Behaviour Research and Therapy 98 (2017) 58–75

Contents lists available at ScienceDirect

Behaviour Research and Therapy

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

An introduction to using Bayesian linear regression with clinical data

Scott A. Baldwin [a, *], Michael J. Larson [b]

[a] Department of Psychology, Brigham Young University, USA
[b] Department of Psychology and Neuroscience Center, Brigham Young University, USA

ARTICLE INFO

Article history:
Received 26 August 2016
Received in revised form
20 December 2016
Accepted 20 December 2016
Available online 31 December 2016

Keywords:
Bayesian methods
MCMC
R
Stan
Prediction
Event-related potential
Error-related negativity (ERN)

ABSTRACT

Statistical training psychology focuses on frequentist methods. Bayesian methods are an alternative to standard frequentist methods. This article provides researchers with an introduction to fundamental ideas in Bayesian modeling. We use data from an electroencephalogram (EEG) and anxiety study to illustrate Bayesian models. Specifically, the models examine the relationship between error-related negativity (ERN), a particular event-related potential, and trait anxiety. Methodological topics covered include: how to set up a regression model in a Bayesian framework, specifying priors, examining convergence of the model, visualizing and interpreting posterior distributions, interval estimates, expected and predicted values, and model comparison tools. We also discuss situations where Bayesian methods can outperform frequentist methods as well has how to specify more complicated regression models. Finally, we conclude with recommendations about reporting guidelines for those using Bayesian methods in their own research. We provide data and R code for replicating our analyses.

Mandatory statistical training in psychology largely consists of training in analysis of variance (ANOVA) and linear regression. Some students also take advanced courses in structural equation modeling, multilevel modeling, or psychometrics (Aiken, West, & Millsap, 2008; Schwartz, Lilienfeld, Meca, & Sauvigné, 2016). Regardless of the specific topics, most statistical training will be from a frequentist perspective, where frequentist refers to a particular perspective on probability. Specifically, frequentist methods allow for long-run probability statements or probability statements about repeated sampling from a population (McElreath, 2016). Imagine a study comparing cognitive therapy and behavioral activation for depression. The null hypothesis for a *t*-test comparing the conditions after treatment is: The post-treatment mean for cognitive therapy does not differ from the post-treatment mean for behavioral activation in the population. Suppose the mean difference between the two treatments has a two-tailed *p*-value of 0.01. The correct interpretation of this *p*-value is: In the long-run, the probability of observing a difference as extreme or more extreme than the difference in this study is 0.01, if the null hypothesis is true. Said another way, if researchers repeatedly sampled from a population where cognitive therapy and behavioral activation are equally effective, the proportion of results across the samples that are as or more extreme than this study would be *p*. Likewise, a 95% confidence interval for the difference between these two treatments is interpreted as: Over repeated samples from the population, 95% of intervals constructed will contain the population difference. The interpretation does not describe the probability that a parameter is within an interval, but rather the performance of the method over many samples.

Frequentist methods are powerful and useful in many contexts; however, psychology's adoption of parts of frequentist methodology have not necessarily born fruit and may hinder scientific progress (cf. Meehl, 1978). The field's reliance on *p*-values and null hypothesis significance testing has been heavily criticized. Example problems include: (a) *p*-values are probabilities *assuming the null is true* and researchers often want to know the relative probability of the null as compared to an alternative (i.e., how much evidence is there for particular hypotheses; cf. Cohen, 1994); (b) flexibility in analysis (e.g., *p*-hacking, garden of forking paths) can heavily distort *p*-values (Gelman & Loken, 2014; Simmons, Nelson, & Simonsohn, 2011); (c) authors, reviewers, and editors privileging statistically significant results over non-significant results may distort the published literature (Greenwald, 1975) as well as create incentives that can lead to poor data analysis practices (cf. Rosenthal, 1994); (d) a focus on *p*-values leads to a binary decision regarding whether an effect is scientifically important (Gelman &

* Corresponding author. 285 TLRB, Brigham Young University, Provo, UT, 84602, USA.
E-mail address: scott_baldwin@byu.edu (S.A. Baldwin).

http://dx.doi.org/10.1016/j.brat.2016.12.016
0005-7967/

Carlin, 2014); (e) privileging p-values has reduced attention to precise predictions (Meehl, 1978); and (f) *p*-values do not necessarily help establish whether an effect is true or valid (Ioannidis, 2005).

Bayesian methods are an alternative to null hypothesis testing. They are useful tools that can help us learn about data and they can help us place more emphasis on important issues, such as uncertainty in estimates. However, thoughtful data analysis does not require Bayesian methods. Bayesian methods do not necessarily fix the problems listed above—they do not in and of themselves prevent problems with researcher flexibility just as null hypothesis testing did not produce the problems with research flexibility. McElreath's (2016) perspective is useful here: "This audience accepts that there is something vaguely wrong about typical statistical practice in the early 21st century, dominated as it is by *p*-values and a confusing menagerie of testing procedures …. The problem in my opinion isn't so much *p*-values as the set of odd rituals that have evolved around them, in the wilds of sciences, as well as the exclusion of so many other useful tools" (p. xi-xii). Bayesian methods can be a useful tool that helps researchers move beyond hunting for statistical significance and instead focus on other aspects of statistical models such as prediction, model fit, data visualization, and uncertainty. None of these things are unique to the Bayesian methods, but they are a natural outgrowth of the Bayesian perspective. A further advantage of Bayesian methods is that the tools available for evaluating and understanding simple models generalize fairly easily to more complex models. That is, as we move from normal to non-normal data or single-level to multilevel data, the methods and ideas we use to fit and evaluate the models remains the same.

The primary aim of this paper is to introduce clinical researchers to the fundamentals and foundational ideas of Bayesian models. No attempt is made to be exhaustive or to give readers all the tools needed to transition their analyses to Bayesian methods. Rather we aim to "get the ball rolling" by introducing Bayesian concepts with an accessible statistical model—linear regression. Given that most readers are familiar with regression, this will allow readers to easily identify how the Bayesian approach is similar and how it is different from traditional approaches. Where relevant, we have included references to texts and other resources where readers can find more information.

This paper is divided into five parts. First, we provide necessary background information about Bayesian methods. Second, we discuss an example dataset and show how to build a Bayesian model. Third, we examine the results of the analyses and show how we can extend the model. Fourth, we discuss how additional kinds of models can be constructed. Fifth, we provide Minimum Practice Guidelines that we recommend for researchers using and reporting Bayesian methods. Finally, to aid readers in learning the material, we have included an online appendix that contains the data and R code we used to perform the analyses and create the figures we report. Likewise, given that we introduce new terms, we have included an Appendix with a glossary of potentially unfamiliar terms.

1. Background

1.1. Bayes' theorem

Bayesian inference is straightforward. We start with a prediction about the parameters in the model (e.g., the difference between two groups or the correlation between X and Y). Specifically, we make predictions about the probability of specific parameter values—for example, are positive correlations more plausible than negative correlations or are all correlations equally plausible? Then one uses data to update the predictions about the probability of the parameters. Simply put, Bayesian analysis produces information about the probability of the parameters in the model that is the combination of the predictions about the parameters and what is learned about the parameters from the data (Kruschke, 2015; McElreath, 2016).

The prediction about the probability of the parameters is known as the "prior" because it represents the predictions about the parameter prior to seeing the data. Suppose we begin a study to evaluate the effectiveness of a new psychotropic medication for depression. Effectiveness is defined as the probability that someone will recover and not have clinically significant symptoms after 16 weeks of treatment. We do not know anything about the effectiveness of the treatment; therefore, we believe, before seeing the data, that the probability of recovery is evenly distributed between 0 and 1 (see the solid line in the top panel of Fig. 1)[1]. This is the prior for the analysis of treatment effects.

Researchers new to Bayesian methods may be uncomfortable with priors because priors appear to introduce subjectivity into the analyses. That is, if two researchers can obtain different results with the same dataset by choosing different priors, then which can be trusted? On the face of it, this seems like a reasonable concern. However, it is likely overblown for at least four reasons. First, subjectivity is part of any research project. The measures, design, participants, questions, and review process are all subjective and influenced by the biases, experience, and knowledge of researchers. For example, researchers may choose a particular statistical analysis method such as an ANOVA not because it is the best tool for the particular situation but because that is what they know or have used in previous publications. Likewise, researchers may select measures because they believe they are the most psychometrically sound or the best representation of the constructs of interest. Although these decisions can be carefully thought out and reasonable, the decisions are subjective—they are based on researchers' understanding and interpretation of the literature.

Second, researchers often know a lot about a topic that can influence their choice of prior distributions. This knowledge can include simple things like the range of the outcome variable, which will put limits on possible values parameters such as treatment effects. This knowledge can also include more complicated information such as plausible sizes of correlations or treatment (cf. Baldwin & Fellingham, 2013). Third, all statistical methods, frequentist or Bayesian, make assumptions that are not objective (Greenland, 2006). Fourth, choices about the likelihood for the data (e.g., are the data normally distributed? Binary? Count? Highly skewed?) are often far more important than the choice of the prior (Atkins & Gallop, 2007; Baldwin, Fellingham, & Baldwin, 2016).

Some researchers distinguish between objective and subjective priors. Objective priors aim to make minimal assumptions. Subjective priors incorporate all information available to the researcher about the parameters of interest (Rouder, Speckman, Sun, Morey, & Iverson, 2009). As noted above, we believe it is scientifically defensible to incorporate knowledge about parameters into models. Indeed, if prior information is ignored, one should explain why. In the end, all scientific decisions are evaluated by the research community, both before and after publication. Likewise, priors can and should be evaluated by the research community.

Returning to the example, after specifying the prior, we collect data on 10 participants and just 1 of the 10 recovers after 16 weeks

[1] Priors do not need to be flat. We likely know something about the average recovery rate of many drugs or even placebo, so a flat prior like this isn't particularly convincing. However, we use a flat prior at this point to help solidify understanding of the concept.