Non-Parametric Analysis of NHANES Health Dataset

Scott Phillips

Abstract
This report utilises data provided by the NHANES[2] package in R to answer the overarching question: what variables affect mental health and can this be effectively modelled. After an initial exploratory analysis of the NHANES dataset, the data is split into two subsets, male and female, due to a lack of statistically significant variables for a regression analysis in the combined dataset. Initially a linear model is calculated with a predictor variable corresponding to bad mental health. Analysis of this exploratory run will then lead to the application of scaling or transformation methods to the variables to aid in the process of finding the best linear model that will yield statistically significant results. The relevance of each variable is investigated via a series of bootstrapping methods: Non-Parametric Bootstrap(NPB), Semi-Parametric Residual Bootstrap(SRB) and the Bayesian Bootstrap(BB). After confirming which variables are suitable for a regression analysis, potential clustering is explored within the data using DPM models. Applying this to a regression analysis and producing a variety of results is critical for practical inference of the relationships of the variables. Finally, applying a regression model to the data using similar DPM models based on all the accrued information and producing results will answer the initial question and provide an accurate model for future predictions.

1 Introduction

1.1 Preliminary Data Description

To begin the discussion, it is necessary to address some potential issues within the NHANES dataset. Typically, the NHANES database applies weighted resampling to the data to accurately reflect the nature of the population. This analysis will be working with a resampled subset of the data so it will not require resampling. Additionally, the data is also not to be used for research purposes.

Now, evaluating the dataset itself, beginning with a description of the variables of the NHANES dataset the analysis will utilise: Gender, Age, Poverty, BMI, Pulse, Testosterone, DaysMentHlthBad, DaysPhysHlthBad, SleepHrsNight, Weight, AlcoholYear, SexNumPartnLife, TotChol, BPSysAve, BPDiaAve. These variables are the cornerstones of health in modern society and will aid in determining what affects mental health. All of these, bar age, are numerical, which will enhance the insightfulness of the statistical analysis. Unfortunately, some of the variables have undesirable properties and subjective recording methods, which need to be addressed. DaysMentHlthBad is the recorded number of days in the last 30 where an individual has felt "Bad", this is of course going to be interpreted in different ways. Fortunately, by the principle of averaging[1], the effects of subjectivity decrease for increasing sample size. SleepHrsNight presents other difficult properties to deal with, represent hours of sleep per night and ranges from 2 to 12 hours, the discrete nature and limited range of values could make the analysis challenging. Transformations of these variables may be performed if the nature of the variable is lost in the analysis. Additionally, the recorded values of poverty are important to inspect, smaller values indicate higher levels of poverty. Ideally, a much larger dataset would be utilised

to overcome the subjectivity and large amount of NULL results before cleaning the data.

1.2 Preliminary Data Analysis

Continuing the discussion, it is essential to investigate the predictor variable in detail. Figure 1 displays a histogram of the predictor, it is clear to see that the predictors density is left skewed. Applying a log transformation will reduce the effects of outliers, compress the y-axis and subsequently enhance model estimation. This results in the log transformed histogram where the overlaid density is significantly less skewed and will yield much more effective results.

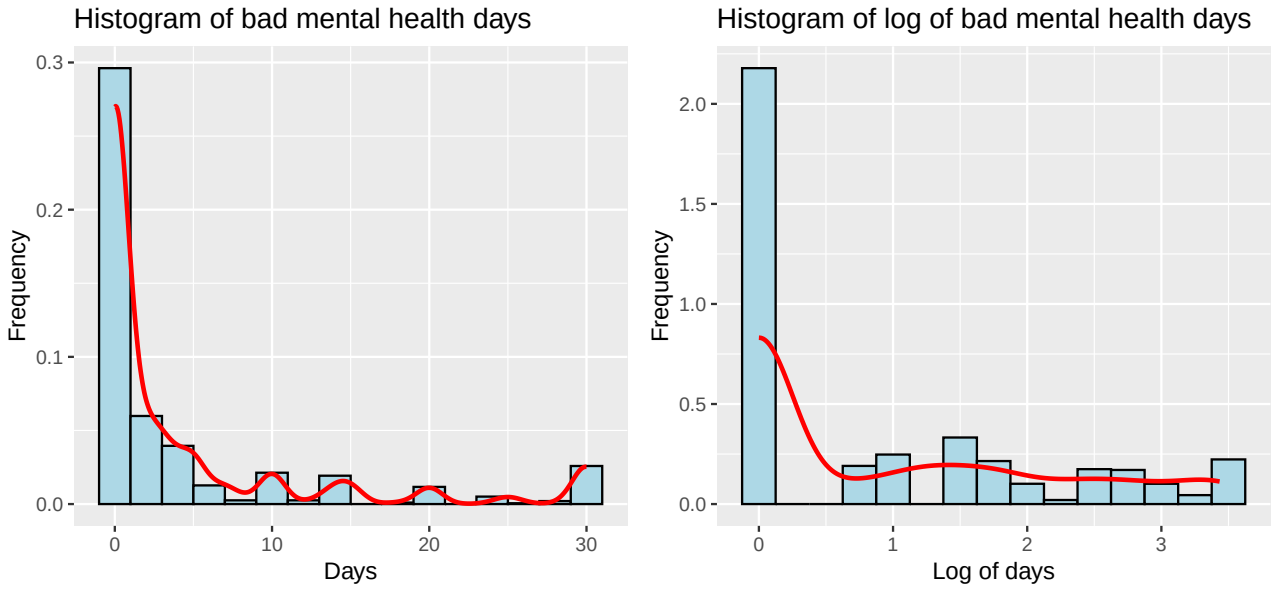

Figure 1: Histograms of Predictor

Figure 2 is a table of summary statistics for bootstrapped samples that will be utilised in the analysis for the DPM's prior information. Note the variables have not been altered, this is as interpretability is essential for answering the overarching question, implying the use of raw data for later cluster and regression analysis. Clearly, for a linear model, scaling is beneficial due to the order of magnitude difference in some of the variables.

Variable	Mean_of_Means	Variance_of_Means	Mean_of_Variances	Variance_of_Variances	Gender
DaysMentHlthBad	4.769759	0.141722873	56.644375	4.448069e+01	Female
SleepHrsNight	6.885404	0.004504815	1.765854	1.887741e−02	Female
Poverty	3.009293	0.007222506	2.810629	1.029658e−02	Female
Pulse	74.774223	0.320171771	123.535557	7.673448e+01	Female
SexNumPartnLife	14.415350	3.200528051	1292.776933	7.837623e+05	Female
DaysMentHlthBad	4.213438	0.120173478	69.812631	4.781403e+01	Male
DaysPhysHlthBad	3.587449	0.097883066	57.156085	4.311437e+01	Male
SleepHrsNight	6.493083	0.002715608	1.599258	9.054470e−03	Male
Testosterone	419.085862	59.965523683	34563.690406	8.018027e+06	Male

Figure 2: Table of Values

2 Analysis of Male Subset

2.1 Multivariate Linear Regression

As described previously, a linear model is applied to the dataset with the log of BadMentHlth-Days as the predictor variable. Before proceeding with any subsequent analysis of the data, the hypothesis test for the test statistics must be stated:

$$H_0 : \beta_i = 0$$
$$H_1 : \beta_i \neq 0$$

Where β_i represents the regression coefficient of the i-th parameter, evaluating this test at the 99th percent level. Applying scaling to the response variables will enhance interpretability of the resulting confidence intervals for the hypothesis test. This is due to the orders of magnitude several of the variables operate under. Calculating the resulting linear model with this criteria leads on to the next step in finding a best fit linear model.

To ensure that the linear model accurately resembles the data, a test for multicollinearity must be carried out, upon doing this, the correlation between the variables BMI and Weight is seen to be approximately 0.9. Constructing a new linear model without the response variable BMI results in finding the linear model that best fits the scaled data and is representative of what factors affect mental health. Upon closer investigation, the statistically significant variables are found to be SleepHrsNight, BadPhysHlthDays and Testosterone, as seen in Figure 3.

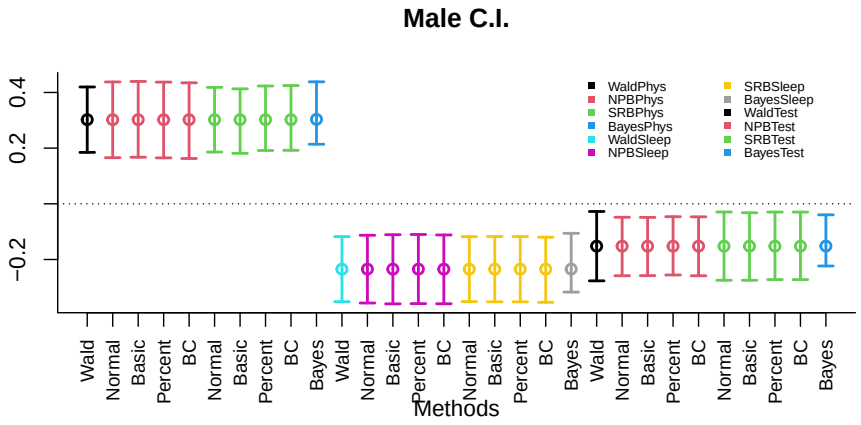

Figure 3:

Figure 3 displays confidence intervals for the relevant hypothesis test via the Wald, NPB, SRB and BB methods. Inside the NPB and SRB methodology, normal, percentile, bias-corrected, and Bootstrap-t confidence intervals are constructed. The benefit of using all these methods is it showcases the various strengths of each method. The resulting confidence intervals in Figure 3 show clear similarities in the range of values the confidence intervals occupy. This is as a result of the normal and percentile placing heavy assumptions on normality and introducing bias. Fortunately, in scaled data this is not much of an issue. This is demonstrated in the corrected equivalent of these intervals, the t and bias corrected confidence intervals. These show a very similar range of values to the uncorrected intervals, loosening the assumptions and making corrections for bias introduced. This equates to the variables being approximately normal and with limited amounts of bias. More importantly, giving the statistically significant variables to apply to the next stage of the analysis.

2.2 Cluster Investigation

For the cluster investigation, the following model will be used with prior parameters utilised from Figure 2.

$$
\begin{aligned}
y_i \mid \mu_i, \Sigma_i &\overset{\text{ind}}{\sim} N_q(\mu_i, \Sigma_i), \\
(\mu_i, \Sigma_i) \mid G &\overset{\text{i.i.d.}}{\sim} G, \\
G &\sim \text{DP}(\alpha, G_0(\mu, \Sigma)), \\
G_0(\mu, \Sigma) &\equiv G_0(\mu \mid \Sigma) G_0(\Sigma), \\
G_0(\mu \mid \Sigma) &\sim N_q(\mu_0, k_0^{-1}\Sigma), \\
G_0(\Sigma) &\sim \text{IW}(\Sigma_0, \nu_0).
\end{aligned}
\qquad
\begin{aligned}
\mu_0 &= (4.213, 3.587, 6.493, 419.1) \\
k_0 &= 0.01 \\
\Sigma_0 &= \text{diag}(69.8, 57.2, 1.60, 1000)
\end{aligned}
$$

Large computation issues were prevalent within the cluster algorithm, including cluster degeneration and non-valid matrices in the mcmc simulations. The issues required a change in sampling method from ICS to SLI for the mcmc sampler. This is due to the SLI sampling method being more efficient at exploring high dimensional parameter space than ICS and it lacks a rejection step, ensuring that there are far less issues with the sampler. Within the priors itself, k_0 was set to be diffuse to allow for more exploration of the clusters around the prior means stated and Σ_0 to be diagonal so there is no correlation between variables for simplicity.

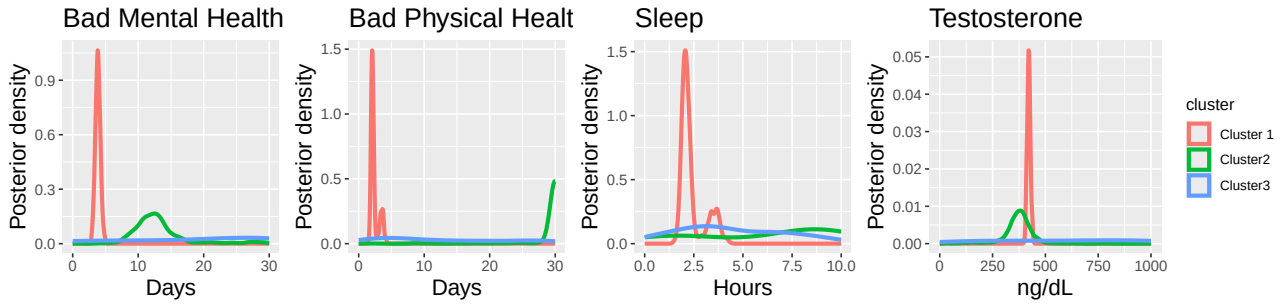

Figure 4: Cluster Densities of Male Variables

Figure 4 displays the resulting densities for the dominant clusters in the male dataset. The relationship between the clusters and the range of values they occupy within their variables domain is immediately noticeable. However, there are issues with cluster production, namely that the densities for Bad Physical Health and Sleep are identical. Upon closer examination, it is clear to see that the way the clusters interact over the range of values they occupy reflects what is expected for how these variables affect mental health. Specifically, based on the clusters, mental health gets worse as you become more ill, get less sleep and have lower testosterone levels. The knowledge gained from this cluster analysis can now be applied to the prior information of the regression model.

2.3 DPM Regression Model

The final stage of analysis of the male dataset is to assign the regression model described below to the data. In this model, the value of some parameters were adjusted to incorporate non-informative priors, such as in a_0, b_0 to keep the regression models results similar to the previous cluster analysis.

$$
\begin{aligned}
y_i \mid x_i, \beta_i, \sigma_i^2 &\sim N(x_i^\top \beta_i, \sigma_i^2), \\
(\beta_i, \sigma_i^2) \mid G &\sim G, \\
G &\sim DP(\alpha, G_0(\beta, \sigma^2)), \\
G_0(\beta, \sigma^2) &= G_0(\beta) G_0(\sigma^2), \\
G_0(\beta) &\sim N_p(m_0, S_0), \\
G_0(\sigma^2) &\sim \text{InvGamma}(a_0, b_0).
\end{aligned}
\qquad
\begin{aligned}
m_0 &= \text{c}(4.213, 3.587, 6.493, 419.1), \\
a_0 &= \text{c}(2), \\
b_0 &= \text{c}(1.5), \\
S_0 &= \text{diag}(15, 10, 1, 100).
\end{aligned}
$$

As shown in Figure 4, the probability that an individual is placed in cluster 3 is extremely unlikely for the range of values the variables occupy. This suggests operating under the assumption of two clusters would be beneficial and reduce computational issues due to the decrease in complexity. In Figure 5 the output of the regression model is evaluated at quantiles of interest for the response variables, namely the median of DaysPhysHlthBad and all subsequent quantile values are clearly shown. There is a clear relationship between testosterone levels, hours of sleep, and mental health which aligns with the patterns formed in the cluster model. The ultimate goal of the analysis is to model mental health based on these variables. Therefore, posterior densities for the regression coefficients are produced. It is immediately clear why scaling was used when approximating a linear model, as linear modelling methods cannot handle when values are orders of magnitude different in scale. This concludes the discussion of the male dataset.

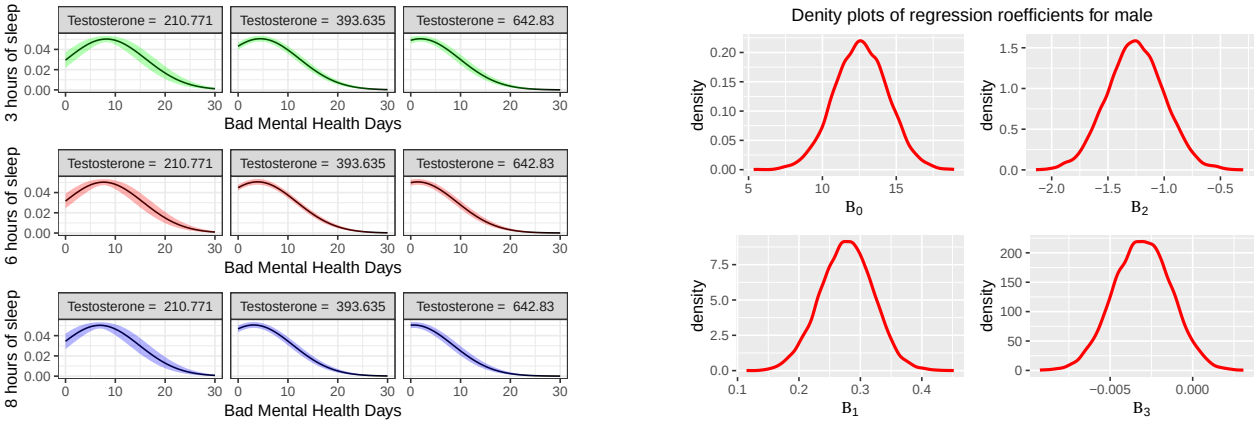

Figure 5: Plots of Regression densities and posterior regression coefficient densities

3 Analysis of Female Dataset

3.1 Multivariate Linear Regression

The analysis of the female dataset begins with a multivariate linear analysis, utilising the same response and predictors as in the male dataset. Including the corresponding log transformation and scaling of the response variables. After testing for multicollinearity and deleting all highly correlated variables, the resulting linear model contains the variables SexNumPartnLife, SleepHrsNight, Pulse and Poverty which reject the null hypothesis of the following hypothesis test at the 99 percent level.

$$H_0 : \beta_i = 0$$
$$H_1 : \beta_i \neq 0$$

Evaluating the hypothesis test as before and evaluating the significance of the linear model's variable via NPB, SRB and BB methods for which the following confidence intervals are calculated: Wald, normal, percentile, bias-corrected and t-bootstrap. These are displayed in Figure 6

It is apparent from Figure 6 that all variables are statistically significant, however, the variable NumSexPartnLife has both the normal and percentile intervals failing to reject the null hypothesis. Fortunately the corrected version of these reject the null hypothesis. This highlights how the nature of the data can influence the confidence intervals and how important it

is to take this into account and apply suitable methods. As the corrected intervals suggests, SexNumPartnLife is still statistically significant, the analysis may therefore proceed with all afore mentioned variables.

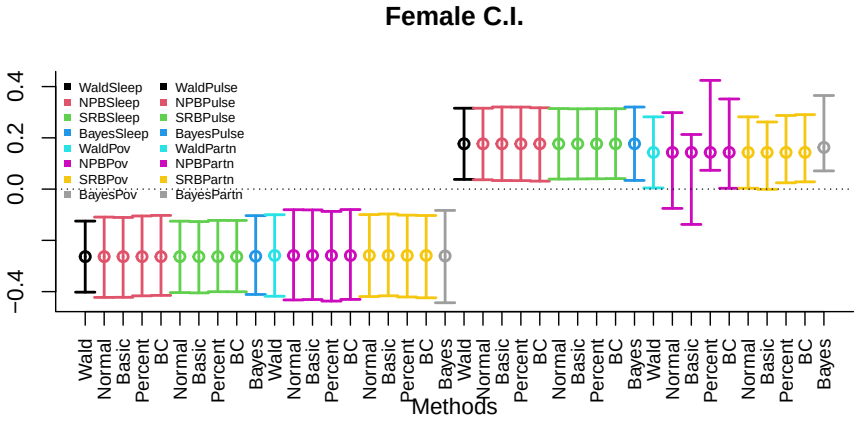

Figure 6:

3.2 Cluster Investigation

The clustering algorithm used is identical to that employed in the male dataset, but with the following parameters values taken from Figure 2: $\mu = (4.770, 6.885, 3.001, 74.78, 14.42), k_0 = 0.01, \Sigma_0 = diag(56.64, 1.766, 2.811, 123.5, 1293)$

The coefficients of the model are adjusted to reflect the diffuse prior beliefs about the female dataset. This allows the model to efficiently explore the entire probability space for any clustering. The model is the same DPM applied in the male dataset with the stated parameter adjustments. This assignment is chosen over a HDPM due to the lack of grouping in the variables. This simulation suffered from far fewer computational challenges, partly due to the more continuous nature of some of the variables.

Figure 7 shows the formation of the cluster densities. These clusters demonstrate very distinct patterns, namely that the first and second clusters illustrate the detrimental effects of poverty, high pulse rate, low levels of sleep and high number of sexual partners on mental health. The third cluster takes some careful reasoning to arrive at the conclusion that this cluster represents mentally ill women in the dataset. This is due to the high poverty index, very high levels of sleep, high number of sexual partners that contrast with what you would expect in society for someone living with these variables. Unless other excluded variables shed light on this particular cluster, this seems a fair and logical conclusion to the clustering shown.

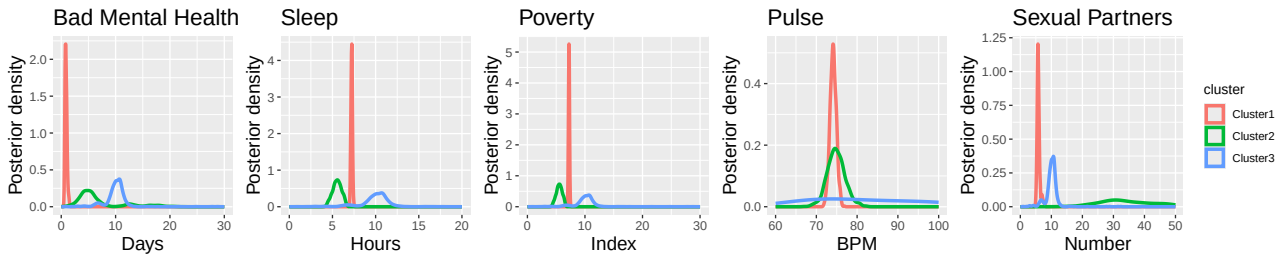

Figure 7: Cluster Densities of Female Variables

The clustering here is very insightful and allowed for some very interesting interpretations. Due to the absence of scaling, the direct relationship between variables are visible, effectively

modelling expected real-world interactions. With the information from the clustering, the analysis now proceeds to the regression model

3.3 DPM Regression Model

For the creation of a regression model, the parameters are set as
$m_0 = (4.770, 6.885, 3.001, 74.78, 14.42), S_0 = diag(10, 1, 1, 20, 50), a_0 = 2, b_0 = 1.5$

These parameters are chosen to ensure that the regression model closely aligns with the extremely intuitive clustering results. Equipping the DPM as used in the male dataset with the non- diffuse parameter specification's, should ensure a tighter fit and less exploration by the regression algorithm. Figure 8 shows the corresponding density plots for the quantiles of interest for the regression algorithm.

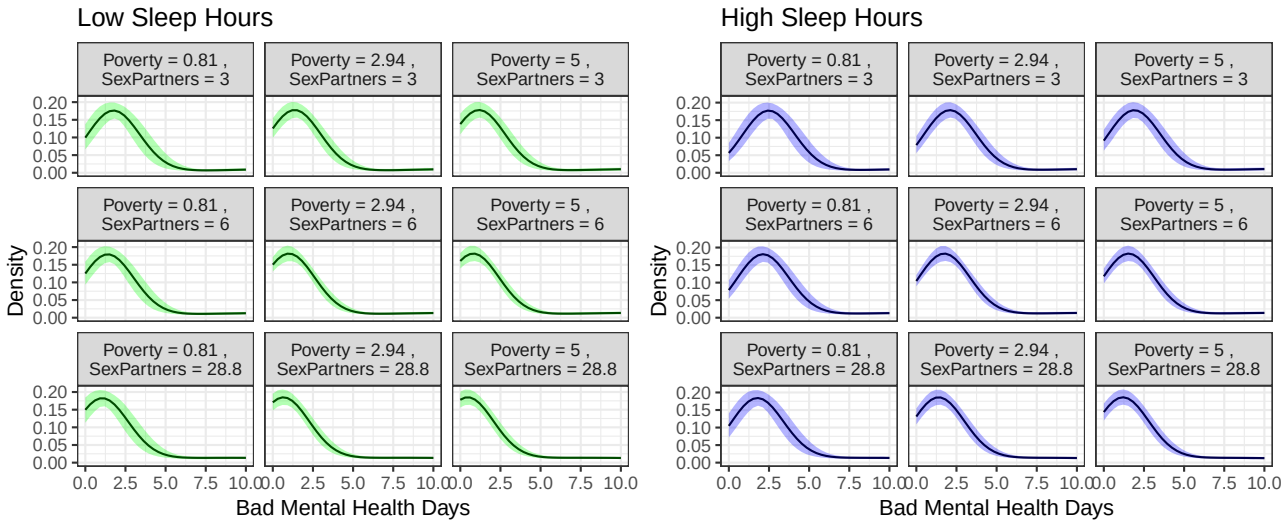

Figure 8: Regression Densities for Female Variables

The regression model clearly closely follows the same patterns illustrated by the clustering algorithm. The aim of the analysis is to produce a regression model. In Figure 9, the posterior densities for the regression coefficients are presented. Overall the results from this analysis have provided interesting and intuitive relationships for the models of the data.

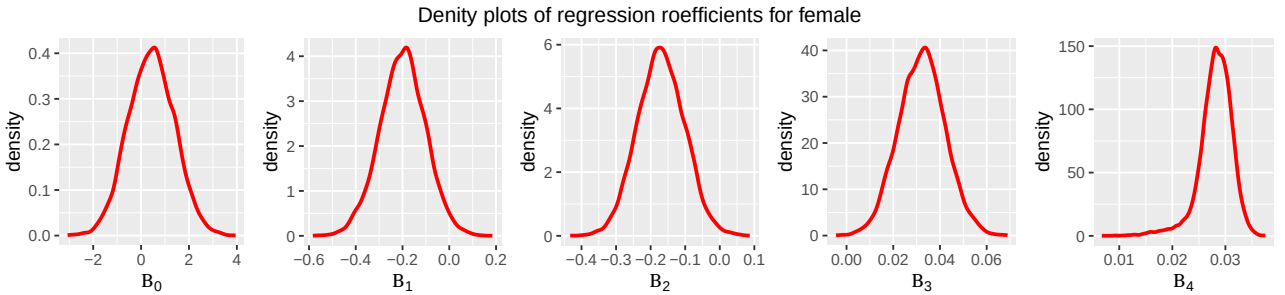

Figure 9:

4 Conclusions

The non-parametric analysis of both datasets has been completed, utilising a variety of methods to analyse the data whilst placing minimal assumptions on the structure of the data. A combination of bootstrap methods and Markov Chain Monte Carlo samplers were used to construct statistical tests and models to find statistically significant variables within the dataset. Identifying how they interact and ultimately developing regression and clustering models to classify the data points.

In the male dataset, the clustering and regression plots illustrate three distinct clusters. These clusters are indicative of factors that affect mental health due to the range of values the clusters occupy on the domain of the variables. This provides insight into individuals with the worst and best possible mental health. Thus, demonstrating clear correlation between good mental health, high levels of testosterone, good levels of sleep and good physical health. It is worth noting, unlike the female dataset, there are no meaningful clusters representing any signs of mental illness in the interviewed male group. The results of the regression analysis, whilst providing valid results, is corroborated by what individuals would interpret as factors that would affect mental health in the ways described by the analysis. However, it is not perfect as poverty, based off human intuition, would correlate with bad mental health but in this dataset this is not the case.

In the female dataset, the clustering and regression outputs provide insights into individuals with mental illness and those without. It is important to note this is based purely off the output of the data. There are undoubtedly many other interpretations for the apparent cluster, this however, is the most interesting and is closely linked with the theme of the analysis. As discussed earlier, the mentally ill cluster breaks from the norm of what has been shown in all the other clusters found in all datasets. This clearly indicates something else occurring that is not being accounted for in the analysis. For future analysis, it would be worth including a categorical variable that classifies mental illness and then apply a HDPM with parameter groupings based off the categories.

Given additional space, further analysis could involve applying a HPDM to the joint dataset using hierarchical priors based on the genders, and if possible, the presence of mental illness. Despite this limitation, the analysis has been fruitful, uncovering informative and applicable models for predicting and modelling mental health of individuals based off the variables stated. The most significant discovery, and the fundamental reason for splitting the dataset is the stark differences in variables that affect mental health in men and women. The reason behind the differences is open to interpretation, there are a plethora of additional factors that needed to be considered before attempting to answer this. However, our initial question has been thoroughly investigated and answered within the constraints of this report.

References

[1] Helen Noble and Joanna Smith. Issues of validity and reliability in qualitative research. *Evidence-Based Nursing*, 18(2):34–35, 2015.

[2] Randall Pruim. *NHANES: Data from the US National Health and Nutrition Examination Study*, 2015. R package version 2.1.0.