Journal of Memory and Language 94 (2017) 28–42

Contents lists available at ScienceDirect

Journal of Memory and Language

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

# Empirical logit analysis is not logistic regression

Seamus Donnelly [a,b,c,*], Jay Verkuilen [c]

[a] Research School of Psychology, The Australian National University, Australia
[b] ARC Centre of Excellence for Dynamics of Language, Australia
[c] Program in Educational Psychology, City University of New York Graduate Center, United States

ARTICLE INFO

*Article history:*
Received 16 September 2015
revision received 27 October 2016
Available online 3 December 2016

*Keywords:*
Generalized linear mixed models
Empirical logit
Visual-world eye tracking
Logistic regression

ABSTRACT

Many recent psycholinguistic studies have used empirical logit analysis as a substitute for mixed-effects logistic regression. In this paper, we describe the differences between empirical logit analysis and mixed-effects logistic regression and highlight three interacting sources of bias in empirical logit analysis. We then report on two sets of simulations comparing logistic regression and empirical logit analysis. We show that relative to logistic regression, empirical logit analysis generally yields biased parameter estimates when proportions are close to 0 or 1, especially when the number of observations underlying a proportion is very low. We also show that, in some circumstances, this bias can create spurious interactions, leading to unacceptable Type I error rates. While these two models may provide similar answers to some questions, we encourage readers to interpret empirical logit parameters cautiously.



## Introduction

In the last decade, cognitive psychologists and psycholinguists have begun using mixed-effects logistic regression rather than traditional ANOVA methods for analyzing proportions. Indeed, a special issue of the *Journal of Memory and Language* contained several excellent papers illustrating the negative consequences of using linear regression and related techniques to analyze proportions (Barr, 2008; Dixon, 2008; Jaeger, 2008; Mirman, Dixon, & Magnuson, 2008). Similar concerns have been raised about the analysis of reaction time data (Lo & Andrews, 2015). Since then, interest in these models has increased dramatically. This is a welcome development as these models are markedly more sensitive to the nature of data gathered in cognitive psychology.

Unfortunately, ordinary logistic regression can be challenging to fit to new users, and mixed-effects logistic regression is even more so. Many recent publications have used a similar alternative technique called empirical logit analysis (Gollan, Schotter, Gomez, Murillo, & Rayner, 2014; Tanner, Nicol, & Brem, 2014; van de Velde, Meyer, & Konopka, 2014; Veríssimo & Clahsen, 2014). This approach entails aggregating observations into bins and calculating proportions, transforming the proportions using the empirical logit transformation, and analyzing the transformed proportions with a weighted mixed-effects linear regression model. The proportion for a given bin can be estimated as $\hat{\pi} = y/n$ where $y$ is the number of successes and $n$ is the total number of observations. Because proportions are not normally distributed, they are transformed using the empirical logit transformation

$$\text{elogit}(\hat{\pi}) = \log\left(\frac{\hat{\pi} + 1/2n}{1 - \hat{\pi} + 1/2n}\right). \quad (1)$$

The empirical logit transformation alters the logit transformation by adding $1/2n$ success and failures to each propor-

* Corresponding author at: Research School of Psychology (Building 38), The Australian National University, Acton, ACT 2601, Australia.
*E-mail address:* Seamus.donnelly@anu.edu.au (S. Donnelly).

http://dx.doi.org/10.1016/j.jml.2016.10.005
0749-596X/

tion. This is done because the logit function goes to $-\infty$ for 0 successes and $\infty$ for 0 failures, but doing so also shrinks the estimate towards 0. As per McCullagh and Nelder (1989), the logit transformation provides a good normalizing transformation for proportions, leading one to suppose it is reasonable to analyze them using linear regression provided weighted least squares is used with weights

$$w(n,y) = \frac{1}{\frac{1}{Y+.5} + \frac{1}{n-y+.5}}. \quad (2)$$

Some example values of the weights computed with this equation are shown in Table 1. As proportions may reflect different numbers of observations, it is sensible to weight those with more observations more heavily than those with few observations. In addition the logit transformation does not render the variance homogeneous, and the weights mitigate this problem. McCullagh and Nelder (1989) note, however, that the argument that supports the use of the empirical logit transformation is asymptotic in nature, and thus that "the transformation is useful only if all the binomial indices are fairly large" (p. 107). By this they mean *both* $y$ and $n$.

We have seen two justifications for this approach in the psycholinguistics literature. First, it is often used with eye-tracking data, where eye-tracking samples are considered events (the $n$ above), in which an area of interest can be either active or inactive (the $y$ above). Because fixations are infrequent relative to the sampling rate of most eye-trackers, successive samples are not independent, even once participant-level and item-level dependencies have been modeled with random effects. Barr (2008) suggested binning samples into time bins and treating groups of samples, rather than individual samples, as the unit of analysis. One way to analyze these groups of samples is to calculate proportions for each time bin, making empirical logit analysis seem like a suitable choice. However, many mixed effects model packages, including *lme4* (Bates, Maechler, Bolker, & Walker, 2015) and *glmmadmb* (Skaug, Fournier, Bolker, Magnusson, & Nielsen, 2016), now allow the user to analyze grouped binomial data using mixed-effects logistic regression, so a binned analysis could be done directly, assuming it was justified. A second justification is that mixed- effects logistic regression often fails to converge when ceiling or floor effects are present. Adding half a success and half a failure pulls proportions away from the ceiling or floor (Mirman, 2014). Given how often mixed-effects logistic regression fails to converge, this may seem like an attractive feature.

Many studies do not explicitly state why empirical logit analysis was conducted rather than logistic regression. Indeed, each of the co-authors of this paper has been approached by colleagues asking if they should conduct empirical logit analysis rather than logistic regression, even when the appropriate logistic regression model converged. Indeed, we have gotten the impression that many colleagues think empirical logit analysis and logistic regression are the same thing or that empirical logit analysis is better.

While empirical logit analysis may be useful in some cases, it differs from logistic regression in important ways and may lead to different inferences. Importantly, empirical logit analysis contains three sources of bias, which may interact in unpredictable ways. First, empirical logit analysis is a linear model on logit-transformed data whereas logistic regression is a generalized linear model. These models can differ substantially, especially when observed proportions are based on few observations or are near the ceiling or floor. Second, adding fractional successes and failures, while justified in some contexts, biases parameter estimates toward 0 on the logit scale (.5 on the probability scale). Third, by reducing the influence of extreme observations, the weights also bias parameter estimates toward 0 on the logit scale.

**Table 1**
Weights in empirical logit analysis.

| $\hat{\pi}$ | $w, n = 10$ | $w, n = 50$ | $w, n = 100$ |
|---|---|---|---|
| 1 | 0.48 | 0.50 | 0.50 |
| 0.9 | 1.28 | 4.90 | 9.40 |
| 0.8 | 1.88 | 8.30 | 16.30 |
| 0.7 | 2.28 | 10.70 | 21.20 |
| 0.6 | 2.48 | 12.10 | 24.10 |
| 0.5 | 2.48 | 12.50 | 25.00 |

In what follows, we consider the differences between empirical logit analysis and logistic regression. First, we review Maximum Likelihood Estimation and show how the empirical logit model differs from logistic regression. Second, we discuss how adding pseudo-observations and weights may mitigate the consequences of floor and ceiling effects but further bias parameter estimates. Third, we report on simulation studies comparing empirical logit analysis with logistic regression.

An important caveat is in order. Throughout this article we assume data that have been generated by a binomial process. While this is commonly assumed in eye-tracking experiments, we are not convinced that it is the case. As noted above, since eye-tracking samples are not independent, their variance may differ from that implied by the binomial distribution. We return to this point in the discussion.

## Theory

### *Maximum likelihood estimation*

One major difference between empirical logit analysis and logistic regression is that the former is a linear model applied to logit-transformed data whereas the latter is a generalized linear model. To see the difference between these two, it is necessary to understand how parameters for the two models are estimated. Most common statistical models are estimated via maximum likelihood estimation. Indeed, linear regression and ANOVA parameters are maximum likelihood estimates when assuming Gaussian errors.

Maximum likelihood estimation involves two steps. First, the user specifies the likelihood function of the problem—the hypothesized population probability distribution conditional on the independent variables. Second, the user finds the parameter values that maximize the likelihood function. Consider, for example, a linear regression predicting weight from height with the form