Acta Psychologica 144 (2013) 83–96

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/ locate/actpsy

Modeling accuracy as a function of response time with the generalized linear mixed effects model

D.J. Davidson [a,*], A.E. Martin [b]

[a] Basque Center for Cognition, Brain, and Language, Donostia, Basque Country, Spain
[b] University of Edinburgh, Edinburgh, Scotland, United Kingdom

ARTICLE INFO

Article history:
Received 7 December 2012
Received in revised form 11 April 2013
Accepted 24 April 2013
Available online 14 June 2013

PsycINFO codes:
2240
2340
2720

Keywords:
Speed–accuracy tradeoff
Generalized linear mixed effects model

ABSTRACT

In psycholinguistic studies using error rates as a response measure, response times (RT) are most often analyzed independently of the error rate, although it is widely recognized that they are related. In this paper we present a mixed effects logistic regression model for the error rate that uses RT as a trial-level fixed- and random-effect regression input. Production data from a translation–recall experiment are analyzed as an example. Several model comparisons reveal that RT improves the fit of the regression model for the error rate. Two simulation studies then show how the mixed effects regression model can identify individual participants for whom (a) faster responses are more accurate, (b) faster responses are less accurate, or (c) there is no relation between speed and accuracy. These results show that this type of model can serve as a useful adjunct to traditional techniques, allowing psycholinguistic researchers to examine more closely the relationship between RT and accuracy in individual subjects and better account for the variability which may be present, as well as a preliminary step to more advanced RT–accuracy modeling.

1. Introduction

Response time and accuracy are both common dependent measures in experimental psycholinguistics and cognitive psychology. Most often these two variables are analyzed separately, with the implicit (and untested) assumption that they represent two independent response measures, though they issue from the same underlying process or processes. However, the existing literature shows that they are often not statistically independent of each other, and moreover the relationship is sometimes complex, and subject to individual differences. Since RT and accuracy are variables derived from the same decision process with an unknown and dynamic criterion, it seems conceptually difficult to regard them as independent variables, and ideally, statistical models of RT and accuracy should reflect this (Fitts, 1966; Pachella & Fisher, 1969; Pachella & Pew, 1968; Pew, 1969; Ratcliff, 1985; Ratcliff & Hacker, 1981; Ratcliff & Rouder, 1998; Wickelgren, 1977). Many researchers consider qualitatively whether a tradeoff between RT and accuracy is present in their data, at least at the level of group or condition averages. Indeed, this represents an additional "researcher degree of freedom" in the analysis of many data sets (Simmons, Nelson, & Simonsohn, 2011), because researchers can choose whether to emphasize either the results of the RT or the error analysis in support of their claims, when in fact, the two are often not independent sources of evidence. However, even when a tradeoff is not present, an accurate model of the relationship between RT and accuracy may improve the statistical analysis of a given data set. Recent work in psychometrics by Loeys, Rosseel, and Baten (2011), building on earlier work by Van der Linden (2007), has shown how to construct a joint linear mixed effects model for the RT–accuracy relation using a Bayesian approach. The present paper provides a simplified mixed effects model that can be used as a building block for these more elaborate analyses. We argue that classifying individual subjects' relationship between error rate and RT, as well as the group-level pattern, is an important first step in data analysis that can offer critical insights for common psycholinguistic paradigms.

Broadly speaking, there may be one of three (simple) relationships between the time it takes for participants to respond, and the probability that they make an error on a given trial. First, it can be that the more accurate subjects are, the earlier they respond (or, equivalently, with decreasing accuracy they respond later). In this situation lower error rates and earlier RTs both indicate better performance in some sense. This means that participants are not trading response time for accuracy.

A second type of relationship is for subjects to become more accurate at the *expense* of response time. That is, the more accurate subjects are, the slower they respond. This is more commonly known as a speed–accuracy tradeoff, and this pattern is particularly problematic when two or more experimental conditions are to be compared. If participants are more accurate but also later in one condition compared to another, one must entertain the possible explanation that there is not a simple effect of condition on accuracy or RT, but rather a more complex effect of condition on the RT–accuracy relation. This

* Corresponding author. Tel.: +34 943 309 300 209.
E-mail address: d.davidson@bcbl.eu (D.J. Davidson).

0001-6918/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.actpsy.2013.04.016

does not invalidate RT or accuracy as response measures, but depending on the magnitude and direction of the tradeoff, it can be difficult to draw conclusions about a dataset.

Finally, there might be no systematic relationship between RT and accuracy. In this case, a curve relating accuracy to RT will be essentially flat. Although there are important exceptions, this relationship appears to be the most commonly-assumed scenario for psycholinguistics, cognitive psychology and cognitive neuroscience researchers, at least implicitly, because it is currently the most commonly accepted practice for the analysis of RTs and errors to be presented as if the effects are independent of each other. However, it is still relatively rare for researchers to formally test whether this is the case.

In all three of these scenarios, the relationship between accuracy and RT can be defined at either the *subject* level or the *trial* level of analysis. At the level of subject averages, some subjects may be faster as well as more accurate, while others may be faster only when they are less accurate, or there may be no systematic relationship between average RT and average accuracy. At the trial level the relation is defined between the probability of responding correctly on an individual trial and the individual trial RT, rather than the average accuracy versus average RT. These two levels of analysis need not have the same relation. Even if it is the case that (on average) fast subjects are not any more likely to be more accurate, it can be the case that each subject shows a systematic relation between RT and accuracy around their individual subject-level averages. This important distinction between these two levels of analysis is discussed in more detail in Appendix B.

Besides the three possibilities described above, another reason the relationship between RT and accuracy is complex is that it is not always a linear relationship, in the sense of a straight line. When participants are making relatively many errors or relatively few errors, they may still take a short or a long amount of time to respond. That is, large differences in response time may correspond to relatively small changes in proportion correct (and vice versa). The probability of responding correctly or incorrectly is constrained between 0 and 1, but the time taken to respond in a task is typically constrained only by instructions or a response deadline, if at all. The result of this is that often the RT–accuracy relation has the form of a curve, and it is not well modeled using ordinary linear regression, without transforming the variables in some way. This RT–accuracy curve can, however, be modeled with logistic regression as we will outline below.

In sum, the relationship between accuracy and RT, when present, is sometimes not a simple linear function, and there are multiple levels to the relation. Most studies treat accuracy and RT as independent response measures, or arrange the experimental situation so that participants have a relatively high accuracy rate. However, in cases where participants have relatively low or relatively high accuracy, small changes in accuracy can correspond to large differences in response time.

1.1. RT–accuracy tradeoff functions

The notion that people can trade response time for accuracy in any task has been well-documented (Fitts, 1966; Garret, 1922; Hick, 1952; Ollman, 1966; Pachella & Pew, 1968; Pew, 1969; Schouten & Bekker, 1967; Wickelgren, 1977; Woodworth, 1899). At the heart of the problem is the fact that individual participants respond per an unknown internal criterion that is likely to be dynamic over time. Thus, participants can trade the speed of response for accuracy of response based on unobservable changes or differences in internal criterion. In order to study the timecourse of information processing, it is therefore more informative to obtain a full RT–accuracy function for an individual performing a given task, of which an RT would yield only one point in time. Wickelgren (1977) outlines various experimental procedures to derive the function (payoffs, deadlines, instructions, response binning or partitioning, and lastly the application of response signals) and argues that the only way to prevent speed–accuracy tradeoff is to use the specialized response-signal interruption paradigm (Reed, 1973, 1976; Schouten & Bekker, 1967). Unfortunately, the specialized design and the procedure needed to implement such a paradigm are not always feasible, nor desirable to many researchers. Furthermore, the analysis strategies are specialized — requiring special designs or statistical estimation techniques. For example, there are limitations as to the interpretation of partitioned responses, and often problems with sparse data in early bins of short reaction times — see Wickelgren (1975), Wickelgren (1977) and Schouten and Bekker (1967). Here, we advocate a simpler approach to diagnosing whether there is a tradeoff, or not, between RT and accuracy in a given dataset, without the application of specialized designs or procedures. Our aim is to give the user a straightforward and simple method for assessing the statistical relationship between RT and accuracy in a dataset — our approach is agnostic regarding the model of the underlying decision process that leads to performance and to the particular relationship between the two variables. However, we note that our analysis approach shares the same core assumptions seen in the extensive literature on computational and theoretical models of two-alternative forced-choice decision processes (e.g., Ratcliff, 1978; Ratcliff, Gomez, & McKoon, 2004; Ratcliff & McKoon, 2008) — namely that response time and performance are inextricably linked, that the relationship between the two must be included in any statistical model of the data, and that the former and the latter points are crucial for interpretation of the data. Note that the existence of any computational and/or statistical relationship (in our case, only the latter) between RT and accuracy in no way implies that changes in RT cause changes in accuracy, or vice versa, if RT is modeled as a function of accuracy.

An important observation to make is that both response time and accuracy can be modeled as random effects in the sense that a typical sample of response times or response choices will have a statistical *distribution*. This distribution can depend strongly on the particular subject who has been sampled. If this is the case, then at the level of individual trials of an experiment, there should be a strong relationship between the response time and the response choice because the same subject variability is affecting both, but is independent of other subjects. That is, RTs should be informative about accuracy at the trial level because both dependent measures will reflect individual variation in participants. At the same time, there may also be a systematic relation between RT and accuracy at the group level. The next section describes our approach to modeling RT as a regression input at these multiple levels.

1.2. Linear and generalized linear mixed effects models

In this paper we will model the proportion response as a function of RT, where the RT enters the model as either a fixed and/or a random effect. This is analogous to the situation in many datasets where the performance outcome variable y is binary. Examples include correct/error response, present/absent decisions about a stimulus, recalled/not-recalled in a memory experiment, or fluent/disfluent in a production experiment. All of these examples share the essential characteristic that the response y takes on one of two values. This situation is different from a response measure like RT, because the binary response is not accurately modeled as a Gaussian distribution at the trial level (e.g., in cases where the response is actually distributed as a binomial). In logistic regression (see Jaeger, 2008 for an introduction, also Jaeger, Graff, Croft, & Pontillo, 2011; Quené & van den Bergh, 2008), we instead model the probability that response = 1 for some regression input x in terms of the inverse logit: $Pr(y = 1|x) = \exp(\beta x)/(1 + \exp(\beta x))$ where x is parameterized with the coefficients β to represent the effect of experimental variables, as well as variables like RT. Here, we use the inverse logit because usually one wants to go from a calculated coefficient in our model to proportions,