Empirical answers to four questions in analysis of time series data[1]

Daniel Mirman, Ph.D.
Moss Rehabilitation Research Institute

Introduction

Cognitive neuroscience and cognitive rehabilitation research frequently involve the collection of time series data. Whether such data track the course of recovery over days or years, or the course of processing of a single stimulus over seconds or even milliseconds, time series data provide unique and important information. Unfortunately, they also pose substantial data analysis challenges, particularly concerning the temporal dependencies in the data (i.e., observations at time *t* are highly correlated with observations at time $t \pm 1$) and the substantial nesting of data (e.g., individual observations within participants, participants within groups). Multilevel regression modeling provides a set of tools for addressing these challenges (for detailed introductions see: Singer & Willett, 2003; Raudenbush & Bryk, 2002). We have particularly focused on applying these tools to the analysis of data from eye-tracking experiments (Mirman, Dixon, & Magnuson, 2008), which have become increasingly common in studies of language and cognition in a wide range of populations, including individuals with language or other cognitive deficits.

Despite the wide acceptance and use of eye-tracking and other time-course methods, there is no concomitant agreement on the appropriate statistical method(s) for analyzing the resulting data. Many investigators continue to use t-tests and ANOVAs, even though these are known to be ill-suited to such data. Among regression analysis methods, there is little agreement about which regression approaches are best or how they should be implemented. In part, these differences persist because abstract statistical considerations do not find much traction with behavioral scientists – in the absence of concrete evidence, behavioral scientists are disinclined to change their statistical methods just because someone says that some other method is "better". To address this lack of empirical evidence, we have recently undertaken an empirical investigation of four open questions in analysis of time-series data, focusing on the domain of eye-tracking data:

1. How to aggregate data across trials of different durations?
2. Do bin-wise t-tests and full time course regression methods differ in robustness to individual differences?
3. Is lack of dynamic consistency a problem for non-linear regression models?
4. Should participants be treated as fixed or random effects?

Below, I summarize the issues standing behind each of those questions and the results that led us to a particular answer.

[1] This is a short summary of a full report that was co-authored with Allison E. Britt and Pyeong Whan Cho: Mirman, D., Britt, A.E., and Cho, P.W. (in preparation). *Empirical answers to four questions in analysis of "Visual World Paradigm" and other time course data.*

Question 1: How to aggregate data across trials of different durations?

In time series studies, the individual time series may not always have the same lengths. For example, if individual trials are terminated by a participant's response, then they will necessarily vary in their lengths, both within-participant and between-participants. When aggregating data across trials, how should one deal with time periods when some trials have ended and other trials have not? Some researchers argue that since there is no data for trials that have ended, they should not be counted during that time period. That is, for each time bin, only those trials that have data for that time bin should be counted. However, that would be a form of selection bias, systematically biasing the later time points to over-represent the slow trials. In many experiments slow response times are causally related to cognitive processes (e.g., competition slows down responses), so this selection bias could lead to serious theoretical consequences.

To understand why this is the case, imagine that we want to evaluate the response rate over time to a drug for a deadly disease. We enroll 100 participants in the trial and administer the drug. At first, only 50% of the participants respond to the drug. As the trial progresses, the non-responders begin to, unfortunately, die. After 6 months, only 75 participants are alive and participating in the trial and the same 50 are responding to the treatment. At this point, is the response rate the same 50% or has it risen to 67%? Would it be accurate to conclude that responsiveness to the treatment increases after 6 months? This example makes it intuitively clear why excluding terminated trials is a form of selection bias that will systematically distort the results. In other words, unbiased data aggregation requires that the denominator of the proportion calculation remain the same over the full time course. Instead, data should be imputed for terminated based on what is reasonable for the study context or task. In an eye-tracking study this could be the final fixation location, or target object fixation (assuming only correct response trials are included in the analysis), or neutral object fixation (e.g., central fixation cross, which is not of interest to the study).

Question 2: Do the analysis methods differ in robustness to individual differences?

As mentioned above, t-tests and ANOVAs remain the most common analysis methods for eye-tracking (and other time series) data, even though they pose several practical problems. In particular, conducting repeated time window tests requires defining those time windows, which introduces experimenter bias (i.e., defining time windows such that the predicted effects will emerge) and presents a trade-off between temporal resolution (smaller time windows) and power (less data in each time window). Furthermore, conducting independent analyses of time windows ignores the inherent and non-arbitrary continuity between time windows and discretizes the continuity of the underlying processes. That is, a t-test may detect a statistically reliable difference in time bin *t* and not in *t-1*, which could be interpreted as a fundamental difference in the processes operating at *t* compared to *t-1*; however, the same result could arise from a continuous and gradually evolving cognitive process that simply failed to reach the arbitrarily-set statistical significance threshold at *t-1*.

We examined these general concerns in the context of a specific experimental factor: individual variability. We used Monte Carlo simulation to generate eye-tracking-like data with a constant condition effect size and differing amounts of individual variability between participants. That is, the condition effect (what the simulated experimenter was investigating) was constant in the population, but the individual simulated participants differed in parameters reflecting their overall processing speed, tendency to look around the screen, etc. The results revealed, not surprisingly, that a multilevel regression approach was more robust to individual differences than bin-wise t-tests. As individual variability increased, the t-test approach became very inconsistent across simulated "experiments". Even when the underlying effect size and population variability were held constant, some simulated experiments showed no condition differences at all, some showed early effects (significant differences in early time bins and not later time bins), some showed late effects (significant differences in late time bins and not early time bins), and some showed condition effects in almost all time bins. In contrast, regression analysis showed quite consistent results at different levels of individual variability and for replications at a particular level of variability. These results concretely and empirically demonstrate that regression is both more powerful and less likely to produce spurious time course differences.

Question 3: Is lack of dynamic consistency a problem for non-linear[2] regression models?

Broadly speaking, regression methods can be based on linear models (including polynomial models, which are linear in their parameters) or non-linear models. Since many time series are asymptotic (learning curves, forgetting curves, fixation proportion curves, etc.), non-linear models have an intuitive appeal. However, non-linear models are not *dynamically consistent*: the model of the average is not equal to the average of the individual models. At the heart of much of statistical inference in the behavioral sciences is the assumption that the mean represents the central tendency of a sample of individual observations. As a result, violating this assumption (i.e., lacking dynamic consistency) could undermine the very foundation of how we think about our data and make inferences about it. We empirically examined the issue of dynamic consistency by comparing the average of individual models with the model of the average for data drawn from a typical "Visual World Paradigm" eye-tracking experiment (Mirman & Magnuson, 2009) that tested a relatively large sample (*N* = 38) from a relatively homogenous population (undergraduate college students at the University of Connecticut).

We used a Logistic Power Peak model as a representative dynamically inconsistent non-linear model that has been used to fit eye-tracking data in previous studies (Scheepers, Keller, & Lapata, 2008). The first thing we noticed was that non-linear models did not converge reliably for individual participants. Indeed, this has been mentioned by researchers who used non-linear models by way of explaining why they analyzed average data from sub-groups of participants or used jackknife/bootstrapping techniques in their analyses. Second, we found that there were, indeed, substantial deviations between the model of the average data and the average of the models of individual participants, which are graphically illustrated in Figure 1. The model of the average was a very good fit to the average data (thick solid line) and there were similarly good fits to individual participant data by the individual participant models

[2] We are specifically referring to non-linear models that are not dynamically consistent. Some "non-linear" models are dynamically consistent (e.g., polynomial models and some formulations of sigmoid models) and are, obviously, exempt from this possible problem.

(thin gray lines). However, when the parameters from those individual participant models were averaged together, the resulting curves did not look like the models of the average (differences between the thick lines).

In sum, lack of dynamic consistency poses a serious problem for non-linear regression models of time-series data. The logic of comparing sample means is invalid when the model of the average is not equal to the average of the individual models. Analyzing only overall average data can avoid this problem, but this approach conflates the sample mean and the population mean (i.e., the variability across individuals is not considered when evaluating condition differences). Analyzing subsamples, as in jackknifing or other resampling techniques, may reduce the severity of dynamic inconsistency while including individual participant variability in the analysis, but this approach precludes analysis of individual differences. Because of these weaknesses, non-linear regression appears to be a substantially less optimal approach than polynomial regression for analysis of gaze (and other time-series) data.

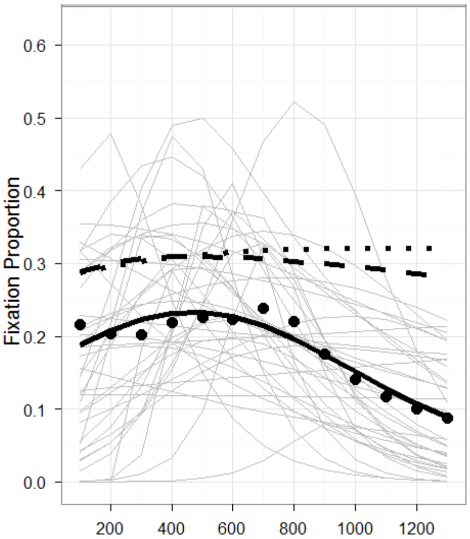

Figure 1. Comparison of model fits. The symbols indicate the mean of the gaze data. The thick black lines correspond to the three model fits: the model of the average data (solid lines), the average of all of the individual models (dashed lines), and the average of just the fully converged models (dotted lines). The thin gray lines show the individual participant model curves.

Question 4: Should participants be treated as fixed or random effects?

In our previous work on adapting multilevel regression models to eye-tracking data (Mirman et al., 2008) we treated both experimentally-controlled factors (e.g., word frequency) and participants as fixed effects. However, the traditional logic is that if levels of a factor are fixed in the world and reproducible (i.e., experimentally-controlled factors such as word frequency, participant's native language, etc.), they should be considered fixed effects; if the levels correspond to randomly-sampled observational units (e.g., individual participants from some population or words from a set with particular properties), then they should be considered random effects.

The mathematical difference is that when participants are treated as a fixed effect, each participant's fixation proportion curve parameters are estimated independently. When participants are treated as a random effect, each participant's fixation proportion curve parameters are constrained to be random deviations from the population mean curve parameters, with the deviations assumed to conform to a normal distribution with mean equal to 0. This additional constraint means that each individual's parameter estimates from a random participant effects model are weighted averages of the parameter estimates from a fixed participant effects model and the group-level parameter estimates. Put simply, the parameter estimates reflect both the individual participant's data and the whole group data. As a result, they tend to "shrink" toward the population mean. This shrinkage can have positive and negative consequences. When individual participant estimates are allowed to be fully independent (i.e., treated as fixed effects), they provide better (that is, independent) estimates of differences between individual participants, but the resulting model can overfit the data.

Researchers deciding whether to treat participants as random or fixed effects need to consider their research goals and their confidence in the homogeneity and normality of the sample population. If the goal is to generalize, then the researcher is essentially forced to assume that the sample is drawn from a homogeneous population and should treat participants as a random effect. However, this form of generalization is not always the goal. For example, neurological case studies inform cognitive theories by showing what must be possible (as in an existence proof) and generating new hypotheses. In such contexts, the goal is to describe the observed data as well as possible and treating participants as fixed effects may be more appropriate. Since participant fixed effect parameters better capture individual differences, they may provide a better approach for studying individual differences (e.g., Mirman, Yee, Blumstein, & Magnuson, 2011; and the individual differences example in Mirman et al., 2008). In such cases, it may be advantageous to acquire independent parameter estimates for the participants by treating them as fixed effects rather than random effects. Finally, for hypothetically homogeneous populations like typical college students, treating participants as random effects may the better approach; but for clearly non-homogeneous populations like neurological patients (who all have unique clinical and neurological presentations, even if their diagnosis is the same) treating participants as fixed effects may be more appropriate. We have also observed that model convergence appears to be somewhat more robust when participants are treated as fixed effects rather than as random effects.

Summary

We considered four questions regarding analysis of time-series data from an empirical perspective. Our results showed that: (1) Terminated trials need to remain part of the analysis to avoid selection bias. (2) Regression is more robust to individual variability than bin-by-bin t-tests and less likely to produce spurious time course differences. (3) Lack of dynamic consistency poses a serious problem for interpretation of non-linear regression models. (4) Participants should be treated as random effects when the statistical goal is primarily generalization, but can be treated as fixed effects when the sample is non-random, non-homogenous, or non-normal, or when the statistical goal is description.

References

Mirman, D., & Magnuson, J. S. (2009). Dynamics of activation of semantically similar concepts during spoken word recognition. *Memory & Cognition*, *37*(7), 1026-1039. doi:10.3758/MC.37.7.1026

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, *59*(4), 475-494. doi:10.1016/j.jml.2007.11.006

Mirman, D., Yee, E., Blumstein, S. E., & Magnuson, J. S. (2011). Theories of spoken word recognition deficits in aphasia: Evidence from eye-tracking and computational modeling. *Brain and Language*, *117*(2), 53-68. doi:10.1016/j.bandl.2011.01.004

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models*. Thousand Oaks: Sage Publications.

Scheepers, C., Keller, F., & Lapata, M. (2008). Evidence for serial coercion: A time course analysis using the visual-world paradigm. *Cognitive Psychology*, *56*(1), 1-29.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal analysis: Modeling change and event occurrence*. New York: Oxford University Press.