Manuscript for Markup Language course - Using bayesian statistics to prove a point

Inan Bostanci

1/15/2022

1 Introduction

In this report, I want to inspect whether protein-consumption is a better predictor of the average height of a country than the general wealth of that country. A few weeks ago, I had a discussion with a friend about determinants of height. I argued that the average height in most western countries is larger than in other countries, mostly because western countries consume the largest amount of protein, mostly in the form of meat. My friend argued that western countries are wealthier and therefore have better overall circumstances (i.e. health, lifestyle) which, together with genetics, are the biggest contributors and that protein plays little to no role. I do believe that genetics and overall circumstances have some effect, but I believe that protein is the more important determinant. After all, animals kept in mass stocks, which live under horrible circumstances but get fed tons of soy seem to grow well and fast for profit-oriented industries. Because this assignment came up soon after our discussion, I figured that it would be a good idea to be the biggest know-it-all of all times by reminding my friend about our discussion months later and showing him evidence for my claim.
To answer the research question, I will compare several linear models by means of the Deviance information criterion (DIC) and Bayes Factor. Parameter estimates will be acquired in a Bayesian way, using a Marov Chain Monte Carlo (MCMC) method, namely Gibbs sampling and Metropolis-Hastings (MH) algorithms. The question will be answered by both the parameter estimates and the model comparisons. This report is outlined as follows: First, I will present the dataset and report descriptive statistics. Next, I will elaborate on the method and explain the algorithm. I will then assess the convergence of the model, test an assumption of the model by means of a posterior predictive p-value (PPP) and report the posterior means, standard deviations (SD) and the 95%-central credible interval (CCI). After interpreting the estimates, I will create two competing models and compare all models by means of the DIC and Bayes Factor. Because of the less known process of Bayesian statistics, I will also compare this method to a Frequentist approach. Finally, I will make a conclusion on the research question.

2 Hypotheses

H_1: Protein-consumption has a larger effect-size than wealth.
This will be tested by comparing the parameter estimates in a standardized linear regression model containing both parameters.
H_2: A model that contains only protein has a better model fit than a model that contains only GDP.
It can be well-argued that protein-consumption and wealth themselves are linearly related. Because protein is an expensive food, citizens in wealthier countries are more likely to be able to afford protein. However, if protein was just a noisy proxy for wealth, then a model with wealth only should fit better than a model with protein only. If both models fit equally well, it indicates that they are perfect substitutes. I expect that a model with protein as the only predictor has a better model fit than a model with wealth as the only predictor.

3 Data

The data was drawn from several datasets on the website Our World in Data (Our World in Data 2021a, Our World in Data 2021b) and merged using the country and year. The data contains information on each country for GDP per capita (in dollars, adjusted for purchasing power), mean male height (in cm) and the mean of consumed calories from animal protein. Observations were made for each year from 1861 to 2019 although the majority of years have missing values. Because the year 1996 contained the lowest number of missing values, it was chosen for this study. This dataset includes 171 countries. After removing rows with missing observations, the final sample size is 158. Mean male height ranges from 159.9 to 182.5 with a mean of 171.3. GDP ranges from 262.5 to 101754.3 with a mean of 12676.6, median of 6806.4. The mean number of calories from protein consumed per country ranges from 15.84 to 305.60 with a mean of 127.37.

4 Method

4.1 Acquiring the model

The first hypothesis will be inspected by means of the linear regression model

$$MeanHeight_i = \beta_0 + \beta_1 * Protein_i + \beta_2 * GDP_i + \varepsilon_i,$$

where $\varepsilon \sim \mathsf{N}(0, \sigma^2)$.
Model estimates will be acquired using an MCMC-method that combines the Gibbs sampling algorithm and the MH-algorithm. The Gibbs algorithm iteratively samples a parameter from a conditional posterior distribution of said parameter given the sampling distribution. The MH algorithm does not sample from a conditional distribution. In short, it iteratively samples from a proposal distribution that loosely approximates the conditional posterior distribution and compares the density of the conditional posterior distribution at the point of the sampled value to the density of the conditional posterior distribution at the point of the previously sampled value, while taking the likelihood of drawing that sample given the proposal distribution into account. The MH algorithm is applied for β_1 (the coefficient for protein), while the Gibbs sampling algorithm is applied each for β_0 (the coefficient of the intercept), β_2 (the coefficient for GDP) and the σ^2 (the residual error variance).
Because at the point of doing this study I only have an intuition of the direction of the effects and their proportionality but no actual historical data or grounded knowledge, the prior probabilities of the coefficients will be uninformative. Therefore, they will each be normally distributed with a mean of zero and a large variance of 1000:

$$b_0 \sim \mathsf{N}(0, 1000)$$
$$b_j \sim \mathsf{N}(0, 1000)$$

Since the outcome variable is also normally distributed, the priors are conjugate, and the conditional posterior probabilities are also normally distributed.
My prior the distribution of the residual error variance is proportional to an inverse-gamma-distribution with shape and scale parameters each approximating 0:

$$p(s^2) \propto \frac{1}{s^2} \approx IG(0.001, 0.001)$$

To sample one estimate from a conditional distribution, the parameters of that conditional distribution have to be known. Therefore, these MCMC methods require starting values. I will specify starting values for b_1, b_2 and s^2 and use these to sample the first value for b0, which will then be used for the first sampled value of b_1, etc. To rule out that the starting value influences the estimated parameters, the sampling procedure will be run twice with equal amounts of iterations, each run being a separate chain. The starting values will be considered as non-influential if the chains converge. Convergence will be assessed by means of a visual inspection, a comparison of the mean and standard deviation of each chain, autocorrelation and the Monte Carlo error.

4.3 Regression assumptions

One assumption of a linear regression model is homoscedasticity, i.e. that the variance of the outcome variable is independent of the predictors. A violation of this assumption can be pictured as a heteroscedasticity in the residuals. I will test homoscedasticity by comparing the correlation between the residuals and the predictor variables with a PPP. First, I will compute the residuals of each set of sampled parameters on the observed dataset. In an iterative procedure, I will then simulate a dataset with each set of sampled parameters. Next, I will correlate the fitted values with the residuals of each simulated dataset and of the observed dataset, respectively. This correlation will be the discrepancy measure of the test. If there is no clear trend about the proportion of the correlation with the observed vs. with the simulated data, the residuals of the observed data seem to behave as would be expected with the proposed model (represented by a PPP-value close to 0.5). If the correlation appears to be larger in the observed data than in the simulated data with most of the sets of parameters, the variance of the outcome variable seems to be more dependent on the predictors than would be expected with the proposed model (represented by a PPP-value below 0.5).

I will also test the assumption of the absence of outliers with a PPP. For this, I will use the same simulated datasets to compute the difference between the largest and the smallest value as the test-statistic. If the expected difference is smaller in the simulated data, this indicates that the observed data is spread wider than expected by the proposed model. If this is the case, I will perform two additional tests using the difference between the mean and the largest and lowest value, respectively, to inspect in which direction the outliers are.

4.4 Testing the model parameters

To test if the parameters are meaningful, the Bayes Factor (BF) will be used for the hypothesis H_1 of both coefficients being equal to zero vs. the complementary hypothesis H_u that the coefficients are not equal to zero ($H_1 : \beta_1 = 0, \beta_2 = 0; H_u : \beta_1, \beta_2$). The BF can be interpreted in the following way: A BF of 0.5 indicates that there is twice as much support for H_u than for H_1, while a BF of 2 indicates that there is twice as much support for H_1 than for H_u.

4.5 Testing the hypotheses

Finally, I will test my claims about the effect of protein and GDP on height in several ways. First, I will test the hypothesis that the effect of protein is larger than the effect of GDP using the BF with standardized coefficients (denoted by the subscript $stz; H_2 : \beta_{1stz} > \beta_{2stz}; H_u : \beta_{1stz} \leq \beta_{2stz}$). Next, I will use the MCMC-method to model the data with a one-predictor-model using just the protein or GDP-variable. I will then compare all three models by means of the DIC, which is a function of the model's fit and complexity. I expect the DIC of the model containing protein only to be lower than the DIC of the model containing GDP only. Furthermore, I expect the model containing protein only to have about the same DIC as the model containing both parameters. This is because I expect the fit to be about as good, since I believe that protein is by far the stronger predictor, and the complexity to be just minorly different, since the full model only has one added parameter given a large sample size.

5 Frequentist methods vs. Bayesian methods

This study does not use classical statistical methods to answer the research question and instead draws upon Bayesian methods. This might surprise my friend and I will therefore highlight the advantages for this specific research question.

For once, it uses a PPP to test a model assumption. This procedure allows for a much more interpretable inspection of how data should behave if the proposed model is true, which is crucial to understand if the proposed model is actually applicable to the scenario that was observed.

It also uses the BF instead of a hypothesis significance test. The largest advantage of this is that the BF

allows to test more differential hypotheses in a convenient way. In this instance, it allows to test if one predictor has a larger effect than another, which would not have been doable with the classical t-test that is being used for the majority of linear regressions. Therefore, it can become much clearer who of us is right about predictors for height, and I can be an even bigger know-it-all than with simple significance tests.
A second advantage of the Bayesian approach lies in the sampling procedure. When coefficients are very close to zero, researchers often look at confidence intervals and inspect weather zero falls within the interval. However, confidence intervals are not very meaningful in single studies. The applied sampling procedure allows for the use of a CCI and a distribution of the parameters, which does give an actual indication of how likely the value zero is for a given coefficient. Although I do not have a precise idea of how the parameters will look, the range of human height is much smaller than the range of calories consumed and of GDP, which might lead to such small coefficients close to zero. Here, a look at the CCIs and histograms can be very informative.

6 Results

Figure 1 depicts the trace plots of the parameters for each chain after the burn-in period. Visually, both chains seem to have an equal width, which indicates that the sampled values have the same range and mean, and thus the algorithm seems to follow the same trend despite different starting values. For the case of the first beta-coefficient (β_1), which was sampled using the MH-algorithm, fluctuations point at a large autocorrelation.
Figure 2 shows autocorrelation plots for each parameter. Except for the case of the residual error variance, the sampling procedure seems to suffer from a high autocorrelation and therefore requires a large number of samples to cover the entire posterior distribution. This is also reflected by the relatively low acceptance rate of 21.1% for the MH-step in the sampling algorithm. However, the 50.000 iterations that were set for this study should be enough to cover the posterior distribution.

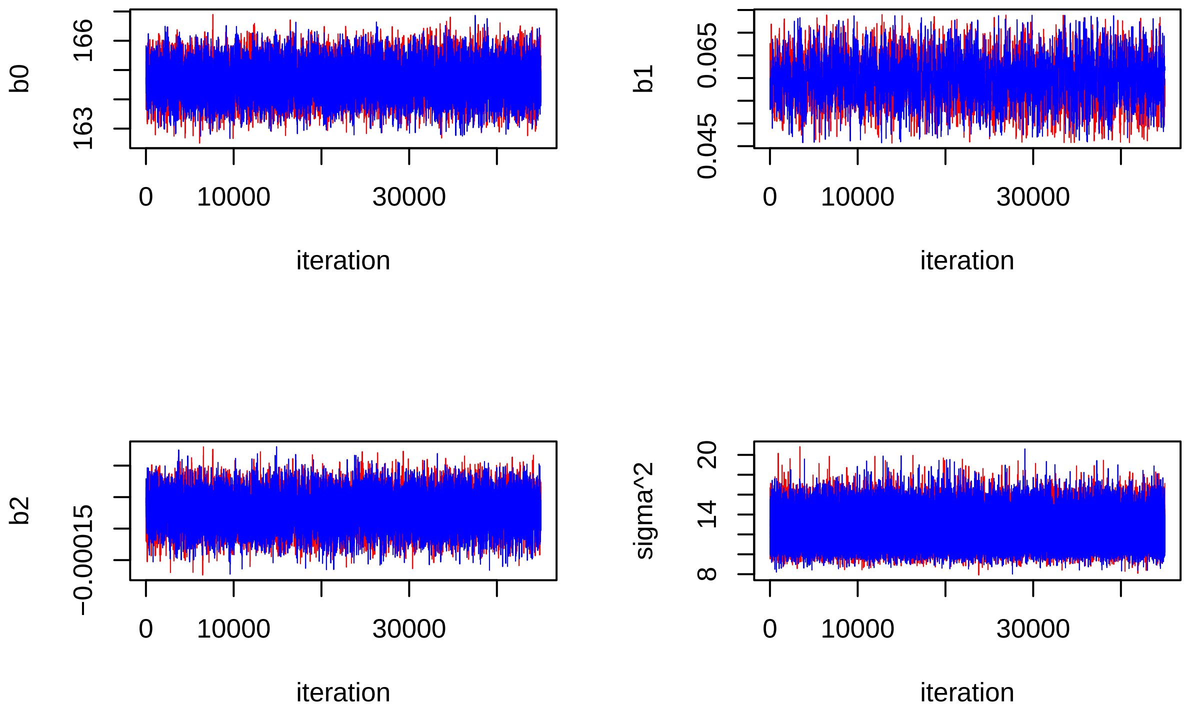
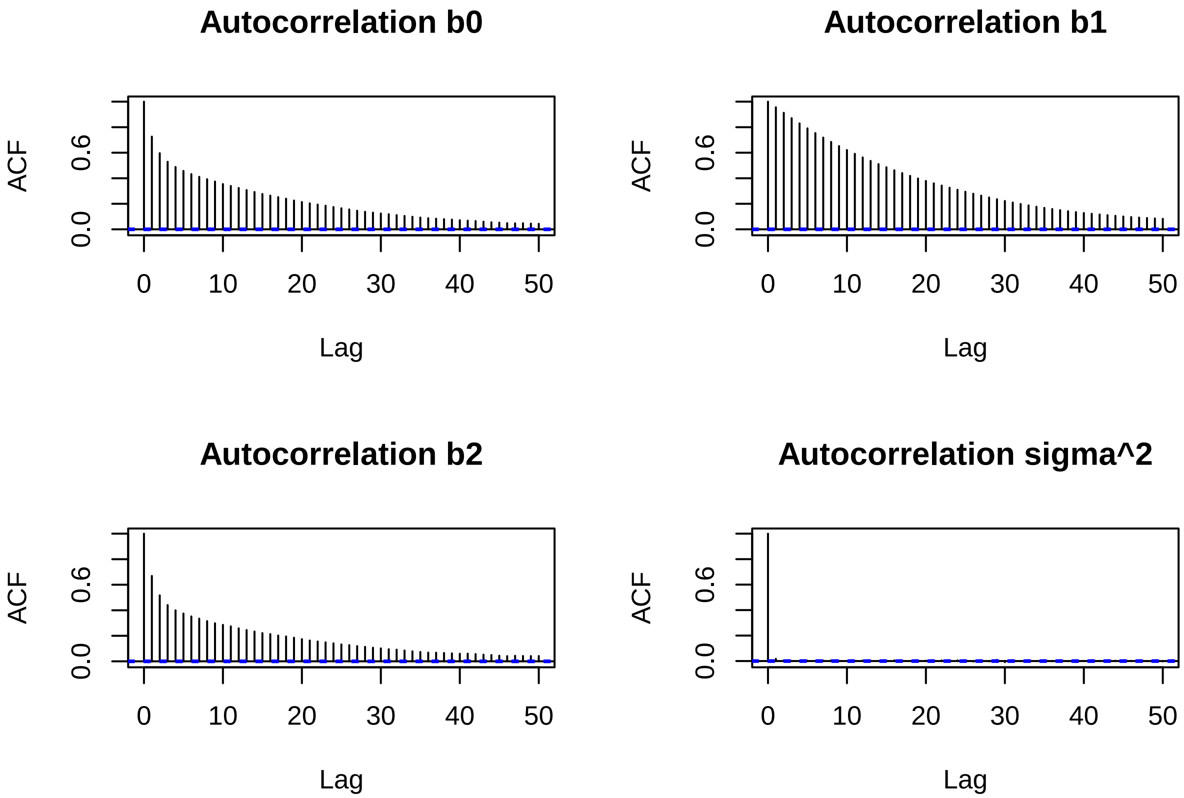


Figure 1: Figure 1: Trace plots of both chains for each parameter.

Table 1 shows the sample statistics of the pooled chain for each parameter, as well as the MC-error. The latter further does not speak against convergence, because each MC-error is much smaller than its corresponding



Autocorrelation b0

Autocorrelation b1

Autocorrelation b2

Autocorrelation sigma^2

Figure 2: Figure 2: Autocorrelation plots for the pooled sampled parameters.

standard deviation. Because no diagnostic criterion speaks against convergence, I will assume that the sampler converged. The posterior mean of the intercept-coefficient is at 164.66 (SD = 0.541, 95%-CCI = [163.606; 165.721]), while the coefficient for protein has a mean of 0.0596 (SD = 0.00478, 95%-CCI = [0.05; 0.0691]) and the coefficient for GDP has a mean of -0.000072 (SD = 0.000024, 95%-CCI = [-0.00012; -0.000025]). respectively. The mean of the residual error variance is 12.533 (SD = 1.439, 95%-CCI = [10.028; 15.65]).

Table 1: Table 1: Sample statistics for the posterior distributions of the parameters. 2.5% and 97.5% represent quantiles to obtain the 95%-CCI.

Pooled chain	Mean	SD	2.5%	97.5%	MC Error
b0	164.658876	0.541385	163.605810	165.720534	0.001712
b1	0.059568	0.004779	0.050359	0.069109	0.000015
b2	-0.000072	0.000024	-0.000120	-0.000025	0.000000
sigma2	12.533275	1.439044	10.027727	15.649456	0.004551

The PPP-value for the assumption of homoscedasticity is 0.448, which means that the residuals are slightly less homoscedastic in the observed data than in the simulated data. However, given that this value is very close to 0.5, this difference seems to be very minor and the observed data seems to behave almost as would be expected with the proposed model in the majority of the instances.

The PPP-value for the assumption of the absence of outliers is 0.048. This means that the proposed model expects a wider spread than present in the data, and it also means that there are no outliers in the observed data.

The bayes factor for $H_1(\beta_1 = 0; \beta_2 = 0)$ has a value of 0.00, meaning there is virtually infinitely more support for the hypothesis that both coefficients are not 0, which indicates that the model fits the data. The bayes factor for $H_2(\beta_{1stz} > \beta_{2stz})$ with the standardized coefficients has a value $>$ 1,000,000, which indicates that there is an unseizable amount of larger support for the hypothesis that the standardized coefficient for protein is larger than the standardized coefficient for GDP.

The DIC for the model containing both parameters is 2537. For the model containing protein only, it is 2559 and for the model containing GDP only, it is 2856.

7 Interpretation

The given data was successfully modelled with the MCMC-method, which was indicated by the first BF. The posterior means of the parameters of interest indicate that the calories coming from protein do have an effect on the mean height of men in a given country. Curiously, the posterior mean of the beta-coefficient for GDP is negative (although very small). This might be because money is often distributed unequally within countries and therefore, in some countries with a large GDP, the wealth is not accessible to most citizens resulting in a shorter mean height. Another possible reason is the collinearity between protein and GDP, resulting in unreliable distributions. However, it might also be that the distribution of GDP is so skewed that estimates become unstable. This can be solved by log-transforming GDP, which will be discussed at the end of this report.

The BF's point in favor of my expectations. There are effects of both GDP and protein on the mean height of men within a country, but the effect of protein is most definitely larger than the effect of GDP. The incredibly large BF almost screams that I should not even have had this discussion with my friend. The DICs for the one-parameter models, which do not suffer from collinearity, further show that protein seems to explain the mean height better than GDP does. Altogether, all the evidence points in favor of my argument.

8 Discussion

In this report, I investigated whether protein-consumption is a better explanation for the mean height of a population than overall wealth. Using a MCMC-method, the DIC and BF, I showed that protein has a a larger effect than GDP. However, the strong relationship between GDP and protein might cause predictors to be unreliable. To circumvent this, I inspected both predictors individually, showing that a model containing protein only has a better fit than a model containing DIC only.

Besides the collinearity between protein and GDP, another problem is the very skewed distribution of GDP. Typically, this is circumvented in economic studies by log-transforming GDP. After writing the sampler, I also ran it with the log-transformed GDP-variable. Figure 3 shows the trace plots of both chains for that model after 10000 iterations on the left and histograms for the posterior distributions of the parameters on the right. It is very obvious that the chains did not converge yet. The most striking observation in the histograms is the very wide distribution of the coefficient for protein and that the first three plots look "less" normal than they should. If I specify a shorter range for the protein-coefficient, the plots actually start to resemble uniform distributions or become skewed or have two modes. Using the log(GDP)-variable in a one-predictor-model, this striking difference does not appear, but the sampler needs quite a while to converge, too (See Appendix).

I assume that the difference stems from the fact that the log of GDP is even stronger correlated with protein and this collinearity results in autocorrelation and weird distributions and in a wider distribution of the protein-coefficient. Strong collinearity seems to cause slower convergence. However, it is still puzzling why the sampler converges more slowly when using the log(GDP)-variable individually.

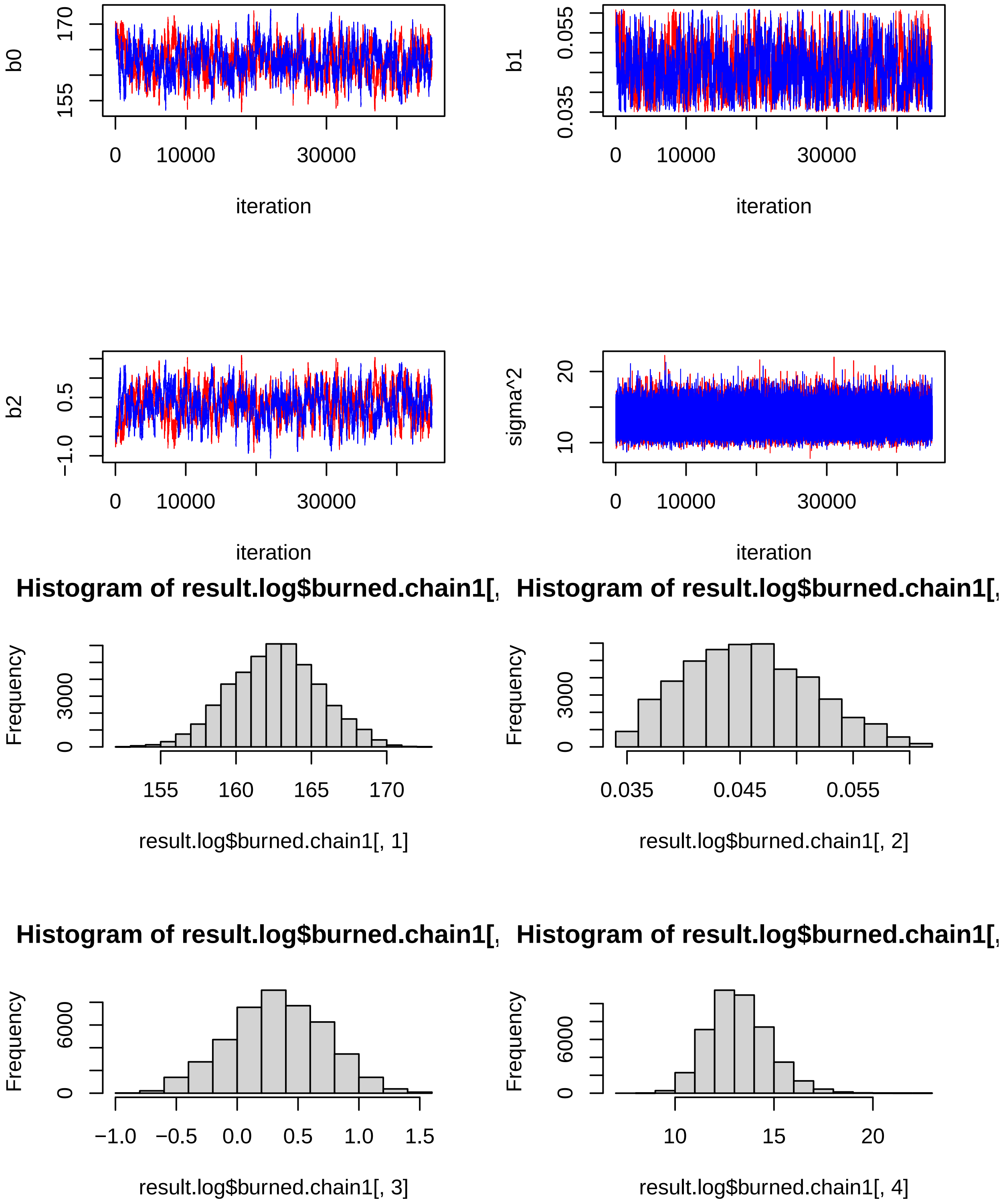


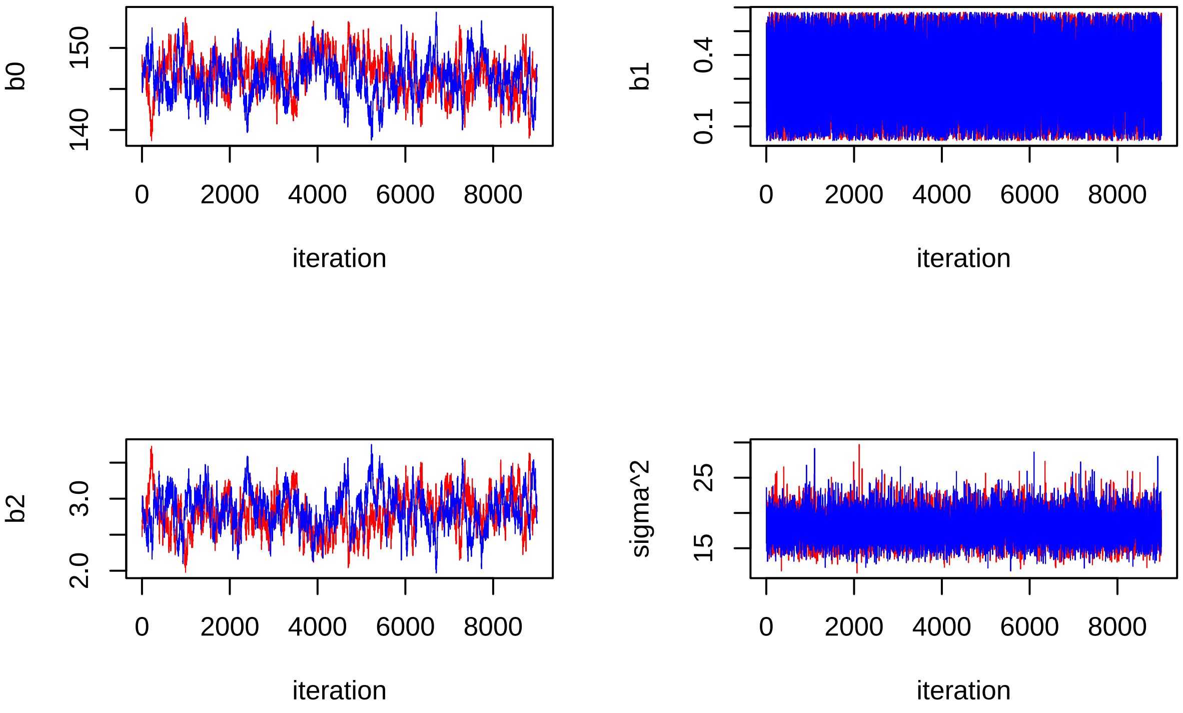
Figure 3: Trace plots and histograms for the model using log(GDP).

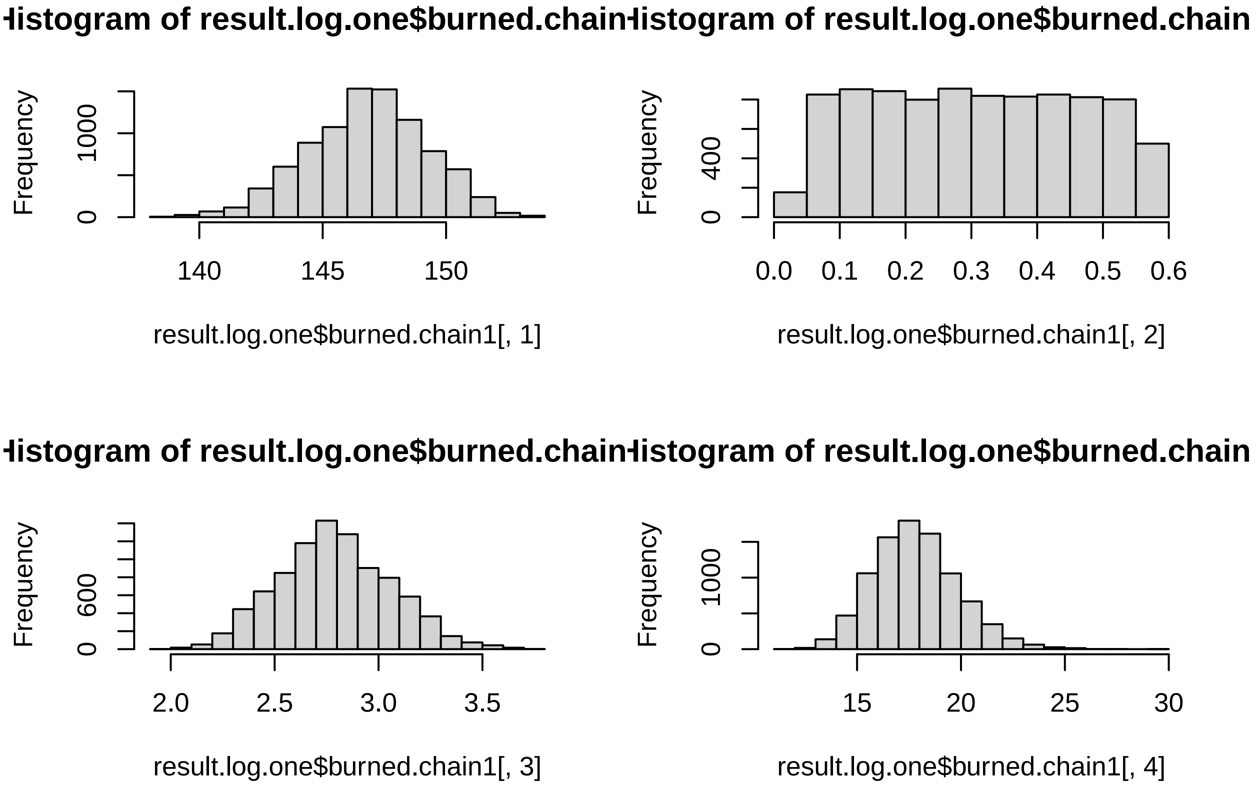
Sources

Our World In Data 2021 a: Total calories from animal protein vs. Mean male height, 1996. https://ourworldindata.org/grapher/share-of-calories-from-animal-protein-vs-mean-male-height. (last accessed 2 June 2021).

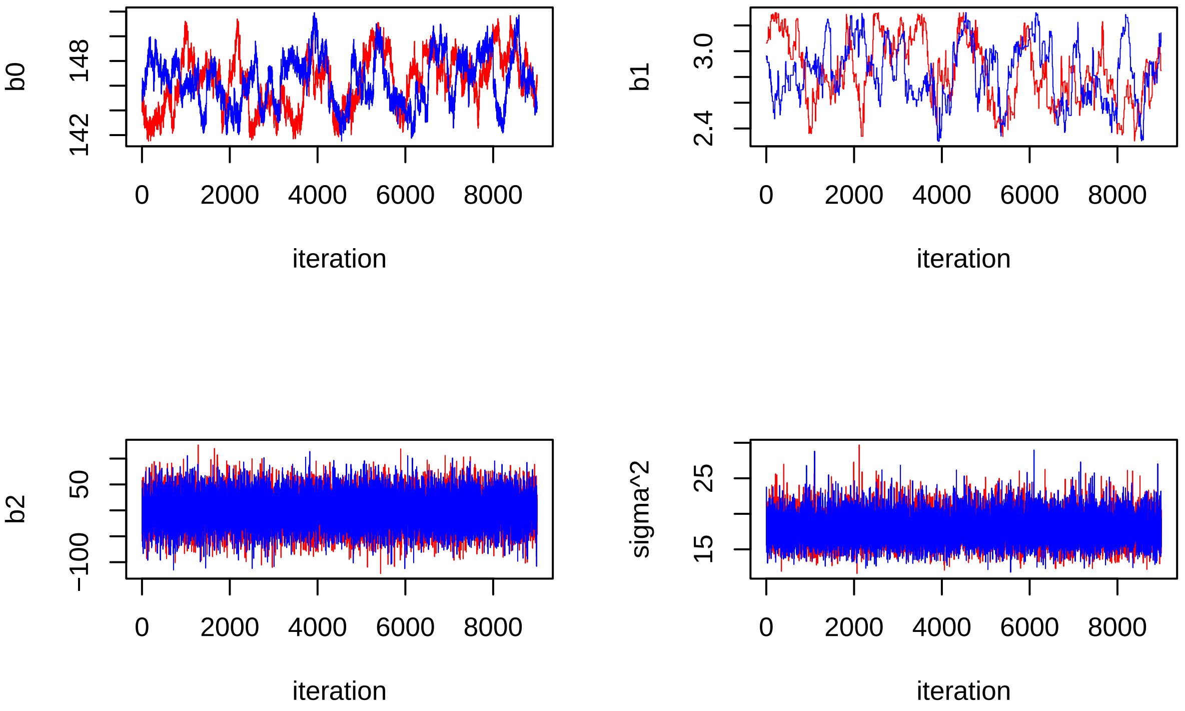
Our World in Data 2021 b: Meat consumption vs. GDP per capita, 2017. https://ourworldindata.org/grapher/meat-consumption-vs-gdp-per-capita. (last accessed 2 June 2021).

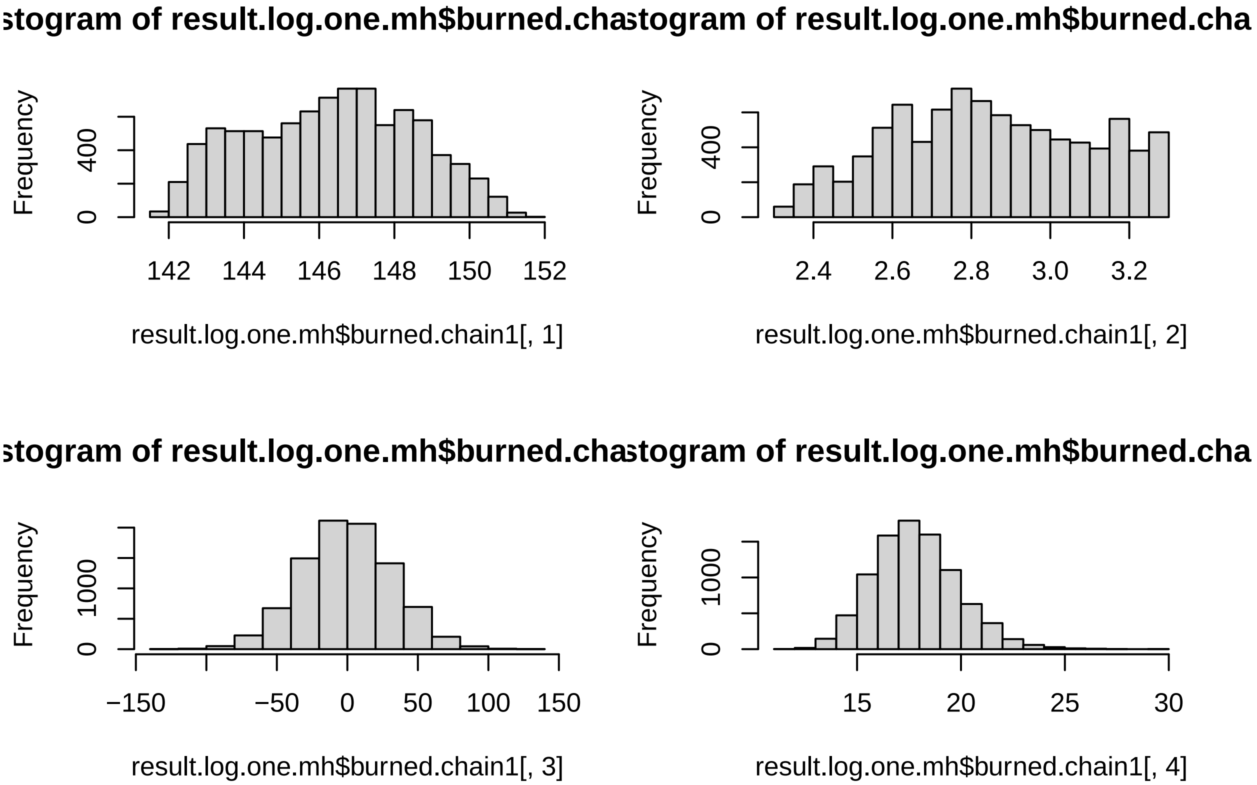
Appendix





Appendix I: Trace plots and histograms for the model using log(GDP) as the only predictor, first coefficient empty (i.e. x_1 is a vector of 0's).





Appendix II: Trace plots and histograms for the model using log(GDP) as the only predictor, second coefficient empty (i.e. x_2 is a vector of 0's).