exactly how long they played. But some people will run into the two hour limit, and so you don't know how long they would have kept playing if you'd been able to continue the study. As a consequence, your data are systematically *censored*: you're missing all of the very long times. How do you analyse this data sensibly? This is the problem that survival analysis solves. It is specifically designed to handle this situation, where you're systematically missing one "side" of the data because the study ended. It's very widely used in health research, and in that context it is often literally used to analyse survival. For instance, you may be tracking people with a particular type of cancer, some who have received treatment A and others who have received treatment B, but you only have funding to track them for 5 years. At the end of the study period some people are alive, others are not. In this context, survival analysis is useful for determining which treatment is more effective, and telling you about the risk of death that people face over time.
- **Repeated measures ANOVA** When talking about reshaping data in Chapter 7, I introduced some data sets in which each participant was measured in multiple conditions (e.g., in the drugs data set, the working memory capacity (WMC) of each person was measured under the influence of alcohol and caffeine). It is quite common to design studies that have this kind of repeated measures structure. A regular ANOVA doesn't make sense for these studies, because the repeated measurements mean that independence is violated (i.e., observations from the same participant are more closely related to one another than to

observations from other participants. Repeated measures ANOVA is a tool that can be applied to data that have this structure. The basic idea behind RM-ANOVA is to take into account the fact that participants can have different overall levels of performance. For instance, Amy might have a WMC of 7 normally, which falls to 5 under the influence of caffeine, whereas Borat might have a WMC of 6 normally, which falls to 4 under the influence of caffeine. Because this is a repeated measures design, we recognise that – although Amy has a higher WMC than Borat – the effect of caffeine is identical for these two people. In other words, a repeated measures design means that we can attribute some of the variation in our WMC measurement to individual differences (i.e., some of it is just that Amy has higher WMC than Borat), which allows us to draw stronger conclusions about the effect of caffeine.

- **Mixed models** Repeated measures ANOVA is used in situations where you have observations clustered within experimental units. In the example I gave above, we have multiple WMC measures for each participant (i.e., one for each condition). However, there are a lot of other ways in which you can end up with multiple observations per participant, and for most of those situations the repeated measures ANOVA framework is insufficient. A good example of this is when you track individual people across multiple time points. Let's say you're tracking happiness over time, for two people. Aaron's happiness starts at 10, then drops to 8, and then to 6. Belinda's happiness starts at 6, then rises to 8 and then to 10. Both of these two people have the same "overall" level of happiness (the average across the three time points is 8), so a repeated measures ANOVA analysis would treat Aaron and Belinda the same way. But that's clearly wrong. Aaron's happiness is decreasing, whereas Belinda's is increasing. If you want to optimally analyse data from an experiment where people can change over time, then you need a more powerful tool than repeated measures ANOVA. The tools that people use to solve this problem are called "mixed" models, because they are designed to learn about individual experimental units (e.g. happiness of individual people over time) as well as overall effects (e.g. the effect of money on happiness over time). Repeated measures ANOVA is perhaps the simplest example of a mixed model, but there's a lot you can do with mixed models that you can't do with repeated measures ANOVA.
- **Reliability analysis** Back in Chapter 2 I talked about reliability as one of the desirable characteristics of a measurement. One of the different types of reliability I mentioned was inter-item reliability. For example, when designing a survey used to measure some aspect to someone's personality (e.g., extraversion), one generally attempts to include several different questions that all ask the same basic question in lots of different ways. When you do this, you tend to expect that all of these questions will tend to be correlated with one another, because they're all measuring the same latent construct. There are a number of tools (e.g., Cronbach's α) that you can use to check whether this is actually true for your study.
- **Factor analysis** One big shortcoming with reliability measures like Cronbach's α is that they assume that your observed variables are all measuring a *single* latent construct. But that's not true in general. If you look at most personality questionnaires, or IQ tests, or almost anything where you're taking lots of measurements, it's probably the case that you're actually measuring several things at once. For example, all the different tests used when measuring IQ do tend to correlate with one another, but the pattern of correlations that you see across tests suggests that there are multiple different "things" going on in the data. Factor analysis (and related tools like principal components analysis and independent components analsysis) is a tool that you can use to help you figure out what these things are. Broadly speaking, what you do with these tools is take a big correlation matrix that describes all pairwise correlations between your variables, and attempt to express this pattern of correlations using only a small number of latent variables. Factor analysis is a very useful tool – it's a great way of trying to see how your variables are related to one another – but it can be tricky to use well. A lot of people make the mistake of thinking that when factor analysis uncovers a latent variable (e.g., extraversion pops out as a latent variable when you factor analyse most personality questionnaires), it must actually correspond to a real "thing". That's not necessarily true. Even so, factor analysis is a very useful thing to know about (especially for psychologists), and I do want to talk about it in a later version of the book.
- **Multidimensional scaling** Factor analysis is an example of an "unsupervised learning" model. What this means is that, unlike most of the "supervised learning" tools I've mentioned, you can't divide up your variables in to predictors and outcomes. Regression is supervised learning; factor analysis is unsupervised learning. It's not the only type of unsupervised learning model however. For example, in factor analysis one is concerned with the analysis of correlations between variables. However, there are many situations where you're actually interested in analysing similarities or dissimilarities between objects, items or people. There are a number of tools that you can use in this situation, the best known of which is multidimensional scaling (MDS). In MDS, the idea is to find a "geometric" representation of your items. Each item is "plotted" as a point in some space, and the distance between two points is a measure of how dissimilar those items are.
- **Clustering** Another example of an unsupervised learning model is clustering (also referred to as classification), in which you want to organise all of your items into meaningful groups, such that similar items are assigned to the same groups. A lot of clustering is unsupervised, meaning that you don't know anything about what the groups are, you just have to guess. There are other "supervised clustering" situations where you need to predict group memberships on the basis of other variables, and those

group memberships are actually observables: logistic regression is a good example of a tool that works this way. However, when you don't actually know the group memberships, you have to use different tools (e.g., k-means clustering). There's even situations where you want to do something called "semi-supervised clustering", in which you know the group memberships for some items but not others. As you can probably guess, clustering is a pretty big topic, and a pretty useful thing to know about.

- **Causal models** One thing that I haven't talked about much in this book is how you can use statistical modeling to learn about the causal relationships between variables. For instance, consider the following three variables which might be of interest when thinking about how someone died in a firing squad. We might want to measure whether or not an execution order was given (variable A), whether or not a marksman fired their gun (variable B), and whether or not the person got hit with a bullet (variable C). These three variables are all correlated with one another (e.g., there is a correlation between guns being fired and people getting hit with bullets), but we actually want to make stronger statements about them than merely talking about correlations. We want to talk about causation. We want to be able to say that the execution order (A) causes the marksman to fire (B) which causes someone to get shot (C). We can express this by a directed arrow notation: we write it as $A \rightarrow B \rightarrow C$. This "causal chain" is a fundamentally different explanation for events than one in which the marksman fires first, which causes the shooting $B \rightarrow C$, and then causes the executioner to "retroactively" issue the execution order, $B \rightarrow A$. This "common effect" model says that A and C are both caused by B. You can see why these are different. In the first causal model, if we had managed to stop the executioner from issuing the order (intervening to change A), then no shooting would have happened. In the second model, the shooting would have happened any way because the marksman was *not* following the execution order. There is a big literature in statistics on trying to understand the causal relationships between variables, and a number of different tools exist to help you test different causal stories about your data. The most widely used of these tools (in psychology at least) is structural equations modelling (SEM), and at some point I'd like to extend the book to talk about it.

Of course, even this listing is incomplete. I haven't mentioned time series analysis, item response theory, market basket analysis, classification and regression trees, or any of a huge range of other topics. However, the list that I've given above is essentially my wish list for this book. Sure, it would double the length of the book, but it would mean that the scope has become broad enough to cover most things that applied researchers in psychology would need to use.
