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 (source). In the current paper, we will investigate the predictive ability of Social Economical Status (SES) and Verbal Ability (VA) on intelligence as measured by the Intelligence Quotient (IQ) through Bayesian methods and statistics.

Research Questions

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

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 SES will 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 will 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 will have higher predicted IQ scores.

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

- The effect of Verbal Ability on IQ depends on self-reported SES, where people who have a higher self-reported SES will have a greater effect of Verbal Ability on IQ than people with lower self-reported SES scores.

$$H_3 : \beta_1, \beta_2, \beta_3 > 0$$

Method

Description of the data

For my dataset, I 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. Furthermore, 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 hypotheses, we will run a total of three Bayesian linear regression analyses. In the first two analyses, we will use Gibbs Sampling with both an uninformative and informative prior for the estimation of our regression coefficients. For the final analysis, regression coefficients will be estimated through an independent Metropolis-Hastings (MH) Sampler, 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

IQ (b1): known to be a variable that has a mean of 100, sd of 15. SES (b2): check categorical distribution in society and try to mimic that in the self-report way variance (vari): intercept??

- nvm, need priors for mu/sigma.. not their distribution (LEL)

Model Diagnostics

First, we will try to assess whether the different methods of sampling gave different results through visual inspection. If they do not differ, we will only continue with the results gained from Gibbs Samplers (reasons).

Second, we will assess convergence of the three analyses through visual inspection of both traceplots and autocorrelation plots. Because the analyses consist of only one chain, we can not use the Gelman-Rubin statistic.

Finally, we will assess whether the residuals of our model are normally distributed through use of the posterior predictive p-value and discrepancy measures with two different test statistics.

The first test statistic is an original statistic with the following formula:

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

where μ is the mean and m_1 is the median of the distribution of the residuals. (reasons why this gives proof).

Furthermore, we will also use 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 asymmetry of a distribution. Therefore, we will use it to see whether the results are similar to the results of our original test statistic.

Model comparison

Besides answering the hypotheses, we also want to see which hypothesis has the most support. To do this, we will compare the hypotheses (and their corresponding statistical models) through the use of the Bayes factor. When comparing the hypotheses, 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 fraction, J is the number of independent constraints in the hypothesis and N is sample size. However, we will conduct a sensitivity analysis with varying numbers of J to see the influence of the fraction.

Results

Model diagnostics

Parameter estimates

Here go the parameter estimates, credible intervals and interpretation

Model comparison

Discussion

- Differences between Bayesian and Frequentist analyses (relevant to the context of the research questions)