Final Assignment - Bayesian Statistics

Nina van Gerwen (1860852)

15/06/2022

Introduction

Bayesian Statistics is a field of statistics that is based on another interpretation of probabilities. In Bayesian Statistics, a probability can be viewed as a quantification of personal belief in an event (de Finetti, Bruno (2017). Theory of Probability: A critical introductory treatment. Chichester: John Wiley & Sons Ltd.).

Methods used in the field of Bayesian statistics have seen a rise 21st century due to more powerful software and algorithms like Markov Chain Monte Carlo (Fienberg, Stephen E. (2006). "When Did Bayesian Inference Become"Bayesian"?". Bayesian Analysis. 1 (1): 1–40.). However, Bayesian Statistics is also often seen as a possible answer to the replication crisis 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 part of Bayesian statistics.

The goal of the current paper is to get more acquainted with Bayesian methods by performing a multiple linear regression through Bayesian methods. To be more specific, we will investigate the predictive ability of Social Economical Status (SES) and Verbal Ability on intelligence as measured by the Intelligence Quotient (IQ) in a Bayesian setting.

Hypotheses

In this paper, we will 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 four 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

For our dataset, we chose one gained from Multivariate Statistics. The dataset contains a total of three variables from 400 high school students and contains 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 will run two Bayesian linear regression analyses. In the first analyses, we will use Gibbs Sampling with uninformative prior distributions (LR1) for the estimation of our regression coefficients. Below, the choice for prior distributions will be further explained. For the second analysis, regression coefficients will be estimated through an independent Metropolis-Hastings (MH) Sampler (LR2), where the proposal distribution is a Student's t-distribution with 1 degree of freedom in order to maximize the uncertainty in the tails.

Prior distributions – iets meer informatie over waarom uninformative priors.

As stated above, we will use uninformative priors for the Gibbs Sampler. Although there is historical data on the effect of verbal comprehension on intelligence, research in this area tends to be of a more psychometric nature (e.g., the g factor model, where general mental ability is supposedly divided in 7 subfactors of intelligence - one of which is verbal comprehension (Spearman, C.E. (1904). General intelligence', Objectively Determined And Measured. American Journal of Psychology. 15 (2): 201–293)). Due to the different methods, we were unsure how to use this information in the specification of our priors and instead chose for uninformative priors. This means that for the intercept and regression coefficients, the prior distributions will have an extremely large variance and a mean that then becomes obsolete. Hence, we will simply set the mean to 1. Finally, for the variance coefficient, which follows a gamma distribution, the scaling and rating parameter will be set close to 0.

Model diagnostics

First, we will assess the convergence of all analyses through visual inspection of both traceplots and autocorrelation plots.

Second, we will assess which of the two different sampling methods performed better. To compare the two models and choose the best model, we will use the Deviance Information-Criterion (DIC).

Finally, for the chosen model, we will assess whether the residuals of our model are 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 will be. 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, we will compare the results to a posterior predictive p-value with 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 and σ the standard deviation of the residuals. Skewness is known to measure the asymmetry of a distribution. Therefore, we will use it to see whether the results are similar to the results of our original test statistic.

Hypothesis support

Besides answering the hypotheses, we also want to see which hypothesis has the most support. To do this, we will compare the three hypotheses (and their corresponding statistical models) through the use of both the Bayes' Factor. When comparing the hypotheses through the Bayes' Factor, the priors distributions will be fractional, where the size of the fraction is 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 hypothesis and N is sample size. Furthermore, we will conduct a sensitivity analysis with varying numbers of J to see the influence it has on the Bayes' Factor.

Results

Parameter estimates

For our LR1 analysis, 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]).

– interpretation –

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]).

– interpretation –

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 regression coefficient for SES, 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 traceplot 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,

however, showed that there might have been some issues as the plot shows that somehow it wanted to accept negative values.

REWRITE THIS PART: Furthermore, some very extreme values were accepted, which might lead to issues if we were to use this model for model assumption checks through the posterior predictive p-value. The autocorrelation plots all showed constant values close to 0. Hence, taken together, we assume that the model converged adequately. However, due to the extreme values, we refrained from using this model to check the normality of the residuals through use of the posterior predictive p-value.

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 are stuck at a local maximum. However, considering that the analyses gave similar results that 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, from now on we will only continue discussing the results that were gained from this statistical 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 difference between the mean and median of these residuals when using our observed dataset compared to using 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, for which the value of J matter more. 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 the xxx using Bayesian methods.

- answer of the research question
- Differences between Bayesian and Frequentist analyses (relevant to the context of the research questions –> praten over hoe het programmeren van een simpele regressie mij doet beseffen hoeveel theorie en wetenschap er eigenlijk achter zit (ook al word het nu overal gebruikt met een simpele line of code / 3 klikjes) –> bayesian kwam op als een soort antwoord op de replicatie crisis in meerdere sectoren, maar wellicht zit het antwoord niet per se in welke statistiek je doet maar in het goed gebruiken ervan (zie hierboven)