Bayesian Statistics

Kalyango Jovan (7014430)

09-03-2022

Introduction

The sale price (in thousands of dollars) and the floor size (in squared meters) for a random sample of 27 single-family houses sold in a given town during 2005 are stored in the file house2.txt.In particular, for each house, the file contains:Taxes the amount of property taxes (thousands of dollars), paid by the previous owner during 2004,Beds number of bedrooms,Baths number of bathrooms,Size floor size (squared meters),Price sale price (thousands of dollars).It has a rows-by-columns structure, where each row corresponds to a unit, each column to a statistical variable. Differently from a matrix,it can contain strings, numerical and logical values (thus allowing for the presence of both categorical and numerical variables).The aim of the report is to fit a multiple linear regression model, in order to assess the impact of the four predictors on the sale price.The second research question is aiming to know which of the predictors has a larger impact on the Price of the house. Hence, finding the most important or useful predictor(s). The research questions are to be answered using Bayesian statistical analysis techniques such as Gibb's sampler and metropolis-Hastings algorithm as a data set of my choice fits the analysis type. Furthermore, DIC to assess the better combination of fit(misfit) and complexity of model for the data set. Bayes factor will be used last in this report to test the hypothesis that all the predictors(four) are greater than zero against the alternative hypothesis that one of then of the predictor is zero. This report is represented in form of sections, these include parameter estimation, Convergence, interpretation(these two are connected to parameter estimation), Posterior predictive P-value, model selection, model comparison and conflicting ideas. The last section is going to be on reflecting the difference between the bayesian and frequetist thinking, this will more of a summary of what is going to be done throughout this article. This report is made of mainly programmed functions that i have derived from lecture notes mainly(mainly lecture number 3) has influenced the gibbs sampler, but i have searched a lot of skeletons of functions in question4(assessing convergence on internet to build my own). The codes to in this report are fitted in exactly 499 lines(this is because the instructions limited me to 500 line), So, for model comparison and selection sections i have tested four models but only one is displayed, but I have labored to add possible adjustments in the corresponding blocks of the code.

Parameter Estimation

In Bayesian setup, which is used in this report, parameter estimation is done from the imagining the choice of priors to checking of whether there is convergence of the model. For the prior choice, I choose uninformative priors, this is because with no concrete previous information about the effect of variable or a combination for predicting the price of houses(the outcome variable). I will use non-informatitive(uninformative) priors for the regression coefficients of the regressors (predictors) in the dataset(houses). Thus, the assumption to be respected is that the predictors are normally distributed. As a consequence,the resulting estimates will be very similar to what the frequentist regression would yield. This report address an applied case, so, we face a challenge that the multivariate posterior distribution(product of likelihood and priors) reached at has an unknown or a very complicated form and this complicates finding of parameters. To solve this Bayesian approach thinking use MCMC motheds, for this report I programmed a function **m_gibbs**, which follows *Gibbs sampler* and a spacial case of *Metropolis-Hastings*. In detail, to generate posterior samples by moving through each variable (or block of variables) to sample from its conditional distribution with the

remaining variables that are fixed to their current values.In this report there is four predictors and total of six parameters to be estimated so,applying the multivariate approach for a Gibbs sampler of multiple linear regression with non-informative priors is the Bayesian technique that is deployed in this report.The linear model is: $y_i = x_i\beta + e_i$, where $e_i = \mathcal{MVN}(\mu, \sigma^2 I_{27})$. Here, μ is a zero-vector with length 5 (the number of predictors, including a 1-vector for the intercept), and I_{27} is a 27-dimensional Identity matrix (where 27 is the number of observations/ houses). The covariance matrix of the distribution of errors is thus a diagonal matrix where the elements on the diagonal σ^2 are all identical (representing the homoscedastictiy assumption) and the off-diagonal elements are all zero (representing the independence of error assumption).The density of the data, given the model, is:$P(x_i, y_i|\beta, \sigma^2) = \mathcal{MVN}(\mu = \beta, \Sigma = \sigma^2 I_{27})$ These starting values are completely arbitrary in both chains,they fits my choice of prior, i will take no previous information into consideration with my starting values. The values are going to be used in my function **gibbs__spl** and it is the outer, inside that function there is another one **gibbs__spl__chain** that goes thrOUgh chains (this is because the gibbs sampler always gives better result with number of chain greater than 1). The estimation is done on more than one variable(four predictors, intercept and residual variance), so i have used matrix algebra to define predictors. To test my Gibbs function mentioned above **gibbs__spl** is stored in another function **m__gibbs** which is used throughout this report to report my estimates.

To test the function, I set chains to 2 chains, model(formula) as it would be in lm format(full model), burnin set at 1000, number of iterations equal to 10000.

For, the metropolis-hastings(MH) a function is programmed to specifically respond to the third question in this report, it is worth noting that the algorithm is a general case of Gibbs sampler, and in the applied case at hand, i have decided to include only size as my predictor, and the most impontant output to look at is the acceptance rate, if the rate is near 1, it means the we have not reached optimal rate as it is argued by different pieces of evidence, the optimal rate is between 20% to 30%, in this case I get 0.98. In the function **run__metropolis__MCMC**, **posterior** function is defined as sum of functions: **prior** and **likelihood**, to sample from the posterior density.The target function **proposalfunction** aims to jump around parameter space.

To sum up the function, the estimates are visualized in **Figure 1**. The upper row shows posterior estimates for slope (a), intercept (b) and standard deviation of the error(sd). The lower row shows the Markov Chain of parameter values.

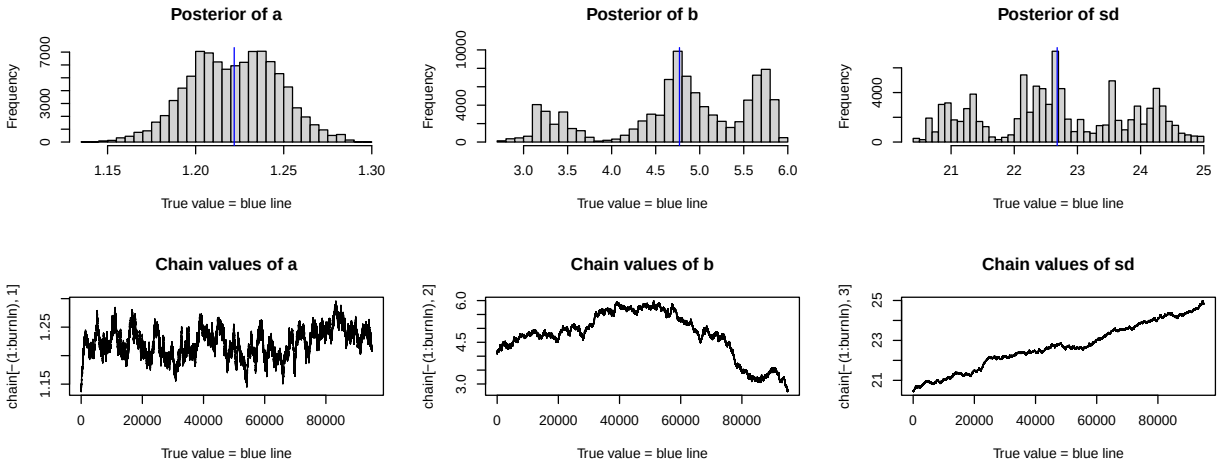

Figure 1: Posterior and Chain values of parameters

Convergence

The convergence of model is assessed by known procedures, including *Trace plots*, *Density plots*, *autocorrelation*, *Gelman and Rubin statistics*, and *naive error* using both Table 1 (*Gelman and Rubin statistics*, and *naive error*) and figures(2 to 4) I display or check the convergence of the model parameters, what is must state is that for this section I have used information given by the **m_gibbs** as the data set contains all the predictors.

Trace plots

Trace plots **Figure 2** show that they follow the desired form that resembles ' caterpillar' for all the parameters display the caterpillar like shape. This indicates that the parameters converged at the same joint posterior distribution of the model parameters.If this was not case we could, we can increase the number of iterations or introduce centering of the parameters around their means to in order the parameters to converge, but in this case there is no need to do that.

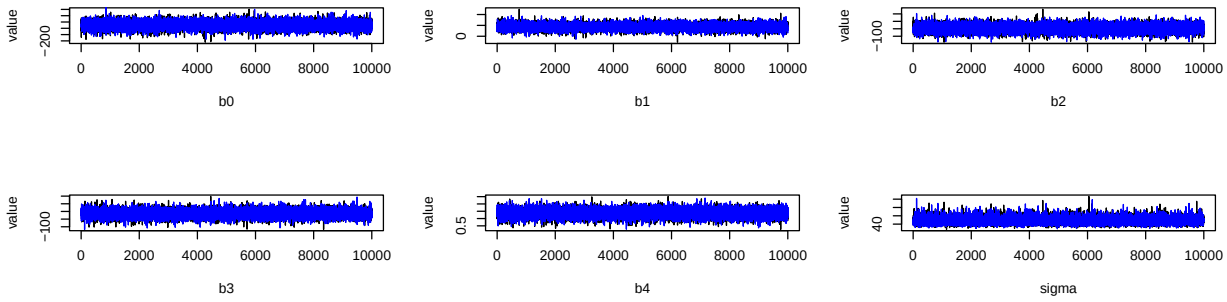

Figure 2: Traceplots of Model Parameters

Density plots

Density plots **Figure 3**, which indicated that our the parameters are as desired, normally distributed around their expected a priori (EAP) accept the 6th parameter that is slightly display left skewness but still the shift deviates within the credible interval.The blue vertical lines indicate the lower and upper bounds of the Predictive Credible Interval and all parameters in the model are within the intervals.

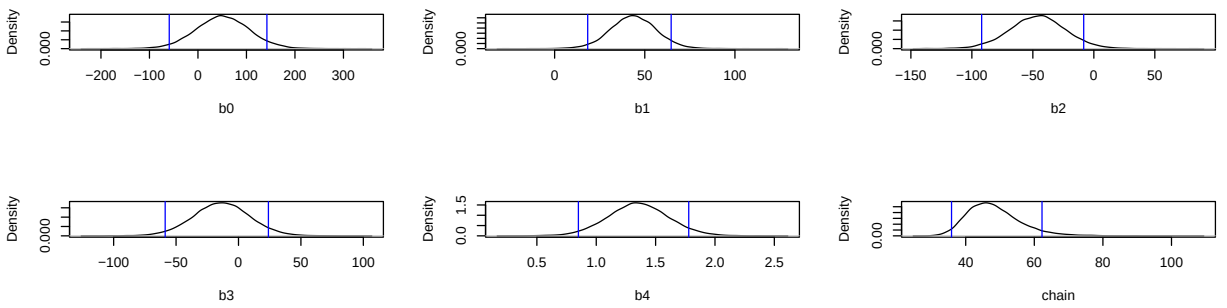

Figure 3: Density Plots of Model Paramters

Autocorrelation

I computed the autocorrelation of the model parameters per chain. The plots of the autocorrelation from lag0 to lag1000 of chain 1 are displayed in **Figure 4**. The autocorrelation of the second chain are approximately

the same and can be inspected using the `acfm()` function that I have programmed, with two main inputs model and n_lags(set to 1000). The autocorrelation converge to zero immediately at the first lag for all parameters. This indicates rapid mixing. Hence, the sampler not only appears to have reached the joint posterior distribution, but it also appears to move through it in an efficiently. Because of this, subsequent values show independence from one another.

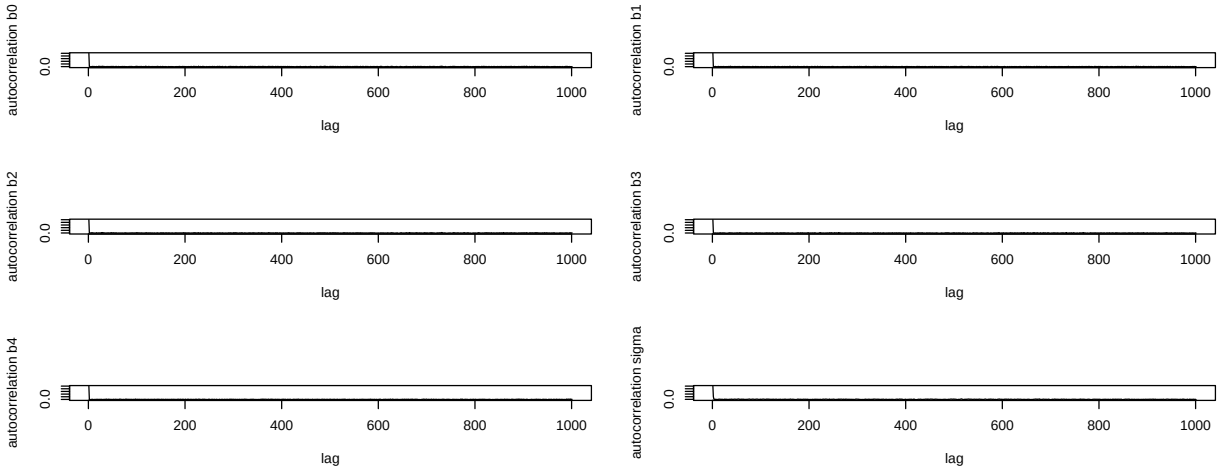

Figure 4: Autocorrelation Plots of Chain 1

Gelman and Rubin statistics

The function **gelman_rubin**, compare variance within chain(s) to total pooled variance, has input as model, the chain to be used is specified , in this function, the inner function **gelman_rubin_inner** has more inputs: para(*parameters*), m(length of the chain but eliminates the any repetitions), n(number of rows as a specified chain). The Gelman and Rubin statistics of all parameters were equal to 1 for example, b3 and b4. This suggests that there were no issues with convergence,hence, the parameters are independent.

Naive error

In the dataset there are some MC errors which are larger than 5% of the sample standard deviations(largest is 39.5% of intercept), which indicates that there are is issues with convergence.

In conclusion, most of the procedures could not find evidence that there were issues with convergence accept for the naive error MCMC. However, we cannot find proof that the sampler(Gibbs) did converge at the right posterior distribution(non-informative priors were used).

Interpretation

To interpret the parameter estimates, in this report i refer to **Table 1** and for the credible interval **Table 2** is referred to. The posterior parameter of the intercept(**b0**) is **50.674** in Table 1, corresponds to a credible interval(*CI*) between $[-59.192, 160.975]$ in Table 2, meaning based on the data, I believe that the probability of the intercept being in the mentioned *CI* is **95%**, hence it is not different from zero. Predictors that are not different from zero according to my belief are **b2** and **b3** see Tables(1 and 2). For, **b4** Size, the posterior parameter is **1.353** in Table 1 has a *CI* of $[0.850, 1.863]$ in Table 2, lead to conclusion that based on my belief the probability of the coefficient to be within the *CI* above is **95%**, thus, this is different from zero and on average a unit increase in *Size* of the house, leads to **1.353** unit increase in *Price* of the house. This is the same interpretation used for both **b1** and **sigma** that are different from zero with a probability of **95%** to be found in the credible interval. The parameter estimates, which were derived by computing the mean and standard deviations of the posteriors of the two chains combined ($N = 20,000$) are displayed in **Table 1**, their corresponding credible interval is showed in **Table2**

Table 1: Parameter Estimates

	b0	b1	b2	b3	b4	sigma
EAP	50.6740000	43.681000	-46.053000	-13.5830000	1.3530000	48.223000
SD	55.8330000	12.828000	23.131000	22.9300000	0.2580000	7.758000
Gelman Rubin	1.0000000	1.000000	1.000000	1.0000000	1.0000000	1.000000
MC error	0.3947971	0.090708	0.163563	0.1621366	0.0018237	0.054856

Table 2: Posterior Quantiles

	2.5%	25%	50%	75%	97.5%
b0	-59.192	13.805	50.454	87.697	160.975
b1	18.346	35.298	43.598	52.156	68.823
b2	-91.972	-61.069	-45.817	-30.938	-0.515
b3	-58.718	-28.705	-13.539	1.409	31.277
b4	0.850	1.183	1.351	1.523	1.863
sigma	35.865	42.740	47.260	52.543	66.145

Posterior Predictive P-value

An important assumption of multiple linear regression is that the residuals of my model are normally distributed. In order to test the hypothesis, a posterior predictive check is conducted in this report through **steps** (1 to 5). In step 1: the null-hypothesis is specified, which is residuals of the model are normally distributed, is formally specified as:$Y_i = X_i\beta + e_i$.Step 2: Sampling from the posterior distribution of the model parameters is done, where , sample parameters(θ), i.e. β and σ^2 10,000 times, using my `gibbs_spl` function.

In step 3: the Posterior Predictive Distribution is generated and each of sampled values of current parameters(θ) from θ_1 to $\theta_{10,000}$ are used to simulate data, by using the linear predictor, X is fixed. Thus,10,000 datasets simulated, where $\hat{Y}_t = \beta_t X$. We thus end up with:$[\hat{Y}_1, X], [\hat{Y}_2, X], ..., [\hat{Y}_{10000}, X]$

Step 4: choosing a discrepancy measure.A suitable test statistic to assess deviation from normality the absolute difference of the mean and median of the error: $D = |\bar{e} - \tilde{e}|$. From step1 the Null-Hypothesis that the residuals are normally distributed, this parameter should be distributed normally around zero. The residuals are computed first, as $e_t = X\beta_t$ and then the test-statistic $D_t = |\bar{e_t} - \tilde{e_t}|$.The *discrepancy measure*, depends on varying part, the posterior model parameters θ_t **and** a fixed part, the data X.

In Step 5, the *posterior-predictive (Bayesian) p-value* is compute, as:$p = P(D([Y_t, X], \theta_t) > D([Y, X], \theta_t)|[Y, X], H0)$, this represents the proportion of *discrepancy-measures*in the simulated data under the Null-Hypothesis were larger than the *discrepancy-measure* in the observed data. A p-value of *0.54* would suggest that my approximately half of the discrepancy measure that were generated using a posterior fall . This means, that in the posterior predictive distribution, an absolute difference between the mean and the median of 0 is representative, which larger differences becoming less and less likely. Therefore, a posterior predictive p-value of approximately 0.50 indicate that the residuals of my model are normally distributed. The findings are displayed in **Figure 5** see in Rmarkdown.

Model Selection-DIC

The dic function, which is more like a calculator suggests that a model without baths has the lowest value($DIC = 289.384$),this implies that a combination of three parameters[taxes, sizes and beds] produce

the best model. but the difference in *dic* of this model and the one that has all the four predictors($DIC = 291.119$) is less than 2, so with that in mind, am skeptical if indeed the baths did not affect the price. So, for future research it should be investigated, it should be kept as a predictor. I know that *DIC* is a Bayesian information criteria used to evaluate model fit or misfit, this is done by evaluating both model deviance and complexity. I would like to add a table in my report but the space is limited and I have compared four models with **dic** but due to space i have only added the other 3 as comments, so by adjusting the comments the models can be compared using *dic*.

Model comparison- Bayes Factor

Bayes factor as the change from the prior to the posterior odds: $BayesFactor = P(Posterior)/P(Prior) = 0.37$ This BF indicates that the data provide $1/0.37 = 2.7$ times more evidence for including all the predictors being included all affect the sale price, practically nothing than they do for the Price having some statistically significant effect. Thus, although the center of distribution has shifted away from 0, and the posterior distribution seems to favor a non-null effect of the predictor, it seems that given the observed data(*houses*), the probability mass has overall shifted closer to the null interval, making the values in the null interval more probable, hence

Since all the BF of less than one i.e, $BF < 1$, this indicates that contains values whose credibility has not been impressively decreased by observing the data. Testing against values outside this interval will produce a Bayes factor larger than 1/BF in support of the alternative(s). But to choose the best model, we look at the value of the function **BF_function**. I have computed, I conclusion that Baths and Size of the house influences the price more, hence more useful predictors.

Frequentist Vs Bayesian

There are two schools of statistical inference: Bayesian and frequentist. Both approaches allow one to evaluate evidence about competing hypotheses. I will finalize this report by reviewing and comparing the two approaches, starting from Prior, this is a requirement in Bayesian setup but not in Frequentist, however, when the used priors are non-infotmative the estimates will be very close or hard to differentiate from Frequentist estimates. Interpretation of estimates and the credible interval(question 6), is a consequence of the idealogical or philosophical differences,in Bayesian parameters are random variables while in Frequentist they are fixed in the population and point estimation is a used procedure to estimate parameter(s). For confidence interval and credible interval(Bayesian), confidence is expressed in terms of repeated sampling from the population, but credible is interpreted as a probability of a given estimate to be laying within it or not. Bayesian framework, for model misfit deployes DIC instead of AIC which is used in Frequentist to do the same job. Model comparison in Bayesian setup is done by means of BF, this quantify null hypothesis under investigation not only two, but multiple and updating is done(Bayesian Updating) which is not possible in Frequentist setup.

References

I have Used last year's group discussions I had with colleagues(so if some of my wording sound like theirs it is because of the highlighted reason) and notes but also general statistics formulaes more in Posterior p-value and MH sections.Futhermore, I have been constrained by the requested space(6 pages), My functions are mainly inspired by formulaes

Andrieu, C.; de Freitas, N.; Doucet, A. & Jordan, M. I. (2003) An introduction to MCMC for machine learning Mach. Learning, Springer, 50, 5-43