

ASSIGNMENT 1

Property Prices in Melbourne Using Bayesian Methods

Tarana Tabassum
S3802509

Math 2269: Applied Bayesian Statistics (2050)
Date of Submission: 30 August 2020

Abstract

Prediction of property prices is an important issue and active research area nowadays. This report conducts Bayesian analyses to predict property prices in Melbourne using a fabricated dataset with real data. Using two apps, Gibbs sampling app for normal-gamma model & Gamma Distribution Specified by Mean and Standard Deviation, the Bayesian estimate of the mean sale price (μ) & variance of the sale prices (σ^2) of properties in Melbourne were found – based on both informative and non-informative prior. Two hypotheses tests have been run on this analysis. Using each of the 95% HDI of the mean sale price and standard deviation of sale prices for both informative and noninformative priors, two hypothesis tests were run with null hypotheses that the mean sale price in Melbourne is 850,000 AUD & standard deviation of sale prices in Melbourne is 300,000 AUD. The hypotheses tests were statistically significant and different mean sale price and standard deviation of sale prices were found.

Table of contents

1. Introduction
2. Objective
3. Methodology
4. Results and Interpretation
 - 4.1 Bayesian Analysis for Informative priors (Task 1.2 & 3.2) & Hypothesis Test (Partly task 2&4)
 - 4.1.1 Descriptive look and the type of data
 - 4.1.2 Mathematical Model
 - 4.1.3 Prior Specification
 - 4.1.4 Posterior Analysis
 - 4.1.5 MCMC diagnostics
 - 4.1.6 Hypothesis Test for informative mean and standard deviation (Task 2 & 4)
 - 4.2 Bayesian Analysis for Non-informative priors (Task 1.1 & 3.1) & Hypothesis Test (Partly 2&4)
 - 4.2.1 Descriptive look and the type of data
 - 4.2.2 Mathematical Model
 - 4.2.3 Prior Specification
 - 4.2.4 Posterior Analysis
 - 4.2.5 MCMC diagnostics
 - 4.2.6 Hypothesis Test for non-informative mean and standard deviation (Task 2 & 4)
5. Conclusion
6. References

1. Introduction

Prediction of property prices is an important issue and active research area nowadays. Data scientists are coming up with different approaches for the prediction of property prices in Australia using regression methods based on either statistical methods or deep/machine learning approaches. We are working in a progressive way to conduct Bayesian analyses to predict property prices in Melbourne using a fabricated dataset with real data after a number of other analyses for the population distributions of the variables included in the dataset. This is done to comply with the rules around the use and distribution of the real data.

2. Objective

In this study we find the Bayesian estimate of the mean sale price (μ) & variance of the sale prices (σ^2) of properties in Melbourne and their 95% HDI with both informative and non-informative prior.

Using each of the 95% HDI of the mean sale price and standard deviation of sale prices for both informative and noninformative priors, we also check if the mean sale price in Melbourne is 850,000 AUD & standard deviation of sale prices in Melbourne is 300,000 AUD.

3. Methodology

A Bayesian analysis has been carried to find the Bayesian estimate of the mean sale price (μ) & variance of the sale prices (σ^2) of properties in Melbourne and their 95% HDI, using both informative and non-informative prior. Two apps used in this analysis are –

- (i) Gibbs sampling app for normal-gamma model
- (ii) Gamma Distribution Specified by Mean and Standard Deviation

Two statistically significant hypotheses tests have been run on this analysis. Using each of the 95% HDI of the mean sale price and standard deviation of sale prices for both informative and noninformative priors, we ran hypothesis tests that the mean sale price in Melbourne is 850,000 AUD & standard deviation of sale prices in Melbourne is 300,000 AUD.

4. Results & Interpretation

The analysis undertaken here has been divided into two major parts –

- (i) Bayesian Analysis for Informative priors & Hypothesis Test (Task 1.2, 3.2 & partly 2,4)
- (ii) Bayesian Analysis for Non-Informative priors & Hypothesis Test (Task 1.1, 3.1 & partly 2,4)


The results & interpretation part of this report consists description of these analyses stating every step and outcome below –

4.1 Bayesian Analysis for Informative priors (Task 1.2 & 3.2) & Hypothesis Test (Partly task 2 & 4)

The steps of this Bayesian analysis for informative priors are described below -

4.1.1 Descriptive look and the type of data

The data file “PropertyPrices” includes only sale price of properties in Melbourne in "y" column in 100,000 of AUD which is continuous data. A screenshot of the first few values of this column via R Studio is following –



	y
1	3.85000
2	5.66000
3	5.67500
4	5.47000
5	4.40000
6	2.50000
7	3.20000
8	6.17000
9	2.47500
10	3.91000
11	5.20000
12	3.18500
13	4.91500
14	6.00000
15	4.40000
16	5.50000
17	3.30000
18	2.14500
19	4.32000

Figure 1: First 19 observations of the Property Prices dataset

A basic summary statistic of the dataset is presented below where we take a look at the mean, median and other summary details of the given dataset.

```
> summary(PropertyPrices)
      y
Min.   : 2.000
1st Qu.: 3.500
Median : 4.500
Mean   : 6.094
3rd Qu.: 6.550
Max.   :70.000
> |
```

Figure 2: Summary statistics of the Property Prices dataset

4.1.2 Mathematical Model

$$\begin{aligned}X &\sim \text{Normal}(\mu, \tau) \\ \mu &\sim \text{Normal}(\mu_0, \tau_0) \\ \tau &\sim \text{Gamma}(\alpha, \beta)\end{aligned}$$

μ here is the mean of sale prices and τ is variance of sale prices of properties.
 μ_0 is mean of the normal prior distribution, τ_0 is variance of the normal prior distribution.
 α and β are parameters of the gamma prior for the population variance.

As we will be modelling two parameters here at the same time – normal and gamma, a Gibbs sampling app for normal-gamma model will be used to implement the Bayesian estimation of the normally distributed mean and variance.

4.1.3 Prior Specification

Here, given is -

mean of variance of sale prices, $\mu = 600,000 = 6$ (converted base)

mean of the normal prior distribution, $\mu_0 = 750,000 = 7.5$ (converted base)

Using the Gibbs sampling app for normal-gamma model, we will find a variance of the normal prior distribution, τ_0 . Since this is an informative Bayesian analysis, we will be assuming a very small variance which will be close to zero, as smaller variance value represents higher belief.

In the Gibbs sampling app, we also need to insert values of α & β parameters. So first, we use the 'Gamma Distribution Specified by Mean and Standard Deviation' app to find these values -

[gamma_2.html](#)

[Download gamma_2.html](#) (819 KB)

Developed by Dr Haydar Demirhan - haydar.demirhan@rmit.edu.au based on the style settings given by Kruschke, J. K. (2015). Doing Bayesian Data Analysis, Second Edition: A Tutorial with R, JAGS, and Stan. Academic Press / Elsevier.

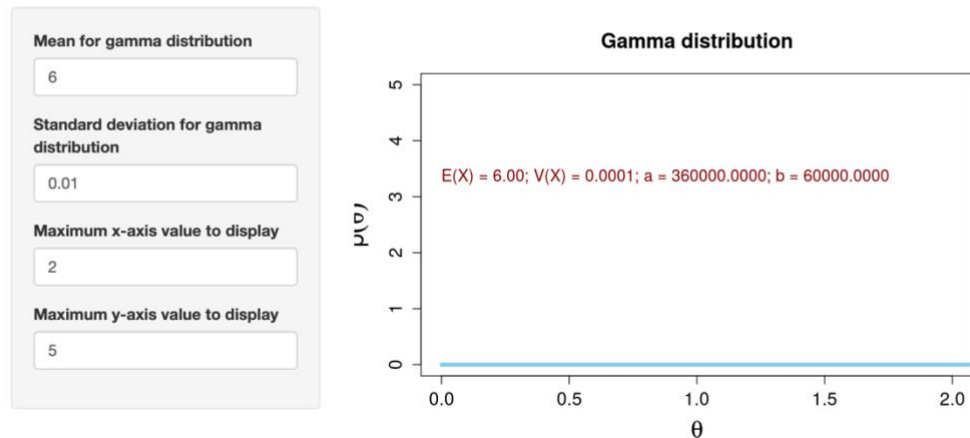


Figure 3: Outcome of the Gamma app as we approach to find α & β

Here,

mean for gamma distribution = 6 which is the mean of variance of sale prices.

Standard deviation = $\sqrt{\text{Small variance}} = \sqrt{0.0001} = 0.01$

We used a variance that is close to zero because this approach is an informative prior.

We find α & β values from this app which are **360000 & 60000** respectively.

4.1.4 Posterior Analysis

Now we insert these α & β values along with the μ_0 value in the Gibbs sampling app for normal-gamma model -

[GibbsNormalGamma.html](#)

[Download GibbsNormalGamma.html](#) (819 KB)

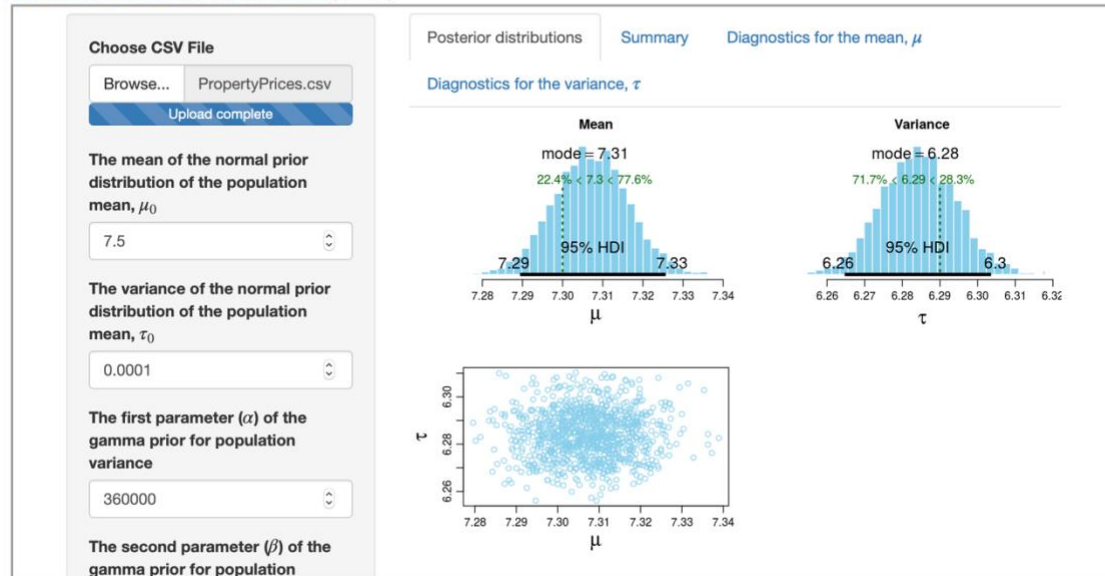


Figure 4: Posterior distributions - Gibbs Sampling App for normal gamma model – informative prior

Here,

$\mu_0 = 7.5$

$\tau_0 = 0.0001$

$\alpha = 360000$

$\beta = 60000$

We find posterior mean, $\mu = 7.31$ and variance, $\tau = 6.28$ on this app.

Here, $\tau_0 = 0.0001$, which was adjusted after multiple trials to bring the posterior mode closer to prior mean. The posterior analysis comes closer to the prior mean only when the variance is small enough for the informative prior and close to zero.

So, for informative prior the mean sale price (μ) was 731,000 & variance of the sale prices (σ^2) of properties in Melbourne was 628,000.

4.1.5 MCMC diagnostics

We will now check the diagnostics for mean and variance. For a good diagnostic, we set number of chains to 2; with 500 burn in steps, 5000 saved steps and thinning to 2.

(i) Diagnostics for mean

GibbsNormalGamma.html

[Download GibbsNormalGamma.html](#) (819 KB)

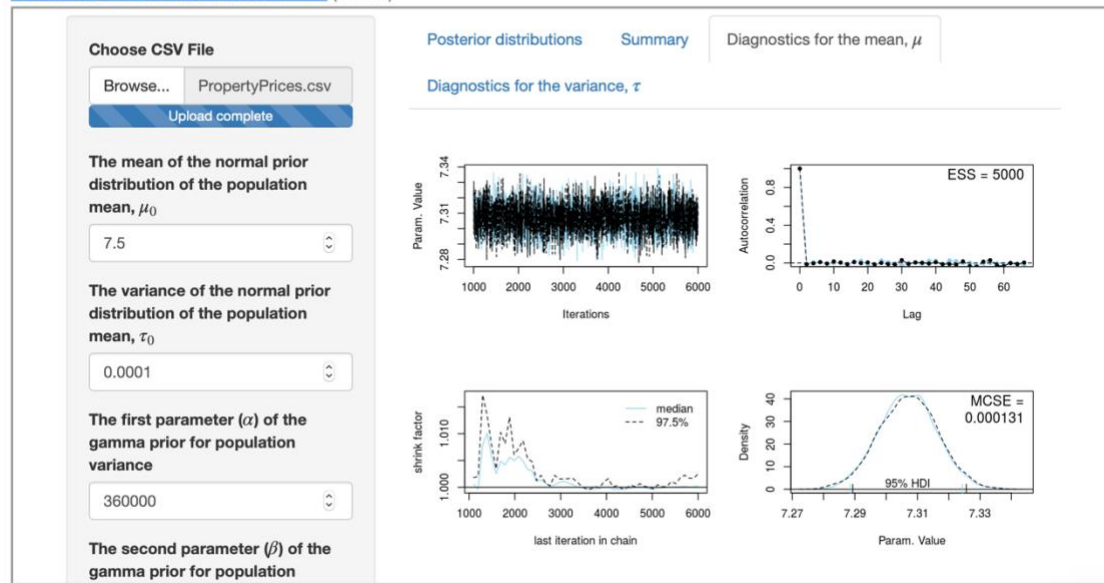


Figure 5: Diagnostics for mean - Gibbs Sampling App for normal gamma model – informative prior

Diagnostic checks confirm a proper posterior. From the figure above, diagnostic for mean can be observed. It shows good amount of overlapping in trace plots, no auto correlation, higher Estimated Sample Size (ESS) 5000, shrink factor (=1) is less than 1.2, overlapping density plot with overlapping HDIs and very low Monte Carlo Standard Error (MCSE) 0.000131.

(ii) Diagnostic for variance

GibbsNormalGamma.html

[Download GibbsNormalGamma.html](#) (819 KB)

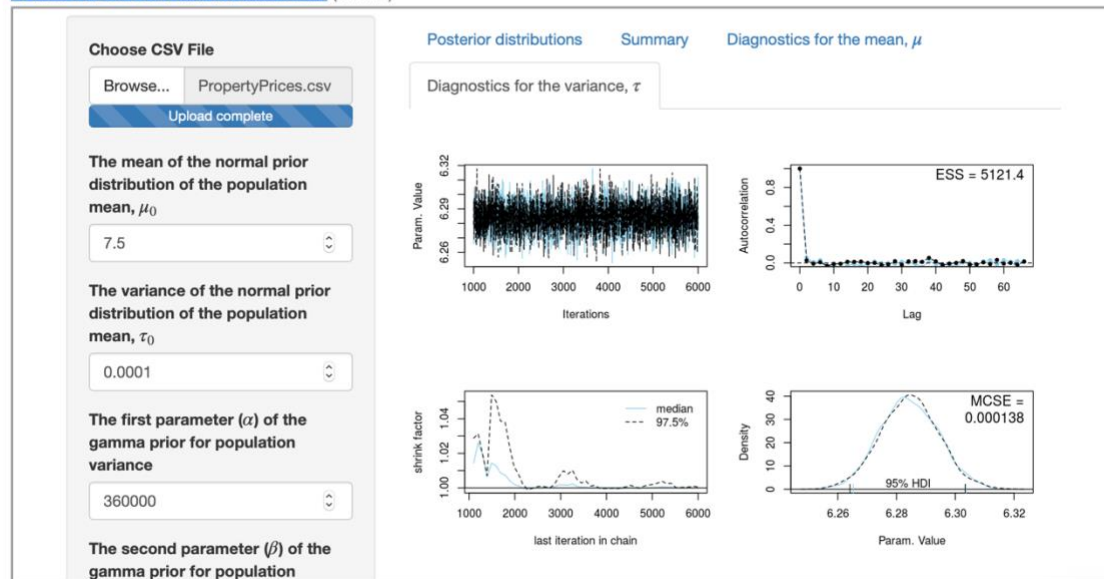


Figure 6: Diagnostics for variance - Gibbs Sampling App for normal gamma model – informative prior

From the figure above, diagnostic for variance can be observed. It shows good amount of overlapping in trace plots, no auto correlation, higher Estimated Sample Size (ESS) 5121.4, shrink factor ($=1$) is less than 1.2, overlapping density plot with overlapping HDIs and very low Monte Carlo Standard Error (MCSE) 0.000138.

4.1.6 Hypothesis Test for informative mean and standard deviation (Task 2 & 4)

Based on our informative model, the summary table for Bayesian estimation is following -

Posterior distributions

Summary

Diagnostics for the mean, μ

Diagnostics for the variance, τ

Summary table for Bayesian estimates

	Mean	Median	Mode	ESS	HDImass	HDIlow	HDIhigh	CompVal
mu	7.307016	7.306986	7.306798	4689.4	0.95	7.289667	7.325410	8.5
tau	6.284349	6.284368	6.282671	4610.1	0.95	6.263271	6.302522	9.0
PcntGtCompVal								
mu		0						
tau		0						

Run Time

	user	system	elapsed
	45.140	0.000	45.228

Figure 7: Hypothesis Test - informative prior

Hypothesis Testing (Mean):

We assume, null hypothesis H_0 : Mean Sale price, $\mu = 850,000$ AUD = 8.5 (Converted base).

From the above summary table, we can see two HDI intervals of μ :

HDI low = 7.289667

HDI High = 7.325410

Our $H_0 = 8.5$ does not fall between the HDI intervals, as a result we can reject the null hypothesis and say that, our mean sale price is not 850,000 AUD.

Hypothesis Testing (Standard Deviation):

We assume, null hypothesis H_0 : Standard Deviation = 300,000 AUD = 3 (Converted base).

So, in this case, the Variance, τ should be = $(SD)^2 = (3)^2 = 9$

From the above summary table, we can see two HDI intervals of τ :

HDI low = 6.263271

HDI High = 6.302522

Our H_0 variance= 9 does not fall between the HDI intervals, so the standard deviation is not 3. So, we can reject the null hypothesis and say that, the standard deviation of sale prices in Melbourne is not 300,000 AUD.

4.2 Bayesian Analysis for Non-informative priors (Task 1.1 & 3.1) & Hypothesis Test (Partly task 2&4)

The steps of Bayesian analysis for non-informative priors are described below.

4.2.1 Descriptive look and the type of data

The same data file “PropertyPrices” is again used for the non-informative Bayesian analysis, which includes only sale price of properties in Melbourne in “y” column in 100,000 of AUD.

4.2.2 Mathematical Model

$$X \sim \text{Normal}(\mu, \tau)$$

$$\mu \sim \text{Normal}(\mu_0, \tau_0)$$

$$\tau \sim \text{Gamma}(\alpha, \beta)$$

μ here is the mean of sale prices and τ is variance of sale prices of properties.

μ_0 is mean of the normal prior distribution, τ_0 is variance of the normal prior distribution.

α and β are parameters of the gamma prior for the population variance.

As we will be modelling two types of parameters here at the same time – normal and gamma, a Gibbs sampling app for normal-gamma model will be used to implement the Bayesian estimation of the normally distributed mean and variance.

4.2.3 Prior Specification

Here, we assume,

mean of variance of sale prices, $\mu = 600,000 = 6$ (converted base)

mean of the normal prior distribution, $\mu_0 = 750,000 = 7.5$ (converted base)

Using the Gibbs sampling app for normal-gamma model, we will find a variance of the normal prior distribution, τ_0 . Since this is a non-informative Bayesian analysis, we will be assuming a large variance compared to the mean value of 7.5.

In the Gibbs sampling app, we also need to insert values of α & β parameters. We use the 'Gamma Distribution Specified by Mean and Standard Deviation' app to find these values -

[gamma_2.html](#)

[Download gamma_2.html](#) (819 KB)

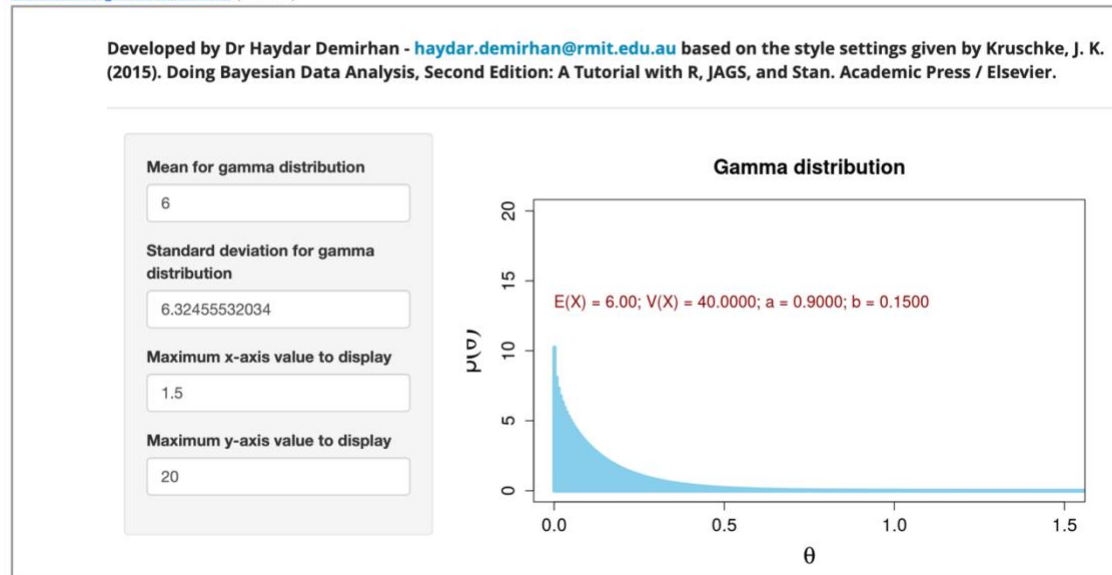


Figure 8: Outcome of the Gamma app as we approach to find α & β

Here,

mean for gamma distribution = 6 which is the mean of variance of sale prices.

Standard deviation = $\sqrt{\text{Large variance}} = \sqrt{40} = 6.32455532034$

We used a large variance because this approach is a non-informative prior.

We find α & β values from this app which are **0.9 & 0.15** respectively.

4.2.4 Posterior Analysis

Now we insert α & β values along with the μ_0 value in the Gibbs sampling app for normal- gamma model.

GibbsNormalGamma.html

[Download GibbsNormalGamma.html](#) (819 KB)

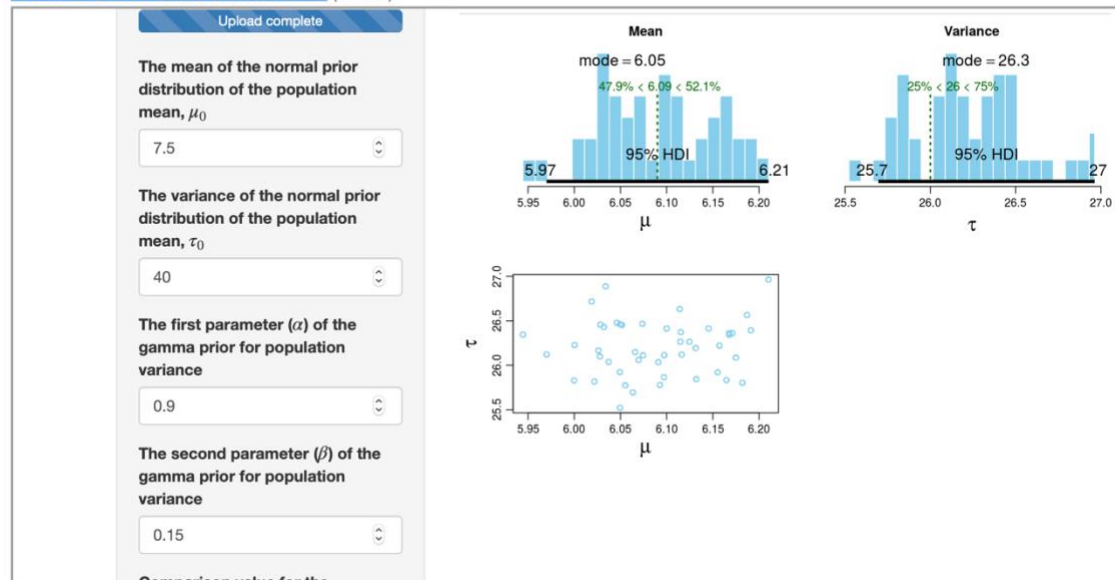


Figure 9: Posterior distributions - Gibbs Sampling App for normal gamma model – non informative prior

Here,

$$\mu_0 = 7.5$$

$$\tau_0 = 40$$

$$\alpha = 0.9$$

$$\beta = 0.15$$

We find posterior mean, $\mu = 6.05$ and variance, $\tau = 26.3$ on this app.

Here, $\tau_0 = 40$, adjusted after several trials, is the large variance which indicates non-informativeness of prior.

So, for non-informative prior the mean sale price (μ) was 605,000 & variance of the sale prices (σ^2) of properties in Melbourne was 263,000

4.2.5 MCMC diagnostics

We will now check the diagnostics for mean and variance. For a good diagnostic, we set number of chains to 2; with 500 burn in steps, 5000 saved steps and thinning to 2.

(i) Diagnostics for mean

GibbsNormalGamma.html

[Download GibbsNormalGamma.html](#) (819 KB)

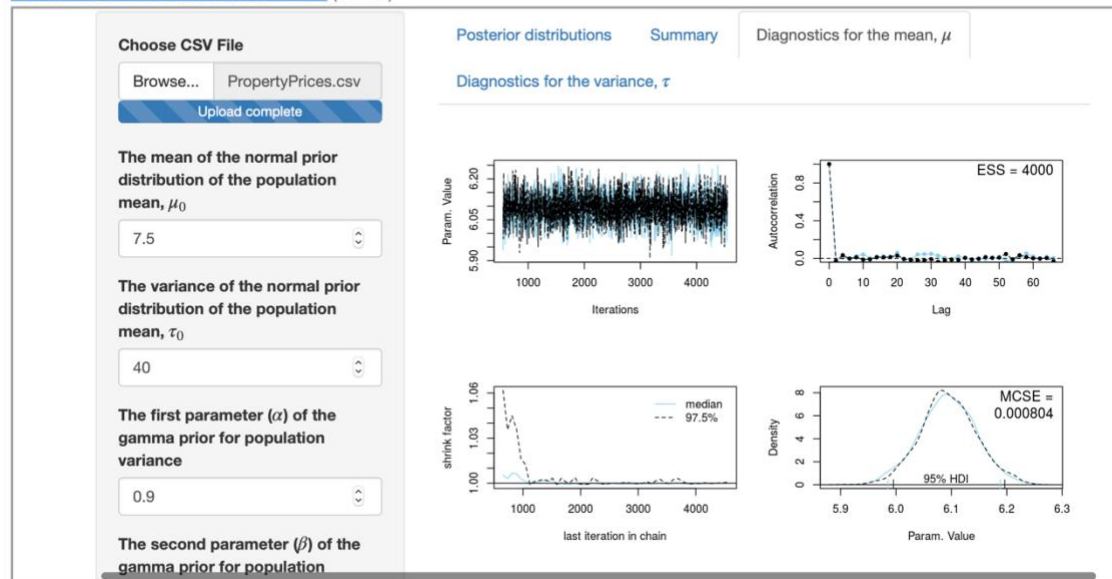


Figure 10: Diagnostics for mean - Gibbs Sampling App for normal gamma model – non informative prior

Diagnostic checks confirm a proper posterior. From the figure above, diagnostic for mean can be observed. It shows that good amount of overlapping in trace plots, no auto correlation, higher Estimated Sample Size (ESS), shrink factor (=1) is less than 1.2, overlapping density plot with overlapping HDIs and very low Monte Carlo Standard Error (MCSE) 0.000804.

(ii) Diagnostic for variance

GibbsNormalGamma.html

[Download GibbsNormalGamma.html](#) (819 KB)

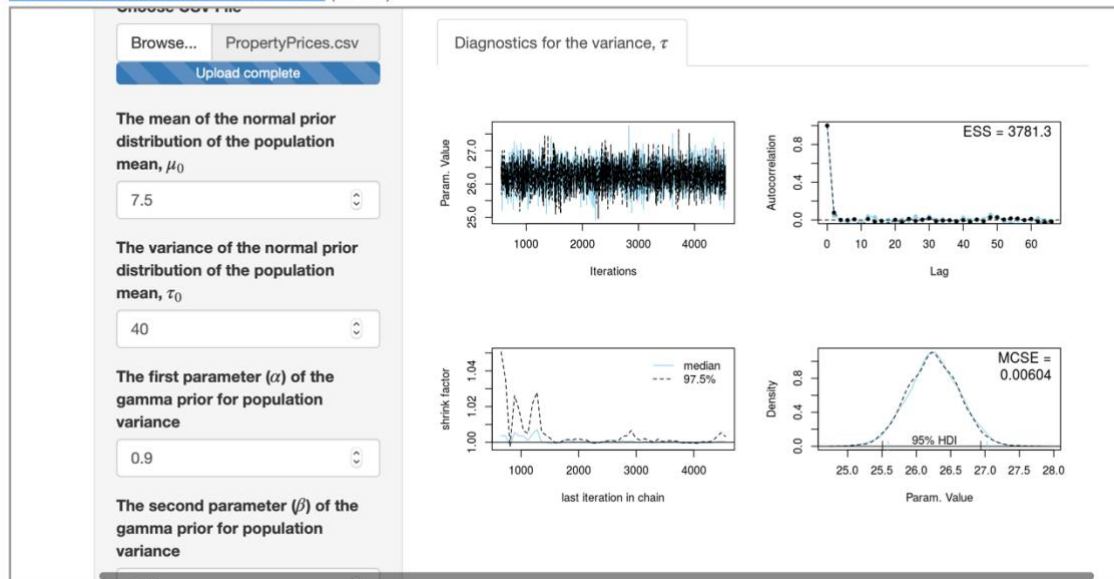


Figure 11: Diagnostics for variance - Gibbs Sampling App for normal gamma model – non informative prior

From the figure above, diagnostic for variance can be observed. It shows good amount of overlapping in trace plots, no auto correlation, higher Estimated Sample Size (ESS), shrink factor ($=1$) is less than 1.2, overlapping density plot with overlapping HDIs and very low Monte Carlo Standard Error (MCSE) 0.00604.

4.2.6 Hypothesis Test for non-informative mean and standard deviation (Task 2 & 4)

Based on our non-informative model, the summary table for Bayesian estimation is following -

Summary table for Bayesian estimates

	Mean	Median	Mode	ESS	HDI _{mass}	HDI _{low}	HDI _{high}	CompVal
mu	6.092866	6.091775	6.087153	5000.0	0.95	5.991737	6.194534	8.5
tau	26.251621	26.246849	26.218263	4278.4	0.95	25.495589	26.943083	9.0
PcntGtCompVal								
mu	0							
tau	100							
Run Time								
	user	system	elapsed					
	47.164	0.008	48.176					

Figure 12: Hypothesis Test – non informative prior

Hypothesis Testing (Mean):

We assume, null hypothesis H_0 : Mean Sale price, $\mu = 850,000$ AUD = 8.5 (Converted base).

From the above summary table, we can see two HDI intervals of μ :

HDI low = 5.991737

HDI High = 6.194534

Our $H_0 = 8.5$ does not fall between the HDI intervals, as a result we can reject the null hypothesis and say that, our mean sale price is not 850,000 AUD.

Hypothesis Testing (Standard Deviation):

We assume, null hypothesis H_0 : Standard Deviation = 300,000 AUD = 3 (Converted base).

So, in this case, the Variance, τ should be = $(SD)^2 = (3)^2 = 9$

From the above summary table, we can see two HDI intervals of τ :

HDI low = 25.495589

HDI High = 26.943083

Our H_0 variance = 9 does not fall between the HDI intervals, so the standard deviation is not 3. So, we can reject the null hypothesis and say that, the standard deviation of sale prices in Melbourne is not 300,000 AUD.

5. Conclusion

In this study we have found the Bayesian estimate of the mean sale price (μ) & variance of the sale prices (σ^2) of properties in Melbourne and their 95% HDI with both informative and non-informative prior. For informative prior the mean sale price (μ) was 731,000 & variance of the sale prices (σ^2) of properties in Melbourne was 628,000. For non-informative prior the mean sale price (μ) was 605,000 & variance of the sale prices (σ^2) of properties in Melbourne was 263,000.

Using each of the 95% HDI of the mean sale price and standard deviation of sale prices for both informative and noninformative priors, we also checked if the mean sale price in Melbourne is 850,000 AUD & standard deviation of sale prices in Melbourne is 300,000 AUD; and both the hypothesis tests were statistically significant stating none of these were predicted right.

6. References

Demirhan, H. 2020, MATH2269, RMIT University, Melbourne.

Demirhan, H. 2020, 'Assignment1PropertyPrices.csv' dataset, MATH2269, RMIT University.