Multilevel probit regression for 3-alternative forced choice audibility testing

Andrew Christian[(1)], Matthew Boucher[(1)], Menachem Rafaelof[(2)], Kevin P. Shepherd[(1)], Stephen A. Rizzi[(1)] and James H. Stephenson[(3)]

(1) NASA Langley Research Center, (2) National Institute of Aerospace

(3) US Army Combat Capabilities Development Center Aviation & Missile Center

Abstract:

A recent psychoacoustic test at NASA Langley generated a dataset of 3-alternative forced-choice (3AFC) responses for 40 subjects that measured the audibility of a tone complex in a shaped broadband masker. The task was completed by 4 subjects at a time in a small theater-like environment using predetermined stimuli levels. These data were subject to 4 forms of probit regression: a "complete pooling" analysis in which all data from the test was fit with one curve, two forms of "no pooling" analyses in which subjects' data were treated individually (using both packaged and custom software), and a "partial pooling" analysis in which multilevel-regression software fit both individual curves as well as population-level parameters at the same time. The results of the analyses are compared in terms of both individual- and population-level parameters. Partial pooling appears to give the most consistent results at both levels, as well as provide the most robustness among the packaged approaches (albeit with added complexity over single-level regression). These results mirror those contained in a recent NASA technical memorandum entitled "Comparisons of Analysis Methods Applied to Alternative Forced Choice Audibility Data."

The Problem:

We are interested in modeling the psychometric functions of a population of individuals for a 3AFC audibility task:

1. We want to know the psychometric function of the ***average*** subject.
2. We also want to know how (much) people ***vary*** from this average.

Taking 3AFC data for an audibility task is relatively straightforward. We will take a limited amount of data from each of a large number of subjects (150 trials from each of 40 subjects) so that our results may be representative of a larger population. This leaves us with the question of how to process the data. The approach must have several attributes:

A. The method needs to produce the statistics that we are interested in.

B. The method needs to deal with the scaled psychometric function that is produced by 3AFC testing.

C. The method needs to work for a dataset that has a low number of responses per-subject.

D. The method, hopefully, should be easy to use and understand.

The Experiment:

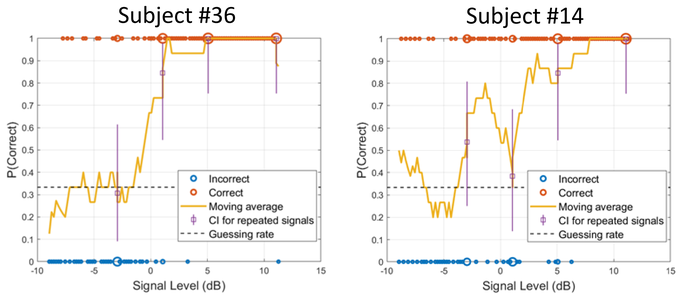
The "ICHIN III" experiment took place in the Exterior Effects Room at NASA Langley Research Center in 2017. The task was for the subjects to try to figure out which interval (out of 3 per-trial) contained a tone complex 'signal' in a shaped-broadband masking 'noise'. The intervals were .5 seconds long. The noise was played throughout the experiment. Subjects were prompted to the timing of the intervals visually via a tablet computer, with which they also registered their response.

The signal levels varied over a 20 dB range, from a clearly audible level to one that was determined to be inaudible during a pilot phase of the test. 106 levels were explored at .2 dB increments. 4 of the levels were repeated a further 11 times, in order to bring the total number of trials up to 150. The trial order was randomized in a way so that there were never long stretches of very-low signal levels.

The subjects were taken from the general population surrounding NASA Langley and were screened for hearing ability. The subjects were tested 4 at a time—hence the lack of ability to have a test design with feedback (such as a staircase).

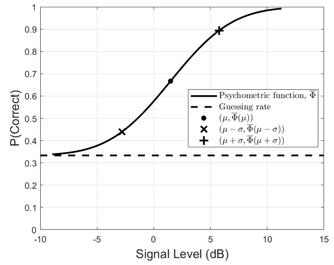
The Data:

The raw results for two subjects are shown on the right. Simple moving averages show the general trends of more correct responses for higher signal level and close to 1/3 correct (3AFC guessing rate) at low levels. It can be seen that the data is not monotone in general, and can sometimes have quite dramatic trend reversals. Therefore it is necessary to model it with a sigmoid.



A probit function was used in this work mainly because its two parameters are easy to interpret and are of the same units as the abscissa (dB). The probit is scaled so that the lower asymptote stops at a value of P(Correct) = 1/3, or the known guessing rate. The formula for this function is shown on the right, in which x is the signal level, μ is the threshold and σ is the spread. The plot on the right shows that a threshold occurs when P(Correct) = 2/3. The spread indicates the change in signal level necessary to increase P(Correct) to 0.89 or decrease it to 0.44.

$$\bar{\Phi} = \frac{1}{3}\left[2 + erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$$



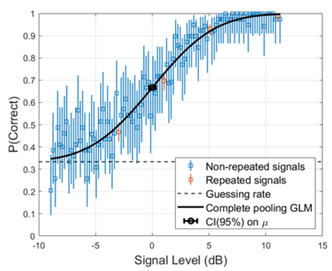
Analysis Approaches:

The scaled probit function can be fit to the data in several ways. These ways differ in how data is aggregated across subjects.

Complete Pooling:

In this approach, the data is effectively treated as if it all came from one (average) subject. One of the benefits of looking at the data this way is that the different subject's responses effectively become repeats at the signal levels which (until now) only had one response at them. Therefore, it is possible to construct vertical confidence intervals for the binomial 3AFC process.

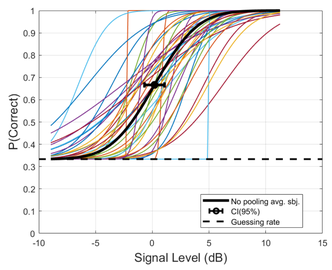
Scaled probit regression is fit to all of the data at once, resulting in a threshold and spread that represent all 40 subjects. For the most part, the curve passes through the binomial confidence intervals at each level of the signal.



No Pooling:

In this approach, each subject's data is fit with a separate probit. Statistics (like the average subject parameters) can then be derived from the sets of the individual fit parameters.

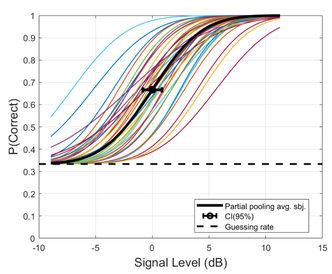
A main issue with this type of analysis is that there is only a limited number of points per subject. Unreasonable steep or shallow psychometric functions can be found (as can be seen at right), as well as cases where the fitting algorithms fail to converge due to the low data density. This should be reflected in large uncertainty estimates for each subject, however it is not simple to incorporate this information in the creation of the population-level statistics.



Partial Pooling:

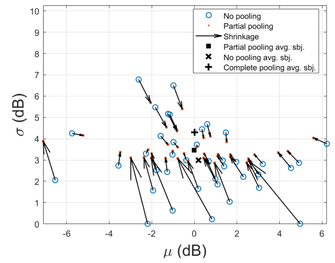
This approach is often called multilevel regression. Here an assumption is made about how the individual-level parameters are distributed (the thresholds are normal around the population average, and the slopes follow a χ^2 distribution), and then the entire system of individual fits and population-level parameters are determined by one algorithm.

It can be seen at right that the unrealistic individual spreads (e.g. the step functions found in the no pooling analysis) are absent.



Shrinkage:

An interesting effect takes place in the partial pooling results called shrinkage. Since the entire model is fit at once, information is shared between the population and individual levels — individuals impact the population average, but the population-level estimates also impact the individual estimates.



The plot at the right shows the differences between the parameters found for the individual subjects between the no pooling and partial pooling results. It can be seen that the arrows point more vertically than horizontally, indicating that the differences in threshold found between subjects in the no pooling analysis are more likely to be significant than those found in spread.

Conclusions:

Conclusions are given here for the 3 approaches. The colored boxes next to each heading indicate how the approach fared in terms of the goals set out in the problem statement.

Complete Pooling:

This approach is deceptively attractive as it produces vertical confidence intervals and is easy to compute. However, one has to take great care in interpreting the results. In this case, the value of the spread (4.3 dB) is a conflation between the spread of the average subject and the differences in threshold between subjects, and is significantly greater than the 3.5 dB found in partial pooling. Also, this method cannot produce estimates of how individual subjects vary around the mean.

No Pooling:

This approach is just as simple as complete pooling computationally, though its tendency to produce confusing results or failing to converge means more user expertise is required (especially when using a scaled psychometric function). The estimates produced do not conflate effects across individual and population levels. Aggregating the individual fits to the population level without considering the individual-level uncertainty can create inconsistency.

Partial Pooling:

This approach is clearly the best performer when it comes to generating consistent results in one step for all of the parameters of interest. Like the no pooling analysis, the outputs that are generated are of the parameters we are interested in and in meaningful units.

However, it is significantly more complex conceptually, and can produce effects in the output that can be counterintuitive. Accordingly, care must be taken when using this method, and application to other datasets may require significant statistical sophistication of the user.