Eric Kernfeld and Paul Sampson

UW Statistical Consulting program, January 2016

Analysis Plans for Transparency Data

This document describes statistical approaches for analyzing results of a human-robot interaction study. This is a technical memo, not a full record; it is intended to be used by those familiar with the project. For more information on the project, contact Leah Perlmutter. If you are a stat/biostat members with access, see the consulting program's winter 2015 records.

Scientific Context

These analysis plans aim to test three hypotheses:

- H1: Adding visualization-based transparency after using only natural transparency will improve task metrics. (During P1, headset and monitor are better than baseline.)
- H2: By improving the user's mental model, visualization-based transparency will improve task metrics even after it is removed. (During P2, headset and monitor are better than baseline.)
- H3: The medium in which visualizations are provided (monitor versus headset) will not impact task metrics. (During P1 and P2, headset and monitor have about the same effect.)

Later in the document, I generate precise *statistical hypotheses* corresponding to these scientific hypotheses.

The main obstacle in our analysis: we need to disentangle the effect of transparency and the effect of learning (i.e. learning from experience alone). The document is broken into two parts:

- Analysis 1 uses uncontroversial assumptions, but it can test only H3, not the others.
- Analysis 2 assumes the effect of learning is constant over time on a logarithmic scale (or log-odds, for binary data). This assumption allows for testing H1 and H2.

Each analysis is presented first for quantitative data. A section at the end discusses modifications for binary data.

Implementation

One easy way to implement the analyses described below is to use the `R` computing language and the packages `lme4` and `car`. This combination of software allows models to be specified concisely in *formula* *syntax*. There is a whole separate document (in progress) laying out the implementation.

Data setup

These data were gathered on 20 subjects. For each subject, some continuous responses were measured:

- number of attempts averaged over a set of tasks
- number of words for each attempt averaged over a set of tasks
- average time taken for each attempt
- average time taken for each task

For count data or time-to-event data, ANOVA-style linear models are more reliable when data are log-transformed. These measurements are all strictly greater than zero, so taking logs is no problem. This also provides a nice interpretation of (functions of the) parameters as percent changes. So, in the descriptions of the analyses, Y_i will denote the log of the task metric. To describe the individual, i will range from 1 to 20. To complicate things slightly, each task metric was measured for baseline, monitor, and headset, and for phase 1 (P1, visualization present) and phase 2 (P2, aftereffects). This will be described as Y_{ij}, so that:

- Y_{21} is the log task metric for person **2**, P1, baseline
- Y_{11} is the log task metric for person **1**, P1, baseline
- Y_{12} is the log task metric for person 1, P2, baseline
- Y_{13} is the log task metric for person 1, P1, monitor, *even if the monitor came after the headset for that person*
- Y_{14} is the log task metric for person 1, P2, monitor
- Y_{15} is the log task metric for person 1, P1, headset, *even if the headset came before the monitor for that person*
- Y_{16} is the log task metric for person 1, P2, headset

For each individual, let X_{ij} be:

- 1 if $j = 3, 4$ and person i used the monitor last
- 1 if $j = 5, 6$ and person i used the headset last
- 0 otherwise

A few binary responses were also recorded:

- whether pointing helped for each command in that phase (5 or 10 binary outcomes)
- whether language helped for each command in that phase (5 or 10 binary outcomes)
- whether each command was successful in that phase (5 or 10 binary outcomes)

Each analysis is presented first for quantitative data. A section at the end discusses modifications for binary data.

Analysis 1

Model and interpretation

The proposed statistical model is a *random-effects* model. It takes the form

$$Y_{ij} = z_i + \mu_j + \beta_{learn} X_{ij} + \epsilon_{ij}.$$

To interpret each component:

- $\epsilon_{ij} \sim N(0, \sigma^2_{phase})$ -- A random "noise" term describing natural variability in a single measurement Y_{ij}.
- $z_i \sim N(0, \sigma^2_{person})$ -- a random effect specific to person i, describing how their initial aptitude with the robot differs from the average.
- β_{learn} -- the average effect of learning that occurs between the first transparency device and the second (rounds 2 and 3; tasks 16 and 31; batches 7 and 13). More info in the section below.
- μ_j -- a fixed, unknown parameter describing:
    - $j = 1$: the average log task metric for P1, baseline
    - $j = 2$: the average log task metric for P2, baseline
    - $j = 3$: the average log task metric for P1, monitor, *having learned from the 15 baseline tasks*
    - $j = 4$: the average log task metric for P2, monitor, *having learned from the 15 baseline tasks*
    - $j = 5$: the average log task metric for P1, headset, *having learned from the 15 baseline tasks*
    - $j = 6$: the average log task metric for P2, headset, *having learned from the 15 baseline tasks*

Capabilities and Assumptions

Assumptions and limitations

This model assumes measurements from individual people are statistically independent. This assumption is crucial.

The model assumes the effect of learning between rounds 2 and 3 is *multiplicative*: on average, the extra practice alters task metrics by a certain percentage. If people learn so much from the headset that their monitor performance skyrockets, but not the other way around, this assumption would be violated.

As a side note, μ_4 and μ_3 are not directly comparable, because the measure performance on different items, and they may also differ due to learning. This is true for any pair of j values where one is odd and the other even.

Estimating effect sizes

It is possible to estimate every parameter in this model, in particular μ_j. This can help tell us which effects are scientifically meaningful. However, the parameters μ_j capture some unknown mixture of the transparency's effect and a learning effect. The parameter β_{learn} can also be estimated. Since β_{learn} is added to the log task metrics, $e^{\beta_{learn}} = 0.7$ means that learning between rounds 2 and 3 decreases predicted task metrics by 30%. This can be used to check whether the effect of learning is scientifically meaningful.

Statistical hypothesis testing

To convert H3 into a statistical hypothesis, we could test the double constraint "$\mu_3 = \mu_5 \textit{ and } \mu_4 = \mu_6$." We initially ran tests for $\mu_3 = \mu_5$ and $\mu_4 = \mu_6$ separately, but "$\mu_3 = \mu_5 \textit{ and } \mu_4 = \mu_6$" corresponds more closely to the scientific hypothesis H3, so the code now tests "$\mu_3 = \mu_5 \textit{ and } \mu_4 = \mu_6$".

This analysis plan does not address H1 or H2.

Analysis 2

The statistical model here is very similar, but it makes one big assumption about the learning curve and it requires one additional covariate. Let X'_{ij} be 1 if $j = 3, 4, 5, 6$ and 0 otherwise. The model is:

$$Y_{ij} = z_i + \nu_j + \beta_{learn} X'_{ij} + \beta_{learn} X_{ij} + \epsilon_{ij}.$$

To interpret the components that differ from Analysis 1:

- β_{learn} -- the average effect of learning that occurs between the first transparency device and the second *or the baseline and the first device.*
- ν_j -- a fixed, unknown parameter describing:
  - $j = 1$: the average log task metric for P1, baseline
  - $j = 2$: the average log task metric for P2, baseline
  - $j = 3$: the average log task metric for P1, monitor, *without the effect of learning from the 15 baseline tasks*
  - $j = 4$: the average log task metric for P2, monitor, *without the effect of learning from the 15 baseline tasks*
  - $j = 5$: the average log task metric for P1, headset, *without the effect of learning from the 15 baseline tasks*
  - $j = 6$: the average log task metric for P2, headset, *without the effect of learning from the 15 baseline tasks*

Capabilities and Assumptions

This model makes the same assumptions outlined in Analysis 1:

- It still assumes measurements from individual people are statistically independent. This assumption is crucial.
- It still assumes the effect of learning between rounds 2 and 3 is *multiplicative*.

One extra assumption is that the effect of learning between rounds 2 and 3 equals the effect of learning between rounds 1 and 2. Earlier, we described this by saying "learning is linear", but since we are on a log scale, this should be reworded. The model assumes *the percent change in average task metrics due to learning between rounds rounds 2 and 3* equals *the percent change in average task metrics due to learning between rounds rounds 1 and 2*.

Since H3 can be tested without this extra assumption, Analysis 2 does not address H3. For the rest, we intially tested $\nu_3 + \nu_5 = 2\nu_1$ to address H1. We tested $\nu_4 + \nu_6 = 2\nu_2$ to address H2. Another option, which is not mathematically equivalent, is to test $\nu_3 = \nu_5 = \nu_1$ for H1 and $\nu_4 = \nu_6 = \nu_2$ for H2. I prefer the second set of options, since they seem closer to our scientific hypotheses. So, the code now tests $\nu_3 = \nu_5 = \nu_1$ for H1 and $\nu_4 = \nu_6 = \nu_2$ for H2.

These tests only assess whether results are due to chance. To assess the *magnitude* of the effects, exponentials of differences are useful. For example, $e^{\nu_3 - \nu_1} = 0.7$ means the estimated *effect* of the *monitor* (compared to baseline) reduces predicted task metrics by 30%. For another example, $e^{\nu_6 - \nu_2} = 0.7$ means the estimated *aftereffect* of the *headset* is to reduce predicted task metrics by 30%.

Analyzing binary outcomes

Model form

This document is primarily intended for Leah's team, but as a brief aside for the consultants, we will analyze the binary data using the same fixed and random effects outlined above. But, we embed them into a quasi-binomial GLM with a canonical link function.

To provide more detail for Leah et al., this technique is closely related to *logistic regression*. We will analyze the binary data by modeling the probability of success for each combination of factors. A simple way to do this is to assume that the success probability maps to the various conditions via a function similar to the linear models in analysis 1.

$$\log(\frac{p_{ij}}{1 - p_{ij}}) = z_i + \mu_j + \beta_{learn} X_{ij}.$$

(For technical reasons, the ϵ_{ij} term is no longer included.)

Intepretation Fundamentals

We can interpret the results using *log odds* or *log odds ratios*. The odds associated with the probability p are $\frac{p}{1-p}$; a probability of 3/4 is the same as 3 to 1 odds. Part of the convenience of this model is that odds can be any nonnegative number, and log odds can be any real number; this means the right hand side above can behave however it wants. Undoing the log shows the odds explicitly:

$$\frac{p_{ij}}{1-p_{ij}} = \exp(\mu_j)\exp(z_i)\exp(\beta_{learn}).$$

The effect measured by β_{learn} can be described as a log odds ratio because if the odds of success with no learning are

$$q' = \exp(\mu_j)\exp(z_i)$$

and the odds of success with learning included are

$$q = \exp(\mu_j)\exp(z_i)\exp(\beta_{learn}),$$

then

$$\beta_{learn} = \log(\frac{q}{q'}).$$

Interpretation Details

Analysis I

To go through the Analysis 1 model in detail:

- $\exp(\mu_j)$ is the baseline odds of success under conditions given by j, which includes the effect of transparency and initial learning.
- $\exp(z_i)$ is the odds ratio associated with subject i.
- $\exp(\beta_{learn})$ is the odds ratio attributable to learning between rounds 2 and 3. No effect coincides to $\beta_{learn} = 0$ and $\exp(\beta_{learn}) = 1$.

Another, equally valid interpretation:

- $\exp(z_i)$ is the baseline odds of success associated with subject i.
- $\exp(\mu_j)$ is the odds ratio attributable to transparency conditions j and initial learning.

Analysis 2

Analysis 2 can be reworked in a similar way: if we assume

$$\frac{p_{ij}}{1 - p_{ij}} = \exp(\nu_j) \exp(z_i) \exp(\beta_{learn} X_{ij}) \exp(\beta_{learn} X'_{ij}),$$

then we might say

- $\exp(\nu_j)$ is the baseline odds of success under conditions given by j.
- $\exp(z_i)$ is the odds ratio associated with subject i.
- $\exp(\beta_{learn})$ is the odds ratio attributable to learning between rounds 2 and 3 or between rounds 1 and 2. The analysis assumes those two odds ratios are equal.

Diagnostics

It is necessary to check some of the modeling assumptions. In particular:

- the level of variability should be constant, not depending on the main effects we want to model. This can be checked by plotting residuals versus fitted values. For binary data, a type of residuals can be obtained.
- To check that the testing procedures will be reliable, residual distributions for quantitative task metrics should be symmetric and unimodal.
- Checking that measurements from different people are independent is impossible, but we can at least plot estimates of z_i ordered by i (does this coincide with the order in which they were measured?) to look for serial correlation.