Bayesian evaluation of inequality-constrained linear regression models and effect size estimation using absolute values

M. Schoenmakers

1 Abstract

Linear regression is an often-used method for assessing relative effects of independent variables on a dependent variable. Two topics are particularly interesting when conducting a linear regression analysis. The first topic is the magnitude of the standardized effect size. The second topic is establishing an order in standardized effect sizes. When we want to order effect sizes on importance rather than sign, a comparison of effect sizes using absolute values is necessary. Current software packages available are unable to perform this task. The present paper thus aims to introduce a new function written in R capable of comparing effect sizes using absolute values and estimating magnitudes of effect sizes. This function utilizes the Bayesian statistical framework. We use data from Carlson and Sinclair (2017) to illustrate the usage of the new function called hyp_test and evaluate the effectiveness of the function.

2 Introduction

Linear regression is a widely used method for assessing effects of a set of independent variables on a continuous dependent variable (Mulder & Ollson-Collentine, 2019). The effect of independent variables on a dependent variable is relevant in many cases, for example policy selection. Suppose a university wants to predict the future success of their students. Are developable factors, like amount of time spent studying indicative of future success? If it turns out amount of time studying is a strong predictor of success, the university may want to implement a policy that promotes student motivation for studying. If the amount of time studying has relatively little impact on future success, this policy would be a waste of time and money. The university may wish to quantify the effect size of amount of time studying on success before they decide whether they should implement the policy.

Linear regression can also be used to answer more scientific questions. Possible applications include studying the relationship between Big five personality traits and inmate recidivism (Clower & Bothwell, 2001), predicting perceptions of discrimination against women (Carlson & Sinclair, 2017) and many more.

When interpreting the results of a linear regression analysis, two questions are of particular interest. Firstly, we would like to estimate the magnitude of an effect. We can do this by verifying whether an effect size is above a certain value, or by checking if an effect size is within an interval. Secondly, we want to establish an order in effect sizes. Ranking effects on their importance is a great asset in theory construction (see Maher & McLachlan, 1995; Darmadi-Blackberry et al., 2004; Vittersø, 2001 for examples). Comparing effect sizes without using absolute values and testing magnitudes of effects can already be done with several software packages, for example lmhyp (Mulder & Olsson-Collentine, 2019), BIEMS (Mulder, Hoijtink, & de Leeuw, 2012) and BAIN (Gu, Mulder, & Hoijtink, 2018). These programs are however unable to compare effect sizes using absolute values.

Comparing effect sizes using absolute values is interesting for several reasons. Imagine a scenario where we have three standardized regression coefficients, where β_1 is 0.2, β_2 is -0.4 and β_3 is 0.6. In this scenario the hypothesis $\beta_3 > \beta_2 > \beta_1$ is false, even though this is the order of coefficients based on importance. Note that the hypothesis $\beta_3 > \beta_2 > \beta_1$ is an inequality constrained hypothesis (We hypothesize that the coefficients are unequal). If we are interested in establishing an order in effect sizes based on importance rather than direction of the effect, using absolute values of effects helps here, as the hypothesis $|\beta_3| > |\beta_2| > |\beta_1|$ would be true. For ordering effect sizes on importance rather than direction when negative predictors are present, we can use absolute values. Absolute values can also be used in an exploratory sense, for example when the magnitude of an effect is known, but we are unsure about the direction of the effect. Finally, we can also use absolute values to compare a negative regression coefficient to a positive regression coefficient based on importance.

For these reasons, the present paper aims to introduce the hyp_test function in R, a free statistical software package. This function is able to estimate magnitudes of effect sizes and comparing effects to each other using both non-absolute and absolute values in R. The following sections describe the theory behind linear regression in more detail, explain why the Bayesian statistical framework was chosen over the frequentist framework, describe the method of the hyp_test function and finally evaluate the effectiveness of the hyp_test function.

2.1 Theory of linear regression

A linear regression attempts to predict values for a dependent variable Y using an intercept α and independent variables X with certain coefficients β (Pallant, 2007).

Example 1

A linear regression model predicting future success based on the amount of time spent studying would look like this:

$$y = \alpha + \beta_1 X_1 + e \tag{1}$$

Where y is the value of success, β_1 is the regression coefficient for amount of time studying, X_1 is the amount of time spent studying and e is the error term. This model means that as time spent studying increases by one unit, our estimated value of success increases by β_1 units. As no prediction is perfect, the true value of success is the estimated value of success plus an error term e.

If we set α to 10, β_1 to 1, the error to 0 and we assume that a certain student spent 300 hours studying, our predicted value for success would be:

$$y = 10 + 1 \times 300 + 0 = 310 \tag{2}$$

We would conclude that this student has an estimated 310 "future success-units". The student would be more likely to be successful than a student with less units, and less likely to be successful than a student with more units. We can also create a model with more than one predictor of success.

Example 2

When we take the previous model and we add IQ as a predictor of future success, the model would be:

$$y = \alpha + \beta_1 X_1 + \beta_2 X_2 + e \tag{3}$$

Where y is the value of success, X_1 is the amount of time spent studying, X_2 is the amount of IQ points and e is the error term. There are two important things to note when interpreting a model with multiple predictors. Firstly, the β-term corresponding to a certain predictor variable is the unit increase in the dependent variable when the independent variable increases by 1 unit, *controlling for all other independent variables in the model.* That is, the β_2 is now the increase in estimated success when hours studying goes up by one when IQ is kept constant (Weisberg, 2005).
Secondly, β-terms (coefficients) cannot be directly compared to assess which of the two variables are more important. Coefficients can only be compared when both predictor variables are on the same

scale. To accomplish this, we may standardize the variables. Standardizing the variables puts all variables on a scale with mean 0 and standard deviation 1. After we do this, direct coefficient comparison is possible. For this reason, all coefficients described in this paper are either standardized or on the same scale. The following section describes the two main statistical frameworks which enable comparison of coefficients.

2.2 Statistical frameworks

In order to compare coefficients, two main statistical frameworks are possible. The first framework is the frequentist framework, which is predominant in psychological research. The second framework is the Bayesian framework. The differences between the frameworks are described in the section below. To evaluate the differences between the frameworks, we use the example of a coin flip.

2.2.1 Frequentist statistics

Frequentists conceptualize probability as a frequency. Frequentism can be divided into finite frequentism and hypothetical frequentism.

2.2.1.1 Finite frequentism

Finite frequentists state that probability is equal to proportion. This means that when we flip a coin, the chance of heads is equal to the number of heads divided by the total amount of coinflips (Hájek, 1996). At first glance, describing probability as a set of observed outcomes may seem like a good idea, as it roots probability theory firmly in empiricism and observable outcomes. The definition unfortunately also creates some serious problems. Under this definition of probability, a coin that is never tossed does not possess a chance of landing heads or tails. Furthermore, a coin that is tossed an uneven number of times can never have an equal chance of landing heads or tails (number of flips/uneven number $\neq 0.5$). Perhaps even worse, a finite frequentist would state that the chance of heads when a coin is thrown once can only be one or zero, as we must divide a frequency of heads (1 or 0) by 1. This is obviously not a very informative answer. Imagine going to the dentist, who informs you that you have a 0 or 1 probability of needing a root canal treatment. Additionally, the very nature of probability means we may observe very unlikely outcomes indeed. To equate frequency to probability may therefore not be the best of ideas.

Another problem of frequentism is the subset problem. As stated earlier, probability is equal to proportion. The question is which proportion we should use for objects that potentially belong to multiple classes (the subset problem). When I want to know the chance that it will rain tomorrow, do I take the frequency of rainy days last year and divide it by the amount of days in that year? Do I take the amount of times it has rained in the current season divided by the amount of days in this season? Or do I take the amount of times it has rained on 24 April divided by the number of times 24 April has occurred? All three of these options will yield different probabilities with no clear indication which is correct (Hájek, 1996).

2.2.1.2 Hypothetical frequentism

Hypothetical frequentism tries to “fix” the definition of frequentism through an adjustment of the conceptualization of probability. When we flip a coin an infinite amount of times, it will show heads 50% of the time, and it will show tails 50% of the time. The hypothetical frequentist chance of a (balanced) coin landing heads is therefore 50%. Hypothetical frequentism requires us to visualize an infinite number of trails for every probability. In doing so, we immediately leave the concept of empiricism behind, which was the main benefit of frequentism to begin with. Hypothetical frequentism also does not solve the subset problem (Hájek, 2012). Philosophically, frequentist statistics do not seem like a good choice as a statistical framework. The following section describes the frequentist procedure of statistical testing and its corresponding problems.

2.2.1.3 Frequentist statistical testing

The traditional, frequentist way of comparing regression coefficients is by doing a null-hypothesis test, where H$_0$: $\beta_1 = \beta_2$ and H$_1$: $\beta_1 \neq \beta_2$. To test this hypothesis, a test statistic is calculated. We reject the null hypothesis when the p-value, the chance of obtaining an equally extreme or more extreme value for the test statistic given that the H$_0$ is true, is less than 0.05. If the H$_0$ is rejected, we conclude that the predictors are significantly different (Paternoster, Brame, Mazerolle & Piquero, 1998).

The usage of p-values is a problematic aspect of frequentist statistics. P-values are dependent on unobserved data, influenced by subjective intentions of the researcher and not a quantification of statistical evidence (Wagenmakers, 2007). The usage of a dichotomous decision with a cut-off value of 0.05 for the rejection of the H$_0$ is also arbitrary. If our p-value is 0.050001, our null hypothesis is not rejected, but if the p-value is 0.049999 we do reject the null hypothesis. A difference of 0.000002 in our p-value could completely reverse our decision. Interpretations of p-values are also problematic, due to the combination of Fisher's and Neyman-Pearson's statistical testing procedures in null-hypothesis testing. In particular, p-values often get confused with type I error rates, leading to incorrect statements such as a p-value of 0.01 being "highly significant" while a p-value of 0.05 is just "significant" (Hubbard & Armstrong, 2005). A final, important problem with frequentist statistics is the inability to obtain the chance of H$_1$ being correct. Frequentist statistics can reject or accept a hypothesis, but it is not possible to compute the chance that H$_1$ is correct. We only get a p-value which has a very problematic interpretation, as stated above. Because of the problems with frequentist statistics outlined above, we turn to a different form of statistics; Bayesian statistics.

2.2.2 Bayesian statistics

Bayesians conceptualize probabilities as a degree of uncertainty regarding the outcome of an event (Hájek, 2012). Before we flip a coin, we are uncertain about the outcome of the coin flip (heads or tails?). We can capture this prior uncertainty in a distribution. If we believe the coin to be approximately balanced, our prior distribution could look like the graph below, where theta > 0.5 signals a belief that the coin is biased towards heads, and theta < 0.5 signals a belief that the coin is biased towards tails. The area under the curve (AUC) is equal to 1. This is a property of a probability distribution.

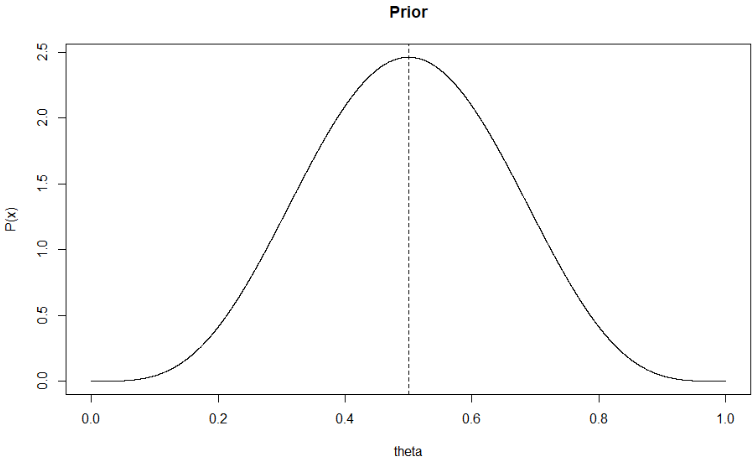

Figure 1. Prior distribution when we flip a coin we believe to be approximately balanced.

To check whether our prior beliefs are accurate, we can use the likelihood. The likelihood depends on the observed data, as opposed to our prior beliefs. Say we flip the coin 10 times, and we observe 7 heads and 3 tails. The likelihood distribution would look like this.

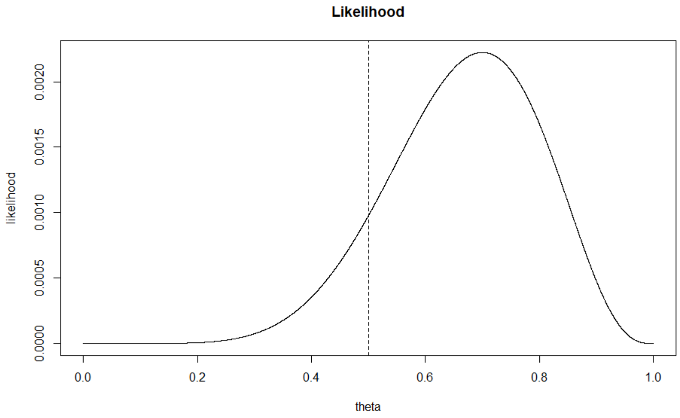

Figure 2. Likelihood distribution after flipping the coin 10 times.

Based purely on the data we observed, we are somewhat inclined to believe the coin has a bias towards heads. Combining the data with our prior beliefs yields the posterior distribution.

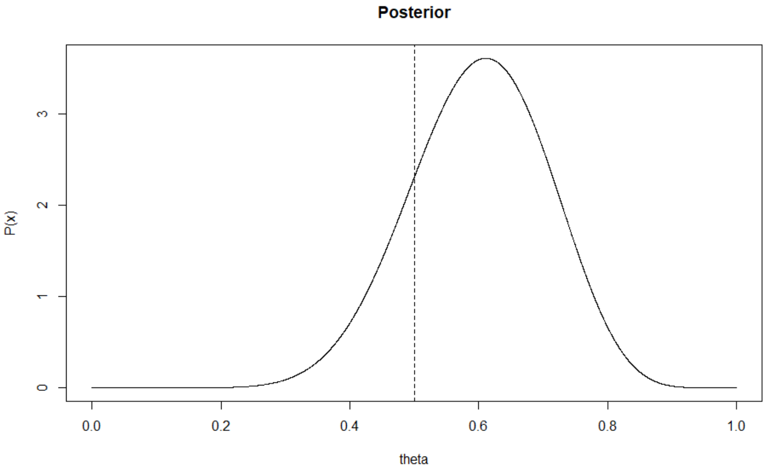

Figure 3. Posterior distribution combining the prior and likelihood distributions.

The posterior distribution is the combination of our prior distribution and our likelihood distribution. After we observed the coinflips resulting in heads more often than tails, we adjusted our prior beliefs about the fairness of the coin using the likelihood. The posterior probability is thus a combination of prior beliefs and the likelihood function. The prior, likelihood and posterior are the key components of Bayesian statistics (Bolstad, 2004).

2.2.2.1 Bayesian hypothesis evaluation

The Bayesian posterior probability of a hypothesis being true is the AUC that satisfies the condition imposed by the hypothesis. For the coinflip example, the posterior probability of hypothesis theta < 0.5 being true (tails being more likely than heads) is equal to the AUC to the left of the dotted line. The posterior probability of theta > 0.5 being true is the AUC to the right of the dotted line. It is also possible to use this method of hypothesis evaluation to compare regression coefficients. To compare whether $\beta_1 < \beta_2$, we can calculate the AUC where the $\beta_1 < \beta_2$ condition is true. This is the first part of the functionality of the hyp_test function. The hyp_test function also calculates the inverse of this AUC. This is the probability of the complementary hypothesis, or HC for short.

2.2.2.2 Bayes factors

Bayes factors are a measure of support for a given hypothesis over another hypothesis. The posterior distribution states there is a more than 50% chance of the coin being biased towards heads. To evaluate the amount of support the hypothesis theta > 0.5 (H_1) receives, we can compare it to an unconstrained hypothesis H_u, where theta is between 0 and 1. For evaluation of inequality hypotheses, we can do this using the formula:

$$BF_{1u} = \frac{fit_1}{complexity_1} = \frac{Posterior\ probability\ of\ theta > 0.5}{Prior\ probability\ of\ theta > 0.5} \tag{4}$$

This formula returns a number with a minimum of 0 and a maximum approaching infinity. A number larger than 1 supports the hypothesis that the coin is biased towards heads. A number equal to 1 favors neither hypothesis. A number smaller than 1 favors the unconstrained hypothesis. As the Bayes factor divides the posterior probability by the prior probability, it inherently rewards hypotheses with a lower complexity. That is, a hypothesis that has a lower prior probability of being true will receive a higher Bayes factor if the posterior probability remains equal.

We can also use the Bayes factor to directly compare hypotheses. To do this, we take the Bayes factor of hypothesis 1 and divide it by the Bayes factor of hypothesis 2. The result is a Bayes factor expressing the support hypothesis 1 receives compared to hypothesis 2.

$$BF_{12} = \frac{BF_{1u}}{BF_{2u}} \tag{5}$$

The function presented in this paper calculates a posterior probability, a complementary hypothesis HC, and two Bayes factors; one comparing the posterior to the prior probability of the hypothesis being true and one comparing the hypothesis to the complementary hypothesis (Hoijtink, 2012). The method is illustrated in the following section.

3 Bayesian hypothesis evaluation in R

To illustrate the use of Bayesian hypothesis evaluation in R, we will use an actual dataset (Carlson and Sinclair, 2017). The dataset attempts to explain perceptions of gender discrimination in hiring for the roles of computer specialist and nurse. The dependent variable was perceptions of discrimination towards women (discW) and the independent variables were belief in discrimination against women (beliefW) , stigma consciousness (stigma) and feminist identification (feminist). Gender (gender) and belief in discrimination against men (beliefM) were added as control variables. The regression equation for the estimated value of discW is:

$$discW' = \alpha + \beta_1 \times beliefW + \beta_2 \times stigma + \beta_3 \times feminist + \beta_4 \times beliefM + \beta_5 \times gender \tag{6}$$

We import the SPSS dataset into R using the read_sav function. After we import the data, we standardize the variables to enable direct comparison of regression coefficients. We estimate the regression coefficients using the lm function in R. The lm function returns a list containing, among other things, the estimated regression coefficients and a variance-covariance matrix.

Table 1. Estimated regression coefficients returned by the lm function.

Intercept	beliefW	stigma	feminist	beliefM	gender
-0.01	0.38	0.05	0.13	-0.09	-0.15

The numbers presented in the table above are point-estimates of the β parameters. It would be very convenient if we were 100% sure that these values are the actual β parameters in the population (this thesis and the hyp_test function would not be needed), but this is unfortunately not the case (Weisberg, 2005). The point-estimates of the coefficients are just that; estimates that we are not certain about. In the Bayesian framework we can quantify the uncertainty about the actual value of the parameters using the posterior distribution which has a multivariate Student t distribution. In R, we need to specify a number of draws per coefficient (5000), a non-centrality parameter delta (the point-estimate of the regression coefficients), a sigma (variance-covariance matrix of the coefficients obtained from the lm function) and finally the degrees of freedom (N minus amount of coefficients). These 5000 draws get saved in the matrix called post_draws. We also take 5000 draws from the same multivariate t-distribution, but this time with a vector of zeros as our centrality parameter, a covariance matrix multiplied by n-k as sigma and 1 degree of freedom. These draws get saved as prior_draws and are used in the computation of the prior. See Appendix 3 for further details.

3.1 Plots and method

We save the 5000 draws per coefficient in a matrix (the data type, not the movie) called post_draws. This matrix has 6 columns (intercept + the 5 coefficients) and 5000 rows. The matrix thus contains 6 × 5000 numbers. When we plot the second and third column of post_draws (corresponding to coefficient 1 and 2) using the plot and density functions (e.g. `plot(density(post_draws[ , 2]))`, we get the following plots of β_1 and β_2:

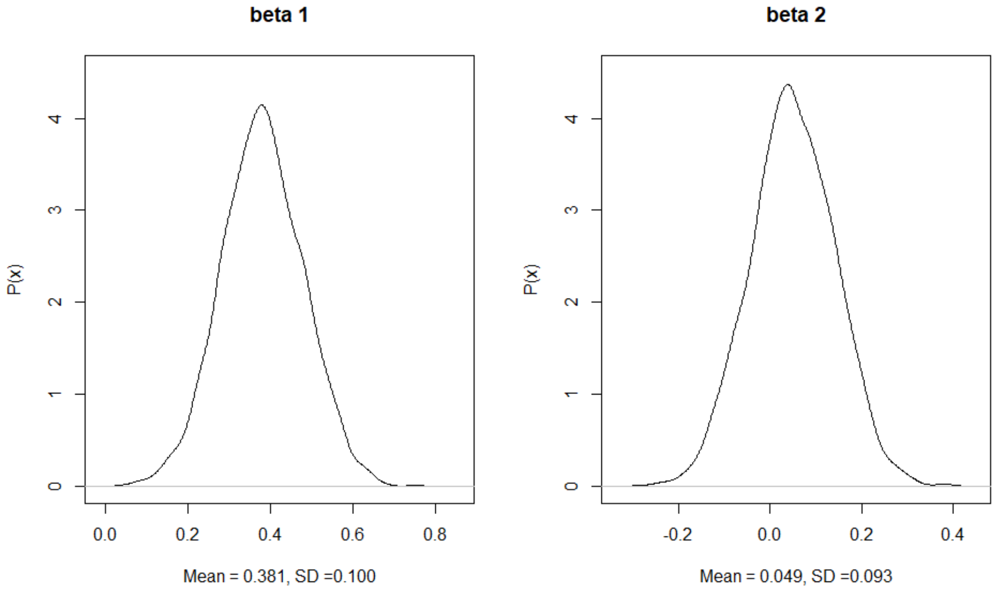

Figure 4. Posterior probability distributions for the first and second regression coefficient.

Note that the AUC is 1 for both curves and that the curve is centered around the point-estimate of the coefficient. If our hypothesis is $\beta_1 > \beta_2$, we can rewrite this hypothesis as $\beta_1 - \beta_2 > 0$. Plotting $\beta_1 - \beta_2$ yields:

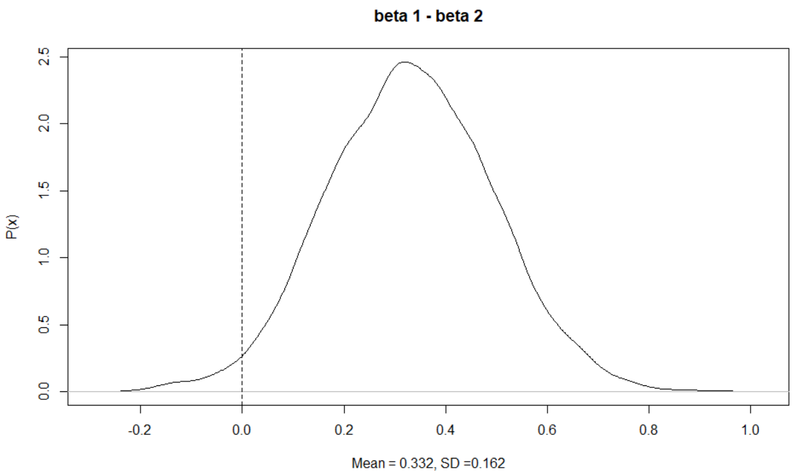

Figure 5. Posterior probability distribution of the first regression coefficient minus the second regression coefficient.

The AUC to the right of 0 is the probability of $\beta_1 > \beta_2$ being true. The AUC to the left of 0 is the probability of the complementary hypothesis, in this case $\beta_1 \leq \beta_2$, being true. The hyp_test function presented in this thesis calculates the AUC's corresponding to certain hypotheses. It requires two inputs: the list generated by the lm function and a certain hypothesis formulated as string (e.g. "$beliefW > stigma$"). This formulation means that beliefW is a more important positive predictor than stigma, which corresponds to $\beta_1 > \beta_2$. There are many possible hypotheses which are accepted by the hyp_test function, listed in Appendix 1.
There are also several optional inputs. We can specify the amount of times we want to draw an estimation of each regression coefficient as reps (default = 5000). We can also input a seed for reproducibility as seed (default = 440738). Finally, we can set return (default = 1) to a value other than 1 to get output useable by the "Simulation" function (for testing purposes, discussed in Appendix 2).

As output, the function returns the AUC of the hypothesis being true as variable PosteriorProbability and the AUC of a hypothesis not being true as variable HC (hypothesis complementary). It also computes the Bayes factor of the specified hypothesis being true compared to an unrestricted model (a model with no constraints) and the Bayes factor of the specified hypothesis being true compared to the complementary hypothesis. The following section explains the inner workings of the hyp_test function in more detail.

3.2 Explanation of the function

The hyp_test function has matrix multiplication using inequality matrices at its core. Further details about matrix multiplication, inequality matrices and the exact computation of the Bayes factors can be found in Appendix 3. The focus of this section will mainly be the comparison of coefficients via absolute values and testing whether an effect is within an interval. Further information about comparing coefficients via non-absolute values can be found in Appendix 7.

3.2.1 Comparing two coefficients via non-absolute values

To highlight the difference between comparing coefficients using non-absolute and absolute values, we will first consider the hypothesis $\beta_2 > \beta_5$, or in words: "$stigma > gender$". Afterwards, we will compare the results of this hypothesis to the results of the hypothesis $|\beta_2| > |\beta_5|$. To evaluate the hypothesis $\beta_2 > \beta_5$, the hyp_test function first converts the hypothesis into an inequality matrix (Appendix 3). After composing the inequality matrix, the function multiplies this inequality matrix with the post_draws matrix mentioned earlier. This yields a matrix containing the values of $\beta_2 - \beta_5$. To obtain the posterior probability of $\beta_2 > \beta_5$, we check how many times out of 5000 $\beta_2 - \beta_5 > 0$.

We obtain a posterior probability of 0.9618. This corresponds to the area of the graph below that is to the right of the dotted line. The hyp_test function also returns the probability of the complementary hypothesis being true (1-0.9618). This probability corresponds to the area in the graph to the left of the dotted line.

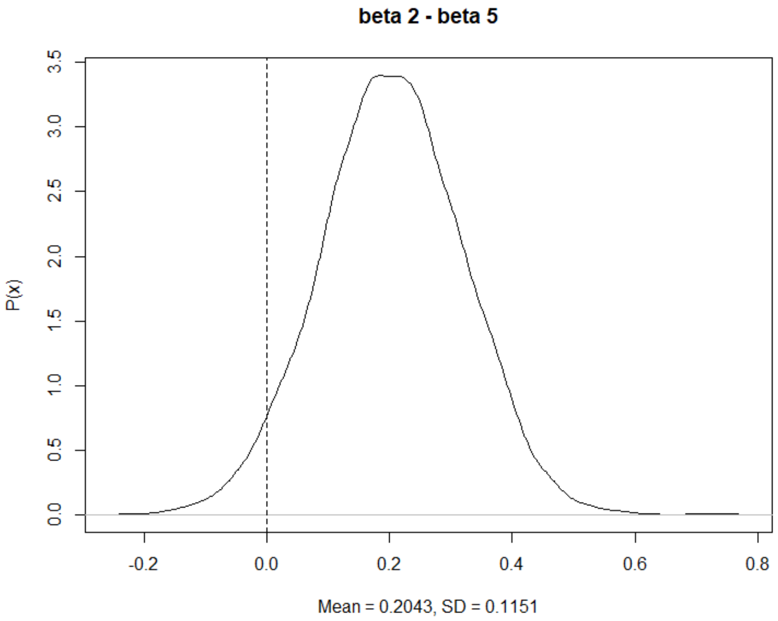

Figure 6. Posterior probability distribution of the second regression coefficient minus the fifth regression coefficient with a dotted line at 0.

The function also returns two Bayes factors. The first Bayes factor, BF_Unconstrained, expresses how much more likely the hypothesis is after observing the data. This Bayes factor is 1.92, meaning that the hypothesis became more likely to be true after observing the data. The second Bayes factor, BF_Complementary, expresses how much more likely the specified hypothesis is compared to the complementary hypothesis. This Bayes factor is 25.18, meaning that

$$BF_{Unconstrained,\ Complementary} = \frac{BF_{Unconstrained}}{BF_{Complementary}} = 25.18 \text{ (see formula 6).}$$

In words, the hypothesis $\beta_2 > \beta_5$ received about 25 times more support than the complementary hypothesis $\beta_2 \leq \beta_5$. The hypothesis $\beta_2 > \beta_5$ thus receives a decent amount of support. This should however not be interpreted as stating that β_2 is a more important predictor than β_5. It merely means that β_2 is likely a more positive predictor than β_5. To see which predictor is more important a comparison of coefficients using absolute values is required, which the following section explores.

3.2.2 Comparing two coefficients via absolute values

We can instruct the hyp_test function to compare our coefficients using absolute values by adding the term “absolute” to our hypothesis. The hyp_test function then takes the absolute values of the post_draws matrix before multiplying the inequality matrix with the post_draws matrix. After the matrix multiplication, this results in a matrix containing $|\beta_2| - |\beta_5|$ instead of $\beta_2 - \beta_5$ (Absolute values are denoted by ||, with the || containing the number or the variable). This results in the following graph:

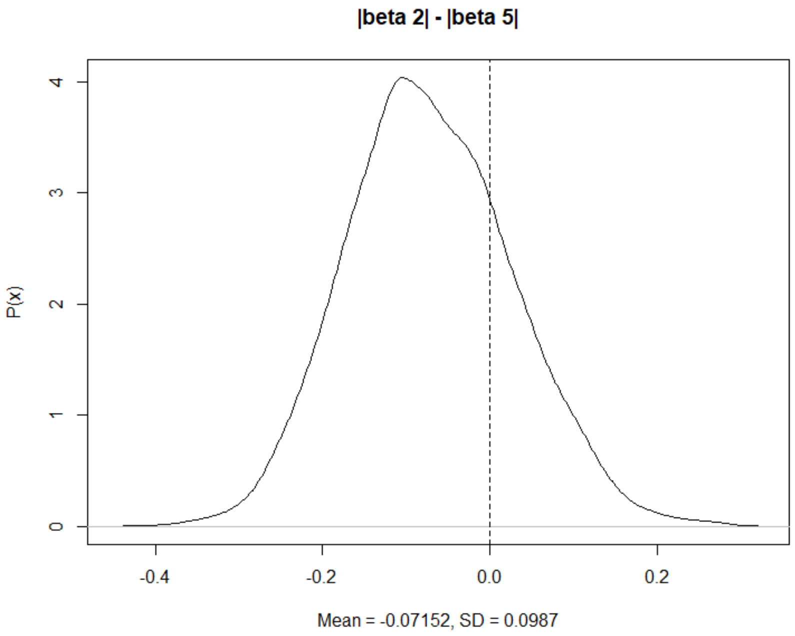

Figure 7. Posterior probability distribution of the absolute value of the second regression coefficient minus the absolute value of the fifth regression coefficient with a dotted line at 0.

The posterior probability of $|\beta_2| > |\beta_5|$ is 0.2324, as $|\beta_2|$ is often smaller than $|\beta_5|$. The HC is 0.7676 (1-0.2324). The BF_Unconstrained is 0.4033 and the BF_Complementary is 0.2227. The hypothesis that $|\beta_2| > |\beta_5|$ thus receives little support. This means that we have a low probability of β_2 being a more important predicator regardless of sign than β_5. If we only test the hypothesis $\beta_2 > \beta_5$ without looking at the sign difference (β_2 has a positive point-estimate and β_5 has a negative point-estimate) we may be misled into thinking β_2 is likely the more important predictor. Using absolute values in hypothesis testing can avoid this problem.

3.2.3 Comparing three or more coefficients

The example above was for a hypothesis where we compare two coefficients, β_2& β_5. We can also compare three or more coefficients. Consider the following hypothesis: $\beta_1 > \beta_2 < \beta_3 > \beta_4 > \beta_5$ or in words; beliefW > stigma < feminist > beliefM > gender. Note that this is the hypothesis our point-estimates of the coefficients point towards, so we expect a reasonably high posterior probability. See Appendix 4 for details on composing an inequality matrix for 2+ coefficients.

Multiplying the inequality matrix with the post_draws matrix yields a posterior probability of 0.513. The complementary probability is 0.487. At first glance, it seems the complementary hypothesis receives about as much support as the specified hypothesis. When we look at the Bayes factors, we see that the BF_Unconstrained is 7.48 and the BF_Complementary is 14.3. The specified hypothesis receives about 14 times more support than the complementary hypothesis. This is caused by the fact that the Bayes factor balances fit and model complexity. The specified hypothesis has less complexity (i.e. is more specific) than the complementary hypothesis, as the complementary hypothesis contains all hypotheses that are not the specified hypothesis.

3.2.4 Comparing a coefficient to two fixed values

To estimate whether a coefficient is within a certain interval, we can compare it to two fixed numbers. An example is $0.2 < \beta_1 < 0.3$, which would indicate β_1 is between 0.2 and 0.3. This hypothesis is first split into two parts; $0.2 < \beta_1$ and $\beta_1 < 0.3$. See Appendix 5 for details on the inequality matrix. The function returns a posterior probability of 0.1758, a HC of 0.8242, a BF_Unconstrained of 7.991 and a BF_Complementary of 9.482.

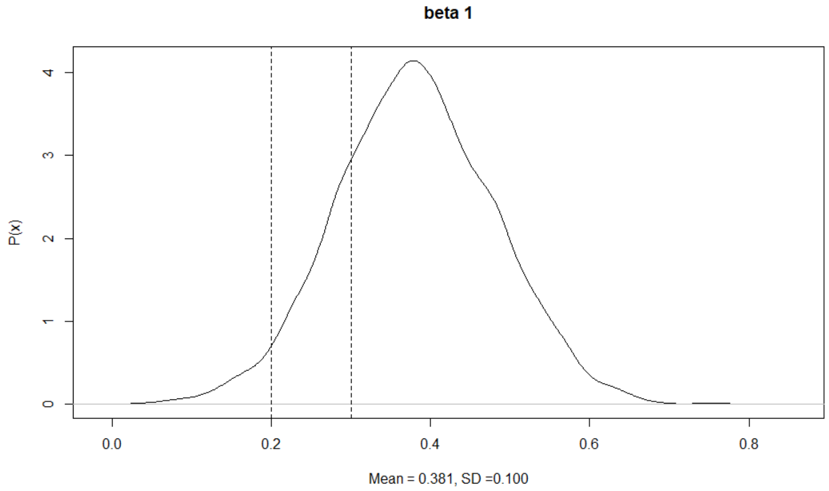

Figure 8. Posterior probability distribution of the first regression coefficient with dotted lines at 0.2 and 0.3.

In graph form: the area between the two dotted lines in the graph is 0.1758 and the area outside the two lines in the graph is 0.8242.

4 Simulation study

It is always a good idea to check if a method works not just in theory, but also in practice and how well it works given different circumstances. The following section is dedicated to this.

To check whether the function works in practice, a simulation study was conducted. Assume we have a population where we estimate a linear regression to predict the estimated value of Y using 4 coefficients. The linear regression equation would look like this:

$$Y' = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \quad (8)$$

Where Y' is the estimated value of Y. Using R, we can generate data that correspond to certain values of regression coefficients in the population (Appendix 2 for details).

4.1 Factors influencing the function outcomes

There are four main factors affecting the function outcomes. The first factor that influences the precision of the outcomes is sample size. Larger samples have a smaller standard error, which means we can be more confident the estimated values for our regression coefficients are close to the regression coefficients in the population. The second factor that influences the outcomes is the effect size. Effect size can in this case be seen as the difference between the standardized regression coefficients. If the coefficients are very close to each other, we will only be able to reliably distinguish between them when the sample size is very large. The third factor that influences the outcomes is the number of variables we compare to each other. A comparison of two variables will always have a higher posterior probability than a comparison of four or six variables, given that all else remains equal. This is caused by almost all coefficients sharing some degree of overlap. The fourth factor influencing the outcomes is the standard deviation of the residuals. Residuals are the prediction errors in the regression model. Large residuals indicate a poor linear regression model fit. The different factors and their impact on the posterior probability will be indicated below.

The inputs for the simulation-function are the sample size, the values of the coefficients in the population, the amount of times we repeat the simulation and the hypothesis we are testing. Say we input the following values for the coefficients:

Table 2. True regression coefficients in the population.

β_1	β_2	β_3	β_4
0.2	-0.4	0.6	-0.8

We also assume we have a sample of 20. We want to test the hypothesis that is true in the population;

$$|\beta_4| > |\beta_3| > |\beta_2| > |\beta_1|$$

When we randomly generate data given these circumstances, our estimates of the coefficients will not be perfectly accurate. They may even be quite bad, given the low sample size. When we let the lm function estimate the coefficients, we get the following results:

Table 3. Regression coefficients estimated by the lm function based on simulated data.

β_1	β_2	β_3	β_4
-0.35	-0.388	1.05	-1.19

When we compare the estimated coefficients to the coefficients in the population, we see that they are quite different. When we run our hyp_test function, we get a value for the posterior probability of 0.351 for the hypothesis that is true in the "population": $|\beta_4| > |\beta_3| > |\beta_2| > |\beta_1|$. To check the consistency of this result, we run the simulation 1000 times and plot the results of posterior probability in a graph. Four graphs of this kind, for N = 20, N = 50, N = 100 and N = 200 are displayed below.

4.1.1 Varying sample size

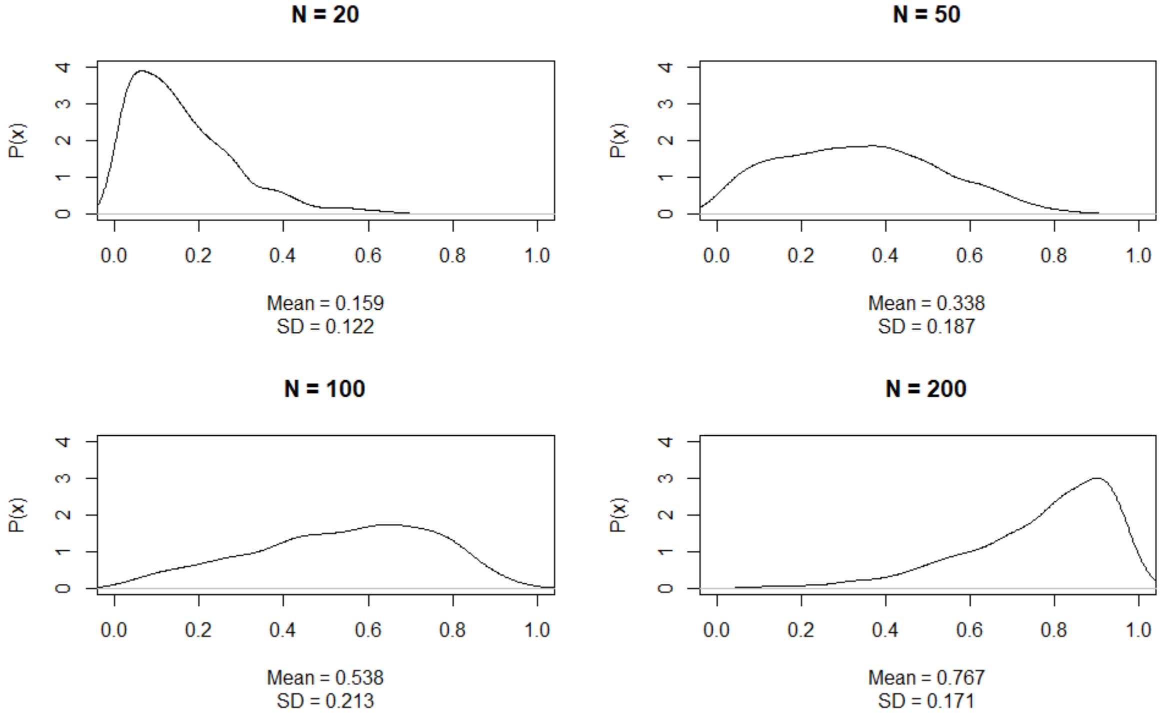

Figure 9. Posterior probability distributions for the hypothesis |x4| > |x3| > |x2| > |x1| for differing sample sizes.

Looking at the graph, we can see that our conclusions are not very reliable when we take a sample of only 20 people. The N = 50 graph already looks a lot better than the N = 20 graph. We draw the right conclusion more often than when we only had a sample of 20, but there is still a lot of room for improvement. Increasing the sample size to 100 yields a better result, where the average PosteriorProbability is 0.538. Finally, increasing the sample to 200 results in a distribution with mean of 0.767. As can be seen in the graphs, larger sample sizes lead to higher posterior probabilities for hypotheses that are true in the population. A sample size of at least 100 when expecting effect sizes of 0.2 seems advisable.

A plot of the Bayes factors also provides some valuable insight in the results of the hyp_test function. Bayes factors for varying sample sizes are presented in the graphs below.

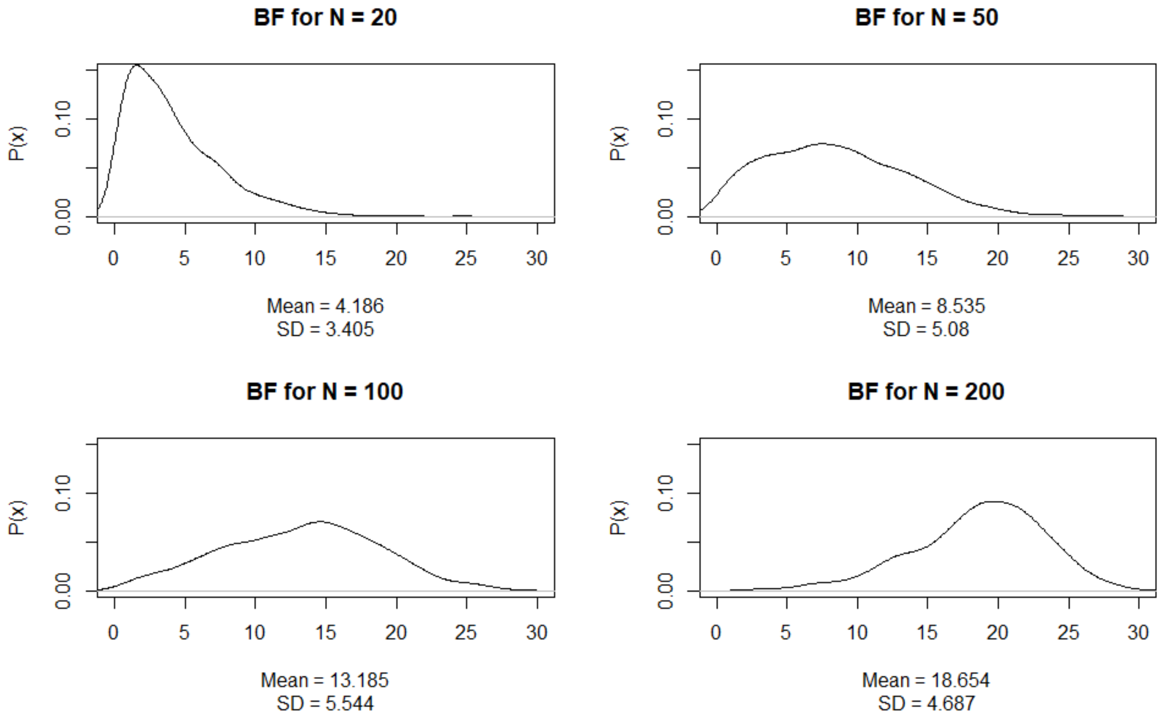

Figure 10. Probability distributions for Bayes factors considering the hypothesis |x4| > |x3| > |x2| > |x1| with varying sample sizes.

The shapes of the Bayes factor distributions somewhat mirror the shape of the PosteriorProbability distributions. To quantify how often the hyp_test function gives a correct outcome, we calculated the proportion of times the Bayes factor was higher than 1. This yields the following results:

Table 4. Proportion of Bayes factors considering the hypothesis |x4| > |x3| > |x2| > |x1| above 1 for varying sample sizes.

BF > 1 for N = 20	BF > 1 for N = 50	BF > 1 for N = 100	BF > 1 for N = 200
0.855	0.960	0.995	1

Looking at the table, we can see that the hyp_test function points us in the right direction for most cases. Even when the sample is as small as 20, the Bayes factor is larger than one 85% of the time. This number only increases for larger sample sizes, to the point where it is larger than one 100% of the time for a sample of 200. The next factor which affects the outcomes is effect size.

4.1.2 Varying effect size

When we fix the N to 50 and we vary the effect size, we get the following plots.

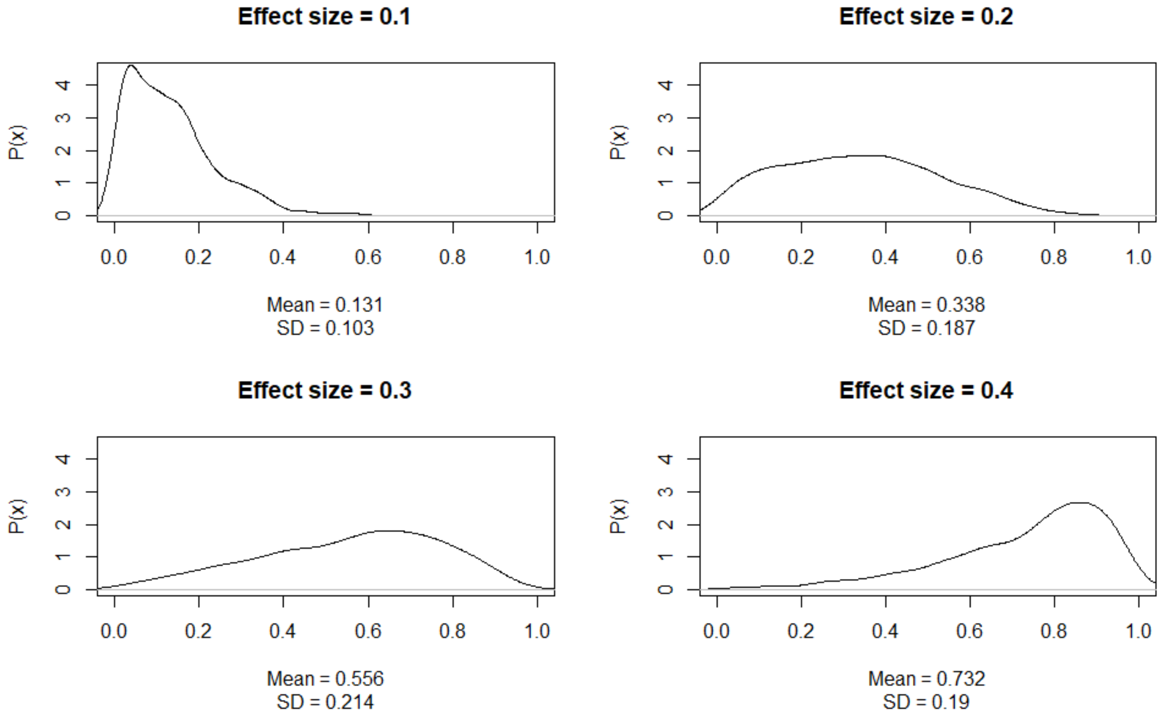

Figure 11. Posterior probability distributions considering the hypothesis |x4| > |x3| > |x2| > |x1| for given effect sizes.

As can be seen in the plots above, larger effect sizes lead to better posterior probability estimations. Expected effect size should be carefully considered when sample size is determined!

Plots for the Bayes factors given varying effect sizes are displayed below.

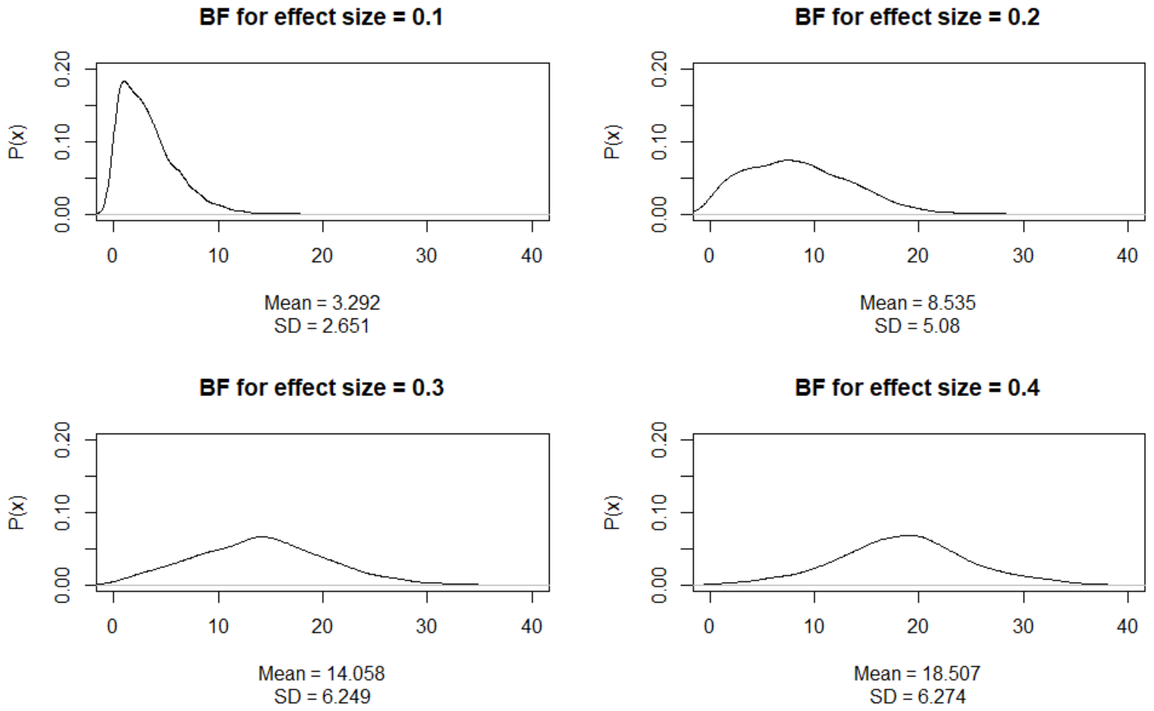

Figure 12. Probability distributions of Bayes factors considering the hypothesis |x4| > |x3| > |x2| > |x1| for given effect sizes.

Just like the posterior probability, the Bayes factor increases as effect size goes up. To again quantify the number of times the hyp_test function points us in the right direction, the proportion of times the Bayes factor was more than 1 was calculated and displayed in the table below.

Table 5. Proportion of Bayes factors considering the hypothesis x4 > x3 > x2 > x1 above 1 for different effect sizes.

BF > 1 for effect size = 0.1	BF > 1 for effect size = 0.2	BF > 1 for effect size = 0.3	BF > 1 for effect size = 0.4
0.786	0.960	0.992	0.999

As we can see in this table, the Bayes factor points us in the right direction most of the time. For very small effect sizes, like 0.1, a large sample can be used to offset the small effect sizes. The next factor to consider is the number of variables that are being compared.

4.1.3 Varying number of compared variables

The number of variables compared is also a factor of consideration. When we test a hypothesis that is true in the population, keep the N constant to 50 and fix the difference between variables to 0.2, we get the following graphs for comparing two, three, four and five variables respectively. PosteriorProbability keeps decreasing when we add a higher number of variables to compare.

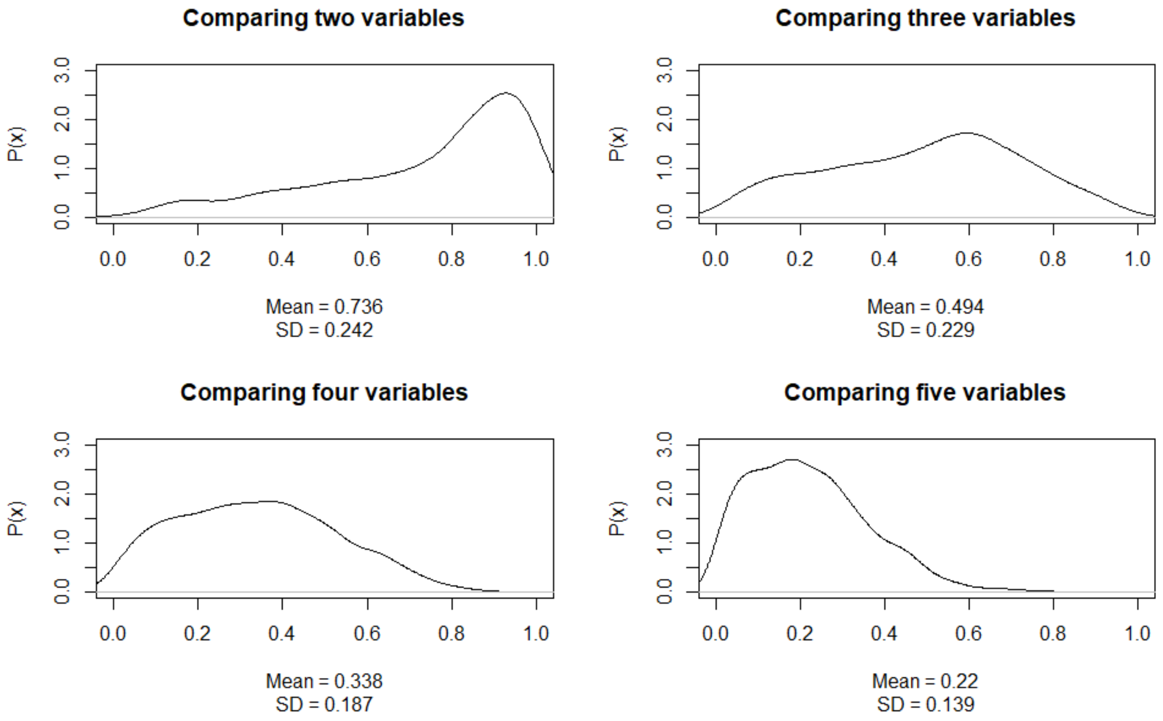

Figure 13. Posterior probability distributions for comparing different amounts of variables using absolute values.

As the PosteriorProbability keeps increasing, the Bayes factor increases along with it. Plots for the Bayes factor and a table for the proportion of times the Bayes factor was above 1 are displayed below.

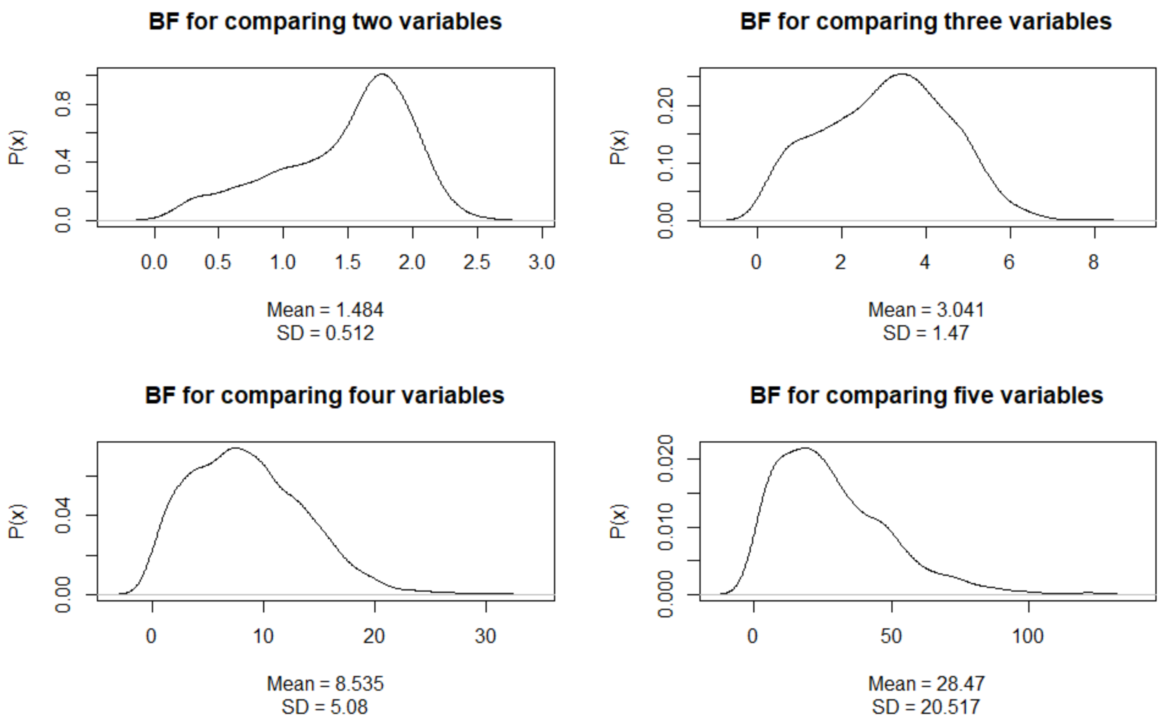

Figure 14. Probability distributions of Bayes factors for different numbers of compared variables using absolute values.

Table 6. Proportion of Bayes factors above 1 for different numbers of compared variables.

BF > 1 for two variables	BF > 1 for three variables	BF > 1 for four variables	BF > 1 for five variables
0.810	0.889	0.960	0.992

Interestingly, the Bayes factor increases when we add more variables to compare, while the posterior probability decreases as we add more variables. The fact that the Bayes factor keeps increasing is due to the decreased complexity of the hypothesis when more variables are added. The decreasing complexity of our hypotheses more than offsets the decrease in PosteriorProbability. A hypothesis with a very low prior probability of being true will receive a high BF even when the posterior probability of the hypothesis being true is very low. This makes it important to interpret Bayes factors in conjunction with posterior probabilities, and not separately. The fact that Bayes factors can be high even for low posterior probabilities raises the question whether the Bayes factor for a (partly) incorrect hypothesis also keeps increasing when complexity of the hypothesis decreases. This depends on several factors, among others the degree of incorrectness of the tested hypothesis, but is generally not the case (discussed in more detail in Appendix 6). Appendix 6 also contains further details about comparing overly complex hypotheses.

4.1.4 Varying sigma

Finally, the standard deviation of the residuals (sigma) should be considered. For N=50, effect size = 0.2, number of variables = 4, PosteriorProbability results for differing sigmas are presented below.

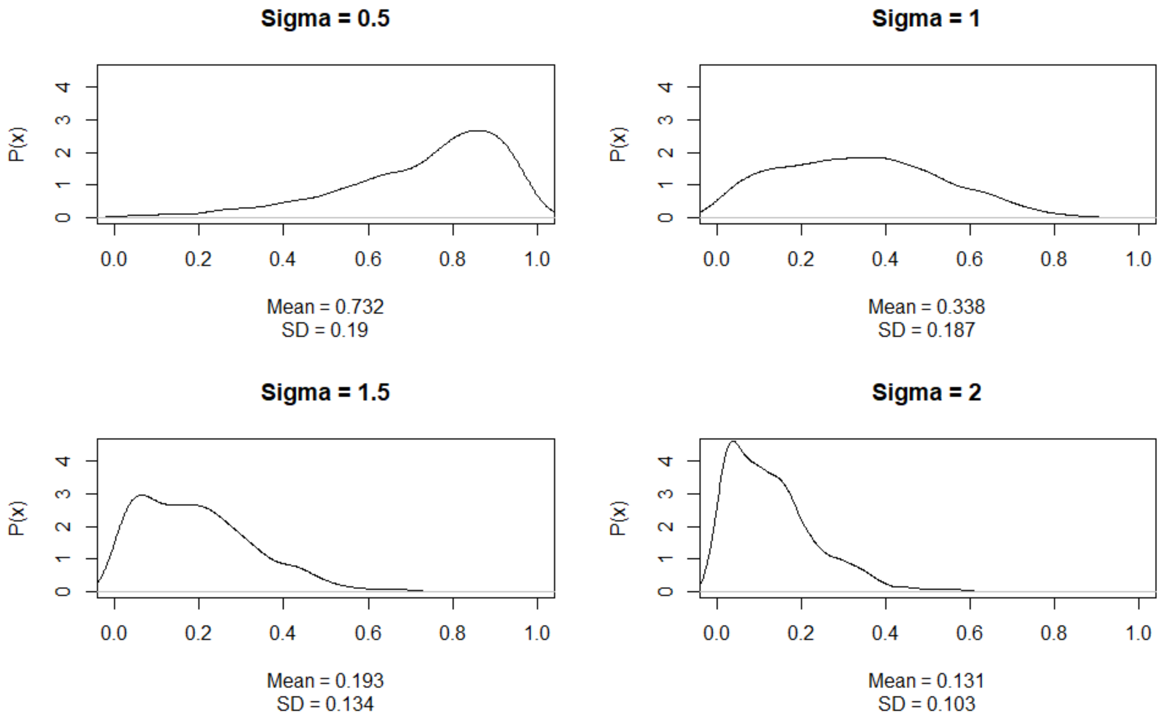

Figure 15. Posterior probability distributions considering the hypothesis |x4| > |x3| > |x2| > |x1| for differing values of sigma.

Plots for the Bayes factors and the table containing the proportion of Bayes factors > 1 are displayed below.

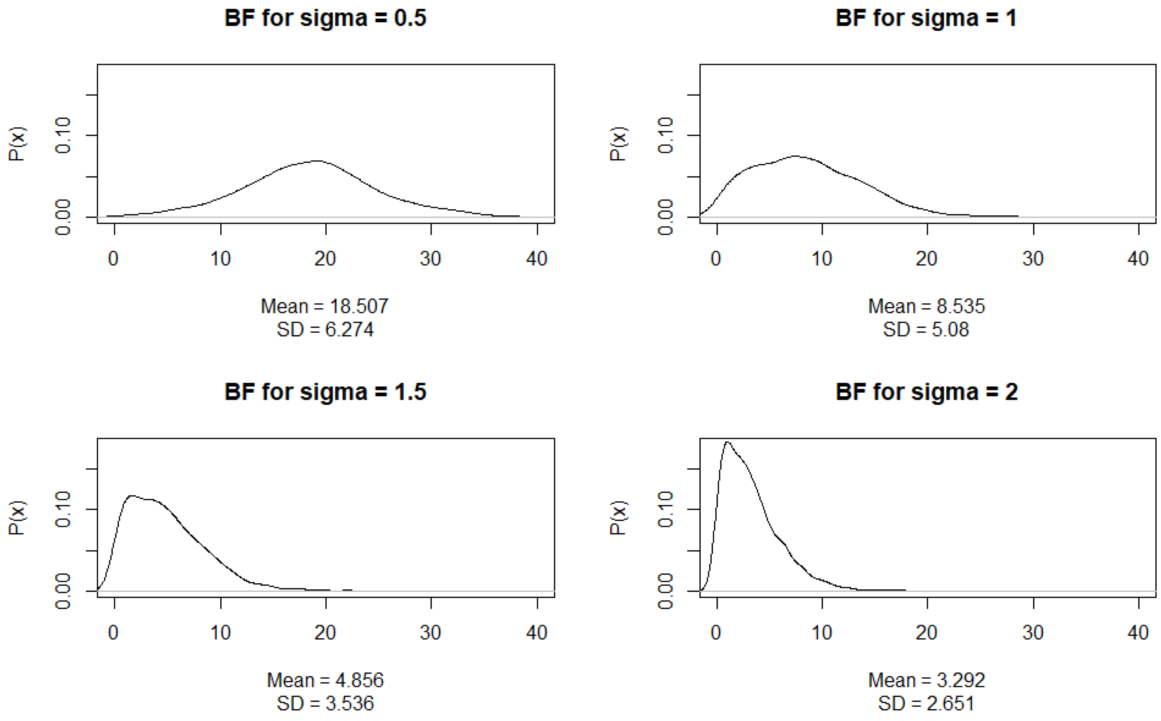

Figure 16. Probability distributions of Bayes factors considering the hypothesis |x4| > |x3| > |x2| > |x1| for differing values of sigma.

Table 7. Proportion of Bayes factors considering the hypothesis x4 > x3 > x2 > x1 above 1 for differing values of sigma.

BF > 1 for sigma = 0.5	BF > 1 for sigma = 1	BF > 1 for sigma = 1.5	BF > 1 for sigma = 2
0.999	0.960	0.876	0.786

The Bayes factor decreases along with the posterior probability for increasing values of sigma. All the comparisons we did here can also be found for non-absolute values in Appendix 8, although the results are very similar.

The answer to the main question of this section (Does the method work?) is a conditional yes. Given that the sample size is adequate for the given effect size, number of variables compared and the amount of variance in the residuals, the method works, or rather, is very likely to work. If the sample size is not adequate for the comparison attempted, the method is very likely to fail.

5 Empirical analysis

Using the dataset mentioned in the Bayesian hypothesis evaluation in R section (Carlson & Sinclair, 2017), we can do an analysis of actual data and interpret the outcomes. See Bayesian hypothesis evaluation in R for a reminder of the properties of this dataset.

5.1 Identifying the most important predictor

Suppose we want to find out which of the variables is the most important predictor of perceptions of discrimination against women, regardless of sign. The first hypothesis could be:

$$H_1: |\beta_1| > |\beta_2|, |\beta_3|, |\beta_4|, |\beta_5|$$

In words:

$$H_1: |beliefW| > |stigma|, |feminist|, |beliefM|, |gender|$$

This hypothesis reflects the possibility that beliefW is the most important predictor of perceptions of discrimination towards female victims, regardless of sign. Alternate hypotheses are:

$$H_2: |\beta_2| > |\beta_1|, |\beta_3|, |\beta_4|, |\beta_5|$$
$$H_3: |\beta_3| > |\beta_1|, |\beta_2|, |\beta_4|, |\beta_5|$$
$$H_4: |\beta_4| > |\beta_1|, |\beta_2|, |\beta_3|, |\beta_5|$$
$$H_5: |\beta_5| > |\beta_1|, |\beta_2|, |\beta_3|, |\beta_4|$$

Each hypothesis reflects the possibility of a certain predictor being the most important predictor. To conduct this analysis, we first import and standardize the data. We then fit the linear regression model to the dataset using the lm function and save the results as fit, using the code:

```
fit <- lm(discW ~ beliefW + stigma + feminist + beliefM + gender, data =
dat21)
```

Where dat21 is the name of the dataset. After this, we specify our hypotheses in the hyp_test function. The hyp_test function can run multiple hypotheses in one function call if we separate the hypotheses by ;. We thus get the lines of code:

```
hyp_test (fit, “absolute beliefW > (stigma & feminist & beliefM & gender);
absolute stigma > (beliefW & feminist & beliefM & gender);
absolute feminist > (beliefW & stigma & beliefM & gender);
absolute beliefM > (beliefW & stigma & feminist & gender);
absolute gender > (beliefW & stigma & feminist & beliefM)”)
```

The function returns the following output (elements with an * were rounded or shortened):

Table 8. Output of the hyp_test function.

Hypothesis*	PosteriorProbability	HC	PriorProbability*	BF_Unconstrained*	BF_Complementary*
$\boldsymbol{H_1}$	0.9244	0.0756	0.2577	3.5873	35.2237
$\boldsymbol{H_2}$	0.0146	0.9854	0.2368	0.0617	0.0478
$\boldsymbol{H_3}$	0.0322	0.9678	0.1757	0.1832	0.1561
$\boldsymbol{H_4}$	6e-04	0.9994	0.2038	0.0029	0.0023
$\boldsymbol{H_5}$	0.0282	0.9718	0.1260	0.2237	0.2012

We can see that the hypothesis that beliefW is the most important predictor regardless of sign is very likely. It receives the highest PosteriorProbability and the highest Bayes factors by a wide margin. Note that the HC of beliefW being the most important predictor is the sum of the posterior probabilities of one of the other predictors being the most important predictor.

5.2 Quantifying the most important predictor

Now that beliefW is established as probably being the most important predictor of discW, we may want to quantify the effect of beliefW on discW. In a scenario where we have an idea of the magnitude of the beliefW predictor, but not of the sign of the effect, we could use absolute values to quantify the magnitude of the effect. A possible set of hypotheses are:

H_1: 0 < |beliefW| < 0.1
H_2: 0.1 < |beliefW| < 0.2
H_3: 0.2 < |beliefW| < 0.3
H_4: 0.3 < |beliefW| < 0.4
H_5: 0.4 < |beliefW| < 0.5
H_6: 0.5 < |beliefW| < 0.6
H_7: 0.6 < |beliefW| < 0.7

In H_1 , the effect of beliefW is between 0 and 0.1. In H_2 ,the effect is between 0.1 and 0.2, etc. In R, the hypotheses are formulated as:

```
hyp_test(fit, “absolute 0 < beliefw < 0.1;
absolute 0.1 < beliefw < 0.2;
absolute 0.2 < beliefw < 0.3;
absolute 0.3 < beliefw < 0.4;
absolute 0.4 < beliefw < 0.5;
absolute 0.5 < beliefw < 0.6;
absolute 0.6 < beliefw < 0.7”)
```

Yielding the results:

Hypothesis*	PosteriorProbability	HC	PriorProbability*	BF_Unconstrained*	BF_Complementary*
$\boldsymbol{H_1}$	0.003	0.997	0.051	0.059	0.056
$\boldsymbol{H_2}$	0.029	0.971	0.056	0.514	0.500
$\boldsymbol{H_3}$	0.176	0.824	0.049	3.587	4.140
$\boldsymbol{H_4}$	0.377	0.623	0.057	6.654	10.069
$\boldsymbol{H_5}$	0.299	0.701	0.049	6.094	8.262
$\boldsymbol{H_6}$	0.103	0.897	0.045	2.278	2.425
$\boldsymbol{H_7}$	0.013	0.987	0.045	0.286	0.277

We can be reasonably certain the effect of beliefW on discW is somewhere between |0.2| and |0.5|. The estimation does not seem to be accurate enough to reliably draw the conclusion that the effect of beliefW is somewhere in a 0.1 interval.

6 Discussion and conclusion

The present thesis demonstrated the hyp_test function. The function can evaluate posterior probabilities of inequality-constrained hypothesis for comparing regression predictors to each other or to fixed numbers. A notable feature of the function is the ability to compare absolute values of coefficients, which is relevant when we want to ignore the directions of effects. Previous programs such as lmhyp (Mulder & Olsson-Collentine, 2019), BIEMS (Mulder, Hoijtink & de Leeuw, 2012) and BAIN (Gu, Mulder & Hoijtink, 2018) were already able to compare coefficients using non-absolute values, but not using absolute values. The ability to compare coefficients using absolute values makes it possible to compare coefficients with different signs based on importance, allowing us to easily identify important predictors (see Maher & McLachlan, 1995; Darmadi-Blackberry et al., 2004; Vittersø, 2001 for examples of studies identifying important predictors). This ability can be useful in statistical modelling and theory construction. Absolute values can further be used in a more exploratory sense, where we have an idea about the magnitude of an effect but not about the direction of an effect. Finally, absolute values enable comparisons of positive and negative coefficients based on importance rather than sign.

A simulation study was conducted to test the effectiveness of the function on actual data, which revealed four major factors which influenced the effectiveness. A high sample size, high effect size differences between predictors, low number of predictors to compare and finally a low residual variance will lead to optimal results when the function is used in practice. The only factor of these we can readily control is the sample size. Expected effect size differences, number of variables to compare and expected variance of the residuals should thus be considered when sample size is selected.

Further optimizations to the function could include a more user-friendly interface and a more efficient method of estimating the posterior and prior probabilities. The method of drawing possible values for coefficients has the drawback of being computationally intensive, especially for less complex hypotheses, where the number of draws required to accurately assess the posterior and prior probabilities increases. For hypotheses which are so complex as to require more than 20 million draws, evaluation is not possible.

The increasing availability of programs and functions which can specify and evaluate Bayesian hypotheses will hopefully lead to a trend towards the use of Bayesian statistics as opposed to frequentist statistics.

Appendix 1: Supported hypotheses types

- Comparing a variable to another variable using < and > (e.g. $beliefW > stigma$)
- Comparing a set of variables using < and > (e.g. $beliefW > stigma < beliefM$)
  Returns the probability of beliefW > stigma while stigma < beliefM. Both conditions need to be true simultaneously. For independent hypothesis evaluation, use ; between hypotheses.
- Comparing a variable to a number using < and > (e.g. $beliefW > 0.1$)
  Numbers in R use dots for decimals, not commas. Using commas will return an error.
- Comparing a variable to two numbers using < and > (e.g. $0 < beliefW < 0.1$)
  Useful for estimating intervals where the regression coefficient may be.
- Comparing a variable to multiple other variables using <, >, () and & (e.g. $beliefW > (stigma \,\&\, beliefM)$)
  Returns the posterior probability of beliefW > stigma & beliefW being bigger than beliefM simultaneously. For independent hypothesis evaluation, use ;. Variables between brackets are separated using &.
- Comparing a fixed number to multiple variables using <, >, () and & (e.g. $0 > (stigma \,\&\, beliefM)$)
  Useful for seeing whether positive coefficients are different from 0. If negative coefficients are present, use absolutes.
- Comparing absolute values of variables using < and > (e.g. $absolute\ beliefW > (stigma \,\&\, beliefM)$)
  The absolute word in the hypothesis affects the entire hypothesis, not just the variable that has absolute in front of it. If the hypothesis is split using ;, absolute will only affect the hypothesis that has absolute in it.
- Comparing "stringed" hypotheses using <, > and ; (e.g. $beliefW > stigma; stigma > beliefM$). This returns two posterior probabilities, one for $beliefW > stigma$ and one for $stigma > beliefM$. ; is used to split the hypothesis.

Appendix 2: Simulating data in R

We can simulate data in R using the self-written "Simulation" function. Inputs required are n (number of people we want to simulate), the values of the coefficients input as a string, reps (how many times we want to repeat the simulation), hypothesis (the hypothesis we are testing input as a string). Sigma is an optional input which sets the standard deviation of the residuals (default = 1).
The simulation function draws n random values out of a normal distribution with mean 0 and standard deviation 1 for each independent variable, and n random values out of a normal distribution with mean 0 and standard deviation sigma for the error variable. The dependent variable Y is then computed as: $Y = 0 + \Sigma\, \beta_p * X_p + e$. After calculating Y, we fit the linear regression model of y being predicted by the independent variables using the lm function. As we know the value of the coefficients in the population, we are now able to test the accuracy of the hyp_test function.

To obtain our results, we import the lm object into the hyp_test function. The posterior probabilities are saved in the results variable, the Bayes factors are saved in the BFH0 and BFHC variables and the prior probabilities of the hypotheses are saved in the Prior variable. The function now redraws the independent variables and error values, recomputes Y using the changed values and refits the linear regression model. The hyp_test function runs again for this new dataset. We repeat this reps times. As the results are returned as matrices, we can choose what we want to do with them afterwards (e.g. plotting, calculating means or standard deviations, etc.).

Appendix 3: Detailed explanation of hyp_test functioning

This Appendix aims to explain the functioning of the hyp_test function in more detail. The Appendix will discuss data types in R, explain the concept of matrix multiplication and its usefulness, apply matrix multiplication to the hyp_test function and finally explain Boolean expressions and their usage in obtaining the outcome variables of the hyp_test function.

Data types in R

Data can be stored in multiple ways in R. The data types relevant for this thesis are strings, vectors and matrices. Strings are combinations of characters between quotation marks. The sentence "Apples == Pears" is a string. Strings are useful because R treats all digits and operators inside a string as characters, which means R will not try to compare Apples to Pears. If we type the sentence Apples == Pears in R outside a string, R tries to compare Apples to Pears and returns an error.
Vectors are one-dimensional series of data (not necessarily numeric). For example, 2, 3 and 4 would be a vector, but a, b and c can also be a vector.
Matrices are two-dimensional series of data. Where vectors have only a length, matrices have a length and "height". $\begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}$ would be a matrix. Matrices are very useful because R can do special operations on them, like matrix multiplication (Venables & Smith, 2003), as shown in the section below.

Matrix multiplication & inequality matrices in general

Matrix multiplication is a special mathematical operation in R that can only be used on matrices or data that is convertible to a matrix. When we have the following matrices: $\begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$ and $\begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix}$ and we multiply the first matrix with the second matrix, you may intuitively expect the result to be $1 * 2 = 2, 2 * 4 = 8, 3 * 6 = 18$, for a matrix of $\begin{bmatrix} 2 & 8 & 18 \end{bmatrix}$. This is not what matrix multiplication does.
When we use matrix multiplication in R, we get a matrix with the number of rows of the first matrix (1) and the number of columns of the second matrix (1). This property is useful, as we can control the number of columns and rows of the resulting matrix. In the present example, we only get one number, as we have only 1 row and 1 column. The number we get is the dot product of the matrices. We get the dot product by multiplying the first row (1, 2 and 3) from the first matrix and multiplying it with the first column (2, 4 and 6) from the second matrix. We then take the sum. $1 * 2 = 2, 2 * 4 = 8, 3 * 6 = 18$. The sum is $2 + 8 + 18 = 28$, therefore the answer is a one-row one-column matrix containing 28, not a one-row three-column matrix containing 2, 8, 18, as one might suspect at first glance. Note that the number of columns of the first matrix must be equal to the number of rows of the second matrix, otherwise matrix multiplication is not possible. Matrix multiplication is also order-sensitive. Multiplying matrix 1 with matrix 2 does not give the same outcome as multiplying matrix 2 with matrix 1 (Venables & Smith, 2003).

Matrix multiplication and inequality matrices in hyp_test

We can use matrix multiplication to make our post_draws matrix fit our hypothesis ($\beta_1 - \beta_2 > 0$). To do this, we require a so-called inequality matrix. The matrix is called an inequality matrix as our hypothesis is one of inequality of β_1 and β_2. The variables of interest are the second and third column of post_draws (the matrix containing the estimates of regression coefficients). All columns that are not of interest get a 0 in the inequality matrix. We need to subtract β_2 from β_1, which means β_2 gets a -1 in the inequality matrix and β_1 gets a 1. We end up with the matrix:

$$\begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 0 \end{bmatrix}$$

We would like to matrix multiply this inequality matrix with the post_draws matrix. The number of columns of the first matrix(6) is not equal to the number of rows of the second matrix(5000). This means we must transpose (flip) the post_draws matrix, creating a matrix with 5000 columns and 6 rows instead of 6 columns and 5000 rows. The matrix multiplication results for the first three columns will be shown here.

$$[0 \quad 1 \quad -1 \quad 0 \quad 0 \quad 0] \text{ and } \begin{bmatrix} 0.1 & -0.1 & 0.03 \\ 0.39 & 0.41 & 0.31 \\ 0.03 & -0.06 & 0.31 \\ 0.18 & 0.05 & 0.17 \\ -0.14 & 0.07 & -0.14 \\ -0.08 & -0.22 & 0.08 \end{bmatrix}$$

The matrix multiplication displayed above will yield a matrix with 1 row and 3 columns. The dot product for the first number is: $0 * 0.1 \ + \ 1 * 0.39 \ + \ -1 * 0.03 \ + \ 0 * 0.18 \ + \ 0 * -0.14 \ + \ 0 * -0.08$. This simplifies to $1 * 0.39 \ + \ -1 * 0.03$, which simplifies to $0.39 - 0.03 \ = \ 0.36$. The number we calculated here is $\beta_1 - \ \beta_2$. The same holds true for the other two numbers $(0.41 - 0.06 \ = \ 0.47 \ and \ 0.31 - 0.31 \ = \ 0)$. The multiplication results in the result-matrix:

$$[0.36 \quad 0.47 \quad 0]$$

When we execute this operation in R with the entirety of post_draws, we obtain a 5000 column, 1 row matrix with 5000 results of $\beta_1 - \ \beta_2$. We call this matrix "result". The only thing left to do is to check how many times out of the 5000 obtained values $\beta_1 - \ \beta_2 > 0$.
To check the amount of times $\beta_1 - \ \beta_2 > 0$, we can use Boolean expressions.

Intermezzo: Boolean expressions

Boolean expressions is the evaluation of statements using TRUE or FALSE. For example, $1 \ > \ 2$ is FALSE. $1 \ < \ 2$ is TRUE. Applied to comparison of the regression coefficients, the statement $\beta_1 - \beta_2 > 0$ is TRUE when $\beta_1 - \ \beta_2$ returns a number larger than 0. $\beta_1 - \ \beta_2 > 0$ is FALSE when $\beta_1 - \beta_2$ returns a number smaller than or equal to 0.
Using the Boolean expression $\beta_1 - \ \beta_2 > 0,$ we make R check whether the values in the result-matrix are larger than 0. R returns TRUE or FALSE for every value in the matrix. The results-matrix will be:

$$[TRUE \quad TRUE \quad FALSE]$$

When we do arithmetic on TRUE or FALSE values in R, TRUE counts as 1 and FALSE counts as 0. The mean of the result-matrix is therefore equal to the proportion of TRUE values in the result-matrix. The mean of the result-matrix gets saved as the PosteriorProbability variable. The HC variable is 1-PosteriorProbability. The hypothesis "beliefW > stigma" yields a PosteriorProbability of 0.9806 and a HC of 0.0194. This means that in 98,06% of the 5000 draws, $\beta_1 - \ \beta_2 > 0$ was true. The hypothesis is therefore very likely to be true. This makes sense, as our point-estimate of β_1 was much higher than our point estimate of β_2.

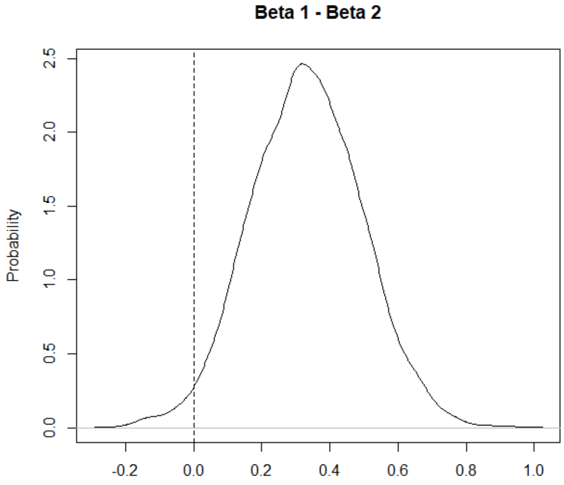

Figure 17. $\beta_1 - \beta_2$ *with dotted line.*

Revisiting the earlier graph, the area to the right of the dotted line is equal to 0.9806, and the area to the left of the line is equal to 0.0194. Now that we have the posterior probability, we can compute the Bayes factors. The hyp_test function computes two Bayes factors; one Bayes factor for the constrained model against the unconstrained model (the hypothesis) and one Bayes factor for the constrained model against the complementary model (the opposite of the hypothesis). The first Bayes factor is calculated by $BF_{Unconstrained} = \frac{PosteriorProbability}{PriorProbability}$, where PriorProbability is calculated using the same method as the PosteriorProbability variable. The only difference between the prior and posterior probability is that the prior probability is generated by a multivariate T-distribution where all centrality parameters are set to zero. The outcome is how much more likely our hypothesis is after observing the data. The second Bayes factor is calculated by $BF_{Complementary} = \frac{BF_{Unconstrained}}{BF_{HC}}$, where $BF_{HC} = \frac{1-PosteriorProbability}{1-PriorProbability}$.

Appendix 4: Composing an inequality matrix for 2+ variables

The hypothesis $\beta_1 > \beta_2 < \beta_3 > \beta_4 > \beta_5$ can be split in 4 parts. $\beta_1 > \beta_2$, $\beta_2 < \beta_3$, $\beta_3 > \beta_4$ and $\beta_4 > \beta_5$. If these four parts are all true simultaneously, the main hypothesis is also true. The hyp_test function first evaluates the parts of the hypothesis individually. The inequality matrix for this hypothesis is:

$$\begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

The first row of the matrix corresponds to $\beta_1 > \beta_2$, the second row corresponds to $\beta_2 < \beta_3$, the third row corresponds to $\beta_3 > \beta_4$ and the fourth row corresponds to $\beta_4 > \beta_5$. When we multiply this matrix with the first three columns of the transposed post_draws matrix, the resulting matrix will be a matrix of 4 rows and 3 columns.

$$\begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} \text{and} \begin{bmatrix} 0.1 & -0.1 & 0.03 \\ 0.39 & 0.41 & 0.31 \\ 0.03 & -0.06 & 0.31 \\ 0.18 & 0.05 & 0.17 \\ -0.14 & 0.07 & -0.14 \\ -0.08 & -0.22 & 0.08 \end{bmatrix}$$

The first row contains $\beta_1 - \beta_2$, the second row contains $-\beta_2 + \beta_3$, also written as $\beta_3 - \beta_2$, the third row contains $\beta_3 - \beta_4$ and the fourth row contains $\beta_4 - \beta_5$. The resulting matrix is:

$$\begin{bmatrix} 0.36 & 0.47 & 0 \\ 0.15 & 0.11 & -0.14 \\ 0.32 & -0.02 & 0.31 \\ -0.06 & 0.29 & -0.24 \end{bmatrix}$$

Booleans are used again to evaluate whether the values in the results matrix are > 0. We get the following matrix:

$$\begin{bmatrix} TRUE & TRUE & FALSE \\ TRUE & TRUE & FALSE \\ TRUE & FALSE & TRUE \\ FALSE & TRUE & FALSE \end{bmatrix}$$

To check whether our main hypothesis is correct, we take the product of the columns. The product yields 1 if all values in a given column are true, and 0 if one or more of the values are false. If the product is 1, the main hypothesis is true in that column; if the product is 0, the main hypothesis is not true in that column. When we do this for all 5000 columns, we get a posterior probability of 0.513 of the main hypothesis being true, and a HC of 0.487. This means we are still quite uncertain about the hypothesis. This makes sense, as the point-estimates of stigma and feminism and beliefM and gender are quite close; we cannot reliably distinguish between these coefficients (when we remove the beliefM and gender comparison, the posterior probability jumps to 0.7126).

Appendix 5: Composing inequality matrices for fixed number comparisons

1. Comparing a variable to one fixed value

As we are not comparing β_1 to another variable, we want to create a result-matrix which contains just the estimated values of β_1. We can do this by using the inequality matrix:

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

When we matrix multiply this inequality matrix with the first three columns of the transposed post_draws matrix, we get:

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \text{ and } \begin{bmatrix} 0.1 & -0.1 & 0.03 \\ 0.39 & 0.41 & 0.31 \\ 0.03 & -0.06 & 0.31 \\ 0.18 & 0.05 & 0.17 \\ -0.14 & 0.07 & -0.14 \\ -0.08 & -0.22 & 0.08 \end{bmatrix} \text{ yields } \begin{bmatrix} 0.39 & 0.41 & 0.31 \end{bmatrix}$$

In words; we just get the estimated values of β_1. We can then compare these values to the value specified in the hypothesis $\beta_1 > 0.1$. This yields $[TRUE \quad TRUE \quad TRUE]$. Do this for all 5000 columns, and we get a posterior probability of 0.997 and a HC of 0.003.

2. Comparing a variable to multiple fixed values

Example hypothesis $0.2 < \beta_1$ and $\beta_1 < 0.3$ yields an inequality matrix of:

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

When we multiply this with the transposed post_draws matrix, we get:

$$\begin{bmatrix} 0.39 & 0.41 & 0.31 \\ 0.39 & 0.41 & 0.31 \end{bmatrix}$$

We check whether the first row of the result-matrix is > 0.2 and whether the second row of the result-matrix is < 0.3. We get:

$$\begin{bmatrix} TRUE & TRUE & TRUE \\ FALSE & FALSE & FALSE \end{bmatrix}$$

To satisfy the hypothesis, both row values in the same column need to be TRUE. We check this by taking the product of the columns. When we do this for all 5000 columns, we get a posterior probability of 0.1758 and a HC of 0.8242. The hypothesis is not very likely to be true.

Appendix 6: Bayes factors for incorrect hypotheses and decreasing complexity

Bayes factors increase as the complexity of hypotheses decreases if the posterior probability remains equal. This raises the question whether an extremely complicated, wrong hypothesis can still receive a Bayes factor that is generally larger than 1. This option is explored in this Appendix.

First, Bayes factors for cases where the hypothesis is true in the population were calculated for N = 50, sigma = 1, effect size = 0.2 and an increasing number of variables. In this instance, comparison using non-absolute values was used, due to absolute values requiring more draws than non-absolute values. The conclusions remain the same for both types of comparison. The proportion of times the Bayes factor was larger than 1 was then calculated and plotted in the graph below.

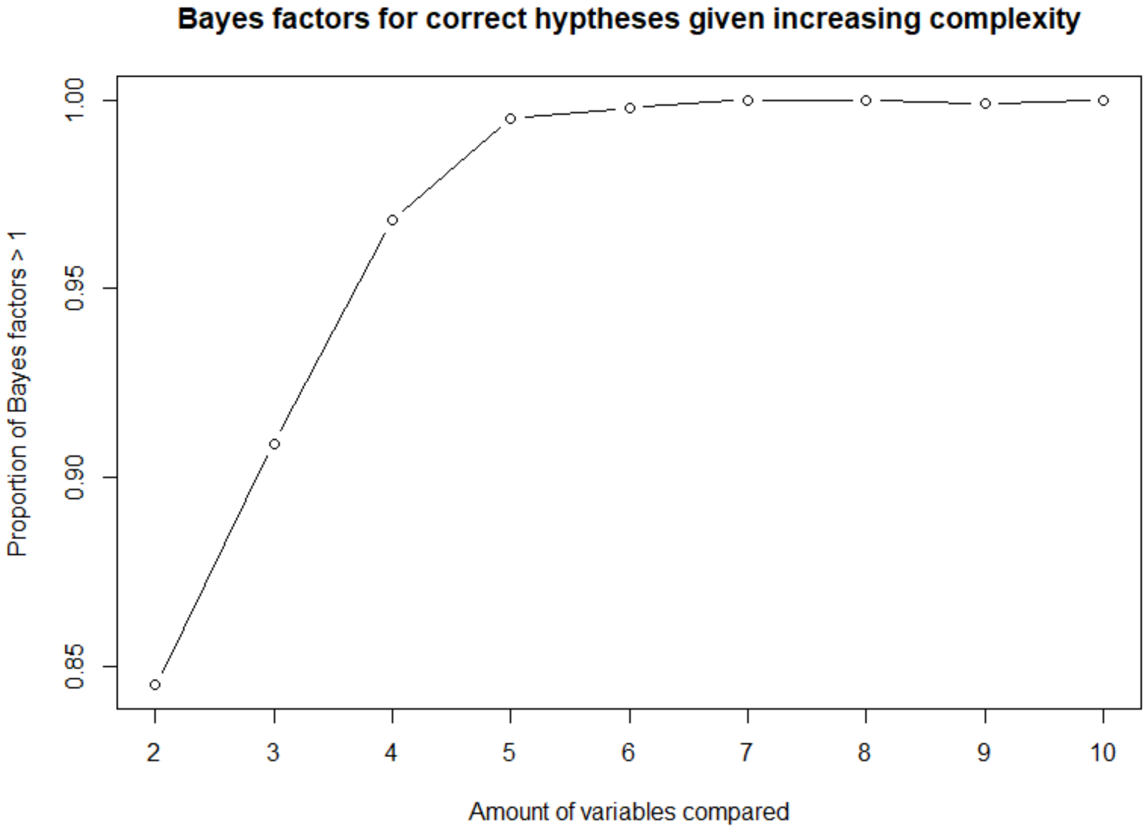

Figure 18. Bayes factors > 1 for differing complexity.

Looking at the graph, it seems the proportion of Bayes factors which point us in the right direction steadily increases until we reach a hypothesis containing five variables. It then stays around 1 for decreasingly complex hypotheses. To evaluate the Bayes factor accurately for complex hypotheses, an ever-increasing number of reps is needed as the complexity of the hypothesis shrinks. A less complex hypothesis therefore takes a longer time to evaluate than a shorter hypothesis.

Secondly, Bayes factors were calculated for hypotheses which are entirely untrue in the population (e.g. if x1 > x2 > x3 in the population, x1 < x2 < x3 is entirely untrue). For hypotheses that are entirely untrue, the Bayes factors follow the pattern displayed below.

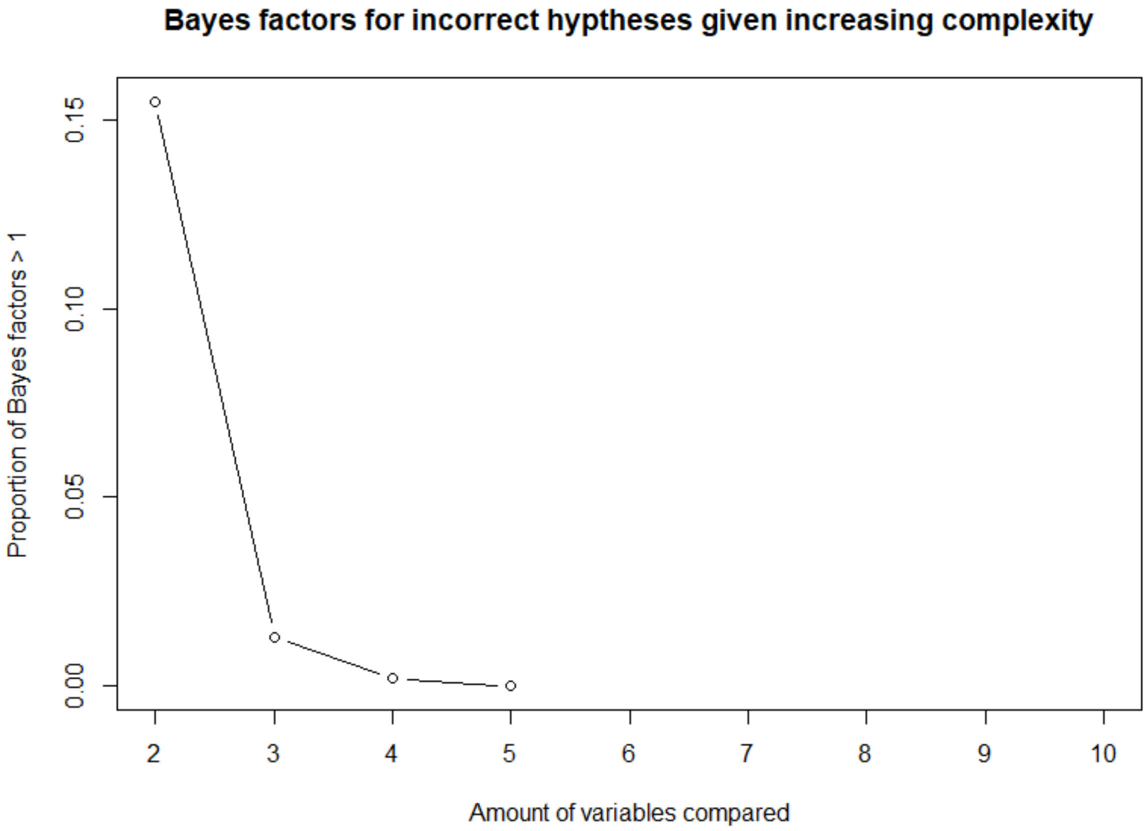

Figure 19. Bayes factors > 1 for differing complexity considering entirely untrue hypotheses.

The Bayes factor for entirely untrue hypotheses is very rarely above 1, and this proportion only proceeds to get lower as the complexity decreases. For more than 4 variables, the posterior probability occasionally reaches 0. The graph therefore stops at this point, as any number divided by zero has an undefined outcome.

Hypotheses can also be partly incorrect. If in the population x1 > x2 > x3 > x4, the hypothesis x1 > x2 > x3 < x4 is less incorrect than x1 < x2 < x3 < x4. The trend of the Bayes factor was also examined for partly incorrect hypotheses. For hypotheses that are only half-incorrect, the proportion of Bayes factors above 1 is displayed in the graph below.

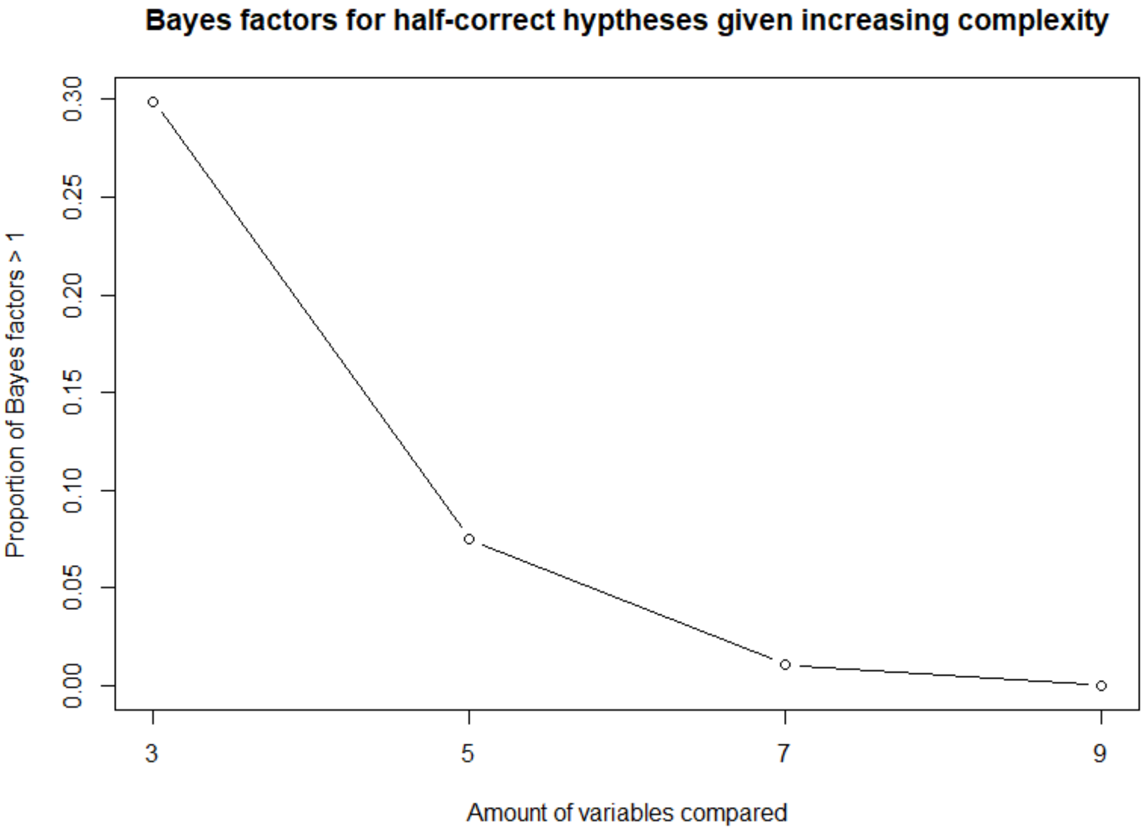

Figure 20. Bayes factors > 1 for differing complexity considering half-untrue hypotheses.

Hypotheses which are half-untrue follow the same pattern as the proportion of Bayes factors > 1 for the entirely untrue hypotheses but take longer to reach a zero point.

Finally, for hypotheses that are almost entirely true (only having 1 misdirected sign), the pattern of Bayes factors is:

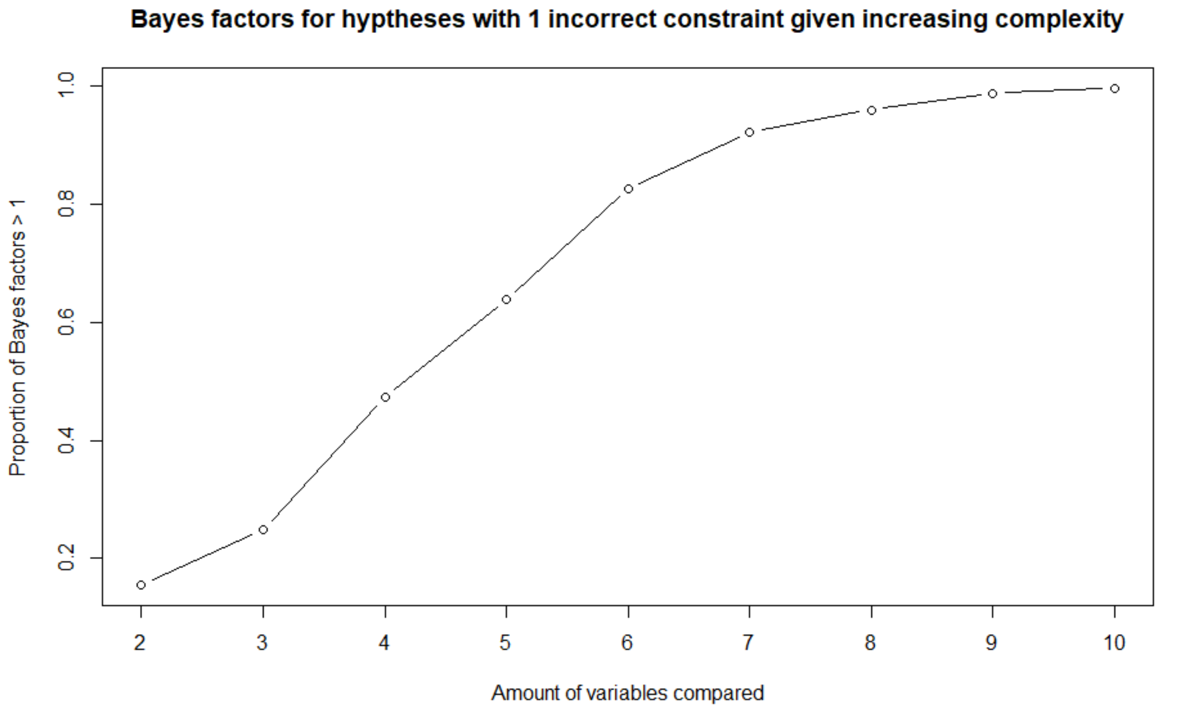

Figure 21. Bayes factors > 1 for differing complexity considering hypotheses with 1 incorrect sign.

Hypotheses which are almost entirely true can produce quite misleading Bayes factors as the complexity of hypotheses decreases. One way to combat this effect is to increase the sample size. Increasing the sample size to 1000 from 50 yields:

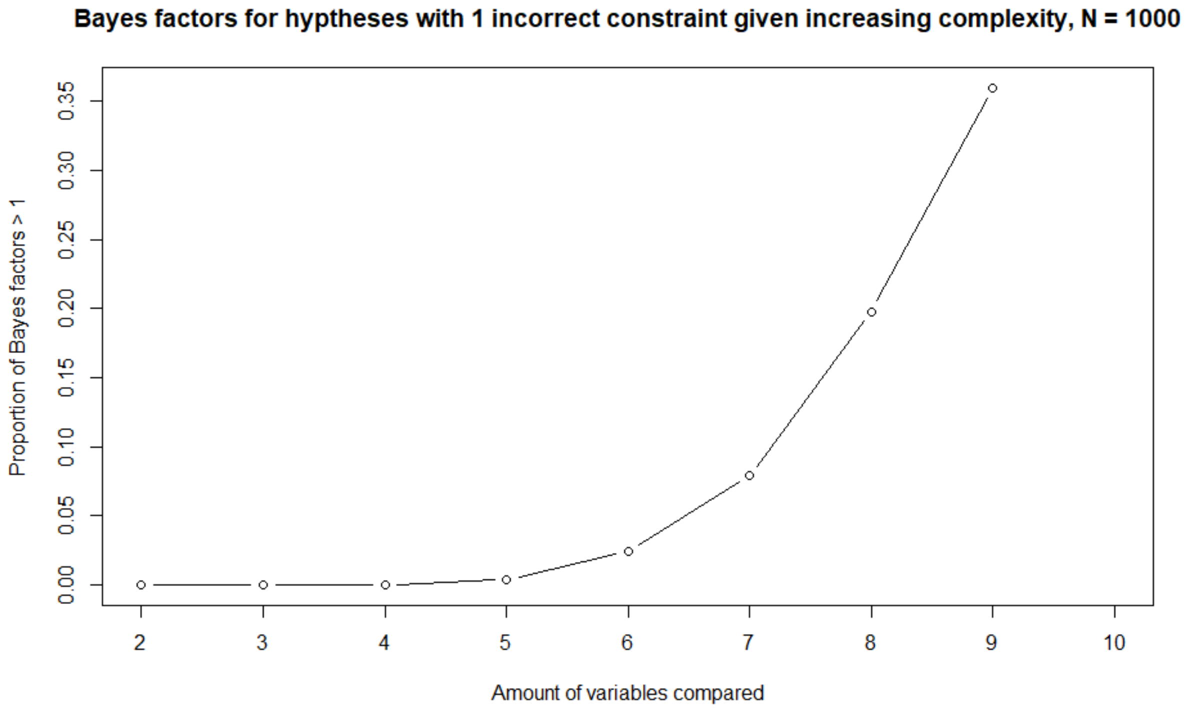

Figure 22. Bayes factors > 1 for differing complexity considering hypotheses with 1 incorrect sign and N = 1000.

In the graph, we can see that the misleading Bayes factors problem was largely solved for comparison of 6 or less variables but arises again when the number of variables rises above this. Further increasing the sample to 2000 yields:

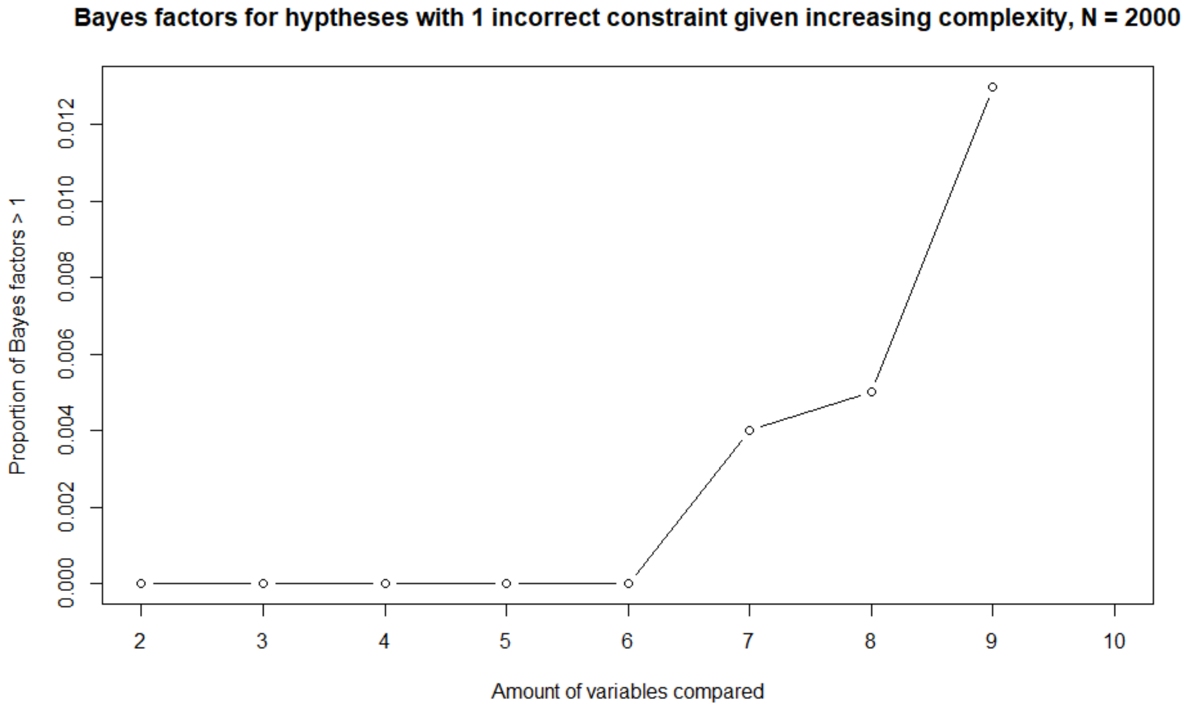

Figure 23. Bayes factors > 1 for differing complexity considering hypotheses with 1 incorrect sign and N = 2000.

As we can see in the graph, a sample size of 2000 is adequate for comparing this number of variables.

To reliably assess whether a complex hypothesis is completely true, multiple options are available. Firstly, we could obtain a huge sample. Taking a huge sample solves the problem but is potentially costly and time consuming.

A second possible course of action is estimating the Bayes factors for all possible orderings of the coefficients and selecting the hypothesis with the highest Bayes factor. The Bayes factor of a correct hypothesis will very often be larger than or equal to the Bayes factor of a hypothesis with one incorrect sign. The proportion of times the Bayes factor for an entirely correct hypothesis with n = 50 and effect size = 0.2 is larger or equal to the Bayes factor for a hypothesis with 1 incorrect sign is presented in the table below.

Table 9. Proportion of times the Bayes factor for a correct hypothesis is larger than or equal to the Bayes factor for a hypothesis with 1 incorrect sign, with N = 50, effect size =0.2 and sigma = 1.

Number of variables compared	2	3	4	5	6	7	8	9	10
Proportion of equal or higher BF's	0.845	0.899	0.912	0.947	0.963	0.957	0.967	0.959	0.961

A third solution is interpreting the Bayes factor only in conjunction with the posterior probability. Low posterior probabilities combined with high Bayes factor can be seen as requiring a larger sample size to obtain a definitive accurate assessment of the truth of the hypothesis.

Overly complex hypotheses

Decreasingly complex hypotheses require an increasing number of draws for the posterior and prior probability estimate to remain accurate. If the number of draws needed for accurate hypothesis evaluation is higher than +- 20 million (around 11 variables compared), R will return an error stating that a vector of this size cannot be allocated. For this reason, overly complex hypotheses are discouraged.

Appendix 7: Other types of comparisons

This appendix aims to describe two types of possible coefficient comparisons which are possible in hyp_test but were not discussed in the main article.

Comparing a coefficient to multiple other coefficients

It is possible we want to compare one coefficient to multiple coefficients at once. This procedure is best used to find the degree of confidence we have in a single predictor being the strongest of a set of positive predictors. A possible hypothesis would be: $\beta_1 > (\beta_2 \,\&\, \beta_3 \,\&\, \beta_4 \,\&\, \beta_5)$. This hypothesis states that β_1 is the most important positive predictor (use absolute values to establish the most important predictor regardless of sign). Composing the inequality matrix and the result-matrix follows the procedure described above for comparing multiple coefficients.

Compare a coefficient to one fixed value

We can also compare a coefficient to a fixed number in order to estimate an effect size. One possible hypothesis would be $\beta_1 > 0.1$ (the effect of beliefW is larger than 0.1). To test this hypothesis, we must first compose our inequality-matrix. See Appendix 5 for details on composing an inequality matrix for comparing coefficients to fixed values. The hyp_test function returns a posterior probability of 0.997 and a HC of 0.003.

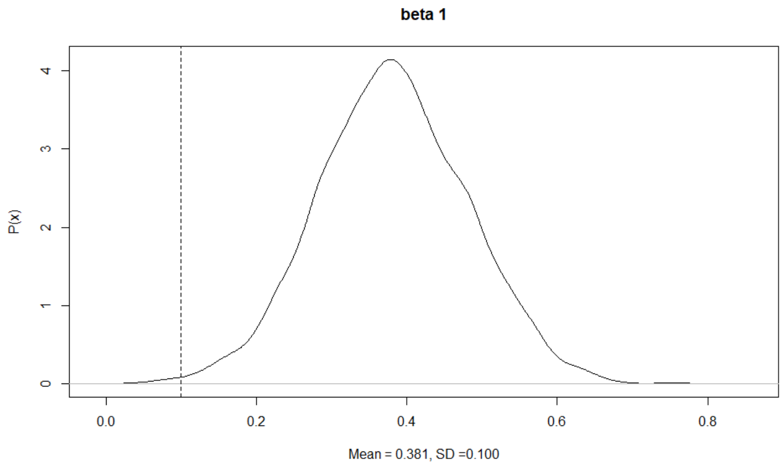

Figure 24. Posterior probability distribution of the first regression coefficient with a dotted line at 0.1.

Appendix 8: Simulation study using non-absolute values

This appendix repeats the simulation study conducted earlier, but this time uses non-absolute values to evaluate the hypothesis $\beta_4 > \beta_3 > \ \beta_2 > \ \beta_1$, where $\beta_1 = 0.2, \beta_2 = 0.4,\ \beta_3 = 0.6\ and\ \beta_4 = 0.8$. The conclusions are very much the same as when we used absolute values. One notable difference is that a comparison using absolute values requires more draws than a comparison using non-absolute values. As the conclusions are identical, only the graphs are presented here.

6.1.1 Varying sample size

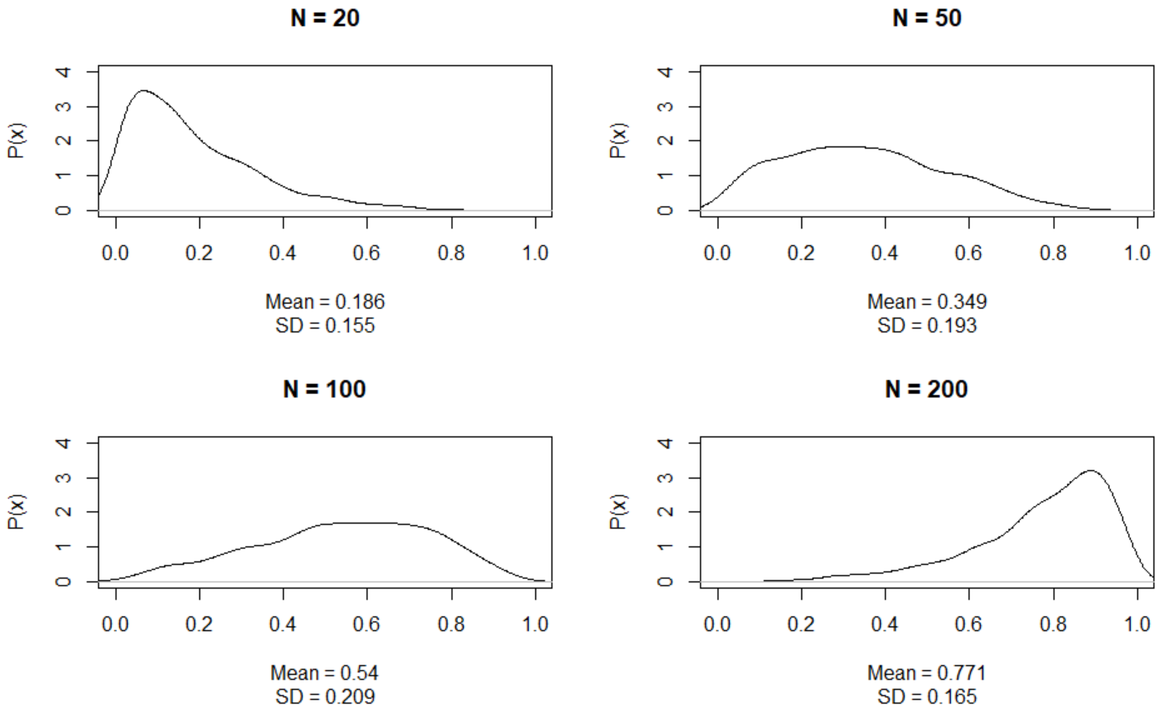

Figure 25. Posterior probability distributions for the hypothesis x4 > x3 > x2 > x1 for differing sample sizes.

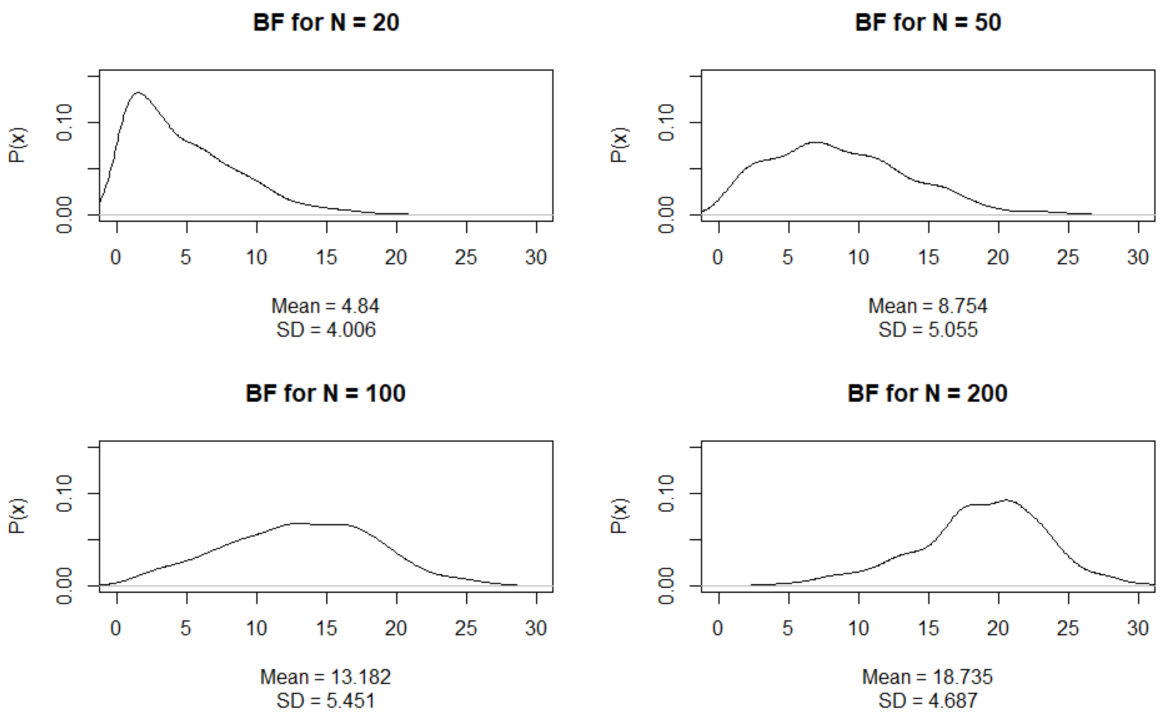

Figure 26. Probability distributions for Bayes factors considering the hypothesis x4 > x3 > x2 > x1 with varying sample sizes.

Table 10. Proportion of Bayes factors considering the hypothesis x4 > x3 > x2 > x1 above 1 for varying sample sizes.

BF > 1 for N = 20	BF > 1 for N = 50	BF > 1 for N = 100	BF > 1 for N = 200
0.853	0.973	0.997	1

6.1.2 Varying effect size

When we fix the N to 50 and we vary the effect size, we get the following plots.

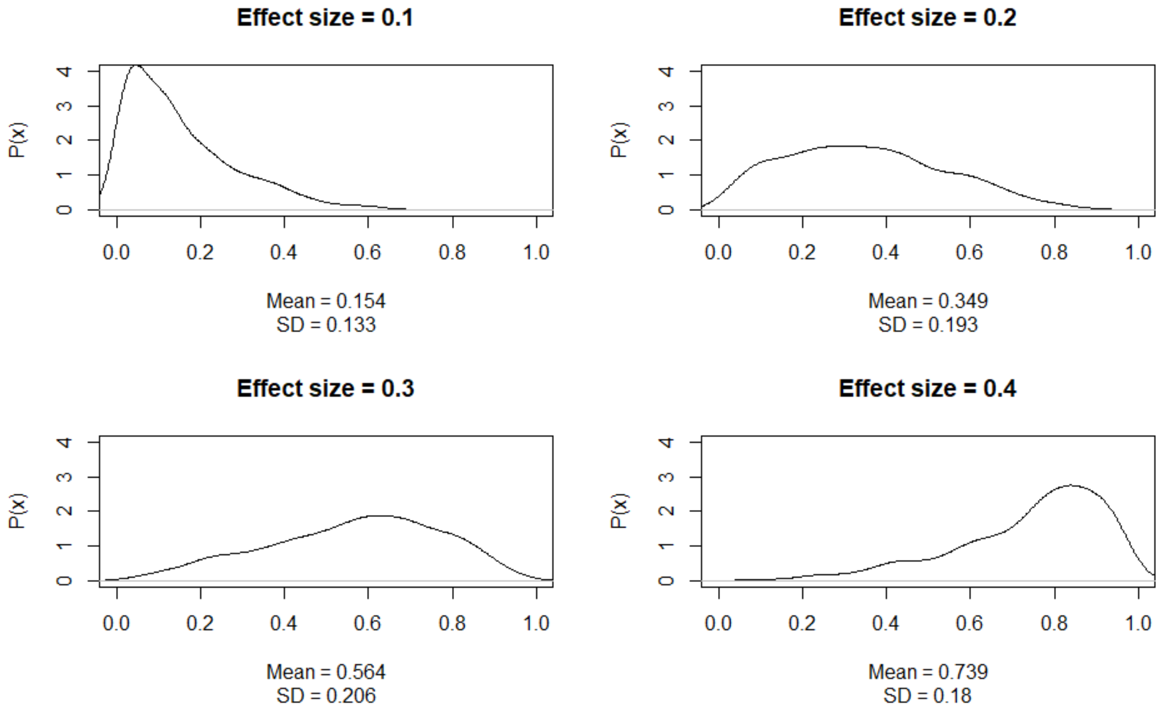

Figure 27. Posterior probability distributions considering the hypothesis x4 > x3 > x2 > x1 for given effect sizes.

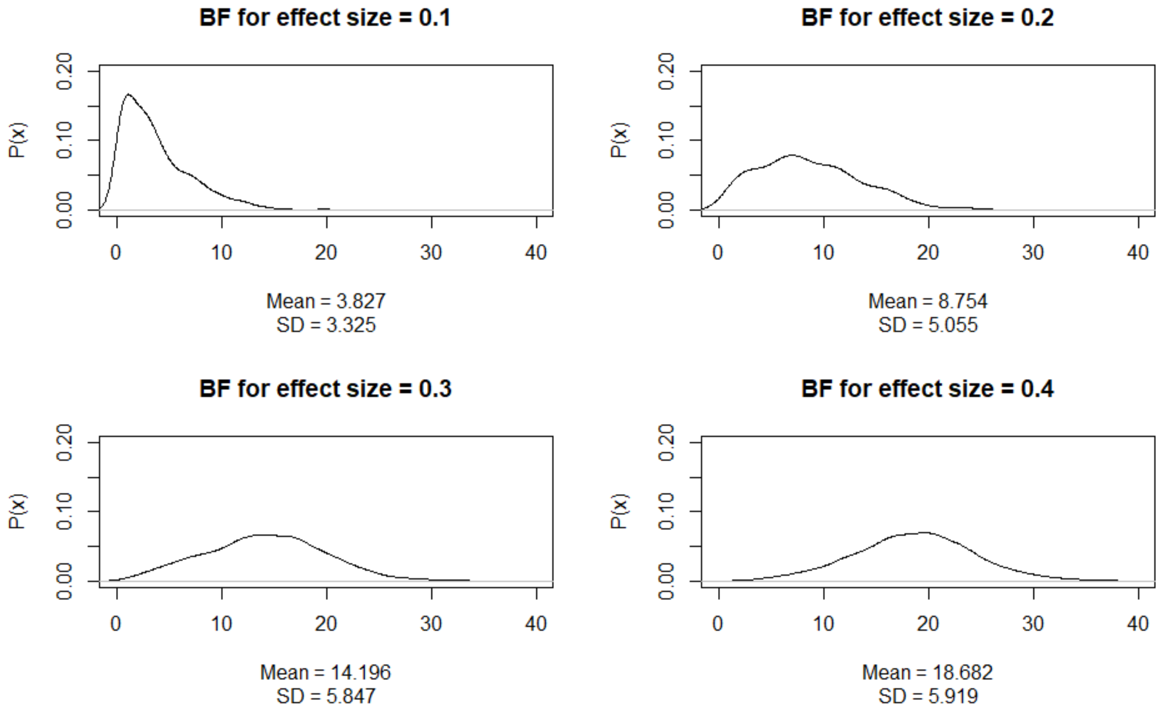

Figure 28. Probability distributions of Bayes factors considering the hypothesis x4 > x3 > x2 > x1 for given effect sizes.

Table 11. Proportion of Bayes factors considering the hypothesis x4 > x3 > x2 > x1 above 1 for different effect sizes.

BF > 1 for effect size = 0.1	BF > 1 for effect size = 0.2	BF > 1 for effect size = 0.3	BF > 1 for effect size = 0.4
0.797	0.973	0.997	1

6.1.3 Varying number of compared variables

The number of variables compared is also a factor of consideration. When we test a hypothesis that is true in the population, keep the N constant to 50 and fix the difference between variables to 0.2, we get the following graphs for comparing two, three, four and five variables respectively.

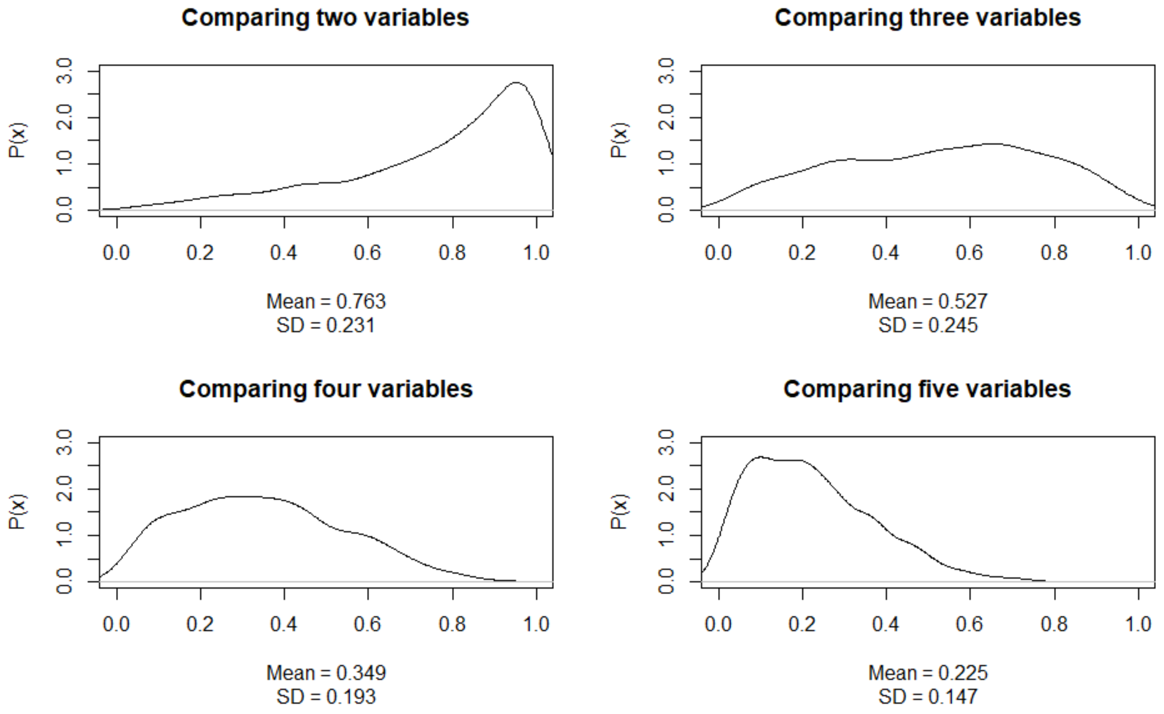

Figure 29. Posterior probability distributions for comparing different amounts of variables.

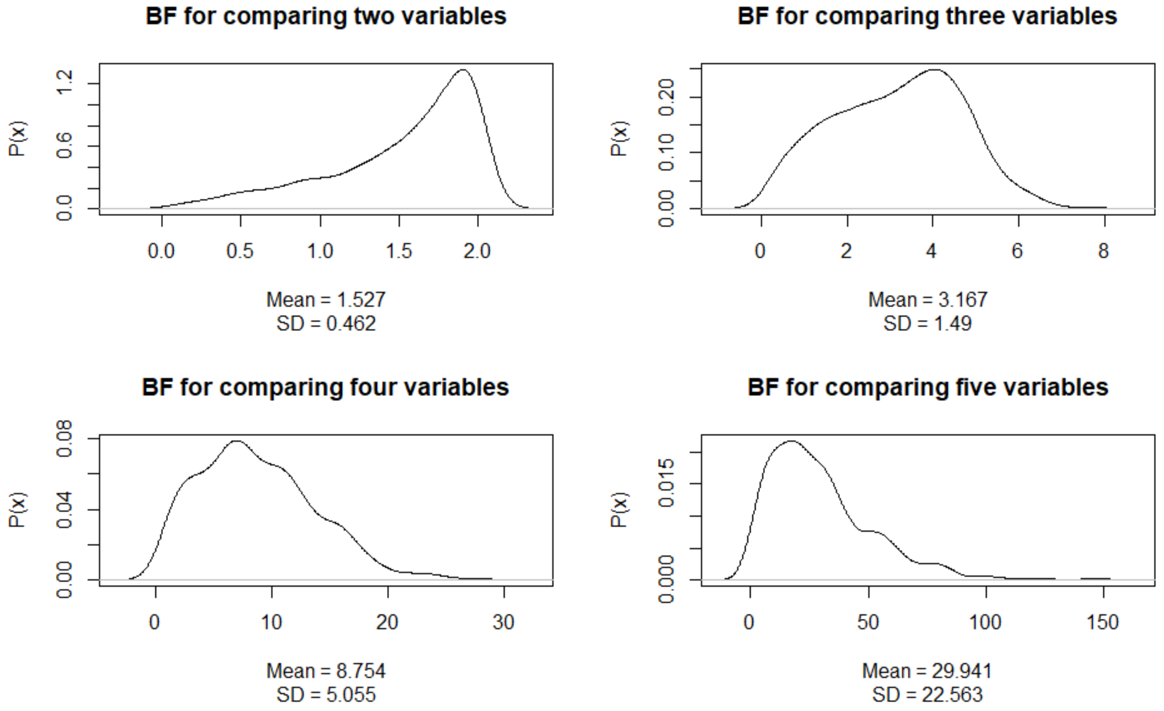

Figure 30. Probability distributions of Bayes factors for different numbers of compared variables.

Table 12. Proportion of Bayes factors above 1 for different numbers of compared variables.

BF > 1 for two variables	BF > 1 for three variables	BF > 1 for four variables	BF > 1 for five variables
0.845	0.909	0.973	0.984

6.1.4 Varying sigma

Finally, the standard deviation of the residuals (sigma) should be considered. For N=50, effect size = 0.2, number of variables = 4, PosteriorProbability results for differing sigma's are presented below.

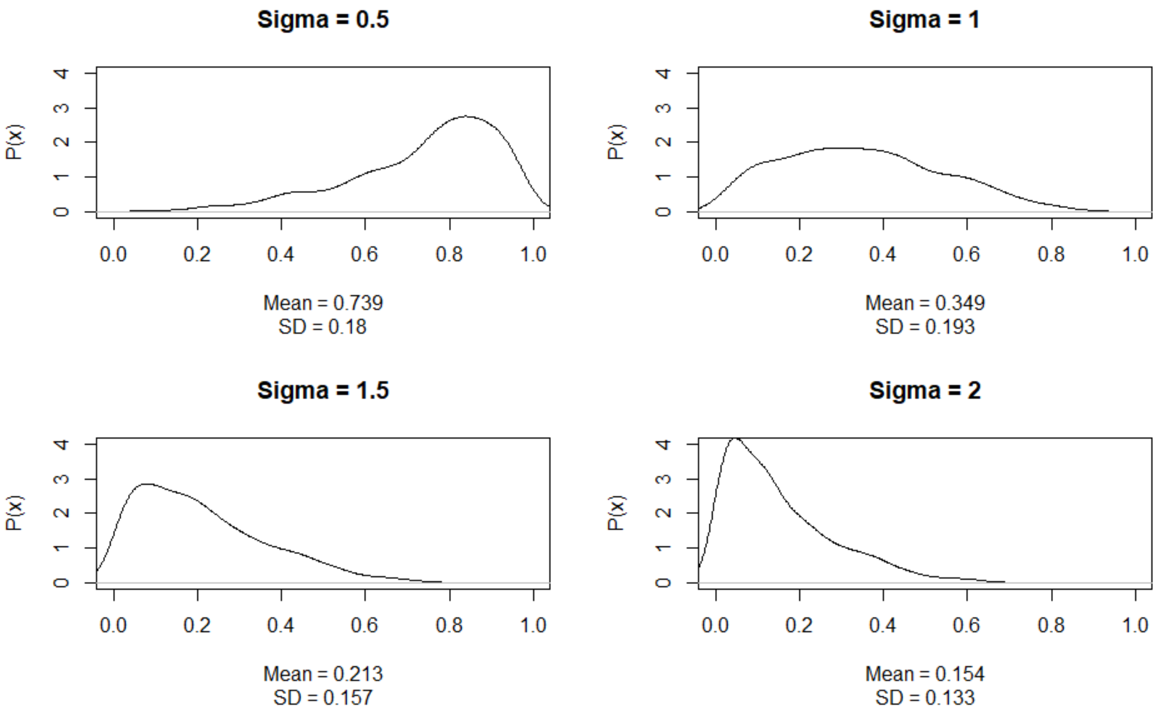

Figure 31. Posterior probability distributions considering the hypothesis x4 > x3 > x2 > x1 for differing values of sigma.

Plots for the Bayes factors and the table containing the proportion of Bayes factors > 1 are displayed below.

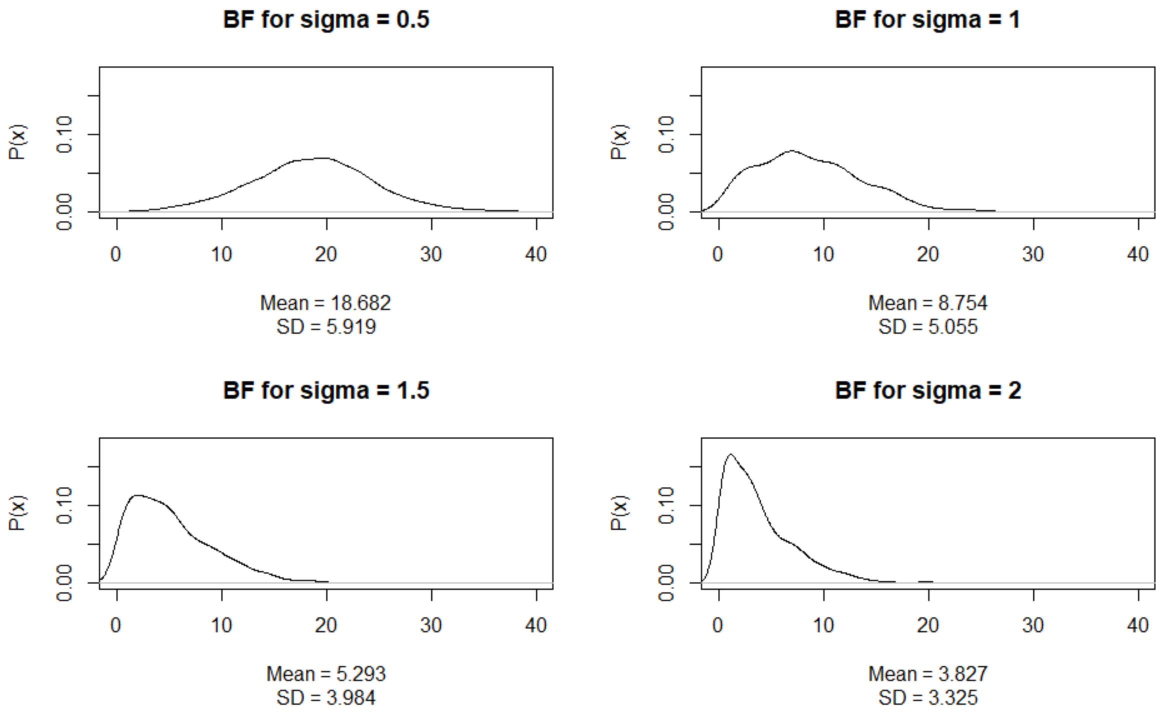

Figure 32. Probability distributions of Bayes factors considering the hypothesis x4 > x3 > x2 > x1 for differing values of sigma.

Table 13. Proportion of Bayes factors considering the hypothesis x4 > x3 > x2 > x1 above 1 for differing values of sigma.

BF > 1 for sigma = 0.5	BF > 1 for sigma = 1	BF > 1 for sigma = 1.5	BF > 1 for sigma = 2
1	0.973	0.883	0.797

References

Bolstad, W., M. (2004). *Introduction to Bayesian statistics.* Hoboken, NJ: John Wiley & Sons Inc.

Carlsson, R., & Sinclair, S. (2017). Prototypes and same-gender bias in perceptions of hiring discrimination. *The Journal of Social Psychology, 158 (3),* 285–297. doi: 10.1080/00224545.2017.1341374

Clower, C. E., & Bothwell, R. K. (2001). An exploratory study of the relationship between the Big Five and inmate recidivism. *Journal of Research in Personality*, *35(2),* 231-237. doi: https://doi.org/10.1006/jrpe.2000.2312

Darmadi-Blackberry, I., Wahlqvist, M. L., Kouris-Blazos, A., Steen, B., Lukito, W., Horie, Y., & Horie, K. (2004). Legumes: The most important dietary predictor of survival in older people of different ethnicities. *Asia Pacific Journal of Clinical Nutrition, 13(2),* 217-220.

Gu, X., Mulder, J., & Hoijtink, H. (2018). Approximated adjusted fractional Bayes factors: A general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology, 71,* 229-261. doi: 10.1111/bmsp.12110

Hájek, A. (2012). Interpretations of probability. Retrieved from https://plato.stanford.edu/archives/win2012/entries/probability-interpret/.

Hájek, A. (1996). "Mises redux"-redux: Fifteen arguments against finite frequentism. *Erkenntnis, 45,* 209-227.

Hoijtink, H. (2012). *Informative hypotheses: Theory and practice for behavioral and social scientists.* Boca Raton, FL: CRC Press.

Hubbard, R., & Armstrong, J. S. (2006). Why we don't really know what statistical significance means: Implications for educators. *Journal of Marketing Education, 28(2),* 114-120. doi:10.1177/0273475306288399

Koufaris, M., Kambil, A., & Priscilla, A. L. (2002). Consumer behavior in web-based commerce: An empirical study. *International Journal of Electronic Commerce, 6(2),* 115-138. doi: https://doi.org/10.1080/10864415.2001.11044233

Maher, J., & McLachlan, R. S. (1995). Febrile convulsions: Is seizure duration the most important predictor of temporal lobe epilepsy? *Brain, 118(6),* 1521-1528. doi: https://doi.org/10.1093/brain/118.6.1521

Mulder, J., Hoijtink, H., & De Leeuw, C. (2012). BIEMS: A Fortran 90 program for calculating Bayes factors for inequality and equality constrained models. *Journal of Statistical Software, 46(2),* 1-39. doi: http://dx.doi.org/10.18637/jss.v046.i02

Mulder, J., & Olsson-Collentine, A. (2019). Simple Bayesian testing of scientific expectations in linear regression models. *Behavior research methods, 1-14*. doi: https://doi.org/10.3758/s13428-018-01196-9

Pallant, J. (2007). *SPSS survival manual.* New York, NY: Open University Press.

Paternoster, R., Brame, R., Mazerolle, P., & Piquero, A. (1998). Using the correct statistical test for the equality of regression coefficients. *Criminology 36(4)*, 859-866.

Venables, W. N., & Smith, D. M. (2009). An introduction to R. Retrieved from https://wiki.math.ntnu.no/_media/drift/stud/r-intro.pdf

Vittersø, J. (2001). Personality traits and subjective well-being: Emotional stability, not extraversion, is probably the important predictor. *Personality and Individual Differences, 31(6),* 903-914. doi: https://doi.org/10.1016/S0191-8869(00)00192-6

Weisberg, S. (2005). *Applied linear regression*. Hoboken, NJ: John Wiley & Sons Inc.