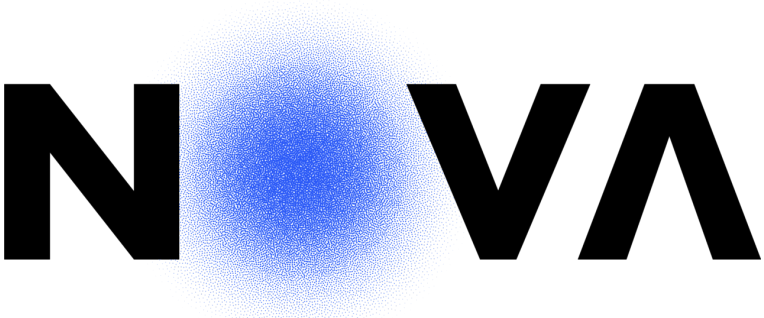


NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

Bayesian Methods Report

Bayesian Methods
AY 2022/2023

Francisco Cardoso 60252
Riccardo Galarducci 66819
Tomaso Castellani 66906

Contents

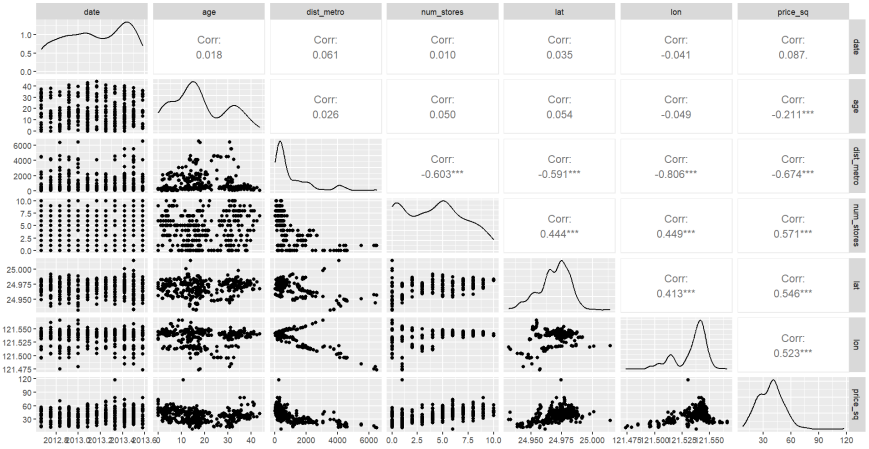
1 Exercise 1

1.1 Data Analysis

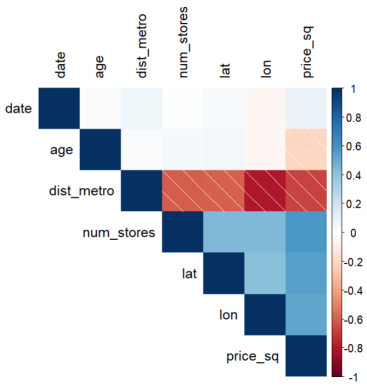
The first task of the project is to tackle a multiple regression problem. The goal is to predict the target variable Y *House Price per square feet* using, as regressors, the other variables that constitute the data set, which are:

- X_1: Transaction date
- X_2: House age (years)
- X_3: Distance to the nearest metro station (feet)
- X_4: Number of nearby convenience stores
- X_5: Latitude
- X_6: Longitude

We began by exploring the relationships between the variables, graphically and numerically through pairs plots (Figure 1a) and Pearson Correlation coefficient (Figure 1b).



(a) Pairs plot



(b) correlation plot

Figure 1: Relationship between variables

From the last row in Figure 1a, we can see a linear dependence between *price_sq* and the other variables. This is confirmed also by 1b in which we can see a relatively high correlation coefficient between *price_sq* and all the other variables, except for the variables *date* and *age* for which the correlation with *price_sq* is close to zero.

1.2 Fitting the Model

The Bayesian model we exploited is the following:

- **Data**

$$Y_i \sim N(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_1 X_{i6}, \sigma^2), i = 1, ..., n \tag{1}$$

- **Likelihood**

$$l(\alpha, \beta_1, ..., \beta_6, \sigma^2 | \mathbf{y}) = \prod_{i=1}^{n} \phi(y_i | \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_6 X_{i6}, \sigma^2) \tag{2}$$

- **Priors**: We assume Uninformative Priors

$$\alpha, \beta_1, ..., \beta_6 \sim Uniform \tag{3}$$

$$log(\sigma) \sim Uniform \tag{4}$$

1.3 Model Diagnostics

We used Monte Carlo Markov Chain (MCMC) to fit the model mentioned earlier. The trace plots shown in Figure 11 indicate that the four generated chains are well mixed, with no apparent seasonal or increasing/decreasing behavior. Moreover, all the effective sample size ratios are greater than 0.1 and the $\hat{R}$ are less than 1.05. This means that we can trust the simulation produced by the MCMC algorithm and that we have stability across parallel chains.

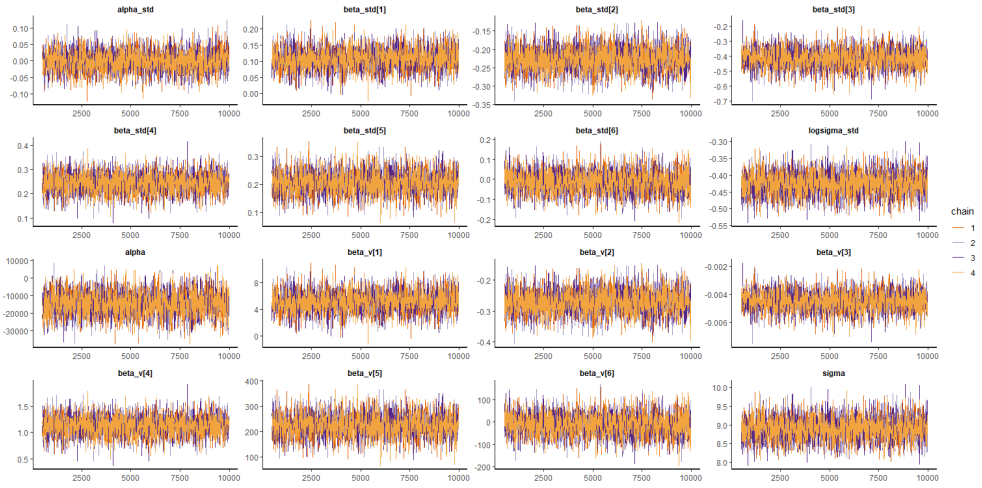


Figure 2: Trace plots

1.4 Model Posteriors

In the following we summarise graphically and numerically the posteriors of the parameters obtained through the simulation.

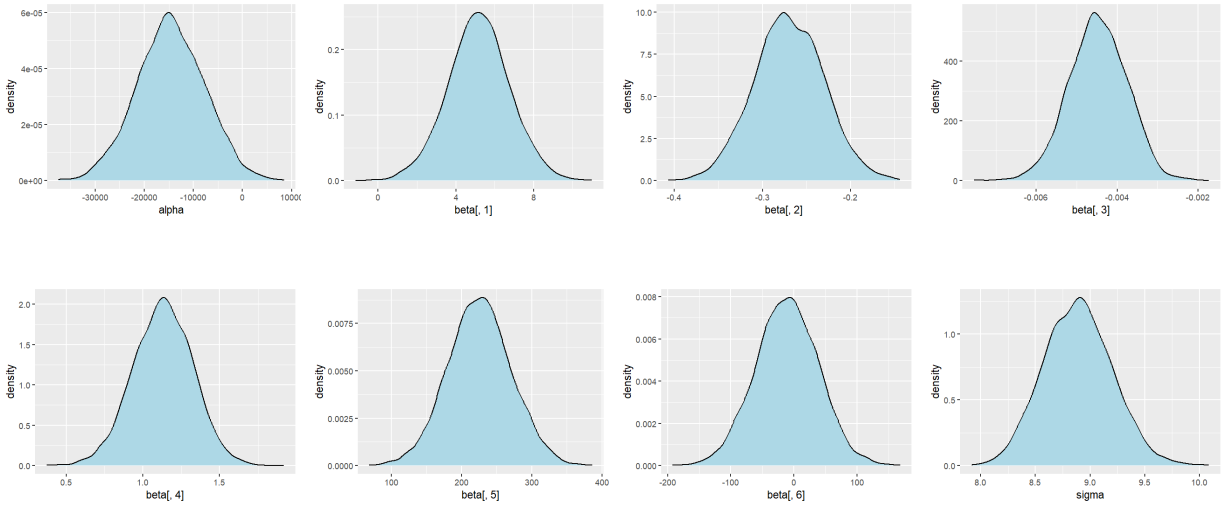


Figure 3: Approximated Posteriors Densities

Parameter	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std
α	-37464	-19259	-14690	-14611	-9929	8504	6877.93
β_1	-1.148	4.140	5.186	5.174	6.204	10.987	1.59
β_2	-0.4072	-0.2965	-0.2708	-0.2701	-0.2433	-0.1438	0.04
β_3	-0.007539	-0.004956	-0.004484	-0.004477	-0.003994	-0.001732	0.00
β_4	0.3740	0.9974	1.1314	1.1292	1.2635	1.9177	0.19
β_5	68.51	195.66	225.19	225.41	254.91	386.16	44.92
β_6	-192.18	-44.77	-10.93	-11.44	22.39	168.26	49.52
σ	7.919	8.669	8.885	8.887	9.096	10.084	0.31

Table 1: Numerical summary of the approximated posteriors densities

As we can see in Table 2, the parameters α, β_5, and β_6 exhibit higher standard deviations compared to the other parameters. This higher uncertainty is reflected in the wider credible intervals we can provide for these parameters (Figure 4).

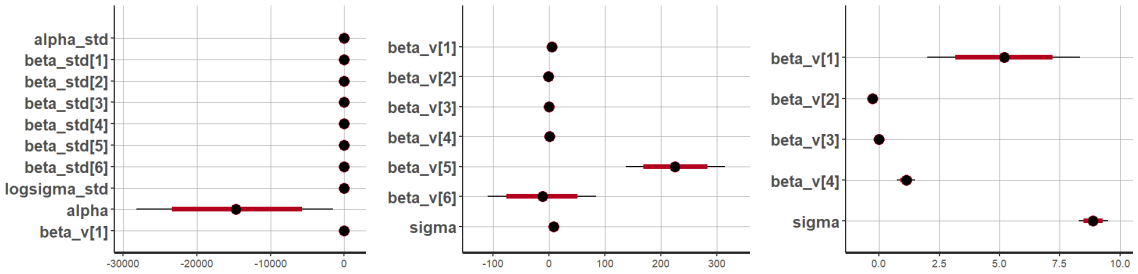


Figure 4: Credible Intervals for the parameters of the model

Before giving an interpretation of the posterior distributions for the model parameters, we want to assess the model's adequacy in fitting the data and determine if it can be simplified by reducing the number of predictors.

1.5 Post Predictive Diagnostics

To assess the adequacy of the model in fitting the data, we examine its ability to accurately capture the summary statistics of the original data.

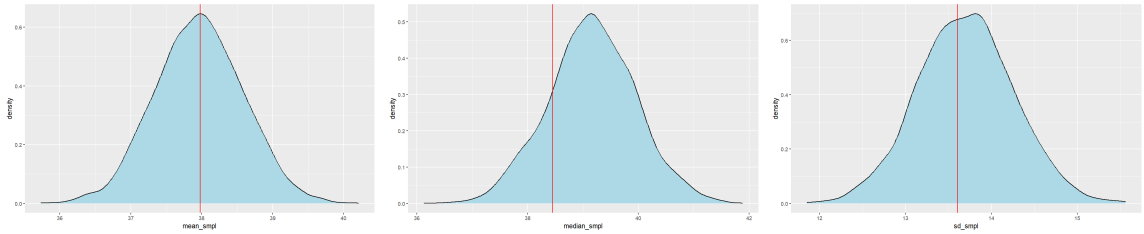


Figure 5: Generated vs Original mean, median and std deviation.

As we can see from Figure 5, the model is able to well capture the mean, median and standard deviation of the response from the original data. However, looking at the minimum and maximum (Figure 6), we don't have the same results:

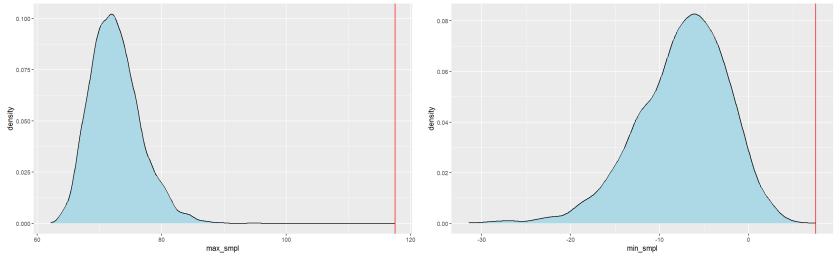


Figure 6: Generated vs Original max and min

The model appears to have difficulty capturing both the minimum and maximum values of the response variable. One reason for this discrepancy is that the model assumes a normal distribution for the prices, which assumes positive probabilities for negative prices, whereas the original data does not follow this assumption. Regarding the maximum value, if we examine the histogram of the original data in Figure 7, we can observe that there is an outlier observation that exceeds the majority of the data points. However, when focusing on the core of the data, the maximum value appears to be around 80 (as our model predicts).

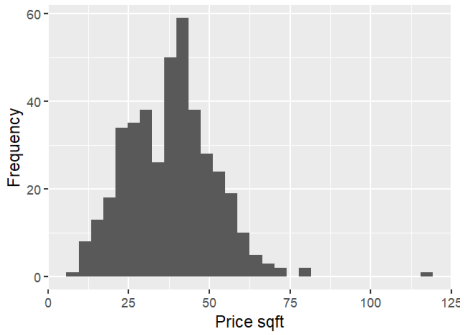


Figure 7: Histogram of the prices in the original data

Finally we check the residuals (Figure 8). They seem to be centered in zero and with constant variance.

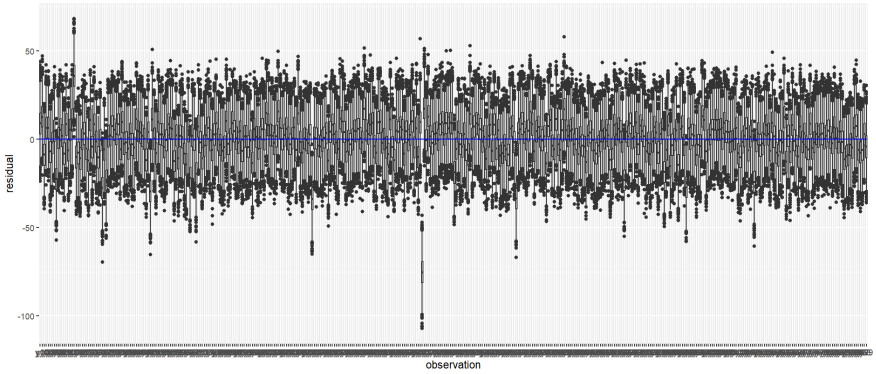


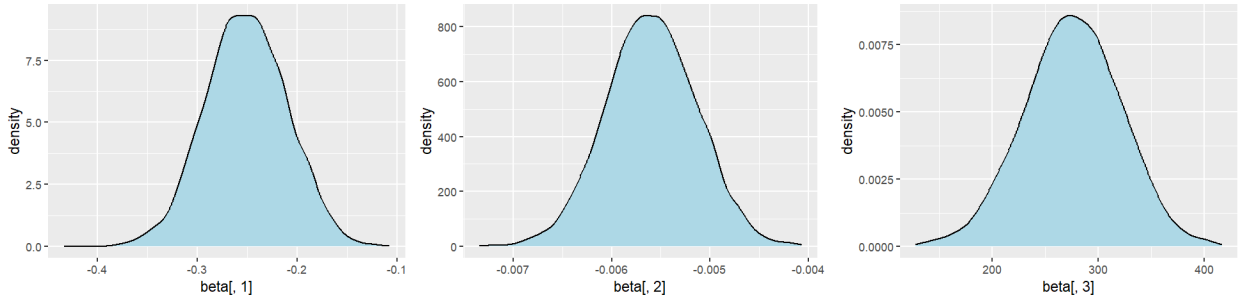
Figure 8: Residuals

1.6 Model selection

In Section 1.1, we noted that the linear correlation between $price_sq$ and $date$, as well as age, is very low. Based on this observation, we attempted to fit various models by excluding different combinations of these two variables. Specifically, we considered models without both age and $date$, only $date$, only age, and compared them to the full model described so far. By utilizing the expected log-pointwise posterior predictive density (ELPD), we determined that the full model and the simplified model without the $date$ variable are almost equivalent, in fact we have:

$$| ELPD_{diff} | < 2 \cdot se_{diff} \text{ , in particular } | -4.8 | < 2 \cdot 2.8$$

This means that $\beta_1 \approx 0$, so we will simplify the model removing $date$ from the regressors.
Since we can notice from Figure 1b that we have high correlation between predictors, we tried to further reduce the model. Using the same technique explained above, the final model we found has $\beta_1, \beta_4, \beta_6 \approx 0$, namely we removed date, number of convenient stores and longitude from the set of predictors. We report in the following the simulated posterior distribution of the coefficients $\beta_2, \beta_3, \beta_5$ in order to interpret them.



(a) Posterior for β_2 (age) (b) Posterior for β_3 (metro distance) (c) Posterior for β_5 (latitude)

Figure 9: Posterior distributions for the simplified model parameters.

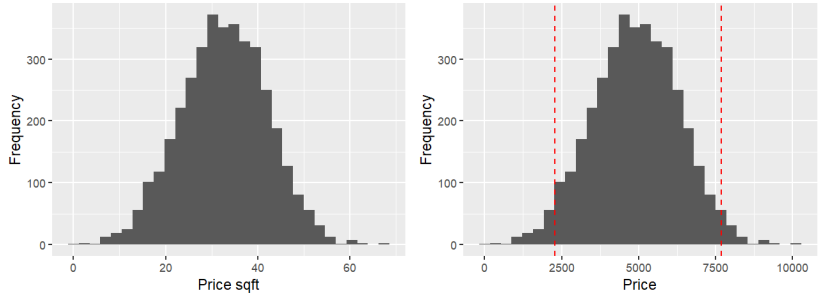
As could be expected, an increase of the age of the house lead to a decrease in the house price per square feet in mean. The same holds for the distance from the nearest metro. On the other hand, an increase in latitude is associated with a rise in house prices (per square feet), in mean.

1.7 Prediction

As last task we compute a 95% highest probability density interval for the price of a house with an area of 150 square feet, sold in the middle of 2013, 17 years of age, 1100 feet away from the nearest metro station, with 4 convenience stores in it's vicinity, situated in latitude 24.95 and longitude 121.50. In order to do this, we generated a new random sample of the response using the simplified model developed in the previous section. In particular:

$$Y_i \sim N(\alpha + \beta_2 X_2 + \beta_3 X_3 + \beta_5 X_5, \sigma^2), i = 1, ..., n_{new}$$

using the posteriors densities for the parameters $\beta_2, \beta_3, \beta_5, \sigma$ and the values reported above for X_2, X_3, X_5. After that, we computed the associated prices multiplying by the square feets of the new house. We obtained $[2281.576, 7698.660]$ as HDI for the price of the new house.



(a) Price per square feet (b) Price for house of area 150 sqft

Figure 10: Prediction

2 Exercise 2

2.1 Fitting the model

The Bayesian model we designed to fit the data is the following:

- Model for *height*: $Y_{it} \sim N(\mu_{it}, \sigma_y^2)$ with $\mu_{it} = \beta_{i0} + \beta_{i1} t$, i index of soil (group), t index of the week
- Priors: $\begin{bmatrix} \beta_{i0} \\ \beta_{i1} \end{bmatrix} \sim N\left(\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \text{DRD} \right)$ with $D = diag(\sigma_1, \sigma_2), \sigma_y^2 \sim Exp(r_y)$
- Hyper-Priors: $\beta_0 \sim N(m_0, v_1^2), \beta_1 \sim N(m_1, v_1^2), R \sim LKJ(l), \sigma_j \sim Exp(r_0)$ with $j = 1, 2$

We chose the following values for the parameters:

- $m_0 = m_1 = 0$ and $v_0 = v_1 = 100$
- $l = 1.00$

- $r_0 = 10$
- $r_y = 10$

We are not assuming independence between β_{0i} and β_{1i} in order to create a more general model.

2.2 Model diagnostics

We fitted the model thorugh a MCMC simulation. In order to check the convergence of the (four) generated chains, we can look at the traceplots reported in figure 11.

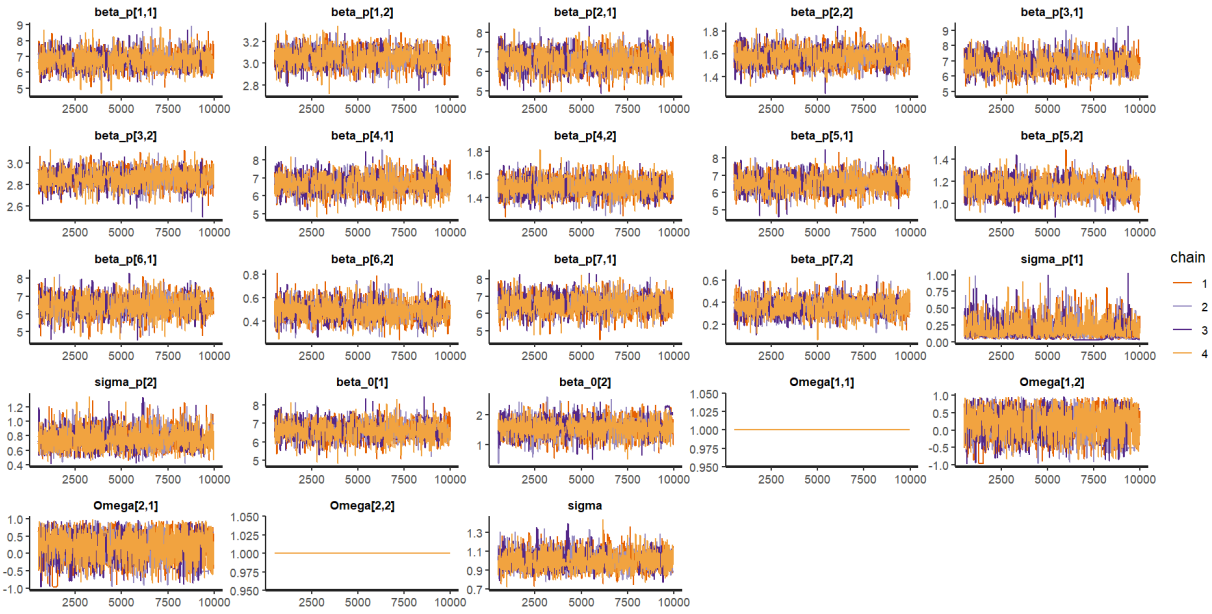
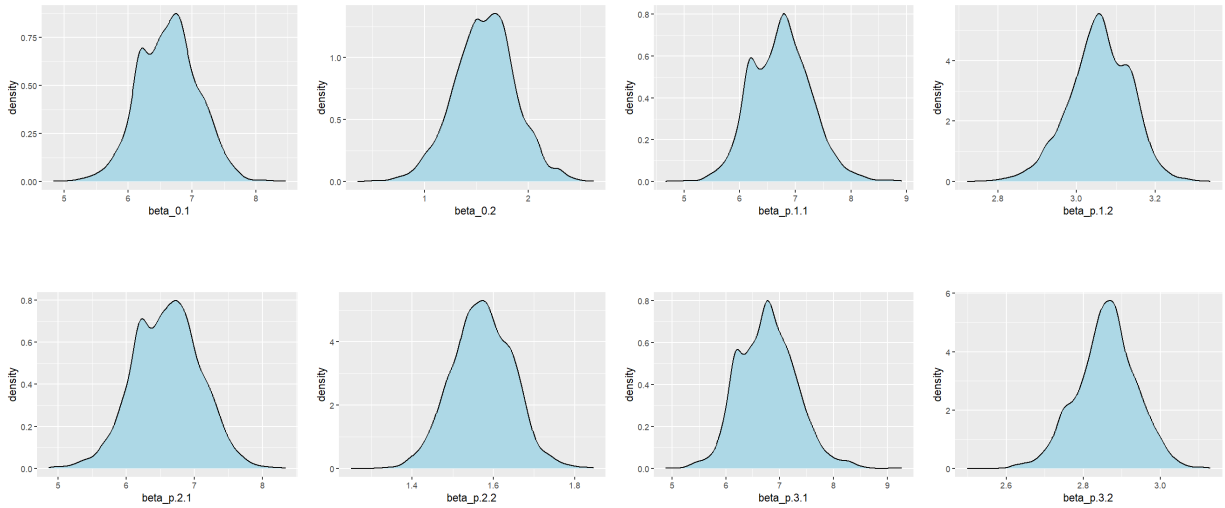


Figure 11: Trace plots

As we can see, the chain are well mixed and they do not show seasonal, increasing or decreasing behaviours. Moreover, we checked that the effective sample size ratios are all greater than 0.1 and the $\hat{R}$ are all lower than 1.05. This means that we can trust the simulation and the chains have converged.

2.3 Model posteriors

In the following we summarise graphically the posteriors of the parameters obtained through the simulation.



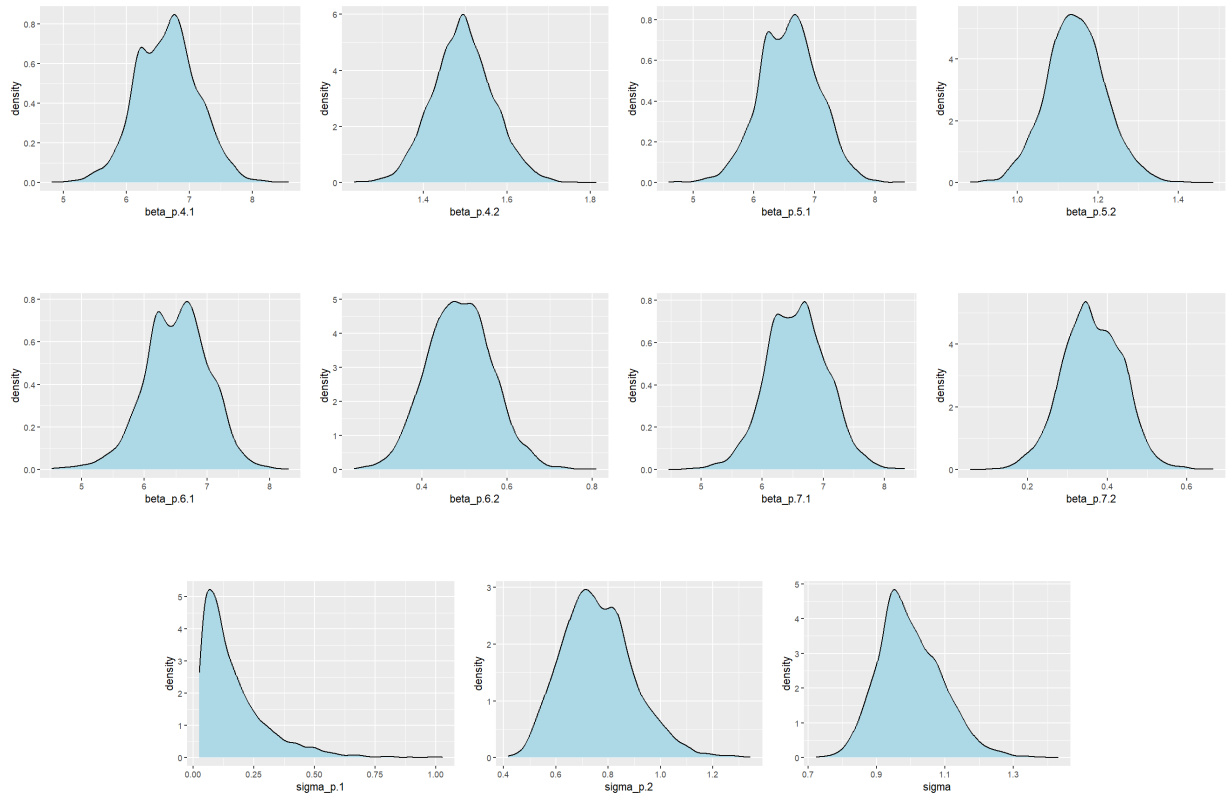


Figure 12: Posteriors

Parameter	Code Parameter	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std
$\beta_{1,1}$	beta_p.1.1	4.668	6.346	6.748	6.739	7.086	8.914	0.5309
$\beta_{1,2}$	beta_p.1.2	2.721	3.009	3.060	3.058	3.116	3.341	0.0788
$\beta_{2,1}$	beta_p.2.1	4.863	6.243	6.601	6.593	6.922	8.338	0.4854
$\beta_{2,2}$	beta_p.2.2	1.250	1.523	1.572	1.572	1.625	1.846	0.0743
$\beta_{3,1}$	beta_p.3.1	4.882	6.384	6.762	6.772	7.120	9.269	0.5343
$\beta_{3,2}$	beta_p.3.2	2.499	2.813	2.862	2.860	2.909	3.130	0.0779
$\beta_{4,1}$	beta_p.4.1	4.816	6.292	6.661	6.650	6.965	8.579	0.4855
$\beta_{4,2}$	beta_p.4.2	1.233	1.447	1.494	1.494	1.541	1.815	0.0728
$\beta_{5,1}$	beta_p.5.1	4.597	6.236	6.590	6.581	6.899	8.485	0.4839
$\beta_{5,2}$	beta_p.5.2	0.8826	1.0954	1.1433	1.1451	1.1928	1.4864	0.0731
$\beta_{6,1}$	beta_p.6.1	4.528	6.179	6.537	6.519	6.867	8.299	0.5101
$\beta_{6,2}$	beta_p.6.2	0.2403	0.4367	0.4875	0.4880	0.5389	0.8094	0.0761
$\beta_{7,1}$	beta_p.7.1	4.467	6.218	6.564	6.565	6.905	8.336	0.4948
$\beta_{7,2}$	beta_p.7.2	0.0554	0.3114	0.3597	0.3624	0.4154	0.6669	0.0747
β_0	beta_0.1	4.828	6.292	6.642	6.631	6.932	8.469	0.4626
β_1	beta_0.2	0.3592	1.3891	1.5886	1.5807	1.7691	2.6287	0.2955
σ_1	sigma_p.1	0.02396	0.07082	0.12281	0.16447	0.21540	1.02613	0.1331
σ_2	sigma_p.2	0.4170	0.6652	0.7503	0.7613	0.8449	1.3442	0.1358
σ_y	sigma	0.7236	0.9353	0.9869	0.9987	1.0604	1.4315	0.0926

Table 2: Numerical summary of the approximated posteriors densities

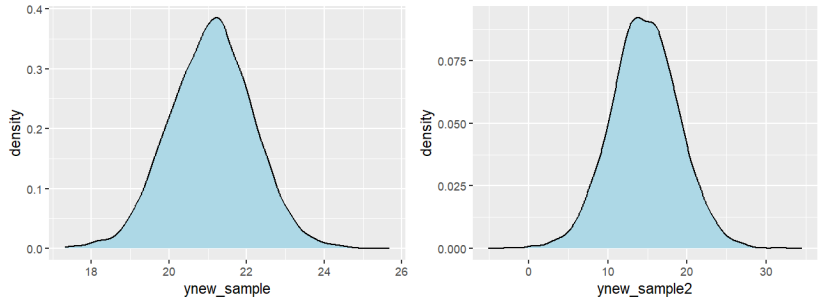
As we can see from the posterior distributions, we have that all the β_{i1} are distributed around 6.5. This means that at time $t = 0$, there is not a significant difference among the various types of soil. This observation is reasonable since, when initially planting a plant, the soil type should not significantly impact the plant's height. On the other hand, if we look at the β_{i2}, they are all positive but with significant differences. For example, comparing the mean values of the slope for two different soil types:

- For $soil_1$ we have the mean of $\beta_{1,2}$ is 3.058.
- For $soil_7$ we have the mean of $\beta_{7,2}$ is 0.3624.

This means that we can expect an higher growth for a plant in $soil_1$ with respect to $soil_7$. This result could be interpreted as a difference in the fertility levels of the soils. Looking at the hyper-parameters, which describe the general behaviour of the constant and the slope for each model of plant growth on site, we can expect to have the mean for the constant around 6.5 and the slope around 1.5.

2.4 Predictions

In this last section we are going to present two different predictions using the model developed so far. In particular we are going to predict the height for two plants at the 5^{th} week. The first one was planted in $soil_3$ and the second one in an external site with an unknown type of soil. Using the (simulated) posteriors of $\beta_{3,1}$, $\beta_{3,2}$ and σ_y we can easily generate the prediction for the first plant. For the second one, using the posteriors of $\beta_0, \beta_1, R, \sigma_1$ and σ_2, we generate a distribution for $\beta_{new,1}, \beta_{new,2}$. Then, we use them to create the distribution for the prediction for the plant in $soil_{new}$.



(a) Plant in $soil_3$ (b) Plant in $soil_{new}$

Figure 13: Prediction of height for plants at the 5^{th} week.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max	Std
$Height_{soil_3}$	17.32	20.38	21.11	21.09	21.80	25.68	1.07
$Height_{soil_{new}}$	-5.06	11.88	14.59	14.63	17.46	34.48	4.31

Table 3: Summary of the Predictions of heights of plants in different soils at 5^{th} week.

As we can see from the results summarized in Table 3, the prediction of the height of the plant in the new soil has much more uncertainty (std equal to 4.31, compared to 1.07 of $soil_3$) due to the fact that we don't have information about that soil and we have to generate the $\beta_{new,1}, \beta_{new,2}$ using the posteriors of $\beta_0, \beta_1, R, \sigma_1$ and σ_2.