ISYE 6420: Bayesian Analysis of Sleep Data

Chitwan Kaudan
ckaudan3@gatech.edu

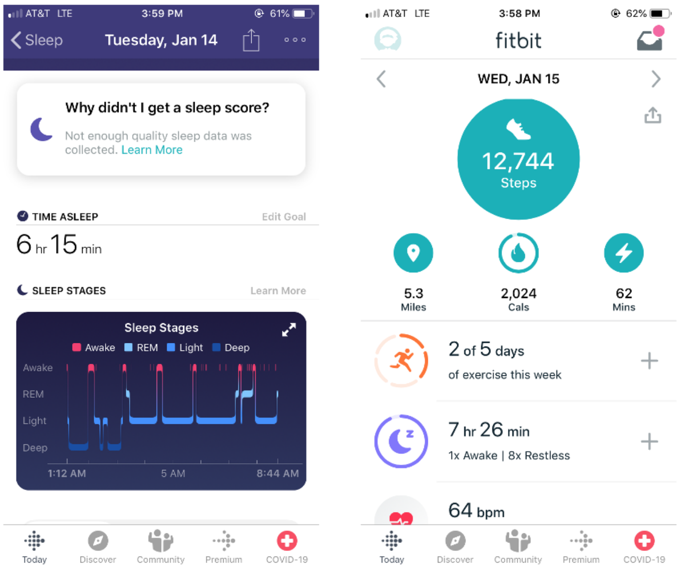
April 20, 2020

1 Problem Statement

In January 2020, I bought FitBit's Inspire HR fitness tracker to monitor my daily physical activity and sleep patterns. The Inspire HR has a 3-axis accelerator that can log active times, perform activity recognition and estimate calories burned. It also has an optical heart rate sensor that tracks sleep stages (i.e. light, deep, REM) data. I was hoping to analyze data the device collects to help build better sleep and exercise habits for myself.

FitBit has a mobile app that syncs with the tracker and provides dashboards for daily activities and sleep data as depicted in Figure 1.

Figure 1: FitBit App Dashboard


Why didn't I get a sleep score?
Not enough quality sleep data was collected. Learn More

These mobile dashboards are not very customizable and do not allow me to compare activities or sleep across time. They also do not provide any descriptive statistics or search for any causal relationships within my activities. So, I decided to use Bayesian methods we learned in class to analyze my sleeping habits in hopes of answering the following question:

1. Everyday, I aim to sleep at 1:00AM and wake by 9:00AM. Based on my historical data, what the probability of meeting these goals?
2. What factors affect sleep duration and sleep quality?

2 Data Collection

For this analysis, I collected data from the following 2 sources:

- **FitBit**
  FitBit provides monthly data exports of some activity, sleep and food logging data. I exported activity and sleep data from 2-1-2020 to 4-15-2020.

- **Canvas**
  I had a hypothesis that my weekly course workload affects my sleep patterns so I exported my Spring 2020 calendar from canvas.

I merged the datasets based on the ”Date” key and engineered features I thought might be affect my sleep patterns. My final merged dataset had 45 data points and the 24 columns noted in Table 1.

Table 1: All feature descriptions

Feature	**Description**	**Source**
Calories Burned	Number of total calories burned	FitBit
Steps	Number of steps taken	FitBit
Distance	Distance travelled	FitBit
Minutes Sedentary	Minutes of sedentary activity	FitBit
Minutes Lightly Active	Minutes of light activity	FitBit
Minutes Fairly Active	Minutes of moderate activity	FitBit
Minutes Very Active	Minutes of intense activity	FitBit
Activity Calories	Total calories burned during active minutes	FitBit
Minutes Very Active	Minutes of intense activity	FitBit
Start Time	Sleep timestamp	FitBit
End Time	Wake timestamp	FitBit
Minutes Asleep	Total minutes asleep during Start Time and End Time	FitBit
Minutes Awake	Total minutes awake during Start Time and End Time	FitBit
Number of Awakenings	Number of awakenings based on heart rate	FitBit
Time in Bed	Time in bed	FitBit
Minutes REM Sleep	Total minutes in REM	FitBit
Minutes Light Sleep	Total minutes of light sleep	FitBit
Minutes Deep Sleep	Total minutes of deep sleep	FitBit
Homework	Number of homework assignments due the next day	Canvas
Project	Number of project assignments due the next day	Canvas
Exam	Number of exams the next day	Canvas
Weekly Assign Num	Number homework, projects, or exam deadlines that week	Canvas
Is Weekend	Indicator for weekend	Engineered
Is Break	Indicator for spring break	Engineered

3 Bayesian Analysis

3.1 Sleep and Wake Time Analysis

To address question 1, I used Start Time and End Time from the merged dataset and created 2 additional indicator datasets: one for sleep time and another for wake time. For sleep time, I enumerated every minute between minimum Sleep Start (12:00AM) and maximum Sleep Start (4:00AM) and added an indicator variable called Asleep that was 0 when I was awake and 1 when I was sleeping. I followed an identical process for wake time using End Time which ranged from 6:00AM to 12:00PM.

I wanted to generate probability distributions of sleeping at a certain time from my indicator data. I decided to use a Bayesian logistic regression model with time as my explanatory variable and asleep indicator as my response. Instead of use the actual timestamp, I used time offset which was the number of minutes from minimum Sleep Start (12:00AM) for the sleep model and number of minutes from minimum Sleep End (6:00AM) for the wake model. I did not have any information to construct priors for my parameters, so I decided to use non-informative priors. My final sleep and wake models looked like the following.

$$\begin{aligned} y_i|\alpha,\beta &\sim Bernoulli(p(t_i)) \\ p(t_i) &= \frac{e^{\alpha+\beta*t_i}}{1+e^{\alpha+\beta*t_i}} \\ p(\alpha,\beta) &\propto 1 \end{aligned} \tag{1}$$

I used adaptMCMC (R code in Figure 9 in Appendix) to generate samples from the posteriors of α and β. The sleep time model had an acceptance rate of 34.9%, effective size of 1062 for α and 1045 for β and Figure 10 in the appendix contains the cumulative plots. Figure 2 and Figure 3 show the results of the MCMC sampling.

Figure 2: Sleep time logistic curve

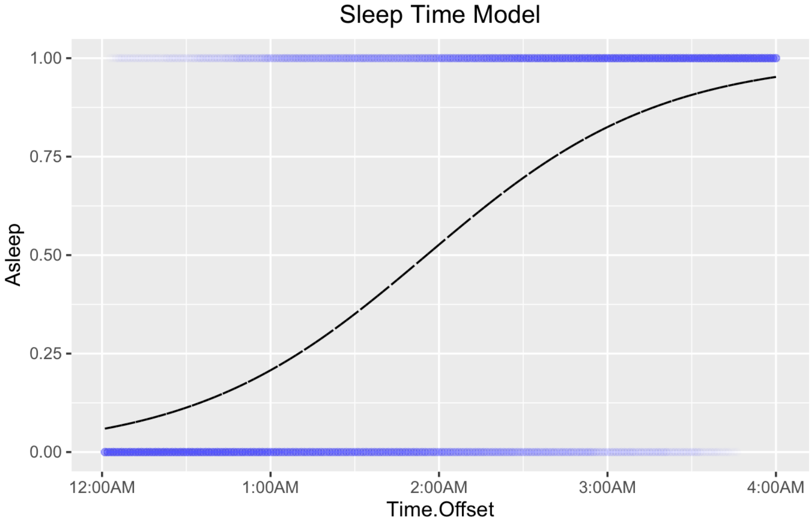
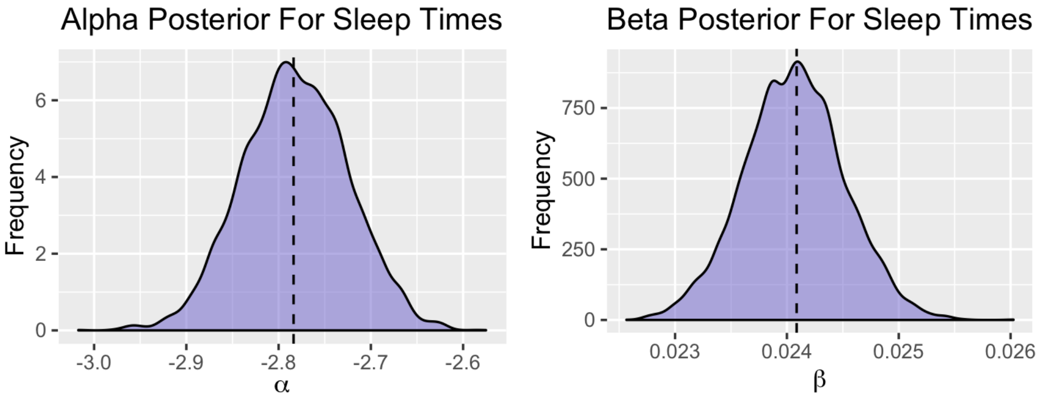


Figure 3: Sleep time parameter posterior densities



Similarly, the wake time model had an acceptance rate of 35.2%, effective size of 1166 for α and 1154 for β, and Figure 11 in the appendix contains the cumulative plots. Figure 4 and Figure 5 show the results of the MCMC sampling.

Figure 4: Wake time logistic curve

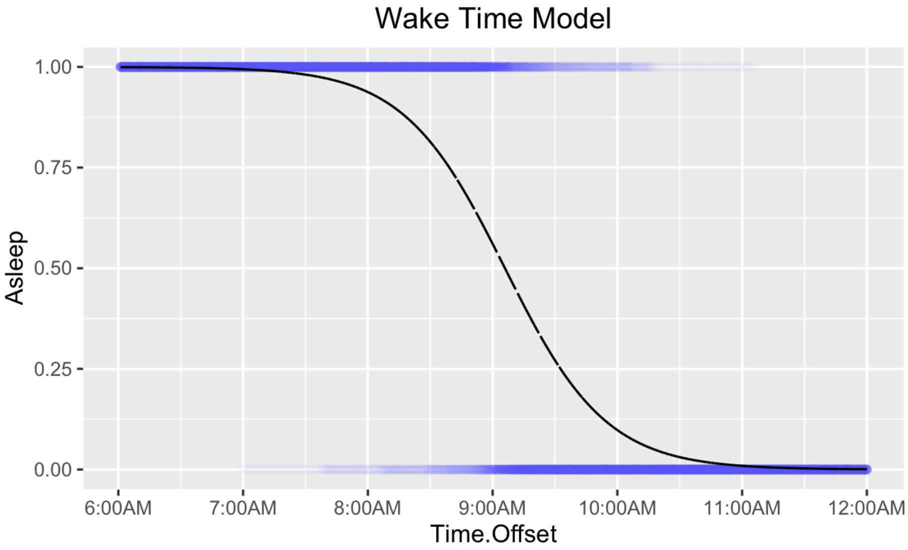
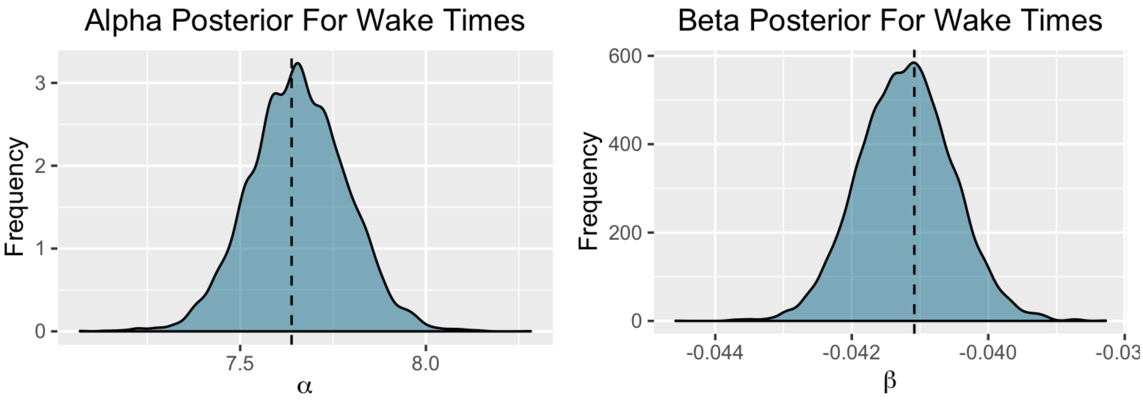
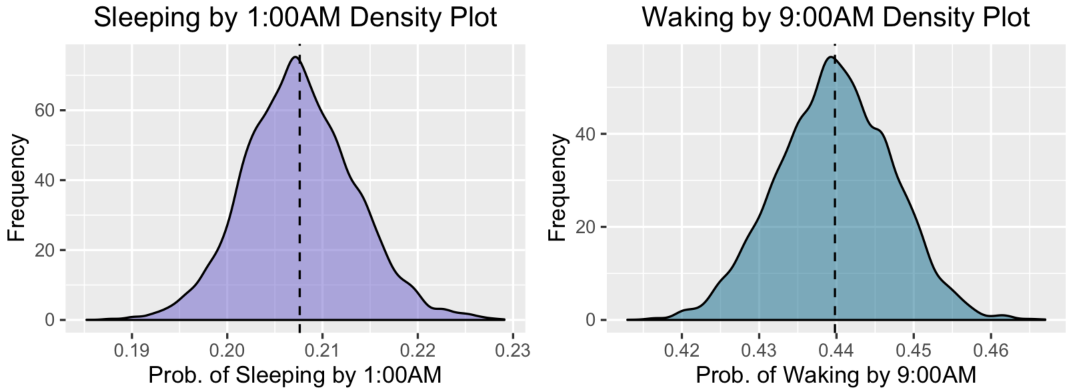


Figure 5: Wake time parameter posterior densities



Finally, Figure 6 shows the probability distribution for sleeping and waking on time.

Figure 6: Sleep by 1:00 AM and wake by 9:00 AM densities



3.2 Sleep Duration and Quality Analysis

To address question 2, I wanted to use Bayesian regression on the merged dataset to see how each feature affects sleep duration and quality. For sleep duration, I used Minutes Asleep as the response and all other non-sleep related features as predictors. For sleep quality, I used Minutes Asleep/Time In Bed as the response and all other non-sleep related features as predictors. Since I did not have any prior information, I used non-informative priors for my parameters.

$$\begin{aligned} y_i|\beta, \sigma^2 &\sim N(X\beta, \sigma^2) \\ p(\beta, \sigma^2) &\propto \frac{1}{\sigma^2} \end{aligned} \tag{2}$$

First, I checked for multicollinearity using a pairwise correlation plot as shown in Figure 12 in the appendix. I computed the Variance Inflation Factors (VIF) and dropped all features with VIF > 20. Next, I checked for heteroskedasticity using the Breusch-Pagan test. At a 0.1 significance level, the test did not reject the null hypothesis that the error variances were equal. Finally, I plotted the pairs plot to look for any obvious non-linearity. I found that both responses had a potentially quadratic relationship (Figure 13 in appendix) with Minutes Sedentary to I added Minutes Sedentary squared variable called Min Sed Sq.

I used the following conditional distributions for the regression parameters that we derived in class and used a MC sampling method to generate 10000 samples from the posterior distributions of the β parameters. In the notation below, p is the number of predictors in the model which in our case was 11.

$$\begin{aligned} \beta|\sigma^2, y &\sim N((X^TX)^{-1}X^Ty, \sigma^2(X^TX)^{-1}) \\ \sigma^2|\beta.y &\sim Inv-Gamma\Big(\frac{n-p-1}{2}, \frac{(y-X\hat{\beta})^T(y-X\hat{\beta})}{2}\Big) \end{aligned} \tag{3}$$

The complete code for the MC sampling is located in Figure 14 in the appendix. Figures 7 and 8 contains the posterior distributions for the sleep duration and sleep quality regression models respectively.

Figure 7: Posterior densities of the coefficients in the sleep duration model

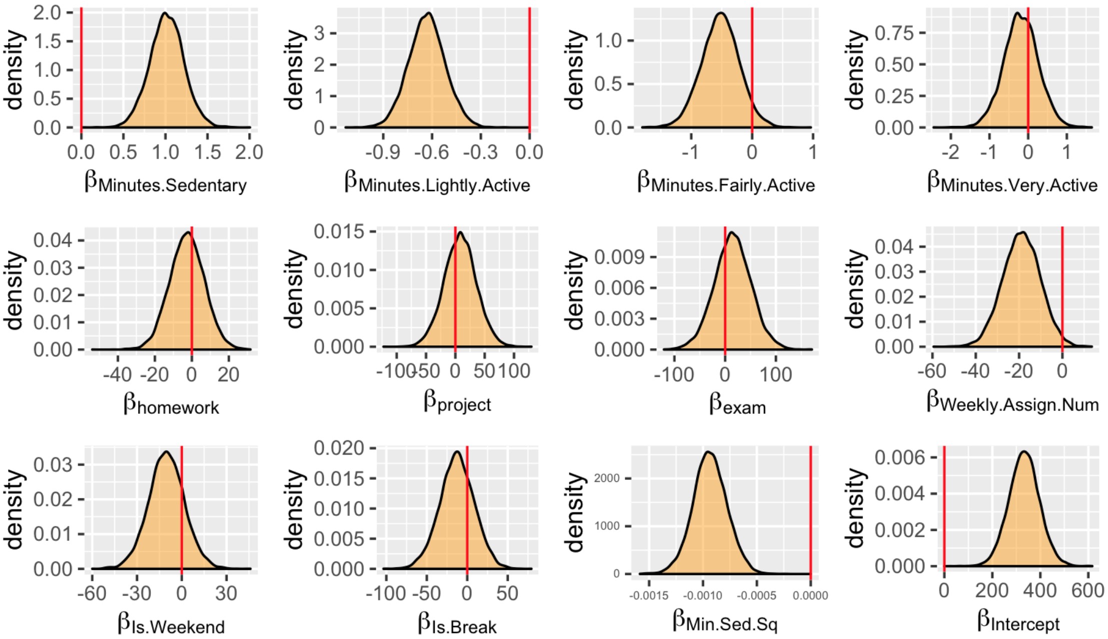
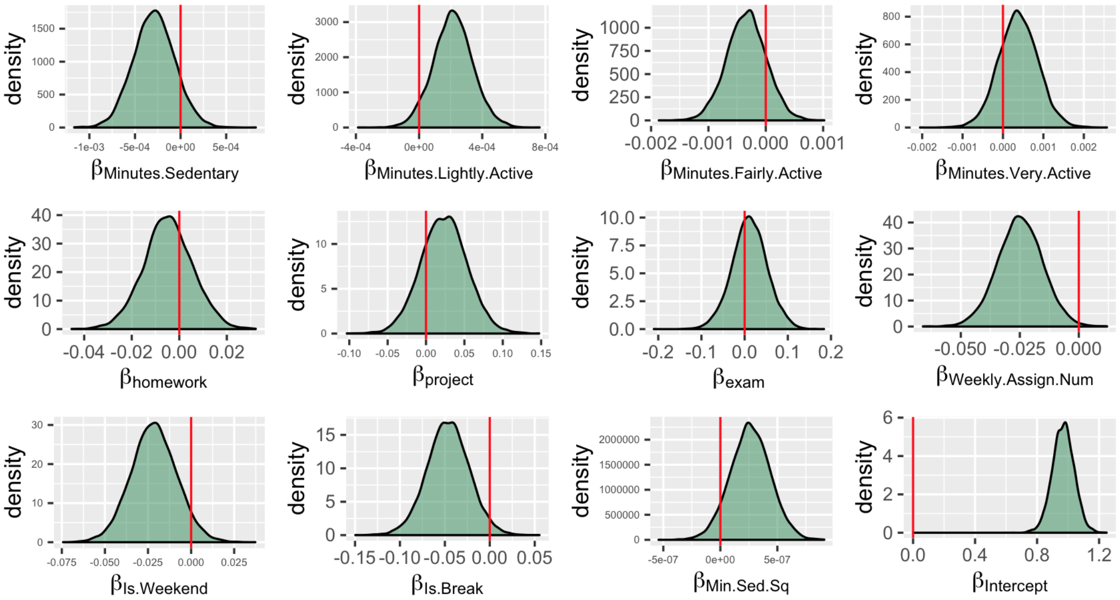


Figure 8: Posterior densities of the coefficients in the sleep quality model



4 Discussion

From Figure 6, it's evident that I'm much more likely to wake by 9:00AM (44%) than I am to sleep by 1:00AM (21%). In a 2 hour time frame from 8:00AM-10:00AM, I go from 6% chance of waking up to 90% chance of waking up. In a similar two hour time frame from

12:00AM to 2:00AM, I go from a 6% chance of sleeping to 53% of sleeping. It seems that I go a better job of waking up around 9:00AM than I do of sleeping at 1:00AM, which makes sense because I have an alarm to wake up to but no similar alarm to sleep.

For the sleep duration analysis, the features significant at a 0.05 significant level were Minutes Sedentary, Minutes Lightly Active, Weekly Number of Assignments, Minutes Sedentary squared, and the intercept. I'm not sure if the quadratic relationship between Minutes Sedentary and Minutes Asleep actually exists or if my model is just over-fitting. I would be curious to collect more datapoints and re-run the analysis to see if this relationship persists. Minutes Lightly Active has a significant negative coefficient, but I have a hunch that number of lectures per day is a confounding factor. Tuesdays are my busiest lecture days and I do a lot of walking from Scheller to ISYE, then back to Scheller. I also don't get time to do much work during the day, so perhaps that's why I tend to stay up later and get less sleep on Tuesday nights. Weekly Number of Assignments has a significant negative coefficient, but homework, project and exam do not. This suggests that during weeks where I have several deadlines I tend to sleep less, but not necessarily the night before the assignment is due.

For the sleep quality analysis, the features significant at a 0.05 significant level were Weekly Number of Assignments, Is Break, and the intercept. Weekly Number of Assignments has a significant negative coefficient which means I get restless sleep when I have impending deadlines. Is Break represents the weeks we had spring break and I was surprised to find it had a significant negative coefficient. I would have assumed I had better quality sleep during break, but that doesn't seem to be the case.

All in all, this analysis was very enlightening for me personally to better understand my sleeping habits. I now know I need to make a more conscious effort to sleep by 1:00AM. In the future, I would like to collect more data points and eventually build a model that is able to predict the time, the duration and quality of my sleep everyday.

5 Appendix

Figure 9: MH Algorithm for Sleep and Wake Time Analysis

```r
library(adaptMCMC)
library(numDeriv)
library(ggplot2)

### Sleep Times Model ###
### Load indicator data
sleep <- read.csv("data/sleepdf.csv")
y <- sleep$Asleep
t <- sleep$Time.Offset
# Define nlogh, logh, h
nlogh = function(theta){
  alpha = theta[1]
  beta = theta[2]
  loglik = ( sum( y*(alpha+beta*t - log(1+exp(alpha+beta*t))) ) +
             -sum( (1-y)*(log(1+exp(alpha+beta*t))) ) )
  return (-1*loglik)
}
logh = function(theta){
  return(-nlogh(theta))
}
# Find good initializations (laplacian)
thetahat=optim(c(-10, 2), nlogh)$par
Sigmahat=solve(hessian(nlogh,thetahat))
# MCMC sampling using adaptMCMC
m=10000
s=(2.4/sqrt(2))^2*Sigmahat #specify covariance matrix of the jumping distribution
out=MCMC(logh,n=m,init=thetahat,scale=s,adapt=T,acc.rate=.35)
theta = out$samples

### Wake Times Model ###
### Load indicator data
wake <- read.csv("data/wakedf.csv")
y <- wake$Asleep
t <- wake$Time.Offset
# Find good initializations (laplacian)
thetahat=optim(c(1, -2), nlogh)$par
Sigmahat=solve(hessian(nlogh,thetahat))
# MCMC sampling using adaptMCMC
m=100000
s=(2.4/sqrt(2))^2*Sigmahat #specify covariance matrix of the jumping distribution
out=MCMC(logh,n=m,init=thetahat,scale=s,adapt=T,acc.rate=.35)
theta = out$samples
```

Figure 10: Cumulative Plots for Sleep Time Model

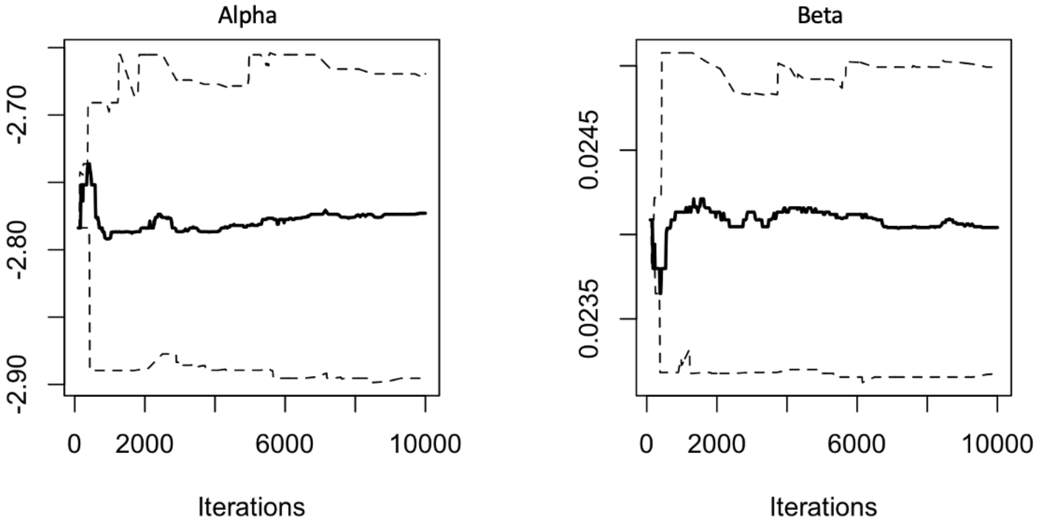


Figure 11: Cumulative Plots for Wake Time Model

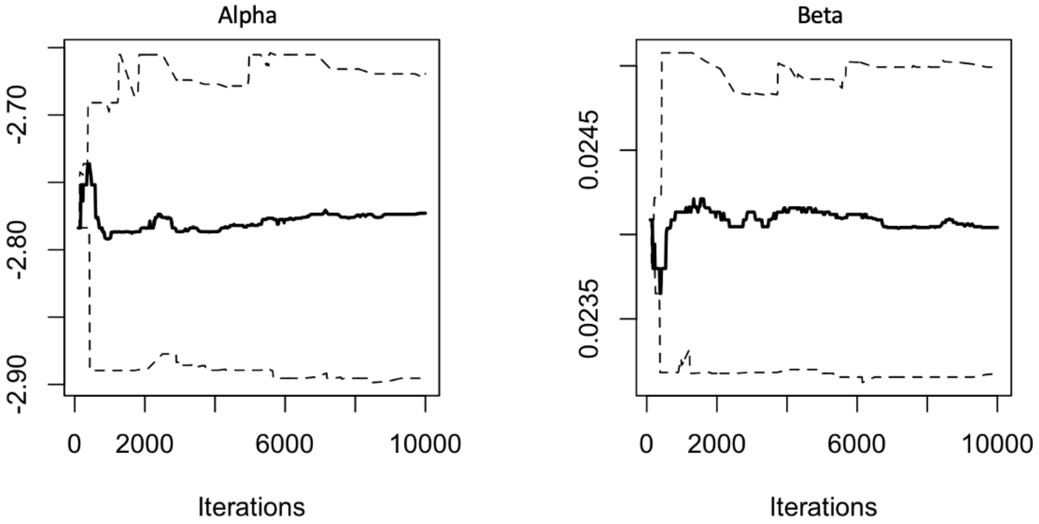


Figure 12: Correlation plot for merged dataset

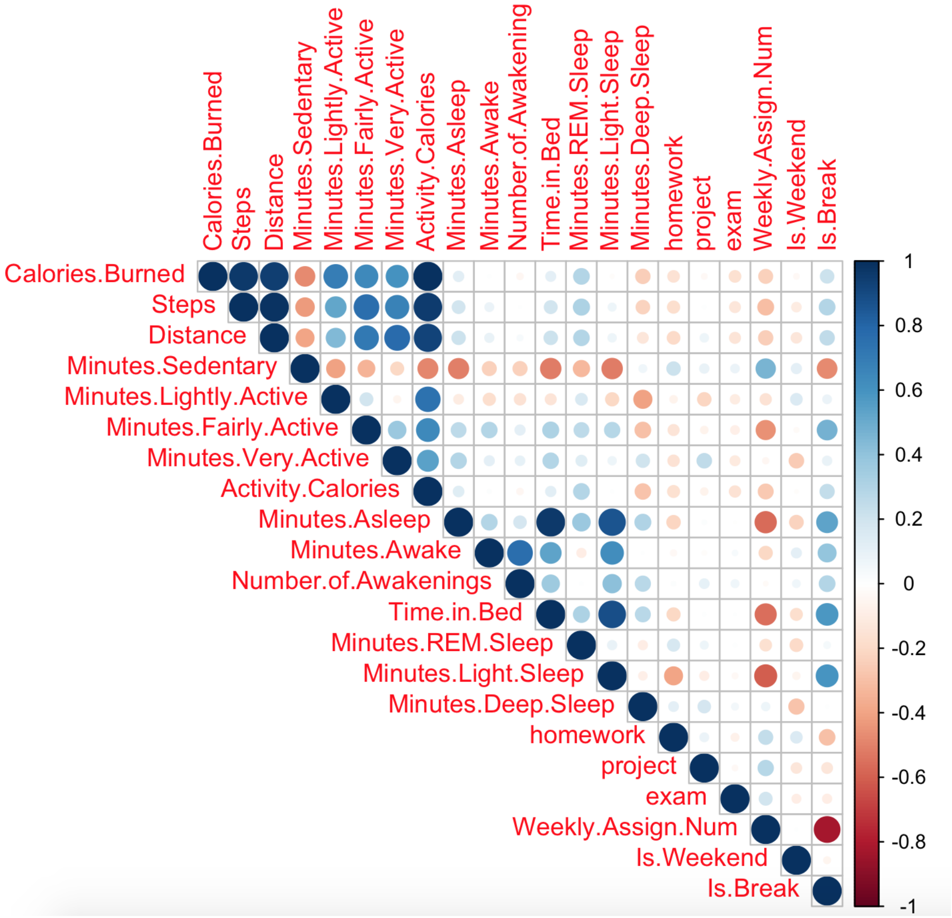


Figure 13: Potentially quadratic relationship between Minutes Sedentary and Response

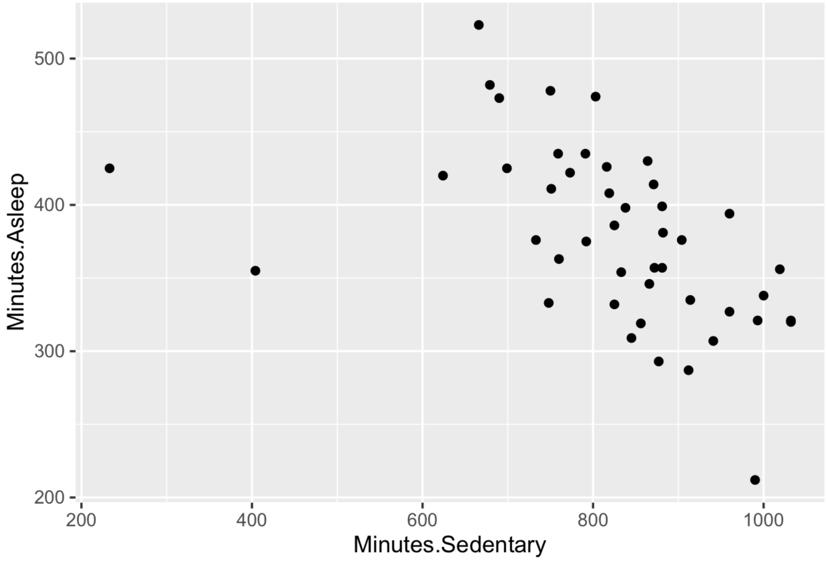


Figure 14: MC Algorithm for sleep duration and quality analysis

```r
### Sleep Duration Regression ###
### Load Data
merged <- read.csv("data/mergeddf.csv")
dropcols = c(1,2,3,8,9,10,11,12,14,15,16,17)
# save small df
sm_merge <- merged[,-dropcols]
sm_merge$Min.Sed.Sq <- sm_merge$Minutes.Sedentary^2 #add quadratic term
### Bayesian regression
# Define design matrix X and y
sm_merge$Intercept <- 1
X=as.matrix(sm_merge[,-5])
y=sm_merge$Minutes.Asleep
n = dim(X)[1]
p = dim(X)[2]-1
m = 10000
# Initializations
freq_mod6 <- lm(Minutes.Asleep~., data=sm_merge[,-13])
summary(freq_mod6)
sigma2=numeric(m)
sigma2[1]=summary(freq_mod6)$sigma^2
beta=matrix(0,nrow=m,ncol=p+1)
# Compute betahat
Sinv=solve(t(X)%*%X)
betahat=Sinv%*%t(X)%*%y
# MC sampling
for(i in 1:m){
  e=y-X%*%betahat
  sigma2[i]=1/rgamma(1,(n-p-1)/2,t(e)%*%e/2)
  beta[i,]=mvrnorm(1,betahat,sigma2[i]*Sinv)
}

### Sleep Quality Regression (# of Awakenings) ###
### Bayesian regression
# Define design matrix X and y
sm_merge$Intercept <- 1
X=as.matrix(sm_merge[,-5])
y=sm_merge$Number.of.Awakenings
# Initializations
freq_mod6 <-  lm(Number.of.Awakenings ~., data=sm_merge[,-13])
summary(freq_mod6)
sigma2[1]=summary(freq_mod6)$sigma^2
# Compute betahat
Sinv=solve(t(X)%*%X)
betahat=Sinv%*%t(X)%*%y
# MC sampling
for(i in 1:m){
  e=y-X%*%betahat
  sigma2[i]=1/rgamma(1,(n-p-1)/2,t(e)%*%e/2)
  beta[i,]=mvrnorm(1,betahat,sigma2[i]*Sinv)
}
```