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.

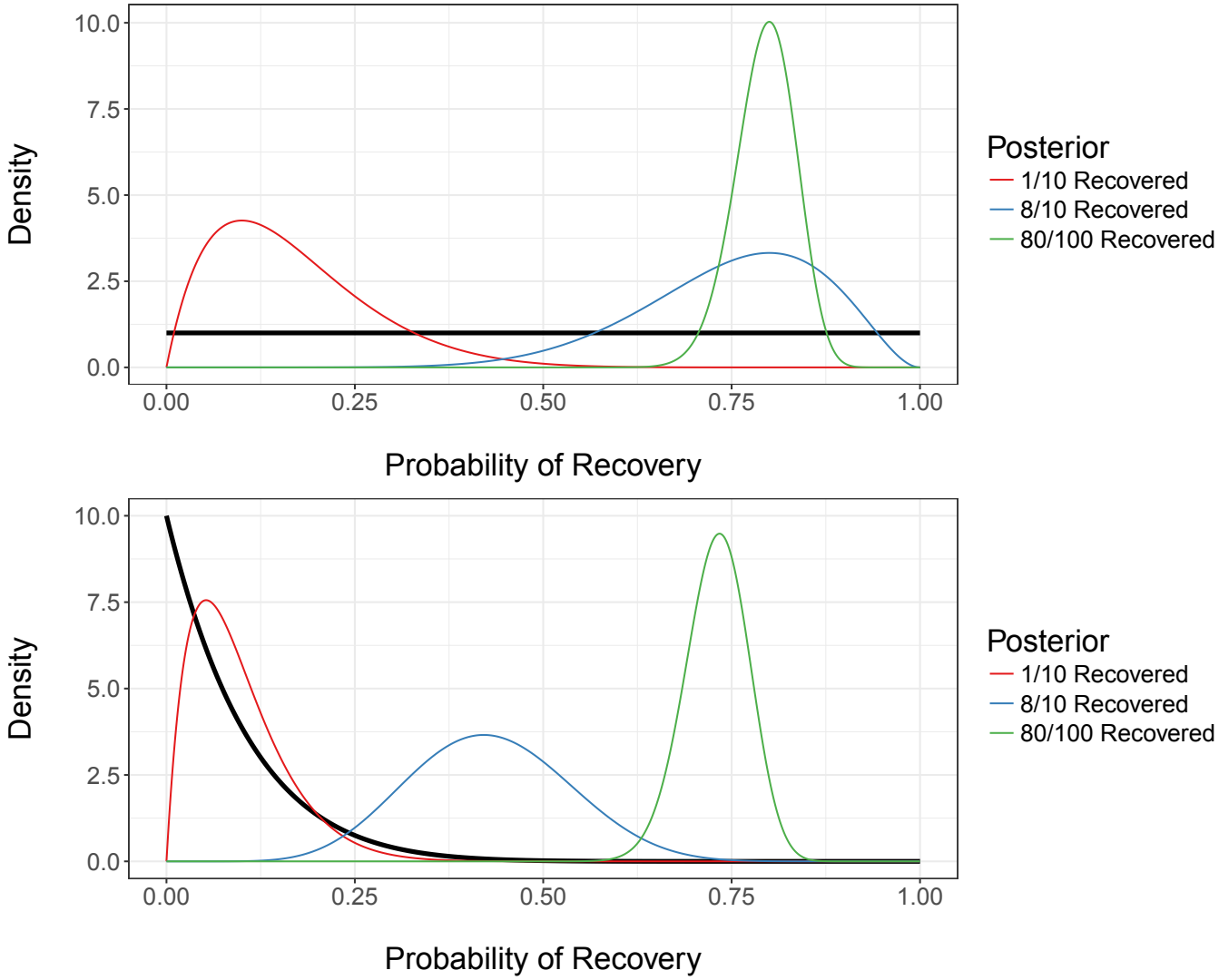

Fig. 1. Posterior probability of recovery for three datasets. The top figure is for a flat prior distribution and the bottom figure is for an informative prior distribution. The solid black line is the prior distribution and the colored lines are the posterior distributions.

of treatment. These 10 observations are the data. Bayesian methods combine the prior probability of recovery, which was evenly distributed from 0 to 1, with the data-based probability of recovery, which is $1/10 = 0.1$. Combining the prior and the data (see Equation 1) produces what is called the posterior distribution because it is created after seeing the data. The red line in the top panel of Fig. 1 is the posterior distribution if only 1 out of 10 participants recovered. The probability of recovery is no longer flat but is peaked at 0.1 and nearly all of the probability is below 0.5. The top panel of Fig. 1 also shows the posterior distribution for two other scenarios, 8 out of 10 recovered and 80 out of 100 recovered. All posterior distributions in Fig. 1 share the same flat prior distribution but differ with respect to the data. Comparing the posterior for 1 out of 10 and the posterior for 8 out of 10 demonstrates how the posterior is heavily influenced by the data, especially when the prior information is not informative.

The final posterior distribution in Fig. 1 is for a scenario where 80 out of 100 participants recovered. Just like when 8 out of 10 recovered, the posterior for 80 out of 100 is centered at a probability of recovery of 0.8. However, with 100 participants in the sample, the posterior distribution is less variable than with just 10 participants. That is, with 100 participants, there is more information about the parameter so the posterior will be narrower. Although inferences are just as valid with 10 as 100, inferences will be more certain with 100.

Now suppose we evaluate a new treatment for depression that uses magnets placed at key parts of the body to regulate magnetic fields in the body. We do not believe that the depression has anything to do with magnetic fields so we predict that the probability of recovery is most likely close to zero. To accommodate a placebo response to the magnet treatment we choose a prior where most of the weight is below 0.25 (see the solid black line in the bottom panel of Fig. 1). This type of prior is called an informative prior because it contributes more information to the posterior distribution than the non-informative, flat prior in the top panel of Fig. 1.

The impact of the prior can be seen by comparing the panels of Fig. 1, given that the datasets are the same in both panels. When the datasets are small, such as 10 participants, the prior has a large influence. With a flat prior, when 8 out of 10 participants recover, the posterior is wide and centered at 0.8. With an informative prior and the same dataset, the posterior is still wide but is centered at about 0.4. The prior pulled the effect down, which is an example of the statistical process called shrinkage (Gelman & Hill, 2007). The impact of the prior is much smaller when the dataset is large, as

evidenced by the fact that the posterior distribution associated with the largest dataset is only slightly affected by the shape of the prior distribution (compare the green distributions to the red and blue distributions in each panel).

The preceding discussion and Fig. 1 conceptually illustrate how Bayesian inference works. The mathematical method for combining priors and data is Bayes' theorem:

$$p(\text{parameters}|\text{data}) = \frac{p(\text{data}|\text{parameters})p(\text{parameters})}{p(\text{data})} \quad (1)$$

where $p(\text{parameters})$ is the prior and $p(\text{parameters}|\text{data})$ is the posterior (read as the probability of the parameters given the data). We have not yet explicitly discussed the other two parts of Bayes' theorem.

The probability of the data given the parameters in the model, $p(\text{data}|\text{parameters})$, is the likelihood. Researchers will choose a probability distribution that best describes their data. In the case of recovered versus not-recovered, which is a binary variable, a common choice for the likelihood is the binomial distribution. For regression, ANOVA, and multilevel models, the likelihood is a normal distribution. For count data, the a common choice is a Poisson or negative-binomial distribution. Although priors get a lot of attention in Bayesian analysis, due to the fact that explicit priors are unique to Bayesian methods, choosing a reasonable likelihood is critical. As we shall see, Bayesian methods require that data analysts make the likelihood explicit, which we see as a major advantage of Bayesian methods because it forces researchers to think carefully about their models.

The denominator of Bayes' theorem, p(data), is the probability of the data. In many applications this probability does not need to be specified. Indeed, for the models we discuss in this paper p(data) can be safely ignored because the analysis methods only require that one specify the likelihood and the priors. Specifically, the methods take advantage of the fact that the posterior distribution is proportional to the likelihood times the prior(s):

$$p(\text{parameters}|\text{data}) \propto p(\text{data}|\text{parameters})p(\text{parameters}) \quad (2)$$

A full discussion of why this is the case is beyond the scope of this introductory article. More complete introductions to Bayes' theorem are available in McElreath (2016) and Kruschke (2015) and a more advanced discussion is available in Gelman et al. (2014). What is critical to understand at this point is that the posterior distribution of a parameter is what is produced by a Bayesian analysis and is a function of the specified prior, the likelihood, and the data.

In fact, the posterior distribution is a major difference between Bayesian and frequentist models. If we use frequentist methods to compute the probability of recovery for a new treatment, the end product is a single number or point estimate. We could also report a standard error or a confidence interval to give a sense of the stability of the probability over repeated samples. In contrast, if we compute this probability from a Bayesian perspective, the end product is a distribution or density, such as those shown in Fig. 1. Because distributions are not always easily summarized by one or two numbers, it is common to graph the posterior distribution. Graphs display a wealth of information (e.g., the most probable values, the degree of skewness, and the degree of uncertainty) in a fairly compact form. We can summarize the posterior using the mean, median, or mode for central tendency of the posterior distribution and the standard deviation or quantiles for dispersion. As sample sizes get large, the mean of a posterior distribution and a frequentist estimate will converge (Gelman et al., 2014). However, point estimates are only summaries of the posterior distribution and interpreting models via tables of point estimates only throws out some valuable information (e.g., uncertainty of the parameters or the relationship between the parameters). This is especially true as models become more complex or include difficult to interpret parameters (e.g., interactions; McElreath, 2016). Consequently, our forthcoming example analysis emphasizes the need to visualize results to aid inferences.

1.2. Markov Chain Monte Carlo

Producing the posterior distribution in a Bayesian analysis is different than producing a point estimate in a frequentist analysis. In frequentist models, point estimates are produced using various algorithms such as least-squares (e.g., regression and ANOVA) or maximum-likelihood (e.g., multilevel models). In Bayesian methods, the method for learning about the posterior distribution depends upon whether the posterior distribution conforms to a known probability distribution. That is, if the posterior distribution is a normal distribution, then one can use what is known about normal distributions to make inferences. The posterior distributions in Fig. 1 correspond to beta distributions (see Lindgren, 1993), so computing them and making inferences from them is straightforward because the mean, variance, and other aspects of the beta distribution are known. Most statistical programs have functions for many probability distributions, making it simple to learn about posterior distributions in cases where the posterior is a known distribution.

Unfortunately, posterior distributions do not typically conform to known probability distributions. In these cases, one can use Markov Chain Monte Carlo (MCMC) methods to simulate random draws from the posterior distribution. The vast majority of papers that use Bayesian methods use MCMC algorithms in their analyses because these algorithms are robust and flexible and can be applied to many analysis problems psychology.

It may seem odd that simulations are at the heart of Bayesian models, especially for those who learned statistics by applying extensive formulas to compute correlations, *t*-tests, and ANOVAs. However, simulation is powerful and we can learn a lot about a distribution from simulation. As Jackman (2009) notes: "Modern Bayesian computation makes extensive use of a simple idea, *the Monte Carlo principle*: anything we want to know about a random variable θ can be learned by sampling many times from $f(\theta)$, the density of θ." (p. 134). For example, in a bivariate regression, θ would be a regression slope (β) and $f(\theta)$ the posterior distribution of β. If we can simulate from $f(\theta)$—if we can take random draws from $f(\theta)$—we can use the random draws to learn about the posterior distribution (cf. Kruschke, 2015, pp. 143–145). For example, we can compute the mean of the random draws to obtain the expected value or the 2.5% and 97.5% to obtain the limits for the middle 95% of the distribution.

For example, suppose we wanted to learn about a parameter θ and we only knew that its posterior is a normal distribution with a mean of 10 and an unknown standard deviation.

$$\theta \sim N(10, ?) \quad (3)$$

Although we do not know the standard deviation of the posterior, suppose we had 5000 random draws from this posterior (obtained using Monte Carlo methods). We can compute the standard deviation of those 5000 draws to estimate the standard deviation. Likewise, we could plot the 5000 draws to visualize the shape of the posterior distribution. The precision of these estimates is only limited by the number of draws from the distribution, but in many instances 1000–2000 draws from distribution will provide an excellent approximation. Of course, if we know that the posterior

distribution is normal, we do not need simulation methods. However, simulations work in this simple case and generalize to more complicated cases where the posterior does not conform to a known probability distribution.

Simulating from a normal distribution in statistical software is trivial as nearly all statistical programs include functions for producing random draws from a normal distribution as well as many other distributions. Simulating from an unknown distribution is more difficult and that is where MCMC comes in. MCMC refers to algorithms that combine the logic of simulation (Monte Carlo methods) with a mathematical random process called Markov chains (Jackman, 2009). It is not critical that you understand the details of Markov chains to follow the rest of the paper. Consequently, we leave the details of how Markov chains are defined and their assumptions to other sources. Both McElreath (2016, see chapter 8) and Kruschke (2015, see chapter 7) provide accessible, conceptual introductions to the logic of Markov chains. Jackman (2009, see chapters 4–5) provides a more technical discussion of Markov chains and how they are combined with Monte Carlo methods in MCMC algorithms.

MCMC algorithms allow us to draw a random value from a known distribution (e.g., a normal distribution), which is called a proposal value, and evaluate whether the proposal came from the posterior distribution. This evaluation requires that you know the form of the posterior density up to a constant, which means that we need to know the likelihood for the data as well as the prior for each parameter (see Equation 2). MCMC algorithms require thousands and sometime tens of thousands of evaluations of the proposal values. The algorithms require that we decide either yes or no for each proposal. If yes, we keep the proposal value as a valid draw from the posterior. If no, the proposal value is discarded and the most recent accepted values replaces the proposal (see Kruschke, 2015, pp. 149–156). This produces a sequence of draws from the posterior distribution—a Markov chain. Baldwin and Fellingham (2013) include a detailed example of a MCMC sampler, including a text description of the algorithm as well as computer code and simulated data for implementing the sampler.

When examining a chain, there will be sequences of iterations where the value of the parameter changes each iteration. That is, where the proposal is accepted several times in a row. Likewise, there will be sequences where the value of the parameter does not change—where the proposal is rejected several times in a row and is replaced by the most recently accepted value. Ideally, a chain should change and not get stuck on the same value for many iterations. If a chain does get stuck, that may indicate a problem with the model. There are a number of diagnostic measures to help determine whether a chain has converged on the posterior distribution, which we discuss below.

There are many kinds of MCMC algorithms available. The algorithms differ with respect to their generality (i.e., how many kinds of models they can accommodate) and efficiency (i.e., the number of iterations needed to obtain useful draws from the posterior distribution). Commonly used algorithms include Gibbs, Metropolis-Hastings, slice sampling, and Hamiltonian Monte Carlo (Gelman et al., 2014). Computer programs can now implement a variety of MCMC algorithms for many kinds of models. Some programs focus on specific types of models, such as structural equation modeling (e.g. Muthén & Muthén, 1998–2015). Other programs are general purpose, allowing the user to fit a huge variety of models. Examples of this type of program are: WinBugs (Spiegelhalter, Thomas, Best, & Lunn, 2003), JAGS (Plummer, 2015), Stan (Stan Development Team, 2016), PROC MCMC in SAS (SAS Institute Inc., 2013), and `bayesmh` in Stata (Stata Corp, 2015b) A recently developed alternative is JASP (JASP Team, 2016), which has a graphical user interface and can estimate commonly used models such as t-tests, ANOVA, and linear regression using Bayesian methods.

In this paper, we focus on estimating models using the Stan (version 2.11.0) software accessed via R (version 3.3.2). We also illustrate the use an R package called `brms` (Bayesian Regression Models using Stan; version 0.9.1.9000), which is a "front-end" to Stan and allows the user to quickly fit models without needing to write the full Stan syntax (Buerkner, 2016). First, Stan uses efficient sampling algorithms (Hamiltonian Monte Carlo and No-U-Turn-Sampler; Stan Development Team, 2016), which can be important as models become more complex. Second, it is open source and free. Third, it can be accessed using a number of programs (e.g., R, Python, Matlab, and Stata). Fourth, it is actively developed and improved, including development of data visualization and model evaluation programs (Gabry, 2016). Fifth, Stan is cross-platform and runs on all major operating systems. Disadvantages of Stan include that it has its own modeling language. The language is R-like, but still has some "start-up" costs to get up and running. Stan is a C++ program, which means that the programs must be compiled prior to running. Consequently, Stan requires a C++ compiler and the developers provide straightforward installation instructions. Nevertheless, installation is more complicated than a simple button-push. Although this can slow down the modeling fitting process, the penalty is usually minimal. Nevertheless, being able to use a compiled language is useful because it tends to speed up the sampling.

2. Example analysis

2.1. Data

As example data, we use a component of the scalp-recorded event-related potential (ERP) called the error-related negativity (ERN). ERPs are collected from the scalp-recorded electroencephalogram (EEG) and consist of averaged waveforms that depict changes in a large population of synchronously active neurons. ERPs can be time-locked to specific stimuli (e.g., the presentation of a picture or sound) or to a participant response (e.g., an erroneous or correct button press). ERPs tend to vary as a function of the strength of the brain's electrical response and differ based on the specific characteristics of the task (i.e., external stimulation) and person (i.e., internal characteristics and traits). For a review and tutorial on ERPs please see Luck (2014). ERP waveforms consist of multiple peaks and troughs that represent the summation of underlying (Luck, 2014). One such component, the ERN, is a negative deflection in the ERP that occurs within 100 ms of a response and is larger following errors than correct responses. Whereas the precise functional significance of the ERN remains a matter of active debate, theories suggest it represents the competition between correct- and error-responses, an emotional response to errors, the motivational salience of an error, or a reinforcement learning signal to improve subsequent performance (see Larson, Clayson, & Clawson, 2014, for a review). The ERN varies with some forms of psychopathology diagnosis and symptom severity. For example, a recent meta-analysis indicates that ERN amplitude is larger (i.e., more negative) in individuals with obsessive-compulsive disorder (OCD) or other anxiety disorders such as generalized anxiety disorder (GAD) or post-traumatic stress disorder (PTSD) compared to individuals without psychopathology (Moser, Moran, Kneip, Schroder, & Larson, 2016). Higher levels of trait anxiety are also consistently related to more negative ERN amplitude (Moser, Moran, Schroder, Donnellan, & Yeung, 2013; Moser et al., 2016). Amplitude of the ERN has even been proposed as a potential neural endophenotype identifying anxiety-related traits and pathology (Kujawa et al., 2016; Olvet & Hajcak, 2008). In contrast, amplitude of the ERN is consistently smaller in individuals with serious

mental illness relative to non-pathology controls (Llerena, Wynn, Hajcak, Green, & Horan, 2016). The relationship between ERN amplitude and other forms of psychopathology, such as depression or autism, is less clear (Hupen, Groen, Gaastra, Tucha, & Tucha, 2016; Moran, Schroder, Kneip, & Moser, 2017). The sample used in this tutorial consists of 81 (22 men and 59 women) people diagnosed by a physician, psychologist, or psychiatrist, with either a depressive or anxiety disorder before presenting to the lab. Diagnoses according to DSM-IV criteria were confirmed upon arrival to the lab using the Mini-International Neuropsychiatric Inventory (MINI; Sheehan et al., 1998). The MINI has strong concordance rates with the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID), but is more rapid to administer. Exclusion criteria included medication changes within the last two months, diagnosis of psychotic or bipolar disorders, history of substance abuse, reported learning disability or attention deficit/hyperactivity disorder, neurological disease, and left-handedness (Larson, Clawson, Clayson, & Baldwin, 2013). Participants completed a modified version of the Eriksen flanker task (Eriksen & Eriksen, 1974) to elicit the ERN. In the flanker task, participants are asked to respond to the direction of the center arrow as quickly and accurately as possible while ignoring flanking arrows. See Larson et al. (2014) for a review of flanker-related electrophysiology. Electroencephalogram (EEG) was recorded from 128 scalp sites using a geodesic sensor net and Electrical Geodesics, Inc. amplifier. For the sake of brevity, please see Larson et al. (2014) for full details of the flanker task and EEG recording/processing and Larson et al. (2016) for the same procedures for ERN extraction. All participants had a minimum of six useable ERPs, though dependability of the waveforms increased with higher trial numbers (Baldwin, Larson, & Clayson, 2015; Clayson, Baldwin, & Larson, 2013). Recent studies suggest the ERN is related to trait anxiety (Moser et al., 2013, 2016), similar to what is measured by the State-Trait Anxiety Inventory (STAI; Spielberger, Gorusch, Lushene, Vagg, & Jacobs, 1983). Given the proposed relationship between ERN amplitude and the trait anxiety, some researchers are suggesting the ERN be used in diagnostic and treatment studies as a potential means for identifying the presence/absence of some forms of trait-like psychopathology and determining if there is neural change following treatment (e.g., Kujawa et al., 2016). Given the current exploration of the ERN as a neural diagnostic and treatment marker and its relationship with some forms of psychopathology, we believed it would be a good test case to examine in a sample of individuals with psychopathology through a Bayesian lens. Based on the continuous nature of ERN amplitude and STAI scores, bivariate regression is an appropriate model to test the relationship. The STAI is a 40-item self-report measure of anxiety symptoms that was developed to measure a patient's anxiety in their everyday life (trait subscale; 20 items), as well as anxiety at the time of the evaluation (state subscale; 20 items). Test-retest correlations range from 0.73 to 0.86 and internal consistency coefficients from 0.89 to 0.92 (Spielberger et al., 1983).

2.2. Model

The model is a bivariate regression, where the ERN is predicted from the trait subscale of the STAI, hereafter referred to as anxiety. For all models, we standardized anxiety prior to including it in the model. This model is:

$$\text{ERN}_i = \beta_0 + \beta_1 \text{Anxiety}_i + \varepsilon_i \tag{4}$$

The residuals, ε_i, are normally distributed with a mean of 0 and an unknown standard deviation.

$$\varepsilon_i \sim N(0, \sigma) \tag{5}$$

To use Bayesian methods for this model, we must define (a) a likelihood for the data and (b) priors for all parameters in the model. As discussed previously, a likelihood is a probability distribution to describe the data. For a linear regression model such Equation 4, the normal distribution is the most common choice. One could fit a robust regression model that uses the *t*-distribution for the likelihood (cf. Kruschke, 2013), but we use the normal here to keep it most consistent with least-squares regression. The likelihood for the data can be written as:

$$\text{ERN}_i \sim N(\mu_i, \sigma) \tag{6}$$

which says that the ERN value for the *i*th person follows a normal (Gaussian) distribution with a mean = μ_i and a standard deviation = σ. The mean, also called the expected value or predicted value, of the ERN is equal to:

$$\mu_i = \beta_0 + \beta_1 \text{Anxiety}_i \tag{7}$$

The standard deviation in Equation 6 is the same standard deviation as in Equation 5. Thus, we can rewrite Equation 6 as:

$$\text{ERN}_i \sim N(\beta_0 + \beta_1 \text{Anxiety}_i, \sigma) \tag{8}$$

Equation 8 indicates that this model has three unique parameters to estimate—β_0, β_1, and σ. Note that μ_i is fully defined by the β_0 and β_1 and is not estimated directly but instead computed after β_0 and β_1 are estimated. Each of these parameters needs a prior distribution. A common choice for regression parameters is a normal distribution because regression parameters can be both positive and negative. We set the prior for β_0 and β_1 to:

$$\beta_0, \beta_1 \sim N(0, 3) \tag{9}$$

A normal distribution with mean at 0 gives equal probability to positive and negative coefficients and a normal distribution with a standard deviation of 3 means that we believe that 95% of the coefficients should be between −6 and 6 (i.e., ±2 standard deviations). Given the typical range of the ERN, this prior is fairly flat and non-informative. By way of comparison, least squares regression can be thought of as having a uniform prior (i.e., all values have the same probability) from $-\infty$ to ∞ for regression coefficients (Gelman & Hill, 2007, p. 346). Thus, our prior is more informative, and more reasonable, given the constraints on the range of the ERN as well as possible anxiety-ERN relationships.

A common choice for a prior on a standard deviation, such as σ, is a half-Cauchy distribution (Gelman, 2006). The full Cauchy distribution is not appropriate for a standard deviation because the Cauchy distribution covers negative values. Thus, a half-Cauchy distribution is needed so that the standard deviation can only take on positive values. The prior for σ was:

$$\sigma \sim \text{half-Cauchy}(0, 2.5) \tag{10}$$

This prior gives the most weight to standard deviations less than 5, but does allow for large estimates if needed. In least squares regression, the implicit prior on σ is a uniform prior from 0 to ∞. Other common choices for priors on standard deviations or variances are uniform priors with a lower-bound of 0 or inverse-gamma priors (because the inverse-gamma distribution is not defined for negative numbers). These priors can be problematic when sample sizes are small because they tend to put too much weight on extreme values (Baldwin & Fellingham, 2013; Gelman, 2006). Consequently, we recommend the half-Cauchy prior,

which can easily be implemented in Stan.

The full regression model, including the likelihood and the prior, is:

$$\begin{aligned}
&\text{ERN}_i \sim N(\mu_i, \sigma)\\
&\mu_i = \beta_0 + \beta_1 \text{Anxiety}_i\\
&\beta_0, \beta_1 \sim N(0, 3)\\
&\sigma \sim \text{half-Cauchy}(0, 2.5)
\end{aligned}$$

Fig. 2 demonstrates the relationship between the data, likelihood, parameters, and priors. The data are represented by the histogram. We use the normal distribution as the likelihood for the data. Moving up Fig. 2 shows that the mean of the likelihood is a function of the regression coefficients, which have normally distributed prior distributions. Likewise, the standard deviation of the likelihood has a half-Cauchy prior distribution. Fig. 2 illustrates that any unknown parameter—μ_i, σ, β_0, and β_1—either needs a prior distribution or needs to have a deterministic relationship with other parameters. In this model, μ_i is completely determined by β_0 and β_1, which means it does not need a prior distribution—uncertainty in β_0 and β_1 will get propagated to μ_i.

2.3. Results

We fit the regression model as described using Stan and `brms`. We sampled four chains for each parameter and each chain had 2000 draws from the posterior. The first 1000 draws in each chain are part of an adaptation phase that Stan uses to tune the sampling algorithm and are discarded (Stan Development Team, 2016). Thus, the analysis provides 4000 draws from the posterior (1000 from each chain) that can be used to draw inferences. There are no fixed rules regarding the number of chains. It is common to use three or four chains and in our experience that has been suitable for our analyses. Furthermore, chains can be run in parallel with multiprocessor computers, which can speed up the analysis.

2.3.1. Convergence

Stan, as well as other software packages, are good at ensuring

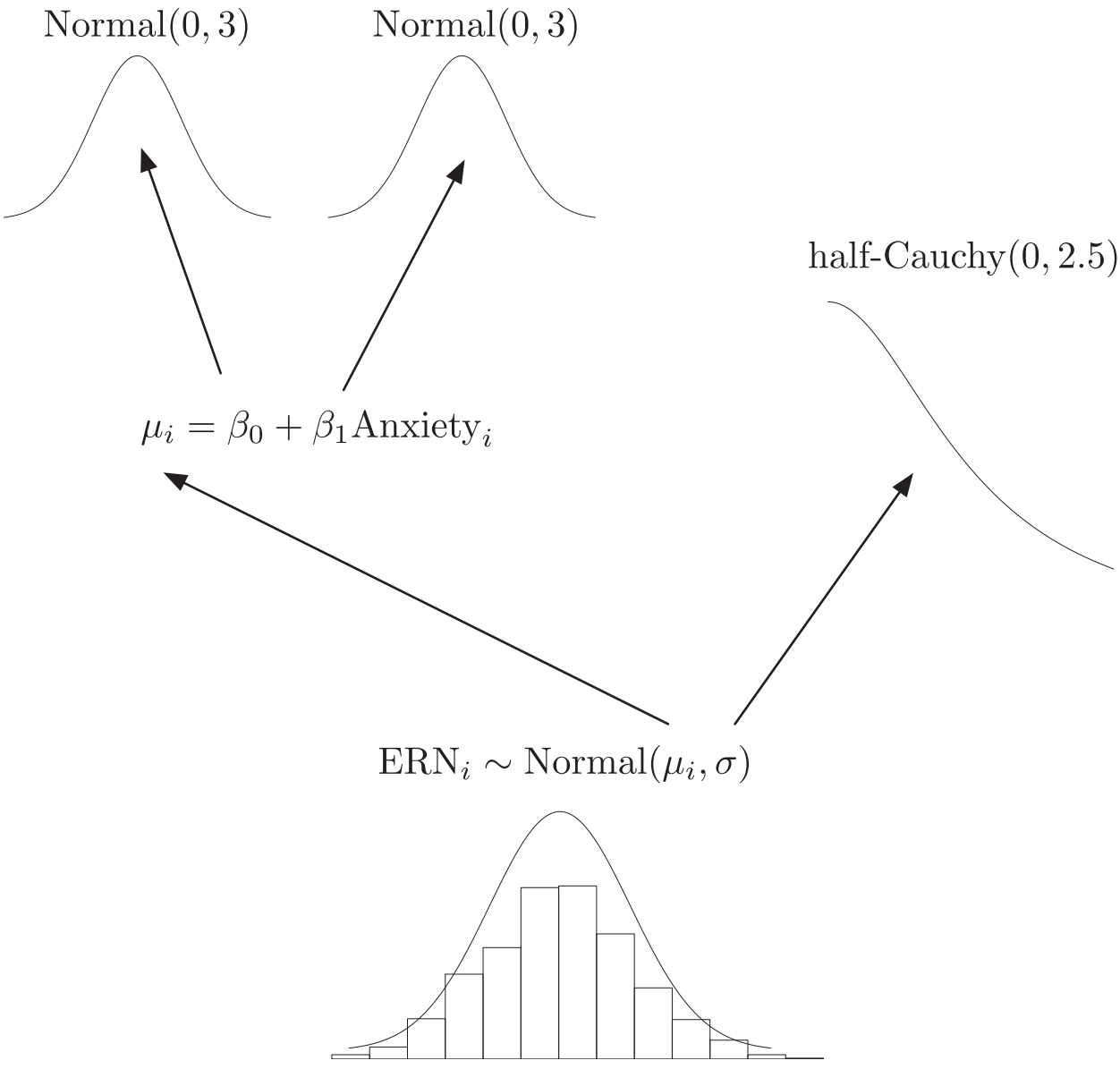

Fig. 2. Graphical representation of the full regression model.

that the samples are from the posterior, especially if we supply it with appropriate likelihoods and priors. However, we have to inspect the chains to ensure that the MCMC chain has converged, which means that the samples came from the posterior. In our experience non-convergence occurs because of an error in specifying the model (e.g., using the wrong type of prior) or because the model is too complex for the data (e.g., trying to fit too many random effects in a multilevel model).

A common, straightforward method for assessing convergence is trace plots. A trace plot is a line plot where the x-axis is the iteration and the y-axis is the sampled parameter value. Fig. 3 includes trace plots from the `brms` package for β_0, β_1, and σ. McElreath (2016) notes that trace plots should have two properties: stationarity and good mixing (p. 253). Stationarity means that the samples stay within the posterior distribution and are representative of the posterior distribution (Kruschke, 2015). That is, the trace plot should not wander outside the posterior distribution, but should appear to stay within the same parameter space across iterations and chains. Good mixing means that within a chain, adjoining samples should not be correlated with one another. Rather, the trace plot should bounce around the posterior, moving up and down. This means that the chain is gathering samples from all parts of the posterior or fully exploring the posterior distribution. All the trace plots in Fig. 3 are stationary and have good mixing (see McElreath, 2016, p. 258 and p. 262 for examples of problematic trace plots).

There were four separate chains in the analysis. MCMC is a stochastic process, meaning that the path a chain takes through the posterior should be random. Starting a MCMC chain requires a starting-value, which is often randomly chosen. Using multiple

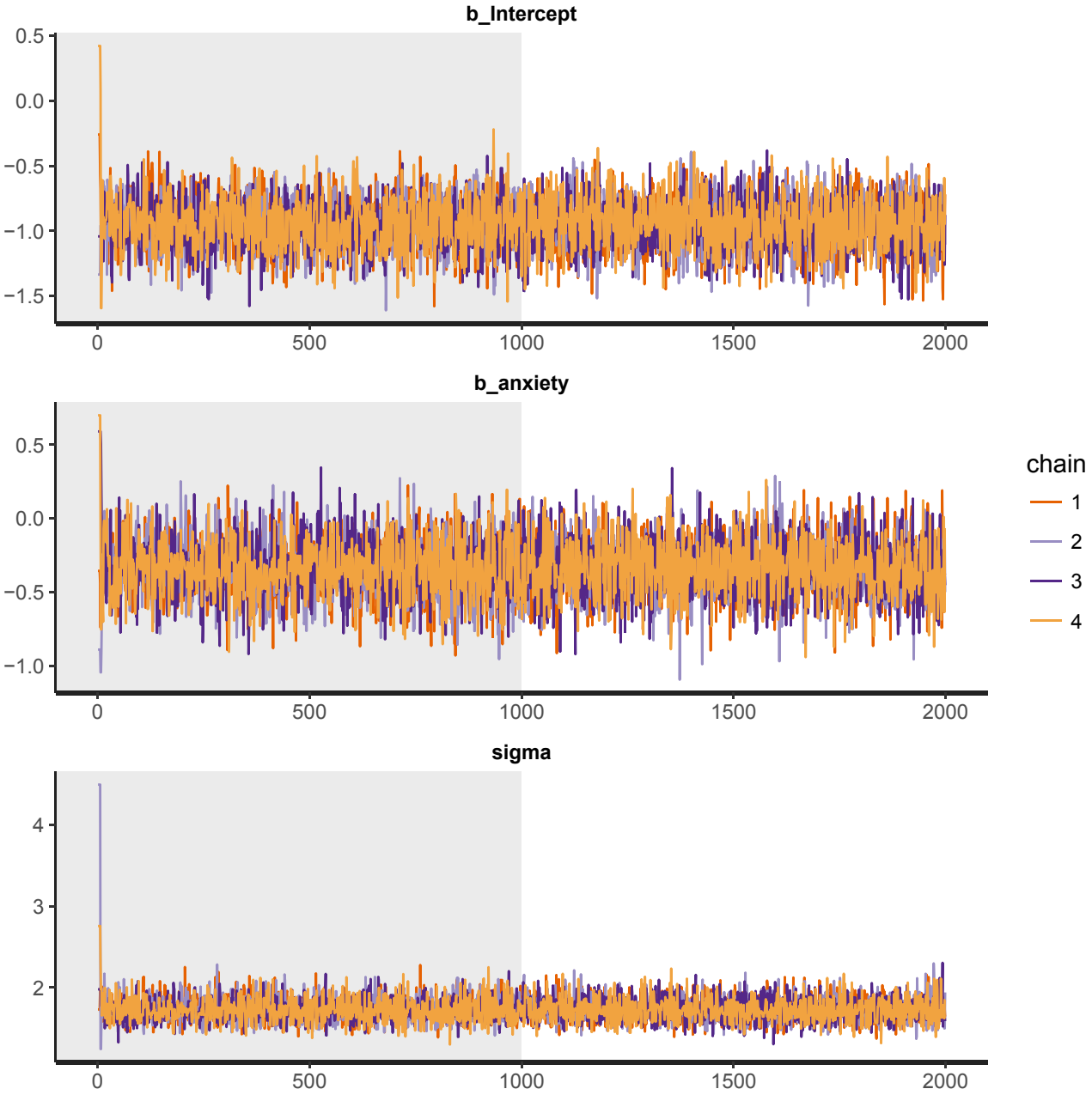

Fig. 3. Traceplot and posterior density for the bivariate regression parameters.

Table 1
Posterior summary and convergence statistics for regression models.

	Model 1				Model 2				Model 3			
Coefficients	Post. Mean	95% CI	ESS	$\widehat{R}$	Post. Mean	95% CI	ESS	$\widehat{R}$	Post. Mean	95% CI	ESS	$\widehat{R}$
Intercept (β_0)	−0.96	−1.34; −0.58	2847	1	−1.13	−1.85; −0.40	2771	1	−1.20	−1.87;-0.51	3041	1
Anxiety (β_1)	−0.33	−0.70; 0.04	3060	1	−0.34	−0.71; 0.03	3053	1	−0.78	−1.30; −0.25	2668	1
Sex (β_2)					0.24	−0.57; 1.09	2667	1	0.30	−0.49; 1.08	3220	1
Anx × Sex (β_3)									0.78	0.10; 1.48	2676	1
Res. SD (σ)	1.74	1.49; 2.03	2620	1	1.74	1.49; 2.04	2541	1	1.7	1.47; 1.98	2836	1
Model Fit												
N	81				81				81			
WAIC	320.72				322.63				319.57			
LOO	320.76				322.73				319.64			

Note. Post. Mean = Posterior mean; 95% CI = 95% credible interval; ESS = Effective sample size; $\widehat{R}$ = Gelman-Rubin Statistic.

chains allows us to examine whether the chains converge on the same space, even though they start at different, randomly chosen places. Fig. 3 illustrates multiple chains for a given parameter converging on the same space even with unique starting values. The Gelman-Rubin statistic ($\widehat{R}$) provides a numerical method for examining whether chains converge on the same posterior (Gelman & Hill, 2007). The Gelman-Rubin statistic compares the variability between chains to variability within chains. If those quantities are roughly equal, meaning $\widehat{R}$ is ≈ 1, then the chains are said to have converged. If there is more variability between chains than within chains, the chains will be sampling values from different places and convergence is in question. The `brms` package, as well as most R packages that interface with Stan, provides $\widehat{R}$ for all parameters. Table 1 indicates that $\widehat{R}$ was 1.0 for all parameters. Once you have determined that chains have converged, you can combine the draws from each chain as all the draws will be from the posterior.

Convergence of the chain tells us whether we have samples from the posterior. It is also important that we have enough samples from the posterior that summaries (e.g., means, 95% quantiles) are accurate and stable (Kruschke, 2015). Ideally, the 4000 draws provide 4000 unique pieces of information about the posterior. Unfortunately, we actually have less information than that because the draws are correlated. That is, there is a relationship between a sampled parameter and the sampled parameter that came before it (right before it or even up to 100 iterations before it). This correlation is called autocorrelation and we want autocorrelation to be small as that indicates we need shorter chains to have a good representation of the posterior. Table 1 presents the effective sample size (ESS) for each parameter, which estimates the number of independent samples we have from the posterior (Kruschke, 2015). There are no hard-and-fast rules for how big the ESS needs to be for stable estimates. However, the 2000+ samples we have for each of our parameters is going to do the job for most purposes. As precision become more important, more effective samples are needed (Jackman, 2009).

Evaluating convergence and efficiency of chains can be involved and we have shown a couple of methods for doing so. Other tools and statistics are available to help researchers ensure their MCMC chains are sufficient for their purposes. See Kruschke (2015, Chapter 7), McElreath (2016, Chapter 8), Jackman (2009, Chapter 6), and Gelman et al. (2014, Chapter 11) for more details.

2.3.2. Understanding the posterior

Having established convergence and determined that we have a reasonable number of samples, we can proceed with using the simulations to make inferences from the model. An excellent place to start is to plot the posterior distribution as a density plot or a histogram. These plots provide a view of the entire distribution, including symmetry, central tendency, and dispersion. Most importantly, it provides a compact overview of the probability of particular values of the parameters. Fig. 4 shows the posterior of the anxiety-ERN slope (β_1).

We can also summarize the distribution. The mean of the simulation samples provides an estimate of the mean of the posterior. We can also compute the median, mode, standard deviation, variance, and quantiles of the samples. Table 1 provides the mean, standard deviation, and the limits for the middle 95% of the posterior distribution. The mean slope for trait anxiety predicting the ERN is $\widehat{\beta_1} = -0.33$ (see the vertical black line on Fig. 4).

Although the mean estimates the most probable value of a parameter (in a symmetrical distribution at least), it is useful to summarize the uncertainty regarding the parameter so as not to place too much confidence in a noisy estimate. Consequently, it is common to provide an interval estimate, which is similar to a confidence interval in frequentist statistics. The interval estimates are often called credible intervals because the estimates provide the upper and lower limits for parameter values with at least a certain amount of probability or credibility. For the example, a 95% credible interval provides the upper and lower limits for the middle 95% of the distribution. That is, 95% of the posterior distribution falls between these two limits and the parameter values within the limits have a higher probability than those outside the limits (Kruschke, 2015). If the intervals are wide, then there is substantial uncertainty regarding the parameter. The shaded portion of Fig. 4 depicts the 95% credible interval for β_1, the slope for anxiety. The limits of this interval are $-0.70, 0.04$; therefore, the parameter values between these limits account for 95% of the posterior distribution[2].

Bayesian credible intervals indicate what parameter values have the most probability, given the data and priors. Frequentist confidence intervals are often incorrectly interpreted in this way. For example, 95% confidence intervals are often interpreted as meaning that there is a 95% probability that the true population value falls between the upper and lower limits. Unfortunately, frequentist intervals do not indicate the probable values of parameters. Instead, the correct interpretation of a 95% frequentist interval is: If we repeated this study many, many times, 95% of intervals will contain the population value. The 95% probability reflects the confidence in the method—the way the interval is constructed—not what values

[2] McElreath (2016) uses 89% interval estimates throughout his book to emphasize that arbitrariness of 95% interval estimates. We have used 95% intervals here but there is not anything particularly compelling about 95% versus 90% versus 80%. That is, what about psychological theories makes evidence from a 95% interval more useful than an 89% or 85% interval? Nothing we can imagine besides convention and habit. A benefit of Bayesian methods is that makes these conventions (e.g., using 95% intervals) a little more obviously conventions rather than necessities.

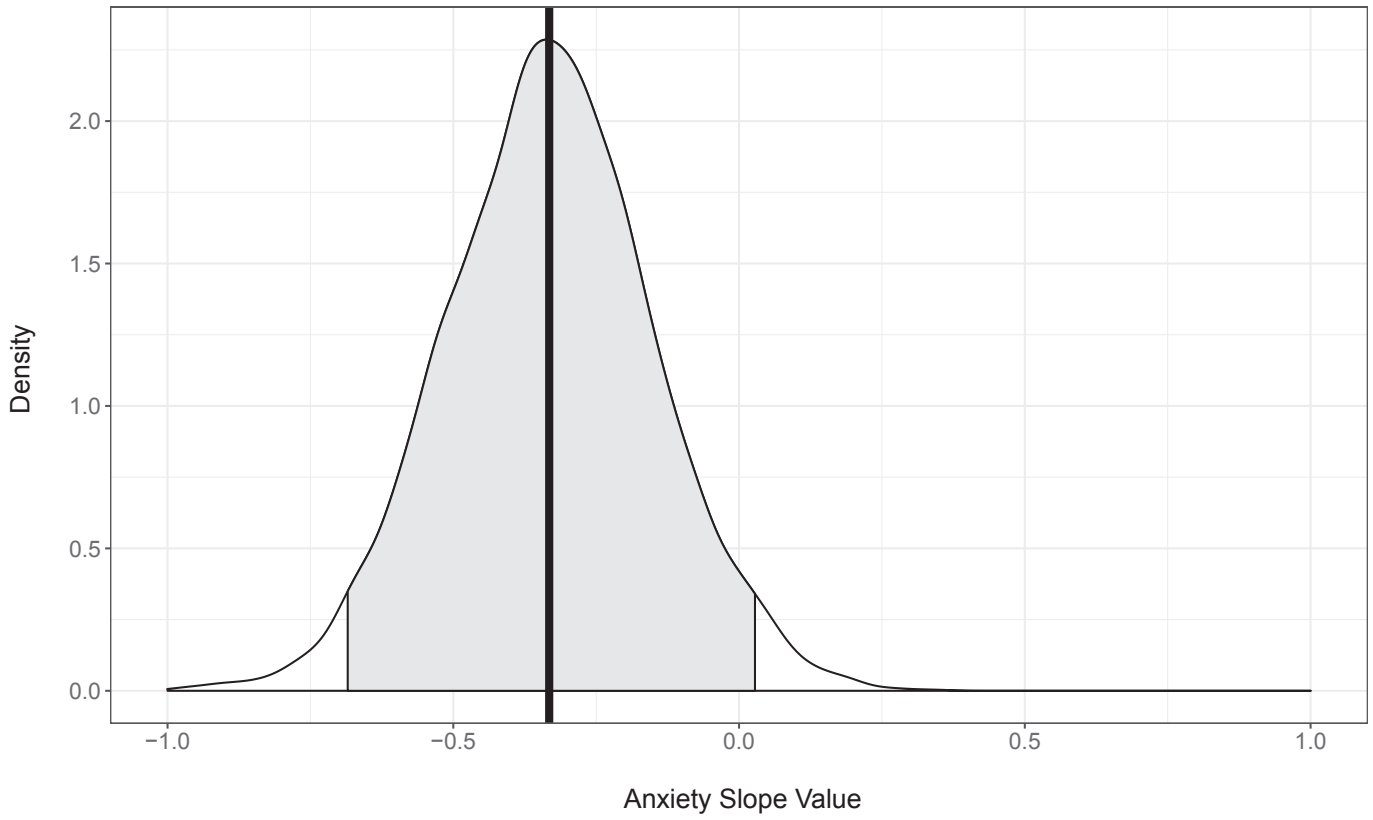

Fig. 4. Posterior density of the error-related negativity and anxiety slope. The vertical black line is the mean of the posterior and the shaded area 95% credible interval.

have the most credibility. As Morey, Hoekstra, Rouder, Lee, and Wagenmakers (2016) note:

> Confidence interval theory was developed to solve a very constrained problem: how can one construct a procedure that produces intervals containing the true parameter a fixed proportion of the time? Claims that confidence intervals yield an index of precision, that the values within them are plausible, and the confidence coefficient can be read as a measure of certainty that the interval contains the true value, are all fallacies and unjustified by confidence interval theory. (p. 118)

The incorrect interpretation of confidence intervals is commonplace both in research and clinical work (e.g., assessments). Bayesian credible intervals provide a useful alternative with a more intuitive, natural interpretation.

In addition to probing the uncertainty in a given coefficient, it can be useful to consider the uncertainty in the entire regression line. It is especially useful to compare that uncertainty to the data so that we can see that the data could be consistent with multiple lines, each with a different level of credibility. Figs. 5 and 6 provide scatterplots visualizing this uncertainty (cf. McElreath, 2016, Chapter 4). Fig. 5 presents the regression line based on the first 25 draws from the posterior. We chose the first 25, but one could easily randomly sample from the draws. In any case, the regression lines help us understand where the uncertainty in the lines comes from—uncertainty is substantial in the extremes of anxiety but less so toward the mean. Fig. 6 shows the mean regression line as well as the 95% credibility interval, which is a more complete summary of the posterior. If understanding relationships for high-levels of a construct are important, such as in psychopathology research, then the uncertainty in the extremes is critical scientifically and clinically. This highlights the need to move beyond just a focus on *p*-values and whether a relationship is positive or negative to a focus on what kinds of inferences are important theoretically and fully unpacking models to understand those inferences. Bayesian methods provide a means for doing so.

The line and dark-shaded area in Fig. 6 also represent the average expected value of the ERN at specific levels of anxiety (line) and the 95% credibility interval for the expected value (dark-shaded area). Thus, the average person reporting anxiety levels one standard deviation above the mean, is expected to have an ERN just below −1. However, due to uncertainty in the regression line (i.e., uncertainty in the intercept and slope), the 95% credible interval for the expected value varies between −0.8 and −1.8.

A posterior distribution of expected values can be created by using the draws from the posterior for β_0 and β_1. A 5-step procedure is:

1. Specify the form of the expected value: $\mu_i = E(\text{ERN})_i = \beta_0 + \beta_1 \text{Anxiety}_i$ (see Equations 6 and 8).
2. Select the first sampled value of β_0 and β_1 from the chain and substitute those specific values. In this case, the first sampled values were: $\beta_0 = -1.11$ and $\beta_1 = -0.57$.
3. Select a specific value of anxiety to produce an expected value. In this case, we selected Anxiety $= z = 1$.
4. Compute the expected value: $\mu_i = -1.11 + -0.57 \times 1 = -1.68$.
5. Repeat Steps 1–4 for all sampled values of β_0 and β_1.

The resulting values are draws from the posterior distribution of the expected values and the tools we have discussed for understanding posterior distributions can be used with this posterior. We can repeat these steps for other values of anxiety. The `brms` package can help automate creating posteriors for predicted values (see the supplemental material).

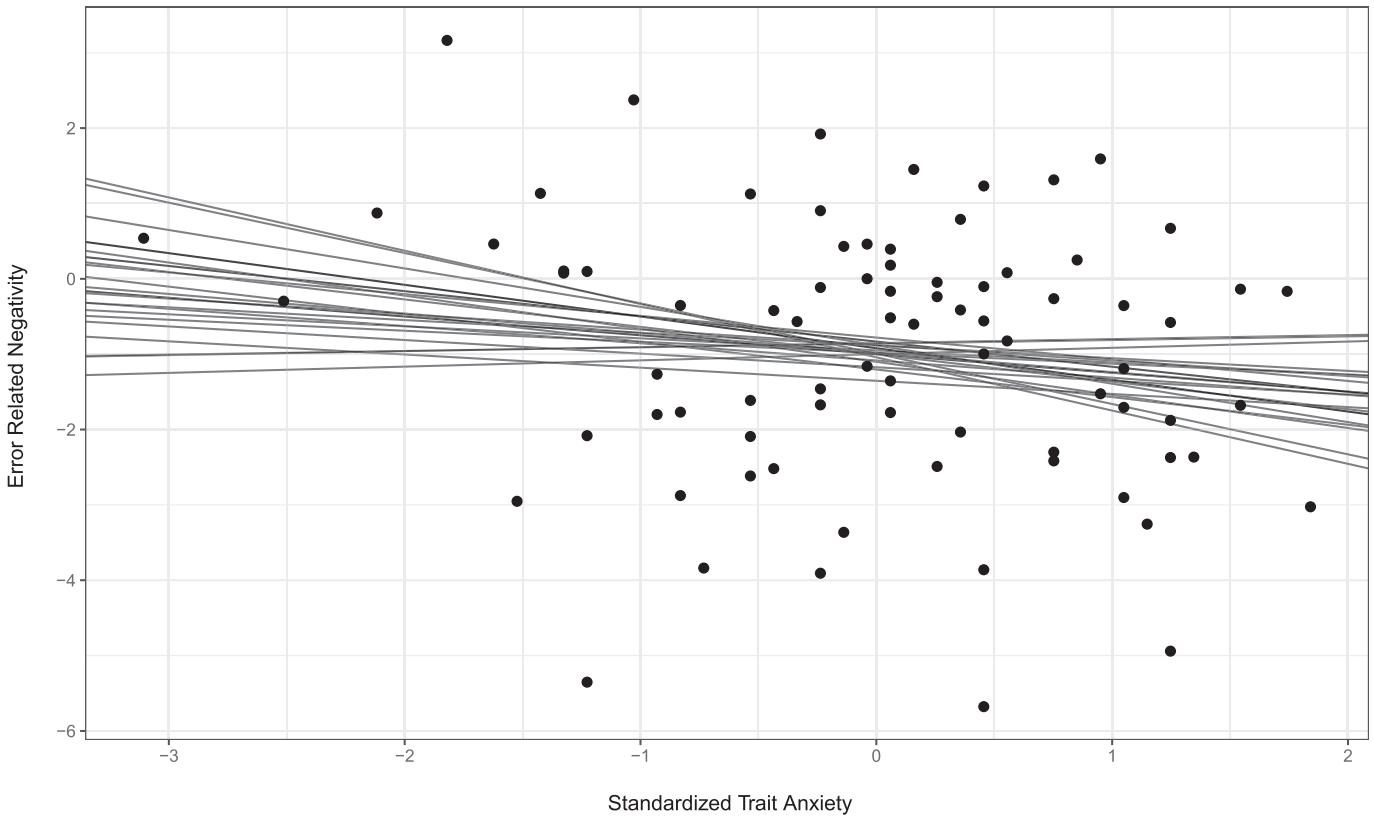

Fig. 5. Regression lines from the first 25 draws from the joint posterior of the intercept and slope.

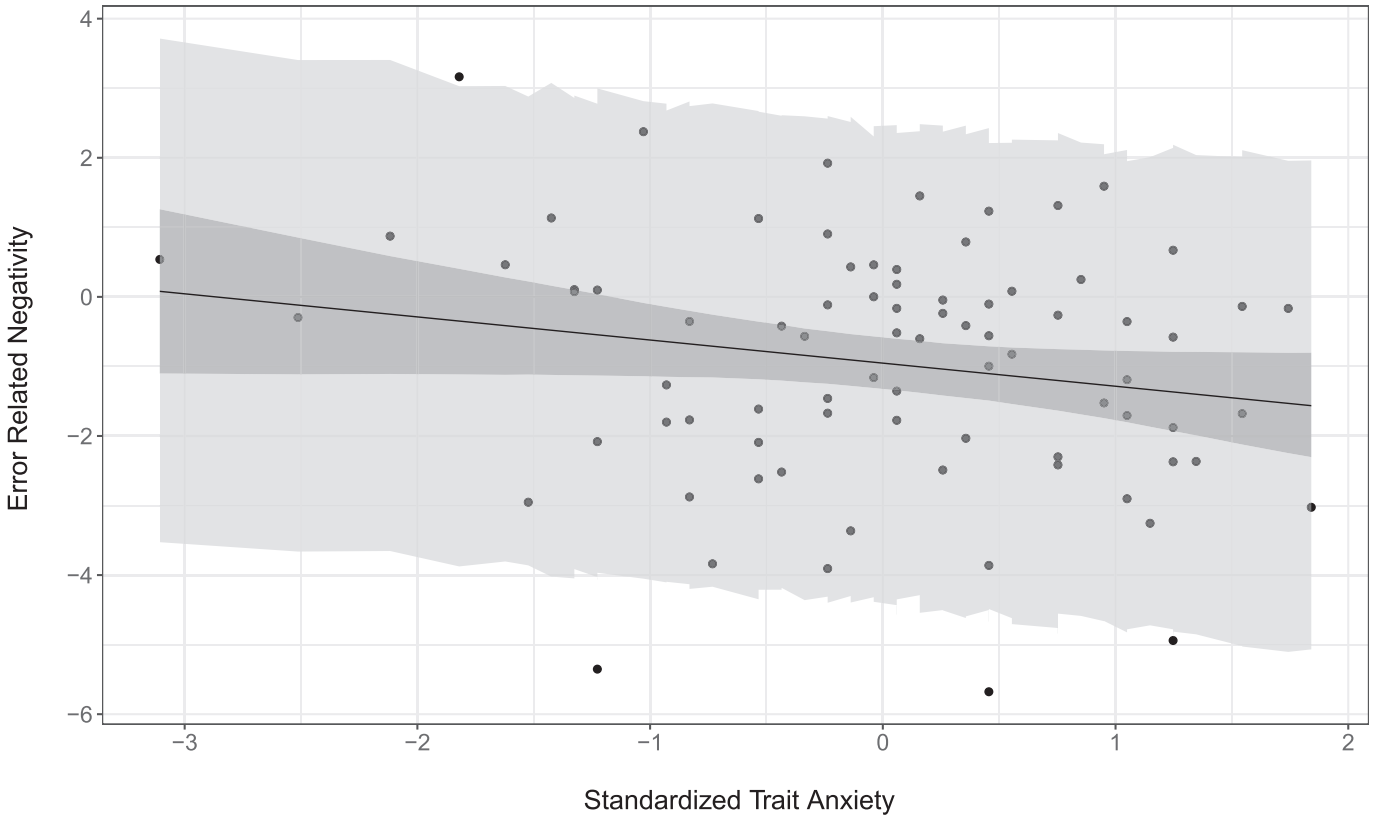

Fig. 6. Scatterplot of anxiety and error-related negativity (ERN). Average expected value for the ERN (line), 95% credible interval for the expected value (dark-shaded area), and 95% prediction interval for the predicted responses (light-shaded area).

In addition to predicted expected values, we may also want to predict new values of the outcome—actual values of the ERN not just the mean. In other words, make predictions about a new or future participant rather than the average participant. Creating a posterior distribution for new responses requires Equation 6 and the following 9 steps.

1 State the likelihood for the data: $\text{ERN}_i \sim N(\mu_i, \sigma)$.
2 Specify the form for μ_i: $\mu_i = E(\text{ERN})_i = \beta_0 + \beta_1 \text{Anxiety}_i$.
3 Select the first sampled value of β_0 and β_1 from the chain and substitute those specific values. In this case, the first sampled values were: $\beta_0 = -1.11$ and $\beta_1 = -0.57$.
4 Select a specific value of anxiety to produce a new data point. In this case, we selected anxiety $= z = 1$.
5 Compute the expected value: $\mu_i = -1.11 + -0.57 \times 1 = -1.68$.
6 Select the first sampled value of σ. In this case, the first sampled value was $\sigma = 1.61$.
7 Substitute the computed expected value and the sampled σ into the likelihood: $\text{ERN}_i \sim N(-1.68, 1.61)$.
8 Draw one new ERN value from the normal distribution specified in Step 7. This is the predicted value for a new participant.
9 Repeat Steps 2–8 for all sampled values of β_0, β_1, and σ.

The posterior of predicted responses will be more variable than the posterior of expected values. Expected values average over sampling variability—average over the fact that people with the same anxiety level will vary with respect to the ERN. Furthermore, the width of the posterior distribution of expected values is a function of the uncertainty in β_0 and β_1. Predicted responses include sampling variability and the width of the posterior of predicted responses is a function of the uncertainty in β_0, β_1, and σ. The light-shaded area in Fig. 6 is the 95% interval estimate for the predicted response. Interval estimates for predicted responses are called prediction intervals or forecast intervals (cf. Stata Corp, 2015a).

Although clinical researchers do not often use predicted responses to understand their models, it does not mean they are not important. In fact, predicted responses are particularly important for research that aims to influence clinical practice. Clinical practice is not about the average patient but specific patients. Thus, to make informed treatment or assessment decisions, it is critical to understand not only how the average patient responds but how much specific patients are expected vary around that average. Prediction intervals are useful in this regard and Bayesian methods make it straightforward to obtain them.

2.3.3. Expanding the model

Up to this point, the model we have used has been small and uncomplicated. Adding predictors to a Bayesian regression model is similar to adding them to a frequentist regression model, with the exception that Bayesian models have the additional requirement of specifying priors for all new parameters. Adding predictors also allows us to illustrate (a) Bayesian measures of model fit that can be used to compare models and (b) how natural it is to interpret interactions, including uncertainty in the interaction, with a Bayesian model.

Moser et al. (2016) argued that models examining the anxiety-ERN relationship should include sex as a moderator because their meta-analysis suggested that the anxiety-ERN relationship was stronger in women than men. Consequently, we fit a model with sex as a covariate and another model with sex as a moderator of anxiety. The model with sex as a covariate (Model 2 in Table 1):

$$\begin{aligned} &\text{ERN}_i \sim N(\mu_i, \sigma) \\ &\mu_i = \beta_0 + \beta_1 \text{Anxiety}_i + \beta_2 \text{Sex}_i \\ &\beta_0, \beta_1, \beta_2 \sim N(0, 3) \\ &\sigma \sim \text{half-Cauchy}(0, 2.5) \end{aligned} \tag{11}$$

The likelihood remains the same, but we have now added sex as a predictor of μ_i as well as a normal prior for the coefficient for sex, β_2.

The model with sex as a moderator of anxiety is a simple extension (Model 3 in Table 1):

$$\begin{aligned} &\text{ERN}_i \sim N(\mu_i, \sigma) \\ &\mu_i = \beta_0 + \beta_1 \text{Anxiety}_i + \beta_2 \text{Sex}_i + \beta_3 \text{Anxiety}_i \times \text{Sex}_i \\ &\beta_0, \beta_1, \beta_2, \beta_3 \sim N(0, 3) \\ &\sigma \sim \text{half-Cauchy}(0, 2.5) \end{aligned} \tag{12}$$

We added the interaction of sex and anxiety as predictor of μ_i as well as a normal prior for the interaction, β_3.

2.3.4. Evaluating model fit

With multiple models we can also ask which model performs the best. The difficulty is to determine what is the best. Some possible ways of defining best are: (a) choosing a model with the most statistically significant predictors; (b) choosing a model with the larges values of R^2 (regression) or η^2 (ANOVA); and (c) choosing the most parsimonious. Unfortunately, there is no one correct way to define best. Furthermore, many decisions we make about models are dictated by conventions ($p < 0.05$) rather than a principled connection between models fit and scientific goals.

An important goal in research, especially research aimed at influencing clinical practice, is how well a model performs in out-of-sample predictions. Statistics such as R^2 or η^2 quantify how well a model predicts the sample-data. However, models are influenced by unique aspects of a given dataset, which means that the model may not fit as well in a new dataset that itself has unique aspects (this is the motivation for adjusted-R^2). Additionally, model fit statistics such as R^2 and η^2 increase as models become more complex (i.e., as we add predictors and interactions), but again the improvements may simply reflect the characteristics of the particular dataset and not be generally applicable. This lack of general applicability is a core issue in the replicability and stability of scientific findings (Open Science Collaboration, 2015).

Researchers could also use cross-validation, where a model is built on a training dataset and then compared to one or more test datasets, to examine out-of-sample predictions. However, datasets are often too small to obtain stable cross-validation estimates. This may be particularly problematic in clinical research where large clinical samples can be challenging to obtain. Consequently, statisticians have developed alternative metrics for approximating a model's out-of-sample performance. Two common metrics applicable to Bayesian methods are the Widely Applicable Information Criterion (WAIC; Gelman et al., 2014; McElreath, 2016) and Leave-one-out Cross-validation (LOO-CV; Vehtari, Gelman, & Gabry, 2016b, 2016a). Both methods can be seen as a way to approximate a model's accuracy in making predictions in a new sample.

WAIC is provided by several software packages, including the `brms` package. WAIC is a generalization of the deviance statistic, which itself is a generalization of the least squares criterion in traditional regression and ANOVA (Singer & Willett, 2003). The deviance statistic indicates how much worse our model is as compared to a perfectly fitting model (i.e., one that completely reproduces the data; Singer & Willett, 2003, p. 117). Consequently, when comparing two models using a deviance, the model with a smaller deviance has better fit. The WAIC approximates the

deviance for new samples (McElreath, 2016) and thus smaller WAIC values indicate better fit. An advantage of WAIC compared to similar metrics, such as the Akaike Information Criterion, is that it uses the full posterior distribution of the parameters, thus taking into account uncertainty in the parameter estimates. A full discussion of how WAIC is calculated is beyond the scope of this paper. Interested readers can consult McElreath (2016, Chapter 6) and Gelman et al. (2014, Chapter 7).

LOO-CV is a cross-validation method where multiple training datasets are made by leaving out a single but different observation from each dataset. Predictive accuracy is estimated by averaging over the performance of the model across the many training datasets. This can be computationally burdensome because we have to run as many models as we have observations. To overcome this limitation, Vehtari et al. (2016b) developed a method that uses Pareto-smoothed importance sampling to estimate the LOO-CV statistic without the need to run the model over and over again. The details of the algorithms can be found in Vehtari et al. (2016b) and the algorithms are implemented in the `loo` package in R (Vehtari, Gelman, & Gabry, 2016a). The `brms` package has a function for interfacing with `loo`. As with WAIC, smaller values of LOO-CV indicate better fit.

Table 1 provides WAIC and LOO-CV values for each model. Model 2 had poorer fit than Model 1—adding sex as a predictor, but not including it as an interaction, decreased the performance of the model. Model 3 fit better than Model 2 or Model 1, indicating that adding the interaction of sex and anxiety improves fit. One might ask whether the models are significantly different from one another. This line of thinking comes primarily from the tradition of significance testing where $p < 0.05$ is a criterion, and sometimes the only criterion, for determining whether a result is scientifically important. McElreath (2016) notes in his book:

> Newcomers to information criteria often ask whether a difference between AIC/DIC/WAIC values is 'significant' …. In general, it is not possible to provide a principled threshold of difference that makes one model 'significantly' better than another, whatever that means. The same is actually true of ordinary significance testing—the 5% convention is just a convention. We could invent some convention for WAIC, but it too would just be a convention. Moreover, we know the models will not make the same predictions—they are different models. So 'significance' in this context must have a very different definition than usual. The attitude this book encourages is to retain and present all models, no matter how big or small the difference in WAIC (or another criterion). The more information in your summary, the more information for peer review, and the more potential for the scholarly community to accumulate information. (p. 200–201)

We echo McElreath and aim for model evaluation and interpretation that is more thorough and nuanced than the binary decision of significant or not. In the case of the ERN-anxiety analysis, the model with the interaction fits better than the model without. Probing the predictions can help us figure out what Model 3 can tell us and then we can judge whether it is scientifically or clinically useful.

2.3.5. Interpreting an interaction

Model 3 includes an interaction between sex and anxiety. Including that interaction allows the relationship between anxiety and the ERN to differ by sex. In other words, the interaction allows for unique regression lines describing the anxiety-ERN relation for men and women. The expected value of the ERN in Model 3 is (see Expression 12):

$$\mu_i = \beta_0 + \beta_1 \text{Anxiety}_i + \beta_2 \text{Sex}_i + \beta_3 \text{Anxiety}_i \times \text{Sex}_i$$

Given that sex is coded as 1 for women and 0 for men, the interpretation of each of these coefficients is:

- β_0: The intercept for men.
- β_1: The anxiety slope for men.
- β_2: The difference in the intercepts for women and men.
- β_3: The difference in the anxiety slope for women and men.

We can take advantage of the fact that sex is a dummy variable coded as 1 for women and 0 for men to construct the distinct lines for women and men. For example, by substituting sex = 0 into Model 3, the equation reduces to the regression line for men:

$$\mu_i|\text{Sex} = 0 = \beta_0 + \beta_1 \text{Anxiety}_i$$

Where $\mu_i|\text{Sex} = 0$ means the expected value of the ERN when sex is 0. Substituting Sex = 1 into Model 3 produces the regression line for women

$$\mu_i|\text{Sex} = 1 = \beta_0 + \beta_1 \text{Anxiety}_i + \beta_2 + \beta_3 \text{Anxiety}_i$$

Rearranging and simplifying to make the intercept and slope for women clearer:

$$\mu_i|\text{Sex} = 1 = \overbrace{(\beta_0 + \beta_2)}^{\text{Intercept}} + \overbrace{(\beta_1 + \beta_3)}^{\text{Slope}} \text{Anxiety}_i$$

To obtain the posterior distribution of the intercept and slope for women, we simply add the draws for β_0 and β_2 together (intercept) and β_1 and β_3 together (slope). We can then analyze the posterior distribution of these parameter combinations in the same way that we have other posterior distributions. For example, Fig. 7 depicts the density plots, including a 95% credibile interval, for the intercept and slope for each sex. The intercepts for the sexes overlap considerably, with less uncertainty for women given the larger number of women. For men, nearly all values for the slope are negative and the entire 95% credible interval is negative. For women, the distribution is centered near zero and positive and negative values are equally credible.

Fig. 8 is a scatterplot of the data as well as the expected ERN values for specific anxiety levels stratified by sex. We included the mean and 95% credible interval for the posterior of expected values. This plot emphasizes that there is no expected relationship among anxiety and the ERN for women but a negative relationship for men. Further, uncertainty in prediction is about twice as large for men than women, as evidenced by the fact that the 95% interval is about twice as large for men and women across all levels of anxiety. This is due to the sample size discrepancy between the two groups; given the uncertainty, more data are needed before firm conclusions can be drawn about sex differences. That is, this interaction may be statistically significant, in the $p < 0.05$ sense, but given the uncertainty of the predictions, caution in drawing strong conclusions is warranted.

Figs. 7 and 8 underscore that quantifying uncertainty is straightforward when using Bayesian methods, even when we want to make inferences about combinations of parameters. For example, obtaining a posterior distribution for a combination of parameters requires only the draws from the posterior of the constituent parts. This makes interpreting interactions, which is aided by computing simple slopes as we did previously, simple and straightforward. Even though we have used a simple example, the concepts are the same even as the data analysis problems become more complex—more complex interactions or non-linear

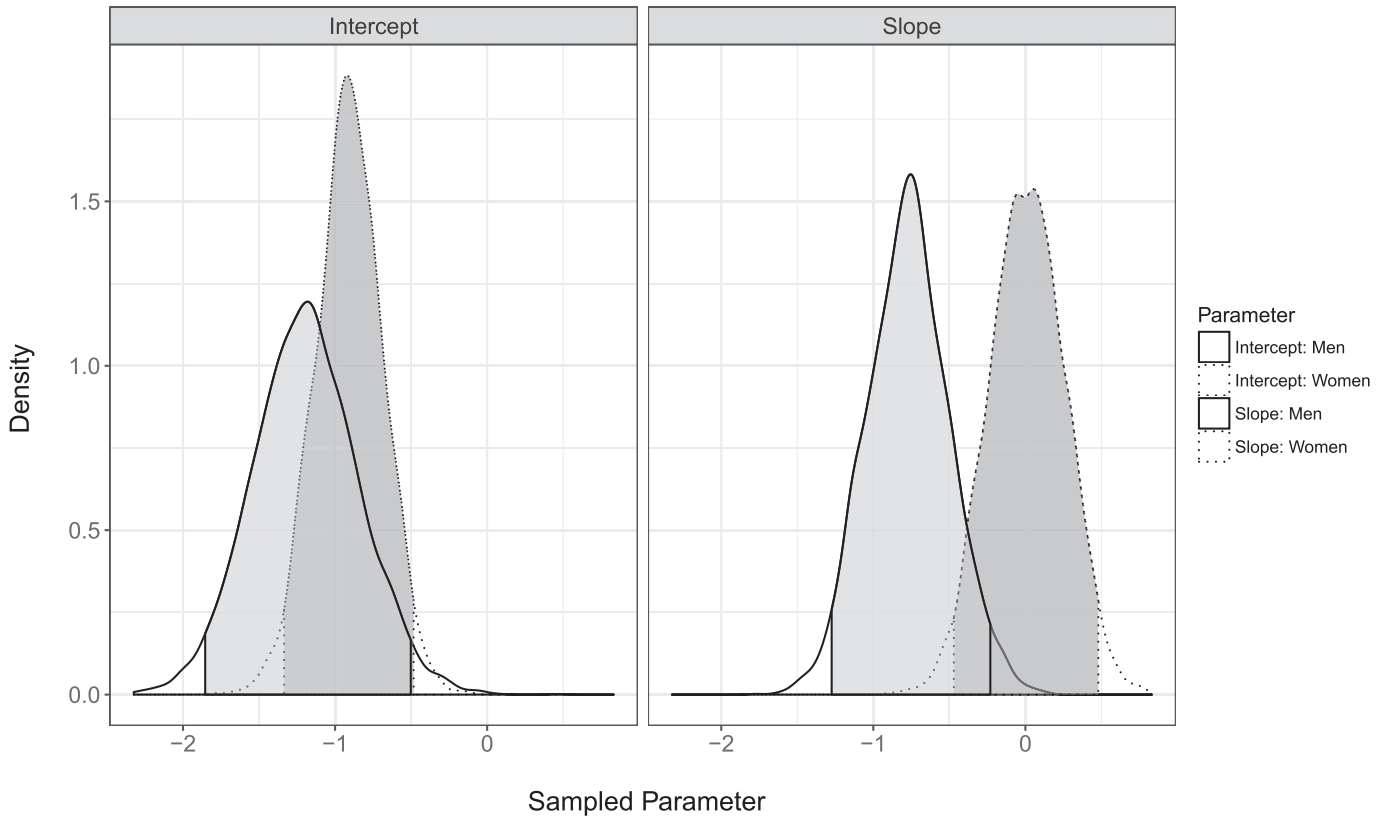

Fig. 7. Posterior density for the intercept and anxiety slope for men and women. Shaded areas represent the 95% credible interval for the parameter (dark-shade for women).

combinations of parameters (e.g., mediation effects, R^2). This is not to say that similar inferences are not possible with frequentist methods. However, computing standard errors (and thus p-values and confidence intervals) for linear and non-linear combinations of parameters can be tricky. For example, entire books have been written on interval estimates for variance components (Burdick & Graybill, 1992) or for decomposing interactions (Aiken & West, 1991).

The ease with which one can obtain standard errors for linear and non-linear combinations of parameters in frequentist models is typically a function of whether the software one uses supports such combinations[3]. In simple cases, obtaining standard errors can be done by hand or with a little coding. For more complex cases, obtaining standard errors can be challenging without the help of additional software. Finally, even when the software computes standard errors, those methods typically rely on the assumption that the sampling distribution of the parameter is normally distributed. This is not always the case—for example, indirect effects in (Yuan & Mackinnon, 2009) or intraclass correlations (Baldwin & Fellingham, 2013). Bayesian methods do not require these assumptions.

3. Other models

Our ERN and anxiety example analyses focused on linear regression. Not surprisingly, Bayesian methods can be used for many kinds of models. In this section, we show how to write out a logistic regression, Poisson regression, and multilevel linear model. We do not discuss model interpretation, the nuances of these models, or even the general background to a model (e.g., what link functions are in generalized linear models or the definition of random effects). Readers needing an introduction to these types of models can consult Gelman and Hill (2007). The aim of this section is to provide a starting point for learning about Bayesian methods for more advanced models and to illustrate how similar these models are to linear regression. Researchers should be able to adapt the equations we provide to their own models. Additionally, the online supplemental materials provide data and `brms` code for fitting similar models.

3.1. Logistic regression

Logistic regression is used to predict binary outcomes, specifically the probability that the outcome y_i is 1 or 0. Examples include premature treatment termination, recidivism, or presence/absence of a diagnosis. The likelihood is a Bernoulli distribution:

$$P(y_i|\pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} \tag{13}$$

where π_i is the probability that $y_i = 1$. In logistic regression, π_i is modeled as a function of regression coefficients. The full probability model for a logistic regression with a single predictor is:

$$\begin{aligned} y_i &\sim \text{Bernoulli}(\pi_i) \\ \pi_i &= \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} \\ \beta_0, \beta_1 &\sim N(0, 3) \end{aligned} \tag{14}$$

[3] For example, the `test`, `lincom`, `nlcom`, and `margins` commands in Stata are quite flexible and can handle many types of parameter combinations (Stata Corp, 2015c).

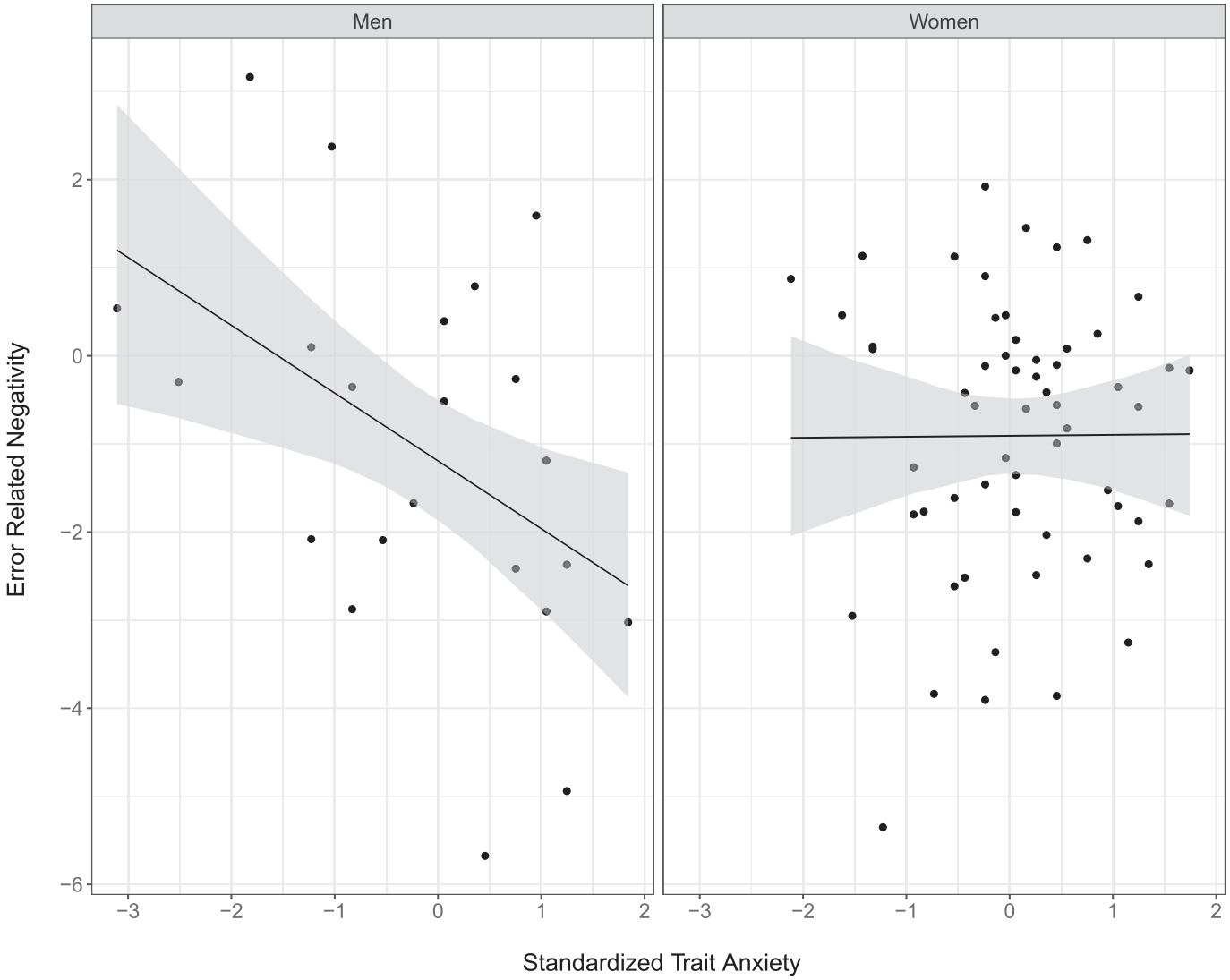

Fig. 8. Scatter plot of anxiety and error-related negativity (ERN) for men and women. Average expected value for the ERN (line) and 95% credible interval for the expected value (light-shaded area).

Thus, with a single predictor, the logistic regression has two parameters, β_0 and β_1. The probability that $y_i = 1$, π_i, is determined by β_0 and β_1 and the inverse-logit function. Consequently, we only need to state priors for the regression coefficients, which should be tailored to the specific situation.

3.2. Poisson regression

Poisson regression is used to model count data, such as the number of alcoholic drinks consumed in a week, number of parasuicidal behaviors, or number of binges in a month. Count data are whole integers and must be greater than or equal to zero. A commonly used likelihood for count models is the Poisson distribution:

$$P(y_i|\lambda_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \tag{15}$$

where λ_i is the expected count and y_i is the observed count. In Poisson regression, λ_i is modeled as a function of regression coefficients. The full probability model for a Poisson regression with a single predictor is:

$$\begin{aligned} \mathrm{y}_i &\sim \mathrm{Poisson}(\lambda_i) \\ \lambda_i &= \exp(\beta_0 + \beta_1 X_i) \\ \beta_0, \beta_1 &\sim N(0, 3) \end{aligned} \tag{16}$$

As with the logistic regression, the Poisson regression with a single predictor has two parameters requiring a prior, β_0 and β_1.

3.3. Multilevel linear model

A multilevel linear model uses a normal distribution as the likelihood. What makes the multilevel model different for linear regression is the addition of random effects to accommodate the fact that observations are nested within clusters. Examples of nested data are patients within therapists, therapists within clinics, or repeated observations within a patient. The outcome variable in multilevel linear model is denoted as y_{ij}, which means the outcome for observation i in cluster j. The full probability model for a multilevel linear model with a single predictor is:

$$
\begin{aligned}
&y_{ij} \sim N\left(\mu_{ij}, \sigma\right) \\
&\mu_{ij} = \beta_0 + \beta_1 X_{ij} + u_j \\
&\beta_0, \beta_1 \sim N(0, 3) \\
&u_j \sim N(0, \sigma_u) \\
&\sigma_u \sim \text{half-Cauchy}(0, 5) \\
&\sigma \sim \text{half-Cauchy}(0, 5)
\end{aligned} \tag{17}
$$

As with the linear regression model, there are priors for β_0, β_1, and σ. The primary difference between this model and the linear regression is the addition of the random effects, u_j. The prior distribution for the random effects is a normal distribution with a mean of 0 and unknown standard deviation, σ_u. Given that σ_u is unknown, it will be estimated and thus needs a prior. Because it is standard deviation, we again use a half-Cauchy distribution.

3.4. Bayes factors

We have focused on model building and parameter estimation, including understanding the uncertainty in parameters and model based predictions. Another common task in clinical research is hypothesis testing, such as testing whether two treatments are equally effective. Null hypothesis significance testing is one way to test this hypothesis, but is limited because significance tests do not provide evidence in support of the null hypothesis (Rouder et al., 2009). What is needed is a method for comparing whether data from a study are more consistent with the null hypothesis than the alternative hypothesis. Bayes factors provide this information.

A Bayes factor can be expressed as (Rouder et al., 2009, p. 228):

$$
\frac{Pr(H_A|\text{data})}{Pr(H_0|\text{data})} = \frac{Pr(\text{data}|H_A)}{Pr(\text{data}|H_0)} \times \frac{Pr(H_A)}{Pr(H_0)} \tag{18}
$$

The ratio on the far right is the prior odds, which quantifies how much we believe in the alternative hypothesis over the null prior to the study. It is the probability of the alternative hypothesis (H_A) relative to that of the null hypothesis (H_0). The ratio on the left is the posterior odds, which quantifies how much we believe in the alternative hypothesis over the null after the study. It is the probability of the alternative hypothesis given the data relative to the probability of the null given the data. The middle ratio is the Bayes factor, which is the probability of the data given the alternative hypothesis relative to the probability of the data given the null hypothesis. In lay terms, the Bayes factor quantifies how our belief in the alternative hypothesis relative to the null changes as a result of the data. The Bayes factor is used to produce the posterior odds (Morey, 2014).

The Bayes factor tells us how much the data are consistent with one hypothesis as compared to the other. For example, in a treatment study, a possible alternative hypothesis is that two treatments differ by half of a standard deviation and the null is that they do not differ at all. A Bayes factor greater than one suggests that the data are more likely given the alternative hypothesis than they are given then null hypothesis. In other words, there is more evidence for the alternative than the null hypothesis. This interpretation is in contrast to the binary decision made with null hypothesis significance testing, where we decide whether to reject the null hypothesis outright. Instead, by using a Bayes factor we quantify the relative evidence for each hypothesis. There are general guidelines for what constitutes meaningful evidence for one hypothesis over (see Rouder et al., 2009), but ultimately these are arbitrary. Bayes factors, like any statistical evidence, ought to be interpreted in light of theory and past evidence.

A full worked example of Bayes factors is beyond the scope of this article. Excellent introductions can be found in Rouder et al. (2009), Morey (2014), and Field (2016). Bayes factors can be estimated in the R package BayesFactor (Morey & Rouder, 2015) and in JASP.

4. Discussion

Up to this point we have considered some benefits of a Bayesian approach: (a) clarifying uncertainty in parameter estimates, (b) focusing on predictions, (c) interval estimates with a more straightforward interpretation than confidence intervals, and (d) easily obtaining posterior distributions by combining parameters. Often the unique benefits of Bayesian methods may become more clear when we consider complex models. In this section, we briefly discuss four benefits of the Bayesian approach when dealing with more complex situations: (a) stable, more realistic estimation when sample sizes are small and/or designs are complex, (b) relaxing model constraints, (c) using existing data as a prior for a model, and (d) solving seemingly intractable computational problems.

A common randomized trial design is what is called a partially-clustered design. In this design, some of the treatment conditions included clustered observations and some do not. An example partially-clustered design is a trial comparing group psychotherapy to individual psychotherapy. In the group psychotherapy condition, patients are clustered within small groups but no such clustering exists in the individual psychotherapy condition. Consequently, the variance structure of the data is complex and frequentist multilevel models can require adjustments to standard errors and degrees of freedom for tests of the intervention effect (Baldwin & Fellingham, 2013; Baldwin, Bauer, Stice, & Rohde, 2011), but these corrections are only available if the outcome is continuous. Bayesian models do not require the adjustments and thus can be used for any outcome type. Furthermore, the Bayesian estimates of the variance components can be more realistic and stable, especially when sample sizes are small (Baldwin & Fellingham, 2013).

Bayesian methods also allow some models to be specified more flexibly. For example, Bayesian estimation can allow for small, but non-zero cross-loadings in confirmatory factor analysis, even when such loadings may lead to a non-identifiable model when traditional estimation methods are used (Muthén & Asparouhov, 2012). Similarly, Bayesian methods allow more flexibility with how correlated residuals are handled in measurement models, which opens up the kinds of models we can evaluate (e.g., better evaluate method effects in measurement models; Muthén & Asparouhov, 2012).

Although we discussed selecting prior distributions based on researchers' evaluation and understanding of a particular literature, priors can also be created from previous data. For example, in cluster-randomized trials, where entire clusters (e.g., clinics, schools, cities, etc.) are randomized to conditions, the number of clusters is often small. Consequently, obtaining a precise estimate of the cluster-level variance when estimating intervention effects can be a problem. A possible solution to this problem is to use existing data to establish a prior distribution for variance components. For example, Turner, Thompson, and Spiegelhalter (2005) showed how existing estimates of an intraclass correlation can be used to create an empirical prior distribution for the analysis of new data in a cluster-randomized trial. This is one of many possibilities of using existing data to enhance a given study.

Finally, Bayesian methods can overcome estimation problems common to maximum likelihood methods. For example, fitting multilevel models for non-normal data (e.g., binary, count) with many random effects or structural equation models with-normal data and four or more latent variables can be difficult, if not impossible (Muthén & Asparouhov, 2012). MCMC works exactly the same when using a normal likelihood as it does with a Poisson or binomial likelihood. Consequently, Bayesian estimation of these

more complex models is feasible. It still may take time to obtain a sufficient number of draws from the posterior, but it will be possible.

Of course, none of these benefits come without a cost; Bayesian methods have their challenges. Here we discuss five challenges. First, for standard models such as regression, ANOVA, and multilevel models, computation time can be longer for a Bayesian model than frequentist models. It simply takes time to obtain draws from the posterior distribution. This may be an annoyance, but we believe that the extra time is worth it. If a Bayesian model provides a better answer or allows one to better understand a model, the extra time is worth it. Second, many Bayesian models require programming. Programs like `brms` in R, JASP, and Mplus are making this less of a hurdle. Third, Bayesian modeling requires more understanding of probability and probability distributions than is typically taught during the training of most psychologists. Fourth, many Bayesian software programs do not require that one specify a prior but will use default priors based on the type of model. These priors are sometimes okay and sometimes not and it can be difficult to tell. Consequently, we recommend that researchers explicitly set the priors for their models. Fifth, to obtain reasonable estimates when the sample size is small, we must carefully choose an informed prior. Likewise, to use existing data to create a prior distribution, we have to make sure the existing data are good enough and similar enough to inform the current analysis. That is, Bayesian analysis does not automatically lead to proper inferences or better estimates. We can fool ourselves with Bayesian methods just as easily as we can with many kinds of analytic techniques (e.g., fiddling with prior distributions until we get the result we want or that is publishable; Simmons et al., 2011). Like any aspect of scientific practice, any use of Bayesian methods should be subject to peer review and criticism.

4.1. Minimum reporting guidelines

As a field, psychology and clinical psychology can improve the way research is reported and discussed. Providing explicit, complete, and transparent reports of the methods and analyses improves communication, makes it possible to fully evaluate the research (both before and after publication), and can assist in replication efforts. To this end, we offer the following reporting guidelines for researchers wishing to use Bayesian methods in *Behaviour Research and Therapy* or similar journal.

1 Provide a complete description of the likelihood and all priors used in the analysis. We recommend that a full probability model be reported. Including a full description of the likelihood and the prior allows others to better evaluate the model and even test alternatives to determine if the model is sensitive to specific aspects of the model. This is especially important given that software programs can estimate models without an explicit declaration of the priors. The defaults may be appropriate, but often they are not the best choices.
2 Provide details of how convergence was assessed. This can include reporting traceplots, $\widehat{R}$, and other convergence statistics (see Jackman, 2009). If including traceplots is not possible based on page constraints, include them in an online supplement or webpage.
3 Note the software, including version number, used to run the analysis. Software programs differ in their algorithms as well as the types of models they can fit.
4 Provide data where possible as well as the scripts used to run the models. This allows reviewers, pre- and post-publication to evaluate the models fully. This includes testing alternative likelihoods and priors. Recently, Shariff, Willard, Muthukrishna, Kramer, and Henrich (2016) demonstrated how poor methodological choices about the coding of independent variables and the likelihood led to the potentially misleading results about religiosity and altruism reported in Decety et al. (2015).
5 When providing summaries of posterior distributions, clearly report the nature of the summaries. That is, if point estimates are means, label them as such. Do not assume that the reader will figure it out.
6 Report the number of draws from the posterior as well as effective sample size.

4.2. Conclusion

We have just scratched the surface of Bayesian methods and their application to clinical research. Bayesian methods are used more than ever in many disciplines and represent a powerful set of tools that can help enhance data analysis in clinical research. Given that they are just tools, there is not guarantee that the reliability and replicability of our research will improve simply by implementing Bayes' theorem. Nevertheless, we have found that thinking about likelihood and priors, visualizing the uncertainty of our estimates, and examining the uncertainty of model predictions, has forced us to think more carefully about our research than we did before. More careful thinking strikes us a worthwhile thing and we hope it aids our clinical research.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.brat.2016.12.016.

Appendix

Glossary of Terms

- **Convergence** = Refers to the whether a Markov Chain has sampled from the posterior distribution.
- **Credible intervals** = Bayesian interval estimates that indicate which parameter values have the most probability given the data and priors.
- **Effective Sample Size** = The number of independent samples from a posterior distribution.
- **Event-related potential**: averaged waveforms that depict changes in a large population of synchronously active neurons
- **Error-related negativity**: Negative deflection in the event-related potential that is larger following errors than correct trials and is associated with error-related brain processes.
- **Leave-one-out Cross-validation** = An index of model fit that estimates the out-of-sample performance of a model. Smaller values indicate better fit.
- **Likelihood** = The probability of the data given parameters in the model.
- **Markov Chain Monte Carlo** = A simulation algorithm used to sample from the posterior distribution.
- **Posterior distribution** = A probability distribution for a parameter. It is the combination of the information from the data and the prior.
- **Prior** = Predictions about a parameter before seeing the data.
- $\widehat{R}$ = The Gelman-Rubin statistic. A numerical method for examining whether multiple chains converge on the same posterior. $\widehat{R}$ values should be between 1 and 1.1.
- **Shrinkage** = When a estimate from the data is influenced by prior information.
- **Widely Applicable Information Criterion** = An index of model fit that estimates the out-of-sample performance of a model. Smaller values indicate better fit.

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Thousand Oaks, CA: Sage.

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist, 63*, 32–50.

Atkins, D. C., & Gallop, R. J. (2007). Rethinking how family researchers model infrequent outcomes: A tutorial on count regression and zero-inflated models. *Journal of Family Psychology, 21*, 726–735.

Baldwin, S. A., Bauer, D. J., Stice, E., & Rohde, P. (2011). Evaluating models for partially clustered designs. *Psychological Methods, 16*, 149–165.

Baldwin, S. A., & Fellingham, G. W. (2013). Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. *Psychological Methods, 18*, 151–164.

Baldwin, S. A., Fellingham, G. W., & Baldwin, A. S. (2016). Statistical models for multilevel skewed physical activity data in health research and behavioral medicine. *Health Psychology, 35*, 552–562.

Baldwin, S. A., Larson, M. J., & Clayson, P. E. (2015). The dependability of electrophysiological measurements of performance monitoring in a clinical sample: A generalizability and decision analysis of the ERN and Pe. *Psychophysiology, 52*, 790–800.

Buerkner, P.-C. (2016). *brms: Bayesian Regression Models using Stan. R package version 0.9.1.9000*. Retrieved from http://github.com/paul-buerkner/brms.

Burdick, R. K., & Graybill, F. A. (1992). *Confidence intervals on variance components. Statistics, textbooks and monographs*. New York: Marcel Dekker.

Clayson, P. E., Baldwin, S. A., & Larson, M. J. (2013). How does noise affect amplitude and latency measurement of event-related potentials (ERPs)? A methodological critique and simulation study. *Psychophysiology, 50*(2), 174–186.

Cohen, J. (1994). The Earth is round (p < .05). *American Psychologist, 49*, 997–1003.

Decety, J., Cowell, J. M., Lee, K., Mahasneh, R., Malcom-Smith, S., Selcuk, B., et al. (2015). The negative association between religiousness and children's altruism across the world. *Current Biology, 25*, 2951–2955.

Eriksen, B., & Eriksen, C. (1974). Effects of noise letters upon the identification of a target letter in a non-search task. *Perception & Psychophysics, 16*, 143–149.

Field, A. (2016). *An adventure in statistics: The reality enigma*. Thousand Oaks, CA: Sage.

Gabry, J. (2016). *shinystan: Interactive visual and numerical diagnostics and posterior analysis for Bayesian models. R package version 2.2.0*. Retrieved from https://CRAN.R-project.org/package=shinystan.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis, 1*, 515–533.

Gelman, A., & Carlin, J. (2014). Beyond power calculations assessing Type S (Sign) and Type M (Magnitude) errors. *Perspectives on Psychological Science, 9*, 641–651.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis (3rd)*. Boca Raton, FL: CRC Press.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*, 460–465.

Greenland, S. (2006). Bayesian perspectives for epidemiological research: I. Foundations and basic methods. *International Journal of Epidemiology, 35*, 765–775.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1–20.

Hupen, P., Groen, Y., Gaastra, G., Tucha, L., & Tucha, O. (2016). Performance monitoring in autism spectrum disorders: A systematic literature review of event-related potential studies. *International Journal of Psychophysiology, 102*, 33–46.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS medicine, 2*, e124.

Jackman, S. (2009). *Bayesian analysis for the social sciences*. Chichester, U.K: Wiley.

JASP Team. (2016). JASP (Version 0.8.0.0) [Computer software].

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General, 142*, 573–603.

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and stan*. London: Academic Press.

Kujawa, A., Weinberg, A., Bunford, N., Fitzgerald, K., Hanna, G., Monk, C., et al. (2016). Error-related brain activity in youth and young adults before and after treatment for generalized or social anxiety disorder. *Progress in Neuro-psychopharmacology & Biological Psychiatry, 71*, 162–168.

Larson, M. J., Clawson, A., Clayson, P. E., & Baldwin, S. A. (2013). Cognitive conflict adaptation in generalized anxiety disorder. *Biological Psychology, 94*, 408–418.

Larson, M. J., Clayson, P. E., & Clawson, A. (2014). Making sense of all the conflict: A theoretical review and critique of conflict-related ERPs. *International Journal of Psychophysiology, 93*, 283–297. http://dx.doi.org/10.1016/j.ijpsycho.2014.06.007.

Larson, M. J., Clayson, P. E., Keith, C. M., Hunt, I. J., Hedges, D. W., Nielsen, B. L., et al. (2016). Cognitive control adjustments in healthy older and younger adults: Conflict adaptation, the error-related negativity (ERN), and evidence of generalized decline with age. *Biological Psychology, 115*, 50–63. http://dx.doi.org/10.1016/j.biopsycho.2016.01.008.

Lindgren, B. W. (1993). *Statistical theory* (4th ed.). New York, NY: Chapman & Hall.

Llerena, K., Wynn, J., Hajcak, G., Green, M., & Horan, W. (2016). Patterns and reliability of EEG during error monitoring for internal versus external feedback in schizophrenia. *International Journal of Psychophysiology, 105*, 39–46.

Luck, S. (2014). *An introduction to the event-related potential technique (2nd)*. Cambridge, MA: MIT Press.

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and stan*. Boca Raton, FL: CRC Press.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806–834.

Moran, T., Schroder, H., Kneip, C., & Moser, J. S. (2017). Meta-analysis and psychophysiology: A tutorial using depression and action-monitoring event-related potentials. *International Journal of Psychophysiology, 111*, 17–32.

Morey, R. D. (2014). *What is a Bayes factor?*. Retrieved from http://bayesfactor.blogspot.com/2014/02/the-bayesfactor-package-this-blog-is.html.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review, 23*, 103–123.

Morey, R. D., & Rouder, J. N. (2015). *BayesFactor: Computation of Bayes factors for common designs. R package version 0.9.12-2*. Retrieved fromhttps://CRAN.R-project.org/package=BayesFactor.

Moser, J. S., Moran, T. P., Kneip, C., Schroder, H. S., & Larson, M. J. (2016). Sex moderates the association between symptoms of anxiety, but not obsessive compulsive disorder, and error-monitoring brain activity: A meta-analytic review. *Psychophysiology, 53*, 21–29.

Moser, J. S., Moran, T., Schroder, H., Donnellan, M., & Yeung, N. (2013). On the relationship between anxiety and error monitoring: A meta-analysis and conceptual framework. *Frontiers in Human Neuroscience, 7*, 466.

Muthén, B. O., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods, 17*, 313–335.

Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus user's guide (7th)*. Los Angeles, CA: Muthén & Muthén.

Olvet, D., & Hajcak, G. (2008). The error-related negativity (ERN) and psychopathology: Toward an endophenotype. *Clinical Psychology Review, 28*, 1343–1354.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*. aac4716–aac4716.

Plummer, M. (2015). *JAGS user manual. Version 4.0.0*. Retrieved from http://mcmc-jags.sourceforge.net.

Rosenthal, R. (1994). Science and ethics in conducting, analyzing, and reporting psychological research. *Psychological Science, 5*, 127–134.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225–237.

SAS Institute Inc. (2013). *Base SAS R 9.4 procedures guide: Statistical procedures* (2nd ed.). Cary, NC: SAS Institute Inc.

Schwartz, S. J., Lilienfeld, S. O., Meca, A., & Sauvigné, K. C. (2016). The role of neuroscience within psychology: A call for inclusiveness over exclusiveness. *American Psychologist, 71*, 52–70.

Shariff, A. F., Willard, A. K., Muthukrishna, M., Kramer, S. R., & Henrich, J. (2016). *What is the association between religious affiliation and children's altruism? Current Biology* (Vol. 26, pp. R699–R700).

Sheehan, D., Lecrubier, Y., Sheehan, K., Amorim, P., Janavs, J., Weiller, E., et al. (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I.): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *Journal of Clinical Psychiatry, 59*, 22–33.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366.

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

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2003). *WinBUGS Version 1.4 user manual*. Cambridge: England: MRC Biostatistics Unit.

Spielberger, C. D., Gorusch, R. L., Lushene, R., Vagg, P. R., & Jacobs, G. A. (1983). *Manual for the state-trait anxiety inventory*. Palo Alto, CA: Consulting Psychologists Press.

Stan Development Team. (2016). *Stan modeling language: User's guide and reference manual. Version 2.11.0*. Retrieved from http://mc-stan.org.

Stata Corp. (2015a). *Stata 14 base reference manual*. College Station, TX: Stata Press.

Stata Corp. (2015b). *Stata 14 Bayesian analysis reference manual*. College Station, TX: Stata Press.

Stata Corp. (2015c). *Stata statistical Software: Release 14*. College Station, TX: Stata Press.

Turner, R. M., Thompson, S. G., & Spiegelhalter, D. J. (2005). Prior distributions for the intracluster correlation coefficient, based on multiple previous estimates, and their application in cluster randomized trials. *Clinical Trials, 2*, 108–118.

Vehtari, A., Gelman, A., & Gabry, J. (2016a). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. R package version 0.1.6*. Retrieved fromhttps://github.com/jgabry/loo.

Vehtari, A., Gelman, A., & Gabry, J. (2016b). *Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC*. Retrieved from http://arxiv.org/abs/1507.04544v4.

Yuan, Y., & Mackinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods, 14*, 301–322.