A Perspective on the Use of Bayesian Methods as a Possible Solution to the Replication Crisis

Nina van Gerwen (1860852)

13/06/2022

Introduction

Bayesian Statistics is a field of statistics where the interpretation of a probability differs from the interpretation that is used in Frequentist Statistics, where a probability is seen as an event's relative frequency over time (Hájek, 2019). Instead, in Bayesian Statistics, a probability is viewed as a quantification of personal belief in an event (de Finetti, 2017).

Methods used in the field of Bayesian statistics have seen a rise in the 21st century due to more powerful software and algorithms like Markov Chain Monte Carlo (Fienberg, 2006). Furthermore, Bayesian Statistics is also often seen as a possible answer to the replication crisis (e.g., Cumming, 2014; Romero, 2019) that plagues multiple disciplinary fields - a few of which are psychology, medicine and economics. This is partly because statistical issues that come from Frequentist significance testing (e.g., p-hacking) are not present in Bayesian Statistics.

The goal of the current paper is to get more acquainted with the Bayesian framework by performing a multiple linear regression through Bayesian methods. To be specific, we will investigate the predictive ability of Social Economical Status (SES) and Verbal Ability on intelligence as measured by the Intelligence Quotient (IQ). Afterwards, we discuss the differences that we experienced by using Bayesian methods versus standard Frequentist methods for our analyses.

Hypotheses

In this paper, we aim to research whether IQ can be predicted by SES and Verbal Ability. To answer this research question, we have come up with a total of three hypotheses.

- IQ in students is predicted by both their Verbal Ability, while statistically controlling for SES, and their SES, while controlling for their Verbal Ability, where people with a higher Verbal Ability or higher SES are predicted to have higher IQ scores.

$$H_1 : \beta_1 > 0, \beta_2 > 0$$

- IQ in students is predicted only by their Verbal Ability, while statistically controlling for SES, where people with a higher Verbal Ability are predicted to have higher IQ scores.

$$H_2 : \beta_1 > 0, \beta_2 \approx 0$$

- IQ in students is predicted only by their SES, while statistically controlling for Verbal Ability, where people with a higher self-reported SES are predicted to have higher IQ scores.

$$H_3 : \beta_1 \approx 0, \beta_2 > 0$$

Method

Description of the data

The chosen dataset, obtained from the course Multivariate Statistics, contains a total of three variables from 400 high school students. Furthermore, there is no missing data. The variables are Verbal Ability, which ranges from 2.26 to 4 ($M = 3.39$, $SD = 0.38$), self-reported SES on a scale from 0 to 60 ($M = 32.78$, $SD = 9.66$) and IQ, which ranged from 85.5 to 143 ($M = 118.2$, $SD = 11.92$). When visually inspecting the histograms of the variables, we find that SES and IQ seem to follow a normal distribution. Verbal Ability, however, seems to be slightly skewed to the left which could be explained by a ceiling effect. For all analyses, we grand mean centered the variable of Verbal Ability in order to aid convergence and interpretation.

Statistical analyses

To answer our research questions and test the hypotheses, we ran two Bayesian linear regression analyses. In the first analysis, we used Gibbs Sampling with uninformative prior distributions (LR1) for the estimation of our regression coefficients. Below, the choice for our prior distributions is further explained. For the second analysis, regression coefficients were estimated through an independent Metropolis-Hastings (MH) Sampler (LR2), where the proposal distribution was a Student's t-distribution with 1 degree of freedom in order to maximize uncertainty in the tails of the distribution. The means and standard deviations of the proposal distributions were gained from the Frequentist linear regression equivalent of our analyses (i.e., a linear regression with both Verbal Ability and SES as the independent variables and IQ as the dependent variable).

Prior distributions As stated before, we used uninformative priors for LR1. There were two main reasons for this. First and foremost, the goal of the paper was to get more acquainted with the Bayesian framework. Hence, the analyses were of secondary importance and we did not wish to bias them with our expectations. Secondly, although there is historical data available on the effect of some of our independent variables on intelligence, research in this area tends to be of a more psychometric nature. For example, the effect of Verbal Ability on IQ can be found in factor analyses such as the g factor model, where general mental ability is divided in 7 subfactors of intelligence - one of which is verbal comprehension (Spearman, 1904). However, due to the different methods, we were unsure how to translate this information to the specification of our priors. Instead, we chose for uninformative priors. This means that for the intercept and regression coefficients, the prior distributions had an extremely large variance and a mean that then becomes obsolete (i.e., we simply set it to 1). Finally, for the variance coefficient, which follows a gamma distribution, the scaling and rating parameter were set close to 0.

Model diagnostics

To investigate the quality of our analyses, we did the following inspections. First, we assessed the convergence of all analyses through visual inspection of both trace- and autocorrelation plots. Second, we evaluated which of the two different sampling methods performed better. To compare the two models and choose the best one, we used the Deviance Information-Criterion (DIC). The DIC is a certain type of Information-Criterion, which are often used for model selection by balancing fit and complexity. In the DIC, complexity is calculated through estimation of the effective number of parameters instead of simply counting the number of parameters (e.g., Alkaike Information Criterion). Finally, for the chosen model, we gauged whether the residuals of our model were normally distributed through use of the posterior predictive p-value and discrepancy measures with the following original test statistic:

$$T = (\mu - m_1)^2$$

where μ is the mean and m_1 is the median of the distribution of the residuals. The reasoning behind the statistic is that the more skewed a distribution is, the larger the difference between the mean and median is. However, this difference can be either negative - in a left-skewed distribution - or positive - in a right-skewed distribution. Hence, we square the difference in order to always gain a positive value. Now, the larger the statistic is, the larger the absolute difference between mean and median, and the more skewed a distribution supposedly is.

To investigate whether our original statistic works correctly, we will compare the results to a posterior predictive p-value that uses the known statistic of skewness:

$$\tilde{\mu}_3 = \frac{\Sigma_i^N (X_i - \bar{X})^3}{(N-1) \cdot \sigma^3}$$

where $\bar{X}$ is the mean of the residuals, σ the standard deviation of the residuals and N is the total number of residuals. Skewness is known to measure the asymmetry of a distribution. Therefore, we used it to see whether the results are similar to the results of our original test statistic.

Hypothesis support

For our hypotheses, we researched which hypothesis had the most support. To do this, we compared the three hypotheses through use of the Bayes' Factor. When comparing the hypotheses through the Bayes' Factor, the priors distributions were fractional, where the size of the fraction was decided by the following formula:

$$b = \frac{J}{N}$$

where b is the denominator of the fraction, J is the number of independent constraints in the hypotheses and N is sample size. Furthermore, we conducted a sensitivity analysis with varying numbers of J to see the influence it has on the Bayes' Factor.

Results

Parameter estimates

For analysis LR1, we found an intercept of 118.32 (95% *C.I.*: [114.47; 122.21]). For the effect of Verbal Ability on IQ, we found a regression coefficient of 11.93 (95% *C.I.*: [9.11; 14.77]) and for the effect of SES on IQ, we found a regression coefficient of 0.003 (95% *C.I.*: [-0.11; 0.11]).

The intercept parameter tells us that when a student has a self-reported SES of 0 and an average Verbal Ability, he has a predicted IQ of 118. With the credible interval, we also know with 95% certainty that the true IQ of the student lies between 114.47 and 122.21. For Verbal Ability, the results mean that for every one unit increase above the mean a student scores on Verbal Ability, their predicted IQ increases on average by 11.93. Furthermore, we can state that there is a 95% probability that the true effect of Verbal Ability on IQ lies between 9.11 and 14.77. As for the effect of self-reported SES on IQ, we found that the point estimate is immensely close to 0 and there is a 95% probability that the true parameter lies between -0.11 and 0.11. These results both indicate that SES cannot predict IQ.

For our LR2 analysis, we found very similar results to the first analysis with an intercept of 115.94 (95% *C.I.*: [105.75; 129.46]) and two regression coefficients of 11.89 (95% *C.I*: [9.84; 14.09]) and 0.08 (95% *C.I.*: [-0.006; 0.17]). These values can be interpreted in the same way as the estimated values in LR1.

Model diagnostics

Convergence When investigating the convergence of the two analyses, we found the following results. For LR1, the traceplots of all coefficients showed very sound convergence. The autocorrelation plots showed two notable results. For the intercept and SES regression coefficient, the autocorrelation started at high values around .90 and then in a monotone decreasing fashion reached 0 at lag 45. The other autocorrelation plots remained constant at values around 0 for every lag. Although the autocorrelation could be improved upon, we feel it is still safe to assume the first model converged properly when also taking the traceplots into account.

For the second analysis, LR2, the traceplots for the coefficients of the independent variables showed sound convergence. The traceplot for the intercept, while still acceptable, showed that sometimes quite extreme values were accepted and convergence of this coefficient could be improved. The traceplot of the variance showed that there were issues as the plot shows only half of what you would expect from a good convergence.

This can be explained by the fact that the algorithm most likely wanted to accept negative values. However, due to a nature of a variance, this is impossible. Furthermore, a few very extreme values were accepted for variances, which might lead to issues if we were to use this model for model assumption checks through the posterior predictive p-value. In contrast to this result, the autocorrelation plots all showed constant values close to 0. This result is indicative of two things. Namely, that convergence is sound and that we chose proper proposal distributions for our regression coefficients. Hence, we still assumed that the model converged adequately.

An important limitation to keep in mind when assessing the convergence for the two analyses, is that they all consisted of only one chain. This means that it is possible that the analyses were stuck at a local maximum. However, considering that the two analyses gave similar results that both seem to agree with the results of a Frequentist linear regression, we feel safe to assume that the analyses were not stuck at local maxima.

Model comparison Comparing the two models through use of the DIC, we calculated that LR1 had a DIC value of 3061.983, whereas LR2 had a DIC value of 3168.71 ($\Delta_{DIC} = 106.73$. With these results, we can conclude that LR1, the analysis that made use of a Gibbs Sampler, performed better than the MH sampler as it has a DIC that is much lower. Hence, we only continue discussing the results that were gained from LR1. This model was also used for the calculation of the posterior predictive p-values, because of this reason and the extreme outliers in the variance coefficient that were present in the other model.

Distribution of the residuals Checking the assumption of normality, we found posterior predictive p-values of $<.001$ for both our original and skewness statistic. These p-values mean the following: for our original statistic, it entails that when using the same sampled coefficients from their posterior distribution to calculate the residuals, we find a larger squared difference between the mean and median of these residuals when used on our observed dataset compared to when it was used on a simulated dataset. For the skewness statistic, it means also means that the calculated residuals in our observed dataset were practically always more skewed than the residuals in a simulated dataset.

From these results, we can conclude two things. Firstly, our original statistic works properly to measure the skewness of a distribution as it gave the same results as the skewness statistic. Secondly, we can conclude that the residuals most likely do not follow a normal distribution and instead are skewed. This result is further strengthened by visual inspection of the residuals, which shows that the direction of the skew is to the right.

Hypothesis support

Calculating the support for each hypothesis gave the following results: H_1 had a Bayes' Factor of 1.99. H_2 had a Bayes' Factor of 10.83. H_3 had a Bayes' Factor of 0. From these results, we can infer the following. H_3, which stated that there was no effect of Verbal Ability on IQ, has practically no support. This means that the other hypotheses also have infinitely more support than H_3. In other words, we can very safely conclude that there most likely is a positive effect of Verbal Ability on IQ. Furthermore, when comparing H_1 to H_2 we discover that H_2 has 5.45 times more support than H_1. Hence, there is more evidence that SES has no effect on IQ than that SES has a positive effect on IQ. However, the sensitivity analyses of the first two hypotheses should be taken into account. Namely, although the Bayes' Factor of H_1 remains constant, the Bayes' Factor for H_2 seems to be monotone decreasing to a limit of 4 with increasing values of J. Such a result was to be expected due to the fact that H_2 uses an equality constraint. This would entail that H_2 has only around 2 - 2.5 times more support than H_1, which affects the strength of our conclusion. Nonetheless, we can still state that there most likely is no effect of SES on IQ.

Discussion

In the current paper, we have investigated whether intelligence can be predicted by Verbal Ability and self-reported SES in students using Bayesian methods. The results have shown in multiple ways the following conclusion: Verbal Ability is able to predict IQ, where a higher Verbal Ability is associated with a higher IQ, whereas self-reported SES was not able to predict IQ, as stated in H_2.

Now, let us hypothesize what the results would be if we had used Frequentist methods. A normal linear regression would have shown a non-significant effect of SES and a significant effect of Verbal Ability. Thus, we would most likely reach a similar conclusion. However, by using Bayesian methods, we were able to find support for this conclusion in multiple ways. Namely, through among other things the size of the credible intervals and the Bayes' factor. In other words, we would argue that the Bayesian methods allowed for much more flexibility than their Frequentist analog for the current paper. This flexibility extended to multiple facets in the analyses. For example, the analyses could have been done with either Gibbs Sampling (using either un- or informative priors) or with MH sampler (which can also be done in multiple ways). All these choices can slightly alter your results and conclusions. However, they also allow you to finetune your analysis in order to fit your research in the best way possible compared to running a simple Frequentist linear regression.

Besides flexibility for gaining evidence for our conclusion, the Bayesian methods also differed in how we ended up forming the conclusion. The Bayesian framework strafes away from the dichotomy of significance testing where there either is a significant effect or there is not. Instead, the Bayesian framework tells us about the relative support every hypothesis had and what would most likely be true in the population, given the data.

Reflecting back to the Introduction, where it was stated that Bayesian Statistics is sometimes seen as a possible answer to the replication crisis. After performing a few Bayesian analyses ourselves, we are unsure whether we agree with this statement. This is partly due to the fact that, although issues such as p-hacking are resolved, other issues could arise. An example would be choosing a posterior predictive p-value after your analyses that gives you the result you are looking for. In both cases, the real answer to the replication crisis would be open science and pre-registration, not the exact field of statistics you use. Nonetheless, by programming multiple Bayesian methods ourselves (e.g., Gibbs Sampler and Bayes' Factor), we did come to realize the amount of theory that is behind all these methods that we had to understand in order to be able to use them and answer a frankly quite simple research question. We believe that if everybody were to dive this deep into the methods required, science as a whole could benefit. Especially when compared to the current situation in Frequentist Statistics, where a linear regression requires only a single line of code with a bare minimum amount of understanding. To conclude, we believe more emphasis should be placed on not only understanding the methods used for analyses, but also ethical science practices such as pre-registration and open science.

References

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29.

De Finetti, B. (2017). *Theory of probability: A critical introductory treatment* (Vol. 6). John Wiley & Sons.

Fienberg, S. E. (2006). When did bayesian inference become "bayesian"? *Bayesian Analysis*, *1*(1). https://doi.org/10.1214/06-ba101

Hájek, A. (2019). Interpretations of Probability. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2019). https://plato.stanford.edu/archives/fall2019/entries/probability-interpret/; Metaphysics Research Lab, Stanford University.

Romero, F. (2019). Philosophy of science and the replicability crisis. *Philosophy Compass*, *14*(11), e12633.

Spearman, C. (1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201–292. http://www.jstor.org/stable/1412107