Sofía Villalobos Aliste - 6060714
Svillalobos.a@hotmail.com
Bayesian Statistics- Assignment

RESEARCH QUESTION

This project aims to examine if Political Self-Positioning in the Netherlands can be predicted from variables concerning Institutional Trust: Trust in the country's parliament (X1), Trust in the legal system (X2), and Trust in Police (X3). This would be addressed by evaluating the following model:

$$Y= \beta0+ \beta1*X1_i+ \beta2*X2_i+ \beta3*X3_i+ \varepsilon$$

DATA

The data for this research had been obtained from the European Social Survey (ESS) 9th round, collected in 2018 in the Netherlands.

The dependent variable that will be used as political self-positioning is "Placement on a left-right scale" and consists of 10 categories from 0 (Left) to 10 (Right). Independent variables Trust in the country's parliament (prl), Trust in the legal system (lgl), and Trust in the police (plc) are categorical variables that go from 0 (No trust at all) to 10 (Complete trust). All these variables will be here treated as continuous.

Two models of multiple linear regression are proposed here that differ in the information used: For the first model we used data from the 8th round European Social Survey conducted in the Netherlands in 2016 as prior information, while the second model was constructed with uninformative priors instead.

The total cases for this survey used in the analysis are 1524 after using listwise deletion.

ESTIMATION

The estimation process was made through Gibbs Sampling, for $\beta0$, $\beta2$, $\beta3$, and σ^2, and with Metropolis Hasting sampling for the estimation of $\beta1$.

The Gibbs sampler consists of specifying a density for the data that describes our model and derives conditional posterior distributions. First, we define the values that will serve as prior information by using the coefficients and standard errors of the model regression with the same variables from the previous round of the ESS. We assigned them a small variance to make it informative. For the uninformative priors, these are constructed with large variances. Then, we define the initial values for each of the two chains that we will be using to sample, and the number of iterations (1000000 for each chain).

The conditional posterior distribution of β0 was obtained by derivating the initial values to obtain the conditional posterior distributions for the mean and variance and then sampling a value from that conditional posterior distribution of β0, and we keep this value for the next iteration and do it again.

β1 was obtained through a Metropolis Hasting step. This is required when we can't find our normalizing constant, but we can compute the function which is proportional to the density. Here the values are drawn iteratively from a proposal distribution, and the values are kept or rejected based on the proportional conditional posteriors. And the acceptance ratio is computed. this is done as a correction in case of obtaining too improbable values. Then, if the proposal value of β1 is rejected for being larger than the ratio, the last value of β1 is retained until there is a proposed value of β1 that is equal to or smaller than this ratio.

Finally, we sample the posterior distribution for the residual variances. Using the sampled values of β0, β1, β2, and β3 to compute the condition posterior of σ^2 by sampling posterior shape parameter alfa and posterior scale parameter beta from where we can compute the sum of squared residuals.

CONVERGENCE

To evaluate the convergence of the chains, obtain with the sampler we look at trace plots, autocorrelation plots, and Montecarlo error.

To evaluate the convergence of these iterations, we first check at trace plots considering the total number of iterations and deleted the first ones where we can see that the chains were far to overlap. The first 900000 iterations were deleted for each chain on each model, resulting in the trace plots where convergence can be observed as a first check in the R code, and that indicates that the chains ended up covering the same area.

Autocorrelation plots indicate that the convergence of the chains will be very slow for B1, which might be because we are using Metropolis Hasting for the estimation of this coefficient, which involves a different process since keeping a different value from the previous one relies on the acceptance ratio. For coefficients B2, this process is less slow, and for B3 and the variance, the convergence will occur very fast. In the case of slow convergence, this is not a problem since we use a large number of iterations.

Tables 2 and 3 show the values of the Monte Carlo error, which is the standard deviation of the estimation of each parameter divided by the squared root of the number of iterations, and it should not be larger than the 5% of the standard deviation of the samples. In these cases, all the MC errors for both models are smaller than that proportion, as observed in the tables, which is expected because MC error decreases as the number of iterations increases.

Table 1

	SD	MC error
B0	0.1676	0.0003
B1	0.0339	0.0000
B2	0.0308	0.0000
B3	0.0089	0.0000
σ^2	0.1386	0.0003

Table 2

	SD	MC error
B0	0.1676	0.0003
B1	0.0339	0.0000
B2	0.0308	0.0000
B3	0.0089	0.0000
σ^2	0.1386	0.0003

We conclude by looking at the evidence that the chains have converged, even though we can never be completely sure.

POSTERIOR PREDICTIVE CHECK

To check the assumption of normality of residuals we first establish the null hypotheses as our proposed linear regression model:

$$y_i = \beta_0 + \beta_1 * X1_i + \beta_2 * X2_i + \beta_3 * X3_i + e_i \quad \text{with } e_i \sim N(\mu, \sigma^2)$$

Second, we sample the posterior distribution of the model parameter, this is for each sampled parameter $\Theta = [\beta_0, \beta_1, \beta_2, \beta_3, \sigma^2]$ computed in every 200000 iterations of our Gibbs sampler, and then we modeled Y as a regression with fixed effect of X1, X2, and X3. So, we end up with a matrix of 304800000 columns and 1524 rows, where every box reflects and simulated outcome per subject of the observed data using a different set of estimates for every column.

Now we compute the residuals for the observed data: This is, we subtract to the observed Y of every subject the estimate predicted with the 304800000 sets of Θ. To compute the residuals for the simulated data, we subtract the same estimated with the 304800000 Θ to the modeled Y that we previously obtained by simulating the outcomes maintaining X1, X2, and X3 fixed. Then, we compute the skewness in the absolute value = |3 * (Mean – Median) / Standard Deviation|, to avoid problems after when calculating the Bayesian p-value for example having a skewness coefficient of -.5 is considered as less skew than a coefficient of 1. Finally, we count how many times the skewness on the simulated data is larger than the skewness in the observed data, and we store the results in a matrix of 304800000 columns with 1 row where every time we will put a 0 if the coefficient of the observed data is larger than the simulated, and a 1 if it is the other way around. This procedure is computed in the same way for both models.

For model 1, there were 191964 occasions where skewness was larger for simulated data, leading to a bayesian p-value of 0.96. For model 2, there were 191741 occasions where skewness was larger for simulated data, leading to a p-value of 0.96. This shows that for both models, the assumption of normality of residuals is met.

In comparison with a frequentist approach, where can be observed the different meanings for the concept of the p-value with the posterior predictive p-value. In the first case, we use a

uniformly distributed and unbiased p-value, where we aim to reach the significance threshold at the alfa level of 0.05, which will be the probability of our data in the case the null hypothesis is true. And for the bayesian p-value, we test our specified model as the null hypothesis, and then we quantify with a t-statistic the proportion of times that our estimated parameters differ from what our data should look like if it comes from the model we specify. We compare this value with a threshold of 0.5, even though this comparison is flexible and does not determine the analysis, we still wish our bayesian p-value to be close to 0.5 so we can consider that there is evidence that our data comes from the model we specified.

RESULTS AND INTERPRETATION

Tables 3 and 4 show the estimated results for each model, this is the average value across the samples.

Table 1

	Mean	SD	Q025	Q975
B0	5.690	0.168	5.353	6.021
B1	0.0861	0.034	0.019	0.152
B2	-0.161	0.031	-0.222	-0.101
B3	0.003	0.009	-0.014	0.020
σ^2	3.818	0.139	3.556	4.100

Table 4

	Mean	SD	Q025	Q975
B0	5.733	0.198	5.377	6.118
B1	0.083	0.036	0.016	0.158
B2	-0.165	0.035	-0.234	-0.097
B3	0.003	0.010	-0.016	0.022
σ^2	3.819	0.139	3.556	4.101

For B0 average results are positive, with a mean of 5.764 and CI [5.3-6.1] in the first model and a mean of 5.683 and CI [5.3-6-1] in the second model. Since this coefficient represents the intercept, we don't pay much attention to it.

For B1 average results are positive, with a mean of 0.086 and CI [0.019-0.152] in the first model and a mean of 0.083 and CI [0.016-0.15] in the second model. In both cases, the credible interval does not include the 0 and that indicates that there is evidence to consider that the effect of these predictors is different from 0. So, as the trust of a person in the country's parliament increases, it will also increase their value on the political self-positioning scale, and they will be more on the right-wing of the scale.

For B2 average results are negative, with a mean of -0.16 and CI [-0.22 - -0.10] in the first model and a mean of -0.165 and CI [-0.23- -0.097] in the second model. In this case, the means indicate that as the trust in the legal system increases, your position in the political self-positioning scale will also decrease, leading you to be more left-wing. The credible intervals also don't include 0, so that indicates that there is evidence that this predictor has an effect different from 0.

For B3 average results are positive, with a mean of 0.0003 and CI [-0.014- 0.020] in the first model and a mean of 0.003 and CI [-0.016-0.022] in the second model. This is interpreted as the trust in the police increases, it will also increase your value in the political self-positioning, leading you to be more right-wing. However, the credible interval of the coefficient for each

model includes the 0, so there is not enough evidence to think that the trust in the police affects different from 0 on predicting the variable political self-positioning.

In comparison with a frequentist approach, a big difference is the use of the credible interval, in the case of the Bayesian approach, versus the confidence interval used in the frequentist approach. The confidence interval is computed as an interval of values that will include the true value of the population in a 95% of the cases where the interval is computed, and that is why they use the term *in repeated samples*, and each of these samples, the value will be or not be included inside the interval. Whereas for the Bayesian approach a credible interval is computed based on a posterior distribution, and a probability can be assigned to the values inside this interval that will indicate how likely is a certain value to occur in the population, and the most probable values that can be true can be observed in this interval.

MODEL SELECTION USING THE DIC

The deviation information criterion (DIC) is a model selection tool that can be considered as an extension of the AIC, but with the difference that the parameters cannot be counted but have to be estimated due to the complexity of the model and the fact of use prior information. The DIC evaluates the likelihood evaluated at the posterior mean of the parameters, and adds the difference between the average likelihood over the posterior distribution of the parameters and the likelihood evaluated at the posterior mean of the estimated parameters, which represents the model complexity as an estimate of the number of parameters,

Information criteria (IC) in general can be advantageous in comparison to the frequentist p-value for example because the latest has a uniform distribution, so they are all equally probable under the null hypothesis. It allows us to test a model vs a no-model. IC is a misfit and complexity measure that allows comparison models to each other. This is relevant for hypotheses evaluation because sometimes the test of effect or not effect is not so interesting, but we are more interested in testing hypotheses about which kind of effect has more evidence than the other and in testing them against each other. In comparison to frequentist and bayesian p-value, both allow testing one model against other models at same the time, while with IC (AIC, DIC, etc.) we can compare more than two models while are not nested, and it is a broader set of tools to pick one IC that allows us to test our ideas of what a good model should achieve.

Here we compare one model estimated with prior information, and one with uninformative priors. DIC for the model with informative priors is 12901,08, whereas for the model with uninformative priors is 12901,65. Since both are equivalent, we will stay with the one with prior information for the next analysis.

MODEL SELECTION USING BAYES FACTOR

The Bayes factor (BF) quantifies the support in the data for a set of tested hypotheses. This is a big difference and an advantage compared to the frequentist approach for testing hypotheses due to: many hypotheses can be tested at the same time, the direction of it can be

also tested, and it can be updated with new information every time by computing the BF again and including the new relevant data available. The BF shows which of the tested hypothesis has more support, but this can also become a drawback because even if we are testing a poor hypothesis we will obtain as a result that one of them is the more "correct".

After selecting the model with prior information, 10 different hypotheses were tested concerning each of the predictors being larger or smaller than 0.

Table 5

Hypotheses	BF.u	BF.c	PMPa	PMPb
$H_0 = \beta1 = 0, \beta2 = 0, \beta3 = 0$	*0.005*	*0.005*	*0.001*	*0.001*
$H_1 = \beta1 > 0, \beta2 > 0, \beta3 > 0$	*0.000*	*0.000*	*0.000*	*0.000*
$H_2 = \beta1 > 0, \beta2 > 0, \beta3 < 0$	*0.000*	*0.000*	*0.000*	*0.000*
$H_3 = \beta1 > 0, \beta2 < 0, \beta3 > 0$	*5.248*	*102.884*	*0.929*	*0.790*
$H_4 = \beta1 > 0, \beta2 < 0, \beta3 < 0$	*0.068*	*0.058*	*0.012*	*0.010*
$H_5 = \beta1 < 0, \beta2 < 0, \beta3 < 0$	*0.002*	*0.002*	*0.000*	*0.000*
$H_6 = \beta1 < 0, \beta2 < 0, \beta3 > 0$	*0.260*	*0.235*	*0.046*	*0.039*
$H_7 = \beta1 < 0, \beta2 > 0, \beta3 > 0$	*0.000*	*0.000*	*0.000*	*0.000*
$H_8 = \beta1 > 0, \beta2 < 0, \beta3 < 0$	*0.067*	*0.057*	*0.012*	*0.010*
$H_9 = \beta1 < 0, \beta2 > 0, \beta3 < 0$	*0.000*	*0.000*	*0.000*	*0.000*

These hypotheses were tested using the package Bain in R, trying different seeds to ensure the stability of the results. From it, we can observe that according to the column BF.u that indicates the Bayes factor as a ratio of fit and complexity, the most strong hypothesis to be supported among the ones tested is the third one.

This hypothesis indicates that among all the tested hypotheses, it is more *likely* that to predict the outcome variable Y- Political Self-positioning, the variable Trust in the country's parliament (B1) would have a positive association with Y, Trust in the legal (B2) system would have a negative one, and Trust in Police (B3) would have a positive one. This is, the more a subject trust in parliament and police, would be placed more to the right on the scale of political self-positioning. And works in the opposite way for trust in the legal system, for each point of trust on it that a person has, it would be placed more to the left.