Utrecht University

Faculty of Social and Behavioural Sciences

Sleep like a Bayesian:

Do healthy lifestyle habits promote sleep?

By

Shannon Dickson,

Methodology and Statistics for the Behavioural, Biomedical, and Social Sciences

Bayesian Statistics: Final Assignment

Submitted: June 2022

1 Introduction

Despite a great consensus that sufficient sleep is important for our general health, many adults do not engage in behaviours that promote it. According to the National Sleep Foundation, adults need between 7 - 9 hours of sleep to function well. How realistic is this? A 2017 National Health Survey reported that 1 in 5 people in the Netherlands have problems getting to sleep and staying asleep and 41% of these individuals reported poor daily functioning. Furthermore, a study at Leiden University found that more than a third of students at Dutch Universities do not get the recommended hours of sleep and that this negatively impacts their grades. This raises an important question: what kind of healthy lifestyle behaviours promote better (longer) sleep? I will focus on two possibilities: daily exercise and evening screen time. Exercise is known to regulate our circadian rhythm in a way that improves sleep. Screen time has the opposite effect on sleep, due to blue light exposure that promotes alertness. I aim to use a Bayesian linear regression analysis to assess the impact of daily exercise and evening screen time on the duration of sleep in adult. Although I expect both factors to be influential for sleep duration, I have a small suspicion that screen time will have a larger effect than exercise. My suspicion is entirely based on my own sleep-deprived status, despite excessive exercise, which is probably due to my habit of doom-scrolling Twitter in the hour before bed.

2 Descriptive Statistics

Data is acquired from the Sleep Quality and Behavioural Health dataset which has been shared on Kaggle (Arora et al., 2022a, 2022b; *SleepQual and b.health Dataset, Version 1*, 2022), which contains information about the sleep quality and behavioural health of 24 university students collected using smartphones and wearable technology. Only the following variables are used in the analysis: duration (daily sleep in minutes), active (daily exercise minutes), screen (evening screen time minutes), and onset (sleep latency in minutes). Sleep duration ranged from 94 - 866 minutes (mean = 413, SD = 128). Daily exercise ranged from 10 - 216 minutes (Mean = 70, SD = 45). Screen time ranged from 23 = 963 minutes (mean = 279, SD = 202). Sleep latency ranged from 0 - 69 minutes (Mean = 16, SD = 13). An interaction variable was created between screen time and sleep latency, as it is expected that greater screen time will lead to longer sleep latency, consequently reducing the time spent asleep.

3 Hypotheses

I have the following beliefs about the relationships in the data:

H_2: Sleep duration is greatly affected by daily exercise and evening screen time.

H_2: Screen time will have a larger affect on sleep duration than daily exercise and latency.

H_3: An interaction between screen time and sleep latency will have a larger effect on sleep duration than daily exercise and screen time alone.

4 Methods

4.1 Models

To assess H_1 and H_2 I will obtain the following model:

Model 1: $Sleep_i = \beta_0 + \beta_1 * Exercise_i + \beta_2 * Screentime_i + \beta_3 * Latency + \epsilon_i$

To assess H_3 I will obtain the following model:

Model 1: $Sleep_i = \beta_0 + \beta_1 * Exercise_i + \beta_2 * Screentime_i + \beta_3 * Interaction_i + \epsilon_i$

4.2 Markov Chain Monte Carlo: Sampling by the Gibbs and Metropolis-Hastings Algorithms

I will use two MCMC sampling methods to obtain the coefficients for different parameters. MCMC is a class of methods whereby we use simulated draws from an approximate posterior distribution and use these draws to calculate quantities of interest for the posterior distribution, such as the mean, standard deviation credible intervals, and MC error. The Metropolis-Hastings algorithm works by sampling from a proposal distribution, with each sampled value accepted givena a certain probability, known as the acceptance ratio. This ratio represents the probability that the proposed sampled value is greater than a value sampled from a uniform distribution. The Gibbs algorithm is a special case of Metropolis-Hastings, where the proposed value is always accepted. The Gibbs algorithm samples iteratively from the joint conditional posterior distribution of a certain parameter, given all other parameters. As such, we use Gibbs sampling when we know the full conditional posterior distribution and otherwise we use Metropolis-Hastings.

In this study, Gibbs sampling is used for β_0, β_1, β_3, and σ^2) and Metropolis-Hastings sampling is used for β_2. Slightly informative priors are specified $\beta_0 \sim N(0, 0.25)$ and $\beta_{1...3} \sim N(0, 0.5)$. The residual

error variance is $\tau \propto \frac{1}{s^2} \sim IG(0.001, 0.001)$. Starting values are chosen arbitrarily for the parameters and ideally do not (heavily) influence the coefficients. The sampling procedure is performed twice, forming two chains with different starting values. I assess convergence of each chain and the pooled chains is by a visual inspection of trace plots, autocorrelation plots, and by comparing the resulting parameter estimates including the Monte Carlo Error.

4.3 Posterior Predictive Checks

Posterior predictive checks are used to evaluate the assumptions of the regression model, by means of outliers, skewness, and a crude measure of homoscedasticity (correlation of residuals and fitted values). These test-statistics and discrepancy measure of the observed data is compared to that of many simulated datasets. Outliers are assessed by computing the difference between the largest and smallest values. P-values around 0.5 are ideal, with extreme values indicating a smaller or wider spread in the simulated data versus observed. Skewness is also computed as an assessment of multivariate normality, and similarly compared in the simulated vs observed datasets. Lastly, a crude, graphical posterior predictive check is conducted by comparing the fitted values vs residuals in the simulated and observed datasets (see Supplementary A).

4.4 Model comparison

I will compare the different models by means of the Deviance Information Criterion (DIC). The DIC is a Bayesian method for model comparison that considers the fit and complexity of a model, formulated as $DIC = D(\bar{\theta}) + 2p_D$, with lower DICs preferred.

4.5 Bayes Factor

The model parameters are compared using the Bayes Factor, according to H_1, H_2, and H_3 specified previously. Each hypothesis will be compared to its complement, such that Bayes Factors are a measure of support for $H_1 : \beta_1 = 0, \beta_2 = 0, \beta_3 = 0; H_c : \beta_1, \beta_2, \beta_3$, $H_2 : \beta_2 > \beta_1; H_C : \beta_1, \beta_2$, and $H_3 : \beta_3 > \beta_2 \& \beta_3 > \beta_1; H_C : \beta_1, \beta_2, \beta_3$.

5 Results

5.1 Convergence

Figure 1 presents the trace plots per chain of each sampled parameter, once the burn-in period has been discarded. The two chains had different starting values, but the range of the chains appears similar to one another indicating they are sampling similar values. However, there appears to be a lot of variability in the samples, which can sometimes indicate autocorrelation, but at least the chains vary similarly. I am comfortable concluding that the convergence according to these trace plots is acceptable and 31,000 iterations should be sufficient.

Figure 1. Trace plots

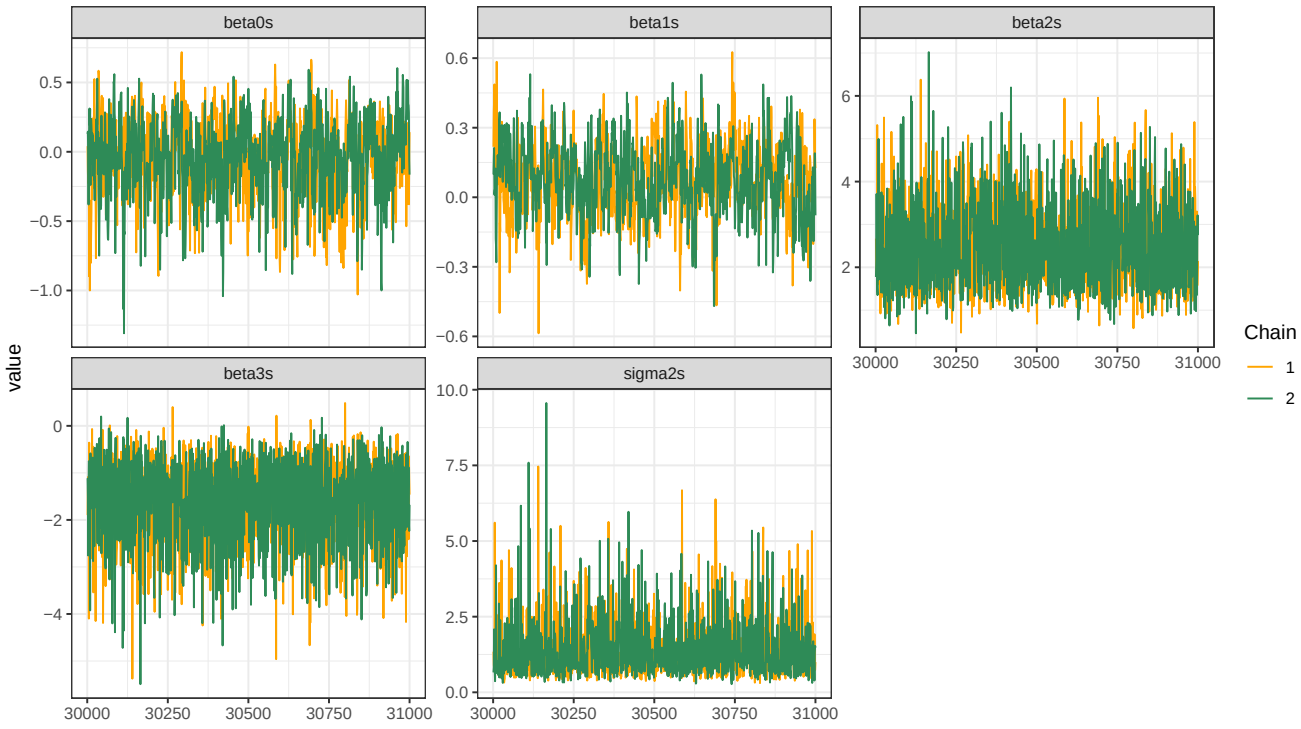
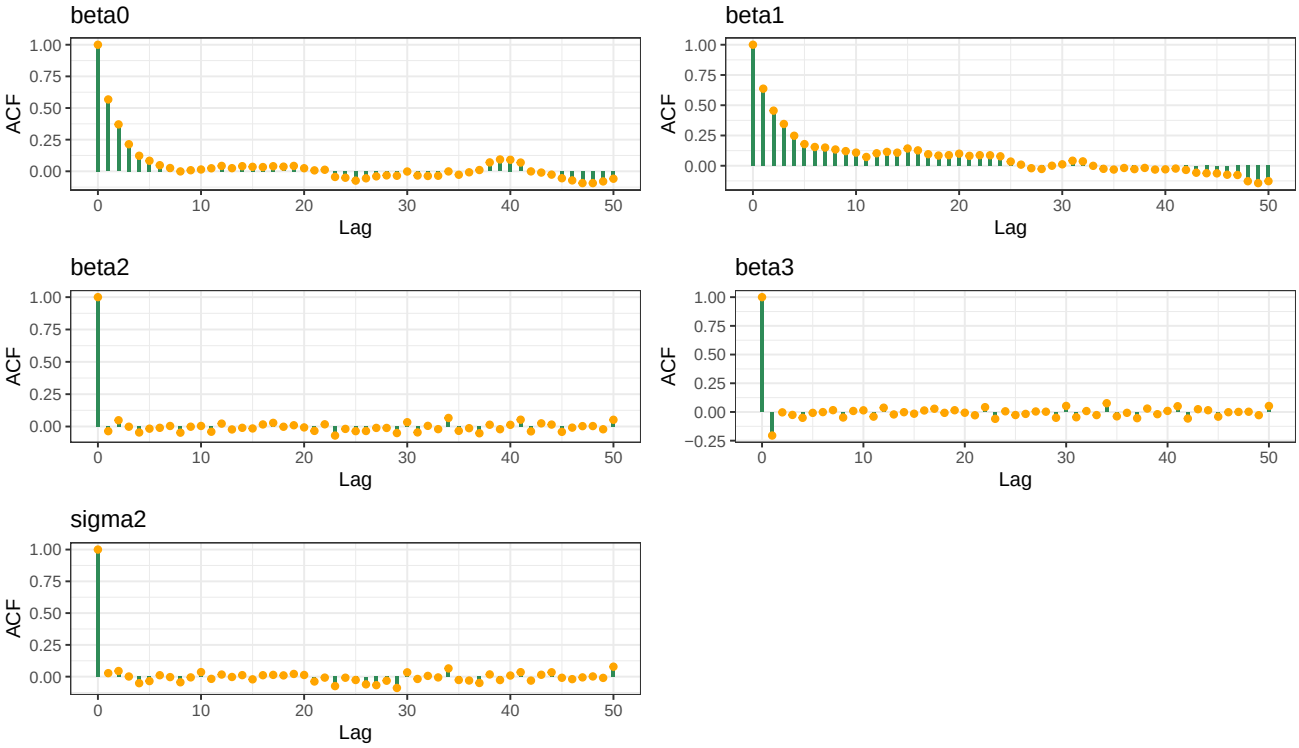


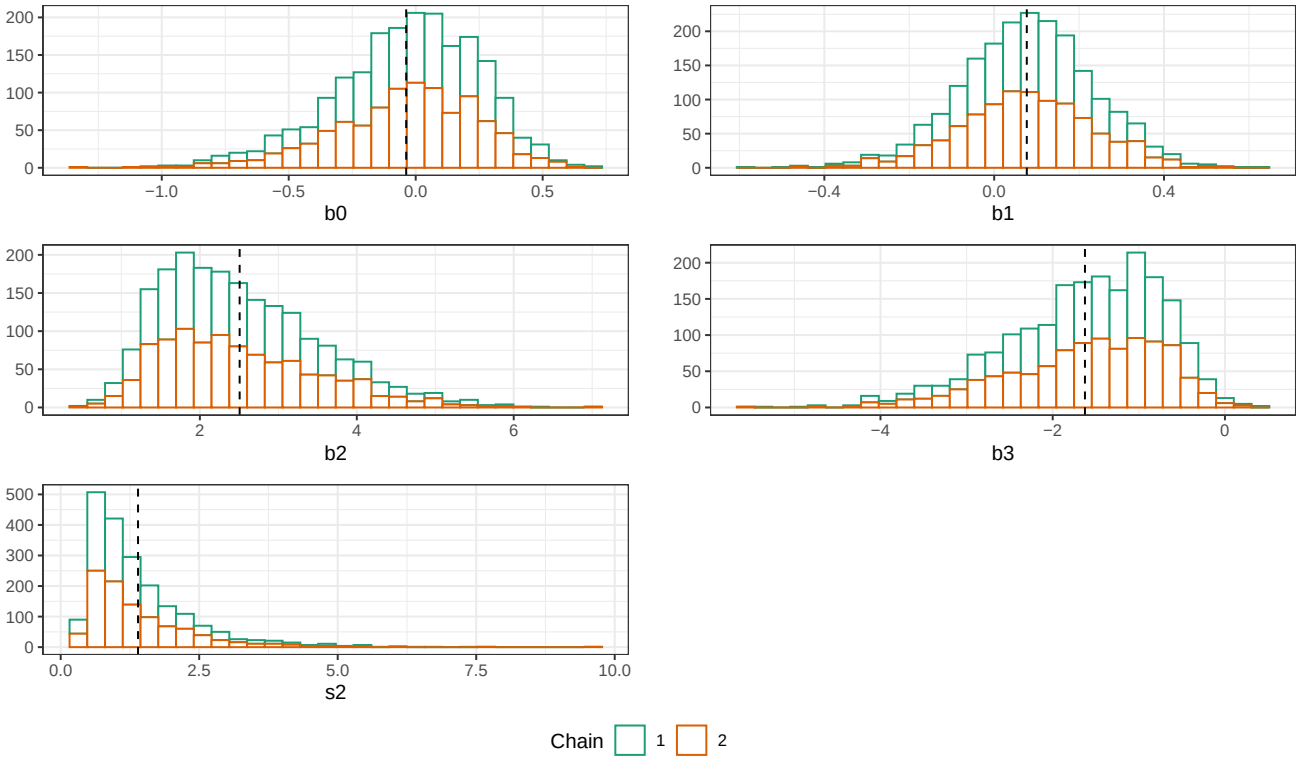
Figure 2. presents a nice overview of the autocorrelation for each parameter (pooled across chains). By visually inspecting these plots autocorrelation does not seem to be a problem, at most for β_1. Given the variability seen in the trace plots I am somewhat surprised that autocorrelation appears so low, at least visually. However, this is good news for β_2 which is sampled with a Metropolis-Hastings step. I am happy with the autocorrelation of this parameter, given the very small tuning step.

Figure 2. Autocorrelation plots



Posterior densities for the two chains are shown in **Figure 3**. Although the posterior densities show that the two chains overlap reasonably well, despite the different starting values, they are not normally distributed. The posterior density for sigma is extremely right skewed and chain 2 exhibits more extreme values in the tails. From this, I expect the model assumptions to be violated, namely outliers and skewness.

Figure 3. Posterior densities



5.2 Sample Statistics

Table 1 displays the sample statistics for the pooled chains for each parameter, including the the (posterior) mean, standard deviation, 2.5% and 97.5% credible confidence intervals, and the MC error. The sample statistics for sigma seem rather large, but the MC error is still much smaller. In fact, the MC errors for all parameter estimates are much smaller than the standard deviations. Overall, I conclude that the sampler has converged.

Table 1: Pooled sample statistics for the posterior distributions of the parameters

	Mean	SD	CI 2.5%	CI 97.5	MC Error
beta0	-0.037397	0.285159	-0.686944	0.442037	0.001620
beta1	0.076763	0.156129	-0.243785	0.371403	0.000887
beta2	2.506897	0.993447	1.037795	4.808925	0.005642
beta3	-1.625667	0.886619	-3.640130	-0.285205	0.005036
sigma	1.396202	0.956962	0.442508	4.031321	0.005435

The posterior predictive p-value (PPV) for the assumption of the absence of outliers is 0.001, which is very different from the ideal 0.5, and indicates that outliers are present in our data. The PPV for skewness is 1, again, indicating that the data is highly skewed. Lastly, the PPV for the crude measure of homoscedasticity is 0.626, meaning there is some heterskedasticity present. However, the graphical check of the homoscedasticity appears not too extreme.

5.3 Bayes Factor and DIC

The Bayes Factor for $H_1 : \beta_1 = 0, \beta_2 = 0, \beta_3 = 0; H_c : \beta_1, \beta_2, \beta_3$ is 0.00, meaning there is a very large amount of support for the hypothesis that all three coefficients are greater than zero. Therefore, we can conclude than the hypothesised model fits the data and that sleep is affected by exercise, screen time, and sleep latency. The Bayes Factor for $H_2 : \beta_2 > \beta_1; H_C : \beta_1, \beta_2$ is also 0.00, again indicating that there is more support than exercise has a larger effect on sleep than exercise. Finally, the Bayes Factor for $H_3 : \beta_3 > \beta_2 \& \beta_3 > \beta_1; H_C : \beta_1, \beta_2, \beta_3$ is 166.59, also supporting the hypothesis that the interaction between screen time and sleep latency has a greater effect on sleep than exercise and screen time alone.

The DIC for model 1 is 7444 and for model 2 is 7857.

6 Discussion and Interpretation

Using a Bayesian regression analysis that combines Gibbs and Metropolis-Hastings algorithms, I was able to model the effect of daily exercise, screen time, and an interaction between screen time and sleep latency, on sleep duration. Furthermore, I computed a DIC for two models, and Bayes Factors for each of my three hypotheses. The Bayes Factors for H_1 corroborated my hypothesised beliefs that these factors *do* affect sleep duration. Additionally, they indicate consistent with H_3 that the interaction term has the largest effect on sleep, followed by exercise and then screen time, the latter being inconsistent with H_2. The DIC was lowest for the model including the interaction term, which also matches my hypotheses.

The positive regression coefficients for screen time is not entirely as expected. I predicted that increased screen time would reduce sleep duration, and I stand corrected. It is interesting to think about possible confounders, for instance, more screen time at night could mean a later bedtime and a much longer sleep-in than usual to compensate. This is, however, speculation. As expected, increased exercise lead to increased sleep duration, and the interaction term lead to reduced sleep duration.

The regression assumptions tested by means of a PPV highlighted some model violations. Outliers, skewness, and slight homoscedasticity are present in the data. I tried to circumvent this issue in two ways: Firstly, I implemented a skew-normal proposal distribution within the MH-sampler. This did help the sampler converge more appropriately. Secondly, I computed the log-conditional posterior t-distribution within the MH step to account for the skewed data in particular. Although not included here, without these two approaches the sampler would not come close to convergence. However, the assumptions remain violated. Future studies could focus on a sensitivity analyses, to inspect the influence of different kinds of priors.

6.1 Frequentist or Bayesian: is that the question?

I only consider a Bayesian approach to this analyses, for the following reasons. Bayes Factors are very desirable, as they allow more flexible testing of several different hypotheses in a way that classic hypothesis testing cannot. This means we can see what is the *most* important predictor of sleep, satisfying our search for meaning. I also prefer the interpretaton of credible intervals over confidence intervals. Credible intervals form a distribution that gives a nice indication of the actual chances that our coefficient is zero. This is the case for exercise for instance, and we can assume it would not be significant in a classical setting, giving the best of both worlds. Lastly, I was able to incorporate prior information into my analyses like the true scientist I claim to be. Priors are also useful for regulating

non-normal data. Although the priors are only weakly informative, they do help the sampler converge and we can use this information going forward (e.g. in a sensitivity analysis).

7 References

Arora, A., Chakraborty, P., & Bhatia, M. P. S. (2022a). SleepQual and B.Health: Smartwatch and Smartphone based Behavioral Datasets of Youth. *2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 340–344. https://doi.org/10.1109/Confluence52989.2022.9734122

Arora, A., Chakraborty, P., & Bhatia, M. P. S. (2022b). Intervention of Wearables and Smartphones in Real Time Monitoring of Sleep and Behavioral Health: An Assessment Using Adaptive Neuro-Fuzzy Technique. *Arabian Journal for Science and Engineering*, *47*(2), 1999–2024. https://doi.org/10.1007/s13369-021-06078-5

SleepQual and b.health dataset, version 1. (2022). https://www.kaggle.com/datasets/anshika1011/sleepqual-and-bhealth-dataset.

8 Supplementary A

Figure A1 and A2 show the fitted values plotted against the residual values, in the simulated and observed data.

Figure A1: Observed Data

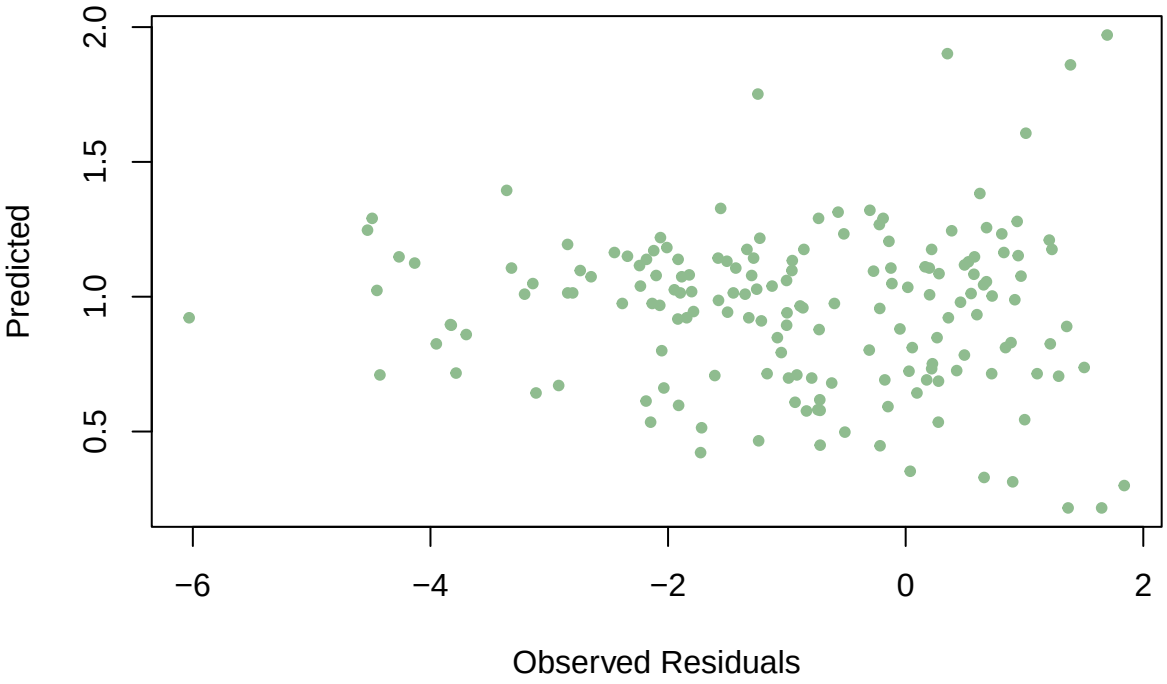
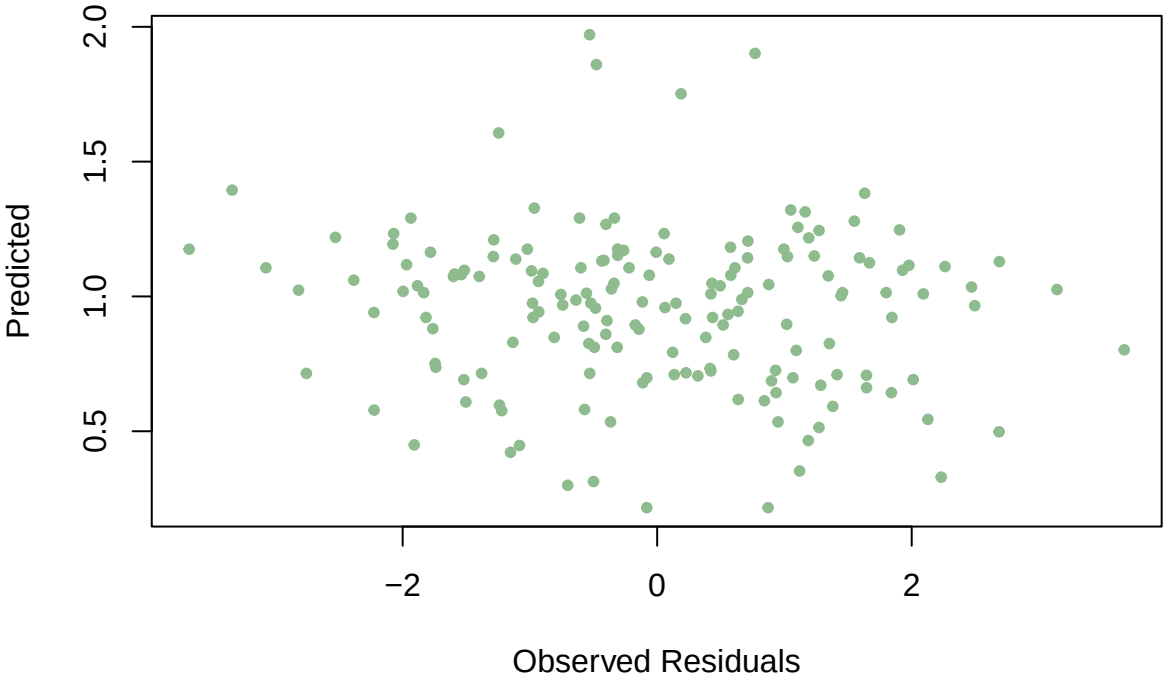


Figure A2: Simulated Data



t