Bayesian analysis project: Vinho Verde

This assessment a logistic regression on the Vinho Verde wine dataset, which describes the quality of wine based off various physiochemical tests.

1. Read the dataset into R. Check if there are missing values (NA) and, in case there are, remove them.

The Dataset is read into R using read.csv. It seemed as though there were not any NA data, which was found using the sum(is.na())) function. In anycase, the line na.omit was used to remove any potential ones.

2. We want to implement a logistic regression, therefore we want a response variable which assume values either 0 or 1. Suppose we consider "good" a wine with quality above 6.5 (included).

This is easily done using an ifelse function within a transform function, where if the quality is greater than of equal to 6.5 it is given the value of 1, and if it is less than 6.5 it is given the value of 0. This is stored under column good_wine.

3. Run a frequentist analysis on the logistic model, using the glm() function. What are the significant coefficients?

Using all the provided variables, the glm is written as shown in figure 1. However it should be noted that quality was not included as that would be predicting the good_wine using the answer.

```
glm(formula = good_wine ~ fixed.acidity + volatile.acidity +
    citric.acid + residual.sugar + chlorides + free.sulfur.dioxide +
    total.sulfur.dioxide + density + pH + sulphates + alcohol,
    family = binomial(link = "logit"), data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9878  -0.4351  -0.2207  -0.1222   2.9869

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           2.428e+02  1.081e+02   2.247 0.024660 *
fixed.acidity         2.750e-01  1.253e-01   2.195 0.028183 *
volatile.acidity     -2.581e+00  7.843e-01  -3.291 0.000999 ***
citric.acid           5.678e-01  8.385e-01   0.677 0.498313
residual.sugar        2.395e-01  7.373e-02   3.248 0.001163 **
chlorides            -8.816e+00  3.365e+00  -2.620 0.008788 **
free.sulfur.dioxide   1.082e-02  1.223e-02   0.884 0.376469
total.sulfur.dioxide -1.653e-02  4.894e-03  -3.378 0.000731 ***
density              -2.578e+02  1.104e+02  -2.335 0.019536 *
pH                    2.242e-01  9.984e-01   0.225 0.822327
sulphates             3.750e+00  5.416e-01   6.924 4.39e-12 ***
alcohol               7.533e-01  1.316e-01   5.724 1.04e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1269.92  on 1598  degrees of freedom
Residual deviance:  870.86  on 1587  degrees of freedom
AIC: 894.86

Number of Fisher Scoring iterations: 6
```

Figure 1. Summary of GLM Model

Using the p-values we can determine which of the coefficients are unreliable. The significant value coefficients at a 95% level are:

- fixed.acidity
- volatile.acidity
- residual.sugar
- chlorides
- total.sulfur.dioxide
- density
- sulphates
- alcohol

Using a smaller significant factor, this list can be further refined down.

4. Estimate the probabilities of having a "success": fix each covariate at its mean level, and compute the probabilities for a wine to score "good" total.sulfur.dioxide, and plot the results.

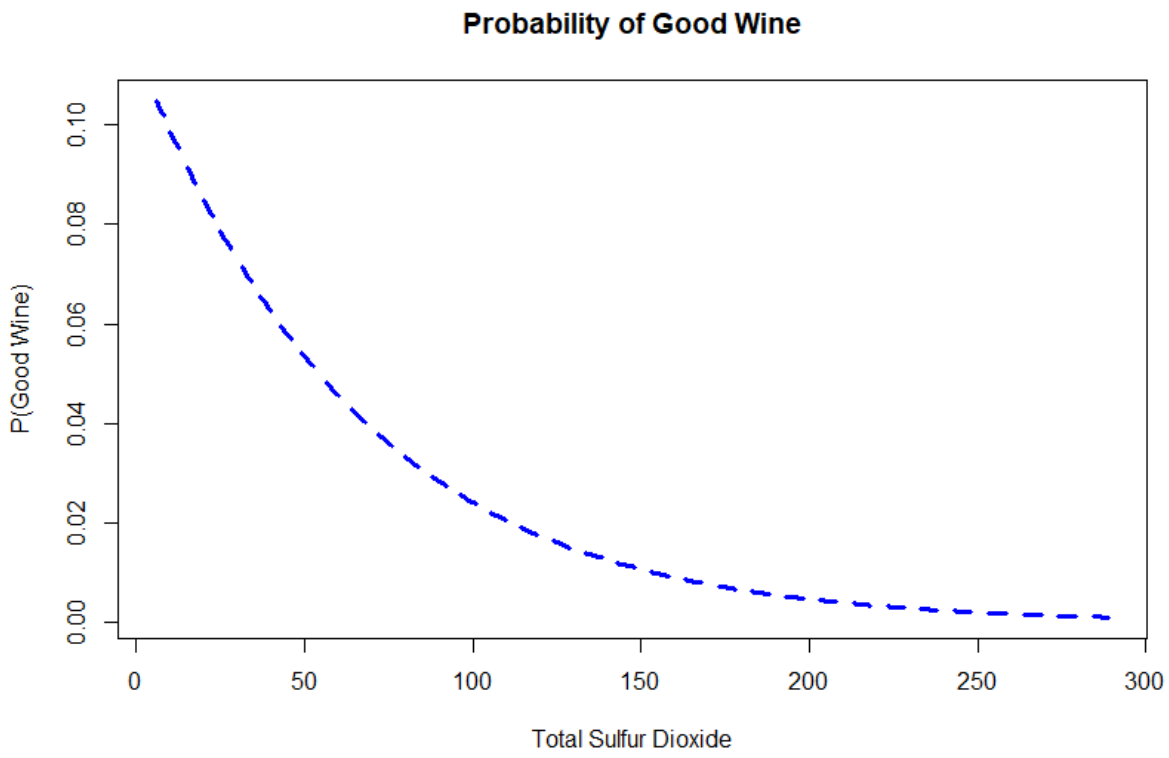


Figure 2 Proability of Good Wine from Total Sulfur Dioxide

Fixing all the covariates with the exception of the Total Sulfur Dioxide to the mean, and calculating the quality for a range of values from minimum to maxium. Using these values the probability is calculated from the following

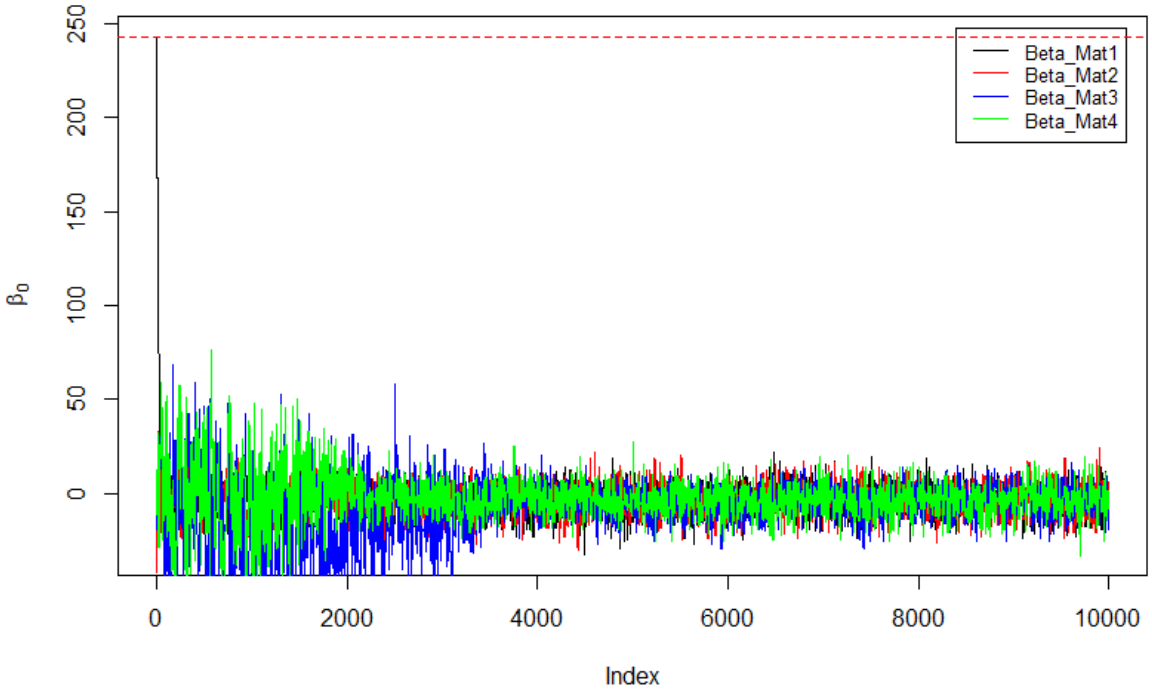
$$P = \frac{\exp(Log\ Odds)}{1 + \exp(Log\ Odds)}$$

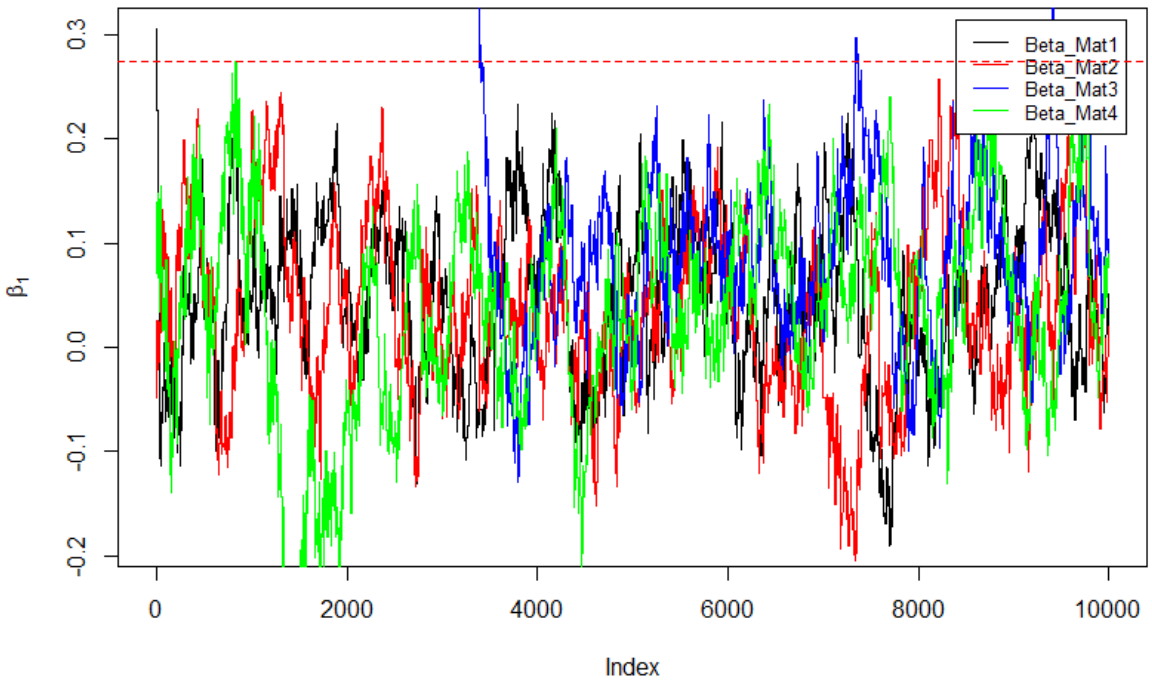
Which derives to the following,

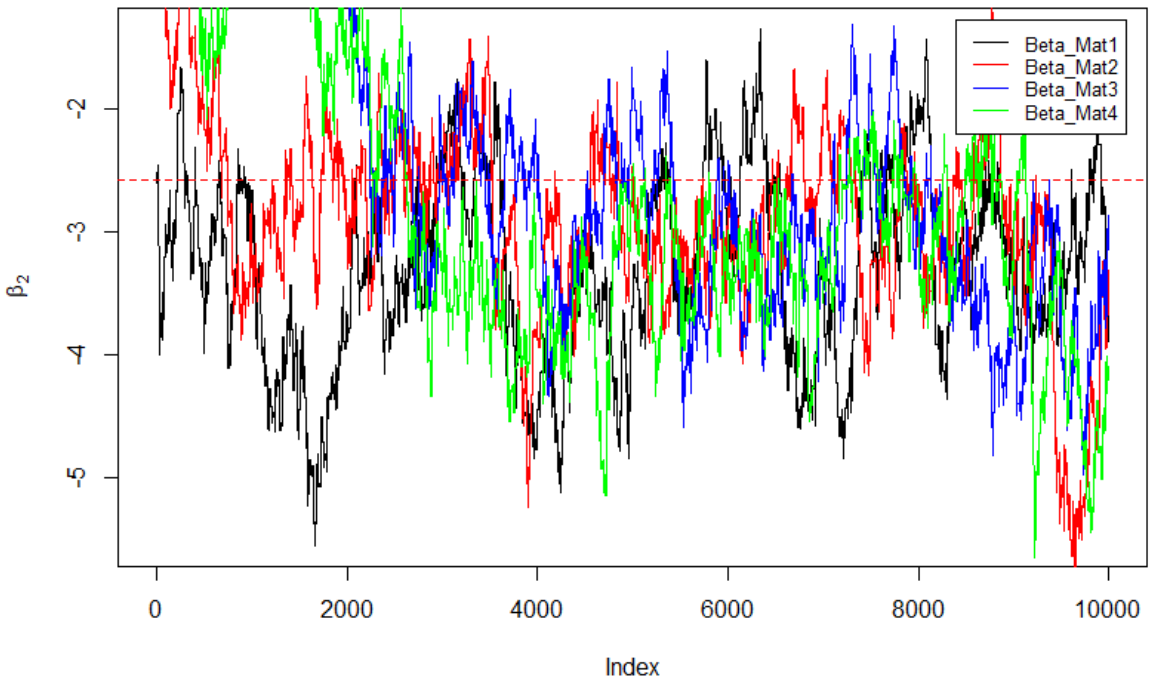
$$P = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots \beta_{11} x_{11})}{1 + \exp(L\beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots \beta_{11} x_{11})}$$

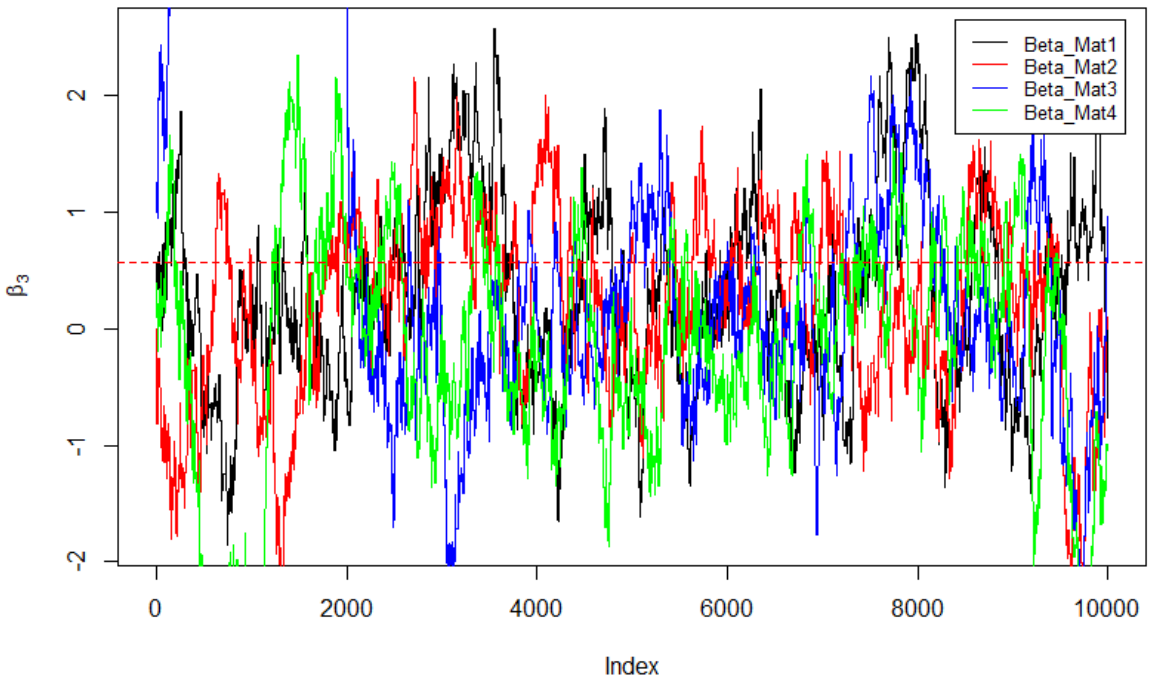
5. Perform a Bayesian analysis of the logistic model for the dataset, i.e. approximate the posterior distributions of the regression coefficients.

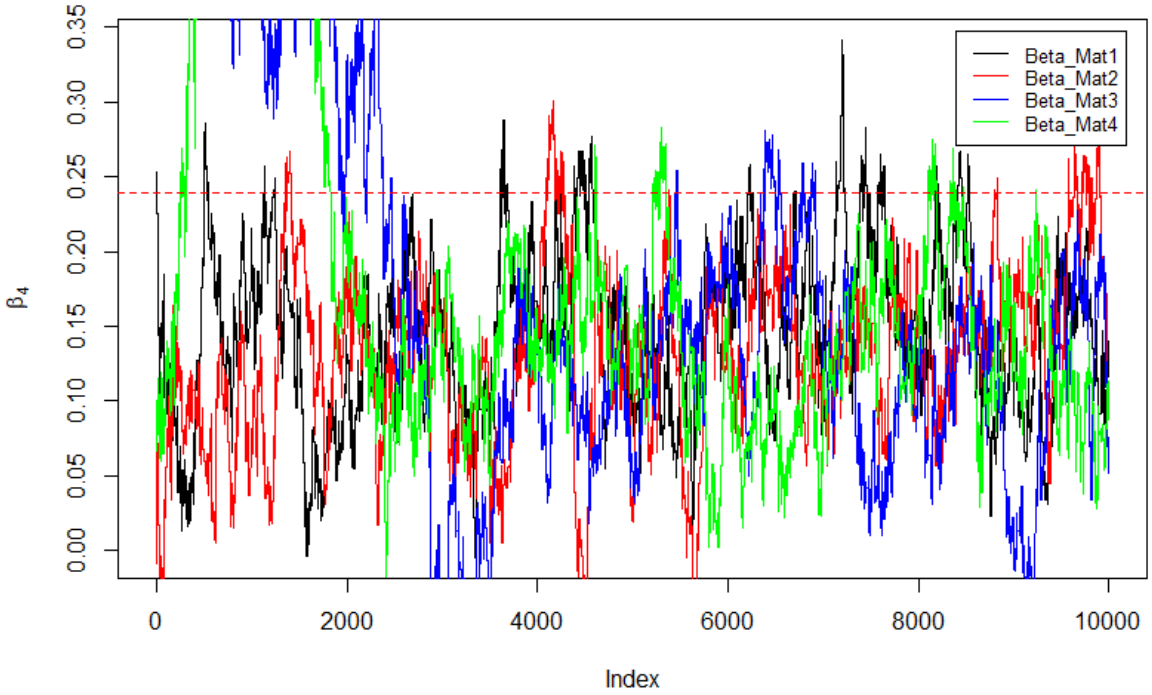
It can be seen in the following graphs that the coefficients do converge, however the starting iteration values, does affect how quickly the convergence happens. For example β_0, when the MLE was used, it oscillated much more prior to stabilising.

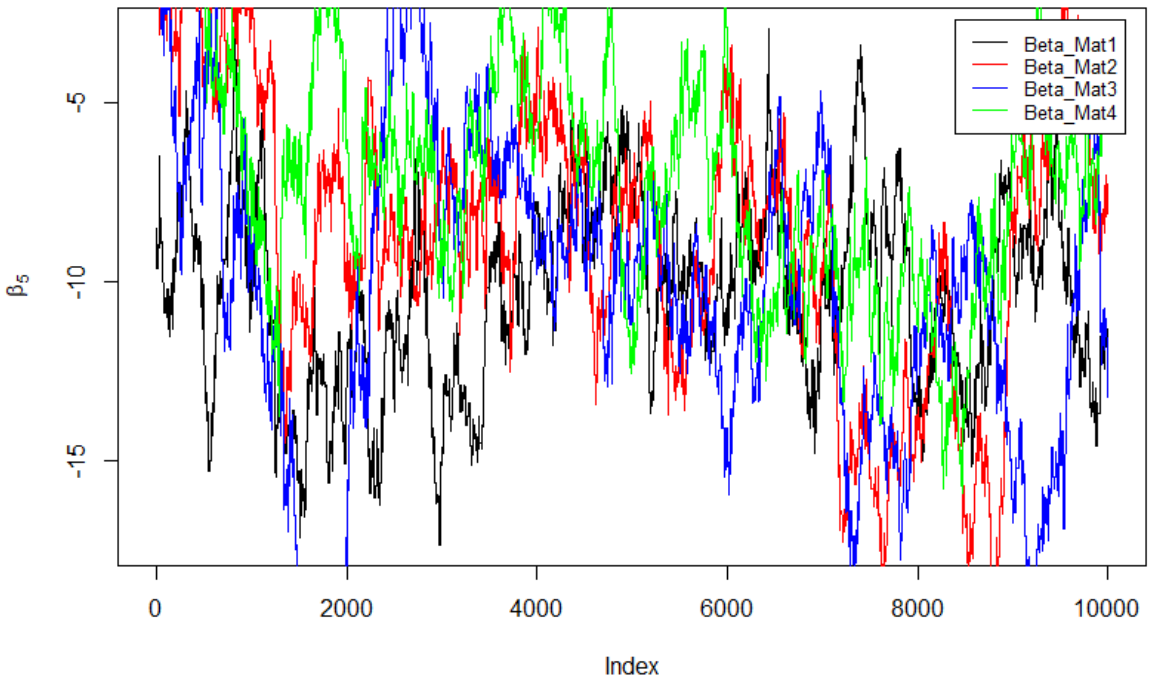


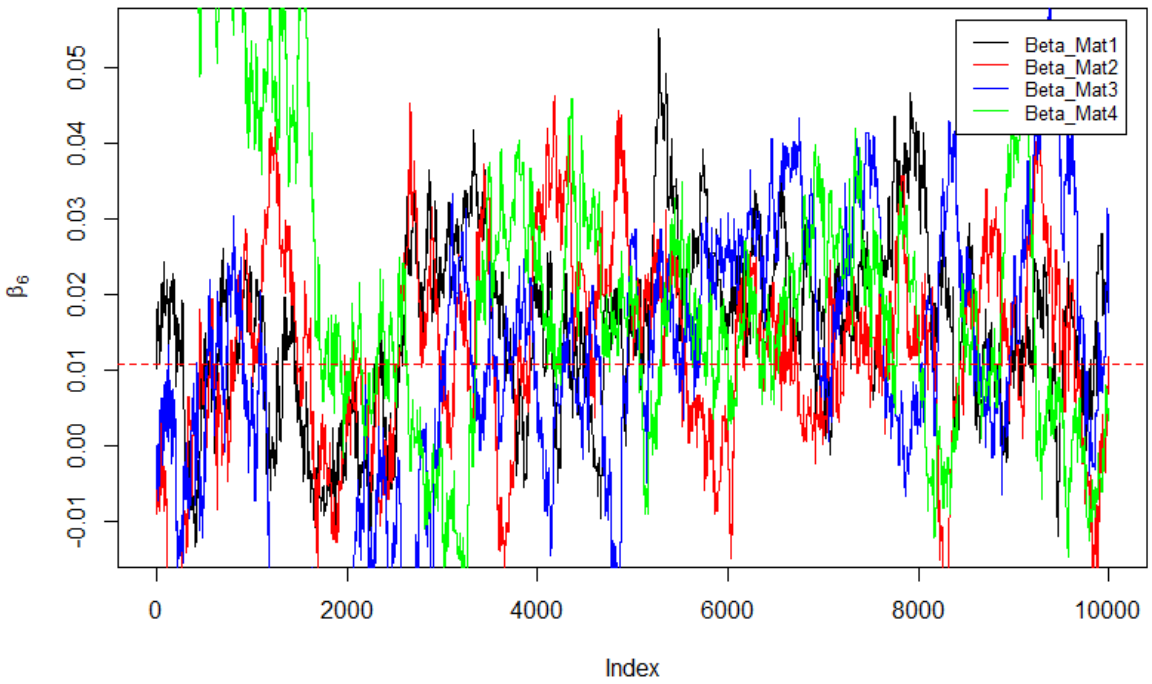

Beta_Mat1
Beta_Mat2
Beta_Mat3
Beta_Mat4
β1
0.3
0.2
0.1
0.0
-0.1
-0.2
0
2000
4000
6000
8000
10000
Index

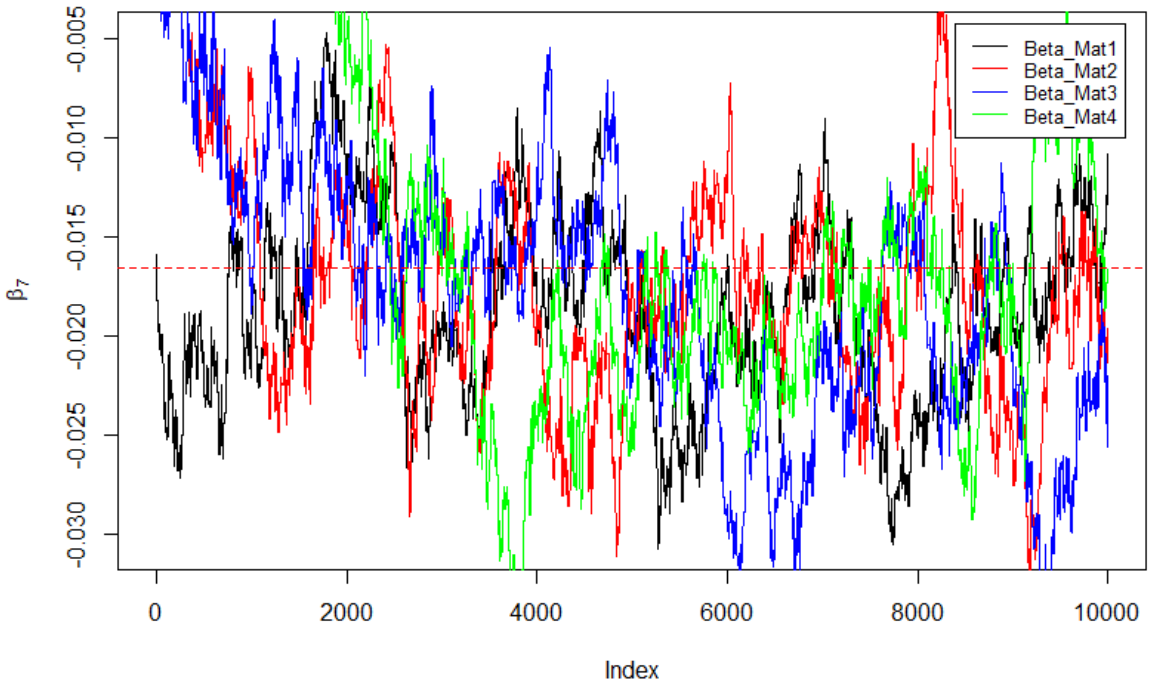

Beta_Mat1
Beta_Mat2
Beta_Mat3
Beta_Mat4
β2
-2
-3
-4
-5
0
2000
4000
6000
8000
10000
Index

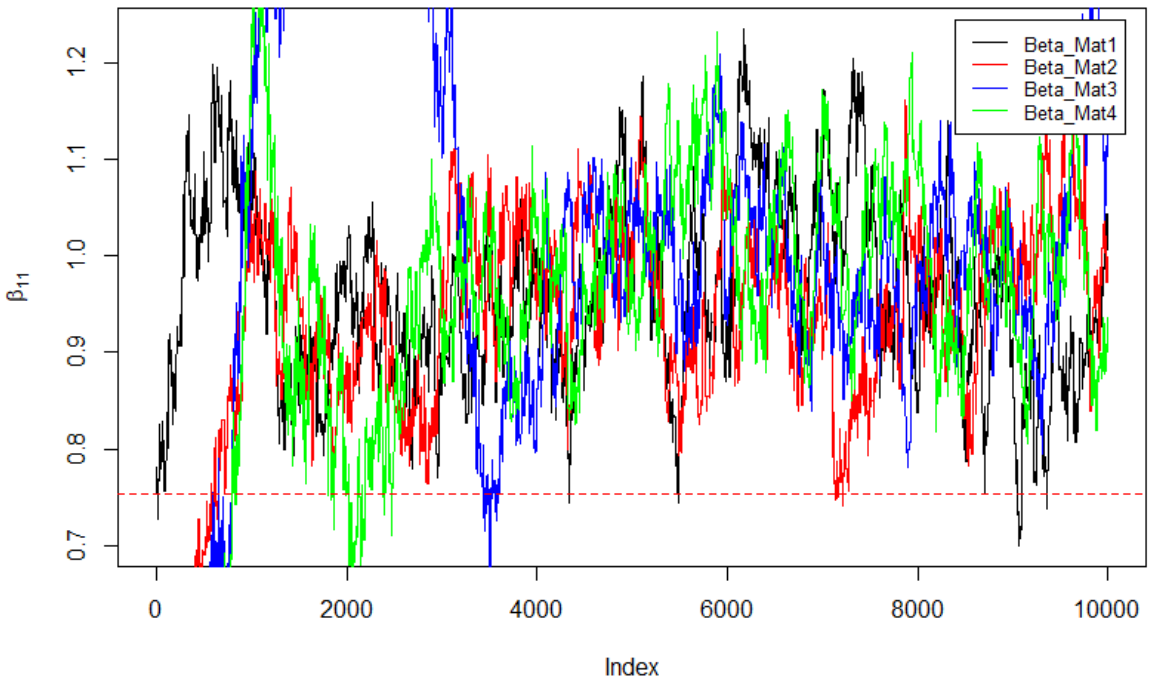

Beta_Mat1
Beta_Mat2
Beta_Mat3
Beta_Mat4
β3
2
1
0
-1
-2
0
2000
4000
6000
8000
10000
Index

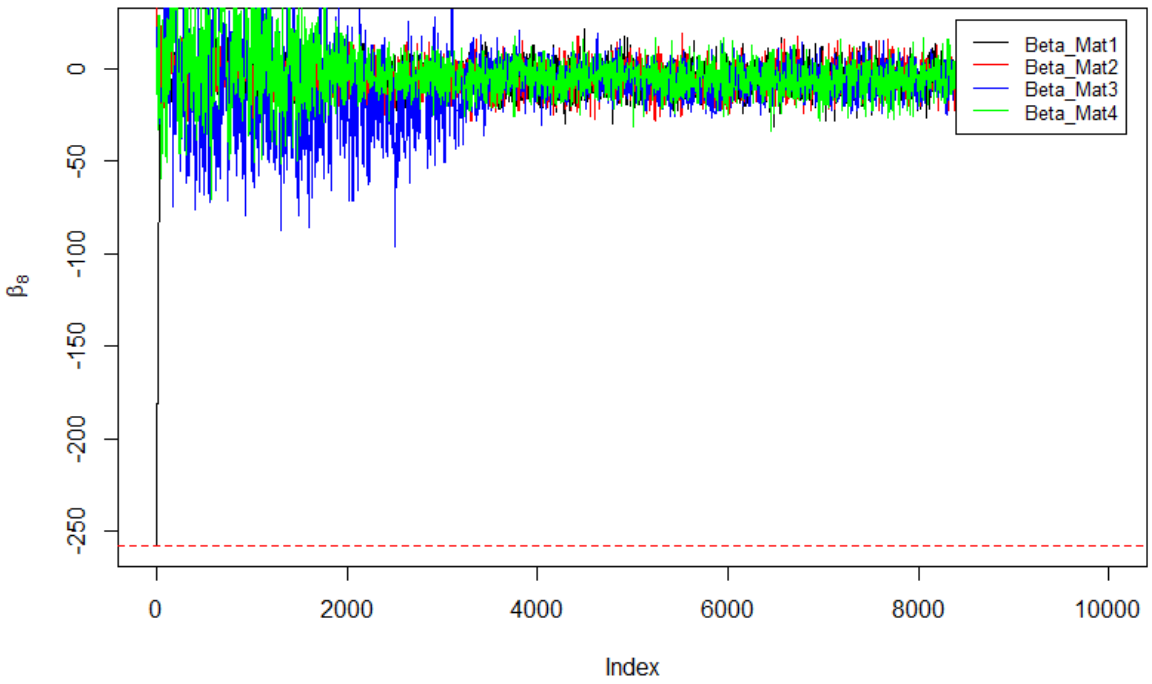

Beta_Mat1
Beta_Mat2
Beta_Mat3
Beta_Mat4
β4
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
2000
4000
6000
8000
10000
Index

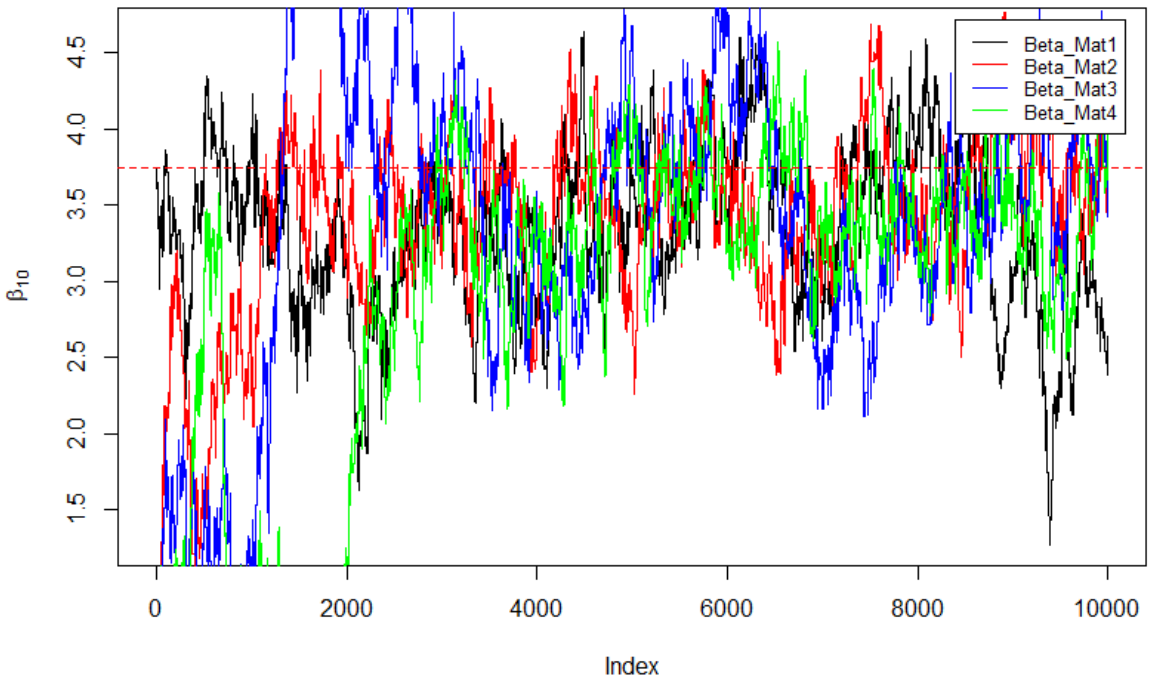

Beta_Mat1
Beta_Mat2
Beta_Mat3
Beta_Mat4
β5
-5
-10
-15
0
2000
4000
6000
8000
10000
Index

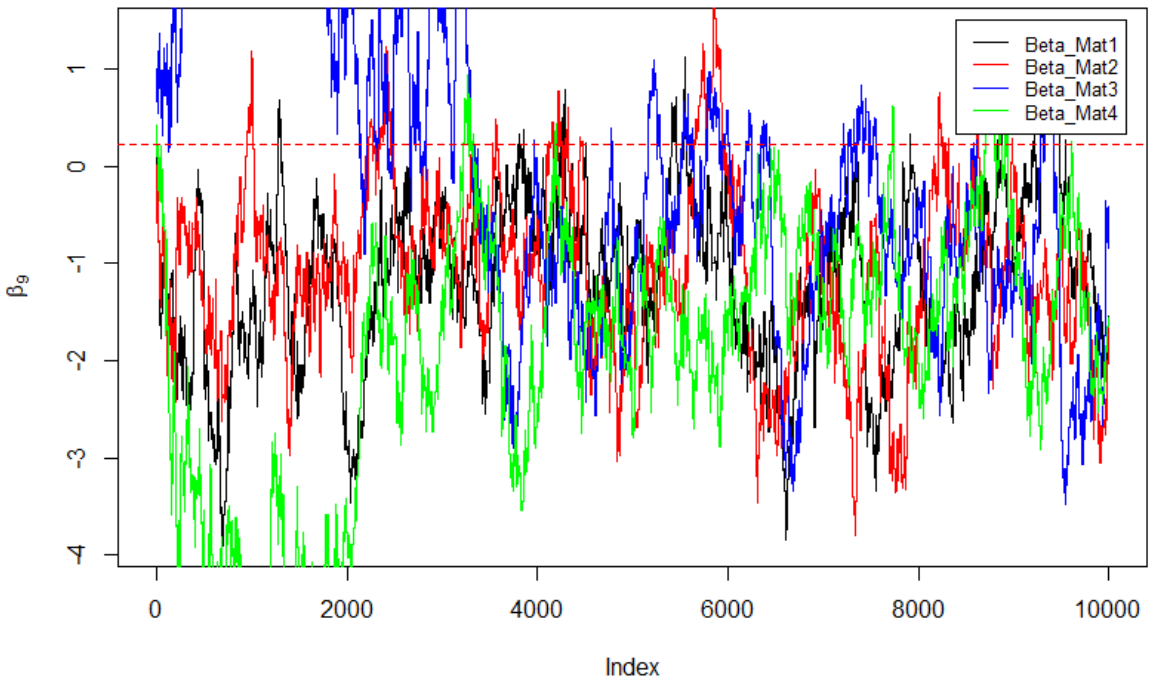

Beta_Mat1
Beta_Mat2
Beta_Mat3
Beta_Mat4
β6
0.05
0.04
0.03
0.02
0.01
0.00
-0.01
0
2000
4000
6000
8000
10000
Index


Beta_Mat1
Beta_Mat2
Beta_Mat3
Beta_Mat4
β7
-0.005
-0.010
-0.015
-0.020
-0.025
-0.030
0
2000
4000
6000
8000
10000
Index


Beta_Mat1
Beta_Mat2
Beta_Mat3
Beta_Mat4
β11
1.2
1.1
1.0
0.9
0.8
0.7
0
2000
4000
6000
8000
10000
Index


Beta_Mat1
Beta_Mat2
Beta_Mat3
Beta_Mat4
β8
0
-50
-100
-150
-200
-250
0
2000
4000
6000
8000
10000
Index


Beta_Mat1
Beta_Mat2
Beta_Mat3
Beta_Mat4
β10
4.5
4.0
3.5
3.0
2.5
2.0
1.5
0
2000
4000
6000
8000
10000
Index


Beta_Mat1
Beta_Mat2
Beta_Mat3
Beta_Mat4
β9
1
0
-1
-2
-3
-4
0
2000
4000
6000
8000
10000
Index