Team Mean, Median, Mode
Rei Leenah Balachandran s3112637
Shubhankar Sanjay Jahagirdar s3793593
Tanmay Madan Shendkar s3735580

Prediction of Legal High Consumption with Logistic Regression

1. Executive Summary

Drug consumption is not an uncommon activity, and there are factors that aid in the prediction of drug consumption. This study used a dataset containing 1885 participants, with their age, gender and 7 personality scores (neuroticism, extraversion, openness to experience, assertiveness, conscientiousness, impulsiveness and sensation seeking) and their status on legal high consumption specifically. JAGS was used to run a robust logistic regression model to obtain the posterior distributions of the model parameters.

It was observed that there were 3 insignificant parameters in the initial model, which was removed from the next model. The second model showed to have improved performance in the prediction of legal high consumption.

2. Introduction

Various studies have put forth an analysis which describes personal demographics as a predominant factor in drug consumptions. These include the factors such as Age, Gender and various personality and circumstantial conditions which diverts an individual towards drug consumption, eventually leading to substance abuse. In the Ghana Medical Journal, a research states that 86.8% of people under the age group 15 -25 are into drug abuse. It also states that the Males, in general, are more towards substance abuse rather than Females. This research information motivates us to concentrate on developing relationships between personality traits and demographic features and the consumption of Legal Highs (Lamptey, 2005).

Several organizations work together with researchers and doctors to understand the psychoactive demographic features which might be provided as a deterministic factor towards drug consumption leading to substance abuse. Based on the research provided by the American Addiction Centres, it was stated that women more rapidly develop a prescription painkiller than men (Thomas, 2020).

Numerous descriptions of drug abuse and substance abuse are used in public health, medical, and criminal justice contexts. In addition to the possible physical, social and psychological harm, the use of some drugs may also lead to criminal penalties. Drugs associated with this term include crack, heroin, amphetamine, ketamine, LSD, meth, legal highs, and many more. To study the pattern and gain relevant insights we consider Legal Highs as the outcome feature of our model. Legal High was first publicly addressed in a BBC news article which describes these drugs as psychoactive substances. The effects of these drugs include paranoia, seizures, coma, and even death. However, medical professionals have discovered patterns in the personality traits of individuals which usually provoke them towards substance abuse. Studying these patterns by analysing respondents' information can produce effective insights and hence detect target susceptible audiences.

In this project, we explore the drug consumption data set which is publicly available on UCI Machine Learning repository. This data set contains records of 1885 respondents. For each respondent, 12 attributes are represented in the data set which includes personality and demographic traits. The data set outlines class variables each representing an illegal drug. To

achieve our aim and produce effective estimations, we reduce the features to the most relevant personality and demographic representation of every respondent and study insights, and produce posterior distributions to gain knowledge about the consumption of Legal Highs.

To achieve our aims, in this project we follow the Bayesian approach and use logistic regression modelling with the assistance of JAGS (Just Another Gibbs Sampler) to run MCMC (Markov Chains Monte Carlo) and produce posterior distributions with point and interval estimates of each individual based on their demographic and personality traits. We further extend our model capabilities and test our model on unseen data that is testing data to ensure productivity and provide assurance.

3. Data and Exploration

This project models data which is publicly available on UCI Machine Learning Repository (Data Link). The data set contains records for 1885 respondents. Every individual's personality and demographic traits are recorded. These include Age, Gender, Education, Country, Ethnicity, N-score (Neuroticism), E-score (Extraversion), O-score (Openness to experience), A-score (Agreeableness), C-score (Conscientiousness), Impulsive (impulsiveness), SS-score (Sensation). Age, Gender, Education, Country, and Ethnicity are categorical features, where Age is ordinal, and others are nominal variables. The scores are numerical real values recorded for every individual. The data set contains 18 classification problems each representing seven classes representing the drug consumption duration, including 'Never Used', 'Used over a decade ago', 'Used in Last Decade', 'Used in Last Year', 'Used in Last Month', 'Used in Last Week', and 'Used in Last Day'.

Since the data set provides a wide range of information for every respondent, we did some initial exploration and decided to model this project for the consumption of Legal Highs. We then dropped all other class variables. In this project, however, we aim to study whether an individual is susceptible to consumption of Legal Highs based only on the recorded personality, Age, and Gender. Hence, we changed the class variable to represent a binary level of information that describes whether an individual consumes drugs, therefore we convert to the class level of 'Never Used' and 'Used over a decade ago' to level 0 and the rest to level 1. Since MCMC is a complex method of modelling and it takes long durations to deliver posterior distributions we limit our independent features to the most specific ones to model the data. These features include age, gender, N-score, E-score, O-score, sensation seeking, and impulsiveness. Since age is a multilevel ordinal categorical feature, we perform one-hot-encoding and convert it into a dummy variable to represent every level on a binary platform. Similarly, we convert the Gender feature and represent the levels in a binary form.

To have a descriptive look at the data and have some basic understanding of the representativeness of the information, we used R language and plot some visualizations to understand the information.

Team Mean, Median, Mode
Rei Leenah Balachandran s3112637
Shubhankar Sanjay Jahagirdar s3793593
Tanmay Madan Shendkar s3735580

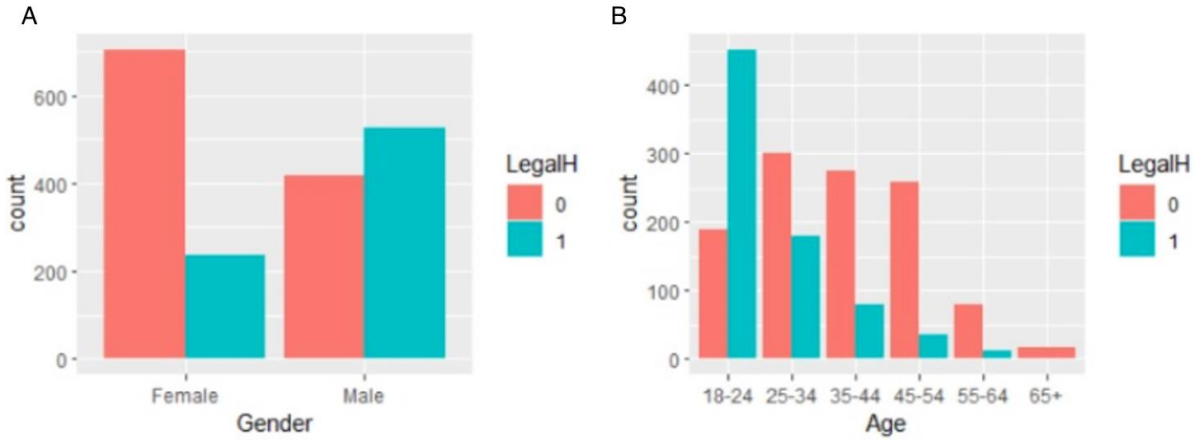

Figure 1: *Bar plots of relationship between A) gender and B) age with consumption of legal highs.*

Based on the above Figure 1 in terms of Gender, we observed that the number of Males that consume legal highs is more than the Females. It can also be observed that there is a significant difference in the number of consumers and non-consumers for Females, however, the difference in consumers and non-consumers of Males is very less with consumers being on a higher side contradictory to females. Considering the Age feature, it can be observed that except for the age group of 18-24 the number of consumers is less as compared to the number of non-consumers of Legal Highs. For the age group 65+, the data set describes no consumers of legal highs. With this information, it is evident that in the provided population information Males in the age group of 18-24 are most vulnerable to the consumption of Legal Highs. However, to have more information and better analysis we use R to plot all the numerical variables as well.

Based on the graphical representations of Figure 2, it can be observed that the consumption of Legal High is associated with lower Agreeableness (A-score) and Consciousness (C-score). It can also be observed that although the Extraversion (E-score) parameter does not provide any valuable information towards the consumption of legal highs, high Neuroticism (N-score), Openness (O-score), impulsiveness, and sensation seeking shows likelihood towards consumption of legal highs. Outliers are clearly represented in Figure 2 boxplots.

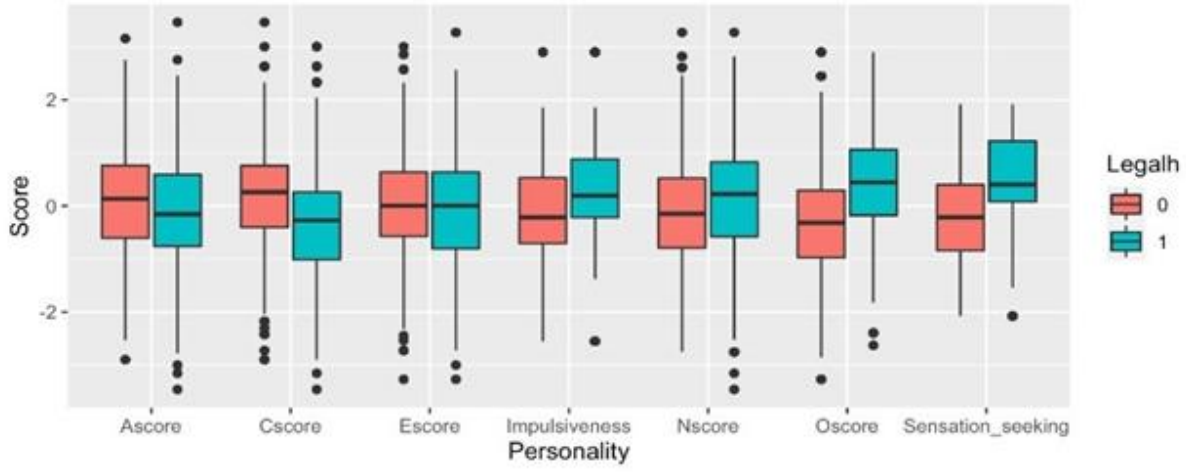

Figure 2: *Relationship between personality traits with legal highs consumption.*

Team Mean, Median, Mode
Rei Leenah Balachandran s3112637
Shubhankar Sanjay Jahagirdar s3793593
Tanmay Madan Shendkar s3735580

4. METHODS

4.1 MATHEMATICAL MODEL AND PRIOR SPECIFICATION

In defining the mathematical model, the class variable - consumption of legal high - is a binary variable and it is dependent on 13 variables. Since the aim is to model the probability of legal high consumption of each subject the likelihood would be Bernoulli, and theta can be modelled by a Logistic Regression model. As shown in figure 2, there are outliers in the 7 personality scores, therefore Robust Logistic Regression was implemented as it is robust to the impact of outliers.

Below is the mathematical model of the robust logistic regression, while figure 3 shows the JAGS model diagram, where each beta is a continuous variable and is represented by a normal distribution.

$$Y_i \sim Bernoulli(\theta)$$

$$\theta = logistic(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_{11} + \beta_{12} x_{12} + \beta_{13} x_{13})$$

Where:

x_1 = Age (18 - 24) Binary variable (0: False 1: True)

x_2 = Age (25 - 34) Binary variable (0: False 1: True)

x_3 = Age (35 - 44) Binary variable (0: False 1: True)

x_4 = Age (45 - 54) Binary variable (0: False 1: True)

x_5 = Age (55 - 64) Binary variable (0: False 1: True)

x_6 = Gender Binary variable (0: Female 1: Male)

x_7 = N-score Neuroticism score of an individual

x_8 = E-score Extraversion score of an individual

x_9 = O-score Openness to experience score of an individual

x_{10} = A-score Agreeableness score of an individual

x_{11} = C-score Conscientiousness score of an individual

x_{12} = Impulsiveness score of an individual

x_{13} = Sensation Seeking score of an individual

In attempt to specify the priors, we have looked upon studies that have been conducted on predictors of drug consumption but were not able to find information in relevant quantifiable terms that can be applied to the priors, hence we modelled with non-informative priors.

Team Mean, Median, Mode
Rei Leenah Balachandran s3112637
Shubhankar Sanjay Jahagirdar s3793593
Tanmay Madan Shendkar s3735580

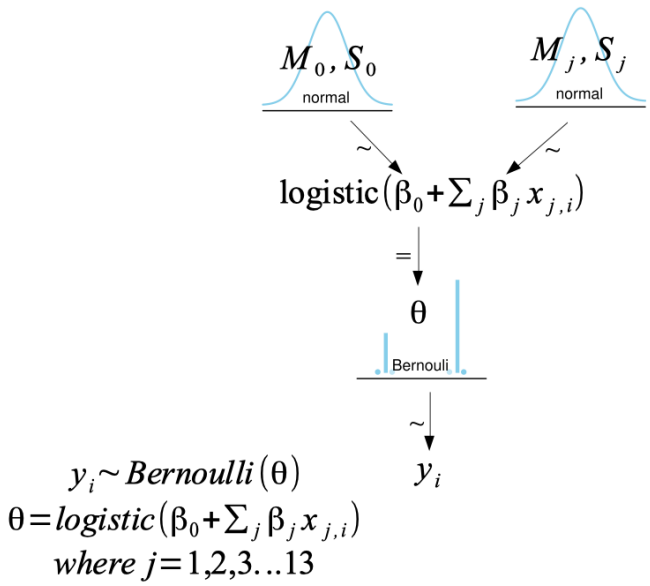

Figure 3. *JAGS model diagram for probability of consumption of legal high.*

4.2. Determining MCMC Setting

In order to obtain an optimal MCMC setting that would produce good diagnostics, a process of trial and error was carried out with the MCMC setting, where the diagnostics of each run was studied, and the setting for the subsequent run was adjusted accordingly until the desired diagnostic is observed. If the trace plot is not overlapping well, and the shrink factor is not below 1.20, the number of burn-in steps will need to be increased, if the autocorrelation is still present at high level, the number of thinning step will need to be increased, and if the density plot does not produce a nice curve and the 95% HDIs does not overlap each other, the number of saved steps will need to be increased. As shown in table 1, the first run had low setting and progressed to a high setting of the fourth run, understandably, the elapsed time also increases with increase setting.

Run number	Adaptation steps	Burn-in steps	Number of chains	Number of Thinning	Number of saved steps	Elapsed time (sec)
1	500	500	3	5	500	1534.80
2	1000	1500	3	50	1000	5105.99
3	1500	2000	3	160	2000	32209.20
4	1500	2000	3	200	4000	79967.39

Table 1. *MCMC run trials and their setting*

A common issue of high autocorrelation is present in the diagnostic when the MCMC chains were run on unstandardised data. We attempted the first run with the 7 personality scores unstandardised and did not observe any autocorrelation in the diagnostic so we proceed with the MCMC runs in such a manner. Interestingly there was very high autocorrelation present for the coefficients of all age groups and gender, and it requires a high thinning value of 200 to reduce the autocorrelation to a minimum.

For the chains to be *representative* of the posterior distributions, the diagnostic has to show that the trace plots, density plots, and shrink factor plots meet acceptable criteria. The chains

in the trace plots need to overlap each other and mix well, this shows that the chains have converged. The histogram plots of the chains need to be mostly smooth in shape and overlapping each other with overlapping 95% HDIs, and the shrink factor plot should have median values below 1.20 to indicate convergence.

The chains also need to be *accurate* of the posterior distributions, which means that the ACF plots need to show minimal or no autocorrelation, the Effective Sample Size (EES) number needs to be high, ideally close to or even above 10,000, and Markov Chain Standard Error (MCSE) value needs to be very low, the closer to zero the better. A high ESS and low MCSE suggests stable and accurate chains.

Figure 4 to 17 shows the diagnostics of the 14 betas from the model. It can be observed from all these diagnostics that the chains overlap each other well in the trace plots (plot A), the median value of the shrink factor is below 1.2 in the shrink factor plot (plot C), and the density plot (plot D) overlap each other nicely, and are smooth in shape, with close or overlapping 95% HDI. All these Suggests that the chains are representative.

At the highest MCMC setting, there was still some amount of autocorrelation in $\beta_0, \beta_1, \beta_2, \beta_3$ and β_4 but it is very minimal, and correspondingly, the ESS values are also below 10,000, while there is virtually no autocorrelation in all the other betas, and correspondingly, their ESS value are above 10,000. The MCSE for all betas are very low. All these suggests that the chains are accurate.

The data was split into 80% for train data and 20% for test data and ran on MCMC. It is worth noting that the diagnostics for all the predictions of the test data are also representative and accurate, however, they will not be included in this report.

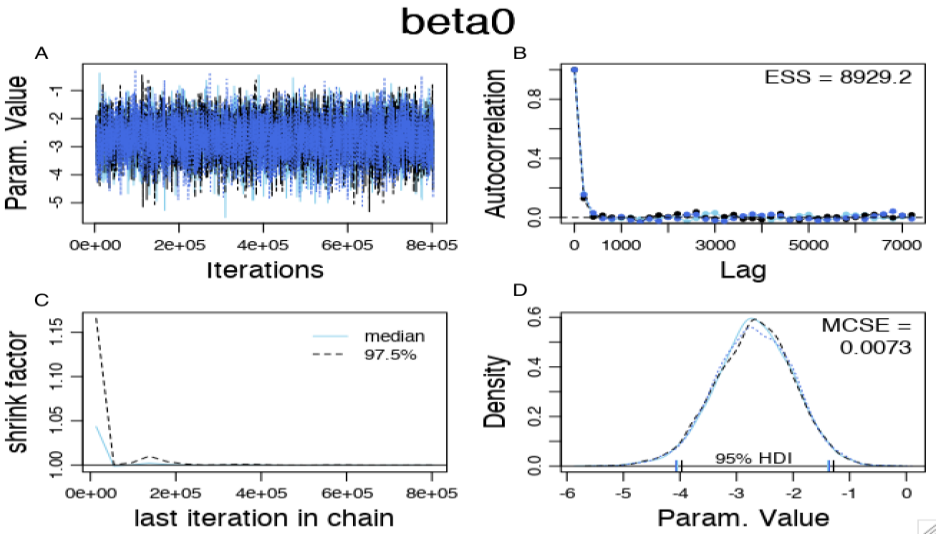

Figure 4. *Diagnostic plots of β_0. A) Trace plot, B) ACF plot, C) Shrink factor plot, D) Density plot.*

Team Mean, Median, Mode
Rei Leenah Balachandran s3112637
Shubhankar Sanjay Jahagirdar s3793593
Tanmay Madan Shendkar s3735580

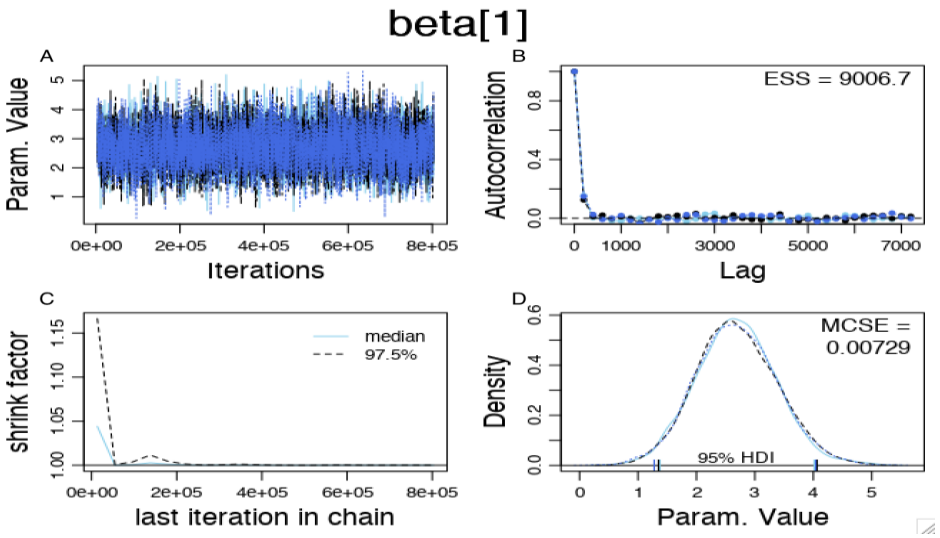

Figure 5. *Diagnostic plots of β_1. A) Trace plot, B) ACF plot, C) Shrink factor plot, D) Density plot.*

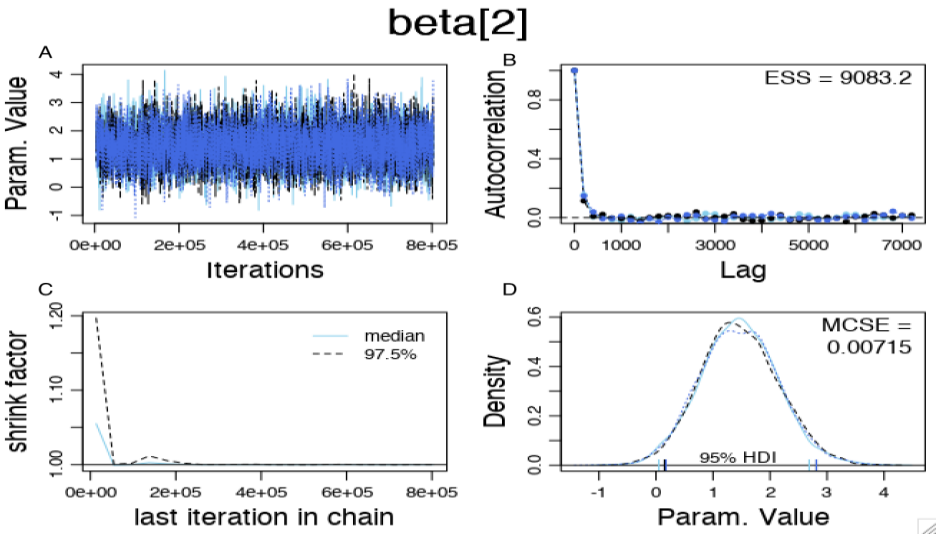

Figure 6. *Diagnostic plots of β_2. A) Trace plot, B) ACF plot, C) Shrink factor plot, D) Density plot.*

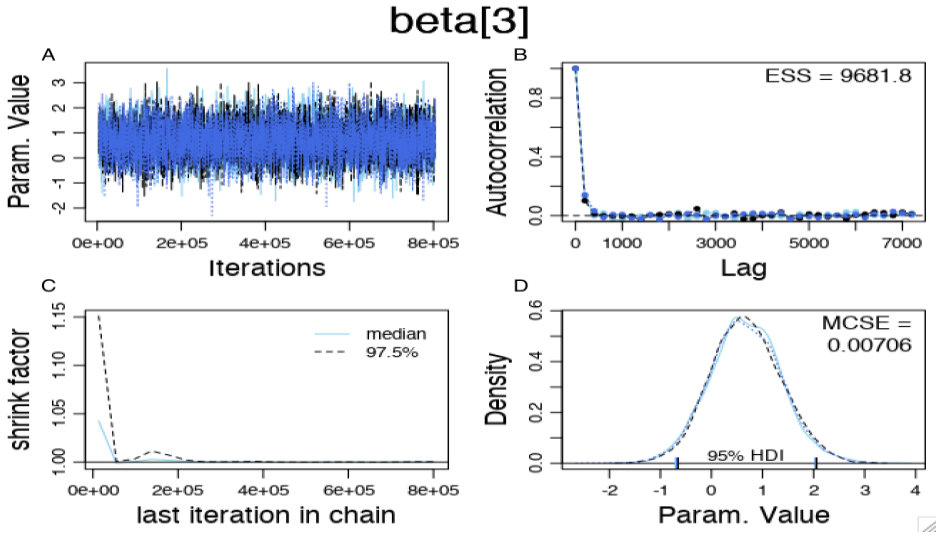

Figure 7. *Diagnostic plots of β_3. A) Trace plot, B) ACF plot, C) Shrink factor plot, D) Density plot.*

Team Mean, Median, Mode
Rei Leenah Balachandran s3112637
Shubhankar Sanjay Jahagirdar s3793593
Tanmay Madan Shendkar s3735580

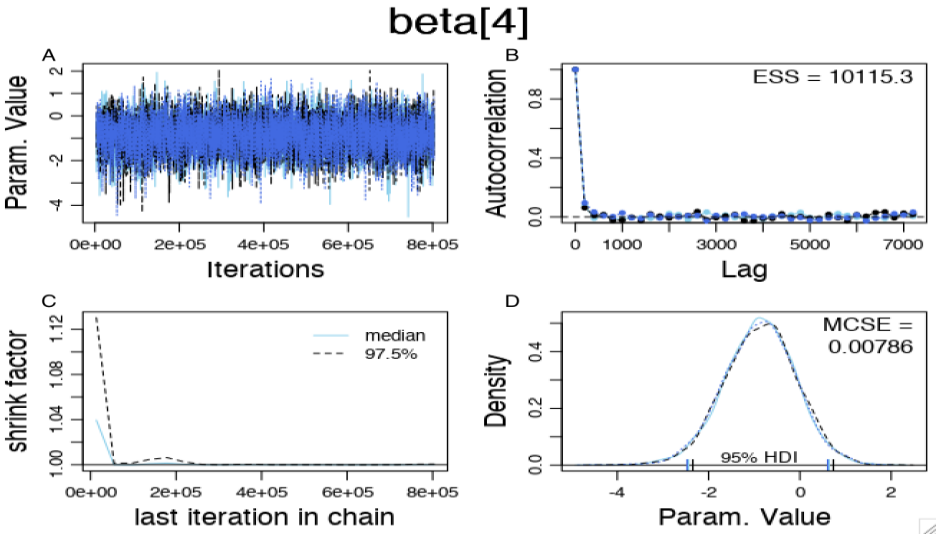

Figure 8. *Diagnostic plots of β_4. A) Trace plot, B) ACF plot, C) Shrink factor plot, D) Density plot.*

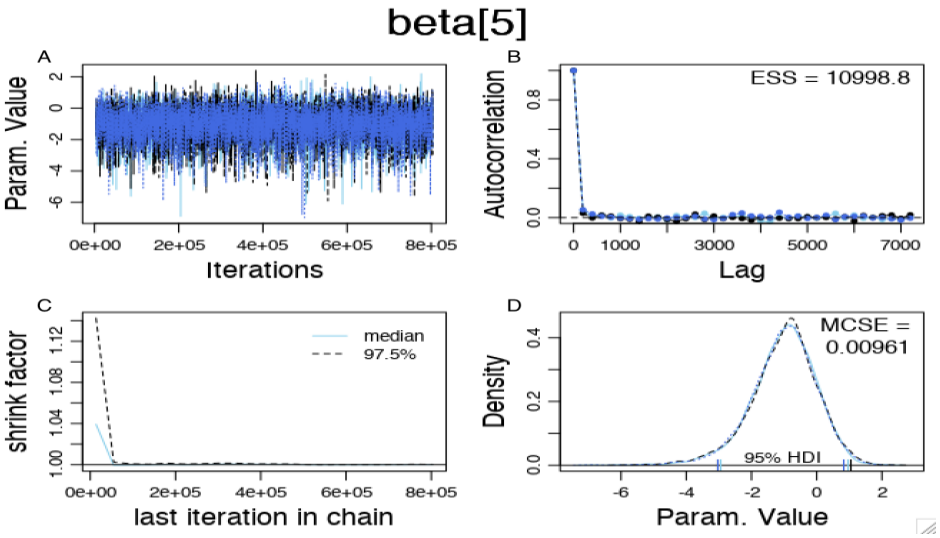

Figure 9. *Diagnostic plots of β_5. A) Trace plot, B) ACF plot, C) Shrink factor plot, D) Density plot.*

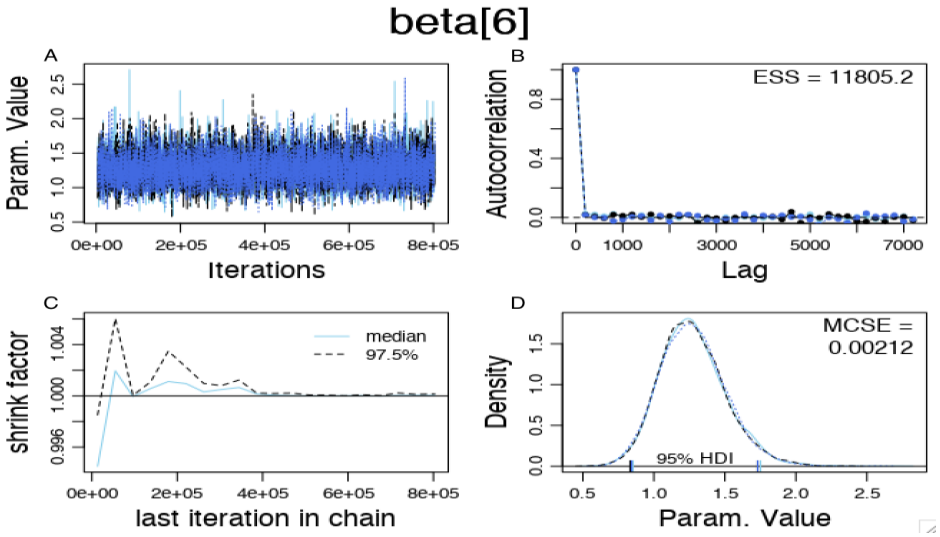

Figure 10. *Diagnostic plots of β_6. A) Trace plot, B) ACF plot, C) Shrink factor plot, D) Density plot.*

Team Mean, Median, Mode
Rei Leenah Balachandran s3112637
Shubhankar Sanjay Jahagirdar s3793593
Tanmay Madan Shendkar s3735580

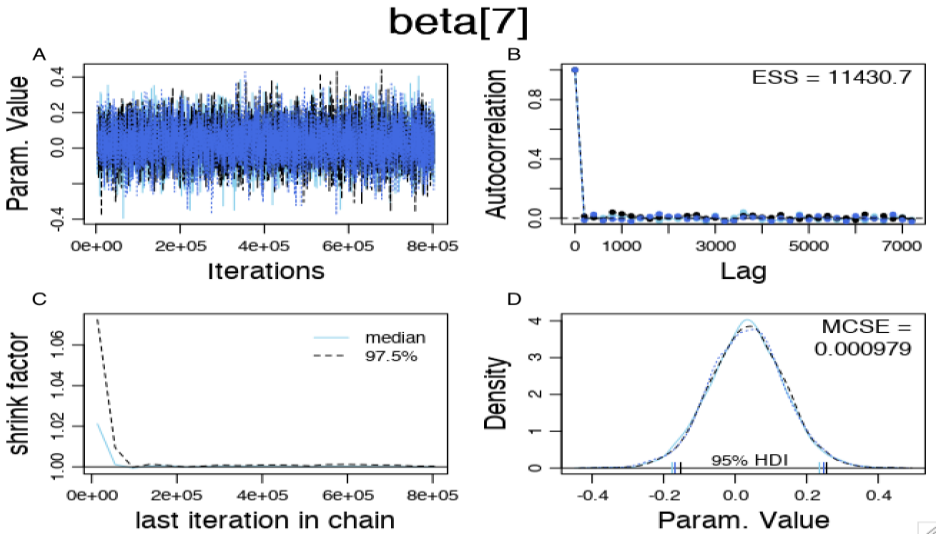

Figure 11. *Diagnostic plots of β_7. A) Trace plot, B) ACF plot, C) Shrink factor plot, D) Density plot.*

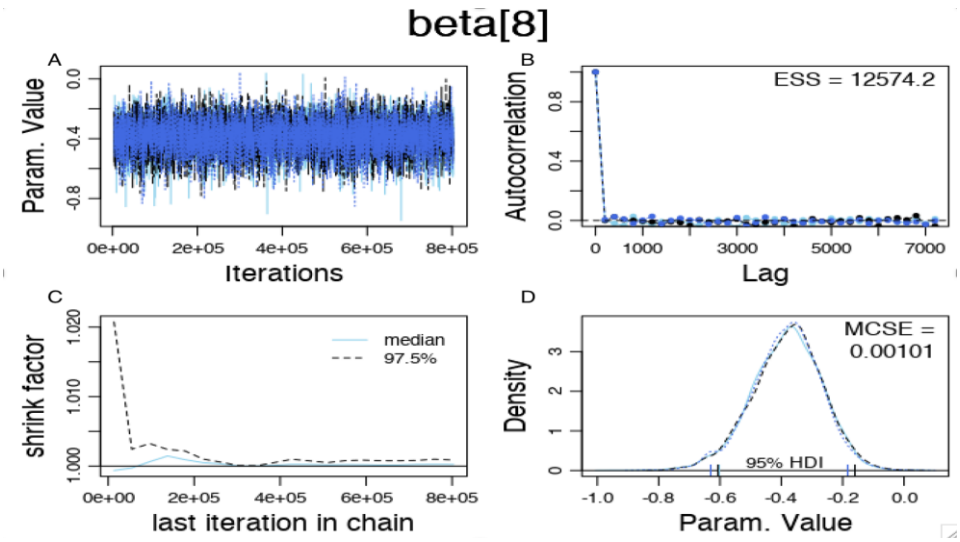

Figure 12. *Diagnostic plots of β_8. A) Trace plot, B) ACF plot, C) Shrink factor plot, D) Density plot.*

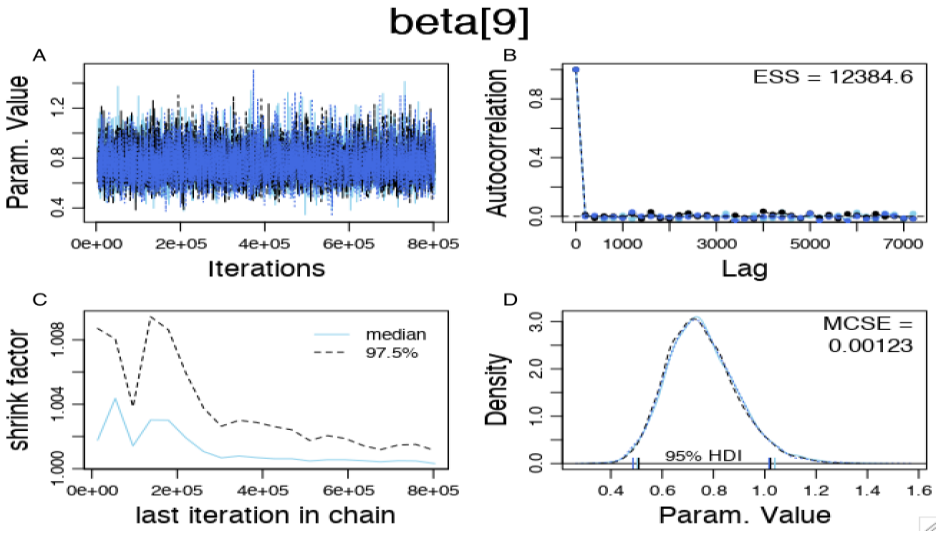

Figure 13. *Diagnostic plots of β_9. A) Trace plot, B) ACF plot, C) Shrink factor plot, D) Density plot.*

Team Mean, Median, Mode
Rei Leenah Balachandran s3112637
Shubhankar Sanjay Jahagirdar s3793593
Tanmay Madan Shendkar s3735580

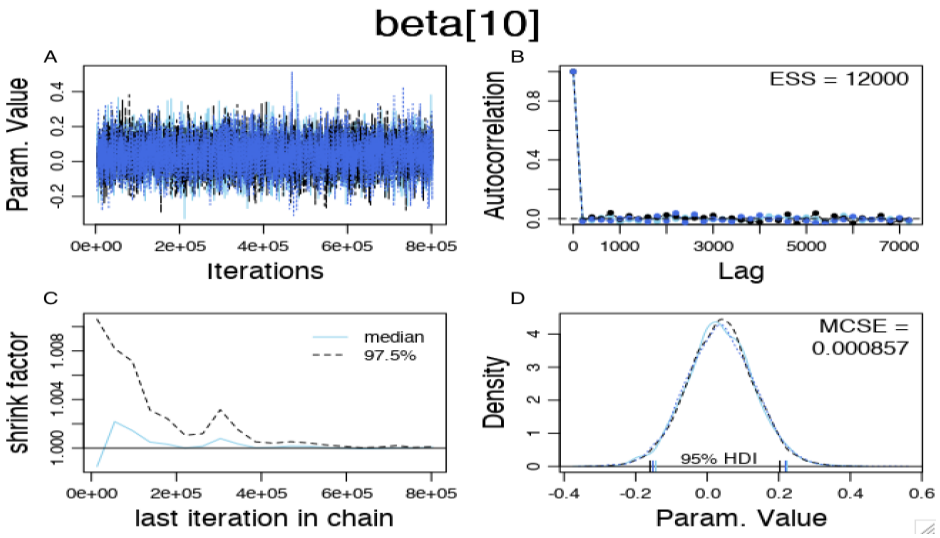

Figure 14. *Diagnostic plots of β_{10}. A) Trace plot, B) ACF plot, C) Shrink factor plot, D) Density plot.*

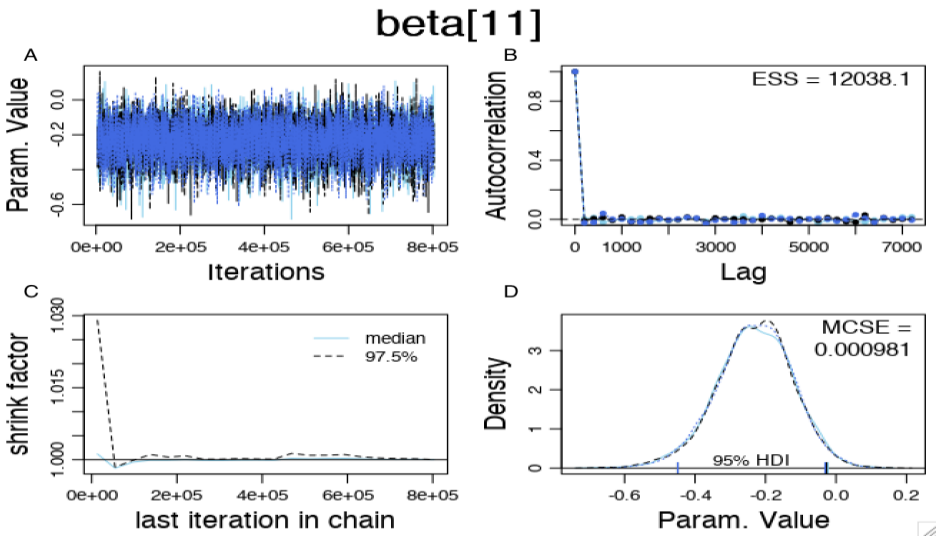

Figure 15. *Diagnostic plots of β_{11}. A) Trace plot, B) ACF plot, C) Shrink factor plot, D) Density plot.*

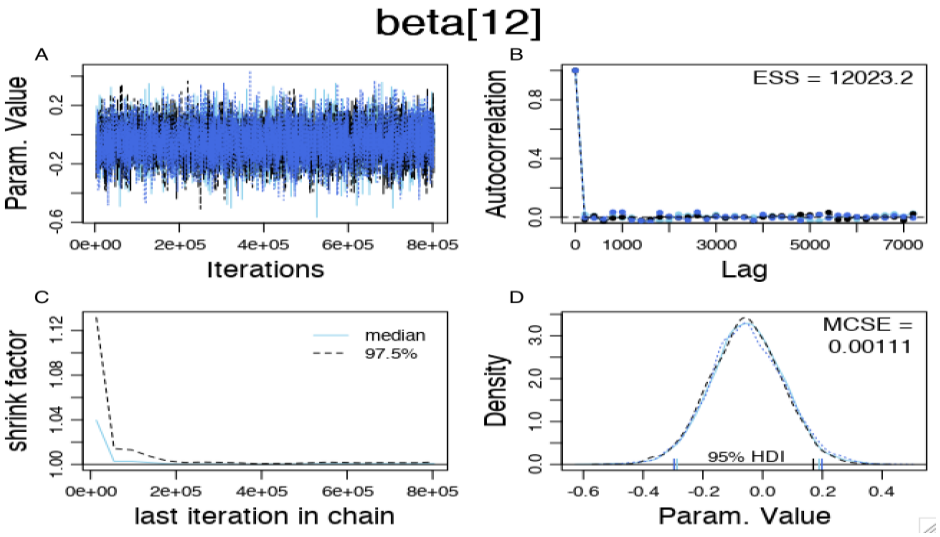

Figure 16. *Diagnostic plots of β_{12}. A) Trace plot, B) ACF plot, C) Shrink factor plot, D) Density plot.*

Team Mean, Median, Mode
Rei Leenah Balachandran s3112637
Shubhankar Sanjay Jahagirdar s3793593
Tanmay Madan Shendkar s3735580

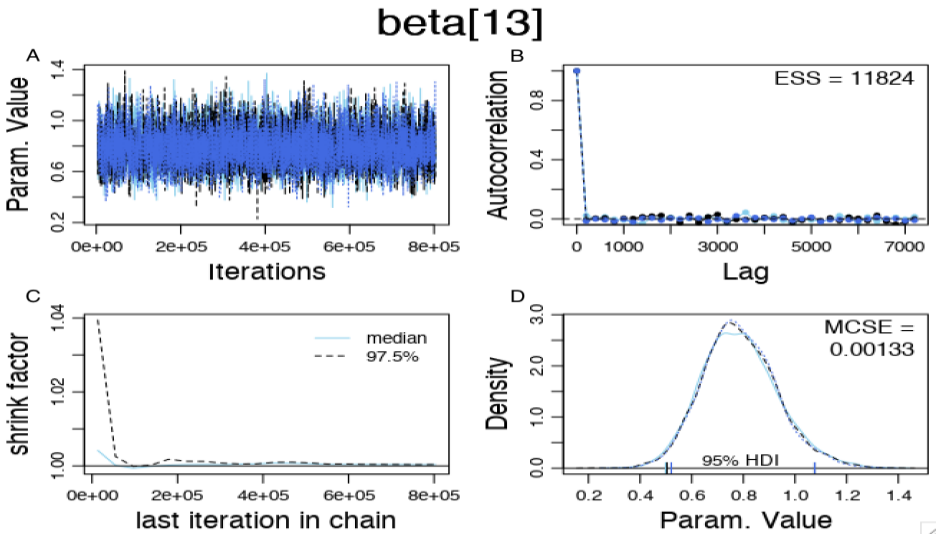

Figure 17. *Diagnostic plots of β_{13}. A) Trace plot, B) ACF plot, C) Shrink factor plot, D) Density plot.*

6. Results and Model Interpretation

The result of the model parameters and comparison between coefficients $\beta_1, \beta_2, \beta_3, \beta_4$ and β_5 corresponding to the 5 age groups are shown in Table 2, and figure 18 shows the posterior distribution of all the model parameters.

The exponential of mode was computed to facilitate the interpretation of the effect of the coefficients on the probability of legal high consumption, as it provides insights in terms of the odds of legal highs consumption over non-consumption.

When the odds approach 0, non-consumption is more probable, logically, when the odds approach 1, there is no difference between the probability of consumption and non-consumption, therefore when the odds approach infinity, consumption is more probable.

Following the above logic, the result shows that relative to the 65+ age group, the probability of legal high consumption decreases with an increase in age groups. This is consistent with our data exploration. Also consistent with our data exploration is that being in the lowest age group, 18-24 years significantly increases the probability of legal high consumption. Being in the 18-24 age group would increase the odds by 13.42 (exp(β_1)), which means that legal high consumption becomes 13.42 times more probable than non-consumption. Being in the next age group, 25-34 reduces the odds to 3.97 (exp(β_2)), there is a further decrease in the odds to 1.71 (exp(β_3)) in the age group 35-44. The turning point began in the age group 45-54, where the odds are 0.48 (exp(β_4)), which meant that non-consumer became approximately 2 times more probable, similarly the odds are 0.46 (exp(β_5)) being in the 55-64 age group.

Also consistent with our data exploration, compared to being a female, being a male makes legal high consumption 3.47 times more probable, since the odd is 3.47 (exp(β_6)).

Contrary to our data exploration, the coefficients for neuroticism score (β_7) assertiveness score (β_{10}) and impulsiveness score (β_{12}) does not have an impact on the probability of consumption, as the odds are very close to one. This is also supported by the location of the zero value of these betas being close to the center of the 95% HDI as seen in figure 18.

Team Mean, Median, Mode
Rei Leenah Balachandran s3112637
Shubhankar Sanjay Jahagirdar s3793593
Tanmay Madan Shendkar s3735580

	Mean	Median	Mode	exp (Mode)	HDImass	HDILow	HDIHigh	CompVal	%GtCompVal
$\boldsymbol{\beta_0}$	-2.6866	-2.6736	-2.7241	0.0656	0.95	-4.0538	-1.3647	-2.6	45.5667
$\boldsymbol{\beta_1}$	2.6870	2.6682	2.5971	13.4245	0.95	1.3366	4.0381	2.6	53.8917
$\boldsymbol{\beta_2}$	1.4623	1.4484	1.3788	3.9702	0.95	0.1551	2.8224	1.320	57.3250
$\boldsymbol{\beta_3}$	0.6750	0.6584	0.5390	1.7142	0.95	-0.6777	2.0502	0.460	61.5250
$\boldsymbol{\beta_4}$	-0.8765	-0.8577	-0.7213	0.4861	0.95	-2.4425	0.6555	-0.830	48.7167
$\boldsymbol{\beta_5}$	-1.0156	-0.9379	-0.7691	0.4634	0.95	-3.0153	0.9387	-0.970	51.4250
$\boldsymbol{\beta_6}$	1.2740	1.2584	1.2440	3.4696	0.95	0.8408	1.7351	1.260	49.6917
$\boldsymbol{\beta_7}$	0.0332	0.0339	0.0398	1.0406	0.95	-0.1656	0.2491	0.0	63.0333
$\boldsymbol{\beta_8}$	-0.3823	-0.3766	-0.3586	0.6987	0.95	-0.6088	-0.1602	-0.410	61.4833
$\boldsymbol{\beta_9}$	0.7564	0.7442	0.7283	2.0715	0.95	0.5087	1.0378	0.730	54.5333
$\boldsymbol{\beta_{10}}$	0.0372	0.0364	0.0351	1.0357	0.95	-0.1494	0.2193	0	65.5333
$\boldsymbol{\beta_{11}}$	-0.2311	-0.2277	-0.2498	0.7790	0.95	-0.4488	-0.0256	-0.166	27.8083
$\boldsymbol{\beta_{12}}$	-0.0528	-0.0540	-0.0554	0.9461	0.95	-0.2958	0.1833	0	32.9583
$\boldsymbol{\beta_{13}}$	0.7864	0.7777	0.7449	2.1062	0.95	0.5021	1.0721	0.780	49.3583
guess	0.0727	0.0712	0.0731		0.95	0.0015	0.1387		
$\boldsymbol{\beta_1 - \beta_2}$	1.2248	1.2147	1.2320	3.4283	0.95	0.7962	1.6868	0	100.0
$\boldsymbol{\beta_1 - \beta_3}$	2.0121	1.9812	1.9062	6.7274	0.95	1.4254	2.6694	0	100.0
$\boldsymbol{\beta_1 - \beta_4}$	3.5635	3.4964	3.3106	27.4003	0.95	2.5405	4.6669	0	100.0
$\boldsymbol{\beta_1 - \beta_5}$	3.7026	3.5444	3.3585	28.0799	0.95	2.1409	5.4687	0	100.0
$\boldsymbol{\beta_2 - \beta_3}$	0.7874	0.7697	0.7323	2.0799	0.95	0.2938	1.3159	0	100.0
$\boldsymbol{\beta_2 - \beta_4}$	2.3387	2.2818	2.2179	9.1881	0.95	1.4234	3.3118	0	100.0
$\boldsymbol{\beta_2 - \beta_5}$	2.4779	2.3248	2.1497	8.5826	0.95	1.0492	4.2221	0	100.0
$\boldsymbol{\beta_3 - \beta_4}$	1.5514	1.5027	1.4024	4.0649	0.95	0.7162	2.5156	0	100.0
$\boldsymbol{\beta_3 - \beta_5}$	1.6905	1.5575	1.4322	4.1879	0.95	0.1969	3.3451	0	100.0
$\boldsymbol{\beta_4 - \beta_5}$	0.1391	0.0582	0.0344	1.0350	0.95	-1.5206	1.9042	0	100.0

Table 2. *Summary table of model parameters.*

Our data exploration shows that there is no difference in extraversion score between consumers and non-consumers, however, this model shows that a subject is 1.43 (exp(β_8) of 0.69) times more probable of being a non-consumer with each unit increase in extraversion score. The relationship between openness to experience, conscientiousness, and sensation seeking and legal high consumption provided by the model follows an obvious logic that an individual with a higher score on openness to experience and sensation-seeking score will be more likely to consume legal high, while a higher score in conscientiousness will be less likely to consume legal high. Therefore, each unit increase in openness to experience score makes legal high consumption 2.07 times more probable, each unit increase in conscientious score makes non-consumption 1.28 (exp(β_{11}) of 0.779) more probable, and each unit increase in sensation seeking score makes consumption 2.11 times more probable.

Since there are 6 levels in the age category, a comparison between the different age groups was also carried out to understand the difference in effect on the probability of legal high consumption.

It can be observed further down table 2, relative to the age group 65+, the difference in odds between age groups 18-24 and 25-34 ($\beta_1 - \beta_2$) is 3.43, and increase to 6.73 between age groups 18-24 and 35-44 ($\beta_1 - \beta_3$). There is a significant jump in the difference between age

groups 18-24 and 45-54 ($\beta_1 - \beta_4$) to the odds of 27.40, and similarly a difference of 28.08 in the odds between 18-24 and 55-64.

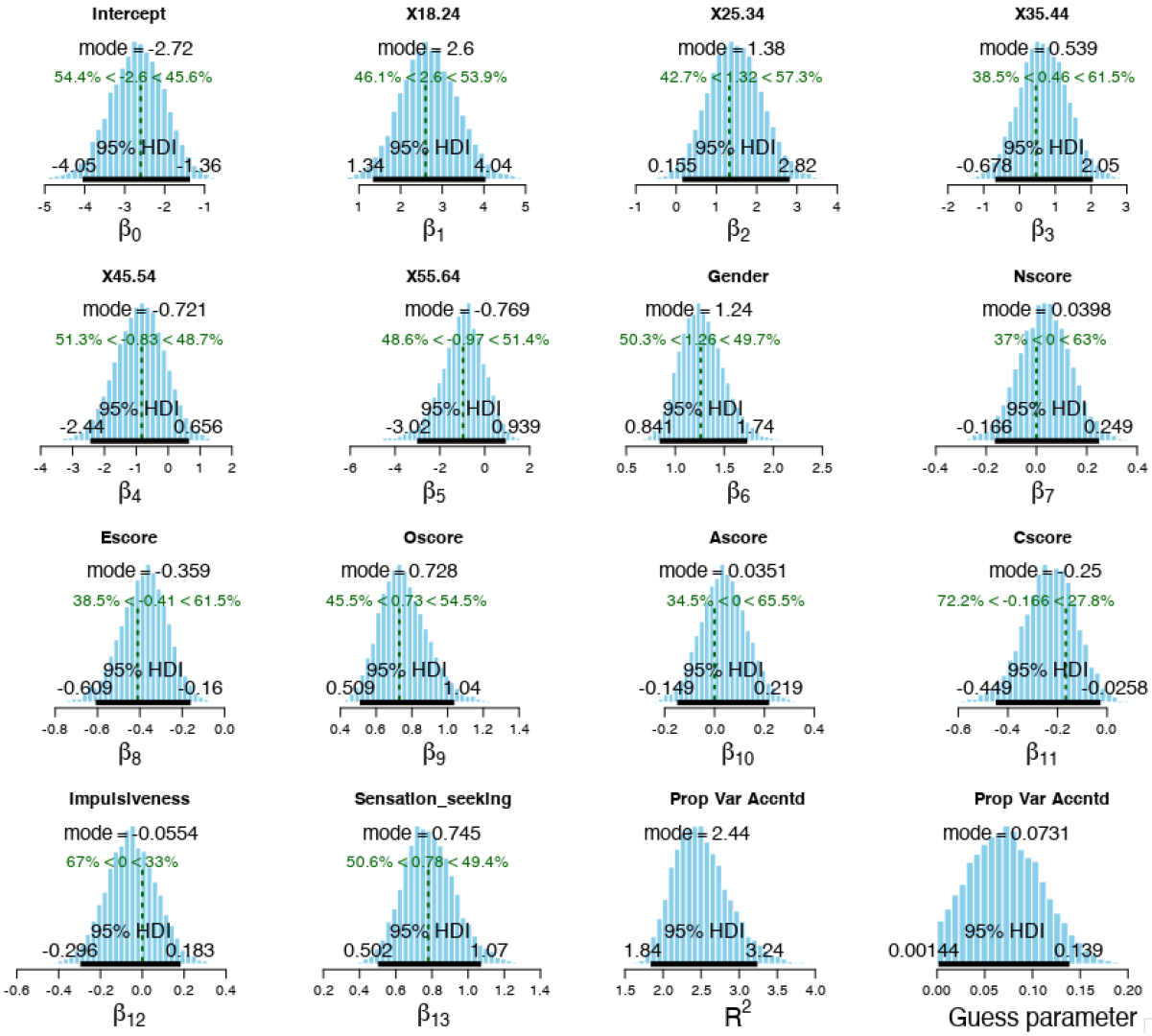

Figure 18. *Posterior distribution of model parameters.*

The difference in odd between age groups 25-34 and 35-44 ($\beta_2 - \beta_3$) is 2.08 and a jump to 9.19 in odds difference between age group 25-34 and 45-54 ($\beta_2 - \beta_4$), the difference in odds between the age groups 25-34 and 55-64 ($\beta_2 - \beta_5$) is similar at 8.59.

The difference in odds between age group 35-44 and 45-55 ($\beta_3 - \beta_4$) is 4.06 and similarly a difference of 4.18 in odd between ($\beta_3 - \beta_5$), and as expected there is no difference in odd of legal high consumption between age group 45-54 and 55-64 ($\beta_4 - \beta_5$). This trend of difference in odds is consistent with the odd of legal highs consumption presented by each age group.

Since we modelled with a robust logistic regression, there is also a result for the guess parameter. The mode of it is low, at 0.0731, with the 95% HDI between 0.0014 and 0.139, suggesting that the outliers did have some effect on the model.

Team Mean, Median, Mode
Rei Leenah Balachandran s3112637
Shubhankar Sanjay Jahagirdar s3793593
Tanmay Madan Shendkar s3735580

7. Model Exploration

We ran another model omitting the three insignificant coefficients of neuroticism score, assertiveness score, and impulsiveness score. The MCMC ran with the same high setting as the first model with the diagnostics for all the parameters can be seen in figure 19 to 29. They are all representative as indicated by overlapping trace plots (plot A), shrink factor being below 1.20 (plot C), with overlapping density plots, and clustered or overlapping 95% HDI as seen in the density plots (plot D). Understandably, the accuracy of the parameters for this model is very similar to those in the first model, as there is also a low amount of autocorrelation present in $\beta_1, \beta_2, \beta_3$ and β_4, with corresponding ESS value below 10,000, while the rest of the betas are above 10,000, and the MCSE value for all betas were low.

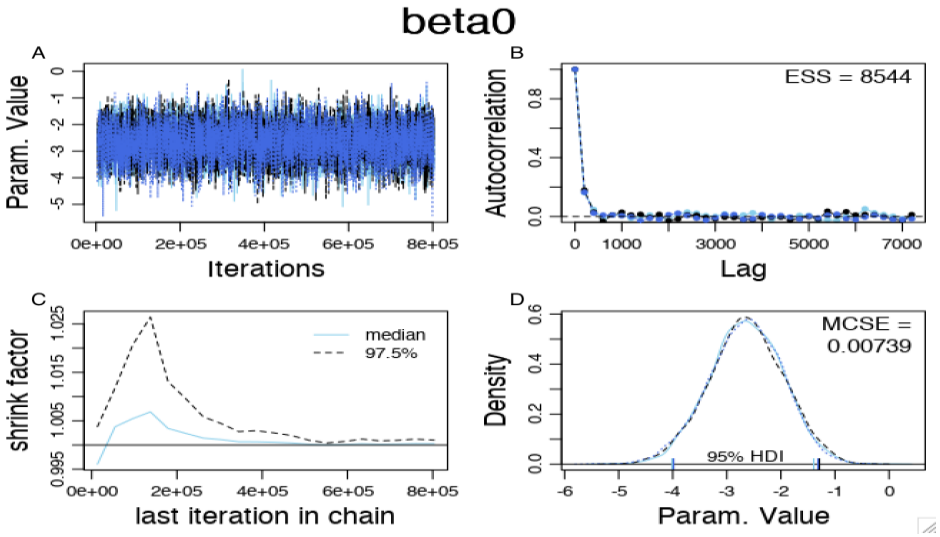

Figure 19. *Diagnostic plots of β_0. A) Trace plot, B) ACF plot, C) Shrink factor plot, D) Density plot.*

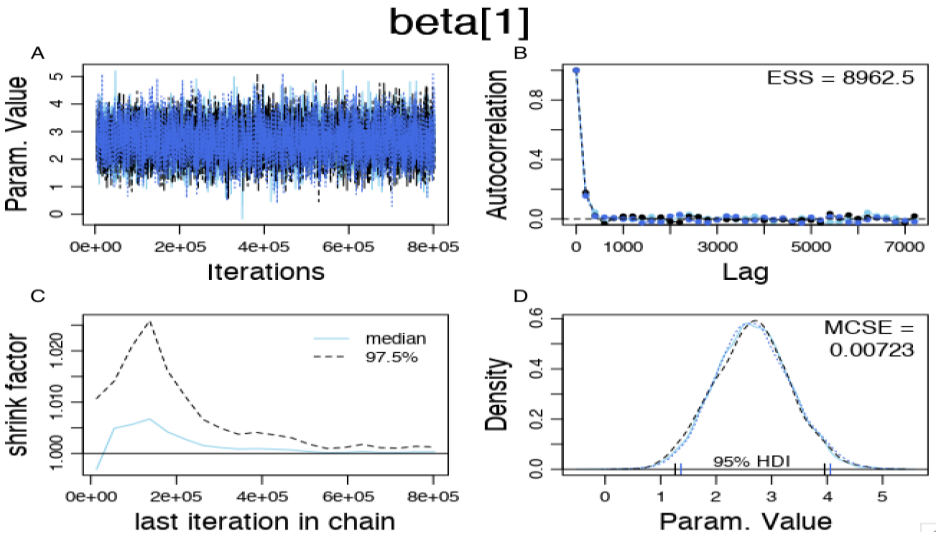

Figure 20. *Diagnostic plots of β_1. A) Trace plot, B) ACF plot, C) Shrink factor plot, D) Density plot.*

Team Mean, Median, Mode
Rei Leenah Balachandran s3112637
Shubhankar Sanjay Jahagirdar s3793593
Tanmay Madan Shendkar s3735580

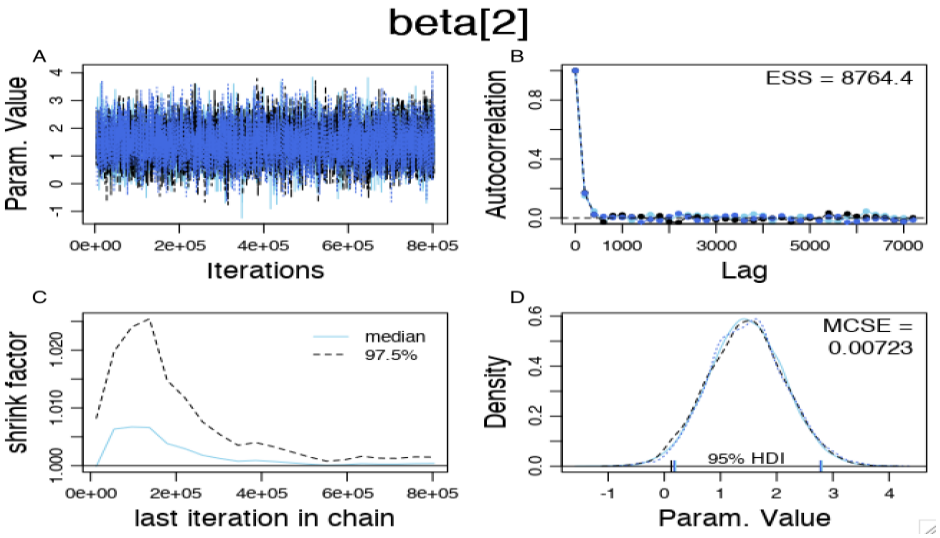

Figure 21. *Diagnostic plots of β_2. A) Trace plot, B) ACF plot, C) Shrink factor plot, D) Density plot.*

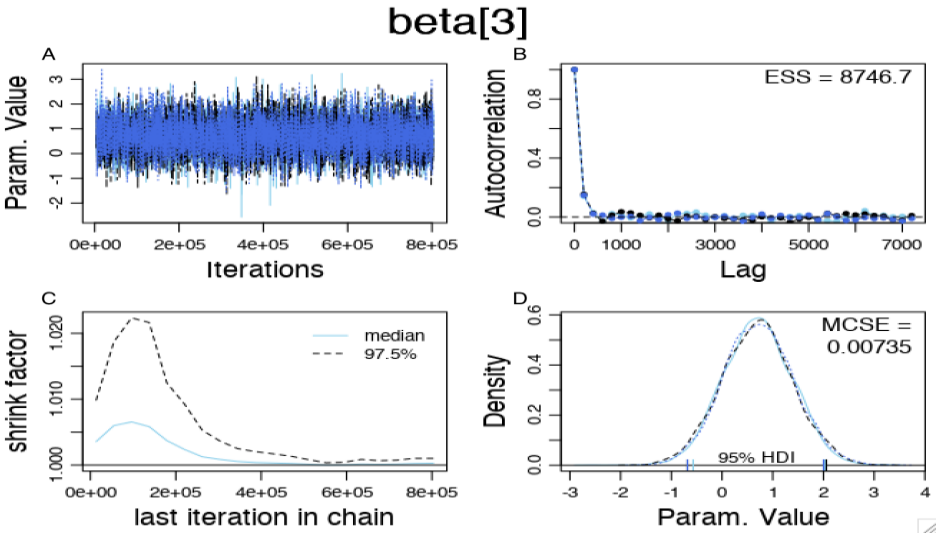

Figure 22. *Diagnostic plots of β_3. A) Trace plot, B) ACF plot, C) Shrink factor plot, D) Density plot.*

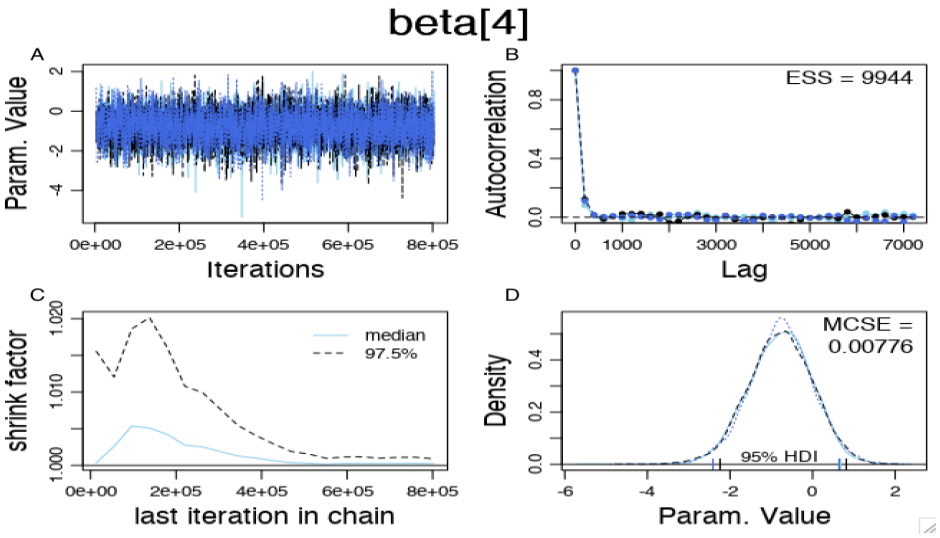

Figure 23. *Diagnostic plots of β_4. A) Trace plot, B) ACF plot, C) Shrink factor plot, D) Density plot.*

Team Mean, Median, Mode
Rei Leenah Balachandran s3112637
Shubhankar Sanjay Jahagirdar s3793593
Tanmay Madan Shendkar s3735580

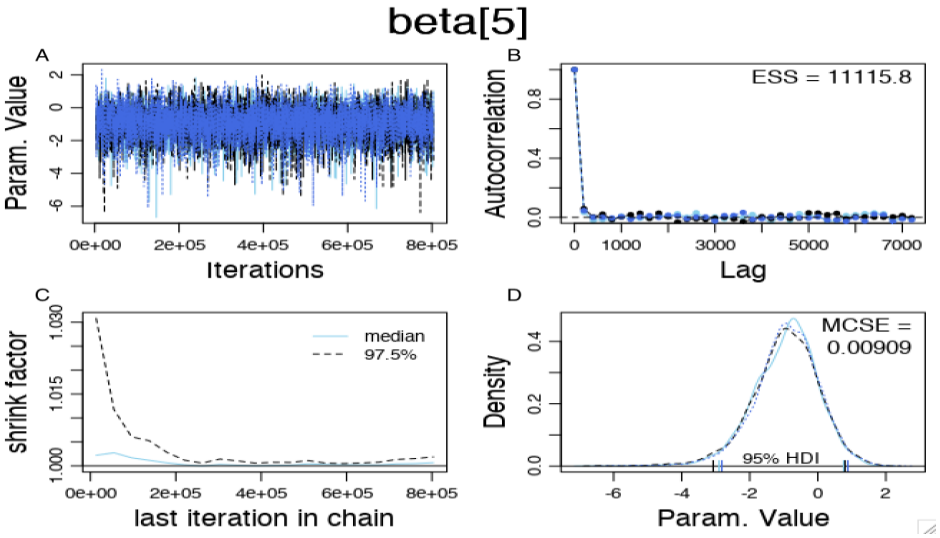

Figure 24. *Diagnostic plots of β_5. A) Trace plot, B) ACF plot, C) Shrink factor plot, D) Density plot.*

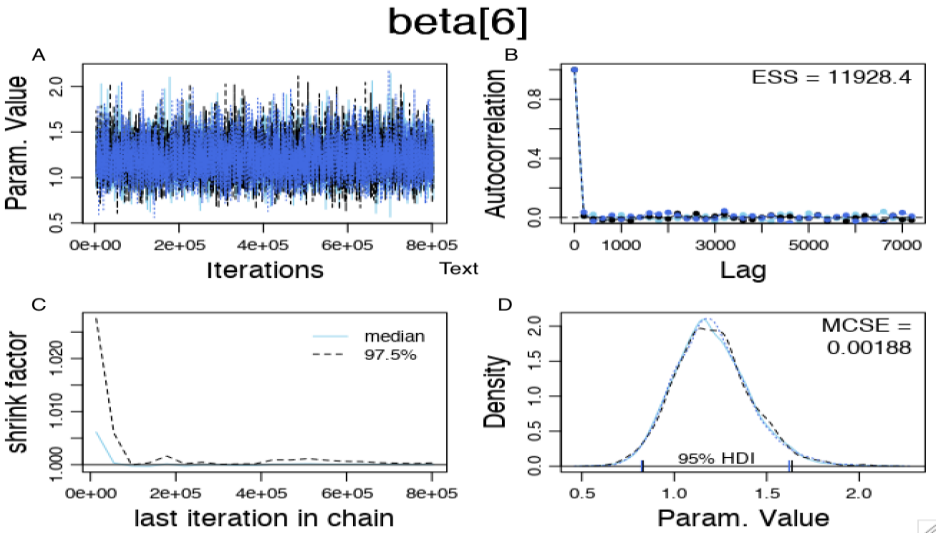

Figure 25. *Diagnostic plots of β_6. A) Trace plot, B) ACF plot, C) Shrink factor plot, D) Density plot.*

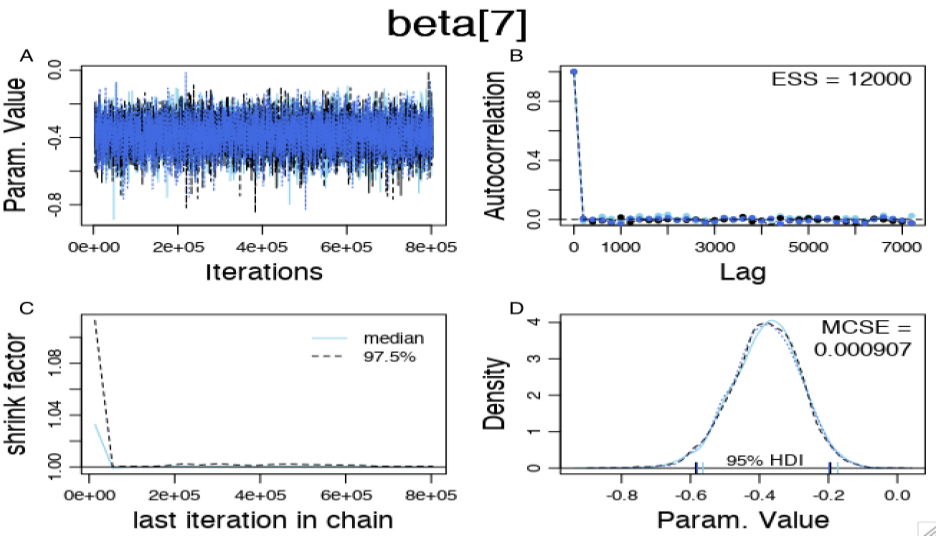

Figure 26. *Diagnostic plots of β_7. A) Trace plot, B) ACF plot, C) Shrink factor plot, D) Density plot.*

Team Mean, Median, Mode
Rei Leenah Balachandran s3112637
Shubhankar Sanjay Jahagirdar s3793593
Tanmay Madan Shendkar s3735580

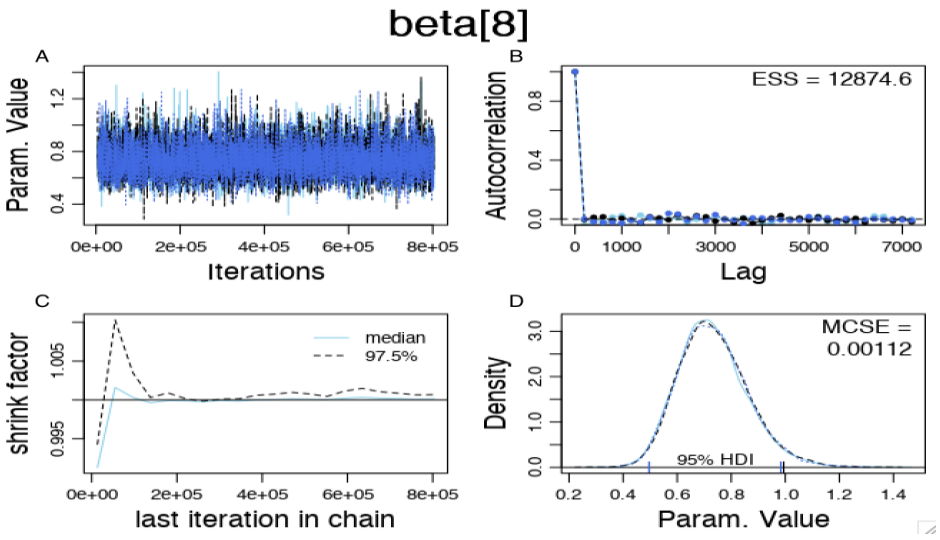

Figure 27. *Diagnostic plots of β_8. A) Trace plot, B) ACF plot, C) Shrink factor plot, D) Density plot.*

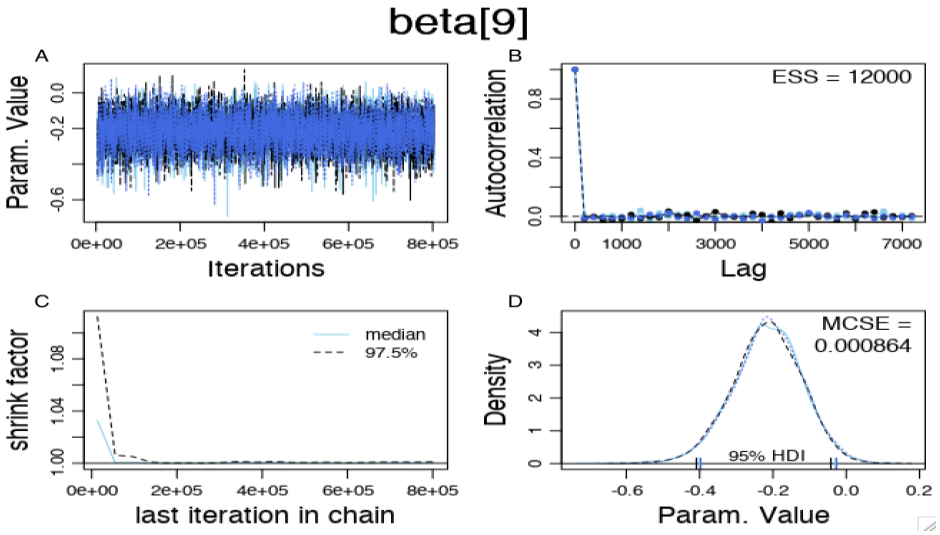

Figure 28. *Diagnostic plots of β_9. A) Trace plot, B) ACF plot, C) Shrink factor plot, D) Density plot.*

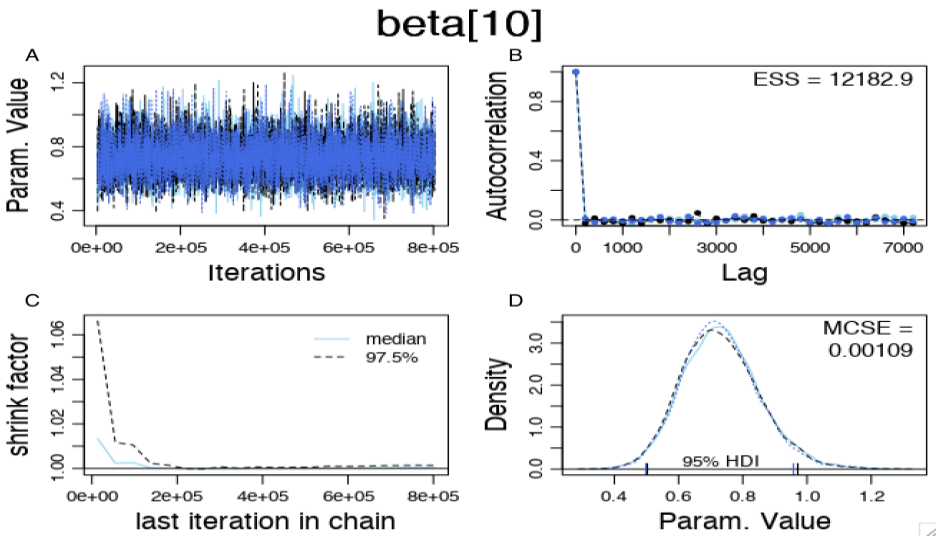

Figure 29. *Diagnostic plots of β_{10}. A) Trace plot, B) ACF plot, C) Shrink factor plot, D) Density plot.*

7.1 Results of Model 2

Team Mean, Median, Mode
Rei Leenah Balachandran s3112637
Shubhankar Sanjay Jahagirdar s3793593
Tanmay Madan Shendkar s3735580

As expected, the summary results and posterior distributions of model 2 parameters are very similar to the equivalent parameters in model 1. This is expected as the MCMC chains were run on the same dataset.

	Mean	Median	Mode	HDImass	HDILow	HDIHigh	CompVal	%Gt CompVal
β_0	-2.6434	-2.6360	-2.6374	0.95	-3.9798	-1.3142	-2.6	47.925
β_1	2.6617	2.6529	2.5920	0.95	1.3226	3.9962	2.6	53.1583
β_2	1.4698	1.4668	1.5866	0.95	-0.5413	2.7937	1.32	58.4417
β_3	0.7072	0.7075	0.7511	0.95	-0.6452	2.0414	0.46	63.0750
β_4	-0.7749	-0.7576	-0.7270	0.95	-2.2366	0.7837	-0.830	53.6667
β_5	-0.9348	-0.8751	-0.8567	0.95	-2.9016	0.8610	-0.970	54.3917
β_6	1.2050	1.1914	1.1572	0.95	0.8272	1.6253	1.260	36.6250
β_7	-0.3808	-0.3769	-0.3738	0.95	-0.5710	-0.1816	-0.410	62.7833
β_8	0.7343	0.7244	0.6990	0.95	-0.4977	0.9878	0.730	48.1333
β_9	-0.2149	-0.2128	-0.2186	0.95	-0.3967	-0.0250	-0.166	30.7083
β_{10}	0.7313	0.7240	0.7080	0.95	0.5014	0.9687	0.780	32.2168

Table 3. *Summary results of model parameters.*

Table 3 shows the summary results and figure 30 shows the posterior distributions of all the parameters.

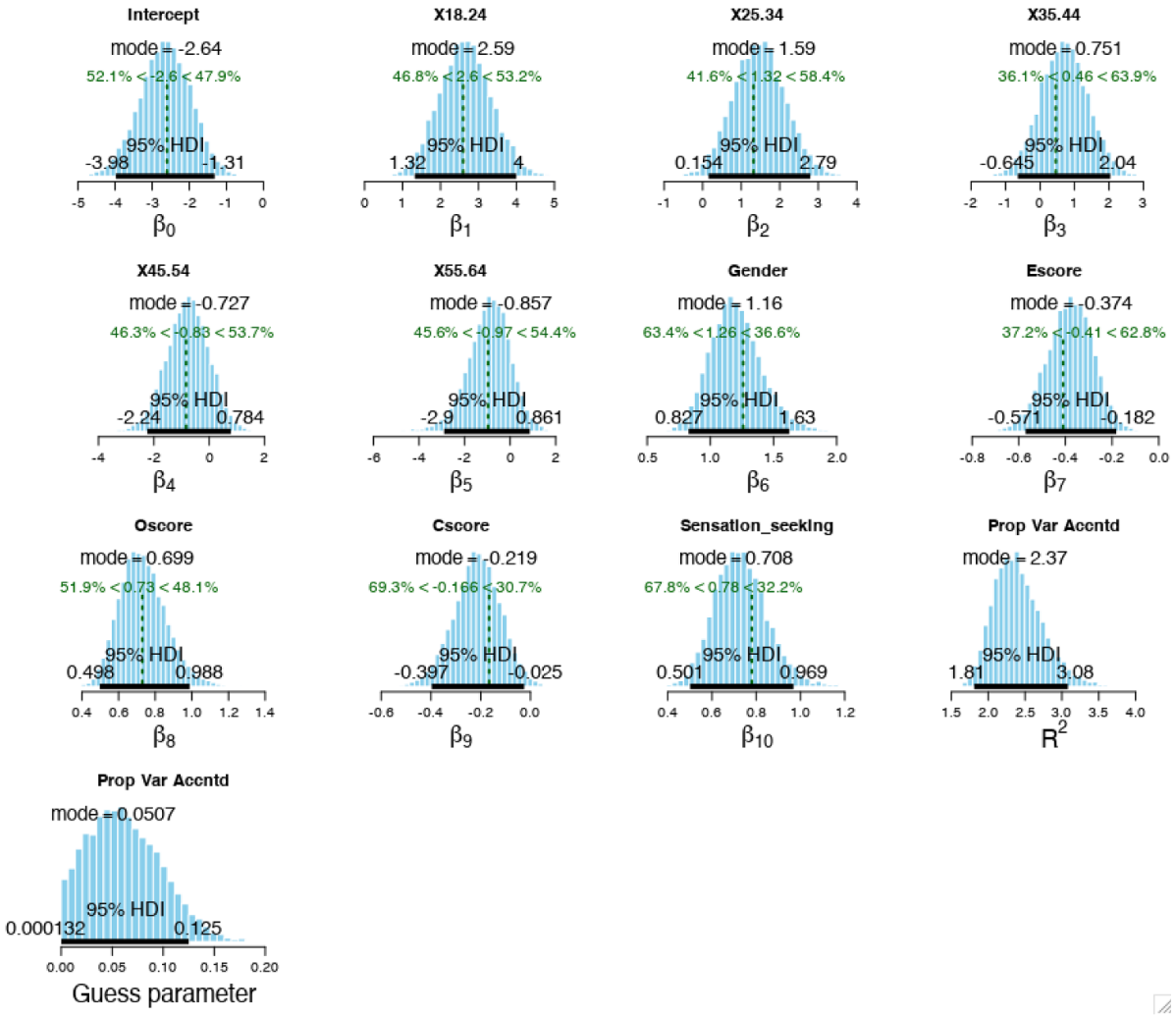

Figure 30. *Posterior distribution of model 2 parameters.*

Table 4 shows the highest metric of model 1 and 2 obtained at a threshold of 0.4. It can be seen that model 1 performed poorly in predicting the consumption of legal high, and although the performance of model 2 is not fantastic, it is apparent that it is significantly better than model 1 by all metrics.

	Model 1	Model 2
Accuracy	0.6207	0.7533
Precision	0.5291	0.6744
Recall	0.5948	0.7581
F-score	0.56	0.7138

Table 4. *Metric values of model 1 and 2.*

8. DISCUSSION AND LIMITATIONS

While conducting this study, we experienced a few limitations that should be discussed. Firstly, as we were unable to find meaningful information to specify the priors, we have carried out the modelling with non-informative priors, therefore we were unable to test for model sensitivity.

Secondly, a model comparison between 5 models was attempted to explore the whole parameter space, with each model containing a different combination of parameters. The output of the MCMC run was peculiar and also unexpected. Model comparison by MCMC is finicky and challenging, and to explore this area it is advisable to use Stan instead, but it is out of the scope of this project.

Although there was an obvious improvement in the performance of model 2 over model 1, the performance of model 2 is outstanding. There may be other more suitable models for this dataset, perhaps they can be trialled for improved performance.

Finally, the original dataset also contains information on education, country, and ethnicity, which were not included in this robust logistic regression model. Perhaps the performance of the model can be improved if these variables were also included. Therefore, it is worth trialling another model with all the variables provided.

9. CONCLUSION

In concluding this study, we were able to carry out a robust logistic regression model with MCMC to predict the consumption of legal high based on predictors including age, gender, and the 7 personality scores on neuroticism, extraversion, openness to experience, assertiveness, conscientiousness, impulsiveness, and sensation-seeking. The first model made use of all these variables and performed with 62% on accuracy, 53% on precision 59% on recall and 56% on F-score.

The coefficients of neuroticism, assertiveness and impulsiveness score were shown to be insignificant, therefore they were omitted in the second model. Model 2 performed significantly better than model 1 with 75% on accuracy, 67% on precision, 76% on recall, and 71% on F-score.

Team Mean, Median, Mode
Rei Leenah Balachandran s3112637
Shubhankar Sanjay Jahagirdar s3793593
Tanmay Madan Shendkar s3735580

10. REFERENCES

Lamptey, J. (2005). Socio-demographic Characteristics of Substance Abusers Admitted to a Private Specialist Clinic. Ghana Medical Journal.

Thomas, S. (2020, October 5). https://americanaddictioncenters.org/rehab-guide/addiction-statisticshttps://americanaddictioncenters.org/rehab-guide/addiction-statistics. Retrieved from https://americanaddictioncenters.org/.

11. APPENDIX

CODE

```
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(beepr)
source("DBDA2E-utilities.R")

#===============PRELIMINARY FUNCTIONS FOR POSTERIOR INFERENCES====================

smryMCMC_HD = function(  codaSamples , compVal=NULL , rope=NULL ,
                         diffSVec=NULL , diffCVec=NULL ,
                         compValDiff=0.0 , ropeDiff=NULL ,
                         saveName=NULL ) {
 mcmcMat = as.matrix(codaSamples,chains=TRUE)
 summaryInfo = NULL

 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
 rowIdx = dim(summaryInfo)[1]

 # differences of beta's:
 if ( !is.null(diffCVec) ) {
  Nidx = length(diffCVec)
  for ( t1Idx in 1:(Nidx-1) ) {
```

Team Mean, Median, Mode
Rei Leenah Balachandran s3112637
Shubhankar Sanjay Jahagirdar s3793593
Tanmay Madan Shendkar s3735580

```
    for ( t2Idx in (t1Idx+1):Nidx ) {
      parName1 = paste0("beta[",diffCVec[t1Idx],"]")
      parName2 = paste0("beta[",diffCVec[t2Idx],"]")
      summaryInfo = rbind( summaryInfo ,
                      summarizePost( mcmcMat[,parName1]-mcmcMat[,parName2] ,
                                compVal=compValDiff , ROPE=ropeDiff ) )
      rowIdx = rowIdx+1
      rownames(summaryInfo)[rowIdx] = paste0(parName1,"-",parName2)
    }
   }
  }
  # save:
  if ( !is.null(saveName) ) {
   write.csv( summaryInfo , file=paste(saveName,"SummaryInfo.csv",sep="") )
  }
  show( summaryInfo )
  return( summaryInfo )
}

#==============================================================================

plotMCMC_HD = function( codaSamples , data , xName="x" , yName="y", preds = FALSE ,
                showCurve=FALSE ,  pairsPlot=FALSE , compVal = NULL,
                saveName=NULL , saveType="jpg" ) {
  # 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$|^zbeta\\[",colnames(mcmcMat))]
  # if ( ncol(x)==1 ) { zbeta = matrix( zbeta , ncol=1 ) }
  beta0 = mcmcMat[,"beta0"]
  beta  = mcmcMat[,grep("^beta$|^beta\\[",colnames(mcmcMat))]
  if ( ncol(x)==1 ) { beta = matrix( beta , ncol=1 ) }
  if (preds){
   pred = mcmcMat[,grep("^pred$|^pred\\[",colnames(mcmcMat))]
  }
  guess = mcmcMat[,"guess"]
  #-----------------------------------------------------------------------------
  # Compute R^2 for credible parameters:
  YcorX = cor( y , x ) # correlation of y with each x predictor
  Rsq = beta %*% matrix( YcorX , ncol=1 )
  #-----------------------------------------------------------------------------
  if ( pairsPlot ) {
   # Plot the parameters pairwise, to see correlations:
   openGraph()
   nPtToPlot = 1000
```

Team Mean, Median, Mode
Rei Leenah Balachandran s3112637
Shubhankar Sanjay Jahagirdar s3793593
Tanmay Madan Shendkar s3735580

```
  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
  }
  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" )
  if ( !is.null(saveName) ) {
   saveGraph( file=paste(saveName,"PostPairs",sep=""), type=saveType)
  }
 }
 #-------------------------------------------------------------------------
 # Marginal histograms:

 decideOpenGraph = function( panelCount , saveName , finished=FALSE ,
                     nRow=4 , nCol= 4 ) {
  # 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:
```

Team Mean, Median, Mode
Rei Leenah Balachandran s3112637
Shubhankar Sanjay Jahagirdar s3793593
Tanmay Madan Shendkar s3735580

```
  panelCount = 1
  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"] ))
  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") )
  histInfo = plotPost( Rsq , cex.lab = 1.75 , showCurve=showCurve ,
                  xlab=bquote(R^2) , main=paste("Prop Var Accntd") , finished=FALSE )

  panelCount = decideOpenGraph( panelCount , saveName=paste0(saveName,"PostMarg") )
  histInfo = plotPost( guess , cex.lab = 1.75 , showCurve=showCurve ,
                  xlab="Guess parameter" , main=paste("Prop Var Accntd") , finished=TRUE )

  panelCount = 1
  if ( pred){

    for ( pIdx in 1:ncol(pred) ) {
      panelCount = decideOpenGraph( panelCount , saveName=paste0(saveName,"PostMarg") )
      histInfo = plotPost( pred[,pIdx] , cex.lab = 1.75 , showCurve=showCurve ,
                  xlab=bquote(pred[.(pIdx)]) , main=paste0("Prediction ",pIdx) )
    }
  }
  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") )

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

#===============PRELIMINARY FUNCTIONS FOR POSTERIOR INFERENCES====================

drug <- read.csv("DrugConsumptionFinal.csv")
head(drug)
drug$Gender <- as.numeric(as.factor(drug$Gender)) - 1 # To get 0/1 instead of 1/2; Female = 0; Male
= 1
head(drug)
attach(drug)
smp_siz = floor(0.80*nrow(drug))
set.seed(123)
train_sample = sample(seq_len(nrow(drug)),size = smp_siz)
```

Team Mean, Median, Mode
Rei Leenah Balachandran s3112637
Shubhankar Sanjay Jahagirdar s3793593
Tanmay Madan Shendkar s3735580

```
trainData = drug[train_sample,]
testData = drug[-train_sample,]

# THE DATA.
y = trainData[,"Legalh"]
x = as.matrix(trainData[,2:14])

PredData = testData
head(PredData)

summary(PredData)

xPred = as.matrix(PredData[-1])

Nx = ncol(x)

# Specify the data in a list, for later shipment to JAGS:
dataList <- list(
  x = x ,
  y = y ,
  xPred = xPred ,
  Ntotal = length(y),
  Nx = Nx,
  Npred = nrow(xPred)
)

# First run without initials!
initsList <- list(
  beta0 = 0,
  beta = c(2.4, 1.5, 0.82, -0.25, -0.16, 1.3, 0, -0.26, 0.69, 0, -0.24, 0, 0.6)
)

modelString = "
model {
  for ( i in 1:Ntotal ) {
    # In JAGS, ilogit is logistic:
    y[i] ~ dbern( mu[i] )
    mu[i] <- ( guess*(1/2) + (1.0-guess)*ilogit(beta0+sum(beta[1:Nx]*x[i,1:Nx])) )
  }
  # Priors vague on standardized scale:
  beta0 ~ dnorm( 0 , 1/2^2 )
  # non-informative run
   for ( j in 1:Nx ) {
     beta[j] ~ dnorm( 0 , 1/2^2 )
   }

  guess ~ dbeta(1,9)

  for ( k in 1:Npred){
    pred[k] <- ilogit(beta0 + sum(beta[1:Nx] * xPred[k,1:Nx]))
  }
}
```

```
Team Mean, Median, Mode
Rei Leenah Balachandran s3112637
Shubhankar Sanjay Jahagirdar s3793593
Tanmay Madan Shendkar s3735580

"
writeLines( modelString , con="TEMPmodel.txt" )

parameters = c( "beta0")
for ( i in 1:Nx){
  parameters = c(parameters, paste0("beta[",i,"]"))
}
for ( i in 1:nrow(xPred)){
  parameters = c(parameters, paste0("pred[",i,"]"))
}

parameters = c(parameters, "guess")

adaptSteps = 1500  # Number of steps to "tune" the samplers
burnInSteps = 2000
nChains = 3
thinSteps = 200
numSavedSteps = 4000
nIter = ceiling( ( numSavedSteps * thinSteps ) / nChains )

startTime = proc.time()
runJagsOut <- run.jags( method="parallel" ,
                    model="TEMPmodel.txt" ,
                    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()
show(stopTime - startTime)

beep()

#======================= MCMC on only significant variables ====================

# First run without initials!
initsList <- list(
  beta0 = 0,
  beta = c(2.4, 1.5, 0.82, -0.25, -0.16, 1.3, -0.26, 0.69, -0.24, 0.6)
)

modelString = "
model {
  for ( i in 1:Ntotal ) {
    # In JAGS, ilogit is logistic:
    y[i] ~ dbern( mu[i] )
```

Team Mean, Median, Mode
Rei Leenah Balachandran s3112637
Shubhankar Sanjay Jahagirdar s3793593
Tanmay Madan Shendkar s3735580

```
    mu[i] <- ( guess*(1/2) + (1.0-guess)*ilogit(beta0 + beta[1] * x[i,1] + beta[2] * x[i,2] + beta[3] * x[i,3]
+ beta[4] * x[i,4] + beta[5] * x[i,5] + beta[6] * x[i,6] + beta[7] * x[i,8] + beta[8] * x[i,9] + beta[9] * x[i,11]
+ beta[10] * x[i,13]))
  }
  # Priors vague on standardized scale:
  beta0 ~ dnorm( 0 , 1/2^2 )
  # non-informative run
   for ( j in 1:10 ) {
     beta[j] ~ dnorm( 0 , 1/2^2 )
   }

  guess ~ dbeta(1,9)

  for ( k in 1:Npred){
    pred[k] <- ilogit(beta0 + beta[1] * xPred[k,1] + beta[2] * xPred[k,2] + beta[3] * xPred[k,3] + beta[4]
* xPred[k,4] + beta[5] * xPred[k,5] + beta[6] * xPred[k,6] + beta[7] * xPred[k,8] + beta[8] * xPred[k,9]
+ beta[9] * xPred[k,11] + beta[10] * xPred[k,13])
  }
}
"
writeLines( modelString , con="TEMPmodel.txt" )

parameters = c( "beta0")
for ( i in 1:10){
  parameters = c(parameters, paste0("beta[",i,"]"))
}
for ( i in 1:nrow(xPred)){
  parameters = c(parameters, paste0("pred[",i,"]"))
}

parameters = c(parameters, "guess")

adaptSteps = 1500  # Number of steps to "tune" the samplers
burnInSteps = 2000
nChains = 3
thinSteps = 200
numSavedSteps = 4000
nIter = ceiling( ( numSavedSteps * thinSteps ) / nChains )

startTime = proc.time()
runJagsOut <- run.jags( method="parallel" ,
                    model="TEMPmodel.txt" ,
                    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 )
```

Team Mean, Median, Mode
Rei Leenah Balachandran s3112637
Shubhankar Sanjay Jahagirdar s3793593
Tanmay Madan Shendkar s3735580

```
stopTime = proc.time()
show(stopTime - startTime)

beep()
graphics.off()

load(file = "rEnvironment_80_20_1500_2000_3_200_4000.RData")
load(file = "rEnvironment_lessVariables_80_20_1500_2000_3_200_4000.RData")

diagMCMC( codaSamples , parName="beta0" )
for ( i in 1:10){
  diagMCMC( codaSamples , parName=paste0("beta[",i,"]") )
}

#=================Analysis of Posterior distributions ===========================

compValfull <- data.frame("beta0" = -2.6, "beta[1]" = 2.6, "beta[2]" = 1.32, "beta[3]" = 0.46, "beta[4]" = -0.83, "beta[5]" =  -0.97, "beta[6]" =  1.26, "beta[7]" = 0, "beta[8]" = -0.41, "beta[9]" = 0.73, "beta[10]" = 0, "beta[11]" = -0.166, "beta[12]" = 0, "beta[13]" = 0.78, check.names=FALSE)
#compVal for model 1

compVal <- data.frame("beta0" = -2.6, "beta[1]" = 2.6, "beta[2]" = 1.32, "beta[3]" = 0.46, "beta[4]" = -0.83, "beta[5]" =  -0.97, "beta[6]" =  1.26, "beta[7]" = -0.41, "beta[8]" = 0.73, "beta[9]" = -0.166, "beta[10]" = 0.78, check.names=FALSE)
#compVal for model 2.

summaryInfo <- smryMCMC_HD(codaSamples, compVal= compVal , diffCVec=c(1,2,3,4,5), compValDiff=0.0)
print(summaryInfo)

#plotMCMC_HD for full variables
plotMCMC_HD( codaSamples = codaSamples , data = trainData, xName = c( "X18.24", "X25.34", "X35.44", "X45.54", "X55.64", "Gender", "Nscore", "Escore", "Oscore", "Ascore", "Cscore", "Impulsiveness", "Sensation_seeking"),
 yName="Legalh", compVal = compValfull, preds = FALSE)

#plotMCMC_HD for only significant variables
plotMCMC_HD( codaSamples = codaSamples , data = trainData, xName = c("X18.24", "X25.34", "X35.44", "X45.54", a"X55.64", "Gender","Escore","Oscore", "Cscore", "Sensation_seeking"), yName="Legalh", compVal = compVal, preds = FALSE)

# Predictions for full records in training set

preds <- data.frame(subjectID = PredData[,1], PredProb = summaryInfo[13:389,3], drug_consumption = PredData[,1])

#summaryInfo[15:391,3] for all variables
#summaryInfo[13:389,3] for only significant variables

dim(PredData)
threshold <- 0.4#
preds[which(preds[,2]<threshold),3] <- 0
```

Team Mean, Median, Mode
Rei Leenah Balachandran s3112637
Shubhankar Sanjay Jahagirdar s3793593
Tanmay Madan Shendkar s3735580

```
preds[which(preds[,2]>threshold),3] <- 1

predsSorted <- preds[order(preds$subjectID),]
table(preds$drug_consumption)

predict_subjects <- myData1[which(myData1$X %in% predsSorted$subjectID),15]

# ============ Predictive check ============

confusionMatrix <- function(resp, pred){
  classRes <- data.frame(response = resp , predicted = pred)
  conf = xtabs(~ predicted + response, data = classRes)

  accuracy = sum(diag(conf))/sum(conf)
  accuracy
  precision = conf[2,2]/(conf[2,1]+conf[2,2])
  precision
  recall = conf[2,2]/(conf[1,2]+conf[2,2])
  recall
  Fscore = 2*((precision*recall)/(precision+recall))
  Fscore
  return(list(accuracy = accuracy, precision = precision, recall = recall, Fscore = Fscore, conf = conf))
}

confusionMatrix(resp = predict_subjects, pred = predsSorted[,3])
```