Caffeine Consumption Regression

Radha Chitgopkar, Muskaan Mahes, Adeenah Choksi, and Lauren Figura

Final Project

STAT 3300: Applied Statistics

Professor Wickersham

30 April 2024

Radha Chitgopkar, Adeenah Choksi, Muskaan Mahes, Lauren Figura
Group 1

Abstract:

This study investigates how much caffeine intake students consume depending on their credit hours, earliest class time, and hours of sleep they get per night, om average. The sample consisted of n=28 participants from STAT 3300 class. From this study it was hypothesised that the various factors of credit hours, hours of sleep, and preferred class time does impact how much caffeine a student intakes. This study could have limitations due to response bias as it may not accurately reflect the genuine preferences of students. Based on the data there is sufficient evidence at the $\alpha = 0.1$ level to conclude the variable hours slept and credit hours have a significant relationship with the consumption of caffeine intake.

Study Design:

In this study, our population of interest is the students within the Data Science and Statistical Science majors and minors at Southern Methodist University. We suspect that credit hours, the amount of sleep one gets on average, whether someone is a morning or night person, and one's earliest class affects the amount of caffeine one consumes. Our sample frame was taken from the SMU STAT3300 roster and our targeted sample size was the 30 students in STAT3300. Our study consists of four explanatory variables: the amount of sleep one gets per night on average, the number of credit hours taken during the present semester, preferred earliest class time, and someone being a morning or night person. One quantitative response variable, being the amount of caffeine one consumed, was also recorded. In addition, the amount of sleep one gets per night on average and being a morning or night person were categorical variables. Preferred earliest class time was also categorical, with 7-8:59 a.m., 9-10:59 a.m., 11-12:59 a.m., etc., being the options. Caffeine intake was quantitative. Our link is listed below to access:

https://docs.google.com/forms/d/e/1FAIpQLSclBVv8EfO5tvrJTVS9hKvdU118KuqFv3eUBgVBdZFJQGk-dA/viewform?usp=sf_link

Data Collection

The data collection method used was an observational, cross-sectional study. This is because the variables in question contain ongoing factors in order to solve for the main question: How much caffeine does the students of STAT 3300 consume? This was a purposive sampling method, as all participants are students of the STAT 3300 class taken during the Spring 2024 semester. The process was simple, as it consisted of a simple, 5-question survey of multiple choice and free-response formatted questions that the participants answer based on what answer is most applicable to them, respectively. This survey was emailed to 32 people from the STAT 3300 class using the class roster. Participants simply had to choose which answer choice best fit their current lifestyle. Out of the 32 people this survey was emailed to, 28 participants responded with their preferences whereas four participants did not respond.

Limitations

The purpose of our study was to measure the amount of caffeine consumed by the population of SMU data science or statistical science majors or minors due to whether they're morning or night people, the number of credit hours they're currently taking, and the hours of sleep they get, on average, per night. Thus, by creating our survey we included several times ranging from seven to eleven AM, where students can pick their preferred class time. From this section of the survey we discovered that there was a potential impact to consider as some students selected the earliest class time as 7-8:59 am. We inferred that the individuals may have unconsciously favoured the earliest available class time due to influence of our current STAT 3300 class time; rather than honestly expressing what optimal timing would be best for these students Therefore, there are limitations to our study as it may not accurately reflect genuine preferences of students.

Additionally, our study may be subject to the response bias which occurs when participants respond inaccurately to questionnaires. It is unlikely that participants accurately recorded how much caffeine they consume daily or how much sleep they get. Their responses are likely estimates instead of exacts. Moreover, participants may skew their answers if their caffeine intake or average sleep amount is unhealthy and they do not feel comfortable admitting that. Furthermore, there was a nonresponse bias present with our data as 28 out of 32 participants responded, with four participants not responding. Due to this, the obtained sample is not representative of the complete population for our study. Moreover, an undercoverage bias is formed as a part of the population is excluded from the sample frame and as such, have no chance to participate in the sample.

A variety of lurking variables may be impacting our study's conclusions. Two potential lurking variables to consider are the student's sleeping pattern and personal activities of preferences. Firstly, for example participant's may be excluding the fact that they could be constantly waking up in the middle of the night due to various factors such as higher levels of stress or anxiety due to altered caffeine consumptions that therefore subtracts from their total amount of sleep. Secondly, students may have daily personal activities or a job that they engage in which could influence their response to be subsequent to whether they are a morning or night person.

Radha Chitgopkar, Adeenah Choksi, Muskaan Mahes, Lauren Figura
Group 1

HoursSlept		CreditHours	PreferredEarliest		MorningNightPerson		CaffeineIntake
3-5	1	18	7 - 8:59 am	1	Night Owl	1	3
8+	3	18	7 - 8:59 am	1	Night Owl	1	50
3-5	1	12	7 - 8:59 am	1	Night Owl	1	200
6-8	2	16	7 - 8:59 am	1	Morning person	2	22.5
6-8	2	15	7 - 8:59 am	1	Night Owl	1	20
6-8	2	15	7 - 8:59 am	1	Night Owl	1	80
6-8	2	15	7 - 8:59 am	1	Night Owl	1	95
3-5	1	15	7 - 8:59 am	1	Morning person	2	595
6-8	2	15	11 am - 12:59 pm	3	Morning person	2	3
6-8	2	17	9 - 10:59 am	2	Night Owl	1	20
6-8	2	15	7 - 8:59 am	1	Night Owl	1	55
6-8	2	18	7 - 8:59 am	1	Night Owl	1	15
6-8	2	15	7 - 8:59 am	1	Morning person	2	40
6-8	2	12	9 - 10:59 am	2	Night Owl	1	45
6-8	2	15	7 - 8:59 am	1	Night Owl	1	30
6-8	2	16	11 am - 12:59 pm	3	Night Owl	1	100
6-8	2	15	9 - 10:59 am	2	Night Owl	1	90
6-8	2	16	9 - 10:59 am	2	Night Owl	1	20
6-8	2	16	7 - 8:59 am	1	Morning person	2	0
6-8	2	16	9 - 10:59 am	2	Night Owl	1	20
6-8	2	14	9 - 10:59 am	2	Morning person	2	120
3-5	1	18	7 - 8:59 am	1	Night Owl	1	0
8+	3	15	9 - 10:59 am	2	Night Owl	1	0
6-8	2	12	9 - 10:59 am	2	Night Owl	1	200
6-8	2	15	7 - 8:59 am	1	Morning person	2	40
3-5	1	16	9 - 10:59 am	2	Morning person	2	200
6-8	2	15	11 am - 12:59 pm	3	Night Owl	1	2.5
6-8	2	15	9 - 10:59 am	2	Night Owl	1	300

The five variables we have are hours slept, credit hours taken, preferred earliest class time, morning or night person, and caffeine intake. The coded values for each variable are:

Hours slept:
1 = 3-5 hours
2 = 6-8 hours
3 = 8+ hours

Preferred Earliest Class Time:
1 = 7-8:59am
2 = 9 - 10:59 am
3 = 11 am - 12:59 pm

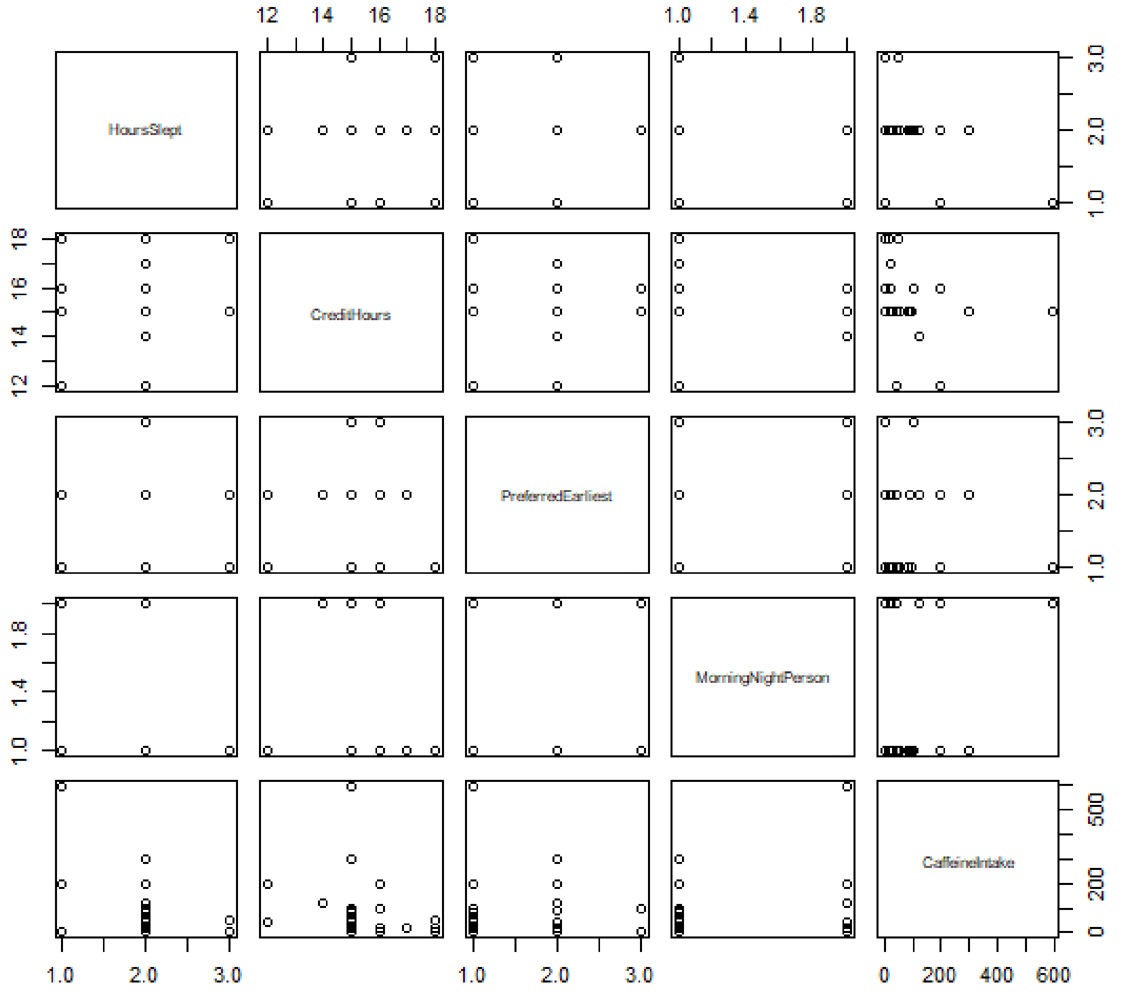
Morning or Night Person
1 = Morning Person
2 = Night Owl

Exploratory Analysis

Below are the scatterplots and correlations between each variable for the Multiple Linear Regression model.

Statistical Inference

We are testing whether hours of sleep, credit hours taken, preferred earliest class time, and if the student is a morning or night person have an effect on the amount of caffeine consumed. The statistical analysis that will be run on this data is **multiple linear regression (MLR).**



```
                  HoursSlept CreditHours PreferredEarliest MorningNightPerson CaffeineIntake
HoursSlept              1.00        0.00              0.18              -0.18          -0.41
CreditHours             0.00        1.00             -0.19              -0.04          -0.31
PreferredEarliest       0.18       -0.19              1.00              -0.07          -0.05
MorningNightPerson     -0.18       -0.04             -0.07               1.00           0.22
CaffeineIntake         -0.41       -0.31             -0.05               0.22           1.00
```

Looking at the scatter plots, there are no obvious high correlations between any of the variables being studied as they are discrete variables. Moreover, the correlation values of all the variables are well below 0.7, therefore showcasing that there is no reason to worry about multicollinearity.

The equation for our model specifically is:

$$\mu_{\text{caffeine}} = \beta_{\text{intercept}} + \beta_{\text{credit hours}} + \beta_{\text{hours of sleep}} + \beta_{\text{earliest class time}} + \beta_{\text{morning or night owl}}$$

Overall ANOVA hypothesis:

H_0: $\beta_{\text{credit hours}} = \beta_{\text{hours of sleep}} = \beta_{\text{earliest class time}} = \beta_{\text{latest class time}} = 0$

H_a: At least one β_j does not equal 0.

Where μy is the population mean for the response variable y, β_0 is the mean response when all x variables equal 0, and each β_j represents the average change in the response variable y whenever the variable x_j increases by a unit if all other variables were held constant.

Our four predictors are hours slept, credit hours taken, preferred earliest class time, and whether the students are morning or night people. Our response variable is the amount of caffeine consumed. Our sample size is $n = 28$ people from the Spring 24 STAT3300 class.

Individual T-test hypotheses:

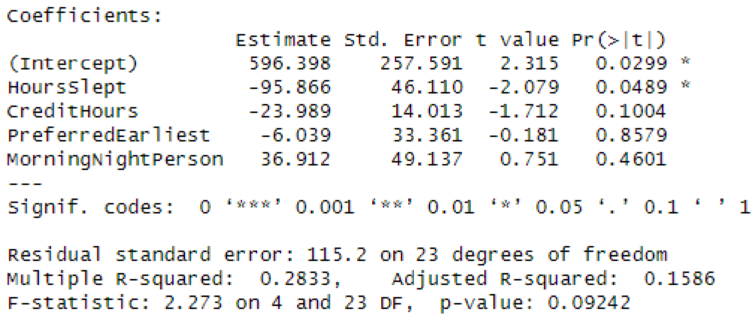
Credit hours: $H_0 : \beta_{creditHours} = 0$ vs $H_a : \beta_{creditHours} \neq 0$

Hours of sleep: $H_0 : \beta_{hoursSlept} = 0$ vs $H_a : \beta_{hoursSlept} \neq 0$

Preferred Earliest Class Time: $H_0 : \beta_{preferredEarliest} = 0$ vs $H_a : \beta_{preferredEarliest} \neq 0$

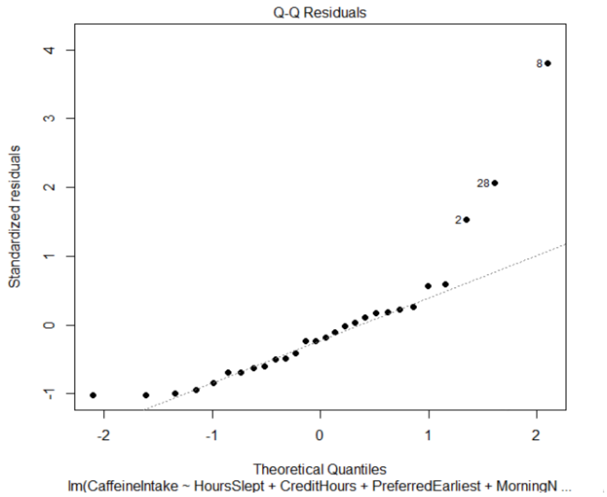
Morning or Night Owl: $H_0 : \beta_{morningNightPerson} = 0$ vs $H_a : \beta_{morningNightPerson} \neq 0$

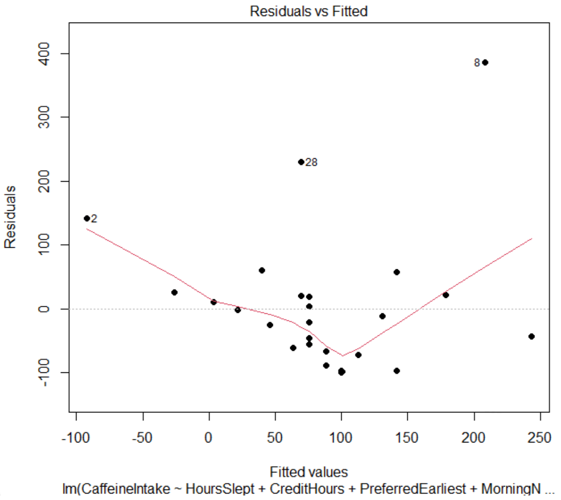
Here is the model containing all the explanatory factors as predictors that was run using R.



```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)  
(Intercept)         596.398    257.591   2.315   0.0299 *
HoursSlept          -95.866     46.110  -2.079   0.0489 *
CreditHours         -23.989     14.013  -1.712   0.1004  
PreferredEarliest    -6.039     33.361  -0.181   0.8579  
MorningNightPerson   36.912     49.137   0.751   0.4601  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 115.2 on 23 degrees of freedom
Multiple R-squared:  0.2833,	Adjusted R-squared:  0.1586 
F-statistic: 2.273 on 4 and 23 DF,  p-value: 0.09242
```

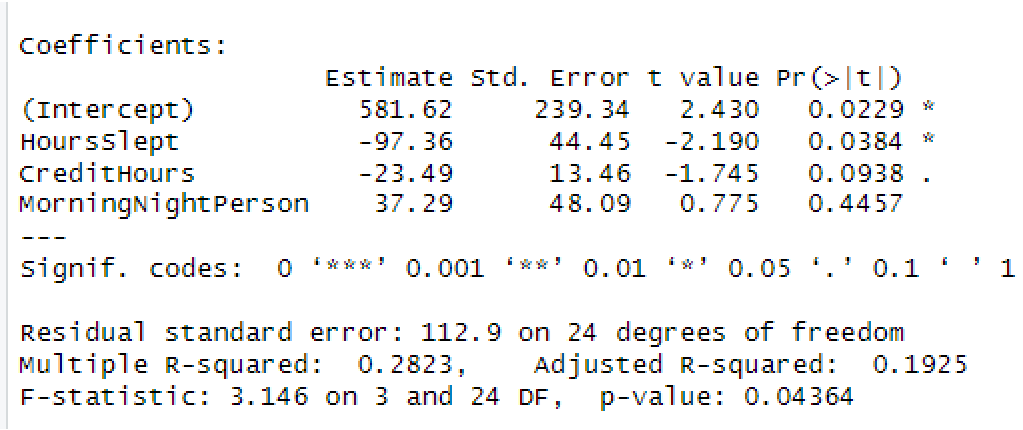




Based on the residual plot, the condition of linearity is not met. There is a curved, exponential line. The independence condition is not met because we attempted a census of our STAT 3300 class that would be considered a convenience sample but is representative of our population so we will continue as if this condition is met. The normality condition is not met as the Q-Q plot is not linear. The equal variance condition is not met as the vertical standard deviations around the line are not constant, and there is a fan shape in the residuals versus fitted plot. This is a preliminary study and the conditions are not met to analyse the data using the linear regression model. However, we are going to continue onward as if the conditions were met. Through this analysis, it was found that the F test was significant at the 0.1 level as it was found to be 2.273 and the p-value is 0.09667, which means there is sufficient evidence to conclude that at least one of the coefficients does not equal zero or that any of the four variables are significant predictors of the amount of caffeine consumed. As the p-value is less than alpha, it is statistically significant, so we reject the null hypothesis.

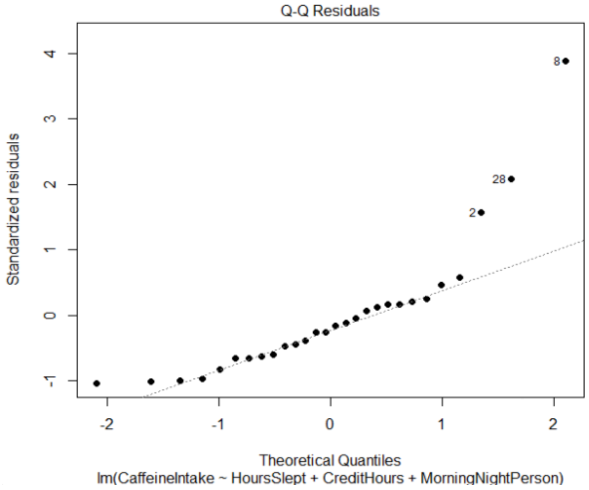
From the individual T-tests, we were able to conclude that the variable hours slept had a significant linear relationship with the amount of caffeine consumed. So, we can reduce this model by removing the preferred earliest class time variable as it has the highest p-value and is not statistically significant.

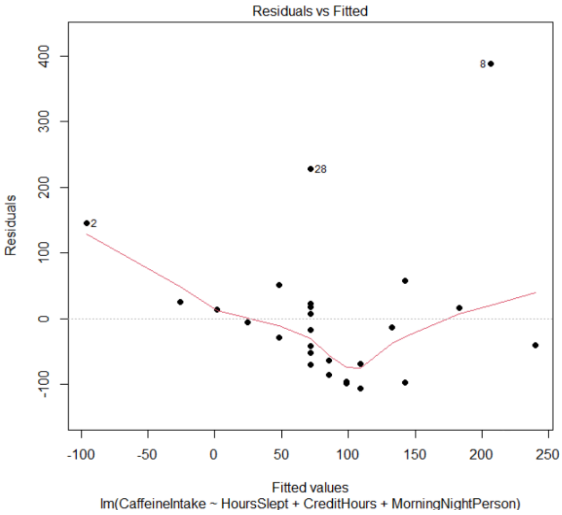
Below is the second model which was run with the remaining three most significant variables - hours slept, credit hours taken, morning or night person.



```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          581.62     239.34   2.430   0.0229 *
HoursSlept           -97.36      44.45  -2.190   0.0384 *
CreditHours          -23.49      13.46  -1.745   0.0938 .
MorningNightPerson    37.29      48.09   0.775   0.4457
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 112.9 on 24 degrees of freedom
Multiple R-squared:  0.2823,    Adjusted R-squared:  0.1925
F-statistic: 3.146 on 3 and 24 DF,  p-value: 0.04364
```

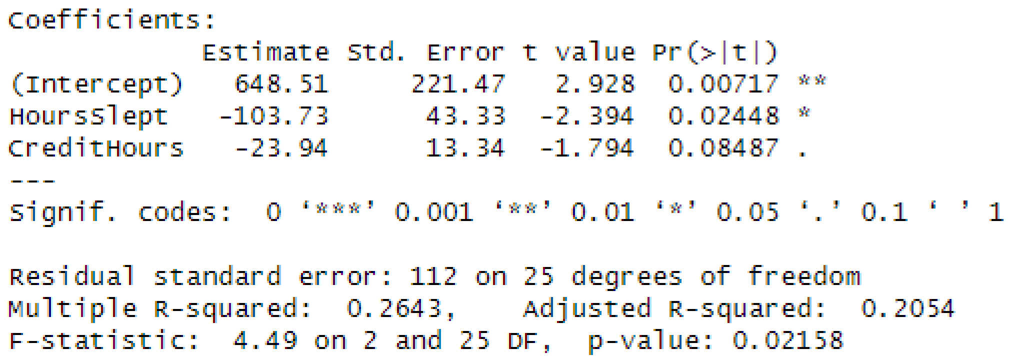




The four conditions of linearity, independence, normality, and equal standard deviation are not met. However, we are going to continue onwards as if the conditions were met as this is a preliminary study. Additionally, we will be removing the morning or night person variable as it has the highest p-value in the predictors in model 2, therefore, it is the least significant.

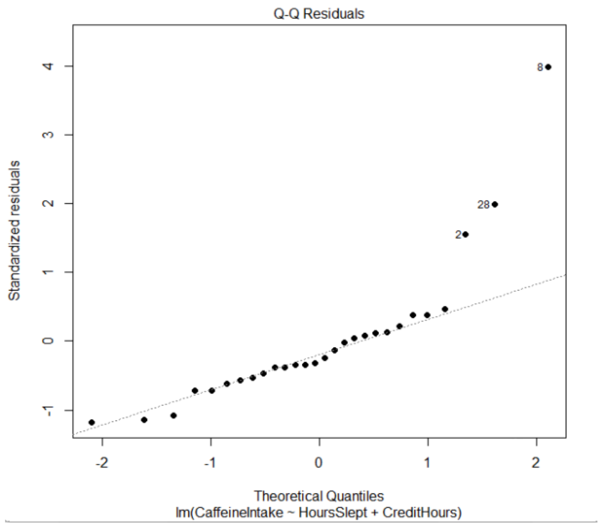
Model 3:

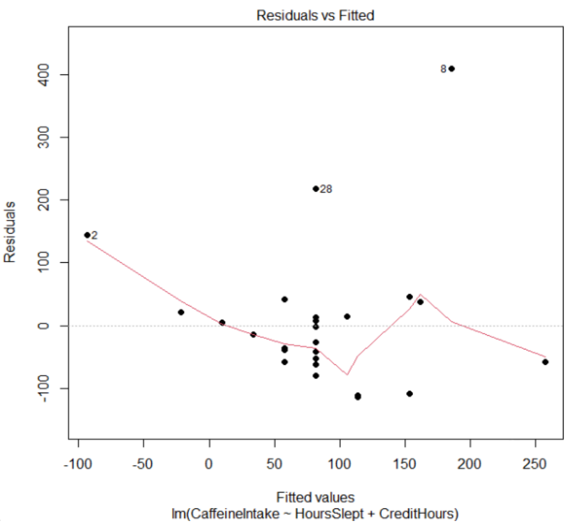
Below are the results from model 3 with the two most significant variables - number of credit hours and hours of sleep.



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)   648.51     221.47   2.928  0.00717 **
HoursSlept   -103.73      43.33  -2.394  0.02448 * 
CreditHours   -23.94      13.34  -1.794  0.08487 . 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 112 on 25 degrees of freedom
Multiple R-squared:  0.2643,	Adjusted R-squared:  0.2054 
F-statistic:  4.49 on 2 and 25 DF,  p-value: 0.02158
```

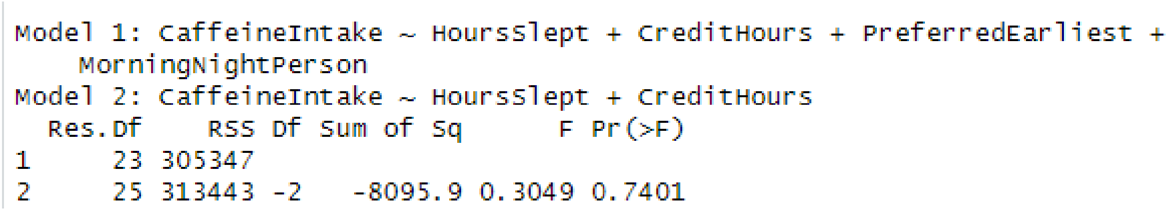




From this model, all of the four conditions are not met. As previously stated, this is a preliminary study and the conditions are not met to analyse the data using the linear regression model so we are going to continue onward as if the conditions were met.

Based on the F and p-value, the factors are statistically significant, there is enough evidence to conclude that hours slept and credit hours have a significant linear relationship with the amount of caffeine consumed.

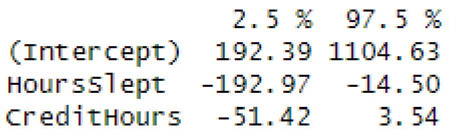
Finally, we will run an ANOVA test to compare the full model (model 1) and model 3 to see which is a better predictor for the amount of caffeine consumed.



```
Model 1: CaffeineIntake ~ HoursSlept + CreditHours + PreferredEarliest +
    MorningNightPerson
Model 2: CaffeineIntake ~ HoursSlept + CreditHours
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     23 305347
2     25 313443 -2   -8095.9 0.3049 0.7401
```

The large p-value demonstrates that there is not enough evidence that the full model is significantly better than the reduced model to predict the amount of caffeine consumed by the students. Therefore, the reduced model is a much more significant predictor to predict the amount of caffeine consumed by students. In cross-validation, the adjusted R-squared value did increase for the reduced model compared to the full model.

Looking at the adjusted R^2 values, the reduced model with the 2 most significant variables, hours slept and credit hours taken has the highest value. Below is the confidence interval for the hours slept and credit hours coefficients from that reduced model. The confidence interval of the coefficients at the alpha = 0.1 level is:



```
              2.5 %  97.5 %
(Intercept)  192.39 1104.63
HoursSlept  -192.97  -14.50
CreditHours  -51.42    3.54
```

The above interval for Hours Slept is (-192.97, -14.50) and for each additional increase in hours slept, holding all other variables constant, the predicted amount of caffeine consumed would decrease by 103.74 milligrams, on average. The above interval for the number of credit hours taken is (-51.42, 3.54) and for each additional increase in the value assigned if the number of credit hours taken increases, the predicted amount of caffeine consumed increases by 23.94 milligrams, on average.

As shown above, this confidence interval for the hours slept predictor does not include zero, thus presenting it as a valuable predictor for predicting the amount of caffeine consumed. The greater the hours slept, the more negative the confidence interval is and the less caffeine consumed, thus making it inversely proportional. However, credit hours taken does contain zero, this indicates that it has no value for predicting the amount of caffeine consumed.

Conclusion:

Based on this data, there is sufficient evidence at the 0.1 level to conclude that hours slept and credit hours were the significant predictors and has a significant linear relationship with the amount of caffeine consumed. Whether the student was a morning or night person and preferred earliest class time were insignificant variables. Since the p-value for these remaining variables are greater than the alpha level of 0.1 it would make these variables not significant; therefore, it would contradict our hypothesis as we fail to reject the null.