# MATH2269 Semester 2, 2020 – Applied Bayesian Analysis

# Using Bayesian multiple linear regression to predict Melbourne properties sales price

## Assignment 2

Millie Woo (s3806940)

## Contents

## Table of Figures

## Table of Tables

# 1. Introduction

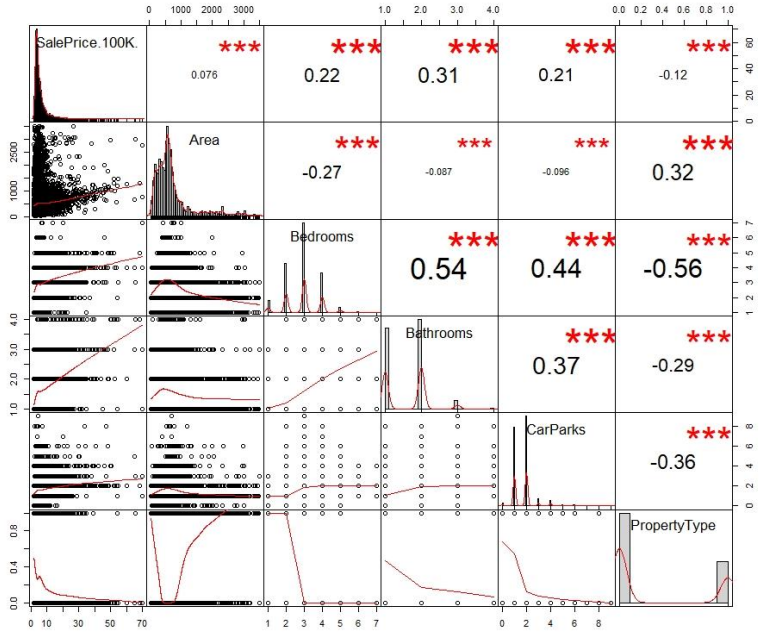
In this assignment, we are using Bayesian analysis to predict property prices in Melbourne. A datafile of several independent variables and one single dependent variable (sale price in AUD $100,000) are given to us, we are using Bayesian multiple linear regression on the provided information to examine which independent variables are relevant predictors, and then draw statistical inference to determine sale price for new entries which only independent variables are presented. In addition, we are also investigating whether prior knowledge has any impact and compare the predictions we made with different prior parameter values on the same dataset. We would document all our findings in this report.

# 2. Visualization of the dataset, sampling, distributions and modelling

## [i] - dataset

First, the dataset is the likelihood of our model, we should look at the relationship between the variables in the dataset, follow by summary statistics or count of each variable. Please refer to [A3] for the coding of this section.

*Figure 1 - PairPlots of all the variables in the dataset*



From Figure 1, it shows that there are following variables in the given dataset of 10,000 records:

1. SalePrice: Sale price in (100,000 AUD), which is our dependent variables, it is continuous and highly right-skewed, with the mean at 6.09, median at 4.5, ranges from 2.00 and the tail extends to 70. (refers to Table 1)
2. Area: Land size in $m^2$ of the sold property, which is the only continuous independent variable and it is also highly right-skewed, with the mean at 690.18, median at 568.00, ranges from 50.00 and the tail extends to 3500.00. (refers to Table 2)
3. Bedrooms: The number of bedrooms, a discrete variable, ranges from 1 to 7, more than half of the records with 3 bedrooms, refers to Table 3 for details.
4. Bathrooms: The number of bathrooms, a discrete variable, ranges from 1 to 4, records mainly with 1 or 2 bathrooms, refers to Table 4 for details.
5. CarParks: The number of car parks, a discrete variable, ranges from 0 to 9, records mainly with 1 or 2 carparks, refers to Table 5 for details.
6. PropertyType: The type of the property, a discrete variable, (0: House, 1: Unit), 68% of the records are for house, and the rest are for unit, refers to Table 6 for details.

*Table 1 - descriptive statistics of sales price*

```
## Descriptive Statistics
## myData$SalePrice.100K.
## N: 10000
##
##                         Mean   Median   Std.Dev     Q1     Q3    IQR    Min     Max
## --------------------- ------ -------- --------- ------ ------ ------ ------ -------
##       SalePrice.100K.   6.09     4.50      5.12   3.50   6.55   3.05   2.00   70.00
```

*Table 2 - descriptive statistics of area*

```
## Descriptive Statistics
## myData$Area
## N: 10000
##
##                Mean   Median   Std.Dev       Q1       Q3      IQR     Min       Max
## ---------- -------- -------- --------- -------- -------- -------- ------- ---------
##       Area   690.18   568.00    564.88   353.00   752.00   399.00   50.00   3500.00
```

*Table 3 - Frequency count and proportion on number of bedrooms*

| | 1 Bedroom(s) | 2 Bedroom(s) | 3 Bedroom(s) | 4 Bedroom(s) | 5 Bedroom(s) | 6 Bedroom(s) | 7 Bedroom(s) |
|---|---|---|---|---|---|---|---|
| Count | 594 | 2517 | 4594 | 2030 | 237 | 22 | 6 |
| proportion | 0.0594 | 0.2517 | 0.4594 | 0.203 | 0.0237 | 0.0022 | 0.0006 |

*Table 4 - Frequency count and proportion on number of bathrooms*

| | 1 Bathroom(s) | 2 Bathroom(s) | 3 Bathroom(s) | 4 Bathroom(s) |
|---|---|---|---|---|
| Count | 4499 | 4976 | 463 | 62 |
| proportion | 0.4499 | 0.4976 | 0.0463 | 0.0062 |

*Table 5 - Frequency count and proportion on number of carparks*

| | 0 CarParks(s) | 1 CarParks(s) | 2 CarParks(s) | 3 CarParks(s) | 4 CarParks(s) | 5 CarParks(s) | 6 CarParks(s) | 7 CarParks(s) | 8 CarParks(s) | 9 CarParks(s) |
|---|---|---|---|---|---|---|---|---|---|---|
| Count | 175 | 4246 | 4814 | 385 | 279 | 48 | 42 | 2 | 8 | 1 |
| proportion | 0.0175 | 0.4246 | 0.4814 | 0.0385 | 0.0279 | 0.0048 | 0.0042 | 0.0002 | 0.0008 | 0.0001 |

*Table 6 - Frequency count and proportion on property type*

| | House | Unit |
|---|---|---|
| Count | 6838 | 3162 |
| Proportion | 0.6838 | 0.3162 |

Correlation among independent variables is considerably low, <most correlated is -0.56 (Bedrooms vs PropertyType)>, we can refer to Figure 2 to examine how the discrete variables are inter-related. For example, in the middle chart, it shows that nearly 3000 records are houses with 2 bedrooms and 2 carparks, as the correlation between bedrooms and carparks is only 0.44, we can picturize there would be a mixture of different number of

bedrooms and carparks, and the chart also confirms there are spread of 3 or 4 bedrooms with low number of carparks and 1 bedroom with high number of carparks. We could then infer there is no collinearity problem with the discrete variables.

From Figure 1, it shows that Area has weak correlation <least correlated is -0.087 (Area vs Bathrooms) to most correlated at 0.32 (Area vs Property Type)> with all the discrete independent variables. This further ensures us there should not be any collinearity issues among all the independent variables.

All independent variables have weak correlation <least correlated is 0.076 (Area vs Sales Price) to most correlated at 0.31 (Bathrooms vs Sales Price)> with the sales price. Thus, individually, there would be a weak linear relationship between each independent variable with sales price. From Figure 3, we can observe the impact on the spread of Area vs Sales Price group by each discrete variable, (e.g. grouping by property type, we can see the data are clearly differentiate into 2 regions, but we are still not sure whether linear relationship exist between "Area + PropertyType" versus sales price). By using JAGS, it would help us to investigate which independent variable(s) are significant to form multiple linear regression with sales price.

*Figure 2 - Relationship of independent variables (only discrete variables)*

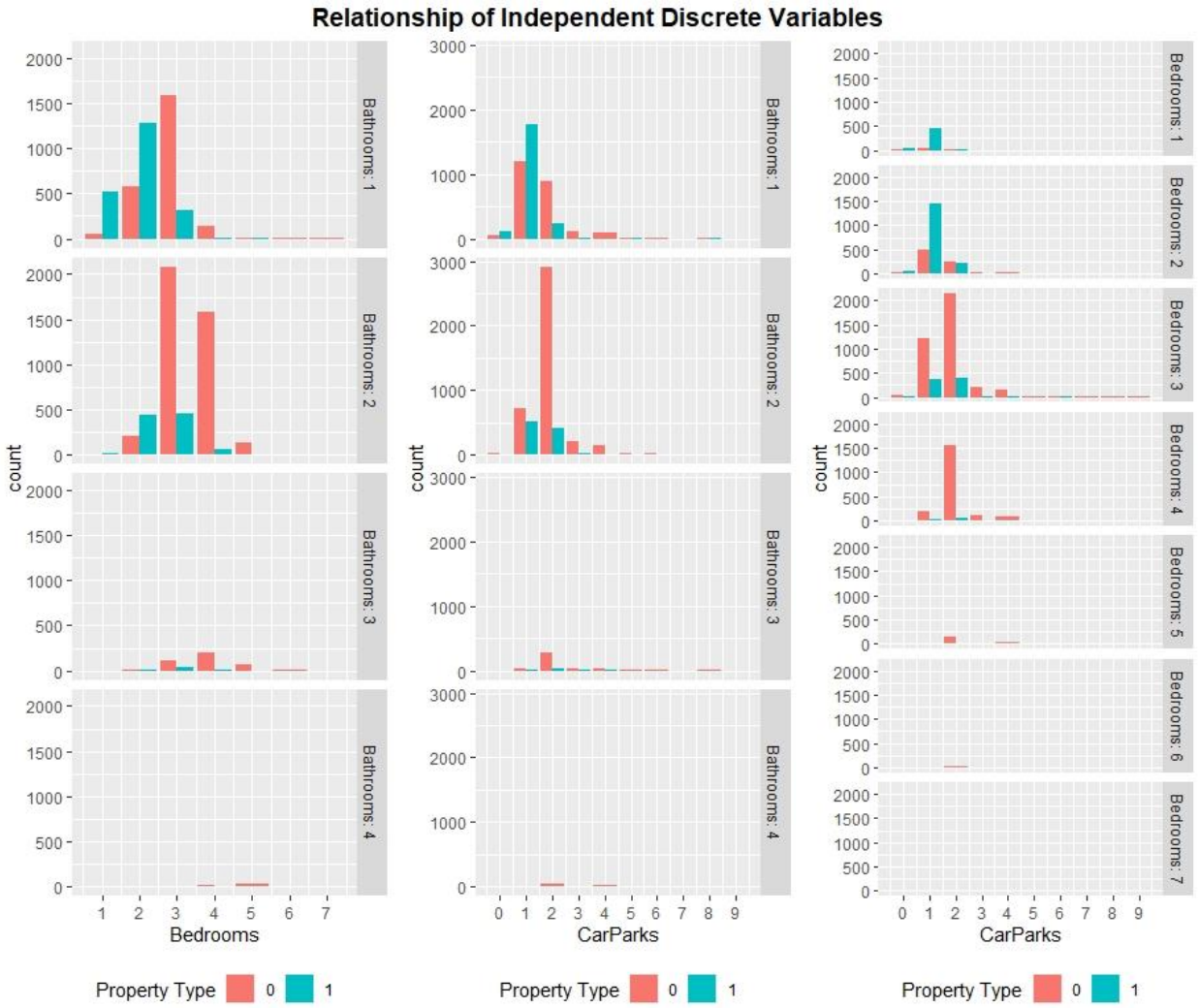
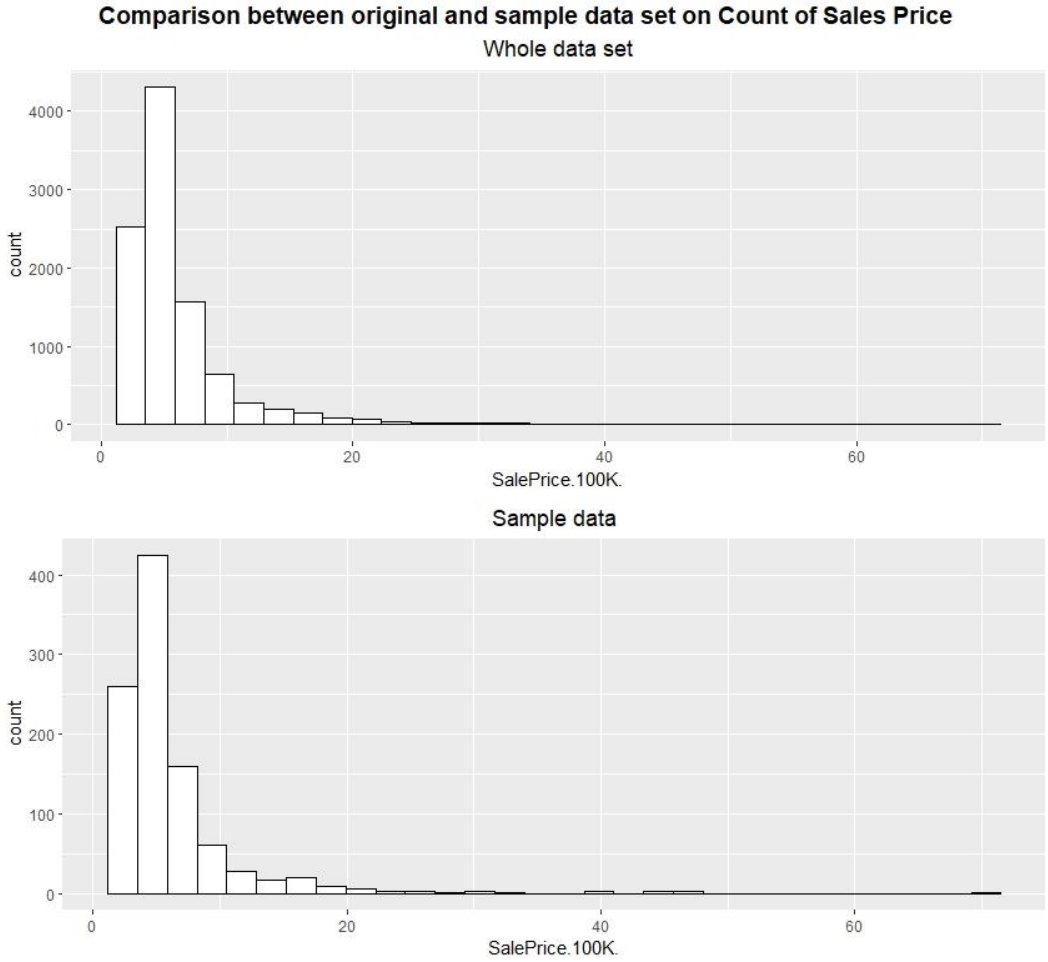


*Figure 3 - Relationships between Sales Price and Area grouped by each Discrete Variable*

## [ii] - sampling

To run JAGS efficiently, we need to sample a subset of records from the dataset, we have been working with 1000 random samples from the given data as the likelihood of our model. Figure 4 shows the resemblance from the sample dataset on the distribution of sales price. Please refer to [A4] for the coding of this section.

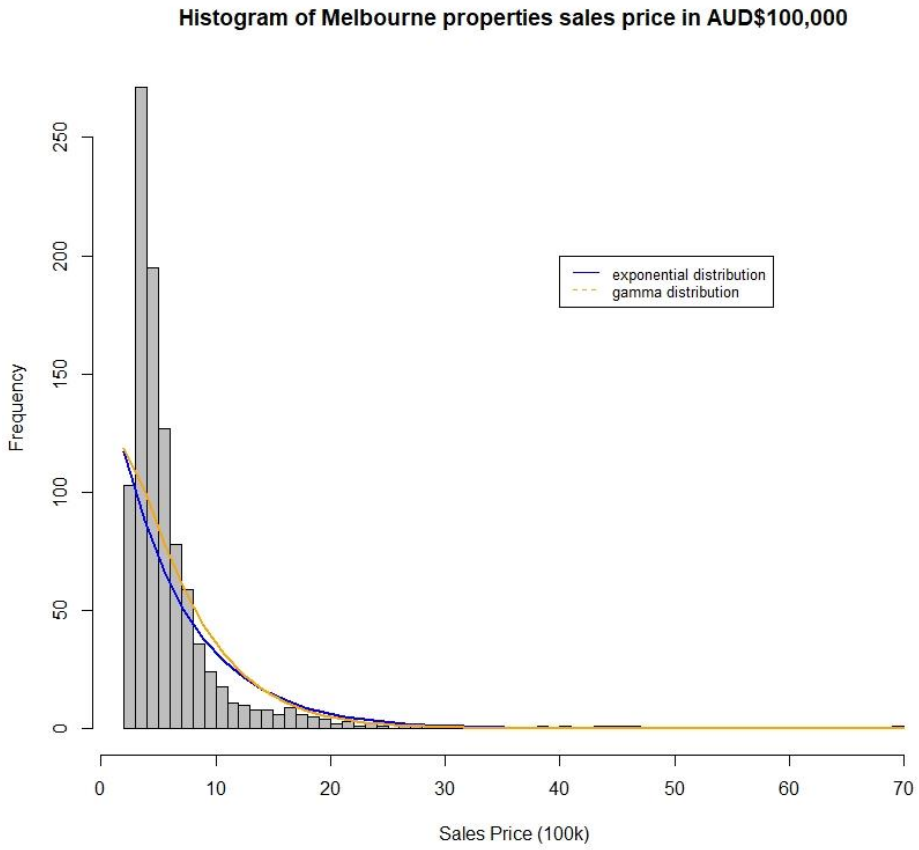
*Figure 4 - Resemblance of the sample data on sales price distribution*



## [iii] – distribution and modelling

The observed data on the dependent variable, sales price, is highly skewed in the sample dataset. As the data is continuous, with the domain of [0, +∞] , exponential and gamma distributions are most suitable to fit as the likelihood of the dependent variable. Refer to [A5], we have used the density of exponential distribution function dexp(x, rate = 1/μ)to draw a corresponding exponential distribution, and used the density of gamma distribution function dgamma(x, alpha=$\mu^2/\sigma^2$, beta= $\mu/\sigma^2$)to draw a corresponding gamma distribution, overlaying on top of the histogram of sales price in Figure 5. (where μ is the mean, σ is the standard deviation of the sales price of the sample data set). We have noticed how both distributions fit the dataset by catering the long tail and intercepting middle of the burst on second bar of the histogram.

*Figure 5 - Histogram of Sales Price overlayed with exponential and gamma distribution*



We then model the entire given dataset using the multiple linear regression formula as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

Where

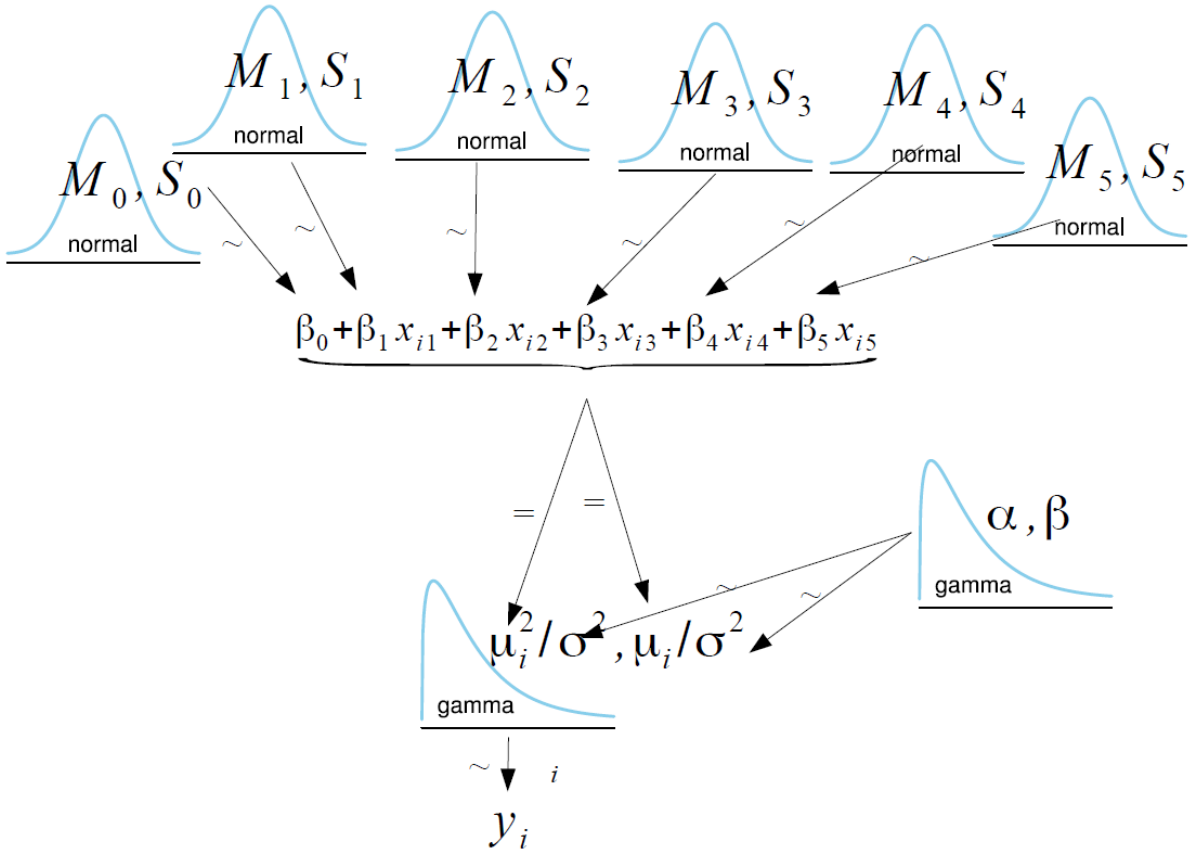
1. Y is the dependent variable, sales price in AUD ($100,000)
2. $\beta_0$ is the intercept of regression model
3. $X_1$ is the independent variable, area (land size in $m^2$ of the sold property), its corresponding slope $\beta_1$ shows the effect of unit change in area on the sales price.
4. $X_2$ is the independent variable, bedrooms (the number of bedrooms of the sold property), its corresponding slope $\beta_2$ shows the effect of unit change in number of bedrooms on the sales price.
5. $X_3$ is the independent variable, bathrooms (the number of bathrooms of the sold property), its corresponding slope $\beta_3$ shows the effect of unit change in number of bathrooms on the sales price.
6. $X_4$ is the independent variable, carparks (the number of carparks of the sold property), its corresponding slope $\beta_4$ shows the effect of unit change in number of carparks on the sales price.
7. $X_5$ is the independent variable, property type (0: House, 1: Unit), its corresponding slope $\beta_5$ shows the effect of the type change from house to unit on the sales price.
8. $\varepsilon$ is the error of the regression, the typical distance that the data points fall from the regression line.

Since we have picturized both gamma and exponential distributions are possible fit to the dataset, we can use 2 different scenarios to model the multiple linear regression:

## Scenario 1 – dependent variable and error of the regression model with gamma distribution

Figure 6 illustrates the hierarchical dependencies of the Bayesian regression model with gamma distribution modelling the dependent variable and error of the regression. We would call this as gamma model in short from this point forward.

*Figure 6 - Model diagram for multiple linear regression with gamma distributed $y_i$ and ε*




This model diagram is explained as follows:

1. $Y_i \sim \text{Gamma}(\mu_i^2/\sigma^2, \mu_i/\sigma^2)$: $Y_i$ , the observed data point of the dependent variable, sales price in AUD ($100,000) is a gamma-distributed random value, with 2 parameters, mean ($\mu_i$ ) and variance($\sigma^2$) . Correspondingly, the error (ε) of the regression model is also distributed with this gamma distribution.
2. $\mu_i = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5$: $\mu_i$ , the central tendency of the gamma distribution in point 1, $Y_i \sim \text{Gamma}(\mu_i^2/\sigma^2, \mu_i/\sigma^2)$. $\mu_i$ is transformed to the regression model $\beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5$.

3. $\sigma^2 \sim \text{Gamma}(\alpha, \beta)$: $\sigma^2$, variance of the gamma distribution in point 1, $Y_i \sim \text{Gamma}(\mu_i^2/\sigma^2, \mu_i/\sigma^2)$. $\sigma^2$ is modelled with another gamma distribution where we have set a low concentrated prior: $\alpha$=0.01 and $\beta$=0.01 (Refer to [A1] for detailed explanation), as in regression context, $\sigma^2$ does not have any meaningful prior.
4. $\beta_0 \sim \text{Normal}(M_0, S_0)$: $\beta_0$ , intercept of the regression model, is modelled with a normal distribution where we have set a board normal prior: $M_0 = 0$, $S_0 = 4$ (Refer to [A1] for detailed explanation), as in regression context, intercept does not have any meaningful prior.
5. $\beta_1 \sim \text{Normal}(M_1, S_1)$: $\beta_1$ , slope of the independent variable, area, is modelled with a normal distribution, it is given with the prior information that every $m^2$ increase in land size increases the sales price by AUD90. (**very strong expert knowledge**) Refer to [A1], we would put $M_1$ and $S_1$ as below:

| | $M_1$ | $S_1$ |
|---|---|---|
| informative prior | 90/100000 | 0.01 |

6. $\beta_2 \sim \text{Normal}(M_2, S_2)$: $\beta_2$ , slope of the independent variable, bedrooms, is modelled with a normal distribution, it is given with the prior information that every additional bedroom increases the sales price by 100,000 AUD. **(weak expert knowledge)** Refer to [A1], we would put $M_2$ and $S_2$ as below :

| | $M_2$ | $S_2$ |
|---|---|---|
| informative prior | 1 | 2 |

7. $\beta_3 \sim \text{Normal}(M_3, S_3)$: $\beta_3$ , slope of the independent variable, bathrooms, is modelled with a normal distribution, as there is no expert knowledge on the bathroom, I would set it with a board normal prior : $M_3 = 0$, $S_3 = 4$.
8. $\beta_4 \sim \text{Normal}(M_4, S_4)$: $\beta_4$ , slope of the independent variable, carparks, is modelled with a normal distribution, it is given with the prior information that every additional car space increases the sales price by 120,000 AUD. **(strong expert knowledge)** Refer to [A1], we would put $M_4$ and $S_4$ as below :

| | $M_4$ | $S_4$ |
|---|---|---|
| informative prior | 1.2 | 0.1 |

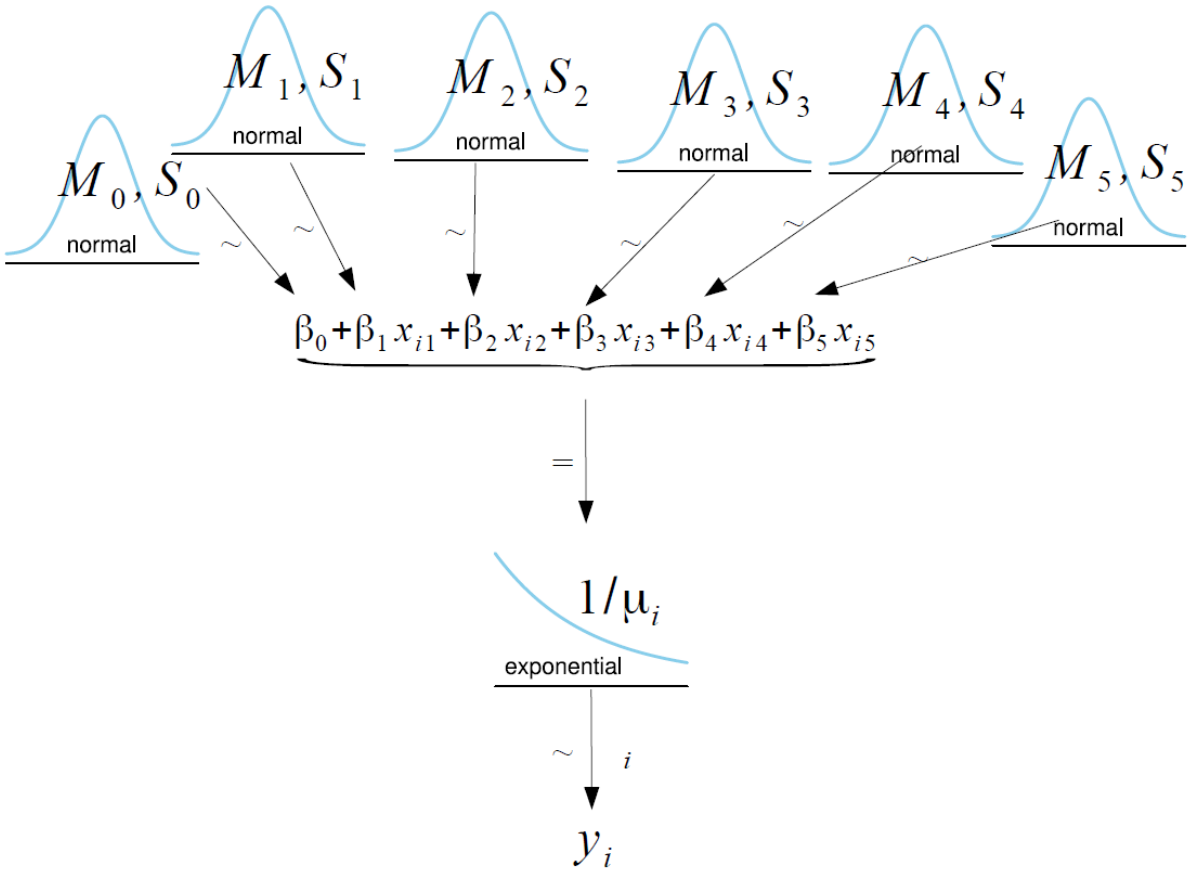
9. $\beta_5 \sim \text{Normal}(M_5, S_5)$: $\beta_5$ , slope of the independent variable, property type (0: House, 1: Unit), is modelled with a normal distribution, it is given with the prior information that the sales price of a unit will be 150,000 AUD less than a house on average. **(very strong expert knowledge)**. Refer to [A1], we would put $M_5$ and $S_5$ as below :

| | $M_5$ | $S_5$ |
|---|---|---|
| informative prior | -1.5 | 0.01 |

## Scenario 2 – dependent variable and error of the regression model with exponential distribution

Figure 7 illustrates the hierarchical dependencies of the Bayesian regression model with exponential distribution modelling the dependent variable and error of the regression. We would call this as exponential model in short from this point forward.

*Figure 7 - Model diagram for multiple linear regression with exponential distributed yi and ε*




This model diagram is explained as follows:

1. $Y_i \sim Exp(1/\mu_i^2)$: $Y_i$ , the observed data point of the dependent variable, sales price in AUD ($100,000) is an exponential-distributed random value, with one parameter, mean ($\mu_i$ ) . Correspondingly, the error ($\varepsilon$) of the regression model is also distributed with this exponential distribution.
2. $\mu_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$: $\mu_i$ , the central tendency of the exponential distribution in point 1, $Y_i \sim Exp(1/\mu_i^2)$. $\mu_i$ is transformed to the regression model $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$.

The rest would be exactly the same as point 4 to point 9 in the gamma model.

# 3. Results from JAGS on the sample dataset

After running MCMC in JAGS (refer to the coding in [A6] and [A7]), we would present the results we have obtained from both scenarios in this section. We would then dive in further to compare the goodness of fit and efficiency on these two models. Diagnostic checks are done before we could review all these results. (Please refer to the section of [Representativeness, accuracy and efficiency] for further details.

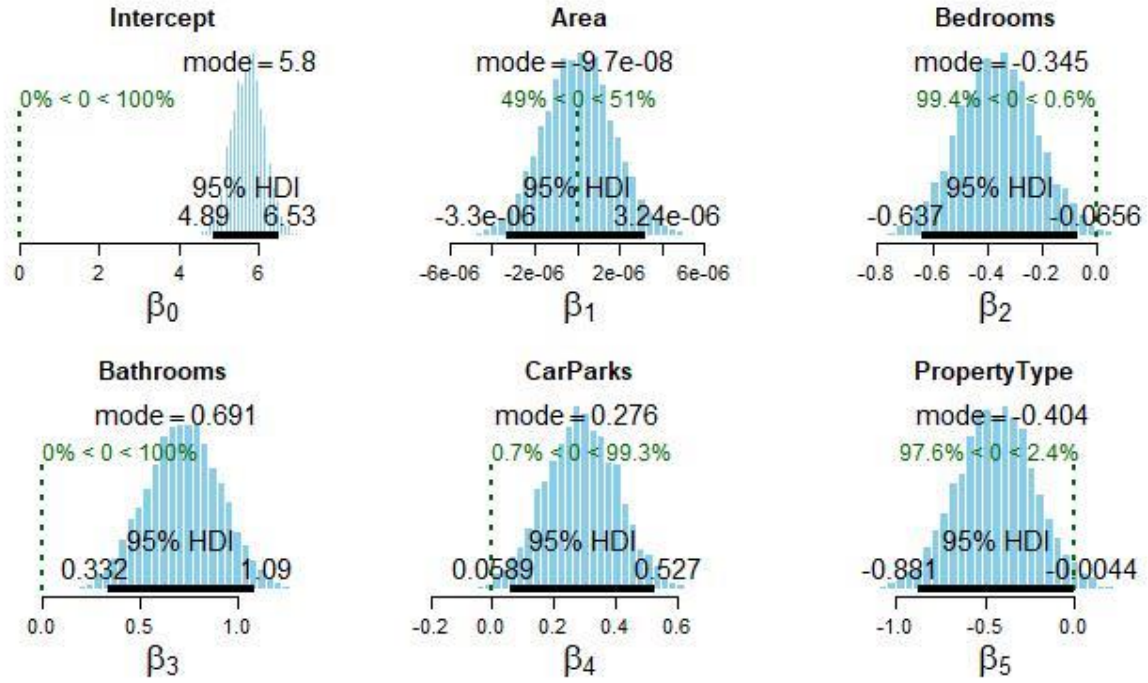
## Scenario 1 – Gamma model

### Posterior distribution plots on Betas (the coefficient of dependent variable)

Figure 8 illustrates that

1) $\beta_1$ 's HDI captures 0 right in the middle of the distribution, this coefficient is insignificant in the model.
2) $\beta_2$ 's HDI does not capture 0, only 0.6% of $\beta_2$ 's posterior distribution will be greater than 0, 99.4% will be less than 0, $\beta_2$ is considered to be significant.
3) $\beta_3$ 's distribution does not capture 0, $\beta_3$ is significant.
4) $\beta_4$ 's HDI does not capture 0, only 0.7% of $\beta_2$ 's posterior distribution will be less than 0, 99.3% will be greater than 0, $\beta_4$ is considered to be significant.
5) $\beta_5$ 's HDI does not capture 0, only 2.4% of $\beta_2$ 's posterior distribution will be greater than 0, 97.6% will be less than 0, $\beta_5$ is considered to be significant.

*Figure 8 - Posterior distribution of Betas in multiple linear regression with gamma distributed yi and ε (with informative prior)*




As the above beta distributions are not skewed, we can use the mode of the distribution to represent the coefficient of our regression model in the following formula:

$$Y = 5.8 - 0.345X_2 + 0.691X_3 + 0.276X_4 - 0.404X_5$$

Where

1) Y is the dependent variable, sales price in AUD ($100,000)
2) $X_1$ is the independent variable, area (land size in $m^2$ of the sold property)
3) $X_2$ is the independent variable, number of bedrooms
4) $X_3$ is the independent variable, number of bathrooms
5) $X_4$ is the independent variable, number of carparks
6) $X_5$ is the independent variable, property type, 0 indicates the property is a House, 1 indicates the property is a Unit.

This model could also be explained as:

1) Every additional number of bedrooms would result a decrease of 0.345 (in AUD 100,000) in the sales price.

2) Every additional number of bathrooms would result an increase of 0.691 (in AUD 100,000) in the sales price.
3) Every additional number of car space would result an increase of 0.276 (in AUD 100,000) in the sales price.
4) If the property is a unit, the sale price will be 0.404 (in AUD 100,000) less than a house on average

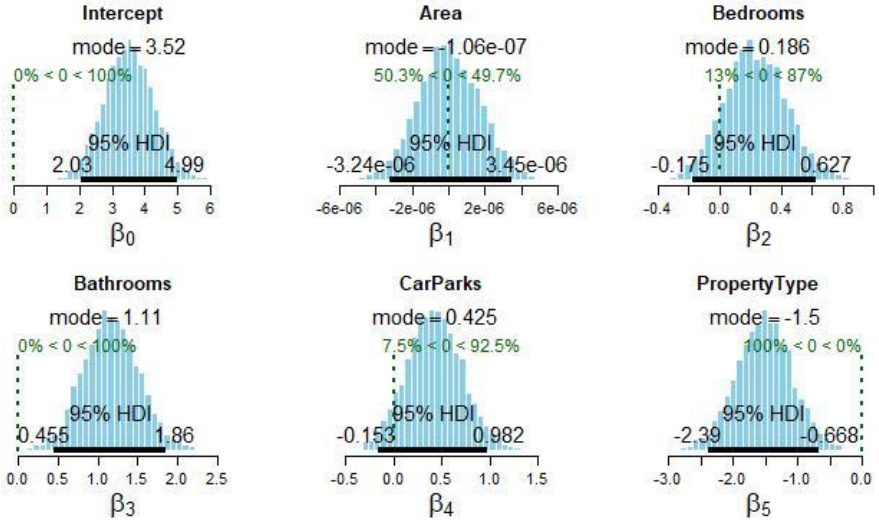
## Scenario 2 – Exponential model

### Posterior distribution plots on Betas (the coefficient of dependent variable)

Figure 9 illustrates that

1) $\beta_1$ 's HDI captures 0 right in the middle of the distribution, this coefficient is insignificant in the model.
2) $\beta_2$ 's HDI captures 0, but only 13% of $\beta_2$ 's posterior distribution will be less than 0, 87% will be greater than 0, $\beta_2$ is still considered to be significant.
3) $\beta_3$ 's distribution does not capture 0, $\beta_3$ is significant.
4) $\beta_4$ 's HDI captures 0, but only 7.5% of $\beta_4$ 's posterior distribution will be less than 0, 92.5% will be greater than 0, $\beta_4$ is considered to be significant.
5) $\beta_5$ 's distribution does not capture 0, $\beta_5$ is significant.

*Figure 9 - Posterior distribution of Betas in multiple linear regression with exponential distributed yi and ε (with informative prior)*



As the above betas distributions are not skewed, we can use the mode of the distribution to represent the coefficient of our regression model in the following formula:

$$Y = 3.52 + 0.186X_2 + 1.11X_3 + 0.425X_4 - 1.5X_5$$

Where

1) Y is the dependent variable, sales price in AUD ($100,000)
2) $X_1$ is the independent variable, area (land size in $m^2$ of the sold property)
3) $X_2$ is the independent variable, number of bedrooms
4) $X_3$ is the independent variable, number of bathrooms
5) $X_4$ is the independent variable, number of carparks
6) $X_5$ is the independent variable, property type, 0 indicates the property is a House, 1 indicates the property is a Unit.

This model could also be explained as:

1) Every additional number of bedrooms would result an increase of 0.186 (in AUD 100,000) in the sales price.
2) Every additional number of bathrooms would result an increase of 1.11 (in AUD 100,000) in the sales price.
3) Every additional number of car space would result an increase of 0.425 (in AUD 100,000) in the sales price.
4) If the property is a unit, the sale price will be 1.5 (in AUD 100,000) less than a house on average

## Comparison of the regression models

### Predictions and $R^2$

After we have generated the regression models, we can use them to predict new entries where independent variables are given as follows:

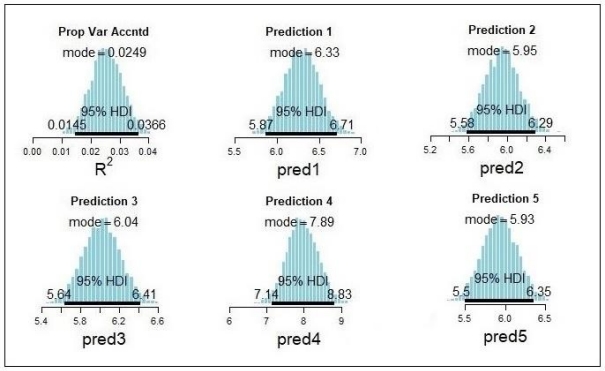
*Table 7 - Prediction entries*

| Prediction | Area | Bedrooms | Bathrooms | CarParks | Property Type |
|---|---|---|---|---|---|
| 1 | 600 | 2 | 2 | 1 | Unit |
| 2 | 800 | 3 | 1 | 2 | House |
| 3 | 1500 | 2 | 1 | 1 | House |
| 4 | 2500 | 5 | 4 | 4 | House |
| 5 | 250 | 3 | 2 | 1 | Unit |

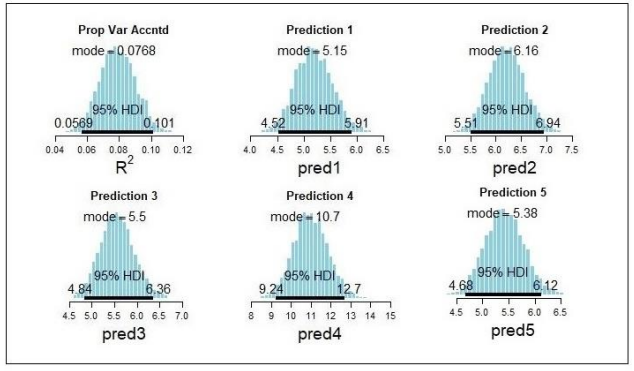
Figure 10 shows us the posterior distributions of the above predictions. For example, prediction 1 of the gamma distributed model could be read as within 95% confidence, prediction 1 is in the range of 5.97 to 6.71 (in 100,000 AUD), its mode (most frequent central tendency) is at 6.33 (in 100,000 AUD).

*Figure 10 – $R^2$ and Predictions from different models*

*Models of gamma distributed $y_i$ and $\varepsilon$*




*Models of exponential distributed $y_i$ and $\varepsilon$*




The top left corner of Figure 10 indicates the mode of $R^2$ of the gamma model is at 0.0249 (with HDI between 0.0145 and 0.0366), which is lower than the exponential model (mode at 0.0768 with HDI between 0.0569 and 0.101), however these figures are all pretty low, which implies that only 2.49%

(gamma) and 7.68% (exponential) of the observed variation can be explained by these models. Thus, even exponential model has a slightly better $R^2$, both of these models have a weak linear relationship with the dependent variables.

If we apply the formula of the regression model, and substitute the values of the independent variables of the predictions accordingly (for example, prediction 1: $X_1 = 600$, $X_2 = 2$, $X_3 = 2$, $X_4 = 1$, $X_5 = 1$, apply to the formula of exponential model: $Y= 3.52 + 0*(600) + 0.186*(2) + 1.11*(2) + 0.425*(1) – 1.5*(1)$), we would find that the results (highlighted in yellow shown in Table 8) are very close to the central tendency (mode) of the posterior distribution in Figure 10. We found that with the same normal prior settings, exponential model evaluates:

1) prediction 4 higher than gamma model
2) prediction 1, 3 and 5 lower than gamma model
3) prediction 2 similarly to the gamma model

Exponential model ➔ $Y= 3.52 + 0.186X_2 + 1.11X_3 + 0.425X_4 – 1.5X_5$

Gamma model ➔ $Y = 5.8 – 0.345X_2 + 0.691X_3 + 0.276X_4 – 0.404X_5$

*Table 8 - Prediction results calculated from the regression model formula*

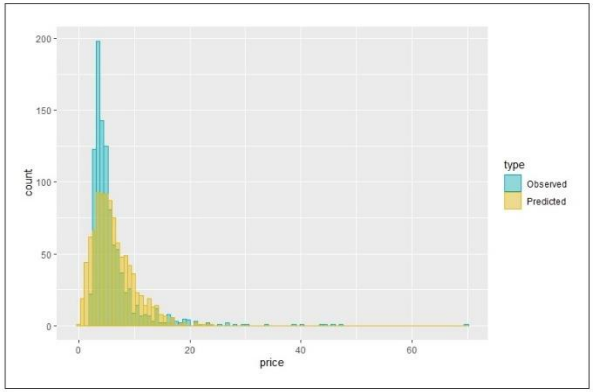
| | Betas | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Model | β0 | β1 | β2 | β3 | β4 | β5 | | |
| gamma | 5.8 | 0 | -0.345 | 0.691 | 0.276 | -0.404 | | |
| exponential | 3.52 | 0 | 0.186 | 1.11 | 0.425 | -1.5 | | |
| | | | | | | | Resutls (in 100,000 AUD) | |
| Predictions | Area | Bedrooms | Bathrooms | CarParks | PropertyType | | Gamma | Exponential |
| 1 | 600 | 2 | 2 | 1 | 1 | | 6.36 | 5.04 |
| 2 | 800 | 3 | 1 | 2 | 0 | | 6.01 | 6.04 |
| 3 | 1500 | 2 | 1 | 1 | 0 | | 6.08 | 5.43 |
| 4 | 2500 | 5 | 4 | 4 | 0 | | 7.94 | 10.59 |
| 5 | 250 | 3 | 2 | 1 | 1 | | 6.02 | 5.22 |

## Goodness of fit

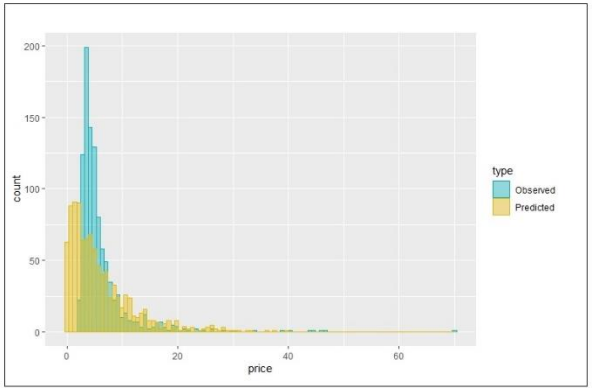
We then generated 1,000 random samples from the gamma or exponential distribution with the shape derived from the regression formula calculated by the coefficients (mode value of all betas, for gamma distribution, we need posterior distribution of variance too). We then compare these distribution to the sample observed data as shown in Figure 11, we could find that gamma model gives a better goodness of fit with the given prior information. Refer to [A9] for relevant coding.

*Figure 11 - Comparison of observe sample data and posterior distribution of $y_i$*

*Models of gamma distributed $y_i$ and $\varepsilon$*




*Models of exponential distributed $y_i$ and $\varepsilon$*




## MCMC run time comparison

However, exponential model runs much faster than gamma model as there is one less component for variance as shown in Figure 6 and Figure 7. Table 9 shows the time taken for running the sample and full dataset (which would be discussed in details in the section [Sample vs whole dataset]) on both models. In the sample dataset, elapsed time of the exponential model is 7.23 mins (434.19 seconds), while the gamma model is 23.89 mins (1433.89 seconds), which is 3 times slower than the exponential model. In the full dataset, elapsed time of the exponential model is 3.45 hours (12394.14 seconds), while the gamma model is 8.23 hours (29647.89 seconds), which is 2.4 times slower than the exponential model.

*Table 9 - Run time for different models with JAGS*

| | Exponential | | | Gamma | | |
|---|---|---|---|---|---|---|
| | user | system | Elapsed | user | system | Elapsed |
| Sample dataset of 1,000 records | 1.69 | 0.52 | 434.19 | 5.55 | 0.89 | 1433.89 |
| Full dataset of 10,000 records | 94.14 | 16.03 | 12394.14 | 194.53 | 66.16 | 29647.89 |

# 4. Comparison and Analysis

## Different parameter values on normal prior of the model

Recapture on the given prior knowledge:

1. every $m^2$ increase in land size increases the sales price by AUD90. (**very strong expert knowledge**)
2. every additional bedroom increases the sales price by 100,000 AUD. **(weak expert knowledge)**
3. No expert knowledge on bathrooms.
4. every additional car space increases the sales price by 120,000 AUD. **(strong expert knowledge)**
5. sales price of a unit will be 150,000 AUD less than a house on average. **(very strong expert knowledge)**.

We have set the following values for our normal prior:

*Table 10 - setting of normal prior for coefficients according to given knowledge on the dataset*

| | $M_1$ | $S_1$ | $M_2$ | $S_2$ | $M_3$ | $S_3$ | $M_4$ | $S_4$ | $M_5$ | $S_5$ |
|---|---|---|---|---|---|---|---|---|---|---|
| values | 90/100,000 | 0.01 | 1 | 2 | 0 | 4 | 1.2 | 0.1 | -1.5 | 0.01 |

$S_i$ is adjusted according to the degree of belief in the prior knowledge. In this section, we are investigating the impact on the regression model when different values are applied to $S_i$

First, we want to see how the regression model will look like without any prior knowledge (Trial 1). Then from Figure 8 and Figure 9, we have observed that $\beta_1$ is insignificant in both models, from an extensive search by applying different values to $S_1$, we found that we can turn $\beta_1$ to significant in the regression model (Trial 7). Similarly, we would also apply different values to $S_2$ (trial 13), $S_4$ (trial 14) and $S_5$ (trial 15) while retaining the same values for other parameters to find a more desired or significant $\beta_2$, $\beta_4$ and $\beta_5$ respectively. We have performed the following trials:

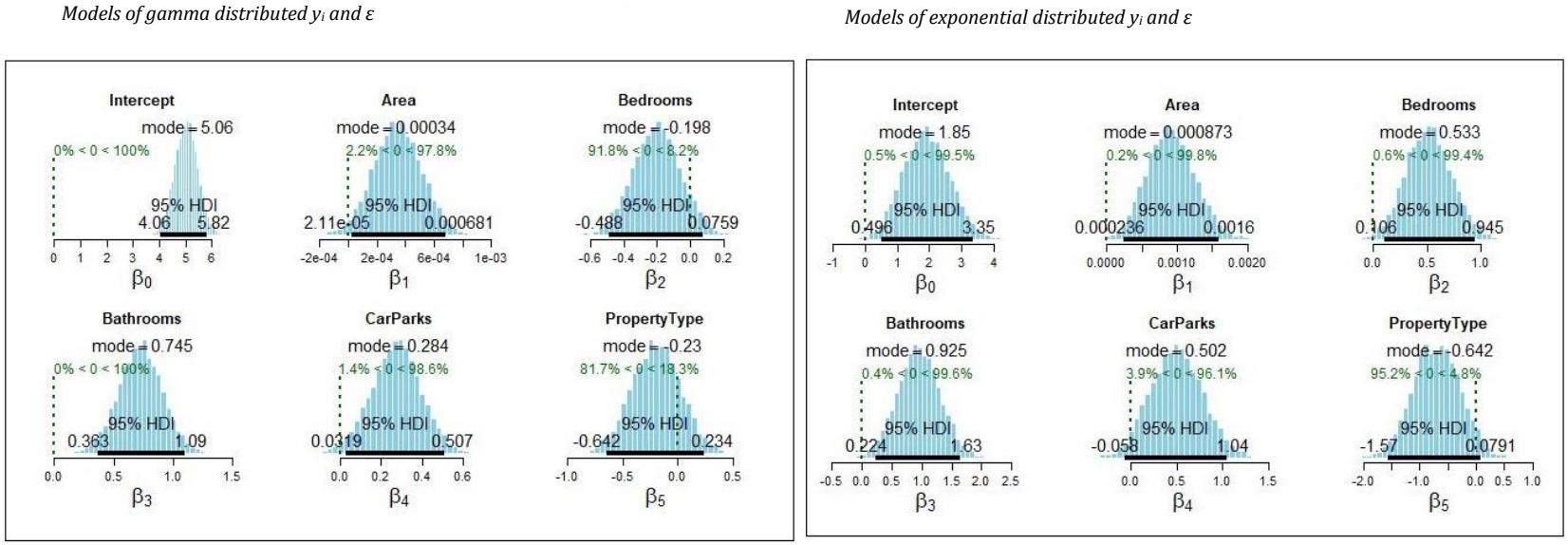
*Table 11 - Trials*

| | $M_1$ | $S_1$ | $M_2$ | $S_2$ | $M_3$ | $S_3$ | $M_4$ | $S_4$ | $M_5$ | $S_5$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Trial 1 | 0 | 4 | 0 | 4 | 0 | 4 | 0 | 4 | 0 | 4 |
| Trial 2 | 90/100,000 | 0.1 | 1 | 2 | 0 | 4 | 1.2 | 0.1 | -1.5 | 0.01 |
| Trial 3 | 90/100,000 | 1 | 1 | 2 | 0 | 4 | 1.2 | 0.1 | -1.5 | 0.01 |
| Trial 4 | 90/100,000 | 2 | 1 | 2 | 0 | 4 | 1.2 | 0.1 | -1.5 | 0.01 |
| Trial 5 | 90/100,000 | 4 | 1 | 2 | 0 | 4 | 1.2 | 0.1 | -1.5 | 0.01 |
| Trial 6 | 90/100,000 | 20 | 1 | 2 | 0 | 4 | 1.2 | 0.1 | -1.5 | 0.01 |
| Trial 7 | 90/100,000 | 100 | 1 | 2 | 0 | 4 | 1.2 | 0.1 | -1.5 | 0.01 |
| Trial 8* | 90/100,000 | 0.01 | 1 | 4 | 0 | 4 | 1.2 | 0.1 | -1.5 | 0.01 |
| Trial 9* | 90/100,000 | 0.01 | 1 | 20 | 0 | 4 | 1.2 | 0.1 | -1.5 | 0.01 |
| Trial 10* | 90/100,000 | 0.01 | 1 | 100 | 0 | 4 | 1.2 | 0.1 | -1.5 | 0.01 |
| Trial 11 | 90/100,000 | 0.01 | 1 | 0.5 | 0 | 4 | 1.2 | 0.1 | -1.5 | 0.01 |
| Trial 12 | 90/100,000 | 0.01 | 1 | 0.01 | 0 | 4 | 1.2 | 0.1 | -1.5 | 0.01 |
| Trial 13 | 90/100,000 | 0.01 | 1 | 0.001 | 0 | 4 | 1.2 | 0.1 | -1.5 | 0.01 |
| Trial 14 | 90/100,000 | 0.01 | 1 | 2 | 0 | 4 | 1.2 | 0.01 | -1.5 | 0.01 |
| Trial 15 | 90/100,000 | 0.01 | 1 | 2 | 0 | 4 | 1.2 | 0.1 | -1.5 | 0.005 |

N.B. trial 8 – 10 were only performed on the exponential model (at that time I was searching for the direction to turn $\beta_2$ more significant, after I found that we need to decrease $S_2$ instead of increasing $S_2$ to turn $\beta_2$ more significant, I didn't perform those trials on increasing $S_2$ for gamma model). There are interesting differences in trial 1, 7 and 13, we would further discuss these trials.

First, lets look at trial 1, refer to Figure 12, without any prior information, in the gamma model, $\beta_1$ - $\beta_4$ are all considered to be significant (0 are all at the edge of the posterior distribution of betas), I would take $\beta_5$ as insignificant, as it is not far away from the middle of the posterior distribution. For the exponential model, all betas are all considered to be significant (as 0 are all at the edge of the posterior distribution of betas).

*Figure 12 - Trial 1*




*As the posterior distribution on betas are not skewed, we could take the mode as the coefficients of the regression formula. If we substitute the values of the independent variables of the predictions, and apply accordingly to both the original (given prior knowledge) and newly derived (trial 1) regression formula (refers to Table 12 and*

Table 13, we would find that there are only small differences for predictions 1-3 and 5. For prediction 4, both models would project a higher sales price in trial 1. The results (highlighted in yellow) are very close to the central tendency (mode) of the posterior distribution in Table 23.

Gamma (given prior info): $Y = 5.8 - 0.345X_2 + 0.691X_3 + 0.276X_4 - 0.404X_5$

*Table 12 - Prediction results calculated from the regression formula (given prior info compare with trial 1) for gamma model*

| | Betas | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Model | β0 | β1 | β2 | β3 | β4 | β5 | | |
| gamma (given prior info) | 5.8 | 0 | -0.345 | 0.691 | 0.276 | -0.404 | | |
| gamma (trial 1) | 5.06 | 0.00034 | -0.198 | 0.745 | 0.284 | 0 | | |
| | | | | | | | Results (in 100,000 AUD) | |
| Predictions | Area | Bedrooms | Bathrooms | CarParks | PropertyType | | Given Prior | Trial 1 |
| 1 | 600 | 2 | 2 | 1 | 1 | | 6.36 | 6.64 |
| 2 | 800 | 3 | 1 | 2 | 0 | | 6.01 | 6.05 |
| 3 | 1500 | 2 | 1 | 1 | 0 | | 6.08 | 6.20 |
| 4 | 2500 | 5 | 4 | 4 | 0 | | 7.94 | 9.04 |
| 5 | 250 | 3 | 2 | 1 | 1 | | 6.02 | 6.33 |

Gamma (trial 1): $Y = 5.06 + 0.00034\ X_1 - 0.198X_2 + 0.745X_3 + 0.284X_4$

Exponential (given prior info) ➔ $Y = 3.52 + 0.186X_2 + 1.11X_3 + 0.425X_4 - 1.5X_5$

*Table 13 - Prediction results calculated from regression formula (given prior info compare with trial 1) for exponential model*

| | Betas | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Model | β0 | β1 | β2 | β3 | β4 | β5 | | |
| exponential (given prior info) | 3.52 | 0 | 0.186 | 1.11 | 0.425 | -1.5 | | |
| exponential (trial 1) | 1.85 | 0.000873 | 0.533 | 0.925 | 0.502 | -0.642 | | |
| | | | | | | | Results (in 100,000 AUD) | |
| Predictions | Area | Bedrooms | Bathrooms | CarParks | PropertyType | | Given Prior | Trial 1 |
| 1 | 600 | 2 | 2 | 1 | 1 | | 5.04 | 5.15 |
| 2 | 800 | 3 | 1 | 2 | 0 | | 6.04 | 6.08 |
| 3 | 1500 | 2 | 1 | 1 | 0 | | 5.43 | 5.65 |
| 4 | 2500 | 5 | 4 | 4 | 0 | | 10.59 | 12.41 |
| 5 | 250 | 3 | 2 | 1 | 1 | | 5.22 | 5.38 |

Exponential (trial 1) ➔ $Y = 1.85 + 0.000873\ X_1 + 0.533X_2 + 0.925X_3 + 0.502X_4 - 0.642X_5$

With the given prior knowledge, $\beta_1$ are both insignificant in gamma or exponential model, however, with extensive search, we found that if we retain the same settings for other parameters and only assign a board variance to normal prior of $\beta_1$, it will become significant in both models. In fact, if we refer to Figure 13, all betas turn into significant. Adding the regression formula of these trials onto Table 14 and Table 15, we would find there are only small differences for all the predictions.

*Table 14 - setting of normal prior for coefficients according to given knowledge on the dataset*

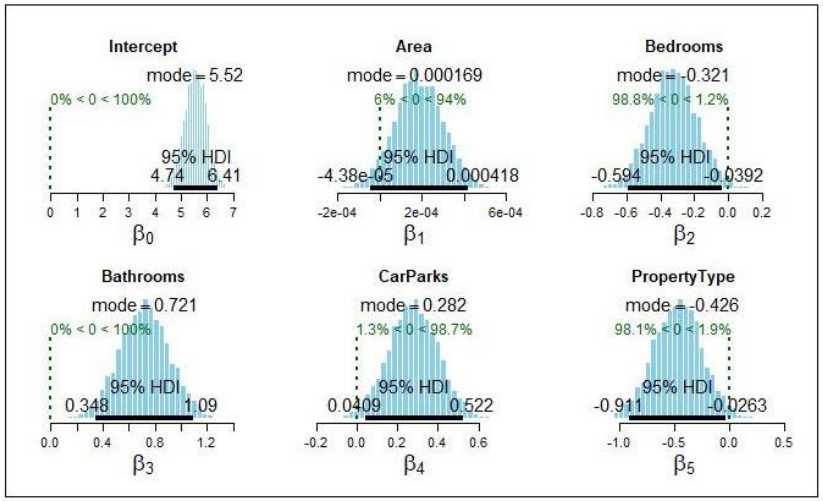
| | $M_1$ | $S_1$ | $M_2$ | $S_2$ | $M_3$ | $S_3$ | $M_4$ | $S_4$ | $M_5$ | $S_5$ |
|---|---|---|---|---|---|---|---|---|---|---|
| values | 90/100,000 | 0.01 | 1 | 2 | 0 | 4 | 1.2 | 0.1 | -1.5 | 0.01 |

*Table 15 - apply a board variance on normal prior of $\beta_1$*

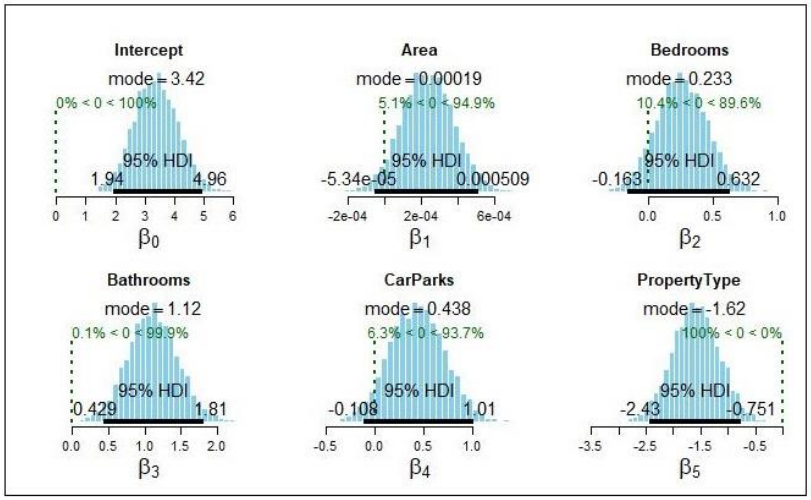
| | $M_1$ | $S_1$ | $M_2$ | $S_2$ | $M_3$ | $S_3$ | $M_4$ | $S_4$ | $M_5$ | $S_5$ |
|---|---|---|---|---|---|---|---|---|---|---|
| values | 90/100,000 | 100 | 1 | 2 | 0 | 4 | 1.2 | 0.1 | -1.5 | 0.01 |

*Figure 13 - trial 7*

*Models of gamma distributed $y_i$ and $\varepsilon$*




*Models of exponential distributed $y_i$ and $\varepsilon$*



Gamma (trial 1): $Y = 5.06 + 0.00034\,X_1 - 0.198X_2 + 0.745X_3 + 0.284X_4$

Gamma (given prior info): $Y = 5.8 - 0.345X_2 + 0.691X_3 + 0.276X_4 - 0.404X_5$

*Table 16 - Prediction results calculated from the regression formula (given prior info compare with trial 1 and 7) for gamma model*

| | Betas | | | | | |
|---|---|---|---|---|---|---|
| Model | β0 | β1 | β2 | β3 | β4 | β5 |
| gamma (given prior info) | 5.8 | 0 | -0.345 | 0.691 | 0.276 | -0.404 |
| gamma (trial 1) | 5.06 | 0.00034 | -0.198 | 0.745 | 0.284 | 0 |
| gamma (trial 7) | 5.52 | 0.000169 | -0.321 | 0.721 | 0.282 | -0.426 |

| | | | | | | Resutls (in 100,000 AUD) | | |
|---|---|---|---|---|---|---|---|---|
| Predictions | Area | Bedrooms | Bathrooms | CarParks | PropertyType | Given Prior | Trial 1 | Trial 7 |
| 1 | 600 | 2 | 2 | 1 | 1 | 6.36 | 6.64 | 6.28 |
| 2 | 800 | 3 | 1 | 2 | 0 | 6.01 | 6.05 | 5.98 |
| 3 | 1500 | 2 | 1 | 1 | 0 | 6.08 | 6.20 | 6.13 |
| 4 | 2500 | 5 | 4 | 4 | 0 | 7.94 | 9.04 | 8.35 |
| 5 | 250 | 3 | 2 | 1 | 1 | 6.02 | 6.33 | 5.90 |

Gamma (trial 7): $Y = 5.52 + 0.000169\,X_1 - 0.321X_2 + 0.721X_3 + 0.282X_4 - 0.426X_4$

Exponential (trial 1) ➔ $Y = 1.85 + 0.000873\,X_1 + 0.533X_2 + 0.925X_3 + 0.502X_4 - 0.642X_5$

Exponential (given prior info) ➔ $Y = 3.52 + 0.186X_2 + 1.11X_3 + 0.425X_4 - 1.5X_5$

*Table 17 - Prediction results calculated from regression formula (given prior info compare with trial 1 and 7) for exponential model*

| | Betas | | | | | |
|---|---|---|---|---|---|---|
| Model | β0 | β1 | β2 | β3 | β4 | β5 |
| exponential (given prior info) | 3.52 | 0 | 0.186 | 1.11 | 0.425 | -1.5 |
| exponential (trial 1) | 1.85 | 0.000873 | 0.533 | 0.925 | 0.502 | -0.642 |
| exponential (trial 7) | 3.42 | 0.00019 | 0.233 | 1.12 | 0.438 | -1.62 |

| | | | | | | Resutls (in 100,000 AUD) | | |
|---|---|---|---|---|---|---|---|---|
| Predictions | Area | Bedrooms | Bathrooms | CarParks | PropertyType | Given Prior | Trial 1 | Trial 7 |
| 1 | 600 | 2 | 2 | 1 | 1 | 5.04 | 5.15 | 5.06 |
| 2 | 800 | 3 | 1 | 2 | 0 | 6.04 | 6.08 | 6.27 |
| 3 | 1500 | 2 | 1 | 1 | 0 | 5.43 | 5.65 | 5.73 |
| 4 | 2500 | 5 | 4 | 4 | 0 | 10.59 | 12.41 | 11.29 |
| 5 | 250 | 3 | 2 | 1 | 1 | 5.22 | 5.38 | 5.22 |

Exponential (trial 7) ➔ $Y = 3.42 + 0.00019\,X_1 + 0.233X_2 + 1.12X_3 + 0.438X_4 - 1.62X_5$

Similarly, With the given prior knowledge, $\beta_2$ is significant, but it is negative in the gamma model, with another extensive search, we found that if we retain the same settings for other parameters and assign a small variance the to normal prior of $\beta_2$, it will become positive and significant in the gamma model, and make $\beta_2$ further positive in the exponential model as shown in **Error! Reference source not found.**. Adding the regression formula of these trials onto Table 20 and Table 21, we would find there are only small differences for all the predictions, except prediction 4 (in exponential model) and 5 (in both models) have been evaluated with higher sales price, as $\beta_2$ coefficient is comparatively high and $\beta_5$ is positive compare to other trials.

*Table 18 - setting of normal prior for coefficients according to given knowledge on the dataset*

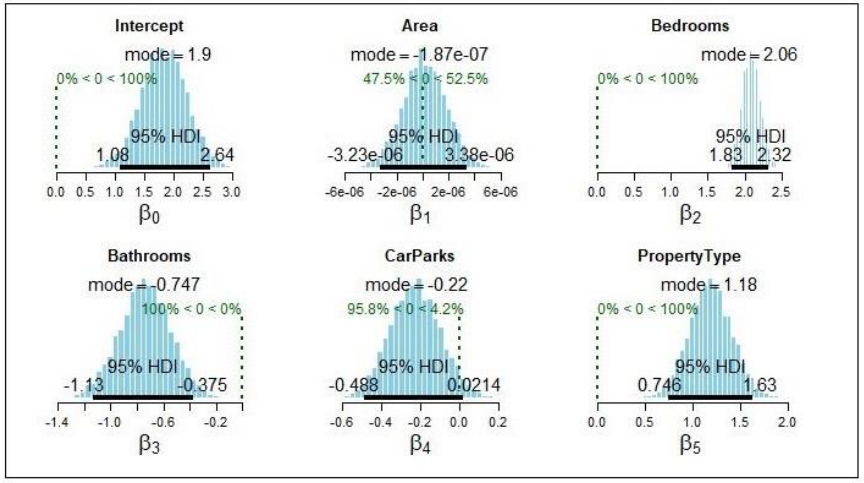
| | $M_1$ | $S_1$ | $M_2$ | $S_2$ | $M_3$ | $S_3$ | $M_4$ | $S_4$ | $M_5$ | $S_5$ |
|---|---|---|---|---|---|---|---|---|---|---|
| values | 90/100,000 | 0.01 | 1 | 2 | 0 | 4 | 1.2 | 0.1 | -1.5 | 0.01 |

*Table 19 - apply a board variance on normal prior of $\beta_1$*

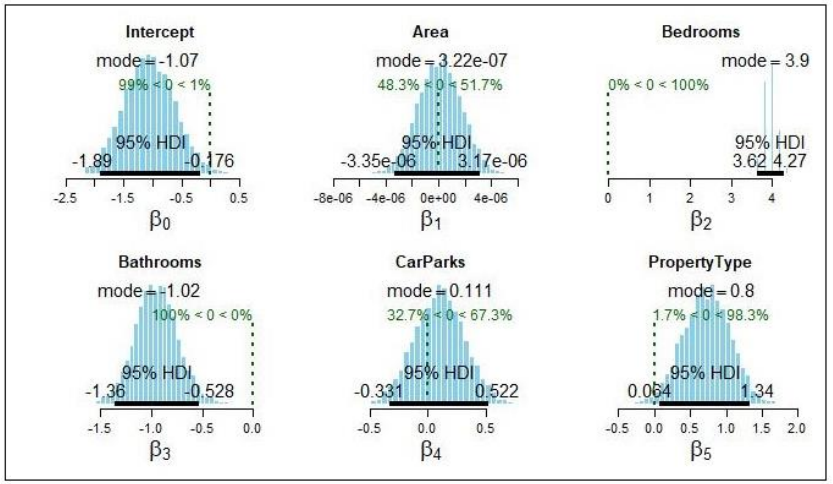
| | $M_1$ | $S_1$ | $M_2$ | $S_2$ | $M_3$ | $S_3$ | $M_4$ | $S_4$ | $M_5$ | $S_5$ |
|---|---|---|---|---|---|---|---|---|---|---|
| values | 90/100,000 | 0.01 | 1 | 0.001 | 0 | 4 | 1.2 | 0.1 | -1.5 | 0.01 |

*Figure 14 - trial 13*

*Models of gamma distributed $y_i$ and $\varepsilon$*




*Models of exponential distributed $y_i$ and $\varepsilon$*



Gamma (trial 1): $Y = 5.06 + 0.00034\,X_1 - 0.198X_2 + 0.745X_3 + 0.284X_4$

Gamma (given prior info): $Y = 5.8 - 0.345X_2 + 0.691X_3 + 0.276X_4 - 0.404X_5$

*Table 20 - Prediction results calculated from the regression formula (given prior info compare with trial 1, 7 and 13) for gamma model*

| | Betas | | | | | |
|---|---|---|---|---|---|---|
| Model | β0 | β1 | β2 | β3 | β4 | β5 |
| gamma (given prior info) | 5.8 | 0 | -0.345 | 0.691 | 0.276 | -0.404 |
| gamma (trial 1) | 5.06 | 0.00034 | -0.198 | 0.745 | 0.284 | 0 |
| gamma (trial 7) | 5.52 | 0.000169 | -0.321 | 0.721 | 0.282 | -0.426 |
| gamma (trial 13) | 1.9 | 0 | 2.06 | -0.747 | -0.22 | 1.18 |

| | | | | | | Resutls (in 100,000 AUD) | | | |
|---|---|---|---|---|---|---|---|---|---|
| Predictions | Area | Bedrooms | Bathrooms | CarParks | PropertyType | Given Prior | Trial 1 | Trial 7 | Trial 13 |
| 1 | 600 | 2 | 2 | 1 | 1 | 6.36 | 6.64 | 6.28 | 5.49 |
| 2 | 800 | 3 | 1 | 2 | 0 | 6.01 | 6.05 | 5.98 | 6.89 |
| 3 | 1500 | 2 | 1 | 1 | 0 | 6.08 | 6.20 | 6.13 | 5.05 |
| 4 | 2500 | 5 | 4 | 4 | 0 | 7.94 | 9.04 | 8.35 | 8.33 |
| 5 | 250 | 3 | 2 | 1 | 1 | 6.02 | 6.33 | 5.90 | 7.55 |

Gamma (trial 7): $Y = 5.52 + 0.000169\,X_1 - 0.321X_2 + 0.721X_3 + 0.282X_4 - 0.426X_4$

Gamma (trial 13): $Y = 1.9 + 2.06X_2 - 0.747X_3 - 0.22X_4 + 1.18X_5$

Exponential (trial 1) ➔ $Y = 1.85 + 0.000873\,X_1 + 0.533X_2 + 0.925X_3 + 0.502X_4 - 0.642X_5$

Exponential (given prior info) ➔ $Y = 3.52 + 0.186X_2 + 1.11X_3 + 0.425X_4 - 1.5X_5$

*Table 21 - Prediction results calculated from regression formula (given prior info compare with trial 1, 7 and 13) for exponential model*

| | Betas | | | | | |
|---|---|---|---|---|---|---|
| Model | β0 | β1 | β2 | β3 | β4 | β5 |
| exponential (given prior info) | 3.52 | 0 | 0.186 | 1.11 | 0.425 | -1.5 |
| exponential (trial 1) | 1.85 | 0.000873 | 0.533 | 0.925 | 0.502 | -0.642 |
| exponential (trial 7) | 3.42 | 0.00019 | 0.233 | 1.12 | 0.438 | -1.62 |
| exponential (trial 13) | -1.07 | 0 | 3.96 | -1.02 | 0.111 | 0.8 |

| | | | | | | Resutls (in 100,000 AUD) | | | |
|---|---|---|---|---|---|---|---|---|---|
| Predictions | Area | Bedrooms | Bathrooms | CarParks | PropertyType | Given Prior | Trial 1 | Trial 7 | Trial 13 |
| 1 | 600 | 2 | 2 | 1 | 1 | 5.04 | 5.15 | 5.06 | 5.72 |
| 2 | 800 | 3 | 1 | 2 | 0 | 6.04 | 6.08 | 6.27 | 10.01 |
| 3 | 1500 | 2 | 1 | 1 | 0 | 5.43 | 5.65 | 5.73 | 5.94 |
| 4 | 2500 | 5 | 4 | 4 | 0 | 10.59 | 12.41 | 11.29 | 15.09 |
| 5 | 250 | 3 | 2 | 1 | 1 | 5.22 | 5.38 | 5.22 | 9.68 |

Exponential (trial 7) ➔ $Y = 3.42 + 0.00019\,X_1 + 0.233X_2 + 1.12X_3 + 0.438X_4 - 1.62X_5$

Exponential (trial 13) ➔ $Y = -1.07 + 3.96X_2 - 1.02X_3 + 0.111X_4 + 0.8X_5$

From the above calculations, we realized that the significance of the coefficients is altered by the variance of its corresponding normal prior in Bayesian regression model. Change of one coefficient (e.g. $\beta_2$ in trial 13) would impact other coefficient values in multiple linear regression. Predictions will change subsequentially, impact are more obvious on predictions with higher or lower than average values in particular independent variables of the affected coefficients.

Table 22 and Table 23 record the mode and HDI for the coefficients and predictions for all the trials. The difference for the predictions is small in all the trials, except:

1) Trial 13, prediction 4 and 5 would be evaluated with higher sales price as the coefficient of bedrooms and property type (for gamma model) become dominant
2) Trial 14, prediction 4 would be evaluated with higher sales price as the coefficient of car space become dominant (for exponential model)

*Table 22 - Mode and HDI limits of coefficients for each trial*

| | | Beta 0 | | | Beta 1 | | | Beta 2 | | | Beta 3 | | | Beta 4 | | | Beta 5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | trials | Mode | HDI Low | HDI High | Mode | HDI Low | HDI High | Mode | HDI Low | HDI High | Mode | HDI Low | HDI High | Mode | HDI Low | HDI High | Mode | HDI Low | HDI High |
| Exponential | given prior info | 3.52 | 2.03 | 4.99 | -0.0000001 | -0.0000032 | 0.0000034 | 0.19 | -0.17 | 0.63 | 1.11 | 0.46 | 1.86 | 0.43 | -0.15 | 0.98 | -1.50 | -2.39 | -0.67 |
| | 1 | 1.85 | 0.50 | 3.35 | 0.0008730 | 0.0002363 | 0.0016008 | 0.53 | 0.11 | 0.95 | 0.93 | 0.22 | 1.63 | 0.50 | -0.06 | 1.04 | -0.64 | -1.57 | 0.08 |
| | 2 | 3.53 | 1.92 | 4.97 | 0.0000001 | -0.0000099 | 0.0000109 | 0.22 | -0.17 | 0.63 | 1.17 | 0.44 | 1.82 | 0.40 | -0.15 | 0.97 | -1.46 | -2.40 | -0.67 |
| | 3 | 3.59 | 2.06 | 5.09 | 0.0000035 | -0.0000301 | 0.0000361 | 0.24 | -0.17 | 0.63 | 1.18 | 0.47 | 1.85 | 0.43 | -0.12 | 0.99 | -1.51 | -2.38 | -0.69 |
| | 4 | 3.46 | 1.97 | 4.99 | 0.0000044 | -0.0000389 | 0.0000526 | 0.23 | -0.18 | 0.62 | 1.17 | 0.48 | 1.89 | 0.36 | -0.14 | 0.98 | -1.55 | -2.41 | -0.69 |
| | 5 | 3.30 | 2.02 | 5.02 | 0.0000078 | -0.0000553 | 0.0000791 | 0.23 | -0.13 | 0.64 | 1.19 | 0.48 | 1.84 | 0.38 | -0.14 | 0.99 | -1.56 | -2.40 | -0.67 |
| | 6 | 3.55 | 2.00 | 4.98 | 0.0000668 | -0.0000798 | 0.0002037 | 0.22 | -0.17 | 0.63 | 1.11 | 0.46 | 1.82 | 0.42 | -0.15 | 0.96 | -1.49 | -2.39 | -0.71 |
| | 7 | 3.42 | 1.94 | 4.96 | 0.0001897 | -0.0000534 | 0.0005091 | 0.23 | -0.16 | 0.63 | 1.12 | 0.43 | 1.81 | 0.44 | -0.11 | 1.01 | -1.62 | -2.43 | -0.75 |
| | 8 | 3.55 | 2.07 | 5.11 | -0.0000001 | -0.0000033 | 0.0000033 | 0.21 | -0.16 | 0.61 | 1.22 | 0.46 | 1.84 | 0.38 | -0.16 | 0.97 | -1.47 | -2.35 | -0.66 |
| | 9 | 3.54 | 1.98 | 4.97 | -0.0000001 | -0.0000033 | 0.0000034 | 0.25 | -0.17 | 0.62 | 1.10 | 0.46 | 1.84 | 0.37 | -0.12 | 0.98 | -1.59 | -2.35 | -0.65 |
| | 10 | 3.36 | 2.06 | 5.05 | 0.0000002 | -0.0000034 | 0.0000033 | 0.23 | -0.18 | 0.62 | 1.17 | 0.45 | 1.83 | 0.43 | -0.15 | 0.97 | -1.52 | -2.42 | -0.73 |
| | 11 | 3.50 | 2.01 | 5.08 | 0.0000002 | -0.0000032 | 0.0000034 | 0.20 | -0.16 | 0.63 | 1.15 | 0.45 | 1.85 | 0.42 | -0.13 | 1.00 | -1.57 | -2.32 | -0.60 |
| | 12 | 2.45 | 0.87 | 3.66 | 0.0000001 | -0.0000033 | 0.0000033 | 0.89 | 0.48 | 1.33 | 0.69 | 0.08 | 1.47 | 0.32 | -0.18 | 0.90 | -0.91 | -1.77 | -0.14 |
| | 13 | -1.07 | -1.89 | -0.18 | 0.0000003 | -0.0000034 | 0.0000032 | 3.96 | 3.62 | 4.27 | -1.02 | -1.36 | -0.53 | 0.11 | -0.33 | 0.52 | 0.80 | 0.06 | 1.34 |
| | 14 | 2.59 | 1.29 | 4.13 | 0.0000002 | -0.0000032 | 0.0000033 | 0.21 | -0.17 | 0.58 | 0.65 | 0.05 | 1.37 | 1.44 | 0.93 | 2.02 | -1.19 | -1.95 | -0.30 |
| | 15 | 4.99 | 3.48 | 6.57 | 0.0000003 | -0.0000032 | 0.0000032 | -0.02 | -0.38 | 0.38 | 1.18 | 0.54 | 1.92 | 0.16 | -0.33 | 0.76 | -2.65 | -3.54 | -1.78 |
| | | | | | | | | | | | | | | | | | | | |
| Gamma | given prior info | 5.80 | 4.89 | 6.53 | -0.0000001 | -0.0000033 | 0.0000032 | -0.35 | -0.64 | -0.07 | 0.69 | 0.33 | 1.09 | 0.28 | 0.06 | 0.53 | -0.40 | -0.88 | 0.00 |
| | 1 | 5.06 | 4.06 | 5.82 | 0.0003399 | 0.0000211 | 0.0006808 | -0.20 | -0.49 | 0.08 | 0.74 | 0.36 | 1.09 | 0.28 | 0.03 | 0.51 | -0.23 | -0.64 | 0.23 |
| | 2 | 5.70 | 4.92 | 6.53 | 0.0000006 | -0.0000098 | 0.0000110 | -0.36 | -0.65 | -0.07 | 0.71 | 0.37 | 1.12 | 0.28 | 0.05 | 0.53 | -0.40 | -0.88 | -0.01 |
| | 3 | 5.81 | 4.96 | 6.65 | 0.0000038 | -0.0000280 | 0.0000380 | -0.38 | -0.63 | -0.08 | 0.74 | 0.35 | 1.09 | 0.28 | 0.02 | 0.51 | -0.45 | -0.87 | 0.01 |
| | 4 | 5.77 | 4.94 | 6.53 | 0.0000097 | -0.0000403 | 0.0000544 | -0.33 | -0.65 | -0.09 | 0.70 | 0.36 | 1.11 | 0.32 | 0.05 | 0.53 | -0.47 | -0.85 | 0.01 |
| | 5 | 5.69 | 4.95 | 6.61 | 0.0000130 | -0.0000496 | 0.0000808 | -0.35 | -0.63 | -0.06 | 0.71 | 0.34 | 1.09 | 0.27 | 0.04 | 0.52 | -0.39 | -0.92 | -0.03 |
| | 6 | 5.73 | 4.80 | 6.49 | 0.0000607 | -0.0000663 | 0.0001994 | -0.34 | -0.62 | -0.06 | 0.72 | 0.34 | 1.09 | 0.26 | 0.05 | 0.53 | -0.47 | -0.90 | -0.02 |
| | 7 | 5.52 | 4.74 | 6.41 | 0.0001694 | 0.0000438 | 0.0004177 | -0.32 | -0.59 | 0.04 | 0.72 | 0.35 | 1.09 | 0.28 | 0.04 | 0.52 | -0.43 | -0.91 | 0.03 |
| | 11 | 5.77 | 4.85 | 6.50 | 0.0000000 | -0.0000032 | 0.0000035 | -0.34 | -0.64 | -0.08 | 0.73 | 0.35 | 1.11 | 0.32 | 0.04 | 0.53 | -0.39 | -0.89 | -0.01 |
| | 12 | 5.18 | 4.32 | 5.93 | 0.0000001 | -0.0000032 | 0.0000033 | -0.02 | -0.30 | 0.26 | 0.51 | 0.16 | 0.91 | 0.22 | -0.01 | 0.47 | -0.23 | -0.68 | 0.20 |
| | 13 | 1.90 | 1.08 | 2.64 | -0.0000002 | -0.0000032 | 0.0000034 | 2.06 | 1.83 | 2.32 | -0.75 | -1.13 | -0.37 | -0.22 | -0.49 | 0.02 | 1.18 | 0.75 | 1.63 |
| | 14 | 5.66 | 4.85 | 6.46 | 0.0000001 | -0.0000033 | 0.0000033 | -0.41 | -0.69 | -0.15 | 0.68 | 0.31 | 1.04 | 0.46 | 0.26 | 0.72 | -0.37 | -0.84 | 0.03 |
| | 15 | 6.05 | 5.29 | 6.91 | 0.0000000 | -0.0000032 | 0.0000033 | -0.47 | -0.72 | -0.18 | 0.72 | 0.34 | 1.08 | 0.27 | 0.03 | 0.50 | -0.71 | -1.18 | -0.30 |

*Table 23 - Predictions of each trial*

| | | Pred 1 | | | Pred 2 | | | Pred 3 | | | Pred 4 | | | Pred 5 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mode | HDI Low | HDI High | Mode | HDI Low | HDI High | Mode | HDI Low | HDI High | Mode | HDI Low | HDI High | Mode | HDI Low | HDI High |
| Exponential | given prior info | 5.15 | 4.52 | 5.91 | 6.16 | 5.51 | 6.94 | 5.50 | 4.84 | 6.36 | 10.74 | 9.24 | 12.68 | 5.37 | 4.75 | 6.18 |
| | 1 | 5.22 | 4.42 | 5.95 | 6.08 | 5.43 | 6.83 | 5.78 | 4.81 | 6.72 | 12.31 | 10.49 | 14.59 | 5.29 | 4.54 | 6.22 |
| | 2 | 5.15 | 4.53 | 5.92 | 6.23 | 5.53 | 6.93 | 5.59 | 4.75 | 6.27 | 11.02 | 9.30 | 12.76 | 5.39 | 4.69 | 6.15 |
| | 3 | 5.15 | 4.46 | 5.88 | 6.21 | 5.50 | 6.89 | 5.53 | 4.81 | 6.36 | 11.03 | 9.36 | 12.71 | 5.40 | 4.71 | 6.16 |
| | 4 | 5.23 | 4.48 | 5.88 | 6.25 | 5.52 | 6.93 | 5.57 | 4.79 | 6.31 | 11.04 | 9.21 | 12.65 | 5.42 | 4.70 | 6.15 |
| | 5 | 5.16 | 4.47 | 5.87 | 6.15 | 5.52 | 6.91 | 5.50 | 4.81 | 6.33 | 10.90 | 9.27 | 12.73 | 5.35 | 4.65 | 6.13 |
| | 6 | 5.12 | 4.45 | 5.87 | 6.16 | 5.55 | 6.93 | 5.61 | 4.86 | 6.37 | 11.07 | 9.35 | 12.79 | 5.31 | 4.63 | 6.09 |
| | 7 | 5.04 | 4.37 | 5.80 | 6.31 | 5.54 | 6.97 | 5.71 | 5.01 | 6.57 | 11.25 | 9.65 | 13.20 | 5.22 | 4.47 | 5.95 |
| | 8 | 5.13 | 4.55 | 5.94 | 6.12 | 5.49 | 6.90 | 5.56 | 4.83 | 6.34 | 10.92 | 9.21 | 12.75 | 5.33 | 4.71 | 6.16 |
| | 9 | 5.11 | 4.48 | 5.87 | 6.11 | 5.47 | 6.90 | 5.57 | 4.80 | 6.31 | 10.93 | 9.35 | 12.76 | 5.38 | 4.73 | 6.16 |
| | 10 | 5.15 | 4.53 | 5.90 | 6.20 | 5.47 | 6.89 | 5.47 | 4.82 | 6.33 | 10.94 | 9.22 | 12.72 | 5.44 | 4.70 | 6.16 |
| | 11 | 5.22 | 4.48 | 5.90 | 6.16 | 5.51 | 6.92 | 5.60 | 4.80 | 6.35 | 10.88 | 9.37 | 12.82 | 5.38 | 4.73 | 6.18 |
| | 12 | 4.96 | 4.32 | 5.66 | 6.42 | 5.69 | 7.12 | 5.11 | 4.50 | 5.89 | 10.98 | 9.49 | 12.87 | 5.84 | 5.15 | 6.66 |
| | 13 | 5.75 | 5.25 | 6.26 | 10.04 | 9.24 | 10.72 | 5.96 | 5.44 | 6.53 | 15.18 | 13.72 | 16.75 | 9.65 | 8.96 | 10.40 |
| | 14 | 4.82 | 4.19 | 5.43 | 6.91 | 6.18 | 7.71 | 5.21 | 4.57 | 5.98 | 12.42 | 10.68 | 14.20 | 5.07 | 4.36 | 5.67 |
| | 15 | 5.04 | 4.37 | 5.73 | 6.57 | 5.92 | 7.48 | 6.44 | 5.62 | 7.32 | 10.73 | 9.14 | 12.49 | 4.99 | 4.33 | 5.69 |
| | | | | | | | | | | | | | | | | |
| Gamma | given prior info | 6.33 | 5.87 | 6.71 | 5.95 | 5.58 | 6.29 | 6.04 | 5.64 | 6.41 | 7.89 | 7.14 | 8.83 | 5.93 | 5.50 | 6.35 |
| | 1 | 6.32 | 5.89 | 6.73 | 5.89 | 5.57 | 6.30 | 6.10 | 5.63 | 6.58 | 8.94 | 7.69 | 9.86 | 5.98 | 5.54 | 6.46 |
| | 2 | 6.28 | 5.87 | 6.71 | 5.96 | 5.59 | 6.32 | 6.01 | 5.64 | 6.41 | 7.90 | 7.09 | 8.77 | 5.93 | 5.51 | 6.37 |
| | 3 | 6.32 | 5.88 | 6.71 | 5.91 | 5.57 | 6.30 | 6.04 | 5.62 | 6.42 | 8.02 | 7.13 | 8.83 | 5.93 | 5.48 | 6.36 |
| | 4 | 6.27 | 5.87 | 6.70 | 5.97 | 5.58 | 6.31 | 6.03 | 5.63 | 6.40 | 7.96 | 7.10 | 8.82 | 5.93 | 5.51 | 6.37 |
| | 5 | 6.29 | 5.87 | 6.71 | 5.95 | 5.59 | 6.32 | 6.00 | 5.63 | 6.41 | 7.92 | 7.13 | 8.84 | 5.91 | 5.49 | 6.35 |
| | 6 | 6.27 | 5.84 | 6.68 | 5.92 | 5.58 | 6.31 | 6.06 | 5.65 | 6.47 | 8.08 | 7.21 | 9.02 | 5.88 | 5.47 | 6.36 |
| | 7 | 6.19 | 5.81 | 6.67 | 6.01 | 5.61 | 6.35 | 6.19 | 5.73 | 6.58 | 8.32 | 7.32 | 9.26 | 5.85 | 5.40 | 6.28 |
| | 11 | 6.24 | 5.85 | 6.70 | 5.97 | 5.61 | 6.32 | 6.04 | 5.63 | 6.40 | 8.13 | 7.11 | 8.79 | 5.90 | 5.51 | 6.38 |
| | 12 | 6.13 | 5.72 | 6.55 | 6.05 | 5.66 | 6.40 | 5.83 | 5.46 | 6.23 | 8.05 | 7.11 | 8.76 | 6.15 | 5.69 | 6.55 |
| | 13 | 5.46 | 5.03 | 5.90 | 6.85 | 6.45 | 7.27 | 5.01 | 4.65 | 5.42 | 8.31 | 7.48 | 9.27 | 7.53 | 7.08 | 8.04 |
| | 14 | 6.24 | 5.81 | 6.63 | 6.02 | 5.66 | 6.37 | 5.96 | 5.60 | 6.37 | 8.22 | 7.30 | 8.97 | 5.83 | 5.36 | 6.22 |
| | 15 | 6.13 | 5.76 | 6.60 | 6.01 | 5.66 | 6.38 | 6.17 | 5.80 | 6.58 | 7.80 | 6.98 | 8.65 | 5.76 | 5.32 | 6.17 |

## Representativeness, accuracy and efficiency

In this section, we would thoroughly discuss how to examine good diagnostic plots of all MCMC chains so that we could provide the above analysis on the posterior distributions. Posterior distributions are only valid resulted from MCMC with good diagnostic plots. I initially started running JAGS with the following settings

```
nChains: 3, adaptSteps: 1000, burnInSteps: 1000, numSavedSteps: 1000, thinSteps: 1
```

in my first ever attempt on the sample dataset and gradually increased to

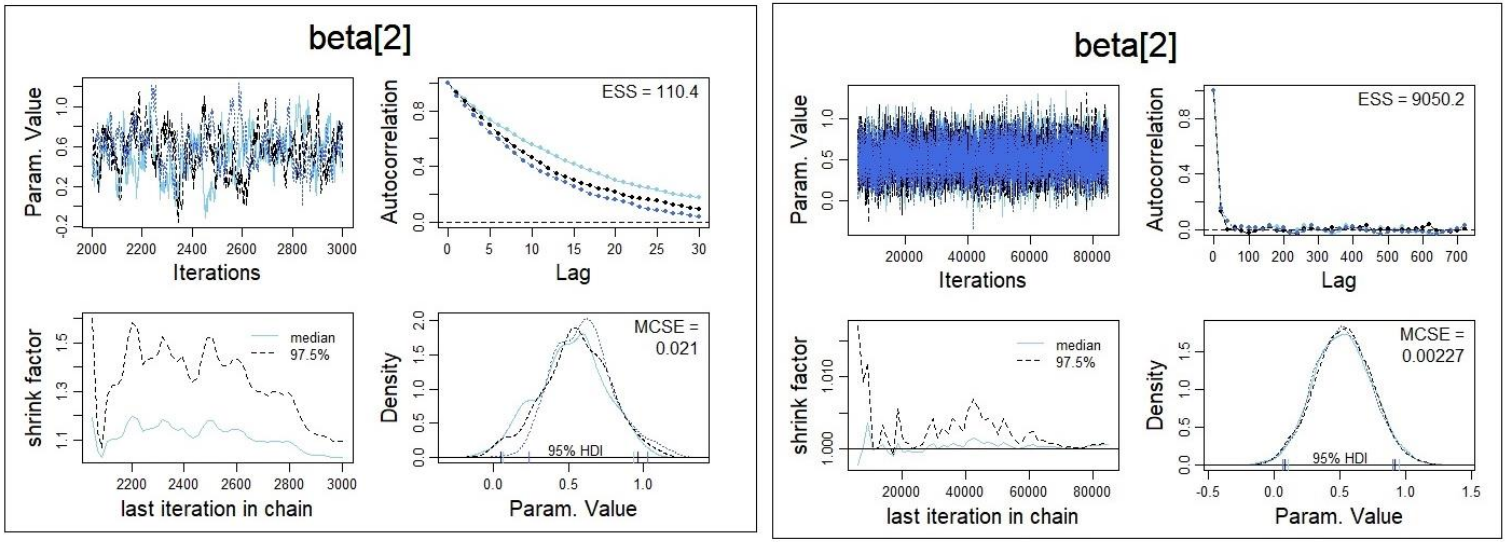
```
nChains: 3, adaptSteps: 1000 burnInSteps: 3000, numSavedSteps: 4000, thinSteps: 20
```

Figure 15 shows how the diagnostic plots have been improved from the initial settings (left) to my satisfactory level (right) on beta2 as an example.

On the Right,

1) 3 chains are all overlapping for the beta2 in the later iterations (top-left plot)
2) there are almost no autocorrelations in the chains (from the top-right plot), ESS is very high, which implies we do not have to change the number of thinning steps.
3) shrink factor is always less than 1.2 (lower-left plot), which implies there are no orphaned or stuck chains
4) In the density plot (lower-right plot), the shape and HDI interval overlap very well. There is no need to adjust any burn-in or saved steps.

*Figure 15 - Comparison of representativeness and accuracy*



On the Left

1) 3 chains do not overlap at all (top-left plot), which indicates we need more saved steps and burn-in steps,
2) there are high autocorrelations in the chains (from the top-right plot), ESS is extremely low, which implies we need to increase the number of thinning steps.
3) shrink factor is higher than 1.2 (lower-left plot), which implies there is orphaned or stuck chains
4) In the density plot (lower-right plot), the shape and HDI interval do not overlap. We need to adjust the burn-in or saved steps.

I put burn-in steps to 3000 in the tuned model, as I found that some of the parameters have high shrink factor in the diagnostic plots all the way in my initial settings when there are 3000 steps in total (`thinning steps = 1, adaptSteps: 1000 + burnInSteps: 1000 + numSavedSteps: 1000 = 3000 steps in total`). As I know that saved steps are not enough, I just put the number of saved steps = adaption + burn-in steps for trial-and-error purpose, and luckily I got a very good tuned model.

With the configuration of `nChains: 3 adaptSteps: 1000 burnInSteps: 3000 numSavedSteps: 4000 thinSteps: 20,` all the parameters would have similar satisfactory diagnostic plots, however JAGS takes much longer to run with these settings. Table 24 shows the runtime of exponential model on sample dataset, it would take 17 minutes on a single run, and I could never obtain any result with these settings on full dataset after an overnight (more than 24 hours) run.

I tried to reduce thinning steps, as I realized the total steps being run would be multiped by number of thinning.

For example, with the satisfactory diagnostic plots, the total steps would be: `adaptSteps: 1000 + burnInSteps: 3000 + (numSavedSteps: 4000 * thinSteps: 20) = 84,000 steps`

However, I have observed some of the parameters do not have good diagnostic plots when thinning go below 20. Thus I tried to reduce the save steps and gradually found that when saved steps=2000, I would have satisfactory plots within an acceptable run time (around 8 minutes for each trial run on sample data set in exponential model as shown in Table 24 ), and I could obtain results on full set of data for both gamma and exponential models with these settings. Refer to [A10] and [A11] for the diagnostic plots for all parameters and predictions when we run full dataset on both models.

*Table 24 - Run time on exponential model*

| Number of Chains | Adaption Steps | Burn-in Steps | Saved Steps | Thining | User Time | System Time | Elapsed time | Total time (in minutes) |
|---|---|---|---|---|---|---|---|---|
| 3 | 1000 | 1000 | 1000 | 1 | 0.36 | 0.89 | 43.05 | 0.72 |
| 3 | 1000 | 1000 | 1000 | 5 | 0.53 | 0.53 | 86.45 | 1.44 |
| 3 | 1000 | 2000 | 1000 | 5 | 1 | 0.46 | 100.57 | 1.68 |
| 3 | 1000 | 2000 | 1000 | 20 | 1.82 | 0.56 | 281.18 | 4.69 |
| 3 | 1000 | 3000 | 2000 | 20 | 2.52 | 0.89 | 467.41 | 7.79 |
| 3 | 1000 | 3000 | 4000 | 20 | 6.11 | 0.78 | 1046 | 17.43 |

## Sample vs whole dataset

Afterall, we would like to validate if the results we found on the sample dataset (with 1000 records) would have true resemblance on the full dataset. With proper tuning as mentioned in the above sections, and putting the initial value as the mode of beta values from our sample run **<this is extremely important, we would not obtain any good diagnostic plots without proper initialization on the full run!>** (refer to [A8]). We could generate posterior results for both models. As the exponential model took 3.45 hours (12394.14 seconds), and the gamma model took 8.23 hours (29647.89 seconds) on a full run. We only attempted full run on the settings with normal prior as follows:

| | $M_1$ | $S_1$ | $M_2$ | $S_2$ | $M_3$ | $S_3$ | $M_4$ | $S_4$ | $M_5$ | $S_5$ |
|---|---|---|---|---|---|---|---|---|---|---|
| values | 90/100,000 | 0.01 | 1 | 2 | 0 | 4 | 1.2 | 0.1 | -1.5 | 0.01 |

Figure 16 and Figure 17 shows the difference between sample and full dataset of the posterior distribution on predictions and $R^2$ in both models. We have obtained the following results on the predictions with full dataset.

*Table 25 - prediction results on full dataset*

| | Prediction 1 | | | Prediction 2 | | | Prediction 3 | | | Prediction 4 | | | Prediction 5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Mode | HDI Low | HDI High | Mode | HDI Low | HDI High | Mode | HDI Low | HDI High | Mode | HDI Low | HDI High | Mode | HDI Low | HDI High |
| Gamma | 6.168 | 6.030 | 6.311 | 5.970 | 5.853 | 6.076 | 5.754 | 5.636 | 5.887 | 7.594 | 7.333 | 7.910 | 6.260 | 6.126 | 6.410 |
| Exponential | 5.721 | 5.470 | 5.948 | 5.587 | 5.398 | 5.799 | 4.823 | 4.627 | 5.046 | 11.191 | 10.657 | 11.709 | 5.970 | 5.730 | 6.220 |

*Compare Table 25 with Table 23, We could summarize the difference in*

Table 26, and we observed that gamma model has smaller difference in mode compare to exponential model, and the range of HDI is much narrower in the full dataset compare to the sample dataset for both models.

*Table 26 - Difference in predictions between full and sample dataset*

| | | Pred 1 | Pred 2 | Pred 3 | Pred 4 | Pred 5 |
|---|---|---|---|---|---|---|
| Gamma Model | Difference in Mode | -0.16 | 0.02 | -0.28 | -0.30 | 0.33 |
| | Range in HDI for sample Data | 0.84 | 0.71 | 0.77 | 1.69 | 0.85 |
| | Range in HDI for full Data | 0.28 | 0.22 | 0.25 | 0.58 | 0.28 |
| Exponential Model | Difference in Mode | 0.57 | -0.57 | -0.68 | 0.45 | 0.60 |
| | Range in HDI for sample Data | 1.39 | 1.43 | 1.51 | 3.44 | 1.43 |
| | Range in HDI for full Data | 0.48 | 0.40 | 0.42 | 1.05 | 0.49 |

The top left corner of Figure 16 and Figure 17 indicate the mode of $R^2$ of the gamma model is at 0.0227 (with HDI between 0.0186 and 0.0264) for the full dataset, which is slightly lower than the sample dataset (mode at 0.0249 with HDI between 0.0145 and 0.0366), the HDI range is also narrower in the full dataset. For exponential model, the mode is at 0.0813 with HDI between 0.0747 and 0.0888, which is slightly higher than the sample dataset. (mode at 0.0768 with HDI between 0.0569 and 0.101), also the HDI range is narrower in the full dataset. However, these figures are all pretty low, which implies that only 2.27% (gamma) and 8.13% (exponential) of the observed variation can be explained by these models, both models still have a weak linear relationship with the dependent variables in the full dataset.

Figure 18 and Figure 19 shows the difference between sample and full dataset of the posterior distribution on coefficients in both models. There are more changes in the Gamma model, as $\beta_2$ becomes positive and $\beta_5$ is insignificant in the full dataset.

With the full dataset, in gamma model, we found that:

1) $\beta_1$ 's HDI captures 0 right in the middle of the distribution, this coefficient is insignificant in the model.
2) $\beta_2$ 's HDI does not capture 0, only 1% of $\beta_2$ 's posterior distribution will be less than 0, 99% will be greater than 0, $\beta_2$ is considered to be significant.
3) $\beta_3$ 's distribution does not capture 0, $\beta_3$ is significant.
4) $\beta_4$ 's HDI does not capture 0, only 0.4% of $\beta_4$ 's posterior distribution will be less than 0, 99.6 % will be greater than 0, $\beta_4$ is considered to be significant.
5) $\beta_5$ 's HDI captures 0 right in the middle of the distribution, this coefficient is insignificant in the model.

With the full dataset, in exponential model, we found that:

1) $\beta_1$ 's HDI captures 0 right in the middle of the distribution, this coefficient is insignificant in the model.
2) $\beta_2$ 's distribution does not capture 0, $\beta_2$ is significant.
3) $\beta_3$ 's distribution does not capture 0, $\beta_3$ is significant.
4) $B_4$ 's distribution does not capture 0, $\beta_4$ is significant.
5) $\beta_5$ 's distribution does not capture 0, $\beta_5$ is significant.

Figure 20 and Figure 21 show the difference of goodness of fit between sample and full dataset. We have noticed the shape of both predicted and observed histograms are similar in full and sample dataset for both models. As there are more data points, the histograms in the full dataset are smoothly connected with each consecutive bars. Gamma model still give a better goodness of fit in the full dataset.

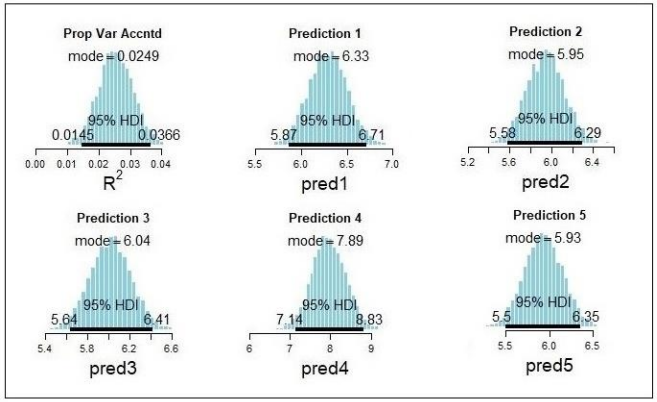
As all the coefficients are influential on the fitted regression lines, there will be correlations between the generated chains for all the betas. If exponential or gamma distribution is used to model $y_i$ and $\varepsilon$, we would use scaling to minimize the effect. Figure 22 and Figure 23 shows the correlations between the scaled betas for both models on sample and full dataset. We have found that the difference of correlation between sample and full dataset are small. High correlated pairs come up similarly in both models and I have summarized as in Table 27 for reference.

*Table 27 - correlation between coefficients*

| | Gamma Model | | Exponential Model | |
|---|---|---|---|---|
| | Sample | Full | Sample | Full |
| zBeta0 (intercept) vs zBeta2 (bedrooms) | -0.61 | -0.61 | -0.56 | -0.64 |
| zBeta0 (intercept) vs zBeta5 (propertyType) | -0.68 | -0.69 | -0.71 | -0.66 |
| zBeta2 (bedrooms) vs zBeta3 (bathrooms) | -0.46 | -0.35 | -0.31 | -0.40 |
| zBeta2 (bedrooms) vs zBeta5 (PropertyType) | 0.47 | 0.45 | 0.43 | 0.44 |

*Figure 16 - Posterior distribution for prediction in Gamma Model*

*Sample Dataset*




*Full Dataset*

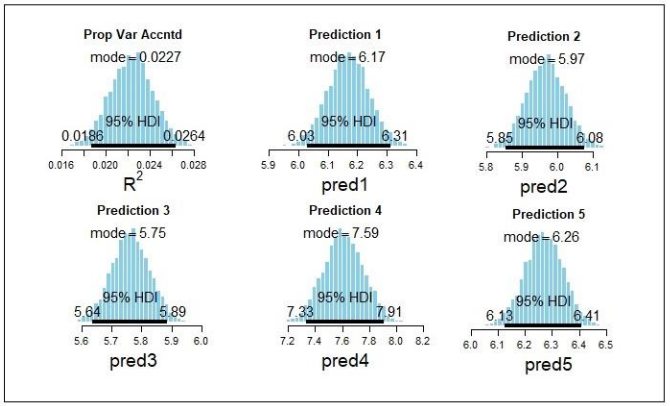
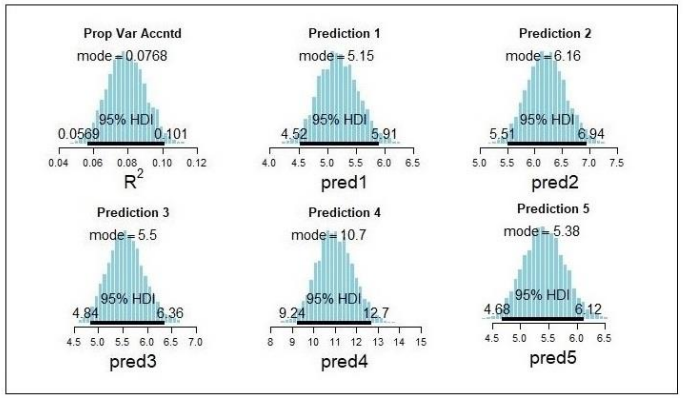



*Figure 17 – Posterior distribution for prediction in Exponential model*

Sample Dataset




Full Dataset

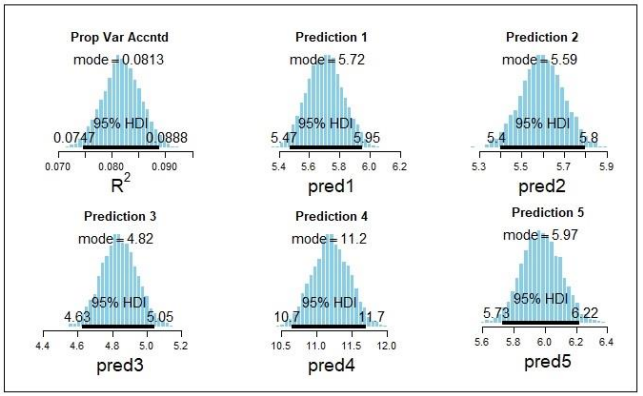
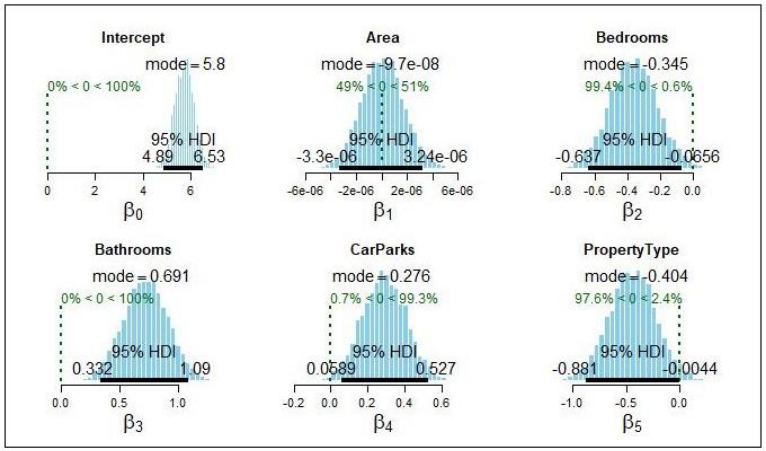


*Figure 18 - Posterior distribution for coefficient in Gamma model*

Sample Dataset




Full Dataset

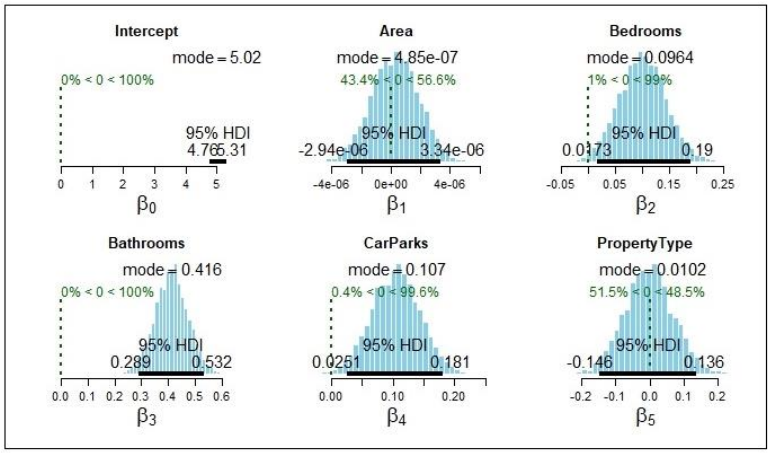
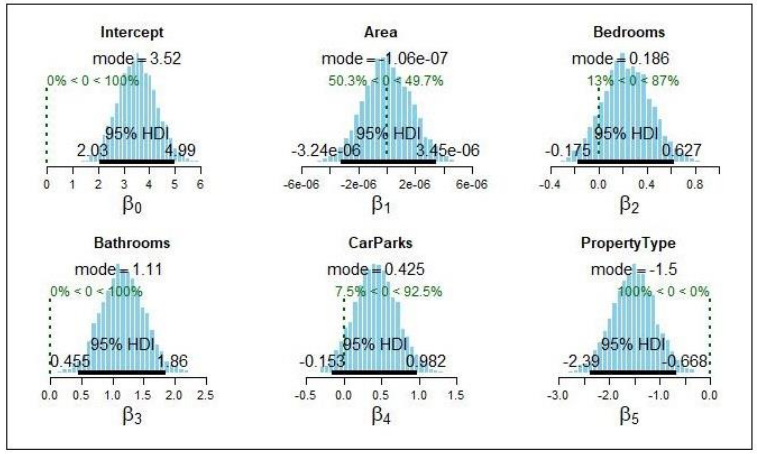



*Figure 19 - Posterior distribution for coefficient in exponential model*

Sample Dataset




Full Dataset

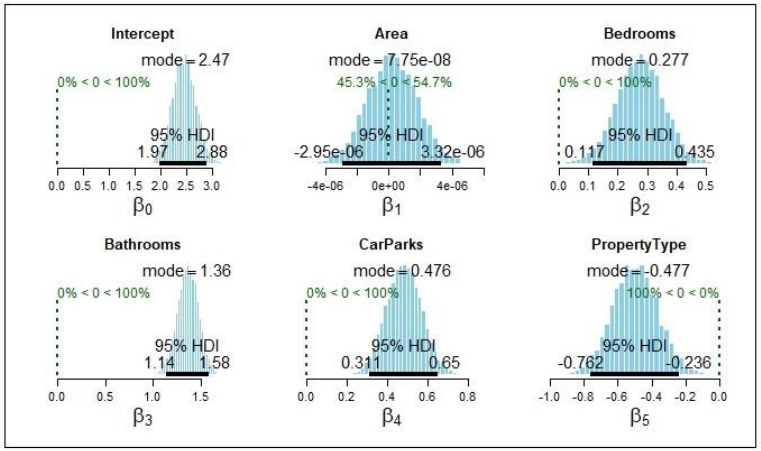
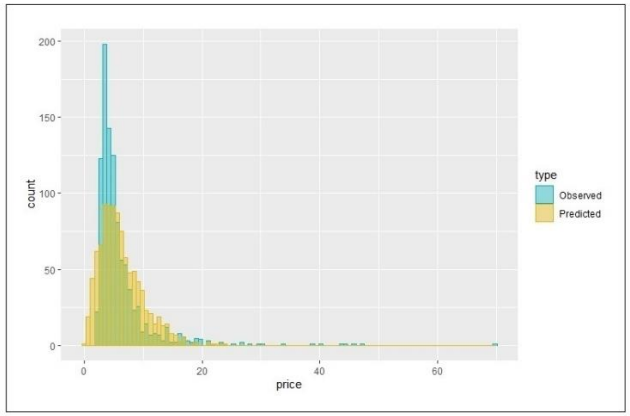


*Figure 20 - Comparison of observed data and posterior distribution of $y_i$ in gamma model*

Sample Dataset




Full Dataset

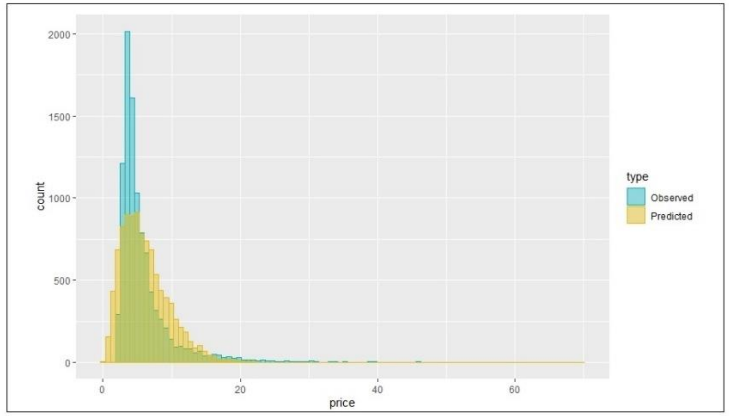
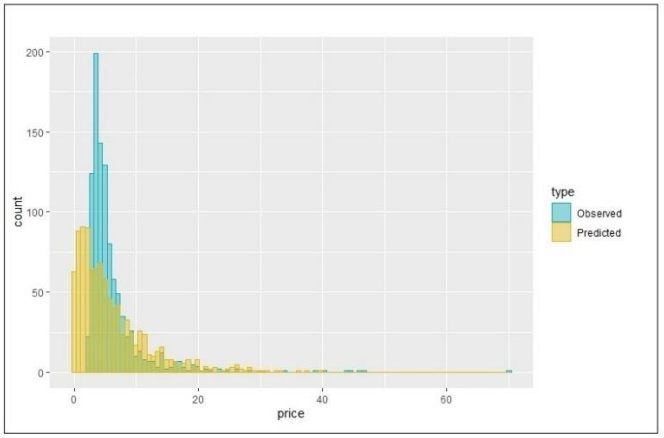



*Figure 21 - Comparison of observed data and posterior distribution of $y_i$ in exponential model*

Sample Dataset




Full Dataset

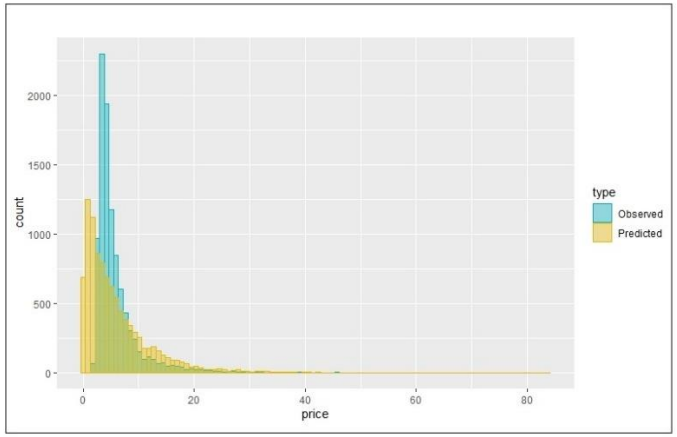
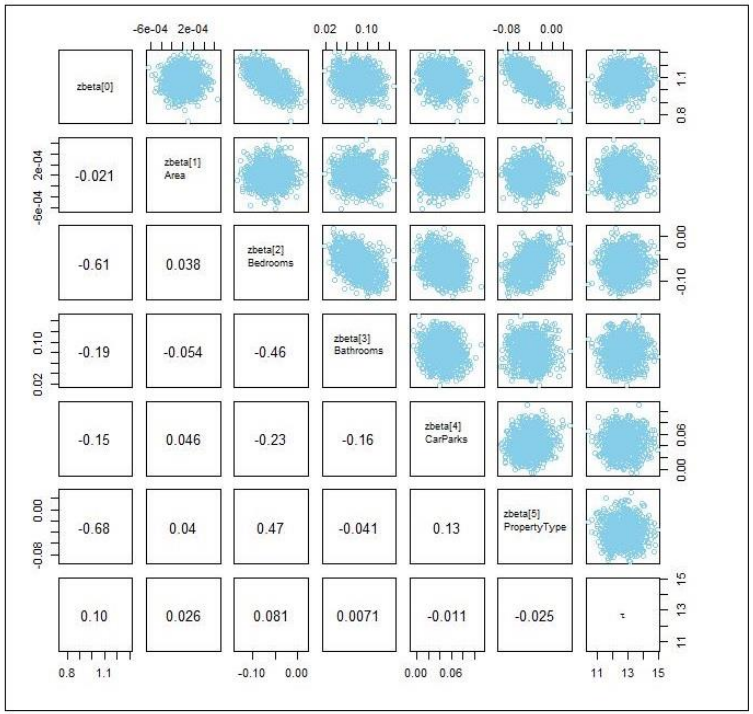


*Figure 22 - Pairplots of gamma model*

Sample Dataset




Full Dataset

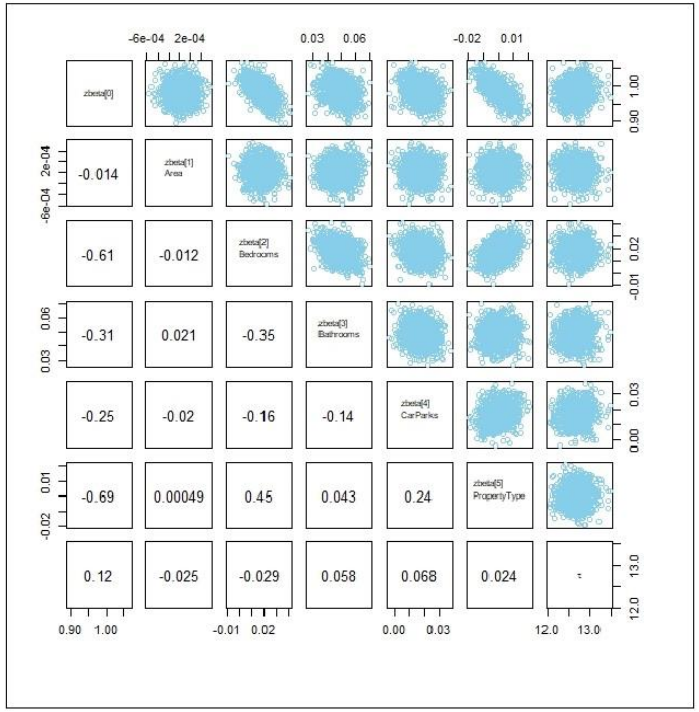
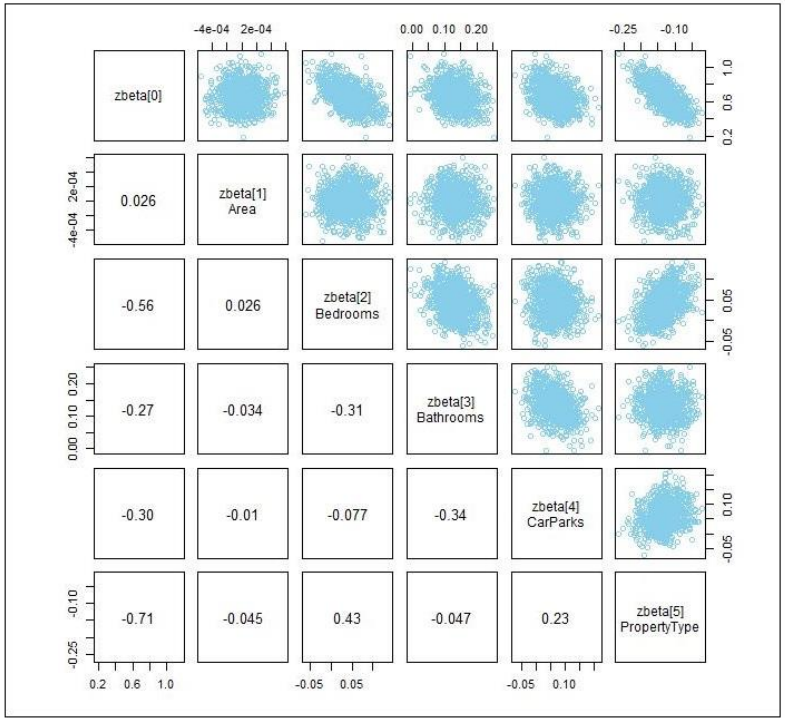
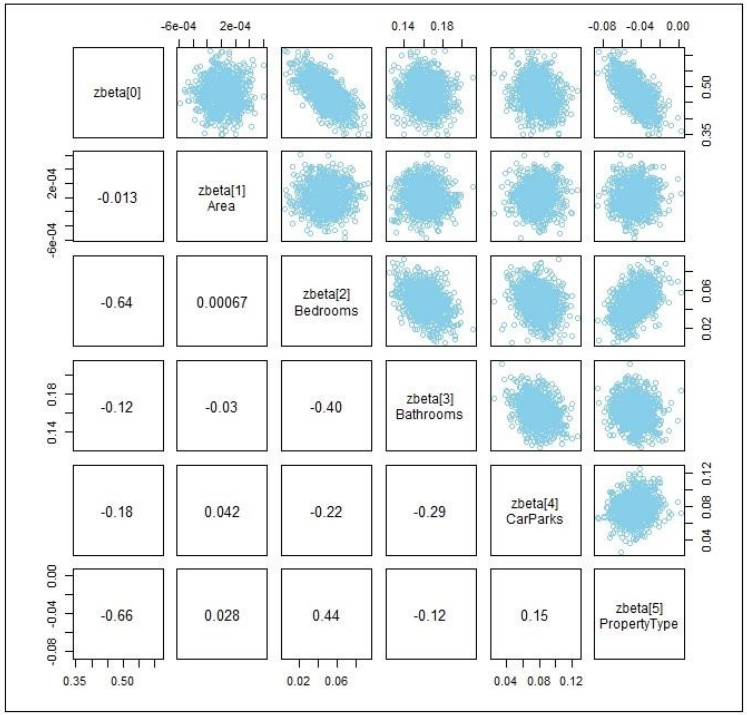


*Figure 23 - Pairplots of exponential model*

Sample Dataset




Full Dataset



# 5. Conclusion

In this assignment, we have learnt how to use JAGS to run MCMC to perform multiple linear regression when dataset and prior knowledge is given. We have worked on 2 linear regression models to perform the analysis. With the given prior knowledge of

1. every $m^2$ increase in land size increases the sales price by AUD90. (**very strong expert knowledge**)
2. every additional bedroom increases the sales price by 100,000 AUD. **(weak expert knowledge)**
3. No expert knowledge on bathrooms.
4. every additional car space increases the sales price by 120,000 AUD. **(strong expert knowledge)**
5. sales price of a unit will be 150,000 AUD less than a house on average. **(very strong expert knowledge)**.

In the gamma model, we found that:

1) number of bedrooms
2) number of bathrooms
3) number of carparks

are significant to predict sales price in the entire dataset. In the exponential model, we found that

1) number of bedrooms
2) number of bathrooms
3) number of carparks, and
4) property Type

are significant to predict sales price in the entire dataset. Gamma model fits better to the dataset, but it runs much slower than the exponential model.

We have also used our regression models to predict the sales price of following properties accordingly

| Prediction | Area | Bedrooms | Bathrooms | CarParks | Property Type |
|---|---|---|---|---|---|
| 1 | 600 | 2 | 2 | 1 | Unit |
| 2 | 800 | 3 | 1 | 2 | House |
| 3 | 1500 | 2 | 1 | 1 | House |
| 4 | 2500 | 5 | 4 | 4 | House |
| 5 | 250 | 3 | 2 | 1 | Unit |

The obtained results are:

| | Prediction 1 | | | Prediction 2 | | | Prediction 3 | | | Prediction 4 | | | Prediction 5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Mode | HDI Low | HDI High | Mode | HDI Low | HDI High | Mode | HDI Low | HDI High | Mode | HDI Low | HDI High | Mode | HDI Low | HDI High |
| Gamma | 6.168 | 6.030 | 6.311 | 5.970 | 5.853 | 6.076 | 5.754 | 5.636 | 5.887 | 7.594 | 7.333 | 7.910 | 6.260 | 6.126 | 6.410 |
| Exponential | 5.721 | 5.470 | 5.948 | 5.587 | 5.398 | 5.799 | 4.823 | 4.627 | 5.046 | 11.191 | 10.657 | 11.709 | 5.970 | 5.730 | 6.220 |

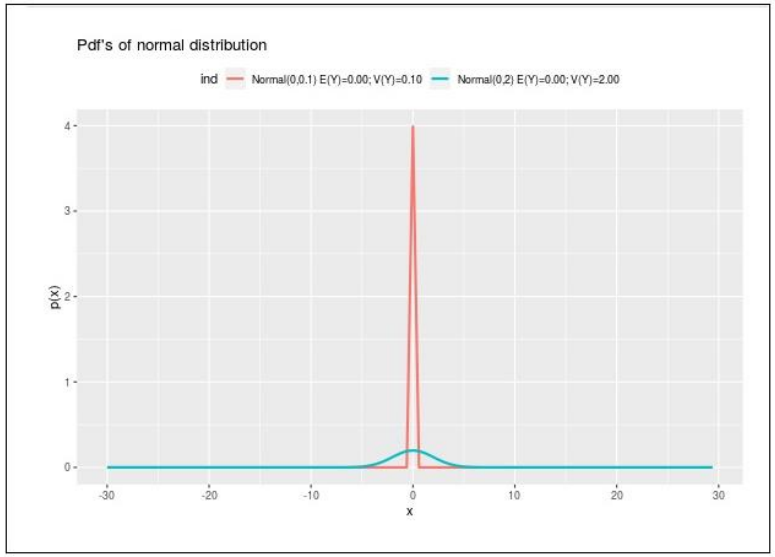
We have also performed extensive analysis to find out that prediction outcomes are subjective to the degree of belief in the prior information.

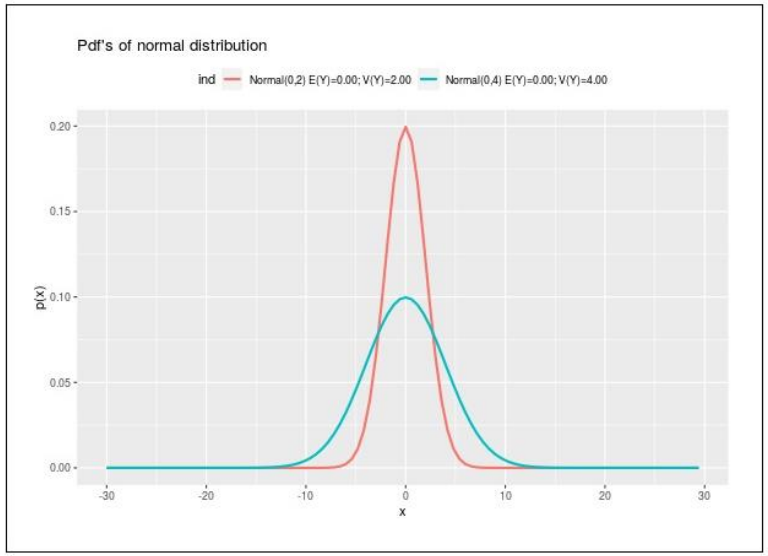
# Appendix

## [A1] Explanation on the degree of belief in normal and gamma prior settings

For all distribution, decrease in variance will increase concentration which will generate a more informative prior. This is more illustrative for normal distribution, left plot in Figure 24 shows that with same mean (location of the prior, 0 in this figure), the curve is more flatten (thus, less concentrated and less informative) when variance is 2 (in blue) compare to variance is 0.1 (in red). Right plot shows the comparison when variance is 2 (in red) with variance is 4 (in blue). Thus for our normal prior, we would set variance as 4 for non-informative, 2 for week belief, 0.1 for strong belief and 0.01 for very strong belief to distinguish the informativeness.

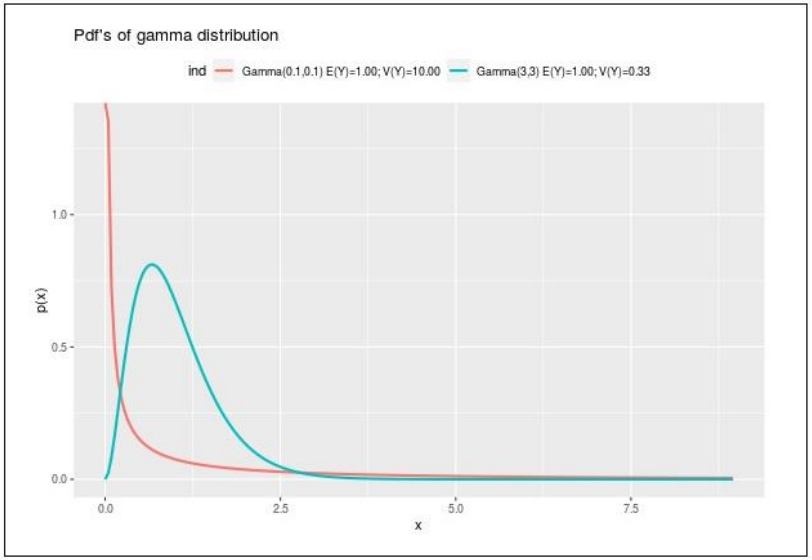
*Figure 24 - Concentration and degree of belief for Normal prior*

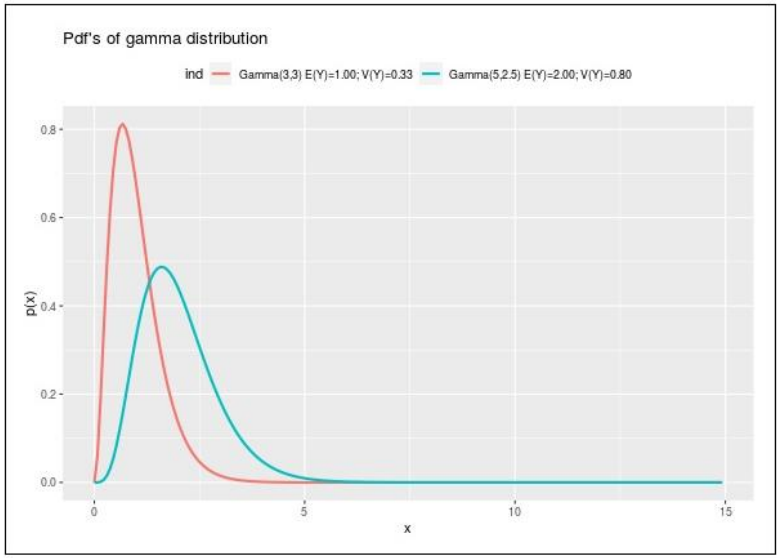


For gamma distribution, it is not as simply illustrated as normal distribution, as the concentration are not shown directly from the plots. For example in the left plot of Figure 25, for Gamma(3, 3) , the curve is more flatten, but actually it has higher concentration compares to Gamma(0.1, 0.1), we are just mislead by the visual effect on the scale of the plot. If we plot Gamma(5, 2.5) with Gamma(3,3) on the same plot as shown on the right of Figure 25, we can see Gamma(3,3) has high concentration when the y-axis drops from 1.5 to 0.8, and x-axis extends from 7.5 to 15. Thus setting Gamma(0.01, 0.01) (This is even further left and less concentrated compare to Gamma(0.1, 0.1)), we are imposing a very non-informative prior in gamma distribution.

*Figure 25 - Concentration and degree of belief for Gamma prior*



## [A2] Import packages and data preparation

```
#The following packages are needed in this assignment:
graphics.off() # This closes all of R's graphics windows.
rm(list=ls())  # Careful! This clears all of R's memory!
library(ggplot2)
library(ggpubr)
library(ks)
library(rjags)
library(runjags)
library(nimble)
library("PerformanceAnalytics")
library(psych)
library(GGally)
library(summarytools)
library(knitr)
library(dplyr)
library(data.table)

setwd("D:/RMIT Master of Analytics/semester 3/MATH2269 Applied Bayesian Statistics/Assignment 2")
source("DBDA2E-utilities.R")

#Read in the datafile
myData <- read.csv("Assignment2PropertyPrices.csv")
```

## [A3] Generate descriptive statistics

```
#Pairplots for all variables
chart.Correlation(myData, histogram=TRUE, pch=19)
```

```r
descr(myData$SalePrice.100K., stats = c("mean", "med", "sd", "Q1", "Q3","IQR", "min", "max"), transpose = T
RUE)

descr(myData$Area, stats = c("mean", "med", "sd", "Q1", "Q3","IQR", "min", "max"), transpose = TRUE)

#Count the frequency of discrete variables
countFreq(myData$Bedrooms)

countFreq(myData$Bathrooms, "Bathroom(s)", start=1, end=4)

countFreq(myData$CarParks, "CarPark(s)", start=0, end=9)

countFreq(myData$PropertyType, isPropertyType = TRUE)

# customized function for counting the frequency and proportion of discrete attributes

countFreq = function(field, str1="Bedroom(s)", isPropertyType=FALSE, start=1, end=7) {
  y = rbind(table(field), prop.table(table(field)))
  if(isPropertyType)
  {
    yColNames <- c('House', 'Unit')
  }
  else
  {
    yColNames <- paste(seq(start, end, 1), str1)
  }
  yColNames
  colnames(y) <- yColNames
  rownames(y) <- c("count", "proportion")
  kable(format(y, digits = 4, drop0trailing = TRUE))

}


#Barplots to explore the relationship of all independent discrete variables

p1 <- ggplot(myData, aes(x=Bedrooms, fill = as.factor(PropertyType))) +
```

```
     geom_bar(stat="count", position=position_dodge()) +
    facet_grid(Bathrooms ~ ., labeller = label_both) + theme(legend.position="bottom") + scale_x_continuous
(breaks=seq(1,7,1)) +
    labs(fill = "Property Type" )


p2 <- ggplot(myData, aes(x=CarParks, fill = as.factor(PropertyType))) +
  geom_bar(stat="count", position=position_dodge()) +
  facet_grid(Bathrooms ~ ., labeller = label_both) + theme(legend.position="bottom") + scale_x_continuous(b
reaks=seq(0,9,1)) +
  labs(fill = "Property Type")


p3 <- ggplot(myData, aes(x=CarParks, fill = as.factor(PropertyType))) +
  geom_bar(stat="count", position=position_dodge()) +
  facet_grid(Bedrooms ~ ., labeller = label_both) + theme(legend.position="bottom") + scale_x_continuous(br
eaks=seq(0,9,1)) +
  labs(fill = "Property Type")


figure <- ggarrange(p1, p2, p3, nrow = 1, ncol = 3)
figure <- annotate_figure(figure,
                          top = text_grob("Relationship of Independent Discrete Variables", face = "bold",
size = 14))


figure


#Scatter plots to explore the relationship of Sales Price vs Area by each categorical variable

p1 <- ggplot(myData, aes(x=Area, y=SalePrice.100K. ,color=as.factor(PropertyType))) +
  labs(color = "PropertyType") +  theme(legend.position="bottom") +
  geom_point()
```

```r
p2 <- ggplot(myData, aes(x=Area, y=SalePrice.100K. ,color=as.factor(Bedrooms))) +
  labs(color = "Bedrooms") + theme(legend.position="bottom") +
  geom_point()


p3 <- ggplot(myData, aes(x=Area, y=SalePrice.100K. ,color=as.factor(Bathrooms))) +
  labs(color = "Bedrooms") + theme(legend.position="bottom") +
  geom_point()


p4 <- ggplot(myData, aes(x=Area, y=SalePrice.100K. ,color=as.factor(CarParks))) +
  labs(color = "CarParks") + theme(legend.position="bottom") +
  geom_point()


figure <- ggarrange(p1, p2, p3, p4, nrow = 2, ncol = 2)
figure <- annotate_figure(figure,
                top = text_grob("Sales Price vs Area group by each Discrete Variable", face = "bold", size
= 14))

figure
```

## [A4] Compare the distribution of dependent variable between sample and full dataset

```r
#Select random samples
set.seed(888)
mySample <- myData[sample(1:nrow(myData), 1000,
                          replace=FALSE),]
```

```
# Histogram for comparing the whole and sample data set on count of sales price

p1 <- ggplot(data=myData, aes(SalePrice.100K.)) +
  geom_histogram(fill="white", color="black") +
  ggtitle("Whole data set") +
  theme(plot.title = element_text(hjust = 0.5))

p2 <- ggplot(data=mySample, aes(SalePrice.100K.)) +
  geom_histogram(fill="white", color="black") +
  ggtitle("Sample data") +
  theme(plot.title = element_text(hjust = 0.5))
figure <- ggarrange(p1, p2, nrow = 2, ncol = 1)
figure <- annotate_figure(figure,
                          top = text_grob("Comparison between original and sample data set on Count of
Sales Price", face = "bold", size = 14))
figure
```

## [A5] *Overlaying gamma and exponential distribution on the histogram of likelihood of dependent variable (sales price of house properties) of sample dataset*

```
# Fitting gamma and exponential distribution as likelihood into the dataset as likelihood

h <- mySample$SalePrice.100K. %>% hist(col="grey",xlab="Sales Price (100k)", main="Histogram of Melbourne
properties sales price in AUD$100,000", breaks=50)

xfit<-seq(min(mySample$SalePrice.100K.),max(mySample$SalePrice.100K.),length=40)

yfit<-dexp(xfit,rate=1/mean(mySample$SalePrice.100K.))
yfit <- yfit*diff(h$mids[1:2])*length(mySample$SalePrice.100K.)
lines(xfit, yfit, col="blue", lwd=2)
```

```r
myAlpha <- (mean(mySample$SalePrice.100K.)^2)/(sd(mySample$SalePrice.100K.)^2)
myBeta <- mean(mySample$SalePrice.100K.)/(sd(mySample$SalePrice.100K.)^2)

yfit<-dgamma(xfit,shape=myAlpha, rate=myBeta)
yfit <- yfit*diff(h$mids[1:2])*length(mySample$SalePrice.100K.)
lines(xfit, yfit, col="orange", lwd=2)
legend(40, 200, legend=c("exponential distribution", "gamma distribution"),
       col=c("blue", "orange"), lty=1:2, cex=0.8)
```

## [A6] Functions to invoke JAGS to run MCMC and relevant diagnostic and summary plots

```r
#==============================================================================
# customized smryMCM function which would print the summary result of the posterior distribution
smryMCMC_HD = function(  codaSamples , compVal = NULL,  saveName=NULL) {
  summaryInfo = NULL
  mcmcMat = as.matrix(codaSamples,chains=TRUE)
  paramName = colnames(mcmcMat)
  for ( pName in paramName ) {
    if (pName %in% colnames(compVal)){
      if (!is.na(compVal[pName])) {
        summaryInfo = rbind( summaryInfo , summarizePost( paramSampleVec = mcmcMat[,pName] ,
                                                           compVal = as.numeric(compVal[pName]) ))
      }
      else {
        summaryInfo = rbind( summaryInfo , summarizePost( paramSampleVec = mcmcMat[,pName] ) )
      }
    } else {
      summaryInfo = rbind( summaryInfo , summarizePost( paramSampleVec = mcmcMat[,pName] ) )
    }
  }
  rownames(summaryInfo) = paramName

  # summaryInfo = rbind( summaryInfo ,
```

```
  #                         "tau" = summarizePost( mcmcMat[,"tau"] ) )
  if ( !is.null(saveName) ) {
    write.csv( summaryInfo , file=paste(saveName,"SummaryInfo.csv",sep="") )
  }
  return( summaryInfo )
}

#=============================================================================
# customized plotMCMC function which would plot the posterior distribution of each parameter

plotMCMC_HD = function( codaSamples , data , xName="x" , yName="y" ,
                        showCurve=FALSE ,  pairsPlot=FALSE , compVal = NULL,
                        saveName=NULL , saveType="jpg", isExp=FALSE) {
  # showCurve is TRUE or FALSE and indicates whether the posterior should
  #   be displayed as a histogram (by default) or by an approximate curve.
  # pairsPlot is TRUE or FALSE and indicates whether scatterplots of pairs
  #   of parameters should be displayed.
  #-----------------------------------------------------------------------------
  y = data[,yName]
  x = as.matrix(data[,xName])
  mcmcMat = as.matrix(codaSamples,chains=TRUE)
  chainLength = NROW( mcmcMat )
  zbeta0 = mcmcMat[,"zbeta0"]
  zbeta  = mcmcMat[,grep("^zbeta$|^zbeta\\[",colnames(mcmcMat))]
  if ( ncol(x)==1 ) { zbeta = matrix( zbeta , ncol=1 ) }
  if(isExp==FALSE)
  {
    zVar = mcmcMat[,"zVar"]
  }
  beta0 = mcmcMat[,"beta0"]
  beta  = mcmcMat[,grep("^beta$|^beta\\[",colnames(mcmcMat))]
  if ( ncol(x)==1 ) { beta = matrix( beta , ncol=1 ) }
  if(isExp==FALSE)
  {
    tau = mcmcMat[,"tau"]
  }
```

```
  pred1 = mcmcMat[,"pred[1]"] # Added by Demirhan
  pred2 = mcmcMat[,"pred[2]"] # Added by Demirhan
  pred3 = mcmcMat[,"pred[3]"] # Added by millie
  pred4 = mcmcMat[,"pred[4]"] # Added by millie
  pred5 = mcmcMat[,"pred[5]"] # Added by millie
  #-----------------------------------------------------------------------------
  # Compute R^2 for credible parameters:
  YcorX = cor( y , x ) # correlation of y with each x predictor
  Rsq = zbeta %*% matrix( YcorX , ncol=1 )
  #-----------------------------------------------------------------------------
  if ( pairsPlot ) {
    # Plot the parameters pairwise, to see correlations:
    openGraph()
    nPtToPlot = 1000
    plotIdx = floor(seq(1,chainLength,by=chainLength/nPtToPlot))
    panel.cor = function(x, y, digits=2, prefix="", cex.cor, ...) {
      usr = par("usr"); on.exit(par(usr))
      par(usr = c(0, 1, 0, 1))
      r = (cor(x, y))
      txt = format(c(r, 0.123456789), digits=digits)[1]
      txt = paste(prefix, txt, sep="")
      if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
      text(0.5, 0.5, txt, cex=1.25 ) # was cex=cex.cor*r
    }
    if(isExp==FALSE)
    {

      pairs( cbind( beta0 , beta , tau )[plotIdx,] ,
             labels=c( "beta[0]" ,
                       paste0("beta[",1:ncol(beta),"]\n",xName) ,
                       expression(tau) ) ,
             lower.panel=panel.cor , col="skyblue" )
    }
    else{
      pairs( cbind( beta0 , beta )[plotIdx,] ,
```

```
            labels=c( "beta[0]" ,
                      paste0("beta[",1:ncol(beta),"]\n",xName) ) ,
            lower.panel=panel.cor , col="skyblue" )
    }
    if (is.null(saveName) ) {
      saveGraph( file=paste(saveName,"PostPairs",sep=""), type=saveType)
    }
    if(isExp==FALSE)
    {

      pairs( cbind( zbeta0 , zbeta , tau )[plotIdx,] ,
            labels=c( "zbeta[0]" ,
                      paste0("zbeta[",1:ncol(zbeta),"]\n",xName) ,
                      expression(tau) ) ,
            lower.panel=panel.cor , col="skyblue" )
    }
    else{
      pairs( cbind( zbeta0 , zbeta )[plotIdx,] ,
            labels=c( "zbeta[0]" ,
                      paste0("zbeta[",1:ncol(beta),"]\n",xName) ) ,
            lower.panel=panel.cor , col="skyblue" )
    }
    if (is.null(saveName) ) {
      saveGraph( file=paste(saveName,"zPostPairs",sep=""), type=saveType)
    }

  }
  #-------------------------------------------------------------------------
  # Marginal histograms:

  decideOpenGraph = function( panelCount , saveName , finished=FALSE ,
                              nRow=2 , nCol=3 ) {
    # If finishing a set:
    if ( finished==TRUE ) {
      if ( !is.null(saveName) ) {
        saveGraph( file=paste0(saveName,ceiling((panelCount-1)/(nRow*nCol))),
```

```
                    type=saveType)
    }
    panelCount = 1 # re-set panelCount
    return(panelCount)
  } else {
    # If this is first panel of a graph:
    if ( ( panelCount %% (nRow*nCol) ) == 1 ) {
      # If previous graph was open, save previous one:
      if ( panelCount>1 & !is.null(saveName) ) {
        saveGraph( file=paste0(saveName,(panelCount%/%(nRow*nCol))),
                   type=saveType)
      }
      # Open new graph
      openGraph(width=nCol*7.0/3,height=nRow*2.0)
      layout( matrix( 1:(nRow*nCol) , nrow=nRow, byrow=TRUE ) )
      par( mar=c(4,4,2.5,0.5) , mgp=c(2.5,0.7,0) )
    }
    # Increment and return panel count:
    panelCount = panelCount+1
    return(panelCount)
  }
}

# Original scale:
panelCount = 1
if (!is.na(compVal["beta0"])){
  panelCount = decideOpenGraph( panelCount , saveName=paste0(saveName,"PostMarg") )
  histInfo = plotPost( beta0 , cex.lab = 1.75 , showCurve=showCurve ,
                       xlab=bquote(beta[0]) , main="Intercept", compVal = as.numeric(compVal["beta0"] ))
} else {
  panelCount = decideOpenGraph( panelCount , saveName=paste0(saveName,"PostMarg") )
  histInfo = plotPost( beta0 , cex.lab = 1.75 , showCurve=showCurve ,
                       xlab=bquote(beta[0]) , main="Intercept")
}
for ( bIdx in 1:ncol(beta) ) {
  panelCount = decideOpenGraph( panelCount , saveName=paste0(saveName,"PostMarg") )
```

```
    if (!is.na(compVal[paste0("beta[",bIdx,"]")])) {
      histInfo = plotPost( beta[,bIdx] , cex.lab = 1.75 , showCurve=showCurve ,
                           xlab=bquote(beta[.(bIdx)]) , main=xName[bIdx],
                           compVal = as.numeric(compVal[paste0("beta[",bIdx,"]")]))
    } else{
      histInfo = plotPost( beta[,bIdx] , cex.lab = 1.75 , showCurve=showCurve ,
                           xlab=bquote(beta[.(bIdx)]) , main=xName[bIdx])
    }
  }
  panelCount = decideOpenGraph( panelCount , saveName=paste0(saveName,"PostMarg") )
  if(isExp==FALSE)
  {
    histInfo = plotPost( tau , cex.lab = 1.75 , showCurve=showCurve ,
                         xlab=bquote(tau) , main=paste("Scale") )
  }
  panelCount = decideOpenGraph( panelCount , saveName=paste0(saveName,"PostMarg") )
  histInfo = plotPost( Rsq , cex.lab = 1.75 , showCurve=showCurve ,
                       xlab=bquote(R^2) , main=paste("Prop Var Accntd") )
  panelCount = decideOpenGraph( panelCount ,  saveName=paste0(saveName,"PostMarg") )
  histInfo = plotPost( pred1 , cex.lab = 1.75 , showCurve=showCurve ,
                       xlab="pred1" , main="Prediction 1" ) # Added by Demirhan
  panelCount = decideOpenGraph( panelCount ,  saveName=paste0(saveName,"PostMarg") )
  histInfo = plotPost( pred2 , cex.lab = 1.75 , showCurve=showCurve ,
                       xlab="pred2" , main="Prediction 2" ) # Added by Demirhan
  panelCount = decideOpenGraph( panelCount ,  saveName=paste0(saveName,"PostMarg") )
  histInfo = plotPost( pred3 , cex.lab = 1.75 , showCurve=showCurve ,
                       xlab="pred3" , main="Prediction 3" ) # Added by Demirhan
  panelCount = decideOpenGraph( panelCount ,  saveName=paste0(saveName,"PostMarg") )
  histInfo = plotPost( pred4 , cex.lab = 1.75 , showCurve=showCurve ,
                       xlab="pred4" , main="Prediction 4" ) # Added by Demirhan
  panelCount = decideOpenGraph( panelCount ,  saveName=paste0(saveName,"PostMarg") )
  histInfo = plotPost( pred5 , cex.lab = 1.75 , showCurve=showCurve ,
                       xlab="pred5" , main="Prediction 5" ) # Added by Demirhan
  panelCount = decideOpenGraph( panelCount ,  saveName=paste0(saveName,"PostMarg") )

  # Standardized scale:
```

```
  panelCount = 1
  panelCount = decideOpenGraph( panelCount , saveName=paste0(saveName,"PostMargZ") )
  histInfo = plotPost( zbeta0 , cex.lab = 1.75 , showCurve=showCurve ,
                       xlab=bquote(z*beta[0]) , main="Intercept" )
  for ( bIdx in 1:ncol(beta) ) {
    panelCount = decideOpenGraph( panelCount , saveName=paste0(saveName,"PostMargZ") )
    histInfo = plotPost( zbeta[,bIdx] , cex.lab = 1.75 , showCurve=showCurve ,
                         xlab=bquote(z*beta[.(bIdx)]) , main=xName[bIdx] )
  }
  panelCount = decideOpenGraph( panelCount , saveName=paste0(saveName,"PostMargZ") )
  if(isExp==FALSE)
  {
    histInfo = plotPost( zVar , cex.lab = 1.75 , showCurve=showCurve ,
                         xlab=bquote(z*tau) , main=paste("Scale") )
  }
  panelCount = decideOpenGraph( panelCount , saveName=paste0(saveName,"PostMargZ") )
  histInfo = plotPost( Rsq , cex.lab = 1.75 , showCurve=showCurve ,
                       xlab=bquote(R^2) , main=paste("Prop Var Accntd") )
  panelCount = decideOpenGraph( panelCount , finished=TRUE , saveName=paste0(saveName,"PostMargZ") )

  #------------------------------------------------------------------------------
}

#==============================================================================
# customized function to generate the Bayesian model with given sensitivity of the prior
# for either gamma or exp model

writeModel <- function(model = c("gamma", "exp"), informative=FALSE, filename="TEMPmodel.txt",
beta1Sens="0.1", beta2Sens="2", beta4Sens="1", beta5Sens="0.1"){

  dataString = "
# Standardize the data:
data {
  ysd <- sd(y)
  for ( i in 1:Ntotal ) {
    zy[i] <- y[i] / ysd
```

```
  }
  for ( j in 1:Nx ) {
    xsd[j] <-   sd(x[,j])
    for ( i in 1:Ntotal ) {
      zx[i,j] <- x[i,j] / xsd[j]
    }
  }
}
"
if (model == "gamma"){
  modelString1 = "
  model {
    for ( i in 1:Ntotal ) {
      zy[i] ~ dgamma( (mu[i]^2)/zVar , mu[i]/zVar ) #gamma likelihood
"
} else if (model == "exp") {
  modelString1 = "
  model {
    for ( i in 1:Ntotal ) {
      zy[i] ~ dexp(1/mu[i])                          #exponential likelihood
"
} else {
  warning('model can only be ("gamma","exp") in this function')
}

modelString2 = "
    mu[i] <- zbeta0 + sum( zbeta[1:Nx] * zx[i,1:Nx] )
  }
"

if (informative == TRUE){
  priorString0=paste("zbeta0 ~ dnorm( 0 , 1/2^2 )", "\n")
  priorString1= paste("zbeta[1] ~ dnorm( (90/100000)/xsd[1] , 1/(", beta1Sens, "/xsd[1]^2) )", "\n")
  priorString2= paste("zbeta[2] ~ dnorm( 1/xsd[2], 1/(", beta2Sens, "/xsd[2]^2) )", "\n")
  priorString3= paste("zbeta[3] ~ dnorm( 0, 1/4 )", "\n")
  priorString4= paste("zbeta[4] ~ dnorm( 1.2/xsd[4] , 1/(", beta4Sens, "/xsd[4]^2) )", "\n")
```

```
  priorString5= paste("zbeta[5] ~ dnorm( -1.5/xsd[5] , 1/(", beta5Sens, "/xsd[5]^2) )", "\n")

  priorString = paste(priorString0, priorString1,priorString2, priorString3, priorString4, priorString5)

}
else{
  priorString = "
  zbeta0 ~ dnorm( 0 , 1/2^2 )  # 1/ variance for normal distribution
  zbeta[1] ~ dnorm( 0 , 1/2^2 ) # 1/ variance for normal distribution
  zbeta[2] ~ dnorm( 0 , 1/2^2 ) # 1/ variance for normal distribution
  zbeta[3] ~ dnorm( 0 , 1/2^2) # 1/ variance for normal distribution
  zbeta[4] ~ dnorm( 0 , 1/2^2 ) # 1/ variance for normal distribution
  zbeta[5] ~ dnorm( 0 , 1/2^2 ) # 1/ variance for normal distribution
"

}
if (model == "gamma"){
  modelString3 = "
  zVar ~ dgamma( 0.01 , 0.01 ) #gamma variance
  tau <- zVar * (ysd)^2 #for gamma

"
} else {
  modelString3 = "
"
}

modelString4 = "

  # Transform to original scale:
  beta[1:Nx] <- ( zbeta[1:Nx] / xsd[1:Nx] ) * ysd
  beta0 <- zbeta0*ysd

  # Compute predictions at every step of the MCMC
  for ( i in 1:Nx){
```

```
    pred[i] <- beta0 + beta[1] * xPred[i,1] + beta[2] * xPred[i,2] + beta[3] * xPred[i,3] + beta[4] *
xPred[i,4]+ beta[5] * xPred[i,5]
  }

}
" # close quote for modelString
# Write out modelString to a text file

finalString = paste(dataString,modelString1, modelString2, priorString, modelString3, modelString4)
writeLines( finalString , con=filename )
if (model == "gamma"){
  parameters = c( "zbeta0" ,  "zbeta" , "beta0" ,  "beta" ,  "tau", "zVar", "pred") # Here beta is a
vector!
}
else{
  parameters = c( "zbeta0" ,  "zbeta" , "beta0" ,  "beta" , "pred") # Here beta is a vector!

}

return (parameters)
}

#==============================================================================
# customized function to call jags with parallel run, the result is saved as an RDS object
# with the filename prefixed with "coda_" in front of the passed-in saveFile


callParaRun <- function(isExp=FALSE, filename="TEMPmodel.txt", saveFile="runsample.RData", parameters,
dataList, initsList, nChains, adaptSteps, burnInSteps, numSavedSteps, thinSteps)
{
  #print configuration
  if(isExp == TRUE)
  {
    print("running Exponential .....")
  }
```

```
  else
  {
    print("running Gamma .....")
  }

  cat(paste("nChains:", nChains, "adaptSteps:", adaptSteps,"burnInSteps:", burnInSteps,"numSavedSteps:",
numSavedSteps, "thinSteps:", thinSteps, "\n"))

  startTime = proc.time()
  runJagsOut <- run.jags( method="parallel" ,
                          model=filename,
                          monitor=parameters,
                          data=dataList ,
                          inits=initsList ,
                          n.chains=nChains ,
                          adapt=adaptSteps ,
                          burnin=burnInSteps ,
                          sample=numSavedSteps ,
                          thin=thinSteps , summarise=FALSE , plots=FALSE )
  codaSamples = as.mcmc.list( runJagsOut )
  stopTime = proc.time()
  elapsedTime = stopTime - startTime
  show(elapsedTime)

  save.image(file=saveFile)
  saveRDS(codaSamples, paste(("coda_"), saveFile))

  diagMCMC( codaSamples , parName="beta0" )
  diagMCMC( codaSamples , parName="beta[1]" )
  diagMCMC( codaSamples , parName="beta[2]" )
  diagMCMC( codaSamples , parName="beta[3]" )
  diagMCMC( codaSamples , parName="beta[4]" )
  diagMCMC( codaSamples , parName="beta[5]" )
  if(isExp == FALSE)
  {
```

```
    diagMCMC( codaSamples , parName="tau" )
  }
  diagMCMC( codaSamples , parName="pred[1]" )
  diagMCMC( codaSamples , parName="pred[2]" )
  diagMCMC( codaSamples , parName="pred[3]" )
  diagMCMC( codaSamples , parName="pred[4]" )
  diagMCMC( codaSamples , parName="pred[5]" )

  diagMCMC( codaSamples , parName="zbeta0" )
  diagMCMC( codaSamples , parName="zbeta[1]" )
  diagMCMC( codaSamples , parName="zbeta[2]" )
  diagMCMC( codaSamples , parName="zbeta[3]" )
  diagMCMC( codaSamples , parName="zbeta[4]" )
  diagMCMC( codaSamples , parName="zbeta[5]" )

  return(codaSamples)
}

#==============================================================================
# customized function to invoke the call parallel run summaryMCMC and plotMCMC function
# for either exponential or gamma function, the result is saved as an RDS object
# with the filename prefixed with “summry” in front of the passed-in saveFile


runAndPlot <- function(isExp=FALSE, informative=FALSE, filename="tempModel.txt",
                       beta1Sens="4", beta2Sens="4", beta4Sens="4", beta5Sens="4",
                       saveFile="runsample.RData", dataList, initsList, nChains, adaptSteps, burnInSteps,
numSavedSteps, thinSteps)
{
  if(isExp==FALSE)
  {
    parameters=writeModel(model ="gamma", informative=informative, filename=filename, beta1Sens=beta1Sens,
beta2Sens=beta2Sens, beta4Sens=beta4Sens, beta5Sens=beta5Sens)
    AcodaSample=callParaRun(isExp=FALSE, filename=filename, saveFile=saveFile, parameters, dataList,
initsList, nChains, adaptSteps, burnInSteps, numSavedSteps, thinSteps)
```

```
    compVal <- data.frame("beta0" = 0, "beta[1]" = 0, "beta[2]" = 0, "beta[3]" = 0, "beta[4]" =  0,
"beta[5]" =  0, "tau" = NA , check.names=FALSE)
  }
  else
  {
    parameters=writeModel(model ="exp", informative=informative, filename=filename, beta1Sens=beta1Sens,
beta2Sens=beta2Sens, beta4Sens=beta4Sens, beta5Sens=beta5Sens)
    AcodaSample=callParaRun(isExp=TRUE,  filename=filename, saveFile=saveFile, parameters, dataList,
initsList, nChains, adaptSteps, burnInSteps, numSavedSteps, thinSteps)
    compVal <- data.frame("beta0" = 0, "beta[1]" = 0, "beta[2]" = 0, "beta[3]" = 0, "beta[4]" =  0,
"beta[5]" =  0,  check.names=FALSE)

  }
  summaryInfo <- smryMCMC_HD(codaSamples=AcodaSample , compVal = compVal)
  print(summaryInfo)
  if(isExp==FALSE)
  {
    plotMCMC_HD( codaSamples = AcodaSample , data = myData,
xName=c("Area","Bedrooms","Bathrooms","CarParks", "PropertyType") ,
                 yName="SalePrice.100K.", compVal = compVal, pairsPlot=TRUE, isExp=FALSE)

  }
  else
  {
    plotMCMC_HD( codaSamples = AcodaSample, data = myData,
xName=c("Area","Bedrooms","Bathrooms","CarParks", "PropertyType") ,
                 yName="SalePrice.100K.", compVal = compVal, pairsPlot=TRUE, isExp=TRUE)
  }
  saveRDS(summaryInfo, paste(("summryInfo_"), saveFile))

  return(summaryInfo)
}
```

## [A7] Running sample dataset

```
y = mySample[,"SalePrice.100K."]
x = as.matrix(mySample[,c("Area","Bedrooms","Bathrooms","CarParks", "PropertyType")])

xPred = array(NA, dim = c(5,5))
xPred[1,] = c(600, 2, 2, 1, 1)
xPred[2,] = c(800, 3, 1, 2, 0)
xPred[3,] = c(1500, 2, 1, 1, 0)
xPred[4,] = c(2500, 5, 4, 4, 0)
xPred[5,] = c(250, 3, 2, 1, 1)

dataList <- list(
  x = x ,
  y = y ,
  xPred = xPred ,
  Nx = dim(x)[2] ,
  Ntotal = dim(x)[1]
)


adaptSteps = 2000  # Number of steps to "tune" the samplers
burnInSteps = 3000
nChains = 3
thinSteps = 20 # First run for 3
numSavedSteps = 2000
```

## [A7i] Running sample dataset for exponential model

```
# First run without initials! - for exponential
initsList <- list(
  zbeta0 = 2,
  zbeta = c(20, 0, 0, 0, 0)
)
```

```
aSummaryInfo=runAndPlot(isExp=TRUE, informative=TRUE, filename="exp_1.txt",
                        beta1Sens="0.01", beta2Sens="2", beta4Sens="0.1", beta5Sens="0.01",
                        saveFile="runexp_1.rds", dataList, initsList, nChains, adaptSteps, burnInSteps, num
SavedSteps, thinSteps)
```

## [A7ii] Running sample dataset for gamma model

```
# First run without initials! - for gamma
initsList <- list(
  zbeta0 = 2,
  zbeta = c(0, 0, 0, 0, 0),
  Var = 1000
)
```

```
aSummaryInfo=runAndPlot(isExp=FALSE, informative=TRUE, filename="gamma_1.txt",
                        beta1Sens="0.01", beta2Sens="2", beta4Sens="0.1", beta5Sens="0.01",
                        saveFile="rungamma_1.rds", dataList, initsList, nChains, adaptSteps, burnInSteps, n
umSavedSteps, thinSteps)
```

## [A8] Running full dataset

```
y = myData[,"SalePrice.100K."]
x = as.matrix(myData[,c("Area","Bedrooms","Bathrooms","CarParks", "PropertyType")])
```

```
xPred = array(NA, dim = c(5,5))
xPred[1,] = c(600, 2, 2, 1, 1)
xPred[2,] = c(800, 3, 1, 2, 0)
xPred[3,] = c(1500, 2, 1, 1, 0)
xPred[4,] = c(2500, 5, 4, 4, 0)
xPred[5,] = c(250, 3, 2, 1, 1)
```

```
dataList <- list(
  x = x ,
  y = y ,
  xPred = xPred ,
  Nx = dim(x)[2] ,
  Ntotal = dim(x)[1]
)


adaptSteps = 2000  # Number of steps to "tune" the samplers
burnInSteps = 3000
nChains = 3
thinSteps = 20 # First run for 3
numSavedSteps = 2000
```

Close to the coefficients of the

## [A8i] Running sample dataset for exponential model

```
initsList <- list(
  zbeta0 = 3.5,
  zbeta = c(0, 0, 1, 0.4, -1.5)
)

aSummaryInfo=runAndPlot(isExp=TRUE, informative=TRUE, filename="exp_whole.txt",
                        beta1Sens="0.01", beta2Sens="2", beta4Sens="0.1", beta5Sens="0.01",
                        saveFile="runexp_whole.rds", dataList, initsList, nChains, adaptSteps, burnInSteps,
numSavedSteps, thinSteps)
```

Exponential model ➔ $Y = 3.52 + 0.186X_2 + 1.11X_3 + 0.425X_4 - 1.5X_5$

From sample dataset

## [A8ii] Running sample dataset for gamma model

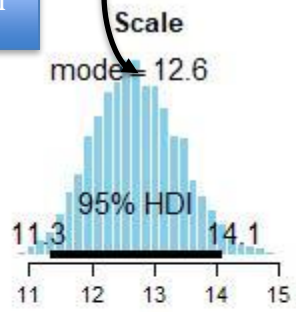
```
initsList <- list(
  zbeta0 = 5.8,
  zbeta = c(0, -0.3, 0.7, 0.3, -0.4),
  Var = 13
)
```

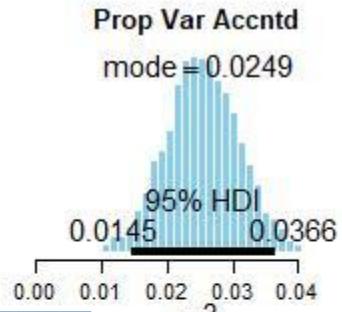
Close to the coefficients of the

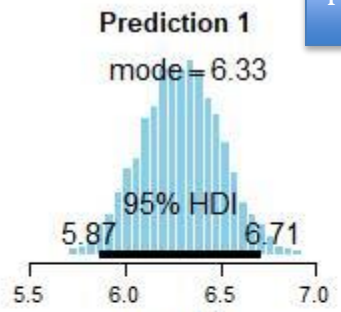
Gamma model ➔ $Y = 5.8 - 0.345X_2 + 0.691X_3 + 0.276X_4 - 0.404X_5$

Close to value of

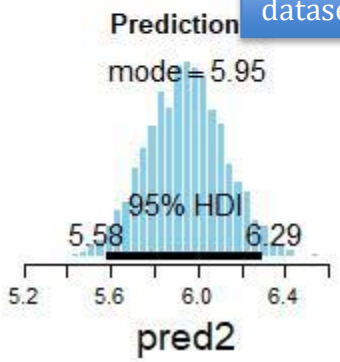
From sample dataset

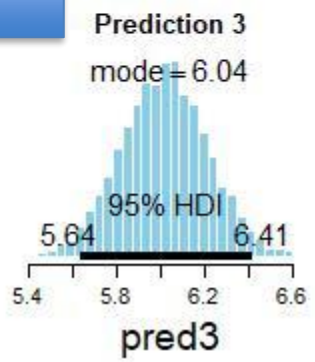


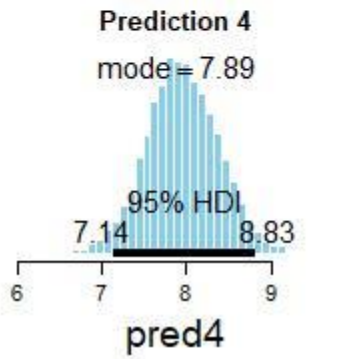




From sample dataset



```
aSummaryInfo=runAndPlot(isExp=FALSE, informative=TRUE, filename="gamma_whole.txt",
                        beta1Sens="0.01", beta2Sens="2", beta4Sens="0.1", beta5Sens="0.01",
                        saveFile="rungamma_whole.rds", dataList, initsList, nChains, adaptSteps,
burnInSteps, numSavedSteps, thinSteps)
```

## [A9] Goodness of fit

### [A9i] exponential model

```
coefficients <- aSummaryInfo[8:13,3] # Get the model coefficients out
Variance <- aSummaryInfo[14,3] # Get the variance out
# Since we imposed the regression model on the mean of the gamma likelihood,
# we use the model (X*beta) to generate the mean of gamma population for each
# observed x vector.
meanExp <- as.matrix(cbind(rep(1,nrow(x)),  x)) %*% as.vector(coefficients)
# Generate random data from the posterior distribution. H ere I take the
# reparameterisation back to alpha and beta.
randomData <- rexp(n= 1000, rate=1/meanExp)


# Display the density plot of observed data and posterior distribution:
predicted <- data.frame(price = randomData)
observed <- data.frame(price = y)
predicted$type <- "Predicted"
observed$type <- "Observed"
dataPred <- rbind(predicted, observed)
hist(observed$price, breaks=40)
hist(predicted$price, breaks=40)


ggplot(dataPred, aes(price, fill = type)) + geom_density(alpha = 0.2)
ggplot(dataPred, aes(price, fill = type)) + geom_histogram(bins=40, color="#e9ecef", alpha=0.4)

ggplot(dataPred, aes(x = price)) +
```

```
  geom_histogram(aes(color = type, fill = type),
                 position = "identity", bins = 100, alpha = 0.4) +
  scale_color_manual(values = c("#00AFBB", "#E7B800")) +
  scale_fill_manual(values = c("#00AFBB", "#E7B800"))
```

## [A9ii] gamma model

```
coefficients <- aSummaryInfo[8:13,3] # Get the model coefficients out
Variance <- aSummaryInfo[14,3] # Get the variance out
# Since we imposed the regression model on the mean of the gamma likelihood,
# we use the model (X*beta) to generate the mean of gamma population for each
# observed x vector.

meanGamma <- as.matrix(cbind(rep(1,nrow(x)),  x)) %*% as.vector(coefficients)
# Generate random data from the posterior distribution. Here I take the
# reparameterisation back to alpha and beta.
randomData <- rgamma(n= 10000,shape=meanGamma^2/Variance, rate = meanGamma/Variance)

# Display the density plot of observed data and posterior distribution:
predicted <- data.frame(price = randomData)
observed <- data.frame(price = y)
predicted$type <- "Predicted"
observed$type <- "Observed"
dataPred <- rbind(predicted, observed)
hist(observed$price, breaks=100)
hist(predicted$price, breaks=100)

ggplot(dataPred, aes(price, fill = type)) + geom_density(alpha = 0.2)
ggplot(dataPred, aes(price, fill = type)) + geom_histogram(binwidth=1, color="#e9ecef", alpha=0.4)

ggplot(dataPred, aes(x = price)) +
  geom_histogram(aes(color = type, fill = type),
                 position = "identity", bins = 100, alpha = 0.4) +
  scale_color_manual(values = c("#00AFBB", "#E7B800")) +
```

```
  scale_fill_manual(values = c("#00AFBB", "#E7B800"))
```

## [A10] Diagnostic plots of gamma model on full dataset.

*Figure 26 - diagnostic plot of beta0 (Intercept) of gamma model on full run*

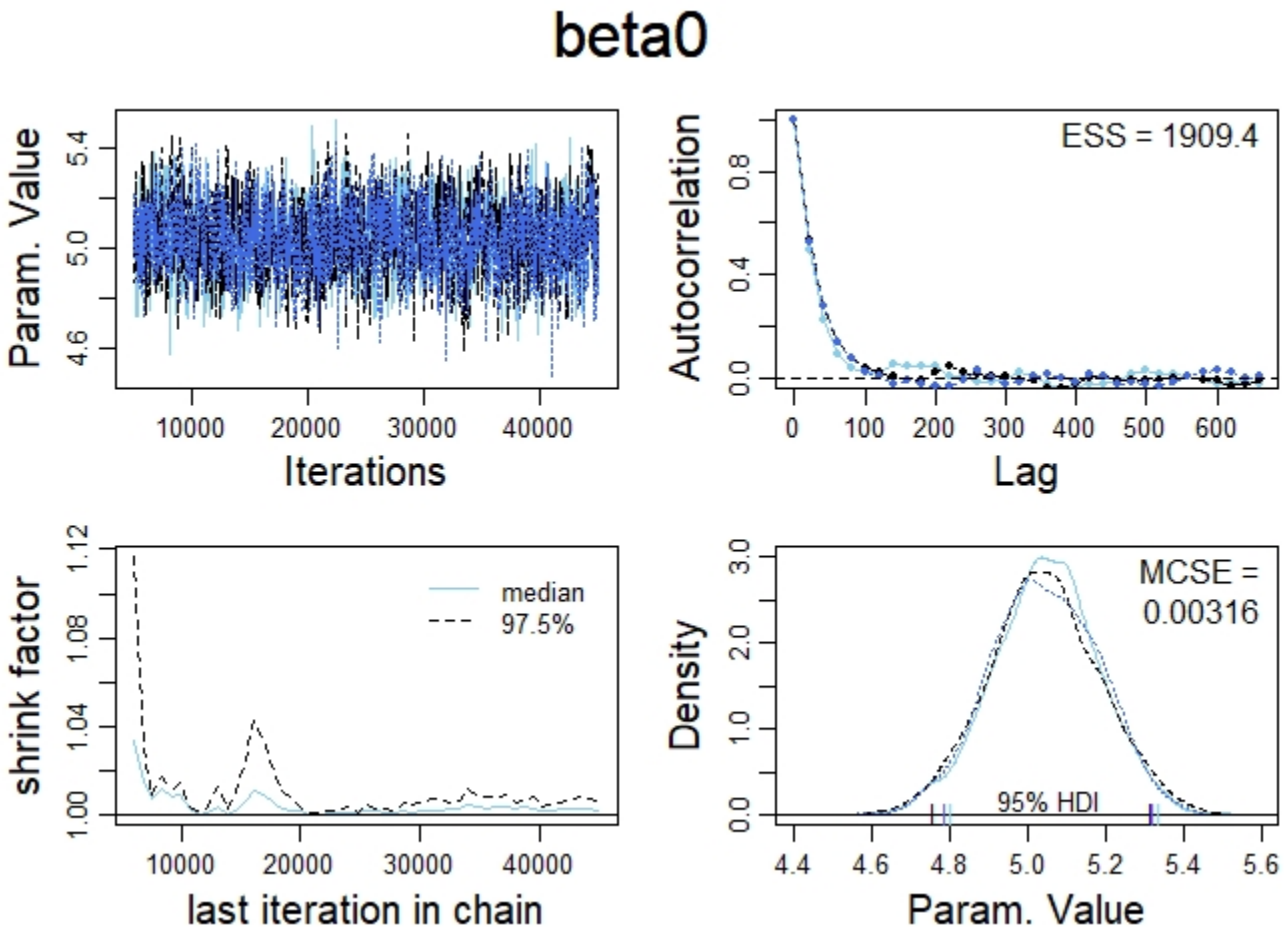


*Figure 27 - diagnostic plot of beta1 (area) of gamma model on full run*

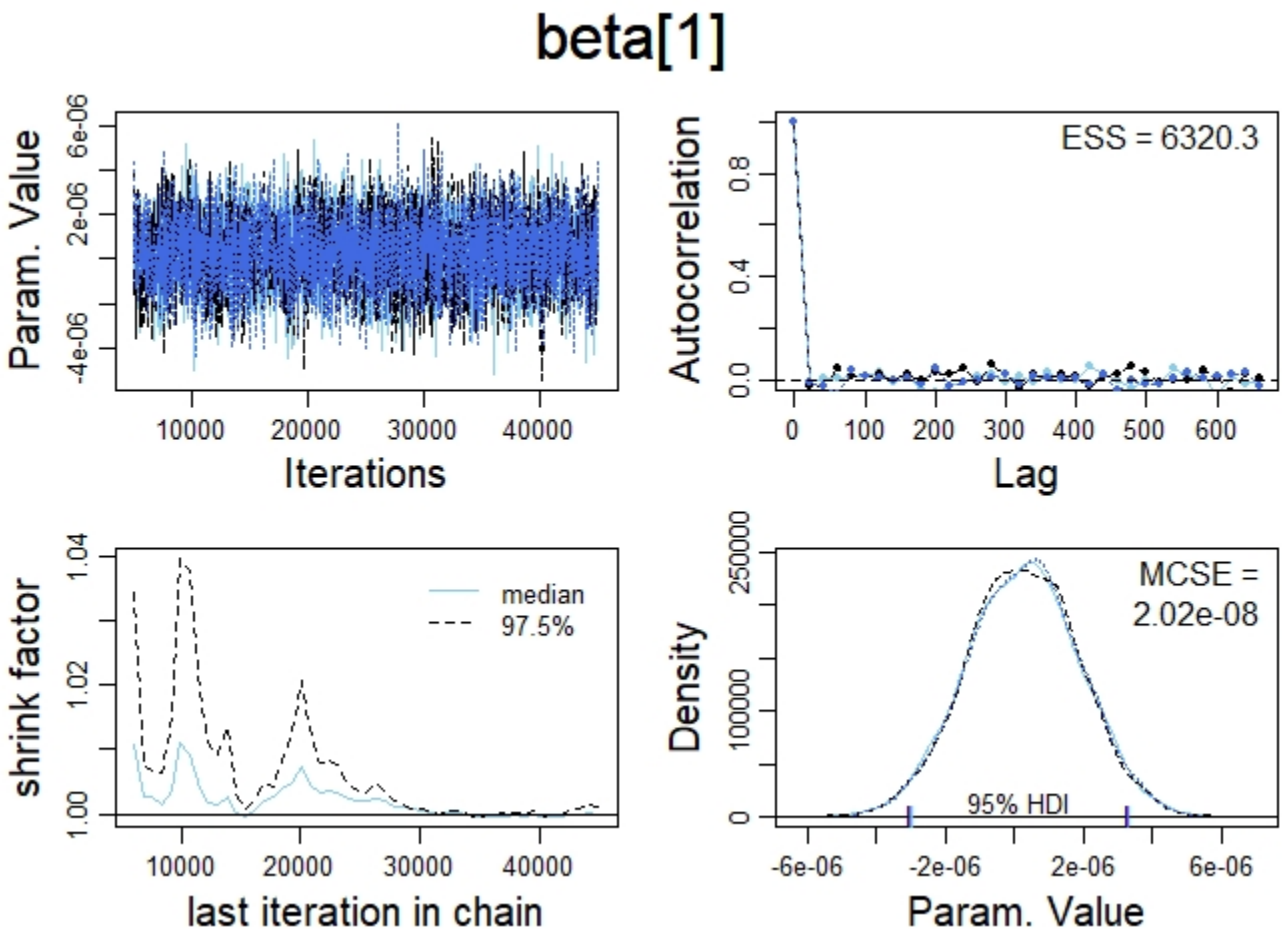


*Figure 28 - diagnostic plot of beta2 (bedrooms) of gamma model on full run*

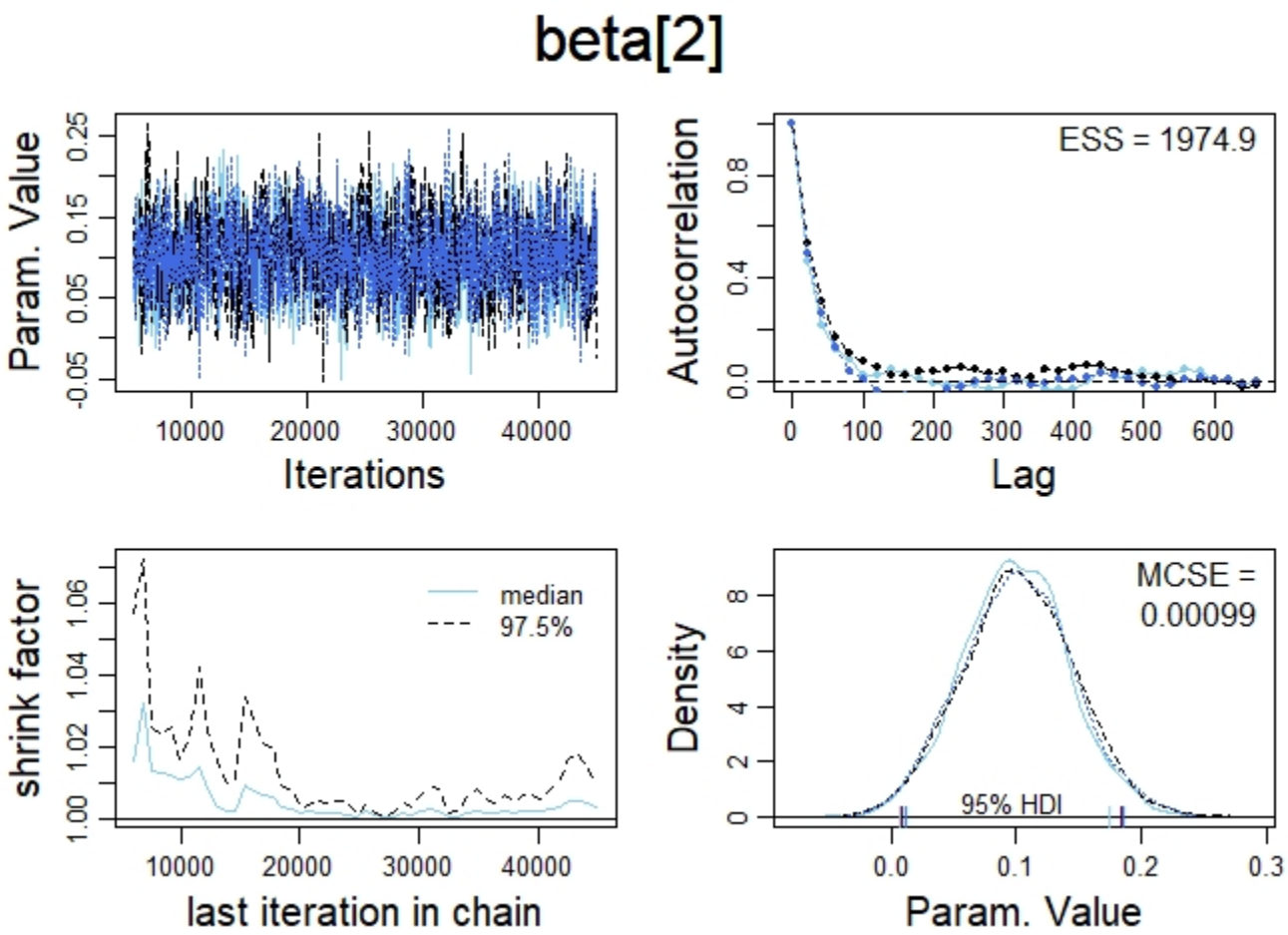


*Figure 29- diagnostic plot of beta3 (bathrooms) of gamma model on full run*

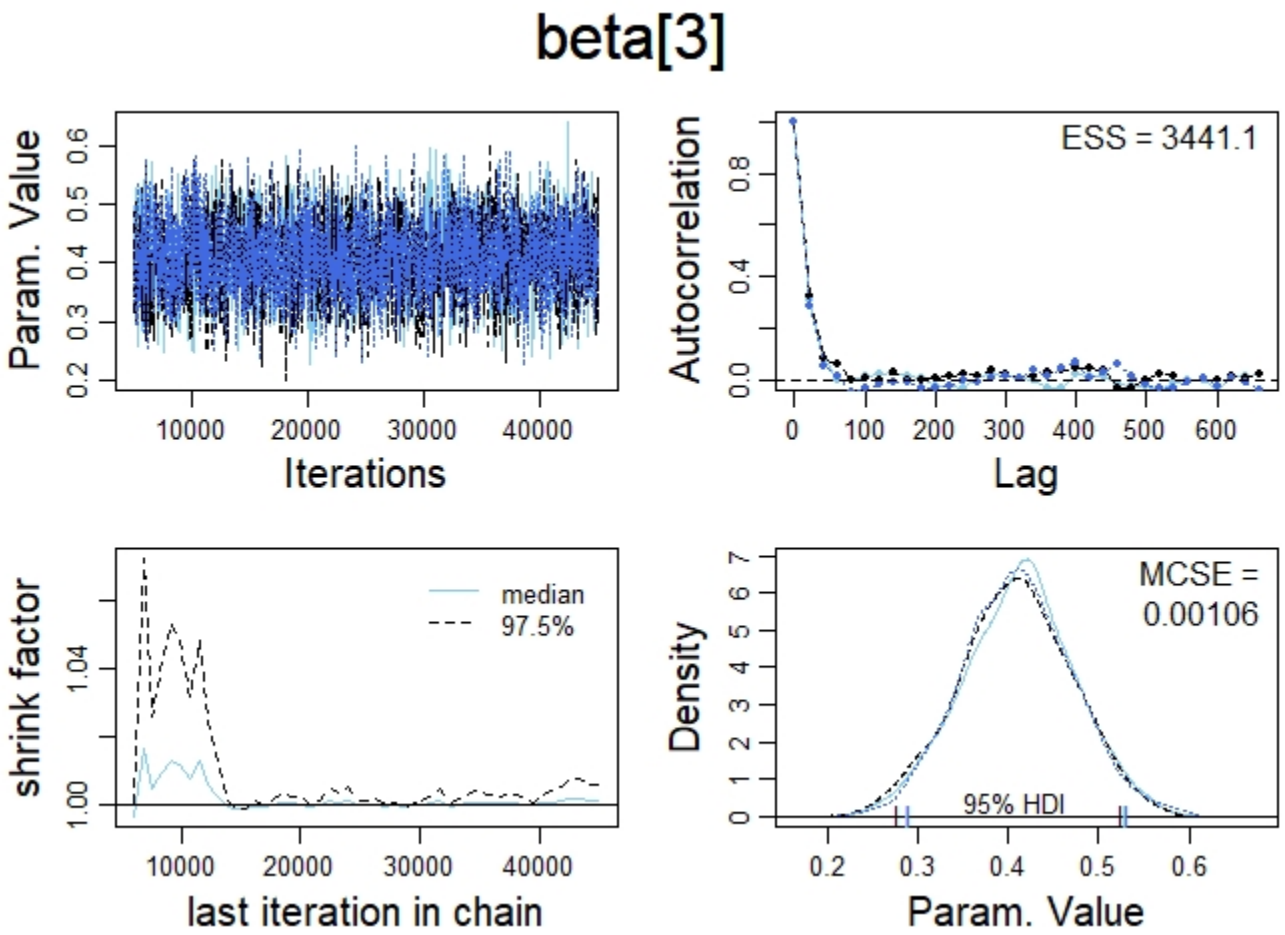


*Figure 30- diagnostic plot of beta4 (carparks) of gamma model on full run*

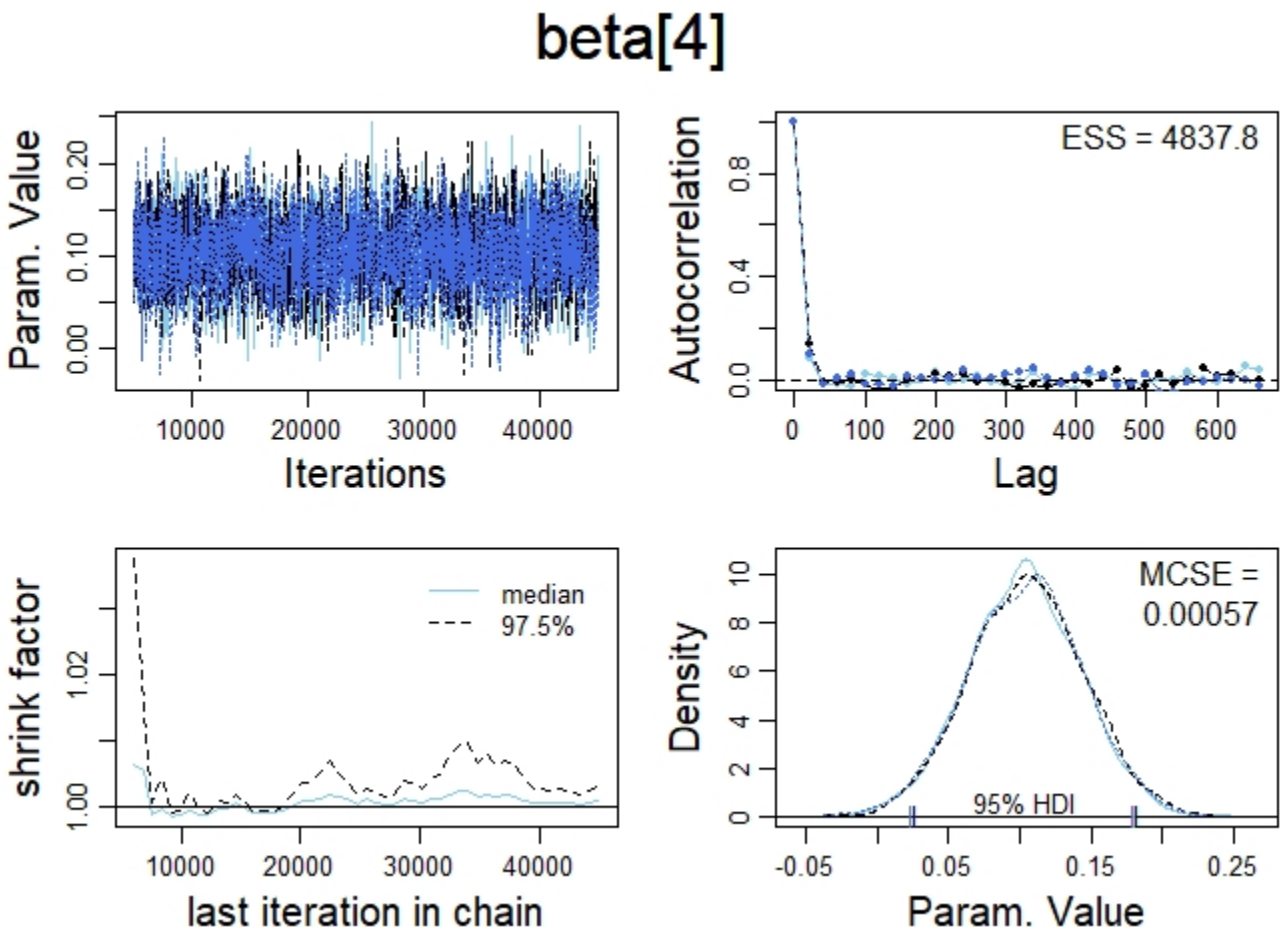


*Figure 31- diagnostic plot of beta5 (property type) of gamma model on full run*

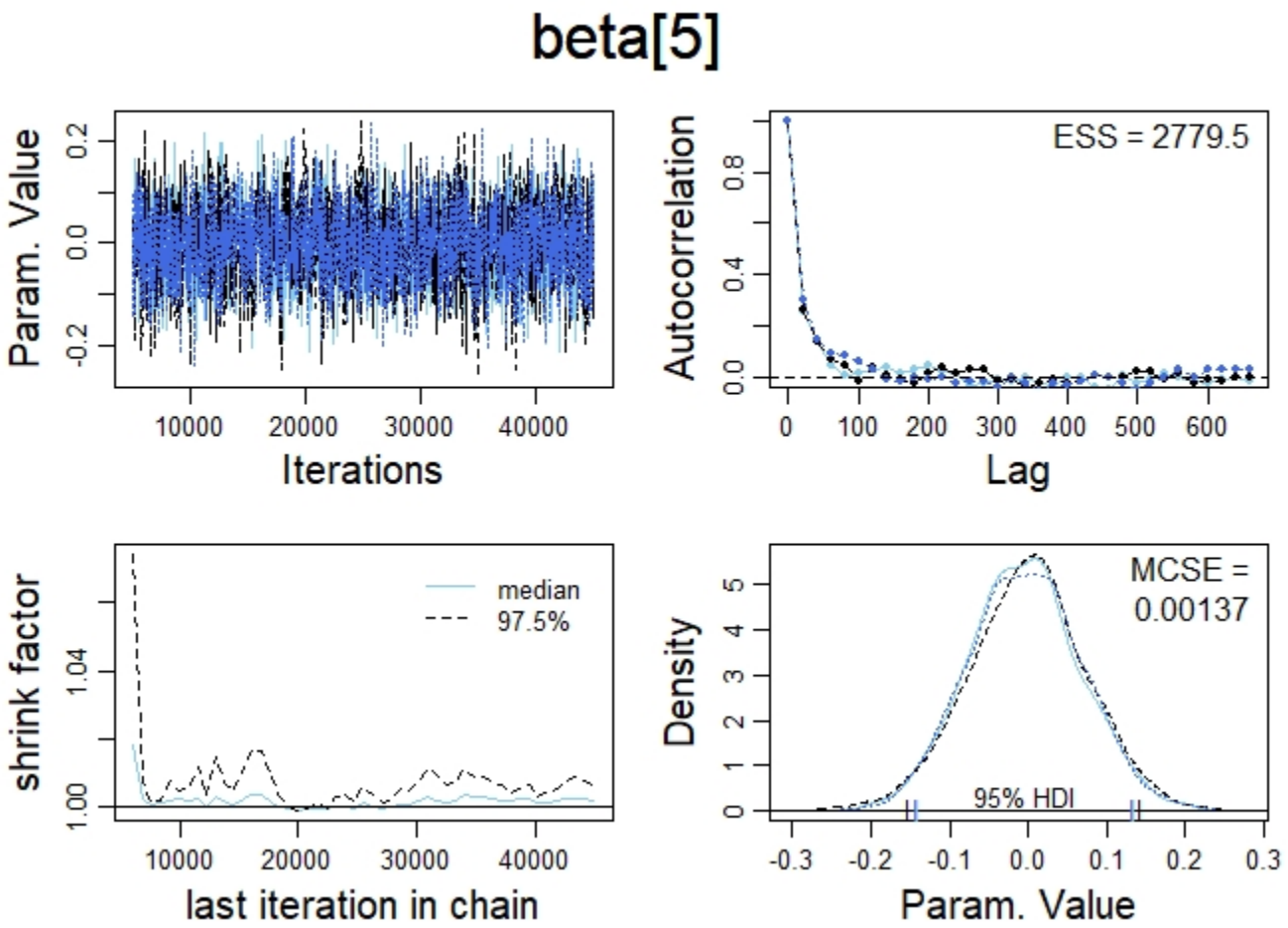


*Figure 32 - diagnostic plot of tau (variance) of gamma model on full run*

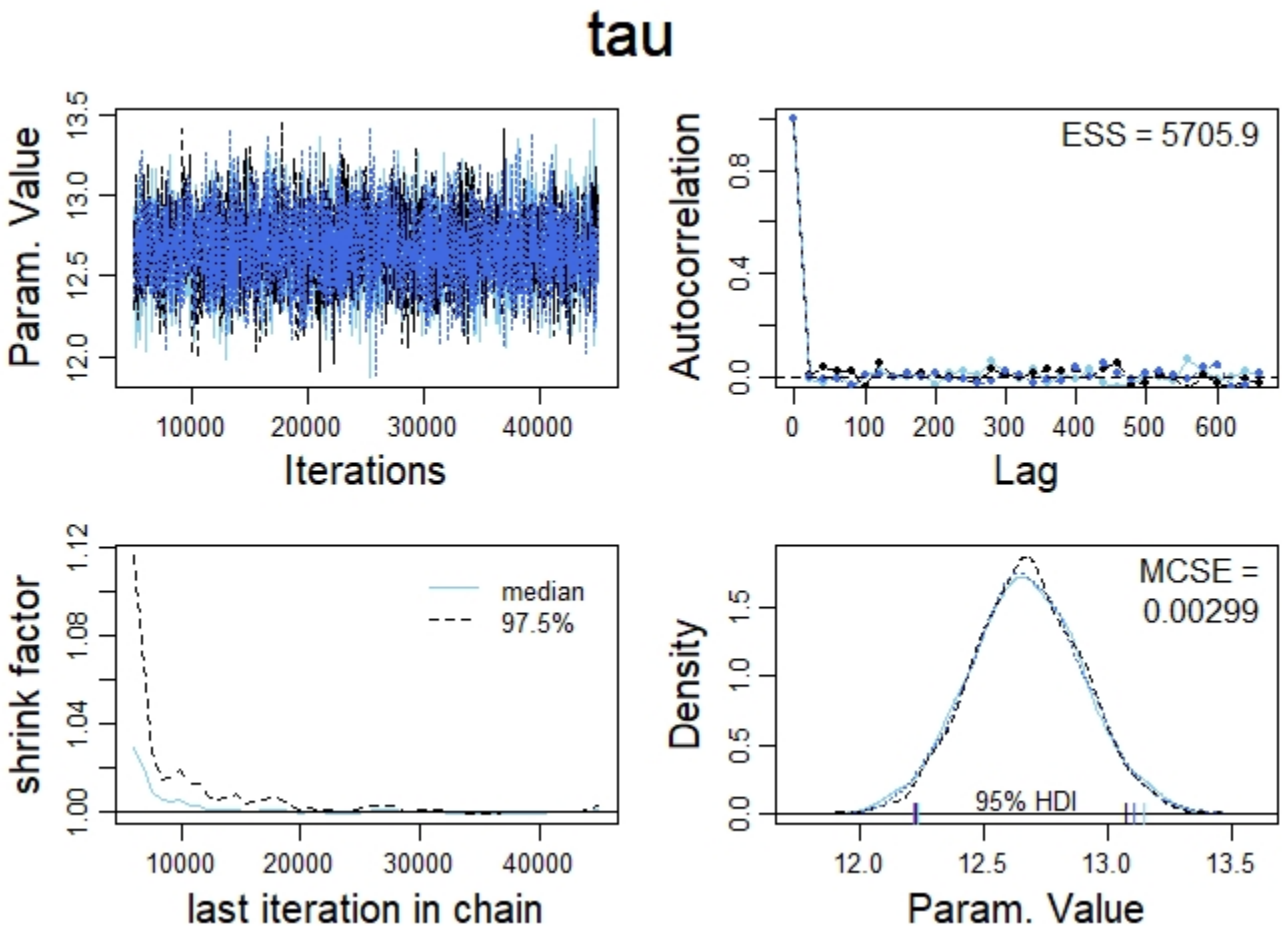


*Figure 33 - diagnostic plot of prediction 1 of gamma model on full run*

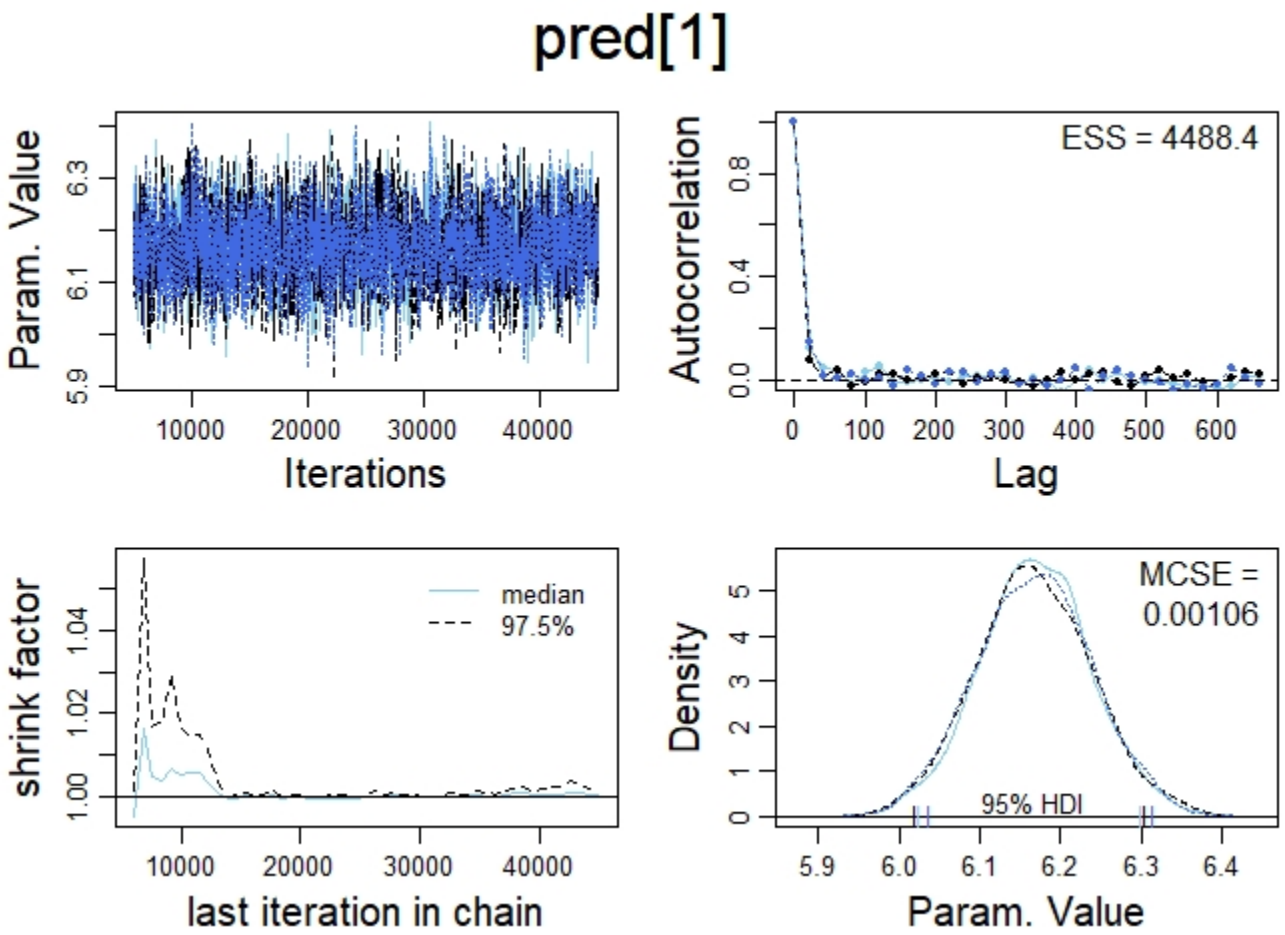


*Figure 34 - diagnostic plot of prediction 2 of gamma model on full run*

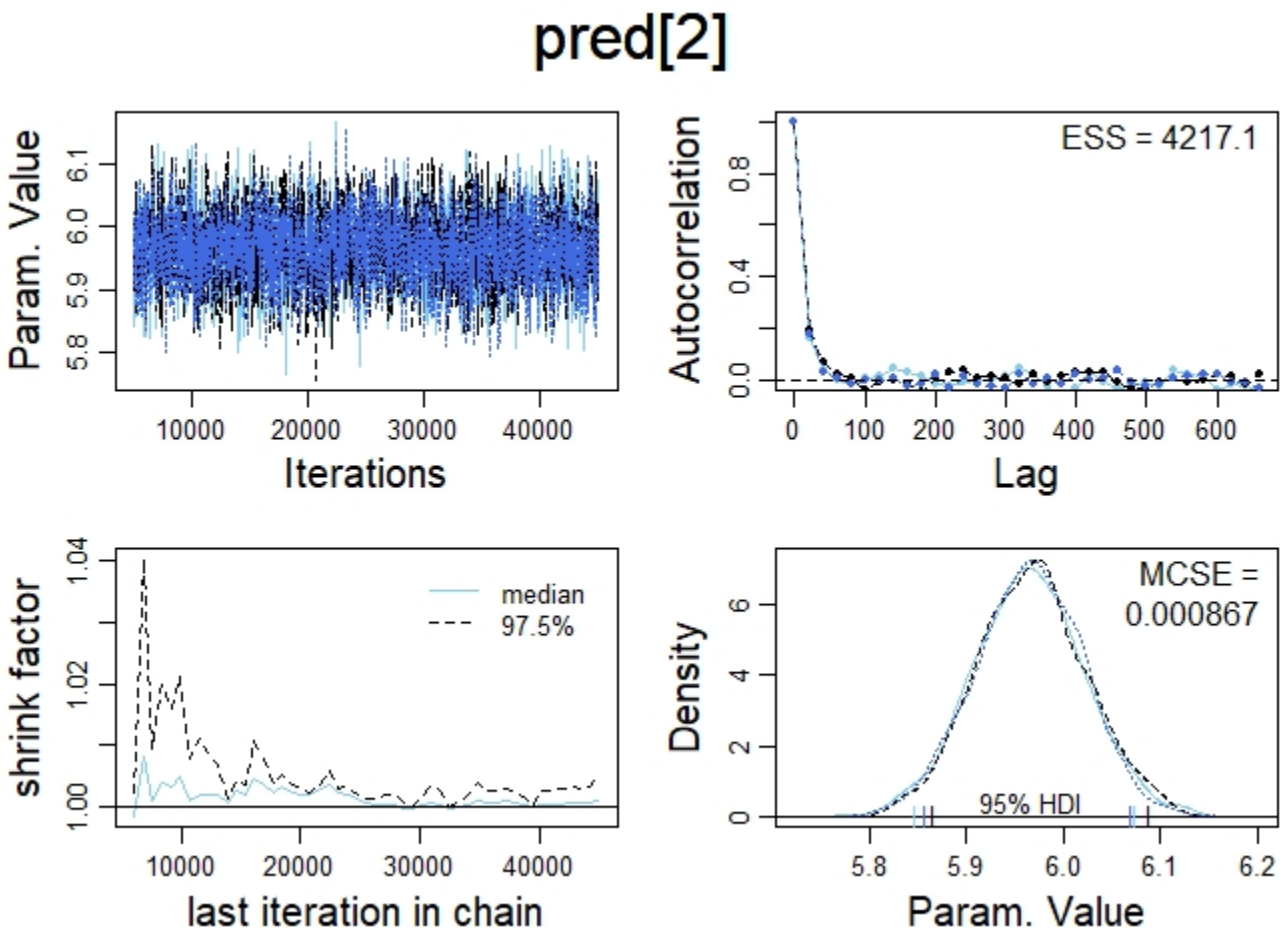


*Figure 35 - diagnostic plot of prediction 3 of gamma model on full run*

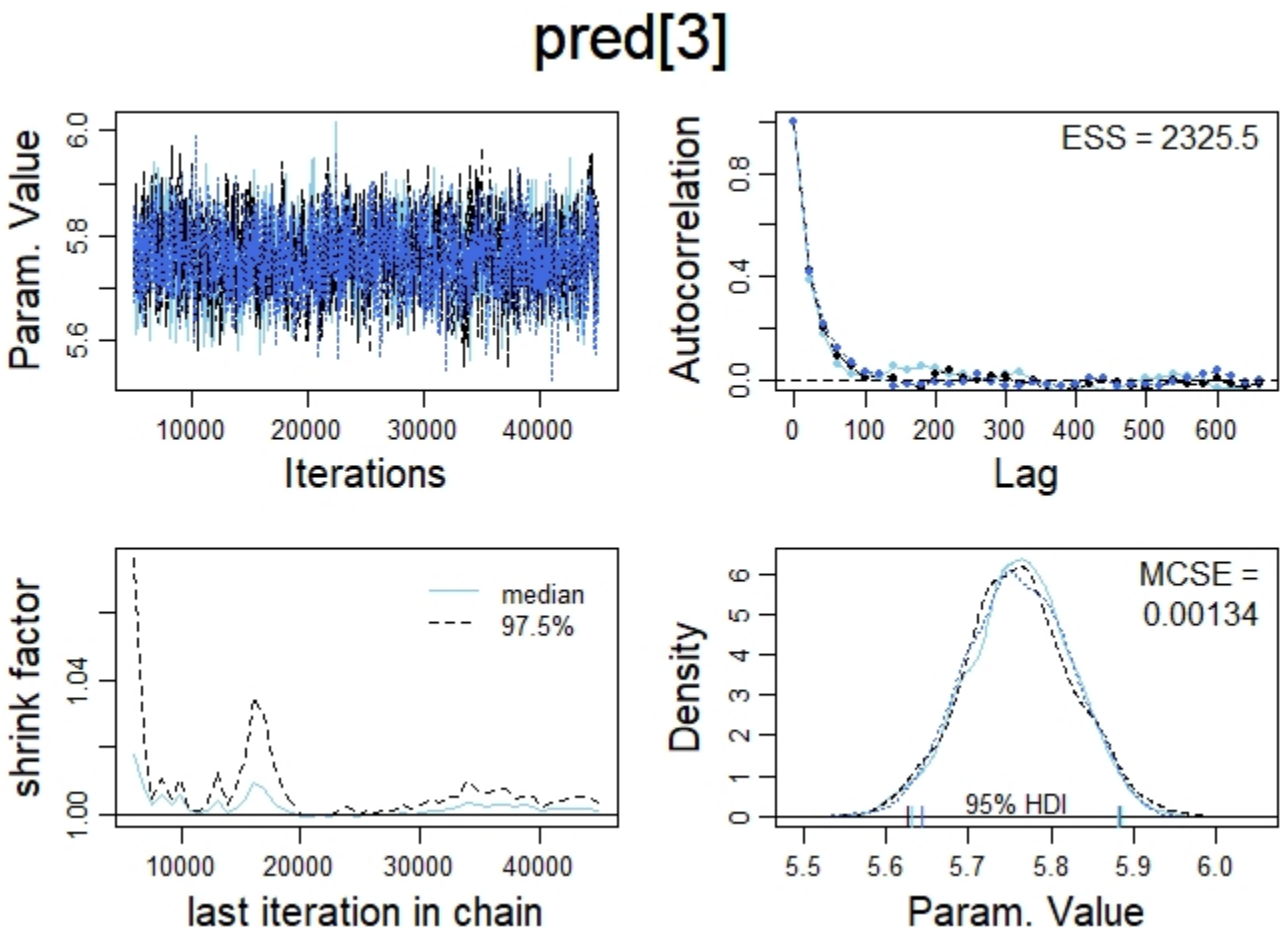


*Figure 36 - diagnostic plot of prediction 4 of gamma model on full run*

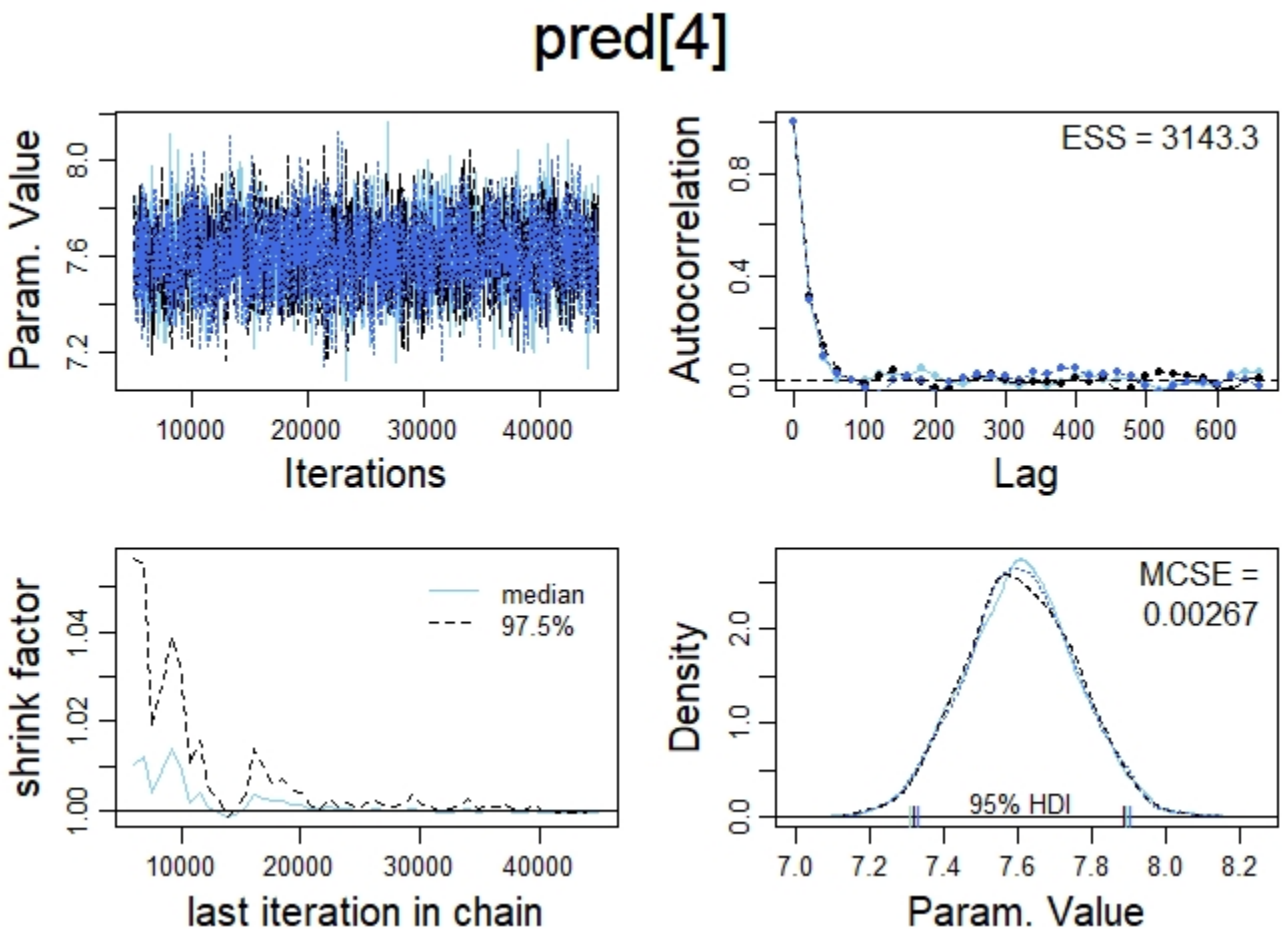
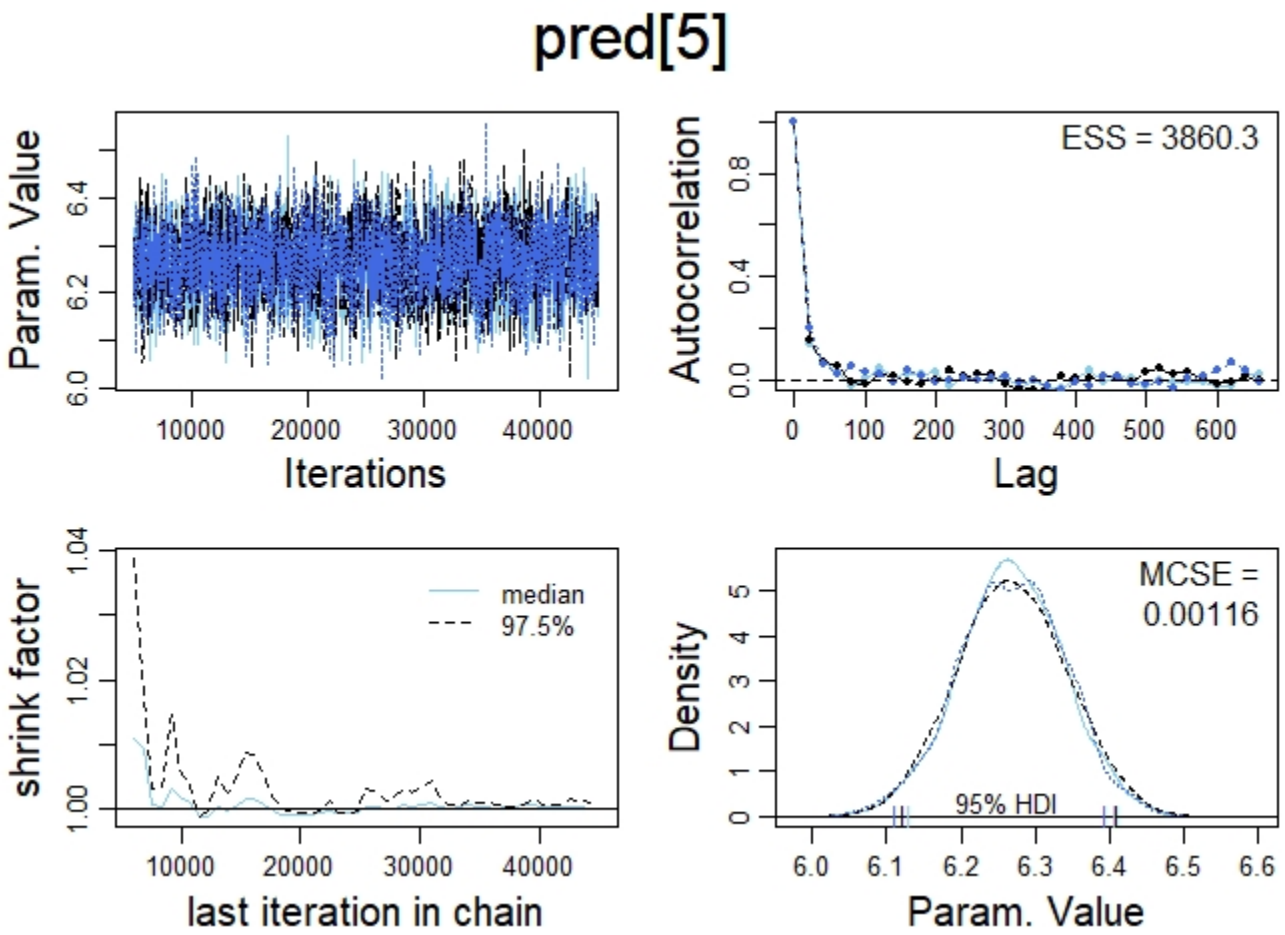


*Figure 37 - diagnostic plot of prediction 5 of gamma model on full run*



## [A11] Diagnostic plots of exponential model on full dataset.

*Figure 38 - diagnostic plot of beta0 (Intercept) of exponential model on full run*

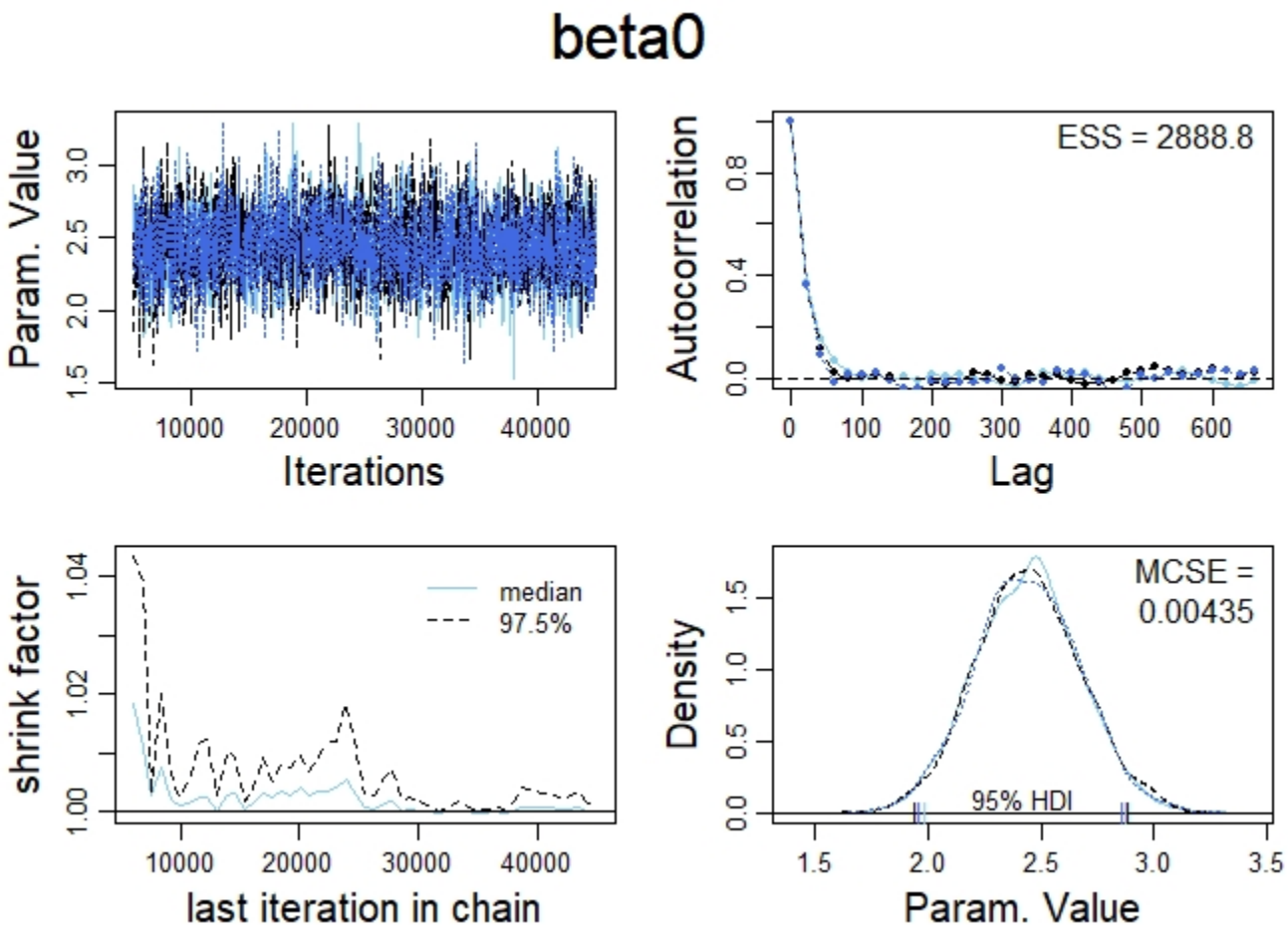



Figure 39 - diagnostic plot of beta1 (Area) of exponential model on full run

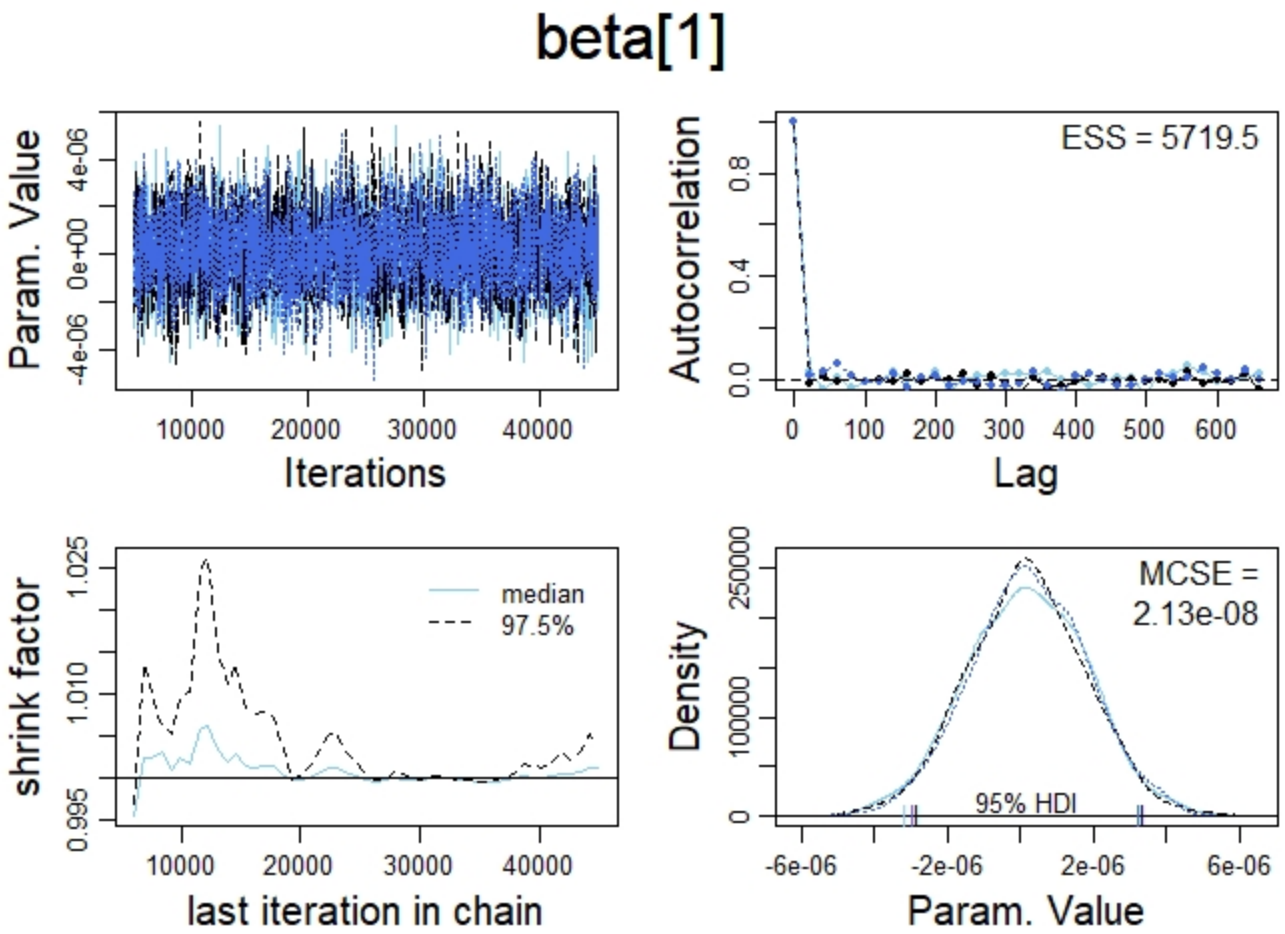



Figure 40 - diagnostic plot of beta2 (bedrooms) of exponential model on full run

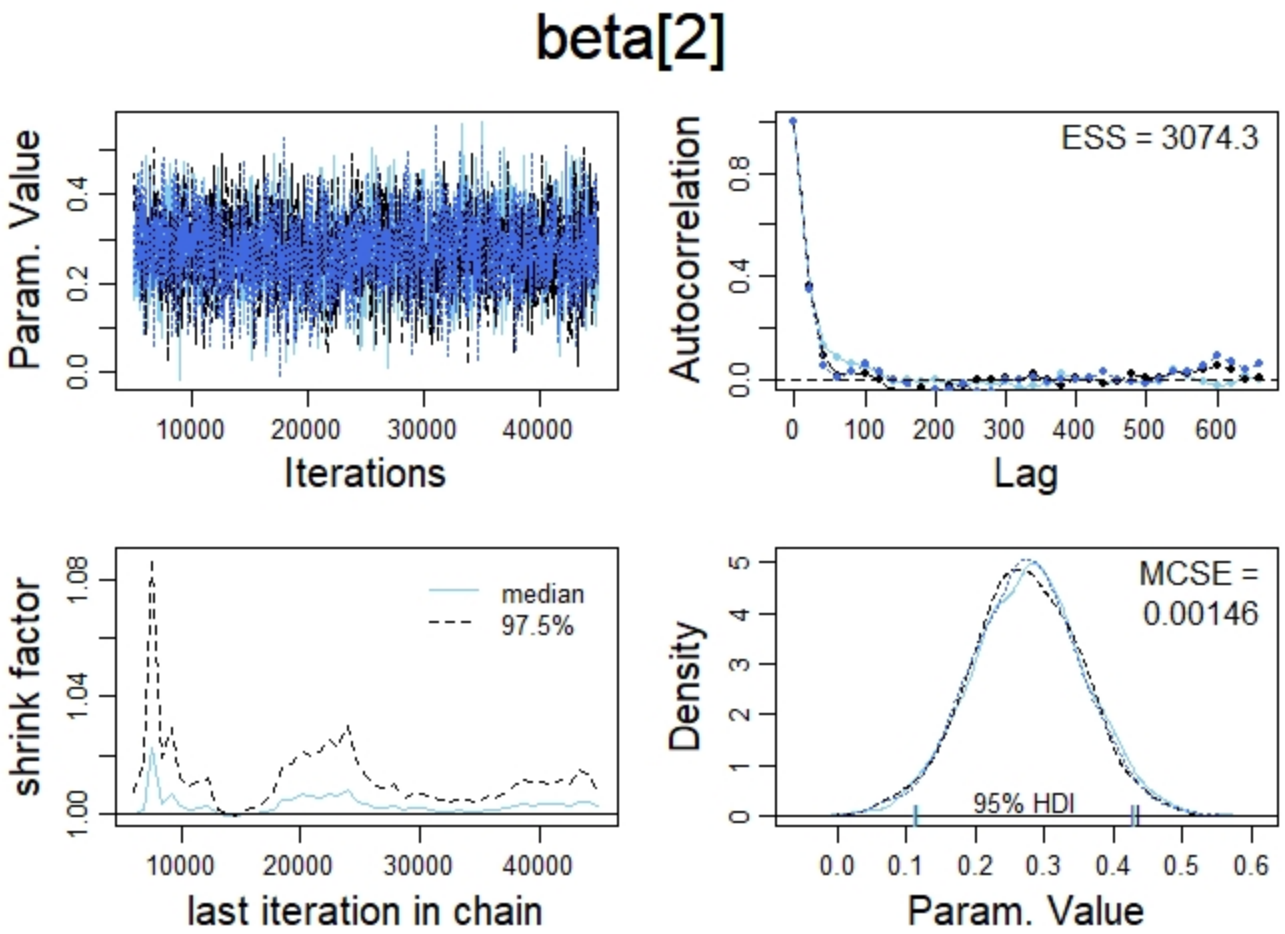



Figure 41 - diagnostic plot of beta3 (bathrooms) of exponential model on full run

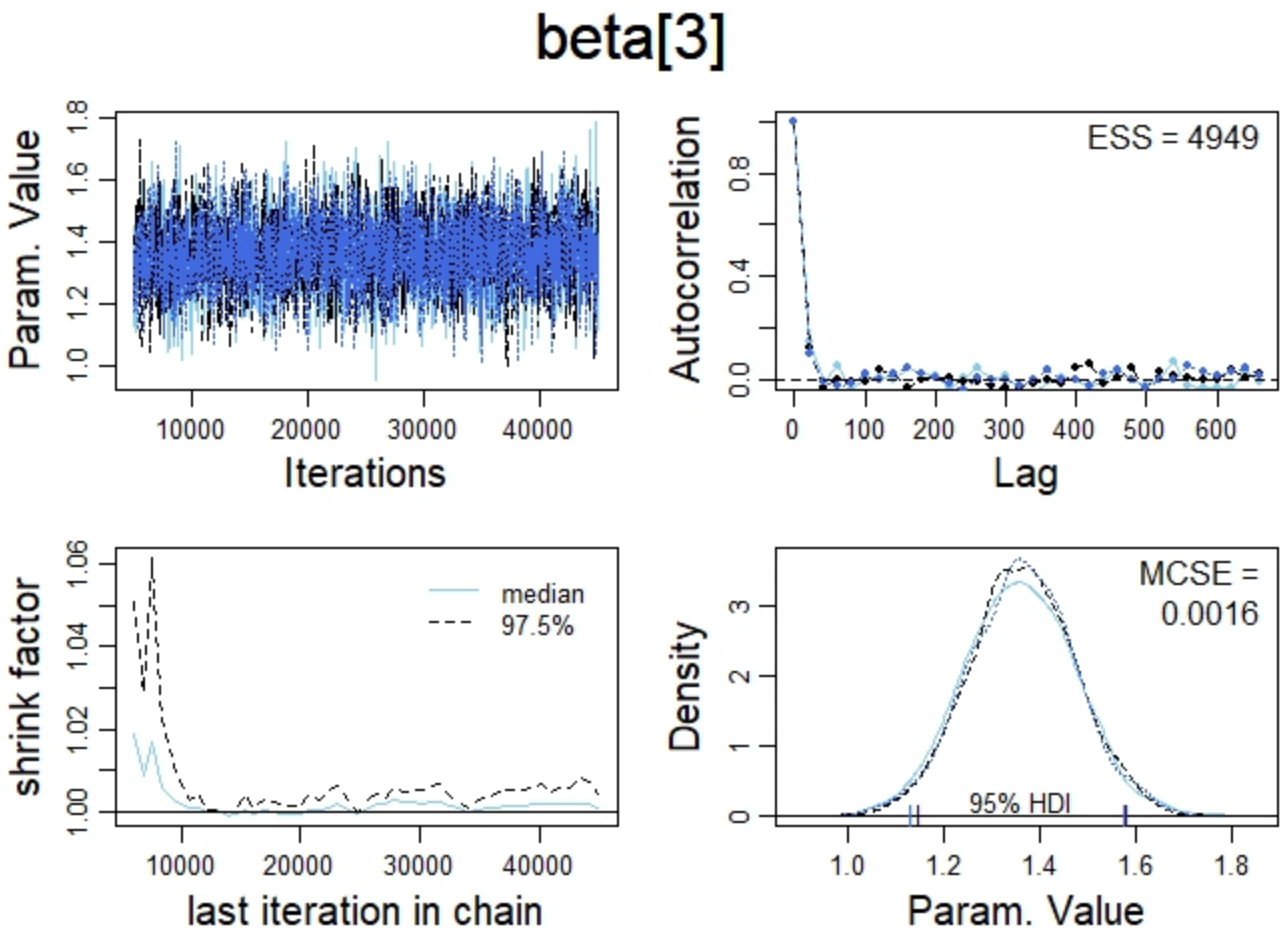
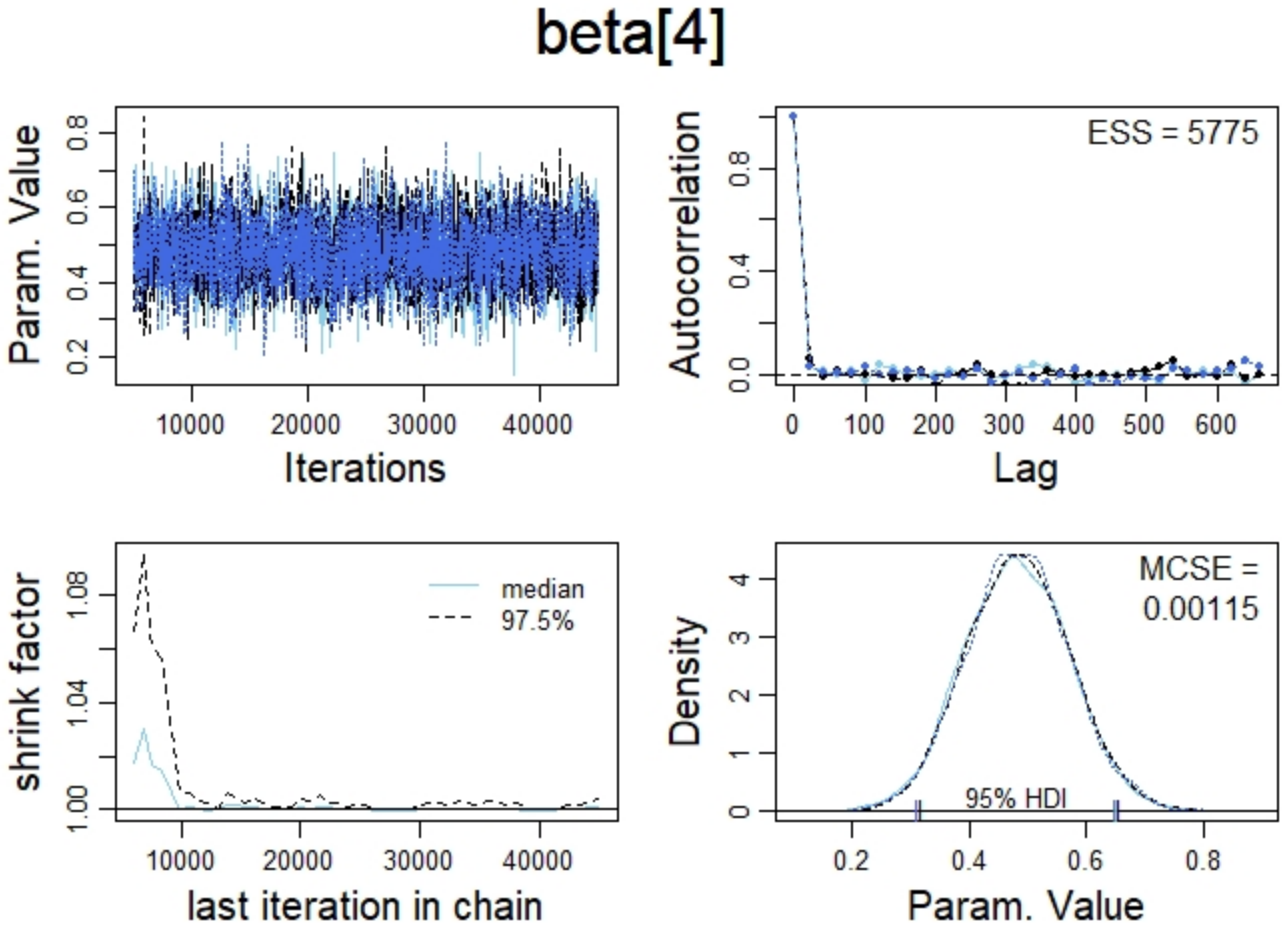



Figure 42 - diagnostic plot of beta4 (Carparks) of exponential model on full run


beta[4]
Param. Value
Iterations
Autocorrelation
ESS = 5775
Lag
shrink factor
median
97.5%
last iteration in chain
Density
MCSE =
0.00115
95% HDI
Param. Value

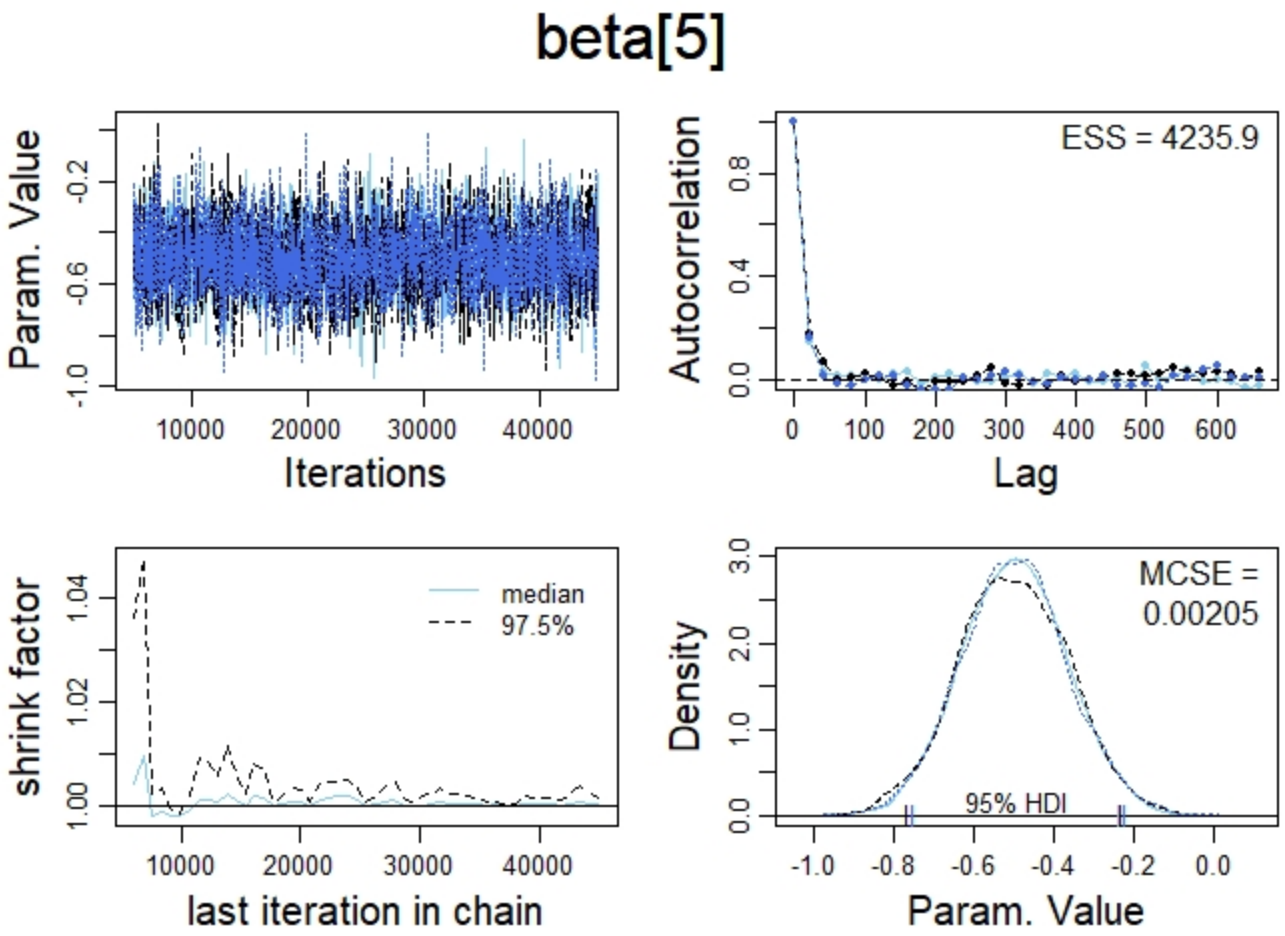



Figure 43 - diagnostic plot of prediction 1 of exponential model on full run

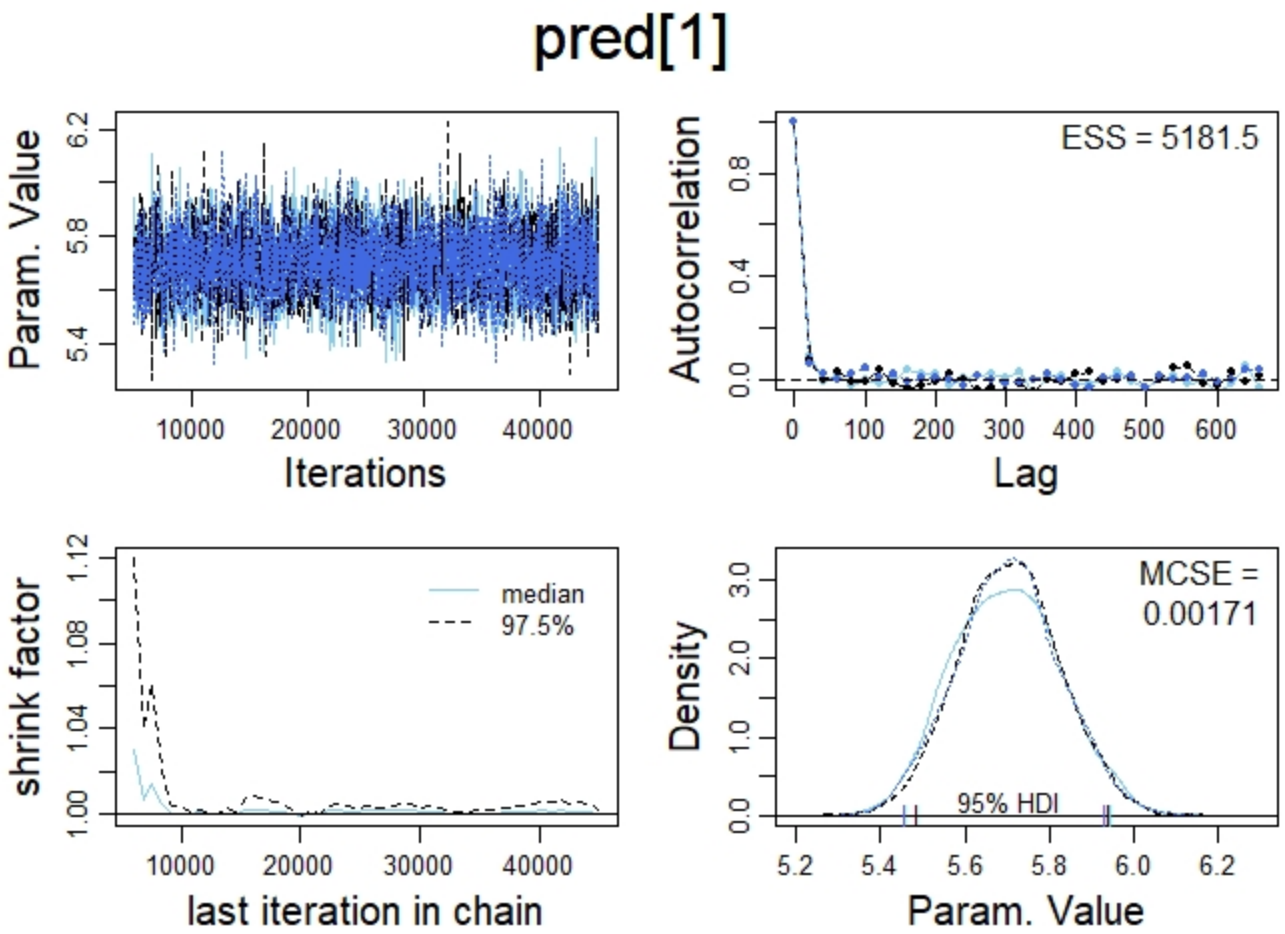



Figure 44 - diagnostic plot of prediction 2 of exponential model on full run

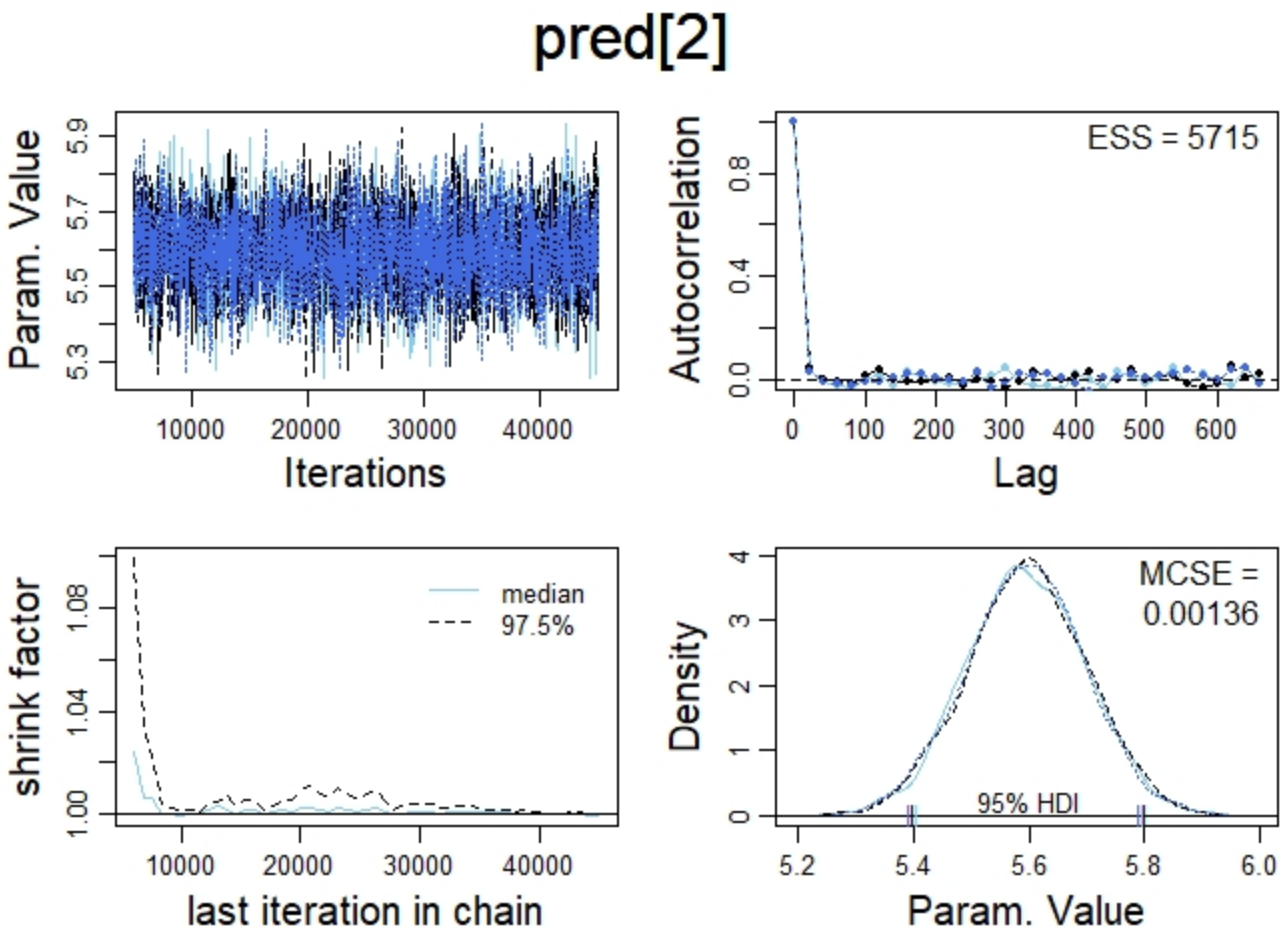



Figure 45 - diagnostic plot of prediction 3 of exponential model on full run

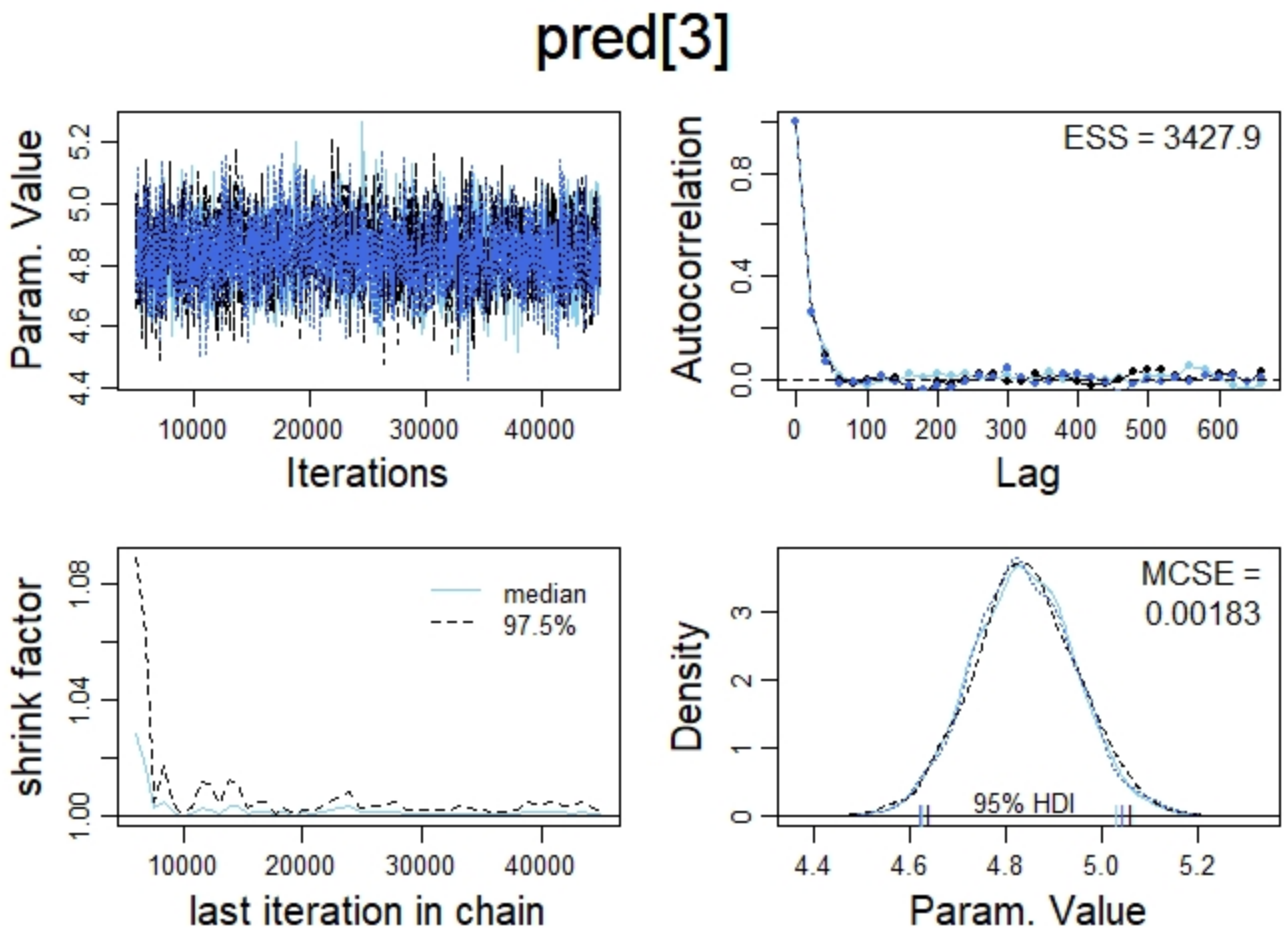



Figure 46 - diagnostic plot of prediction 4 of exponential model on full run

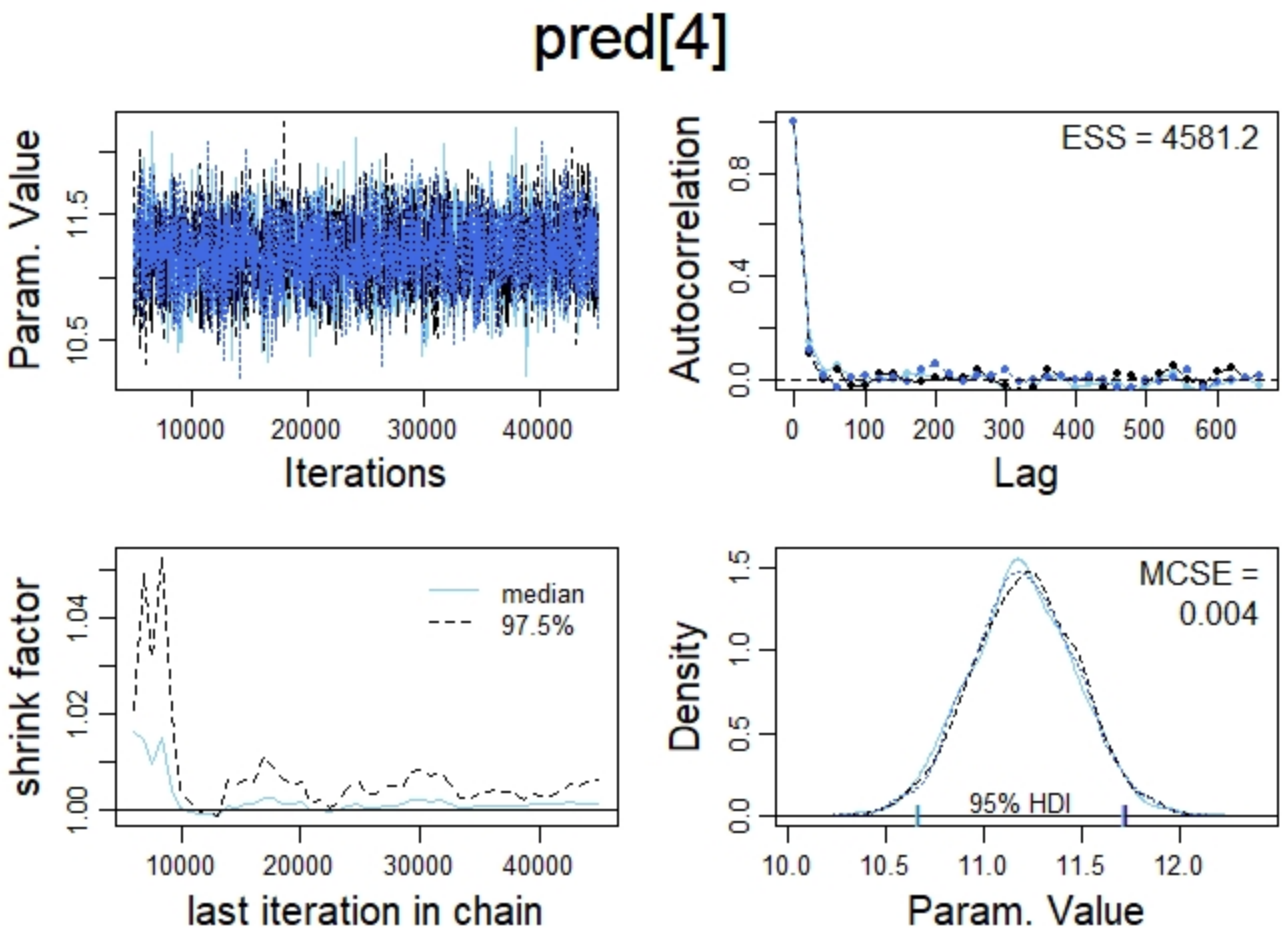
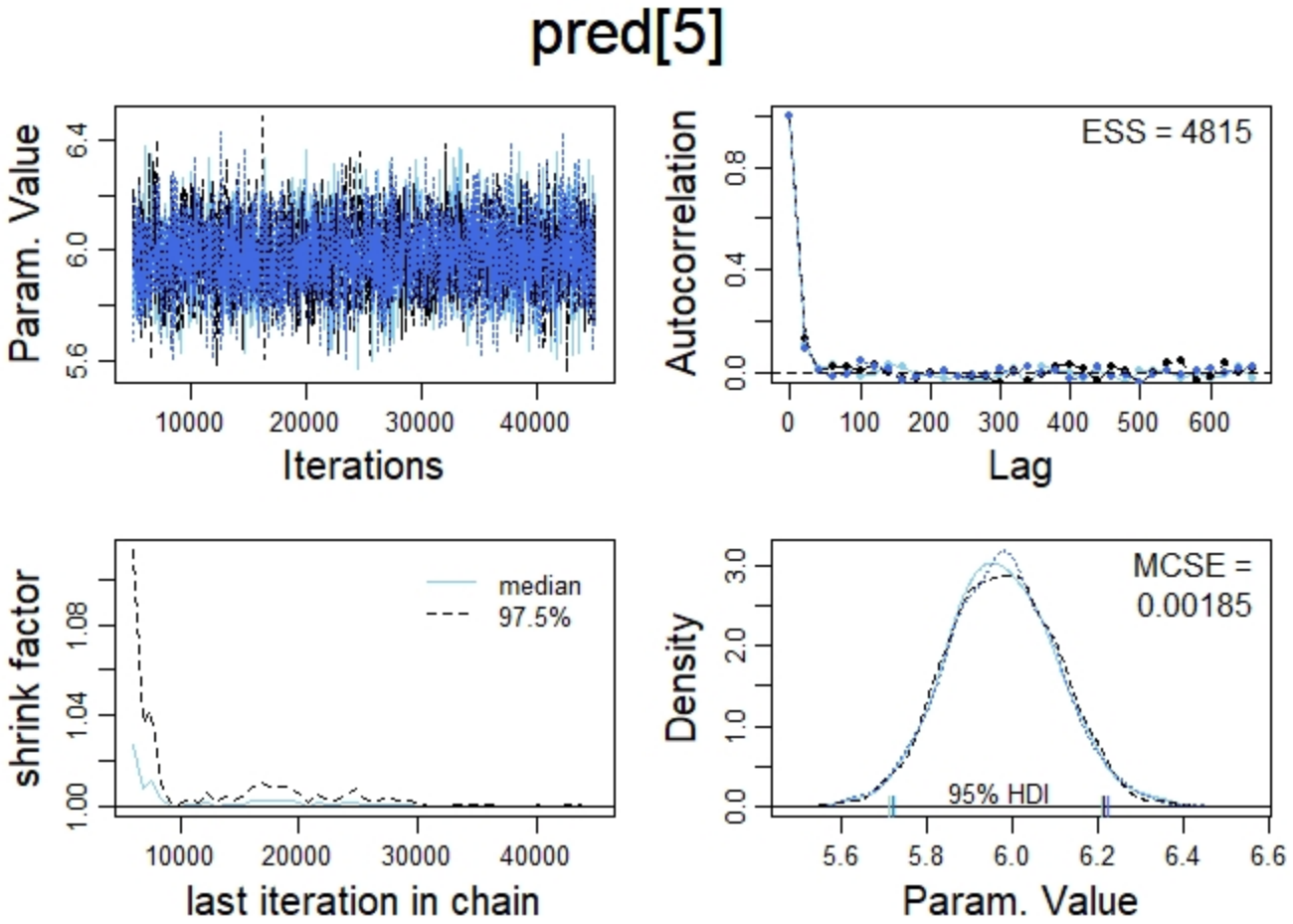



Figure 47 - diagnostic plot of prediction 5 of exponential model on full run


pred[5]
Param. Value
5.6
6.0
6.4
10000
20000
30000
40000
Iterations
Autocorrelation
0.0
0.4
0.8
ESS = 4815
0
100
200
300
400
500
600
Lag
shrink factor
1.00
1.04
1.08
median
97.5%
10000
20000
30000
40000
last iteration in chain
Density
0.0
1.0
2.0
3.0
MCSE =
0.00185
95% HDI
5.6
5.8
6.0
6.2
6.4
6.6
Param. Value