Explaining the Association Between Exercising and Academic Performance

Chhiring Lama, Sakaiza Rasolofomanana Rajery

12/11/2020

1. Background

Students focus on academics might affect their physical well being which can be measured through their exercising habit. This analysis is an attempt to unveil the relationship between the exercising habit of students and their academic performance. Using a dataset of 343 observations from the `Lock5Data` Package, we will rely on bayesian methods, highlighting the importance of balancing prior information with data in order to infer a posterior interpretation, to explain the relationship between these two factors.

2. Data set

```
library(Lock5Data)
data(GPAGender)
```

3. Analysis

3.a) Building the Model

Variables

In this analysis, we used the following variables:

Exercise (hours per week): Response variable, which is continuous

SAT (out of 1600 points) : Explanatory variable, which is continuous

GPA (out of 4 points): Explanatory variable, which is continuous

Defining the likelihood model and parameters

Assuming that there is a linear association for exercise with SAT and GPA, the expected amount of exercise (Y) in hours per week for a student i as a function of GPA (X) and SAT(Z) for that student can be written as:

$$Y_i = \beta_0 + \beta_1 * X_i + \beta_2 * Z_i + \epsilon_i \ N(0, \sigma)$$

where:

β_0 : The amount of exercise in hour when GPA and SAT scores are both 0

β_1 : The change in hours of exercise for every point increase in the student's SAT, when GPA is kept constant.

β_2 : The change in hours of of exercise for every point increase in the student's GPA, when SAT is kept constant

ϵ_i : The measure of the variability in exercising hours per week for students with similar GPA and SAT scores.

Hypothesis

We hypothesized that there is a negative association between hours of exercise and GPA, SAT.

As SAT increases, the number of hours a student exercises per week decreases. Mathematically:

$H_0 : \beta_1 = 0$

$H_a : \beta_1 < 0$

As GPA score increases, the number of hours a student exercises per weak decreases. Mathematically:

$H_0 : \beta_2 = 0$

$H_a : \beta_2 < 0$

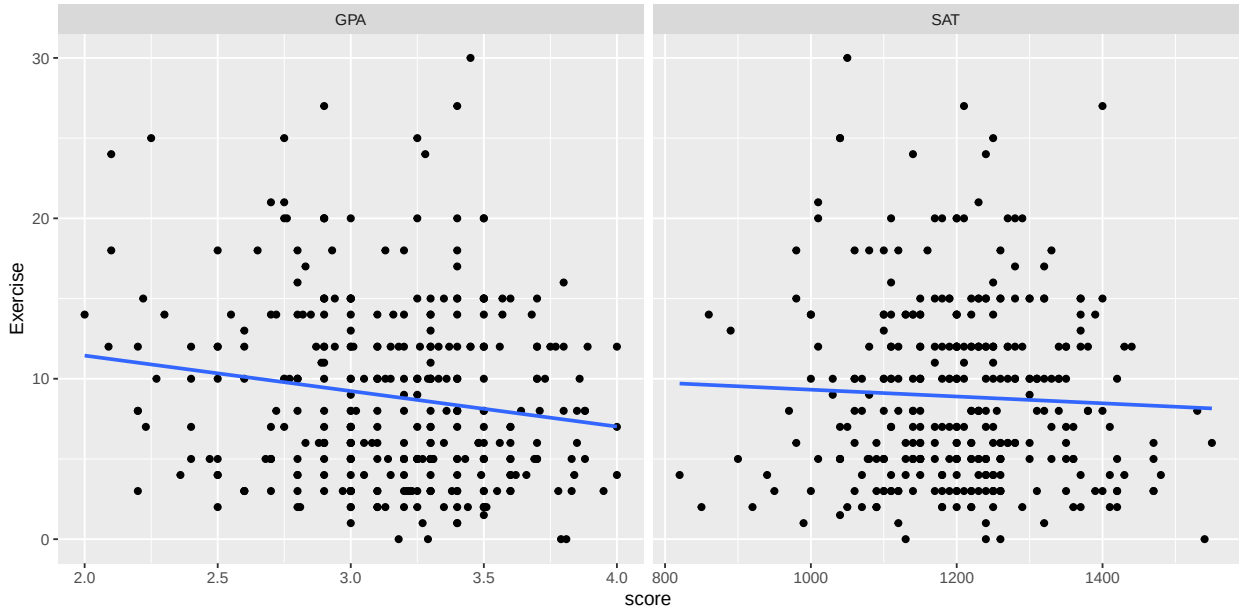
Prior models

Vague prior from the rstan model was used, therefore there was none manually specified.

Visualization

```
new_data <- GPAGender %>%
  gather(key, score, -c(Exercise, Pulse, Piercings, GenderCode))

ggplot(new_data, aes(score, Exercise))+
  geom_point()+
  geom_smooth(method = "lm", se = FALSE)+
  facet_wrap(~key,scales = "free_x" )
```



As shown in the figure above, in general, both GPA and SAT have negative linear correlation with exercise. That means that as the GPA and SAT scores increase, the predicted number of hours exercised per week is estimated to decrease. The decrease is steeper for GPA than SAT, so it can be assumed that the coefficient β_1 will be more negative than β_2.

Defining the Regression Model

To identify the association between the number of hours exercised and academic performance (GPA and SAT score), a multiple linear regression was run with `SAT` score and `GPA` score as the predictors and `Exercise` as the dependent variable.

A bayesian linear regression model was created with **Rstan** package using the `stan_glm` function. The number of hours exercised (`Exercise`) could be predicted using `GPA` and `SAT`, assuming a normal (Gaussian) likelihood model. For this model, it is assumed that a normal-normal model can be used because one can assume that the distribution of exercise as well as that for GPA and SAT scores are bell-shaped. The model contains an MCMC simulation with 4 Markov chains with 10000 iterations for each chain.

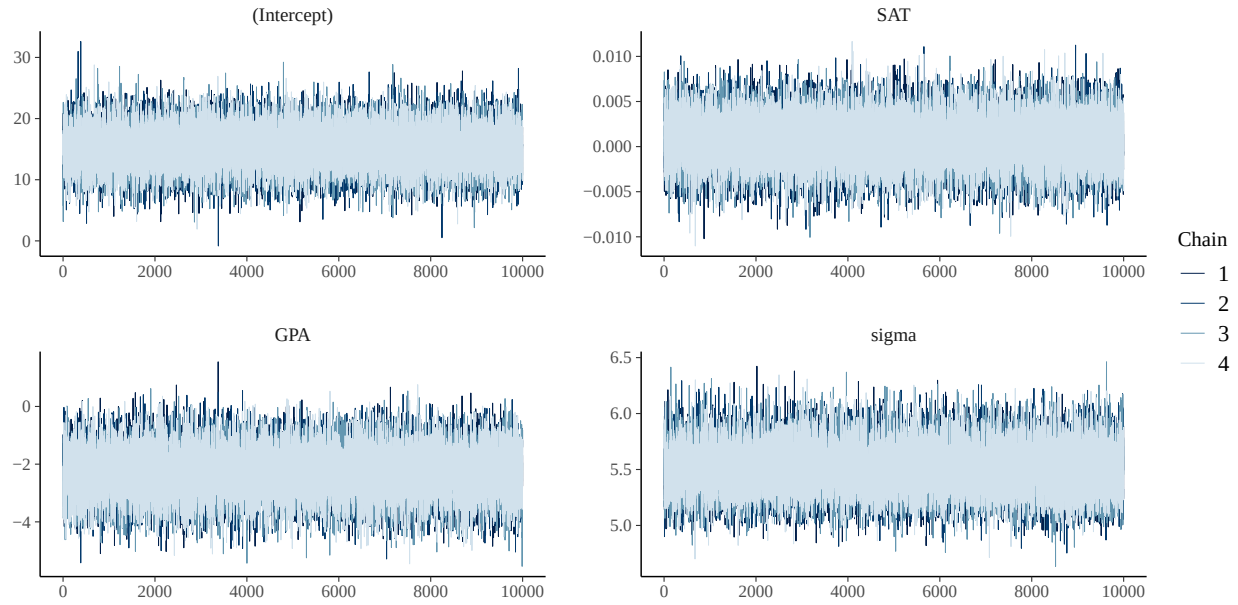
```
set.seed(563453)

gpa_model_sim <- stan_glm(Exercise ~ SAT + GPA,
                          data = GPAGender,
                          family=  gaussian,
                          chains = 4,
                          iter = 10000*2)
```

3.b) Diagnosis

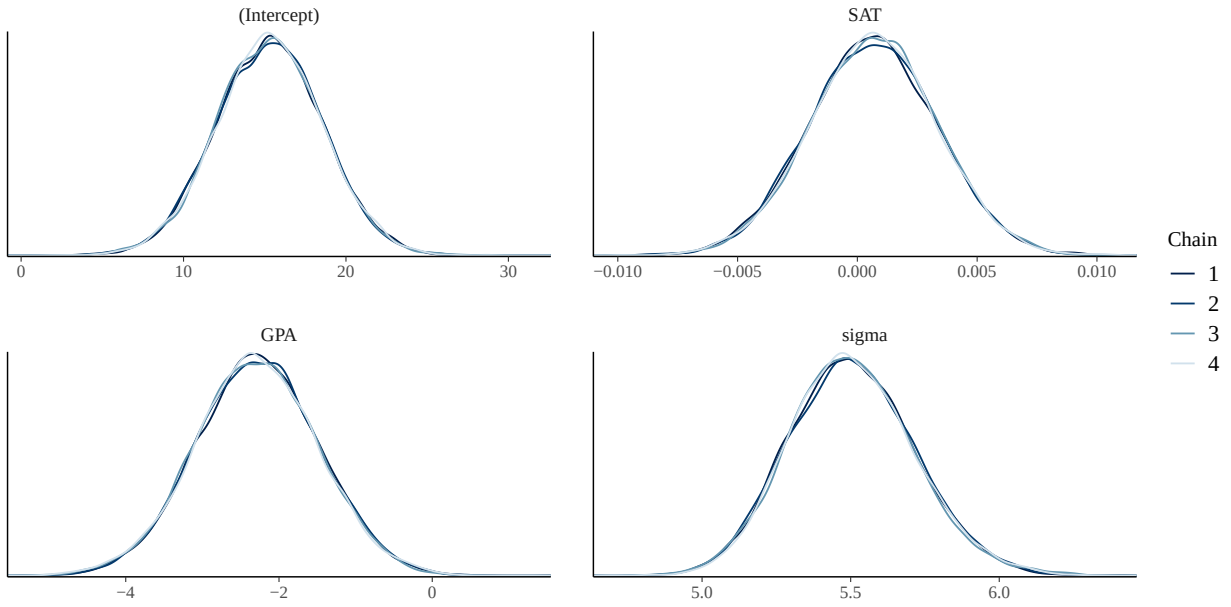
Once the model was created, diagnostics needed to be run to make sure that the Markov chains used for the regression model was large enough (with adequate iterations) and provided more accurate approximation of the posterior. To check if the MCMC simulations meets the criteria, two plots were graphed: MCMC trace plot and density plot overlaying the four chains.

```
mcmc_trace(gpa_model_sim, size = 0.1)
```



The major criteria for a good MCMC simulation according to the trace plot is that the plot resembles white noise without showing any specific overall trends. Looking at the plot above, the graphs for all the coefficients ($\beta_0,\ \beta_1,\ \beta_2,\ \sigma$) meet the criteria.

```
mcmc_dens_overlay(gpa_model_sim)
```



The results from the density plot agree with the trace plots. For each coefficients, even though the points for the chains do not match exactly, which is considerable given the randomness in simulation, the overall trend of the four chains are similar.

```
gpa_model_summary <- summary(gpa_model_sim)
df <- as.data.frame(gpa_model_summary) %>%
  select(n_eff, Rhat)
head(df, -2)
```

```
##             n_eff      Rhat
## (Intercept) 53054 0.9999395
## SAT         38561 1.0000049
## GPA         39073 1.0000592
## sigma       40455 1.0000043
```

Since Rhat for each coefficient is more or less equal to 1, there is stability across the parallel chains. Besides, n_eff(effective sample size) is large enough and close to N (actual sample size). This shows that our posterior approximation is quite accurate.

Hence, utilizing the trace plot, density plot and the summary statistics above, it can be assumed that the MCMC simulation for the model has adequate iterations, the Markov Chains are *good* and the model can be used as an accurate way to approximate the posterior.

3.c) Interpreting the posterior

The posterior summary statistics provide us with the credible interval of each parameter.

```
head(as.data.frame(gpa_model_summary),-2)
```

```
##                       mean         mcse          sd          10%           50%
## (Intercept) 15.2960378827 1.423931e-02 3.279805373 11.110294245 15.3089122266
## SAT          0.0006675531 1.355897e-05 0.002662559 -0.002745957  0.0006726867
## GPA         -2.2857097373 4.065685e-03 0.803662519 -3.306605150 -2.2887135018
## sigma        5.5048969931 1.051700e-03 0.211532659  5.240320743  5.4980539280
##                       90% n_eff      Rhat
## (Intercept) 19.453843977 53054 0.9999395
## SAT          0.004071294 38561 1.0000049
## GPA         -1.250706303 39073 1.0000592
## sigma        5.779217836 40455 1.0000043
```

For Intercept: There is a 80% posterior probability that a student that has a GPA and a SAT score of 0 exercise between 11.04 and 19.5 hours per week.

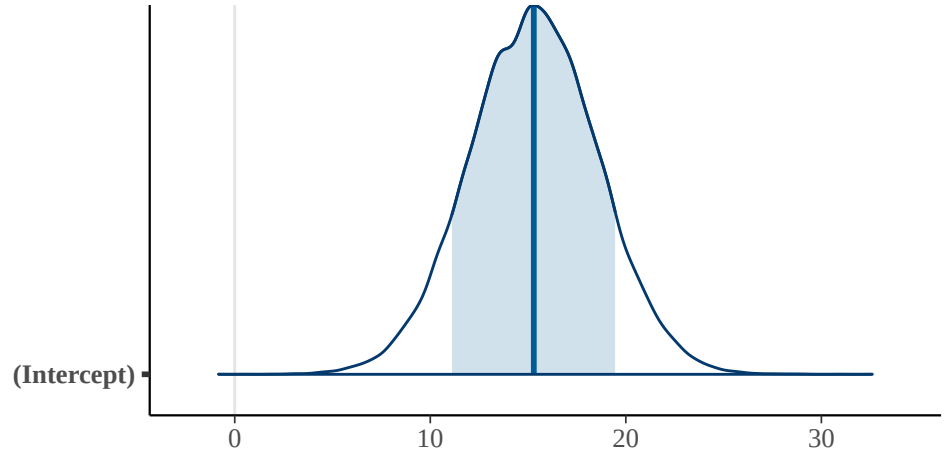
For SAT: There is a 80% posterior probability that for each point increase in SAT, there is a decrease of between 0.003 and 0 in hours of exercise per week for a student.

For GPA: There is a 80% posterior probability that for each point increase in GPA, there is a decrease between 1.27 and 3.30 hours in the amount of exercising a student practices per week .

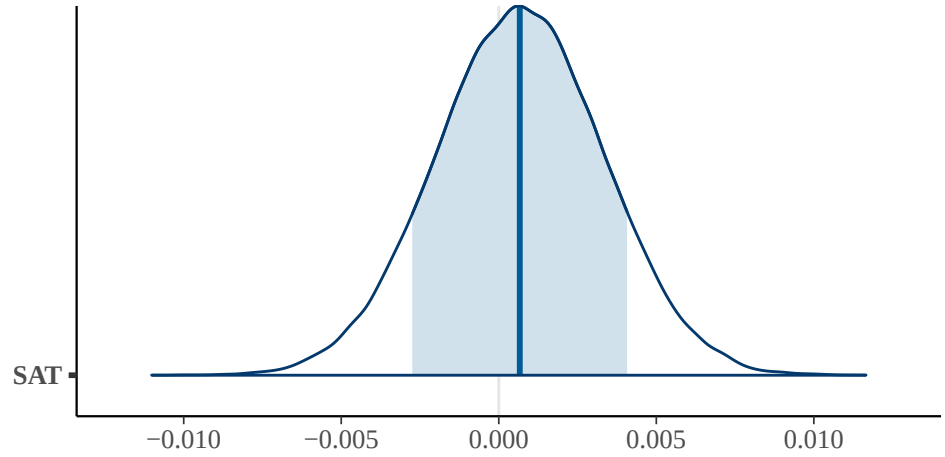
These intervals are visualized by the shaded regions below:

Visual representation of Posterior credible interval

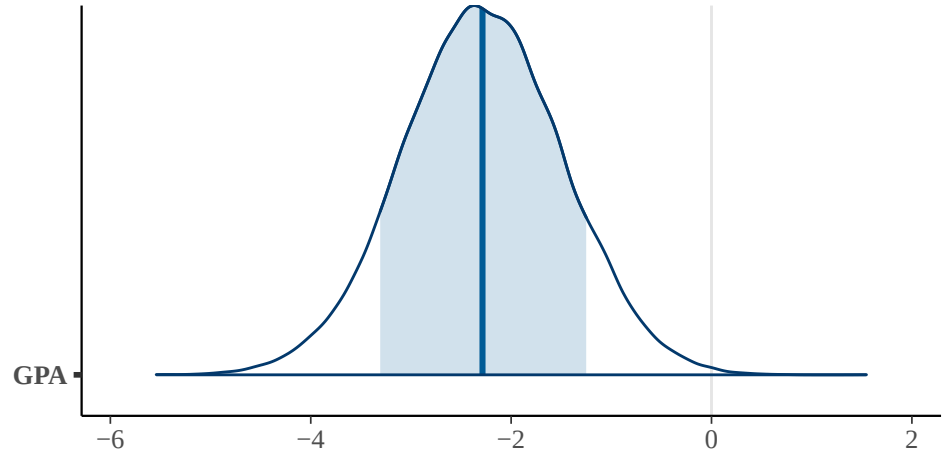
```
mcmc_areas(gpa_model_sim,
           pars = c("(Intercept)"),
           prob = 0.80,
           point_est="mean")
```




```
mcmc_areas(gpa_model_sim,
           pars = c("SAT"),
           prob = 0.80,
           point_est="mean")
```




```
mcmc_areas(gpa_model_sim,
           pars = c("GPA"),
           prob = 0.80,
           point_est="mean")
```



3.d) Hypothesis testing

Since the regression model we built provided us with 40000 alternative scenario for the trend in the relationship between Exercise, SAT and GPA, we can then compute the posterior probability to check if β_1 and β_2 are actually negative.

Data frame creation

```
set.seed(563453)

gpa_model_df <- as.array(gpa_model_sim) %>%
  melt %>%
  pivot_wider(names_from = parameters, values_from = value)
```

Posterior probability that $\beta_1 < 0$

```r
set.seed(563453)

gpa_model_df %>%
  mutate(exceeds_0 = SAT < 0) %>%
  tabyl(exceeds_0)
```

```
##  exceeds_0     n  percent
##      FALSE 24015 0.600375
##       TRUE 15985 0.399625
```

Among 40 000 simulations of the same relationship, 40% display a negative correlation between SAT and exercising. The posterior odds is 0.665 which is less than 1.

Posterior probability that $\beta_2 < 0$

```r
set.seed(563453)

gpa_model_df %>%
  mutate(exceeds_0 = GPA < 0) %>%
  tabyl(exceeds_0)
```

```
##  exceeds_0     n  percent
##      FALSE    77 0.001925
##       TRUE 39923 0.998075
```

Among 40 000 simulations, 99% display a negative correlation between GPA and exercising when controlling for the other variable. The posterior odds is 990 which is much greater than 1.

Because there is 80% chance that β_1,the SAT coefficient, is between -0.003 and 0 and the posterior odds of there being a negative correlation between SAT and number of hours of exercise is less than 1, we fail to reject the null hypothesis that β_1 is equal to zero.

Because there is 80% chance that β_2,the GPA coefficient is between -3.30 and -1.27 and the posterior odds of there being a negative correlation between GPA and number of hours of exercise is much greater than 1, we reject the null hypothesis that β_2 is equal to zero.

4. Conclusion

From the analysis above we can conclude that there is not enough posterior evidence that there is a negative association or even just an association between Exercising hours of students and their SAT scores. However, there is ample evidence that there is a negative association between GPA and exercising hours. Using bayesian methods we went from our prior understanding of the relationship between these variables, to a posterior analysis using 40 000 simulated scenarios, which also account for uncertainties in our posterior. This was made possible by using the data from a dataset of 343 observations. Through this analysis, not only can we see that the bayesian method is a good way to interpret data within human context, but we can also see the power of rtsan in simulating massive information from a small dataset.

5. References:

"Simple Normal Regression." Bayes Rules! An Introduction to Bayesian Modeling with R, by Alicia A Johnson et al., pp. 247–264.