Fitting Models is Like Tetris: Crash Course Statistics #35

Crash Course: Statistics
https://youtube.com/watch?v=QdmG6uUtB4o
https://nerdfighteria.info/v/QdmG6uUtB4o

Hi, I'm Adriene Hill, and welcome back to Crash Course Statistics.

General Linear Models -- like Regression and ANOVA -- let us create a statistical analysis of data for our specific needs. Fitting the right model to our experiments is kind of like Tetris… GLMS are in this analogy tetriminos.

Sometimes you need the skinny-long brick, called the straight sometimes you need the square sometimes you need the left snake. In stats, its similar sometimes you need regression sometimes ANOVA but there's also ANCOVA ---The Analysis of Covariance. And the Repeated Measures ANOVA.

Today we'll look at the shape of those models. And how they might help us level-up! INTRO As a quick review, in a few of our past episodes we covered the fact that ANOVAs and regressions are both General Linear Models.

ANOVAs allow us to analyze the effect of variables with two or more groups on continuous variables. And regressions allow us to analyze two continuous variables. General Linear Models explain the data we observe by building a model to predict that data, and then keeping track of how close the prediction is.

And both regressions and ANOVAs use a similar model setup. It looks just like the equation for a line that you may have seen if you've taken Algebra. The fact that they're set up the exact same way is helpful for two reasons.

One, it means we only have to remember one general mode , and two it allows us to combine these two powerful models to give us the even more flexible ANCOVA. For example, we might want to look at the amount of general anesthesia needed to put a patient under. There have been studies that suggest that redheads require more anesthesia than non-redheads because the gene mutation that causes red hair, also affects pain receptors.

So we have two groups: redheads and non-redheads. Those are categorical variables. But, we also think that weight will have a meaningful impact on the amount of this specific anesthetic that's needed for surgery.

Weight is a continuous variable. To make sure things are relatively equal, we look at only one kind of simple, routine surgery: appendix removal. Working with a hospital, we collect data on 100 randomly selected patients. 50 redheads, and 50 non-redheads.

We record their weight, natural hair color, and the amount of anesthesia needed during their surgery. We can now build a model to predict milliliters of anesthesia based on hair color and weight. Just like its friends, regression and ANOVA, the ANCOVA looks at the overall variation in the data, and uses different variables, like hair color and weight, to explain it.

The overall variation is, as always, measured by the sum of the squared distances between the overall mean amount of anesthesia used, and each dose of anesthesia that was administered. This variation is called the Sums of Squares total. So now we can calculate an ANOVA table that shows us the sums of squares and F-tests for each of our effects.

Even though this is an ANCOVA model, we still usually refer to these as ANOVA tables. And even though this table has both continuous regression factors and categorical ANOVA factors, we read it just like it's a regular ANOVA table. Here we can see that weight is a significant predictor of how much anesthesia you'll need, but hair color isn't .it's really tempting to call hair color "nearly significant" because it's SO close to 0.05.

But our cutoff is strict. It has to be less than 0.05. We now have a tool that allows us to combine categorical and continuous variables into one General Linear Model.

The world as they say is our oyster. We can predict all kinds of things with all kinds of variables. We can also use our new ANCOVA models to make stronger inferences.

In our example,we were interested, mainly, in whether being a redhead significantly increased the dose of a new anesthetic. But we also included weight in the model, since we knew that weight plays a pretty big role in how much anesthetic you need. Weight accounted for a lot of the variation in the model.

Its eta squared is 0.353, which means that it accounts for about 35% of the variation in our data. That's pretty high. And since it "soaked up" all of that variation, our Sums of Squares Error is now smaller.

If we had run a simple ANOVA with JUST hair color, the differences between anesthetic doses due to weight would have just been chalked up to "random variation", or error because it's source--weight--wasn't in our model. For both of these models, the simple case where we ONLY look at hair color, and the more complex case where we look at both hair color and weight, the total variation in the data is the same. Because it's the same data.

Total variation looks only at our outcome variable--like milliliters of anesthetic. So, when we build our models, we're partitioning the same amount of variation into groups. Our simple ANOVA model JUST looks at how much of this total variation is due to being or not being a redhead.

The rest is counted as error, just because "error" refers to variation that our model doesn't account for. When we use the bigger model that includes both hair color and weight, we take some of that variation that was attributed to error, and attribute it to weight instead. This makes our pile of error variation smaller.

For this reason, many researchers will add covariates--continuous variables that are used to explain our outcome variable--not only for inference, but also to reduce the amount of error variation. Let's take another example. Say we want to look at the effect of a new brand of formula on the weight of infants.

We have two randomly assigned groups of infants: those with our new formula and those who get an established brand of formula. But infants grow very quickly, so we want to account for any variation due to age, so we include age in days in our model. If we just ran a model that included formula type, our Sums of Squares for Error is pretty big.

And formula doesn't have a significant effect on infants' weight. But we know that infants weights are strongly correlated with how old they are, so when we include that in a new ANCOVA model, it takes some of the variation that was error variation in our simple model, and accounts for it using age in days. As you can see from this ANOVA table, adding age as a covariate allowed us to explain some of the variation, while making it easier for us to detect the fact that there is actually a significant effect of formula type on babys' weights.

And we're not limited to just one covariate. We can add many, if we want. We could add mother's weight to this ANCOVA, or even another categorical variable, like ethnicity.

Our models are limited only by our ability to collect data. But we have to be careful when we're using covariates to do inference. There are cases when it makes sense to have a bunch of

covariates.

But if someone is adding a bunch of them just to make their p-values significant, that could be considered p-hacking And we can continue to customize our model even further so that we're partitioning our variation more accurately. Previously, we noted that it's difficult to do a statistical test on whether there was a significant difference between the mean ratings of two coffee shops. That's because people's individual coffee preferences add extra variation to our data.

People who hate coffee will always rate it relatively low, and people who love coffee will always rate it pretty high. In that simple case, we did a matched pairs t-test in order to "subtract" the variation due to people's different levels of coffee affinity. Essentially, what we were doing was allowing each person to have their own "baseline" coffee preference.

This allowed us to see whether there was a pattern of one coffee shop getting higher ratings than the other, regardless of whether the people who rated it loved, tolerated, or hated coffee. And we can do that with more than 2 groups as well, using something called a Repeated Measures ANOVA. A Repeated Measures ANOVA asks whether there's a significant difference between 2 or more groups or conditions.

The key to an Repeated Measures ANOVA is that the same experimental unit, whether it's a cell, a person, or an animal, is measured multiple times. Hence "Repeated". And in practice, it works pretty similarly to the match pairs t-test, except it allows you to look at more than 2 groups.

A repeated measures ANOVA lets each experimental unit have its own "baseline". So we could ask whether there's a significant difference between 10 different coffee shops, or whether there's a significant effect of slow, medium, and fast tempoed music on the speed we run. Everyone has a different baseline running speed.

Maybe your friend who injured their knee runs pretty slowly, but your cousin can run a 6 minute mile. But it's still possible to say that on average, people run faster when a bear is chasing them--whether they're fast or slow. We're looking at data from 150 people, and we record how fast they can run a mile listening to slow, medium and fast tempoed songs.

We measure them on different days so that they don't get too tired after all that running (that could affect our data). And we make sure to randomize the order of the music so that not everyone gets slow first, or medium last. If we simple looked at an ANOVA that used music tempo to predict mile pace there's a lot of variation.

And when we ran this simple model, the effect of music tempo is non significant. That may be due in part to the fact that the difference between how fast individual people normally run is counted in the Error Sums of Squares, making it a lot bigger. (That might not be the only reason, though.) So, we tell our model which measurements belong to the same person. And then, we tell our model to let each individual person have their own baseline mile time, and we'll just look at how much music tempo affects the changes from people's baseline running speeds.

So whether you normally run a 5 or 15 minute mile, an increase in 1 minute will be counted the same. Theoretically, it's sorta like centering everyone on their own mean running speed. If you normally run a 6 minute mile, that becomes your 0 baseline.

Same thing if you normally run a 12 minute mile. Since the math of these models--sometimes called Random Effect Models--can get a little intense, we're just going to focus on how to read the ANOVA table output from a Repeated Measures ANOVA. Here, our output shows us that there is actually a significant effect of the music tempo on running time.

Because we allowed everyone to have their own "baseline" speed, we in essence took that variation away, and made our error term smaller. We now have the shapes we need to fit all kinds of situations… We can combine categorical and continuous factors, and we know how to handle data where the same subject is measured multiple times. We can slide these pieces together in all sorts of ways.

We can build a model that looks at how the number of hours of Tetris we play affects how far we go in each game and if expertise level effects how long someone plays. Or we could add statistical rigour to the decade long arguments over which Tetris shapes are the best (it's the straight) and the worst to get. Thanks for watching, I'll see you next time.