PRESENTATION SCRIPT

I'm Jesse. I'm working with Dr. Andrew Holbrook from UCLA biostat, Dr. Nick Tustison from UVA Radiology, as part of a wider network of neuroscientists, radiologists, statisticians and other experts for the Alzheimer's Disease Neuroimaging Initiative.

Let's jump right into the scientific question. The human cerebral cortex is a thin layer of tissue, less than a centimeter thick, wrinkled and folded to fit inside the skull. Each hemisphere contains a region called the entorhinal cortex (shown in red), which plays a role in spatial memory formation. In patients with Alzheimer's, the entire cortex loses mass and literally gets thinner. Neuroscientists believe that the ERC is the first region affected.

Our data set comes from the Alzheimer's Disease Neuroimaging Initiative. It contains 2515 longitudinal observations on 663 unique subjects -- up to 6 per subject.

At the first visit, researchers gave each subject a diagnosis of cognitively normal, mild cognitive impairment, or Alzheimer's Disease, based on his/her performance on a cognitive test. At each visit, researchers conducted an MRI scan of the subject's brain. From the raw images, they used computational algorithms to measure the thickness of 62 different cortical regions, including the entorhinal cortex, which we're looking at.

Our biological question is how cognitive diagnosis relates to ERC thickness. Later we'll see how this gives rise to some questions about imaging, and some questions about statistical methodology.

I haven't been working on this for very long, and I haven't really accomplished anything yet, but what has been most fun for me so far is the process of building up a more and more complex statistical model. I got to be the little boy in the picture, putting the airplane together, and later on, playing with some power tools. A lot of these techniques were actually brand new to me, even if they're standard in the industry. If they're boringly familiar to you, I hope you'll humor me. So I'd like to take you through the process so far.

Here's our starting point, which I'll call Model 1. We'll take a minute to learn what the variables stand for. My preference is to name rather than to number them when possible. You'll see that this a basic linear model with intercept, slope, and random error.

The outcome y\_ij is the ERC thickness for the i^th subject at the j^th time point. T\_ij represents how many months have gone by for the j^th time point for the i^th subject.

CN, or cognitively normal, is the reference group. MCI and AD are indicators for membership in the "mild cognitive impairment" or "Alzheimer's disease" group. Remember, every subject is in exactly one of these three groups.

Beta\_CN is the population intercept, i.e. the mean outcome at the beginning of the study, for the CN group. Beta\_MCI and Beta\_AD are the differences in mean for these groups. So each group has its own intercept.

Beta\_CN\_t is the slope, or change in outcome per month, for the CN group. Beta\_MCI\_t and Beta\_AD\_t are the differences in slope for these groups.

What we have then, is a three different lines, each with its own slope and intercept, for the three different groups.

Finally, epsilon\_ij is the error term. The interpretation that makes the most sense to me is that the error term epsilon\_ij captures all the influences on y\_ij that aren't accounted for by the other variables in the model. It's the "everything else" term. In the Bayesian interpretation, which I like, the definition of random is unknown, so the outcome is the sum of knowns and unknowns. We always assume the mean of the error term is 0 over all the observations, because if it were anything else, we could just change the intercept by that amount. The Central Limit Theorem tells us that the sum of lots of small independent deviations has, roughly, a normal distribution.

Although this first model is not completely naïve, we can improve it. For one thing, it assumes that, not only over all subjects, but even for a given subject i, the average of the epsilon\_ij’s is 0. This usually doesn't fit the data. For a given subject, the differences from the group mean tend to be systematic -- above the mean for all time points, or below. Put another way, the within-subject errors are correlated.

We can account for this by adding "random," i.e. subject-specific, effects. This is what I've done in Model 2. (I'm going to use the convention of showing the changes in green, so you don't have to reread everything.)

Here, gamma\_0i represents the difference in subject i's intercept from that predicted by the covariates. It's random, in the Bayesian sense, because we don't know, in real world terms, what accounts for it.

I just learned about random intercepts. Everybody else here has probably seen them before. But I think it's amazing what this one addition to the model accomplishes.

First, we've decomposed the intercept for the i^th subject into the sum of two components, the overall group intercept and the individual deviation. We can get a better estimate of the group intercepts--our real interest--if they're not distorted by the individual deviations.

Second, we've also decomposed the variance of y\_ij to the variance in gamma\_0i plus the variance of epsilon\_ij. Remember, epsilon represents "everything else" that influences y\_ij. But that no longer includes subject specific effects, so it more closely represents something like measurement error, which we are also probably interested in.

But what about subject specific characteristics that affect change over time, the progression of the disease? This would cause the subject to differ from the mean by an increasing or decreasing amount. So we also add a random slope. Let gamma\_1i be the difference in subject i's slope from that predicted by the covariates.

The random intercept and slope induce a certain covariance pattern on y\_i, even without a covariance pattern in epsilon\_i. We could also allow for covariance \*between\* the random intercept and slope. For our data, the best estimates for this was zero, so we left it out to simplify the model.

My concern at this point was this. The random effects are doing a lot of work. But random effects for subject i are estimated only from subject i, who has at most 6, and as few as 2, observations. This seems too small to get a good estimate.

As you know, if your sample size is very small, and your observation happens to be unrepresentative of the population, your estimate will be very wrong. So what we're going to do is add hierarchy to the model. This is a way of estimating parameters on one subject by borrowing information from other subjects. It may be familiar, but I think that it's brilliant.

To people who don't understand how this works -- not anybody here -- I like to explain it with an extreme example. Suppose you were studying the mean household income in each of the fifty nifty United States, and you only took one observation from each state. And suppose that for most states, your observation was between 50k and 80k, but the household you surveyed in Nebraska answered $500k. What is the "best" estimate of the Nebraska mean household income? Strictly speaking, the maximum likelihood estimate, the best linear unbiased estimate, etc. are all equal to the sample mean, which is equal to that one data point, $500k.

There is nothing illogical about that estimate. But if it seems implausible to you, it's because you know more than just that one data point. You know that Nebraska is one of the United States, and that mean income between states differs but not by THAT much. How do we incorporate this knowledge?

Model each state mean mu\_Alabama, mu\_Alaska, ... mu\_Wyoming as a draw from a distribution of state means N(mu\_USA, sigma2\_USA). Now estimate all of these parameters \*simultaneously\* from the data y\_Alabama, ... y\_Wyoming. The ML estimate for mu\_Nebraska will be less extreme than the observation y\_Nebraska. Why? An extreme observation from a not-so-extreme state mean is more likely than a representative observation from an extreme state mean.

We say that the estimate for mu\_Nebraska will be "shrunk" toward mu\_USA.

So here's Model 4. We borrow information from all 663 subjects to estimate each subject's random effects. Scientifically, we are not really interested in estimates of individual effects, but we might be interested in hyperparameters like sigma2\_gamma\_0, which captures how much the intercepts differ from the mean of their group.

What if the measurement error, or whatever is captured by epsilon, differs by subject? It's also possible to put hierarchy on the epsilons, by making each sigma\_i^2 a draw from an underlying distribution. We have to weigh the benefits against the computational cost. The model already has 1335 parameters. Sometimes we just try adding something like this and see if it makes a difference or not.

Other than making the model more complex, how else can we improve our parameter estimates?

So far there is nothing Bayesian about the model. But now we can use informative priors to incorporate scientific knowledge. There is lots of data already about the thickness of different cortical regions, at different ages, with and without Alzheimer's disease. Part of my job is finding it and translating it into priors.

We can also use "regularizing" priors to keep our estimates conservative. For estimates of beta, this is analogous to lasso or ridge regression. We probably don't need this technique for this model, since we don't have many covariates and variable selection is not the main challenge we're facing.

So now we've added random effects, hierarchy, and priors. There's one more feature we're going to add to the model, which I'll call robustness.

You'll notice that every probability distribution in the model so far is Normal. It would be "a pity" not to consider other distributions.

Think about this. The density of the standard normal distribution at 0 is about .40, and at 3 about .004. So a z-score of 3 is roughly 100 times less likely than a z-score of 0. If we use a Normal model for the likelihood, we are encoding a belief that observations more than a couple sd's from the mean are extremely rare. So outliers tend to pull the mean away from the center of the data and inflate the variance. I'll show you a picture on the next slide.

However, while the Normal has only two parameters, the extended Student t has three: mean mu, scale sigma, and kurtosis nu. If nu is small, we have fat tails; if large, maybe 30 or more, the curve is nearly Normal. So using a t reflects "open mindedness" about outliers. Consider: for a t(1df) distribution, a z-score of 3 is only 10, not 100, times less likely than a z-score of 0.

To show you what I mean, look at this toy example. Under the Normal assumption, it's more likely that all observations are kind of far from the mean than that one is really far. And the sd has to be large, so that no observation is very many sd's from the mean. But the t has the flexibility to consider that y=15 as an outlier, and it affects the kurtosis rather than the mean or scale. I would consider that a better fit, but if you don't, we can agree to disagree.

Here's model 6. All we’ve done is change the error distribution to t, which is equivalent to changing the likelihood to t. When we ran the model, our estimates for nu were low, indicating that t is a better fit.

This is now a pretty sophisticated model, but there’s one more big challenge before launch. We’re going to make one more revision today. Hereafter, write y\_ij = x\_ij T beta + z\_ij T gamma + epsilon\_ij.

Here’s the challenge. We don’t just have one data set. We have seven data sets from what we’re calling different “pipelines.” This is their term. Each represents a different computational algorithm applied to the same raw MRIs. Some of the pipelines were developed by this research team and associates, and if they are shown to be more accurate than the others, that will be an important scientific advance. So that’s a second important scientific question.

The problem is, the measurements differ substantially, and there’s no objective gold standard by which to compare them. There is a truth of the matter about these human beings’ cortical thicknesses at these particular points in time, but we’ll never know what it is. So there’s an open methodological problem as well: how can we determine the relative accuracy of the pipelines? how should we weight them to achieve the best parameter estimates? I think those questions will be answered together.

We have a first step, and it’s pretty clever. We’re going to model the true thickness as a latent variable. Sorry about the change in notation, but it’s unavoidable. Suppose the true, unobserved cortical thickness w\_ij are draws from this normal distribution.

Our observations, the y\_ijk’s, are modeled as the true value plus a random measurement error. Each pipeline has a different t distribution of measurement errors. We estimate all three parameters, so each pipeline has a systematic over/underestimation lambda\_k, and a spread described by tau\_k and nu\_k. A nice feature is that the w\_ij’s -- which again are the true values -- are estimated using *all* the data!

This is a brilliant move, and also hard to understand, so I’m going to let it sit a for a second.

We’ve spent all this time constructing a model to give us the best possible estimates, so I should probably show you some.

The posterior distributions show significant, and significantly different, effects of cognitive category on intercept (starting thickness) and slope (rate of thinning).

T

[]

[]

[]