9.S918: Statistical Inference for Brain and Cognitive Sciences, Pset 4
due 7 May 2025

28 April 2025

1 Held-out testing for multi-level data

Suppose you have a multi-level linear model M of the form (in R formula style):

```
Y ~ X + (X | Group)
```

and you have M groups, each of which has N observations. You are interested in inferences about the effect of `X` on `Y`, so plan on comparing the fit of the above model to the special case of the model nested inside it where the fixed-effect slope for `X` is 0, which we will call M_0.

Here are two strategies for held-out (train/test) evaluation of M versus the M_0 baseline, both of which involves using $(M-1)N$ examples for training and holding out N examples for testing:

1. Sample N examples uniformly at random from the MN total examples, call those your test set, and use the remaining $(M-1)N$ examples as your training set;

2. Hold out *all* N *examples from one group* as your test set set, and use all the examples from the other $M-1$ groups as your training set.

What is the difference between these two strategies in terms of the model comparison? Which strategy more closely corresponds to inferences that you would make using the point estimate and the standard error (or the posterior distribution in a Bayesian setting), of the coefficient for `X` within M, or a nested model comparison between M and M_0, for a more traditional approach where you use all MN observations to fit the data without a train/test split? Make a conceptual case, and demonstrate your point using a simulation. (Potentially useful hint: for a fitted `lme4` model, you can use the `predict()` function with the `re.form = NA` argument to make predictions on observations in a previously unseen group, using the fixed effects estimate only.)

2 Anti-conservativity of OLS regression

Modify the code provided in the first simulation exercise for Monday, April 28 to show that OLS linear regression (i.e., not mixed-effects) on data generated by the linear mixed model specified in the exercise is anti-conservative. For a nominal α-level of 0.05, what is the actual Type I error rate? What is the relationship between that anti-conservativity and the difference in the distribution of standard errors that we observed in class?

3 Small numbers of clusters: treat as random effects, fixed effects, or don't model at all?

Confusion often arises regarding when to treat clusters (that is, groupings of data reflecting repeated observations from a given source, such as multiple data points from the same participant or same stimulus) as "fixed effects" versus "random effects" in hierarchical/multi-level regression modeling, and what is at stake. Sometimes one might hear, for example, that one needs a relatively large number of clusters in order to treat them as a random effect, so that you can effectively estimate the distribution of that random effect. In this example you'll investigate the wisdom of that advice.

Consider a case where you are interested in inferring the relationship between a predictor x and response y using a dataset organized into four clusters of 20 observations each. For example, x might be the number of words in a word list presented to an experimental participant, ranging from 1 to 20, the clusters are four experimental participants, and y is the fMRI BOLD response in a particular brain region of interest. Assume that the number of words in the list varies from 1 to 20, that the design is perfectly balanced, so that each participant gets exactly one list of each length, and that the particular set of words in each list differs from participant to participant (so that there are no by-stimulus repeated measures). Assume a linear relationship between x and y, with by-participant random intercepts and random slopes drawn from $\mathcal{N}\left(0, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$, and trial-level residual error variance of 1. Simulate this dataset and then consider two possible regression analyses:

1. Don't include by-participant effects at all, since there are so few clusters.

2. Treat participants as a **fixed effect**, estimating the main effect of x in the presence of an interaction with participant (as well as a main effect of participant). Make sure you use the appropriate contrast coding for representing participant effects.

3. Treat participant as a **random effect**, using a multi-level model with maximal random-effects specification and fitting the model using maximum likelihood.

Are the point estimates of the effect of x the same or different among the three analyses? What about the standard errors of those estimates? Using Monte Carlo, estimate the Type I error rate of the three approaches when the null is true ($\beta_x = 0$). Interpret your results.