

This article was downloaded by: [173.202.225.227]

On: 06 February 2013, At: 04:29

Publisher: Routledge

Informa Ltd Registered in England and Wales Registered Number: 1072954

Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Multivariate Behavioral Research

Publication details, including instructions for authors and subscription information:

<http://www.tandfonline.com/loi/hmbr20>

Default Bayes Factors for Model Selection in Regression

Jeffrey N. Rouder^a & Richard D. Morey^b

^a University of Missouri

^b University of Groningen

Version of record first published: 17 Jan 2013.

To cite this article: Jeffrey N. Rouder & Richard D. Morey (2012): Default Bayes Factors for Model Selection in Regression, *Multivariate Behavioral Research*, 47:6, 877-903

To link to this article: <http://dx.doi.org/10.1080/00273171.2012.734737>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.tandfonline.com/page/terms-and-conditions>

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages

whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

Default Bayes Factors for Model Selection in Regression

Jeffrey N. Rouder
University of Missouri

Richard D. Morey
University of Groningen

In this article, we present a Bayes factor solution for inference in multiple regression. Bayes factors are principled measures of the relative evidence from data for various models or positions, including models that embed null hypotheses. In this regard, they may be used to state positive evidence for a lack of an effect, which is not possible in conventional significance testing. One obstacle to the adoption of Bayes factor in psychological science is a lack of guidance and software. Recently, Liang, Paulo, Molina, Clyde, and Berger (2008) developed computationally attractive default Bayes factors for multiple regression designs. We provide a web applet for convenient computation and guidance and context for use of these priors. We discuss the interpretation and advantages of the advocated Bayes factor evidence measures.

The old adage “There are several ways to skin a cat,” although gruesome, appropriately describes how researchers draw inferences from data. In today’s literature there is a wide variety of testing and model comparison paradigms, each with its own rationale and corresponding properties. In our view, psychology benefits from this large and diverse methodological toolbox, and researchers can make wise choices that reflect the goals of their research and the types of psychological positions being tested.

Correspondence concerning this article should be addressed to Jeffrey N. Rouder, University of Missouri. E-mail: rouderj@missouri.edu

The topic of this article is Bayes factor, a method of inference first suggested by Laplace (1774/1986), formalized by Jeffreys (1961), and presented to the psychological community shortly thereafter by Edwards, Lindman, and Savage (1963). Bayes factor is highly relevant to applications in which the null hypothesis embeds a substantive regularity or invariance of theoretical interest. Unfortunately, there is a lack of practical guidance and few available software options for Bayes factor.¹ In this article, we provide this practical guidance on how to implement and interpret these Bayes factors in multivariate regression designs and provide a free, easy-to-use web-based applet that computes Bayes factor from the common coefficient of determination statistic R^2 . This guidance is based on the recent work of Liang, Paulo, Molina, Clyde, and Berger (2008), who proposed computationally convenient default priors with desirable theoretical properties. The Bayes factors are easy to use, communicate, and interpret and enable analysts to formally assess evidence in data.

ON ACCEPTING THE NULL HYPOTHESIS

Researchers often find that they are “on the wrong side of the null hypothesis,” that is, their preferred model or explanation serves as the null hypothesis rather than as an alternative. For example, Gilovich, Vallone, and Tversky (1985) assessed whether basketball shooters display hot and cold streaks in which the outcome of one shot attempt affects the outcomes of subsequent ones. They concluded that there is no such dependency. In this case, the lack of dependency serves as the null hypothesis, and, as is commonly known, supporting the null hypothesis is considered conceptually complicated. Conventional significance tests have a built-in asymmetry in which the null may be rejected but not accepted. If the null holds, the best-case outcome of a significance test is a statement about a lack of evidence for an effect. Yet, it would be more desirable to state positive evidence for the invariance of shot outcomes than a lack of evidence for an effect.

In our experience in experimental psychology, being on the wrong side of the null is not a rare occurrence. For example, researchers may hold expectancies of an equivalence of performance across group membership (such as gender, e.g., Shibley Hyde, 2005) or may be interested in the implications of a lack of interaction between stimulus factors (e.g., Sternberg, 1969). In observational context, it may be of great interest that a covariate does not affect a measured variable, and such null statements result in simpler and more parsimonious models. Additionally, models that predict stable relationships, such as the Fechner-

¹Some of the Bayes factors we highlight here may be computed with BAS package in the R language.

Weber Law,² serve as null hypotheses. In summary, being on the wrong side of the null typically corresponds to testing a theoretical position that predicts specific invariances or regularity in data. From a conceptual point of view, being on the wrong side of the null is an enviable position. From a practical point of view, however, being on the wrong side of the null presents difficulties, as conventional testing provides no way of stating positive evidence for the position. The incongruity that null hypotheses are theoretically desirable yet may only be rejected in significance testing has been noted previously by many researchers (Gallistel, 2009; Kass, 1993; Raftery, 1995; Rouder, Speckman, Sun, Morey, & Iverson, 2009).

Consider the following detailed example of how the null may serve as a theoretically useful position. The brain size of hominids grew dramatically over the last few million years. Cranial capacity, which is a surrogate for brain size, grew from 500 cc in the early Australopithecenes to about 1,500 cc in *Homo sapiens*. This rapid growth in brain size occurs on a relatively small evolutionary time frame. The question of why these rapid changes occurred is topical and debated, and possible explanations focus on *ecological*, *climatic*, and *social* selection pressures. Stating evidence for or against these factors as a specific cause of hominid brain evolution is theoretically useful and important.

We use as an example in this article the statistical analysis of Bailey and Geary (2009), who explored which of several variables may have affected the evolution of hominids.³ Bailey and Geary regressed 13 covariates onto the cranial capacity of 175 hominid skulls that varied in age from 1.9 million to 10,000 years. For demonstration purposes, we consider four of these covariates: (a) *local climate variation*, the difference between the average high and low temperatures across a year during the time period; (b) *global average temperature* during the time period; (c) *parasite load*, the number of different types of harmful parasites known to currently exist in the region; and (d) the *population density* of the group the hominid lived within.

Each of these variables corresponds to a specified theory of the evolutionary cause for the rapid brain development in hominids. For example, Potts (1998) hypothesizes that brain-size increases may be driven in part as an adaptation to variability in local climatic conditions. Alternatively, Low (1990) speculates that parasite load is a determining factor. The argument is that as hominids experienced colder climates, they did not have to maintain a metabolically

²The Fechner-Weber Law (Fechner, 1860, republished 1966 in English; Masin, Zudini, & Antonelli, 2009) describes how bright a flash must be to be detected against a background. If the background has intensity I , the flash must be of intensity $I(1 + \theta)$ to be detected. The parameter θ , the Weber fraction, is posited to remain invariant across different background intensities, and testing this invariance is critical in establishing the law.

³We are grateful to Drew Bailey for providing these data.

expensive immune system. These metabolic resources could subsequently be used to support metabolically expensive brain function. Within the context of a multiple regression of several factors, neither local climate variation nor parasite load was significant ($t(170) = .091$, $p \approx .91$ for local climate variation; $t(170) = -1.47$, $p \approx .14$). The conclusions from these p values is that there is a lack of evidence for either effect, perhaps with strained interpretation that there is a more profound lack for local climate variation than for parasite load. We cannot conclude that there is evidence for the lack of an association between cranial capacity and these covariates.

It is our goal to consider methods that allow us to state evidence for either the null or alternative models, depending on which provides a better description of the data. The main difficulty is that because the null is a proper restriction of the alternative, the alternative will always fit the data better than the null. There are several methods that address this problem. Some methods favor models that provide for better out-of-sample prediction, including the Akaike information criterion (AIC; Akaike, 1974) and Mallows's C_p (Mallows, 1973). Covariates that do not increase out-of-sample predictive power are rejected. Other methods, such as *statistical equivalence testing*, expand the null to include a range of small values rather than a sharp point null. Bayes factor is an alternative approach to the same problem that is motivated without recourse to out-of-sample considerations and may be used to state evidence for sharp point nulls.

THE BAYES FACTOR

In this section, we briefly develop the Bayes factor; more comprehensive exposition may be found in Congdon (2006) and Wagenmakers (2007). We begin by defining some notation that is used throughout. Let $\mathbf{y} = (y_1, \dots, y_N)'$ denote a vector of observations, and let $\pi(\mathbf{y} \mid \mathcal{M})$ be the probability (or density) of observing this vector under some model \mathcal{M} . As is common in Bayesian statistics, we use the term *probability of the data* to denote both probability mass for discrete observations and probability density for continuous observations. The probability of data under a model is a Bayesian concept, and we subsequently discuss at length how it is computed.

The Bayes factor is the probability of the data under one model relative to that under another, and the Bayes factor between two models, \mathcal{M}_1 and \mathcal{M}_0 , is

$$B_{10} = \frac{\pi(\mathbf{y} \mid \mathcal{M}_1)}{\pi(\mathbf{y} \mid \mathcal{M}_0)}. \quad (1)$$

The subscript of the Bayes factor identifies which models are being compared, and the order denotes which model is in the numerator and which is in the

denominator. Hence, B_{01} would denote $\pi(\mathbf{y} \mid \mathcal{M}_0)/\pi(\mathbf{y} \mid \mathcal{M}_1)$, and $B_{01} = 1/B_{10}$. The Bayes factor is interpretable without recourse to additional criteria or qualification; for instance, a Bayes factor of $B_{10} = 10$ means that the data are 10 times more probable under \mathcal{M}_1 than under \mathcal{M}_0 .

The Bayes factor is one of a number of Bayesian approaches to inference. Other Bayesian approaches include Aitkin's (1991; see Liu & Aitkin, 2008) marginal-posterior ratios, Kruschke's (2011) use of posterior distributions on contrasts, and Gelman and colleagues' notion of model checking through predictive posterior p values (e.g., Gelman, Carlin, Stern, & Rubin, 2004). The advantages and disadvantages of these methods remain an active and controversial topic in the statistical and social science methodological literatures. Covering this literature is outside the scope of this article. Target articles by Aitkin (1991), Berger and Sellke (1987), and Gelman and Shalizi (in press) and their subsequent commentaries frame the debates clearly. Our view is that none of these alternative approaches offers the ability to state evidence for invariances and effects in as convincing and as clear a manner as does Bayes factors. Additional discussion is provided in Morey, Romeijn, and Rouder (in press).

UNIVARIATE REGRESSION

Consider the previously introduced regression of hominin cranial capacity onto environmental parasite load. Let x_i and y_i denote the parasite load and cranial capacity, respectively, corresponding to the i th skull, $i = 1, \dots, N$. The linear regression model is

$$\mathcal{M}_1 : y_i = \mu + \alpha(x_i - \bar{x}) + \epsilon_i, \quad (2)$$

where μ is the grand mean; α is the slope; \bar{x} is the mean parasite load; and ϵ_i is an independent, zero-centered, normally distributed noise term with variance σ^2 . In the model, there are three free parameters: μ , σ^2 , and α . To make the situation simpler, for now we assume that μ and σ^2 are known and focus on unknown slope α . This assumption will be relaxed subsequently. Model \mathcal{M}_1 expresses a relationship between cranial capacity and parasite load and may be compared with the following null regression model where there is no relationship between cranial capacity and parasite load:

$$\mathcal{M}_0 : y_i = \mu + \epsilon_i. \quad (3)$$

To assess whether data support Model \mathcal{M}_1 or the Null Model \mathcal{M}_0 , we compute the Bayes factor between them. The key is the computation of the probability of the data under the competing models. This task is fairly easy for the null \mathcal{M}_0 because with μ and σ known, there are no parameters. In this case, the

probability of the data is simply given by the probability density of the data at the known parameter values

$$\pi(\mathbf{y} \mid \mathcal{M}_0) = \prod_i \phi([y_i - \mu]/\sigma),$$

where ϕ is the density function of a standard normal.⁴ Likewise, the probability of the data under the alternative is straightforward if the alternative is assumed to be a point hypothesis. For example, suppose we set $\alpha = 1$ in Model \mathcal{M}_1 . Then

$$\pi(\mathbf{y} \mid \mathcal{M}_1) = \prod_i \phi([y_i - \mu - (x_i - \bar{x})]/\sigma).$$

The Bayes factor is simply the ratio of these values,

$$B_{10} = \frac{\pi(\mathbf{y} \mid \mathcal{M}_1)}{\pi(\mathbf{y} \mid \mathcal{M}_0)} = \frac{\prod_i \phi([y_i - \mu - (x_i - \bar{x})]/\sigma)}{\prod_i \phi([y_i - \mu]/\sigma)},$$

and in this case the Bayes factor is the likelihood ratio.

Setting the alternative to a specific point is too constraining. It is more reasonable to think that the slope may take one of a range of possible values under the alternative. In Bayesian statistics, it is possible to specify the alternative as covering such a range. When the slope parameter takes a range of possible values, the probability of the data under \mathcal{M}_1 is

$$\pi(\mathbf{y} \mid \mathcal{M}_1) = \int_{\alpha} \pi(\mathbf{y} \mid \mathcal{M}_1, \alpha) \pi(\alpha \mid \mathcal{M}_1) d\alpha.$$

The term $\pi(\mathbf{y} \mid \mathcal{M}_1, \alpha)$ is the probability density or likelihood and in this case is $\prod_i \phi([y_i - \mu - \alpha(x_i - \bar{x})]/\sigma)$. The probability of the data under the model is the weighted average of these likelihoods, where $\pi(\alpha \mid \mathcal{M}_1)$ denotes the distribution of weights. This distribution serves as the *prior density* of α and describes the researcher's belief or uncertainty about α before observing the data. The specification of a reasonable function for $\pi(\alpha \mid \mathcal{M}_1)$ is critical to defining an alternative model and is the point where subjective probability enters Bayesian inference. Arguments for the usefulness of subjective probability are made most elegantly in the psychological literature by Edwards, Lindman, and Savage (1963), and the interested reader is referred there. We note that subjective probability stands on firm axiomatic foundations and leads to ideal rules about

⁴This density is $\phi(x) = (2\pi)^{-1/2} \exp(-x^2/2)$.

updating beliefs in light of data (Cox, 1946; De Finetti, 1992; Gelman et al., 2004; Jaynes, 1986).

SPECIFYING A PRIOR ON EFFECTS

The choice of prior, in this case $\pi(\alpha \mid \mathcal{M}_1)$, is critical for computing the Bayes factor. One school of thought in specifying priors, known as *objective* Bayesian school, is that priors should be chosen based on the theoretical properties of the resulting Bayes factors. We adopt this viewpoint in recommending priors for regression models. The three properties that the resulting Bayes factors exhibit are as follows:

- *Location and Scale Invariance.* The Bayes factor is location-scale invariant if it is unaffected by the location and scale changes in the unit of measure of the observations and covariates. For instance, if the observations are in a unit of temperature, the Bayes factor should be invariant to whether the measurement is made on the Kelvin, Fahrenheit, or Celsius scales.
- *Consistency.* The Bayes factor is consistent if it approaches the appropriate bound in the large-sample limit. If \mathcal{M}_1 holds, then the $B_{10} \rightarrow \infty$; conversely, if \mathcal{M}_0 holds, then $B_{10} \rightarrow 0$.
- *Consistent in Information.* For the the Bayes factors described here, the data affect the Bayes factor only through R^2 , the coefficient of determination. As R^2 approaches 1, the covariate accounts for all the variance, and the alternative is infinitely preferable to the null. The Bayes factor is considered consistent in information if $B_{10} \rightarrow \infty$ as $R^2 \rightarrow 1$ for all sample sizes $N > 2$.

Although it is common to call priors motivated by these considerations *objective*, the term may be confusing. It is important to note that these priors are subjective and convey specified prior beliefs about the alternative under consideration. To avoid this confusion, we prefer the term *default* prior. The priors we present herein serve as suitable defaults in that they have desirable properties, are broadly applicable, and are computationally convenient.

The aforementioned properties place constraints on priors even in this simple univariate example with known μ and σ . The first property, that the Bayes factor should be invariant to the units of measurement, is met by reparameterizing the model in terms of a standardized effect measure. Model \mathcal{M}_1 may be rewritten as

$$\mathcal{M}_1 : y_i = \mu + \beta\sigma \left(\frac{x_i - \bar{x}}{s_x} \right) + \epsilon_i, \quad (4)$$

where s_x is the (population) standard deviation of \mathbf{x} and β is the *standardized effect* given by

$$\beta = \alpha s_x / \sigma.$$

It is straightforward to show that Equations (2) and (4) are reparameterizations of the same model. The parameter β describes how much a change in standard deviation units of x affects a change in standard deviation units of y . Note that β is simply a rescaling of α into a unitless quantity, and possible values of β include all real numbers. This standardization should not be confused with the more conventional standardization where data and covariates are transformed so that the slope is constrained to be between -1 and 1 (Kutner, Nachtsheim, Neter, & Li, 2004). In the more conventional standardization, the dependent variability is divided by a measure of total variability, whereas here the standardization is with respect to error variability σ .

With this reparameterization, a prior is needed for standardized slope β . Because β may take on any real value, one choice is a normal prior:

$$\beta \sim \text{Normal}(0, g), \quad (5)$$

where g is the variance in β and reflects the prior knowledge about the standardized effect. At first glance, it might seem desirable to set g to a large value, and this choice would reflect little prior information about the standardized effect. For example, if we set g arbitrarily large, then all values of β are about as equally likely a priori. If g is a billion, then the a priori probability that β is 1 million is nearly as large as $\beta = 1$. Yet, such a setting of g is unwise. For any reasonable data, the probability of the data given a slope as large as 1 million is vanishingly small, and placing weight on these values drives lower the average probability of the data given the model (Lindley, 1957). Hence, any model with a g that is unreasonably large will have low Bayes factor compared with the null model. Some authors contend that this dependence on priors is undesirable, but we disagree; we believe it is both natural and reasonable. The dependence is best viewed as a natural penalty for flexibility. Models with large values of g can account for a wide range of data: a model in which g is 1 billion can account for slopes that range over 10 orders of magnitude. Such a model is very flexible and should be penalized compared with one that can account for a more restricted range of slopes. Bayes factor contains a built-in penalty for this flexibility without recourse to asymptotic arguments, counting parameters, or out-of-sample considerations.

Because g affects the flexibility of the alternative, it should be chosen wisely. One approach is to set g to 1.0, and this prior underlies the Bayesian information criterion (BIC; see Raftery, 1995). This choice is computationally convenient, and the resulting Bayes factor obeys location-scale invariance and is consistent.

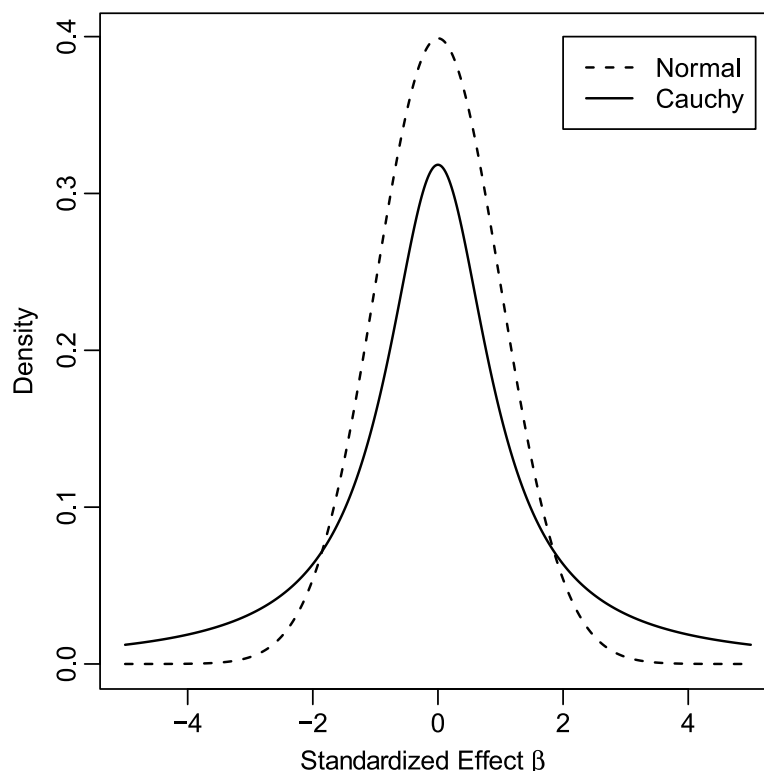


FIGURE 1 A comparison of Cauchy and normal prior densities on standardized effect β .

Unfortunately, the Bayes factor does not satisfy consistency-in-information: in this case, B_{10} asymptotes to a finite value as $R^2 \rightarrow \infty$ (Liang et al., 2008).

Another approach, proposed by Jeffreys (1961), is to place a Cauchy distribution prior on β :

$$\beta \sim \text{Cauchy}(s), \quad (6)$$

where s is the scale of the Cauchy that is set a priori, as discussed later. The Cauchy is a heavy-tailed distribution that encodes little knowledge of the standardized effect.⁵ The Cauchy and normal distributions are shown in Figure 1. With this Cauchy prior, the resulting Bayes factor, presented subsequently, obeys all three desirable theoretical properties (Liang et al., 2008).

In practice, researchers must set s , the scale factor of the Cauchy distribution. This value may be set by a priori expectations. When using the Cauchy prior, s describes the interquartile range of a priori plausible standardized slopes β . We

⁵The probability density function of a Cauchy random variable with scale parameter s is

$$\pi(x; s) = \frac{s}{(s^2 + x^2)\pi}.$$

find that $s = 1$ is a good default, and it specifies that the interquartile range of standardized slopes is from -1 to 1 . To better understand the specification of the Cauchy prior and the role of s , we express it in terms of the total proportion of variance accounted for by the covariate(s). Let R^2 and τ^2 denote the observed and true proportion of variance in y that is not error variance, respectively. Parameter τ^2 is a simple function of parameter β :

$$\tau^2 = \frac{\beta^2}{1 + \beta^2}.$$

Given this relationship, it is straightforward to calculate the implied prior on τ^2 , which is shown for two different values of s in Figure 2A. The solid line represents the prior density for $s = 1$; the dashed line represents the prior density for $s = .5$. When $s = 1$ the prior density is spread throughout the range of τ^2 . Smaller values of s correspond to greater concentration of mass near $\tau^2 = 0$. Panel B shows the corresponding cumulative prior probabilities: for $s = .5$, half of the prior probability is below $\tau^2 = .2$. When $s = 1$, half of the prior mass probability is below $\tau^2 = .5$. Thus, the $s = 1$ prior spreads out the prior probability more evenly across large and small values of τ^2 .

Another familiar quantity is the square root of the R^2 , the Pearson correlation coefficient r . The parameter $\sqrt{\tau^2}$ is in some sense analogous to a true Pearson correlation ρ .⁶ Figure 2C shows the implied prior densities on $\sqrt{\tau^2}$ for $s = 1$ and $s = .5$. Making s smaller concentrates the prior density nearer to $\sqrt{\tau^2} = 0$. As was the case with τ^2 , for $s = 1$ the prior density of $\sqrt{\tau^2}$ is more evenly spread out across the possible range. The corresponding cumulative density plot in Figure 2D shows this dependence clearly.

The Cauchy prior is computationally convenient in the univariate regression case. Unfortunately, the use of independent Cauchy priors on standardized effects proves computationally inconvenient in the multivariate case. Zellner and Siow (1980) made use of the following relationship between the Cauchy and normal to improve computations. The Cauchy distribution results from a continuous mixture of normals as follows: Reconsider the normal prior on standardized effects, but treat g as a random variable rather than as a preset constant:

$$\beta|g \sim \text{Normal}(0, g). \quad (7)$$

Further, let g be distributed as

$$g \sim \text{Inverse Gamma}(1/2, s^2/2). \quad (8)$$

⁶Note that because the x_i are not considered random, the correlation between x and y is not defined; however, we can still interpret $\sqrt{\tau^2}$ as analogous to a correlation.

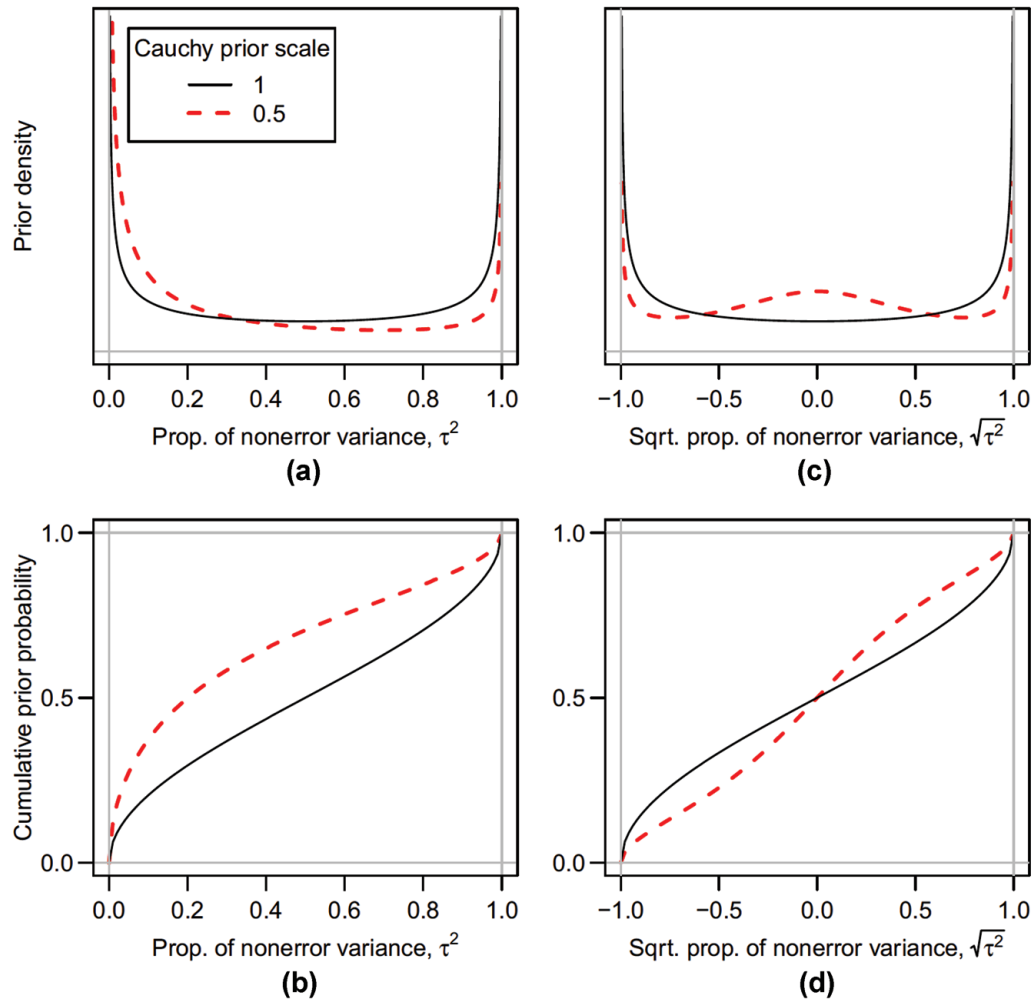


FIGURE 2 Implied prior distributions on the true proportion of variance (τ^2) from the regression. **A & B:** Prior density and prior cumulative distribution function. The solid and dashed lines are for prior Cauchy scales of $s = 1$ and $s = .5$, respectively. **C & D:** Implied prior density and CDF for $\sqrt{\tau^2}$, respectively (color figure available online).

The inverse gamma describes the distribution of the reciprocal of a gamma-distributed random variable.⁷ The two parameters are shape, which is fixed to $1/2$, and scale, which is $s^2/2$. The marginal prior on β may be obtained by

⁷The probability density function of an inverse gamma random variable with shape a and scale b is

$$\pi(x; a, b) = \frac{b^a}{\Gamma(a)x^{a+1}} \exp\left(-\frac{b}{x}\right),$$

where $\Gamma()$ is the gamma function (Abramowitz & Stegun, 1965).

integrating out g , and the result is a Cauchy prior with scale s . Hence, the hierarchical prior defined by Equations (7) and (8) is equivalent to the Cauchy prior in Equation (6). This expression of the Cauchy prior as a continuous mixture of normals is used in the following development for multiple regression.

The prior on effect size, whether expressed as a Cauchy or a mixture of normals, may be termed *weakly informative*. It is not noninformative because the researcher must commit to a scale s a priori. Yet, it is not strongly informative as there are a large range of effect sizes that are compatible with any point setting. It represents vague or diffuse knowledge about the range of alternatives.

THE BAYES FACTOR FOR MULTIPLE REGRESSION

The previous example was useful for discussing the role of priors in Bayes factor but was limited because we assumed known μ and σ^2 . Moreover, the example allowed only a single covariate, whereas most research includes multiple covariates. We now consider the case for more than one covariate and without assuming known intercept or variance parameters. A model for N observations with p covariates is

$$y_i = \mu + \alpha_1(x_{1i} - \bar{x}_{1.}) + \alpha_2(x_{2i} - \bar{x}_{2.}) + \cdots + \alpha_p(x_{pi} - \bar{x}_{p.}) + \epsilon_i, \quad i = 1, \dots, N,$$

where $(\alpha_1, \dots, \alpha_p)$ are slopes and $\bar{x}_{p.}$ is the mean value of the p th covariate across the N observations. It is most convenient to express the model in matrix notation. Let \mathbf{X}_1 be a centered vector of values for the first covariate, $\mathbf{X}_1 = (x_{11} - \bar{x}_{1.}, \dots, x_{1N} - \bar{x}_{1.})'$, and let $\mathbf{X}_2, \dots, \mathbf{X}_p$ be defined similarly for the remaining covariates. Also let \mathbf{X} , the centered design matrix, be $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_p)$. Let $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_p)'$ be a vector of slopes. The model, denoted \mathcal{M}_1 , is

$$\mathcal{M}_1: \quad \mathbf{y} = \mu \mathbf{1}_N + \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \quad (9)$$

where \mathbf{y} is the vector of observations; $\boldsymbol{\epsilon}$ is the vector of independent, zero-centered, normally distributed errors; and $\mathbf{1}_N$ is vector with entries of 1.0 and with length N . We compare this model with a null model with no covariates:

$$\mathcal{M}_0: \quad \mathbf{y} = \mu \mathbf{1}_N + \boldsymbol{\epsilon}. \quad (10)$$

To quantify the support for the models, we compute the Bayes factor between \mathcal{M}_1 and \mathcal{M}_0 . To compute this Bayes factor, appropriate priors are needed for parameters μ , σ^2 , and $\boldsymbol{\alpha}$.

We follow here a fairly standard approach first introduced by Jeffreys (1961), expanded by Zellner and Siow (1980), and studied by Liang et al. (2008). The key motivation behind this approach is that it yields Bayes factors with the

desirable theoretical properties discussed previously. Parameters μ and σ serve to locate and scale the dependent measure. Fortunately, because these location and scale parameters are not the target of inference and enter into all models under consideration, it is possible to place broad priors on them that convey no prior information.⁸ The key parameters for inference are the slopes, which occur in some models and not in others. In the previous example, we placed a weakly informative prior on standardized slope, where the slope was standardized by the variability in the covariate and the variability in the dependent measure. We retain this standardization:

$$\begin{aligned}\alpha|g &\sim \text{Normal}(\mathbf{0}, g\sigma^2(\mathbf{X}'\mathbf{X}/N)^{-1}), \\ g &\sim \text{Inverse Gamma}(1/2, s^2/2).\end{aligned}$$

The term g is the variance of the standardized slope, the term σ^2 scales this variance to the scale of the dependent measure, and the term $(\mathbf{X}'\mathbf{X}/N)^{-1}$ scales the slope by the variability of the covariates. An inverse-gamma (shape of 1/2, scale of $s^2/2$) mixture of g s is used as before, and the marginal prior on α is the multivariate Cauchy distribution (Kotz & Nadarajah, 2004).

The Bayes factor is the integration of the likelihood with respect to the priors. For the null model, there are two parameters, μ and σ^2 , and the integration may be done in closed form. For the alternative, integration must occur with respect to the effects α and the hyperprior g . The integration for all of these may be done in closed form with the exception of g . The resulting expression for Bayes factor between Model \mathcal{M}_1 and Null Model \mathcal{M}_0 is

$$\begin{aligned}B_{10}(s) &= \int_0^\infty (1+g)^{(N-p-1)/2} [1+g(1-R^2)]^{-(N-1)/2} \\ &\quad \left(\frac{s\sqrt{N/2}}{\Gamma(1/2)} g^{-3/2} e^{-Ns^2/2g} \right) dg,\end{aligned}\tag{11}$$

where R^2 is the unadjusted proportion of variance accounted for by the covariates. This formula is relatively straightforward to evaluate. First, note that the data only appear through R^2 , which is conveniently computed in all statistics packages. Second, the integration is across a single dimension (defined by g) and consequently may be performed to high precision by numerical methods such as Gaussian quadrature. We provide a web applet, called the *Bayes Factor*

⁸The location-scale invariant prior is $\pi(\mu, \sigma^2) = 1/\sigma^2$ (Jeffreys, 1961).

Calculator (<http://pcl.missouri.edu/bf-reg>), to compute the Bayes factor in Equation (11). Researchers simply provide R^2 , the sample size (N), and the number of covariates (p); the calculator returns B_{f0} . In practice, researchers will need to choose s , with smaller values corresponding to smaller expected effect sizes. Throughout the remainder of this article, we set $s = 1$.

Figure 3 shows some of the characteristics of the default Bayes factor in Equation (11) for the simple one-predictor case (e.g., the regression of cranial capacity onto parasite load). The figure shows critical R^2 statistic corresponding to Bayes factors of $B_{10} = (1/10, 1, 3, 10)$, respectively, for a range of sample sizes. The dashed line shows the values of R^2 needed for significance at the .05 level. The difference in calibration between significance tests and these default Bayes factors is evident. For small sample sizes, say between 10 and 100, critical significance levels correspond to Bayes factors that are between 1/3 and 3, that is, those that convey fairly equivocal evidence. The situation is even more discordant as sample size increases. For $N < 3000$, R^2 values that would correspond to small p values (indicating a rejection of the null) also correspond to Bayes factor values that favor the null ($B_{01} > 10$). In summary, Bayes factors are calibrated differently than p values. Inference by p values tends to overstate the evidence against the null, especially for large samples. We consider these calibration differences further in the Discussion.

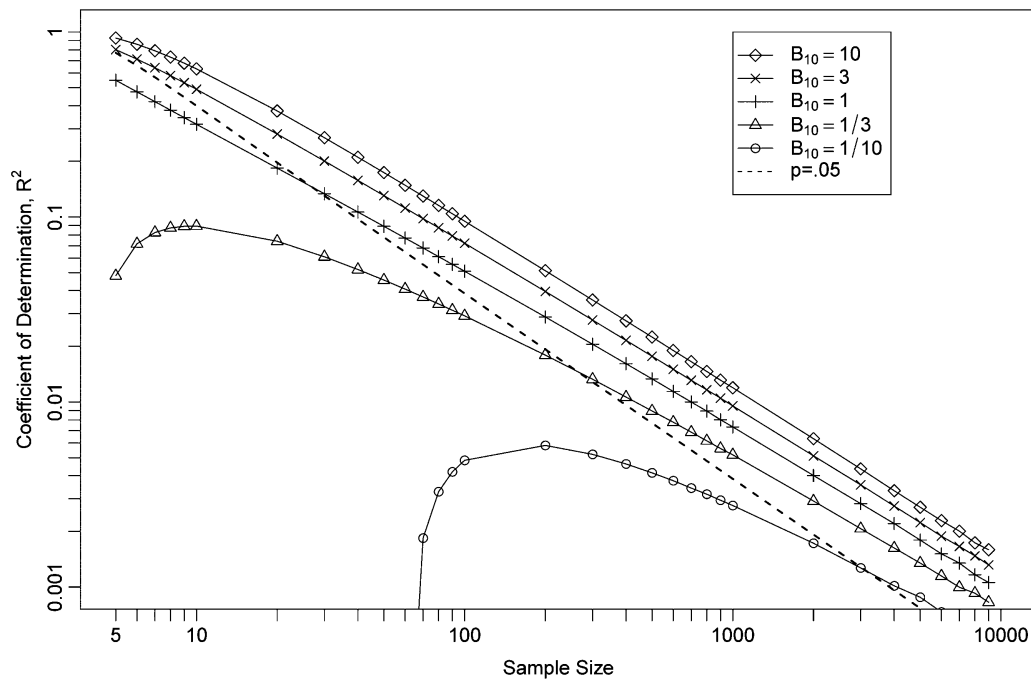


FIGURE 3 Critical values of R^2 for different levels of evidence as a function of sample size. Solid lines are for Bayes factors at specified levels; dashed line is for a p value of .05.

AN APPLICATION

To illustrate the use of Bayes factors, we reanalyze data from Bailey and Geary (2009). We considered four theoretically important covariates: *global temperature* when the hominid lived, *local climate variation*, *parasite load*, and *population density*. A modern approach to multiple regression is model comparison or selection (Hocking, 1976) in which models, represented by sets of covariates, are compared with one another and the best model is identified. There exist a number of strategies for identifying the best among several models, and the process is referred to as subset selection (Miller, 2002). A comparison of different subset selection strategies is outside the scope of this article. Most psychologists take a simultaneous approach, in which they compare a large number of models at once, or a testing approach, in which they sequentially test the effect of specific covariates to answer specific questions. Fortunately, the use of Bayes factors for inference is compatible with either of these approaches as well as with other selection strategies such as top-down or bottom-up stepwise regression. We cover here how Bayes factors may be used in simultaneous comparisons of many models and for testing specific covariates.

Comparing Models Simultaneously

Table 1 shows R^2 for 15 models, formed by considering all the possible submodels of the four covariates. It also provides B_{m0} , the Bayes factor between each submodel and the null model, and these were computed individually with the web applet at pcl.missouri.edu/bayesfactor. We have also computed the Bayes factor of each model relative to the full model B_{mf} , which may be obtained by $B_{mf} = B_{m0}B_{0f}$. For the cranial capacity analysis, the evidence for various models is shown in Figure 4. The model with the greatest evidence is \mathcal{M}_4 , which is comprised of all covariates except for local climate variation.

Testing Covariates

In multivariate settings, psychologists often test covariates, one at a time. Bailey and Geary (2009), for example, computed a t value and corresponding p value for each slope term. In this set, such tests reveal a significant effect of population density ($t(170) = 9.2$, $p \approx 0$) and global temperature ($t(170) = 6.9$, $p \approx 0$) as well as nonsignificant effects of local climate ($t(170) = .09$, $p \approx .93$) and parasite load ($t(170) = 1.47$, $p \approx .14$). We discuss here how analogous comparisons may be performed with Bayes factor. An advantage of the Bayes factor is that one may state positive evidence for a model without a covariate (the null in this case), which is not possible in conventional testing.

TABLE 1
Bayes Factor Analysis of Hominid Cranial Capacity (Data From Bailey & Geary, 2009)

	<i>Model</i>	R^2	B_{m0}	B_{mf}
\mathcal{M}_f	Local+Global+Parasites+Density	.7109	3.54×10^{41}	1
\mathcal{M}_1	Local+Global+Parasites	.567	5.56×10^{27}	1.57×10^{-14}
\mathcal{M}_2	Local+Global+Density	.7072	1.56×10^{42}	4.41
\mathcal{M}_3	Local+Parasites+Density	.6303	3.82×10^{33}	1.08×10^{-8}
\mathcal{M}_4	Global+Parasites+Density	.7109	4.59×10^{42}	12.97
\mathcal{M}_5	Local+Global	.5199	1.02×10^{25}	2.88×10^{-17}
\mathcal{M}_6	Local+Parasites	.2429	1.23×10^8	3.47×10^{-34}
\mathcal{M}_7	Local+Density	.6258	1.84×10^{34}	5.20×10^{-8}
\mathcal{M}_8	Global+Parasites	.5642	4.02×10^{28}	1.14×10^{-13}
\mathcal{M}_9	Global+Density	.7069	2.17×10^{43}	61.03
\mathcal{M}_{10}	Parasites+Density	.6298	4.60×10^{34}	1.30×10^{-7}
\mathcal{M}_{11}	Local	.091	220	6.21×10^{-40}
\mathcal{M}_{12}	Global	.5049	1.10×10^{25}	3.11×10^{-17}
\mathcal{M}_{13}	Parasites	.2221	1.28×10^8	3.62×10^{-34}
\mathcal{M}_{14}	Density	.6244	2.29×10^{35}	6.47×10^{-7}

Note. Local = local climate; Global = global temperature; Parasites = parasite load; Density = population density.

The test of a covariate may be performed by computing a Bayes factor for a model that includes the covariate against one that does not. For example, to test the slope of population density, we compare the full model with one in which density is not present: Model \mathcal{M}_1 in Table 1. The Bayes factor of interest, B_{f1} , is given by B_{f0}/B_{10} . Plugging in the values from Table 1 yields

$$B_{f1} = \frac{B_{f0}}{B_{10}} = \frac{3.54 \times 10^{41}}{5.56 \times 10^{27}} = 6.37 \times 10^{13},$$

meaning that there is overwhelming evidence for a relationship between population density and brain size. The same procedure is applied to test for the effect of the other covariates. For global temperature, the relevant Bayes factor is B_{f3} , which evaluates to 9.26×10^7 . This value indicates overwhelming evidence for a global climate effect. For local climate and parasite load, the relevant Bayes factors evaluate to about 4.5 and 13, respectively, favoring the three-parameter models missing the covariate over the four-parameter model that contains it. Hence there is evidence for a lack of an effect of local climate and parasite load. Note that these statements about the evidence for a lack of an effect are conceptually different from conventional statements with p values about a lack of evidence for an effect.

Psychologists have developed a compact style for stating test results, and this style is easily extended to Bayes factors. For example, the results of the four tests

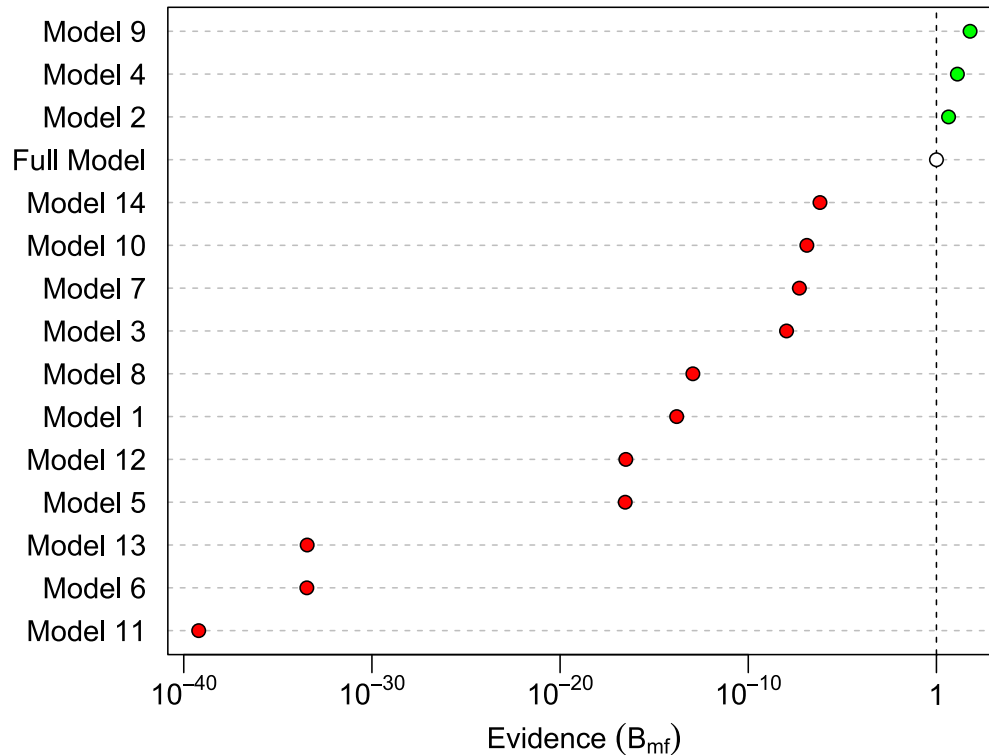


FIGURE 4 Bayes factor evidence for 15 models (see Table 1) (color figure available online).

may be compactly stated as follows: “*Bayes factor analysis with default mixture-of-variance priors, and with reference to the full model with four covariates, indicates evidence for the effect of population density ($B_{10} = 6.4 \times 10^{13}$) and global temperature ($B_{10} = 9.3 \times 10^7$) and evidence for a lack of effect of local climate ($B_{01} = 12.9$) and a lack of effect of parasite load ($B_{01} = 4.4$).*” When comparing individual effects, subscripts may be used to indicate the direction of the comparison, whether the Bayes factor is the evidence for the full model relative to the appropriate restriction (i.e., B_{10}), or the reverse (i.e., B_{01}). We recommend researchers report whichever Bayes factor is greater than 1.0. In our experience, odds measures are more easily understood when the larger number is in the numerator. For example, the statement $B_{10} = 16$ is more easily understood than $B_{01} = .0625$ even though the two are equivalent.

Model comparison by covariate testing should be used with an abundance of caution. A less than ideal view may result when the covariates are not orthogonal. In the current application, covariate testing and a simultaneous comparison of all models lead to the same conclusion that global temperature and population density are needed covariates while local climate and parasite load are not. However, this congruity is not guaranteed and incongruities will occur when

covariates are correlated. Covariate testing fails to account for such correlations and is less desirable than simultaneous comparison of all models when there is a large degree of conlinearity.

ADDING VALUE THROUGH PRIOR ODDS

Bayes factors describe the relative probability of data under competing positions. In Bayesian statistics, it is possible to evaluate the odds of the positions themselves conditional on the data:

$$\frac{\Pr(H_1 | y)}{\Pr(H_0 | y)} = B_{10} \times \frac{\Pr(H_1)}{\Pr(H_0)},$$

where $\Pr(H_1|y)/\Pr(H_0|y)$ and $\Pr(H_1)/\Pr(H_0)$ are called posterior and prior odds, respectively. The prior odds describe the beliefs about the hypotheses before observing the data. The Bayes factor describes how the evidence from the data should change beliefs. For example, a Bayes factor of 100 indicates that posterior odds should be 100 times more favorable to the alternative than the prior odds. If all models are equally probable a priori, then their posterior odds will be numerically equal to the Bayes factors. There is no reason to suppose, however, that all models will always have equal prior odds. A model with covariates that have well-understood mechanisms underlying the relationship between the predictors and the dependent variable should have greater prior odds than one for a covariate in which this mechanism is lacking. Likewise there is little reason to suspect that all readers will have the same prior odds. For any proposed relationship, some readers may be more skeptical than others. Even when researchers disagree on priors, they may still agree on how to change these priors in light of data.

The phenomena of extrasensory perception (ESP) provides a suitable example to highlight the difference between posterior odds and Bayes factors. ESP has become topical with the recent reports of Bem (2011) and Storm, Tressoldi, and Di Rosio (2010). Bem reports nine experiments where he claims evidence that participants are able to literally *feel the future* or have knowledge about future events that could not possibly be known. For example, in Bem's Experiment 1, participants were presented two closed curtains—one concealing an erotic picture and the other nothing—and were asked to identify which curtain concealed the erotic picture. After the participants made their choice, a computer randomly chose where to place the image. Amazingly, participants had above-chance accuracy (53.1%, $t(99) = 2.51$, $p < .01$). Bem concluded that participants' choices were guided to some degree by future events indicating that people could feel the future. Storm et al. concluded that telepathy exists

through a meta-analysis of 67 recent telepathy experiments. They examined experiments in which “senders” had to mentally broadcast stimulus information to isolated “receivers,” who then reported which stimulus was presented. Overall performance was significantly above the relevant chance baseline (Stouffer $Z = 5.48$, $p \approx 2 \times 10^{-8}$ for ganzfeld experiments).

We have performed Bayes factor reanalyses of the data in both of these publications. Our Bayes factor meta-analysis of Bem’s (2011) data yielded a Bayes factor of 40-to-1 in favor of effects consistent with feeling the future (Rouder & Morey, 2011). Likewise our Bayes factor meta-analysis of the data analyzed in Storm et al. (2010) yielded values as high as 330-to-1 in favor of effects consistent with telepathy (Rouder, Morey, & Province, in press). To readers who a priori believe ESP is as likely as not, these values are substantial and important. We, however, follow Bem (2011) and Tressoldi (2011), who cite Laplace’s famous maxim that extraordinary claims require extraordinary evidence. ESP is the quintessential extraordinary claim because there is a pronounced lack of any plausible mechanism. Accordingly, it is appropriate to hold very low prior odds of ESP effects, and appropriate odds may be as extreme as millions, billions, or even higher against ESP. When these low prior odds are multiplied against Bayes factors of 330-to-1, the resultant posterior odds still favor an interpretation against ESP.

The distinction between posterior odds and Bayes factors provides an ideal mechanism for adding value to findings in a transparent manner. Researchers should report the Bayes factor as the evidence from the data. Readers may update their prior odds simply by multiplying (Good, 1979; Jeffreys, 1961). Sophisticated researchers may add guidance and value to their analysis by suggesting prior odds, or ranges of prior odds, much as we do in interpreting Bayes factors from ESP experiments. By reporting Bayes factor separate from posterior odds, researchers ensure transparency between evidence and value-added adjustment.

Finally, researchers (and readers) need not feel obligated to posit prior odds to interpret the Bayes factor. The Bayes factor stands self-contained as the relative probability of data under hypotheses and may be interpreted as such without recourse to prior odds.

GENERAL DISCUSSION

Bayes factors have not become popular, and we routinely encounter critiques against their adoption. Our goal is not to provide a comprehensive defense of Bayes factors (more comprehensive treatments may be found in Berger & Sellke, 1987; Edwards et al., 1963; Wagenmakers, 2007). Instead, we highlight what we consider the most common critiques. This consideration provides a more

complete context for those considering Bayes factors as well as highlighting limitations on its use.

Concerns

Concern #1: The null model is never exactly true. One critique of significance testing rests on the assumption that point null hypotheses are never true to arbitrary precision (Cohen, 1994; Meehl, 1990). According to this proposition, if one collects sufficient data, then the null will *always* be proved wrong. The consequence is that testing point nulls is a suspect intellectual endeavor, and greater emphasis should be placed on estimating effect sizes. Although we are not sure whether the null is truly always false, consideration of the critique helps sharpen the role of Bayes factor as follows:

Bayes factor answers the question which model best describes the data rather than which model most likely holds or which is most likely true. In this spirit, an analyst may speak of the null as being a very good description for the phenomena at hand without commitment to whether the null truly holds to arbitrary precision. Figure 5 highlights the descriptive nature of Bayes factor. The figure shows the default Bayes factor when the observed value of R^2 is .01 (solid line labeled “Point”). As can be seen, the Bayes factor favors the null model for small sample sizes. This behavior is expected as given the resolution of the data, an observed R^2 of .01 is well described by the null. In fact, the evidence for the null increases as sample size increases, but only up to a point.

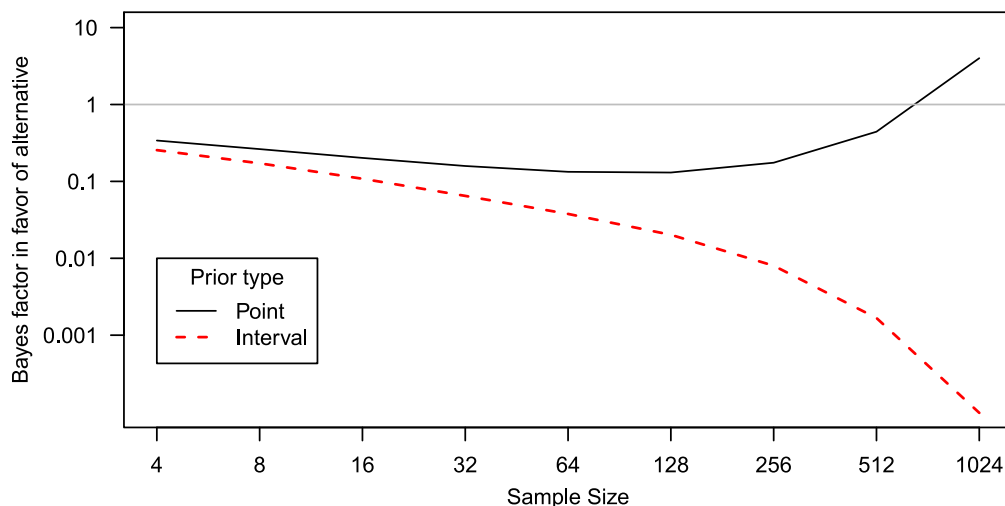


FIGURE 5 Bayes factors as a function of sample size for a small effect ($R^2 = .01$). The solid line shows the case where the null hypothesis is a point; the dotted line shows the same where the null hypothesis is uniformly distributed across a small interval ($\tau^2 < .04$) (color figure available online).

Once the sample size becomes quite large, the data afford the precision to resolve even small effects, and the null is no longer a good description. It is this nonmonotonic behavior that highlights the descriptive nature of Bayes factor. The null may be a good description of the data for moderate sample sizes even when it does not hold to arbitrary precision.

A more conventional testing approach to accommodate null hypotheses that may not hold to arbitrary precision is statistical equivalence testing (Rogers, Howard, & Vessey, 1993; Wellek, 2003). In statistical equivalence testing, the analyst defines a small range of effects around the point null that are to be treated as equivalent to the no-effect null. The usefulness or desirability of equivalence regions are orthogonal to consideration of Bayes factors versus other methods of inference. If an analyst desires these intervals, then null model and alternative models may be recast. Morey and Rouder (2011) offer a range of solutions, including models in which the null has the support of a small interval. Figure 5 also shows the Bayes factor for the small effect $R^2 = 0.01$ under Morey and Rouder's interval null setup (dashed line labeled "Interval"). Here, under the null the true values of τ^2 have support on the interval $\tau^2 < .04$, and under the alternative, there is support for the interval $\tau^2 > .04$ (see Morey & Rouder, 2011, for details). Because the posterior distribution for τ for an observed $R^2 = .01$ is solidly in the equivalence region, the interval Bayes factor favors the null hypothesis. The certainty that τ^2 is within the equivalence region increases as the sample size increases, and the Bayes factor increases in turn. Morey and Rouder's development serves to highlight the flexibility of Bayes factor as a suitable tool for comparing the descriptive value of models even in an imperfect world where nuisance factors are unavoidable.

Concern #2: All the models are wrong. The critique that the null never holds to arbitrary precision may be generalized to the critique that no model holds to arbitrary precision. Consideration of this position leads to a decreased emphasis on testing and selection and an increased emphasis on estimation of effect sizes as well as graphical and exploratory methods to uncover structure in data (e.g., Gelman, 2007; Gelman & Rubin, 1995; Velleman & Hoaglin, 1981). Indeed, the American Psychological Association Task Force on Statistical Inference (Wilkinson & the Task Force on Statistical Inference, 1999) determined that testing is used perhaps too frequently in psychology and that researchers may achieve a better understanding of the structure of their data from these alternative approaches.

Although we are not sure if models are always wrong, we consider this critique useful to potential users of Bayes factors. It is fair to ask about the limits of what may be learned from a comparison of wrong models. There is no global or broadly applicable answer, and the rationale for comparing wrong models will depend critically on the substantive context. In our experience models embed

useful theoretical positions, and the comparison among them provides useful insights for theoretical development that may not be as readily available with graphical methods or with estimation of effect sizes. The gruesome phrase about the multiple ways of skinning cats applies. Not everyone needs to do it the same way, and different methods are better for different cats. Analysts should consider, however, whether the method they chose is best for their cat at hand.

As a rule of thumb, testing and selection seem most warranted when the models faithfully approximate reasonable and useful theoretical positions. These situations that most license testing seem especially conducive to Bayes factor assessment. Our outlook is well captured by the following compact tagline, adopted from a current beer commercial:⁹ "I don't always select among models, but when I do, I prefer Bayes factor."

Concern #3: Bayes factors are subjective, other Bayesian methods are less subjective, and frequentist methods are objective. It may seem a matter of common sense to worry about subjectivity with Bayes factor. In fact, there are a number of Bayesians who worry that the Bayes factor is too sensitive to the prior, and inference is better done with Bayesian analogs of classical methods. Kruschke (2011), for example, recommends plotting posteriors of effects and checking whether they cover null values. Such a method is fairly robust to the prior as the sample size increases; Bayes factor in contrast maintains a dependency on the prior even in the large-sample limit. At first glance it seems that Bayes factor is more subjective than other Bayesian methods or than frequentist methods. We think this critique, however, does not accurately capture the constraints in model selection.

The important question is not one of subjectivity but one of calibration, whether model selection should be calibrated with respect to the null alone, such as in the computation of p values, or should be calibrated with respect to the null and a specified alternative, as in the Bayes factor. Our argument is that model selection should be calibrated with respect to both models rather than just one, and any method that does so, frequentist or Bayesian, will necessarily be subjective.

Let's consider subjectivity in a frequentist context by examining the difference between null hypothesis testing and power analysis. Power is computed by specifying a point alternative. With this specification, the analyst can compute and control Type II error rates and accept the null hypothesis in a principled fashion. Yet, the specification of a point alternative is subjective.¹⁰ A similar adaptation, discussed by Raftery (1995), is that the analyst may choose α

⁹The commercial for Dos Equis brand beer ends with the tagline, "I don't always drink beer, but when I do, I prefer Dos Equis." See <http://www.youtube.com/watch?v=8Bc0WjTT0Ps>

¹⁰This was one of Fisher's arguments against considering alternatives (Fisher, 1955).

based on sample size. For consistent testing, α cannot remain at a constant .05 level in the large-sample limit. Instead, it should asymptotically approach zero. The schedule of this decrease, however, implies subjective knowledge of the alternative. In both cases, consistency in testing is obtained only after subjective specification of an alternative. We think both power and judiciously setting α are vast improvements over null hypothesis testing because consistency may be achieved and positive evidence for the null may be stated.

Conversely, there are Bayesian methods that are seemingly objective yet are poorly calibrated. Consider, for example, the inference underlying one-sample t tests in which one is trying to decide whether or not the mean of a normal is zero. Bayesian analysts may certainly place the noninformative prior on μ and σ^2 and compute a posterior for μ conditional on the observed data. Furthermore, one can compute the q th-percent credible interval, which in this case matches exactly the q th-percent confidence interval. Moreover, inference at a Type I error rate of $\alpha = 1 - q/100$ may be performed by observing whether or not this interval covers zero. Of course, this inference yields results identical to the one-sample t test, and, consequently, inherits its poor calibration (Sellke, Bayarri, & Berger, 2001). In particular, the method provides no principled approach to stating evidence for the null should it hold, nor is it consistent when the null holds.

The critical question is whether inference is calibrated with respect to the null model alone or to both the null and a specified alternative. If a method is calibrated with respect to the null alone, then it tends to overstate the evidence against the null because the null may be rejected even when there is no more evidence for it than for reasonable alternatives (this critique is made by both Bayesians, such as Edwards et al., 1963, and frequentists, such as Hacking, 1965, and Royall, 1997). One example of this miscalibration is the asymmetric nature of inconsistency for methods that calibrate with reference to the null. For instance, consider AIC for a simple univariate regression model. AIC, like significance testing, requires no commitment to a specified alternative. If there is any true slope, then in the large-sample limit, the distribution of difference in deviance grows without bound leading to the correct selection of the alternative. In the case that the null holds, however, the difference in deviance between the null and alternative is a chi-squared distribution with one degree of freedom, and the probability of wrongly selecting the alternative is .157. This error rate holds for all sample sizes and even in the large sample limit.¹¹ Committing to an alternative alleviates these problems by alleviating the asymmetry between null and

¹¹This should not be read as a criticism of AIC. The goal of AIC—to find a model that minimizes the Kullback-Leibler divergence between the predicted data distribution and the true data distribution, assuming all models are wrong—is simply a different goal from Bayes factor, which strives to quantify the relative evidence for two competing models (Burnham & Anderson, 2002).

alternative. Principled and consistent model selection, the type that allows the analyst to state evidence for the null or alternative, requires the commitment to well-specified alternatives. This commitment is subjective regardless of whether one uses Bayesian or frequentist conceptions of probability. One advantage of the Bayesian approach is that it handles this subjectivity in a formal framework, but other approaches are possible.

Bayes factors depend on the choice of prior, and the Bayes factor values will assuredly vary across different prior distributions. The default priors developed here come from the objective Bayesian school where priors are chosen to yield Bayes factors with desired theoretical properties. Nonetheless, there is need for subjectivity even within these priors as the analyst must set the scale parameter s of the Cauchy prior on the standardized slopes (we recommend $s = 1.0$ as a default, but the choice is to an extent arbitrary and matches our subjective a priori expectations). One could study how the choice of s affects the Bayes factor for various sample sizes and various values of R^2 ; this might, for instance, have value if one wished to show what one would have to believe to come to a conclusion different from the one reached. That such an experiment could be performed should not be used, however, as an argument against Bayes factors or subjectivity. Bayes factors are neither “too subjective” nor “not too subjective.” Instead, there is simply a degree of subjectivity needed for principled model selection.

Subjectivity should not be reflexively feared. Many aspects of science are necessarily subjective. Notable aspects of subjectivity include the operationalization of concepts, evaluation of the quality of previous research, and the interpretation of results to draw theoretical conclusions. Researchers justify their subjective choices as part of routine scientific discourse, and the wisdom of these choices is evaluated as part of routine review. In the same spirit, users of Bayes factors should be prepared to justify their choice of priors much as they would be prepared to justify other aspects of research. We recommend the default priors presented here because they result in Bayes factors with desirable properties, are broadly applicable in social science research, and are computationally convenient.

REFERENCES

- Abramowitz, M., & Stegun, I. A. (1965). *Handbook of mathematical functions: With formulas, graphs, and mathematical tables*. New York, NY: Dover.
- Aitkin, M. (1991). Posterior Bayes factors. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(1), 111–142. Retrieved from <http://www.jstor.org/stable/2345730>
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723.

- Bailey, D. H., & Geary, D. C. (2009). Hominid brain evolution: Testing climactic, ecological, and social competition models. *Human Nature*, 20, 67–79.
- Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100, 407–425. Retrieved from <http://dx.doi.org/10.1037/a0021524>
- Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, 82(397), 112–122. Retrieved from <http://www.jstor.org/stable/2289131>
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information theoretic approach* (2nd ed.). New York, NY: Springer-Verlag.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 997–1003.
- Congdon, P. (2006). *Bayesian statistical modelling* (2nd ed.). New York, NY: Wiley.
- Cox, R. T. (1946). Probability, frequency and reasonable expectation. *American Journal of Physics*, 14, 1–13.
- De Finetti, B. (1992). *Probability, induction and statistics: The art of guessing*. New York, NY: Wiley.
- Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70, 193–242.
- Fechner, G. T. (1966). *Elements of psychophysics*. New York, NY: Holt, Rinehart and Winston.
- Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society: Series B (Methodological)*, 17, 69–78. Retrieved from <http://www.jstor.org/stable/2983785>
- Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, 116, 439–453. Retrieved from <http://psycnet.apa.org/doi/10.1037/a0015251>
- Gelman, A. (2007). Comment: Bayesian checking of the second levels of hierarchical models. *Statistical Science*, 22, 349–352. Retrieved from <http://www.jstor.org/stable/27645839>
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). London, UK: Chapman & Hall.
- Gelman, A., & Rubin, D. B. (1995). Avoiding model selection in Bayesian social research. In P. V. Marsden (Ed.), *Sociological methodology 1995* (pp. 165–173). Oxford, UK: Blackwell.
- Gelman, A., & Shalizi, C. R. (in press). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*.
- Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, 17, 295–314.
- Good, I. J. (1979). Studies in the history of probability and statistics: XXXVII. A. M. Turing's statistical work in World War II. *Biometrika*, 66(2), 393–396. Retrieved from <http://www.jstor.org/stable/2335677>
- Hacking, I. (1965). *Logic of statistical inference*. Cambridge, UK: Cambridge University Press.
- Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics*, 32, 1–49. Retrieved from <http://www.jstor.org/stable/2529336>
- Jaynes, E. (1986). Bayesian methods: General background. In J. Justice (Ed.), *Maximum-entropy and Bayesian methods in applied statistics*. Cambridge, UK: Cambridge University Press.
- Jeffreys, H. (1961). *Theory of probability* (3rd ed.). New York, NY: Oxford University Press.
- Kass, R. E. (1993). Bayes factors in practice. *The Statistician*, 42, 551–560.
- Kotz, S., & Nadarajah, S. (2004). *Multivariate t distributions and their applications*. Cambridge, UK: Cambridge University Press.
- Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, 6, 299–312.
- Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2004). *Applied linear statistical models*. Chicago, IL: McGraw-Hill/Irwin.

- Laplace, P. S. (1986). Memoir on the probability of the causes of events. *Statistical Science*, 1(3), 364–378. Retrieved from <http://www.jstor.org/stable/2245476> (Original work published 1774).
- Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103, 410–423. Retrieved from <http://pubs.amstat.org/doi/pdf/10.1198/016214507000001337>
- Lindley, D. V. (1957). A statistical paradox. *Biometrika*, 44, 187–192.
- Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, 56, 362–375. Retrieved from <http://dx.doi.org/10.1016/j.jmp.2008.03.002>
- Low, B. S. (1990). Marriage systems and pathogen stress in human societies. *American Zoologist*, 30, 325–339.
- Mallows, C. L. (1973). Some comments on c_p . *Technometrics*, 15, 661–675.
- Masin, S. C., Zudini, V., & Antonelli, M. (2009). Early alternative derivations of Fechner's law. *Journal of the History of the Behavioral Sciences*, 45(1), 56–65.
- Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, 66, 195–244.
- Miller, A. J. (2002). *Subset selection in regression* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.
- Morey, R. D., Romeign, W.-J., & Rouder, J. N. (2012). The humble Bayesian: Model checking from a fully Bayesian perspective. *British Journal of Mathematical and Statistical Psychology*.
- Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, 16, 406–419.
- Potts, R. (1998). Variability selection in hominid evolution. *Evolutionary Anthropology*, 7, 81–96.
- Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, 25, 111–163.
- Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate the equivalence between two experimental groups. *Psychological Bulletin*, 113, 553–565.
- Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, 18, 682–689. Retrieved from <http://dx.doi.org/10.3758/s13423-011-0088-7>
- Rouder, J. N., Morey, R. D., & Province, J. M. (in press). A Bayes-factor meta-analysis of recent ESP experiments: A rejoinder to Storm, Tressoldi, and Di Risio (2010). *Psychological Bulletin*.
- 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 and Review*, 16, 225–237. Retrieved from <http://dx.doi.org/10.3758/PBR.16.2.225>
- Royall, R. (1997). *Statistical evidence: A likelihood paradigm*. New York, NY: CRC Press.
- Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *American Statistician*, 55, 62–71.
- Shibley Hyde, J. (2005). The gender similarities hypothesis. *American Psychologist*, 60, 581–592.
- Sternberg, S. (1969). The discovery of processing stages: Extensions of Donder's method. In W. G. Kosner (Ed.), *Attention and performance ii* (pp. 276–315). Amsterdam, The Netherlands: North-Holland.
- Storm, L., Tressoldi, P. E., & Di Risio, L. (2010). Meta-analysis of free-response studies, 1992–2008: Assessing the noise reduction model in parapsychology. *Psychological Bulletin*, 136, 471–485. Retrieved from <http://dx.doi.org/10.1037/a0019457>
- Tressoldi, P. E. (2011). Extraordinary claims require extraordinary evidence: The case of non local perception, a classical and Bayesian review of evidences. *Frontiers in Quantitative Psychology and Measurement*. Retrieved from <http://dx.doi.org/10.3389/fpsyg.2011.00117>
- Velleman, P. F., & Hoaglin, D. C. (1981). *Applications, basics, and computing of exploratory data analysis*. Boston, MA: Duxbury Press.

- Wagenmakers, E.-J. (2007). A practical solution to the pervasive problem of p values. *Psychonomic Bulletin and Review*, 14, 779–804.
- Wellek, S. (2003). *Testing statistical hypotheses of equivalence*. Boca Raton, FL: Chapman & Hall/CRC.
- Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals. *American Psychologist*, 54, 594–604.
- Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics: Proceedings of the first international meeting held in Valencia (Spain)* (pp. 585–603). Valencia, Spain: University of Valencia.