Regression Analysis Project Report: Sleep

Exam 3 - Project

Allison Buck, Maria DeMotte, Julia Molinaro

Dr. Zhao

December 4, 2023

Introduction

For this project, we have decided to explore a dataset centered around variables that may affect one's sleep. College students often struggle to get enough sleep at night when they have other stressors and distractions in their lives. We have chosen to explore this topic to understand what factors could be affecting our sleep as we continue through our college journey, as well as possible ways to improve our sleep in the future. Through an analysis of this data, we intend to discover which factors most influence sleep, specifically we want to investigate if anxiety and depression have a large influence on the amount of sleep that college students average each night.

The dataset that we are using for this analysis, "Online Sleep Survey Data", was a voluntary survey found in the source OPENICPSR. The survey was conducted at Appalachian State University throughout several years, and includes 2,218 respondents primarily of the ages 18-22. There is information regarding the respondents demographics, mental health, and self-reported sleep measurements. Of these variables, we have chosen lwsleep as our response variable, and seek to understand what could have affected the average number of hours that a respondent had slept per night the week before they completed the survey.

After analyzing each variables' meaning, we decided to remove the variables mturk, whatdisorder, and chronotype. We chose to remove mturk because we did not have complete confidence in the meaning of the variable, as it was not included in the data description. We removed the chronotype variable because it included redundant information already represented by the rmeq variable. The chronotype variable contained qualitative data regarding how much of a morning or evening person each respondent felt they were based on a scale. The rmeq variable is a score based on a questionnaire about morningness and eveningness, so we felt that including chronotype in our models would not be necessary. Lastly, we removed the whatdisorder variable due to its focus on specific sleep disorders and lengthy descriptions from the respondents. The wordy, open-ended responses made the variable less useful to our data analysis. The use of the sdisorder variable, which is a binary variable that labels whether or not each respondent has a sleep disorder, is sufficient for our analysis in this project. We felt as though solely knowing whether a respondent has a sleep disorder or not was more helpful than having specific descriptions of their disorders.

We needed to make a few transformations and alterations to the dataset to begin our analysis. After looking at the skewness displayed in the histogram of the age vs lwsleep variables, which can be found in the Appendix under the Histograms of Quantitative Variables section, we chose to take a logarithmic transformation of the age variable to make it more normally distributed. We also removed all NA records to focus only on the respondents who provided complete answers under each variable.

Below is the list of variables used from this dataset, as well as their description and datatype.

<u>Response variable</u>

- **Lwsleep**: self-reported average number of hours slept over the last week

<u>Predictor variables</u>

- **Lnsleep**: self-reported number of hours slept last night
- **Depression**: score based on results from PHQ-2 depression questionnaire (0-6 scale)
- **Anxiety**: score based on results from GAD-7 anxiety questionnaire (0-21 scale)
- **Epworth**: score based on results from the Epworth Sleepiness Scale (0-24 scale)
- **Rmeq**: score based on results from the Reduced Morningness-Eveningness questionnaire (4-26 scale)
- **Chronotype**: strong morning-type to strong evening-type scale (with 5 categories) based on self-reported response of how much of a "morning" or "evening" person respondent feels they are

- **Female**: whether or not respondent is female or not (0-no, 1-yes)
- **Minority**: whether or not respondent is a minority or not (0-no, 1-yes)
- **Student**: whether or not respondent is a student or not (0-no, 1-yes)
- **Age**: the age of respondent
- **Optimalsleep**: self-reported number of hours respondent desires to sleep each night for optimal alertness, performance, and functionality
- **Sdisorder**: whether or not respondent has a sleeping disorder or not (0-no, 1-yes)

Regression Analysis (Model Selection & Validation)

After cleaning our data to make it as useful as possible for our analysis, we started building models using a variety of concepts that we have learned throughout our experience with regression analysis thus far. Our first model (Model 1) is a full model that includes every variable listed in the introduction above with the exception of age being replaced with a logarithmic transformation of itself. We felt that it was important to start with a model that included all of the variables that we deemed important from our initial understanding of this dataset. After producing a summary of this model, we discovered many important measures that helped us to create better models. The coefficient of determination, R-squared, revealed that 43.36% of the variation in the data can be explained by Model 1. Also, the s-value of 0.9131 represents the deviation from Model 1's regression line. The lower that this measure is, the better the model. This model also produced an F-statistic of 153.5 which can be used for a hypothesis test determining the overall significance of all of the variables combined in Model 1. When taking a closer look at the variables in Model 1 individually, the t-values, p-values, and significance codes corresponding to each of them helped us to determine their significance. The only variables from this initial model that had high t-values and low p-values were lnsleep, anxiety, epworth, rmeq, and optimalsleep. The variables lnage, student, female, and minority seem to lack significance in this model based on both the Boxplot section of the Appendix and the summary measures. This may be because this data was collected at a university in which most respondents were students in the age range of 18-25. The age of a respondent and whether or not they are a student becomes less significant when the majority of respondents are similar in these categories. The female and minority variables show that gender and being a minority are not necessarily factors that largely affect sleep. It was surprising to us that the depression and sdisorder variables did not have more significance in the full model. Logically, we would assume that having depression or a sleep disorder would affect one's sleep patterns.

From our initial understanding of this data, we assumed that anxiety would have an effect on sleep. When performing a hypothesis test on this model to see if this is true, we did a t-test on the anxiety variable. This t-test can be found in the Appendix under T-test for Anxiety. The null hypothesis says that anxiety is not important to the model's performance, and the alternative hypothesis is that it is important. The t-value corresponding to anxiety from Model 1 is -6.472, and since the absolute value of this is greater than the t-statistic of 2.578, it falls in the rejection region. We can reject the null hypothesis and accept that the anxiety variable is important to Model 1.

When selecting and building models we considered topics such as added variable plots, identification of influential points, collinearity between variables, etc. We decided, based on our unique dataset, that some of these techniques were unnecessary for our analysis while others were important to the development of our models. To look for improvements to our initial model we wanted to see if examining any variables individually with an added variable plot was necessary. An added variable plot is used to show the correlation between a predictor variable and the target holding all other predictors constant. This plot is used when the correlation coefficient for a specific predictor variable is not what we would expect. However, when we examined each of the individual correlations with the target, they all made sense logically. Therefore it was not necessary to explore any variables using this method. The table of correlation coefficients can be found in the Appendix under Correlation Coefficients. We also wanted to look for outliers and influential points. However, when we examined the histograms of all of the quantitative variables along with the boxplots corresponding to each categorical variable, we did not identify any unusual points. All of the plots appeared relatively normal except for age, which we transformed to lnage. The different categories in the tables were all influential to the model as well. Based on these observations, it was not necessary to perform any analysis because no influential points were identified.

After analyzing the p-values, t-values, and significance codes corresponding to each variable in the summary of Model 1, we were able to identify the most significant variables. We then built a second model (Model 2) that includes only the variables that were significant in Model 1. These variables are lnsleep, anxiety, rmeq, epworth, and optimal sleep. Narrowing down the full model to only its significant variables allowed us to see that some of the other variables have little to no influence on the model's performance. Through a comparison of the summary results between Model 1 and Model 2, it is clear that not much changes between the two models' statistics. The R-squared value decreases from 0.4336 to 0.432 from Model 1 to Model 2, but it is important to remember that R-squared does not consider redundancy. When looking at the adjusted R-squared values from each model, 0.4308 and 0.4307, it is clear that the models are almost identical when it comes to explaining variation within the data. Also, the s-values are very close in Model 1 to Model 2. The only big difference between these two models is the F-statistic which increases from 153.5 to 336.5. This jump shows that Model 2 is a better fit than a null model and that each variable's influence on the response is more significant than in Model 1.

Our third model (Model 3) was created based on the forward stepwise selection process. Similarly, the fourth model (Model 4) was created by applying the backward stepwise selection process in our code. Model 4 ended up being identical to Model 3. The forward stepwise selection process fits models by adding one variable at a time starting with the null. First, out of all of the variables, the most statistically significant one is added. Then, this process is repeated with a new variable being added to the model each iteration based on its statistical significance. The downfall of forward stepwise is that once a predictor is added to the model, it stays there forever. In contrast, the backward stepwise selection process fits models by removing the least useful predictors from the full model one at a time. The downfall of backward stepwise is that once a variable is removed from the model, it is gone forever. During the process, we decided to use AIC as our statistic for model comparison and to tell us which model performed the best. We knew that AIC was not as penalizing as BIC and would in turn allow us to keep more variables in our model. When comparing Models 3 and 4 to Model 2, adjusted R-squared increases from 0.4307 to 0.431, s decreases slightly from 0.9132 to 0.9129, and the F-statistic decreases from 336.5 to 280.9. It seems that the models produced by stepwise selection were not much better than Model 2.

To create Model 5 we decided to look into interaction terms. We considered the possibility of an interaction between two variables that could potentially be beneficial for us to incorporate as a new variable in our full model. From our real-life encounters with people and statistics surrounding mental health disorders, people who suffer from anxiety often have depression as well, and vice versa. Knowing this, we thought that it may be interesting to see how an interaction term between these two variables would affect our model. Does having both anxiety and depression affect one's sleep more significantly than if someone just had one or the other? To see our process for creating anxiety*depression, reference the Anova Table for Interaction Term section in the Appendix. From this analysis, we can conclude that the interaction term would be beneficial to add to the model. For our fifth model (Model 5), we added the interaction term of anxiety and depression to Model 1. Model 5 resulted in improvements across all summary measures except for the F-statistic. It has an R-squared of 0.4364 and an adjusted R-squared of 0.4334 which are both higher than any of the models we have made prior. Model 5 also has the lowest s-value of 0.9111. Its F-statistic did decrease slightly from Model 1 which was 153.5 to Model 5 which was 142.3. Overall, it seems as though adding the interaction term to the full model caused less variation within the model and more explanation of the variation. It is also important to mention that while the depression variable was not initially significant in Model 1, adding the interaction term caused it to gain some significance. A respondent who has both anxiety and depression may have a different number of hours slept on average than a respondent with just depression.

Based on the performance of Model 5, we decided to create a sixth model (Model 6) that adds the depression variable, as well as the interaction term mentioned above, to Model 2. Since Model 2 contains only significant variables from the full model, and we have discovered that the interaction between anxiety and depression is significant, we wanted a model that combined the two. All of the variables in Model 6 were significant. The depression variable gained significance from Models 1 and 5. The summary measures for Model 6 were very comparable to Model 5 with the exception of F-statistic. The F-statistic increased from 142.3 in Model 5 to 243.1 in Model 6. A higher F-statistic equates to a better model, and because all of the other measures are similar to those for Model 5, Model 6 appears to be the best model. The regression line that results from Model 6 gives us the best understanding of our data.

After the creation of our models, we ran Anova tests to confirm that Model 6 was the best fit for our data. The first Anova test we did was between Model 1 and Model 2 to test if our simpler model (Model 2) was a better fit than the full model (Model 1). The p-value from the output was 0.3976 which is not significant. This confirms that Model 2 is stronger than the full model. Our next Anova test was between Model 5 and Model 6. These models are the same as Model 1 and Model 2 but they include the interaction term anxiety*depression which we found to be an influential variable when added to the model. The p-value from the test is 0.3575 which is also not significant. Since the p-value was not lower than alpha, the simpler model (Model 6) performs better than the full model (Model 5). Finally, we did an Anova test on Model 2 and Model 6 which are our 2 strongest models to confirm that the interaction term was necessary to include. The p-value 0.00296 is significant at the 0.01 level indicating that the complex model is significantly better than the simpler model. For our analysis, this means that adding the interaction term positively influenced the performance of our model, and confirmed that Model 6 is our strongest model.

It was important for us to interpret the relationships between each predictor and the response variable, lwsleep. We chose to do this for Model 6 because it turned out to be our best-performing model. For every one hour increase in lnsleep, lwsleep increases by approximately 0.4152 hours. For every one hour increase in optimalsleep, lwsleep increases by approximately 0.2639 hours. Both of these relationships make sense. Lnsleep and lwsleep are very similar variables because they represent average hours of sleep for different periods. Also, it is understandable that the value for optimalsleep has a large influence on average hours slept because people will try to sleep for the amount of time that they think is best. For every one scale increase in anxiety, lwsleep decreases by approximately 0.0492 hours. The coefficient for anxiety is logical as it reveals a negative relationship with lwsleep, so people with more anxiety get less sleep on average. For every one scale increase in rmeq, lwsleep increases by 0.0184 hours. A low score on the reduced Morningness-Eveningness Questionnaire corresponds to an evening person who stays up later and often struggles to wake up in the morning. When someone has a higher score on this questionnaire it means they are more of a morning person who goes to sleep earlier and wakes up easily. Based on this model, morning people are more likely to get more sleep than evening people. For every one scale increase in epworth, lwsleep decreases by 0.0197 hours. A low score on the Epworth Sleepiness Scale corresponds to feeling abnormally sleepy throughout the day, and a high score means that the respondent is not abnormally sleepy. A person’s average number of hours slept in a week decreases when they are not abnormally sleepy throughout the day. For every one scale increase in depression, lwsleep decreases by 0.0556 hours. However, in some of the other models, depression had a positive coefficient. This makes slightly more sense because oversleeping can be a symptom of depression in some instances. On the other hand, people with depression may experience fragmented sleep leading, to less sleep throughout the night. How depression affects sleep appears to be more complex than we originally expected. The coefficient for the interaction term, anxiety*depression, is 0.0084. For every one increase in the interaction term, lwsleep increases by about 0.0084.

To test for collinearity we examined the Variance Inflation Factor for each variable in Model 6. We did this for Model 6 because we found that this model resulted in the best summary measures out of all of the other models combined. The interaction term anxiety*depression had the highest VIF of 7.08, and the rest of the variables had VIFs that were less than 5. Since a VIF of above 10 indicates that there is high collinearity, there was nothing to be concerned about.

Discussion and Limitations

While completing our regression analysis, we encountered limitations within the data, including the lack of diversity within the age category. Due to the small amount of staff and students older than the average collegiate age, it is difficult to distinguish whether or not students get less sleep than non-students. To solve this issue, the data could be collected from a pool of only students who are enrolled at the university. This would allow us to avoid the skewness caused by respondents of older ages. Another solution to this issue could include opening the survey up to people who do not attend or work at the university, while also asking if they are a student. This would make the student variable more useful and allow us to better understand the effect that being a student has on sleep. It is also important to note that this survey was conducted from a singular university, so it may not represent the population of students as a whole, as other factors of the specific university that were not recorded may affect the amount of sleep. Many of the variables in this dataset were self-reported, so we cannot be sure of the accuracy of the subjective responses.

Conclusion and Future Work

After conducting our regression analysis, we have concluded the Model 6 resulted in the best fitting model for our chosen dataset. The R-squared value for this model is 0.435, meaning that 43.5% of the variability can be explained by the regression line lwsleep = 1.988 + .042(lnsleep) +.264(optimalsleep) -.049(anxiety) + .018(rmeq) -.020(epworth) -.056(depression) + .008(anxiety*depression). Due to the Anova tests, we were also able to confirm that throughout the progression of our analysis, the models that we created continued to improve and ultimately led us to our final model. With this, we can see that adding the interaction term to Model 5 further improved the F-statistic and overall performance. We recognize that lnsleep and optimalsleep have the most influence on lwsleep, as well as anxiety, which each confirm our intuition. We would expect sdisorder (whether or not the subject has a sleep disorder) to have a greater influence on lwsleep, as this would seem to impact the quality and length of sleep. Seen by the t-test conducted, we were able to reject the null hypothesis that anxiety is not important to the model, and deduce that anxiety has a large influence.

In a future analysis of this subject, it would be beneficial to have more information regarding the respondents that could also affect their sleep patterns. These could include caffeine intake, substance use, or exercise. From personal experience, we understand that these factors could affect the ability to relax before sleep, and would be curious to confirm our intuition on these variables. We also believe that conducting this survey over a period of many years could improve our understanding of the effects of stress and anxiety, and in turn how they affect one’s sleep. It would have been interesting to see the differences in stress levels during years of economic struggle, the COVID-19 pandemic, and times of high unemployment rates. These factors could result in higher depression and anxiety levels, and their correlation to sleep could be more apparent.

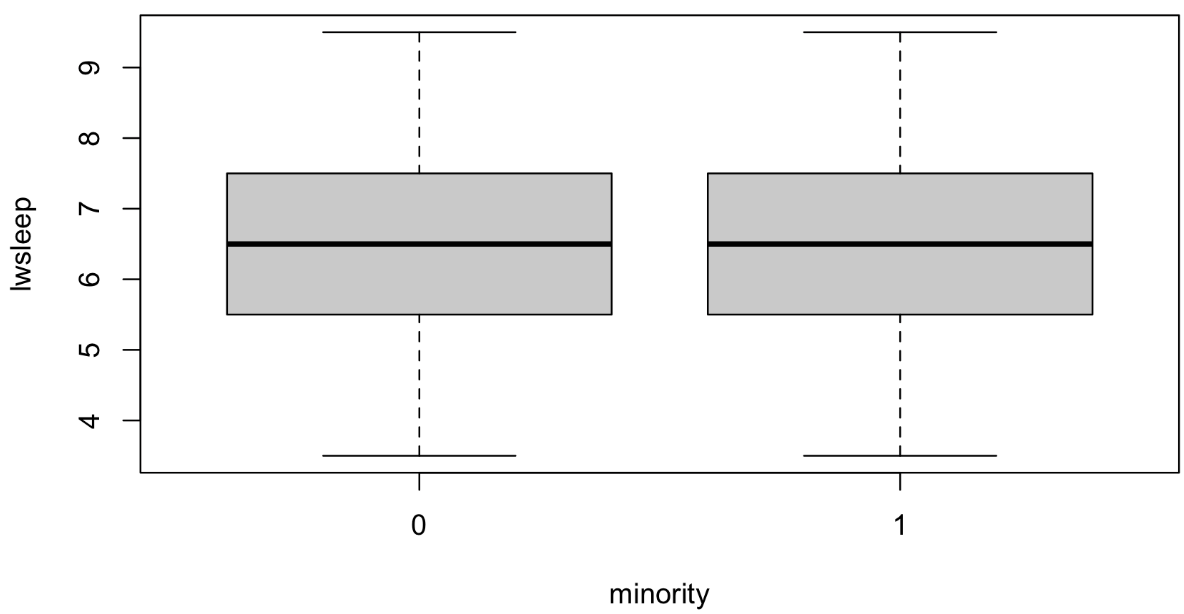
Overall, we were able to learn a lot about factors that are taken into consideration when analyzing the quality and length of sleep that those attending and working at a university get. It was interesting to see the progression of models throughout our analysis. We enjoyed diving deeper into the topic of sleep while being a college student, which is a part of our lives that we strive to improve.

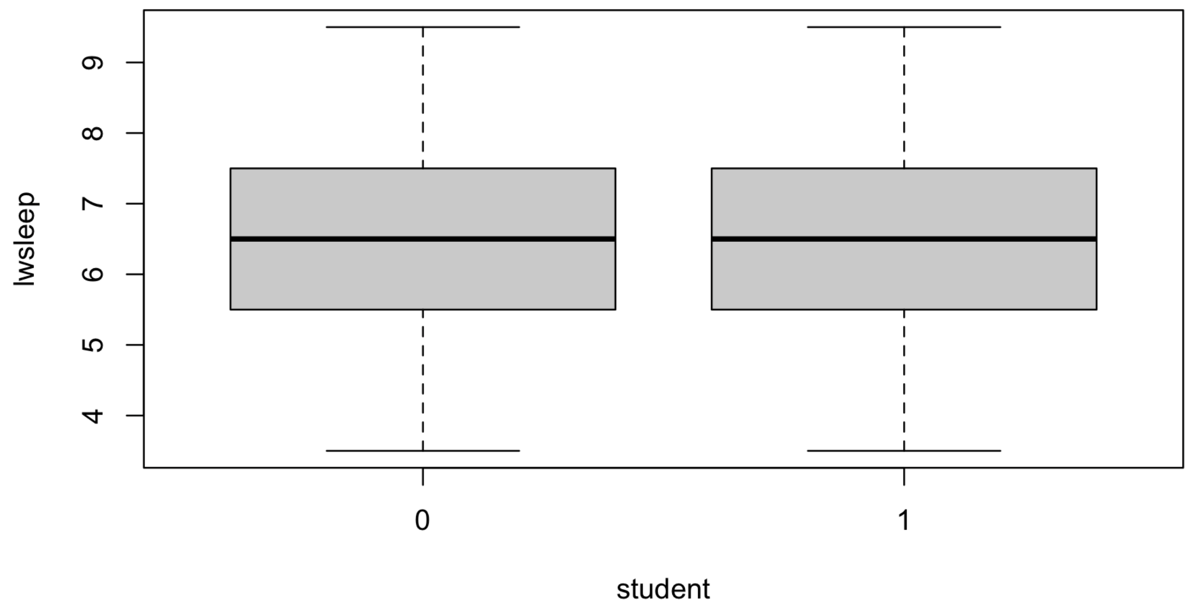
Appendix

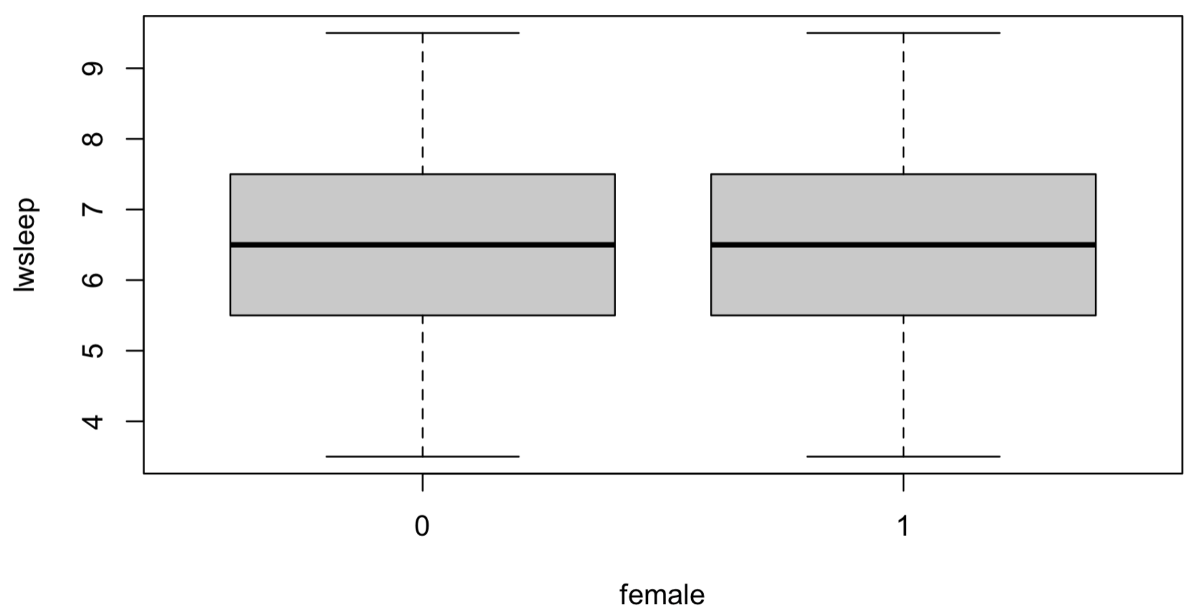
Correlation Coefficients:

	lnsleep	depression	anxiety	epworth	rmeq	female	minority	student	age	optimalsleep	lnage	lwsleep
lnsleep	1.00000000	-0.10414392	-0.188030636	-0.09249110	0.10346857	0.02161144	-0.027992384	0.063548089	-0.06631162	0.242765353	-0.070026091	0.60322239
depression	-0.10414392	1.00000000	0.611869053	0.15076977	-0.13908724	0.01104358	0.013837457	0.033466864	-0.03138730	0.022230756	-0.031098210	-0.12457072
anxiety	-0.18803064	0.61186905	1.000000000	0.17556475	-0.13126700	0.11607197	0.006686886	0.030248743	-0.01609646	0.022373514	-0.009219781	-0.22858213
epworth	-0.09249110	0.15076977	0.175564746	1.00000000	-0.07664575	0.08781532	0.067867292	0.060526789	-0.04379889	0.073526816	-0.042844569	-0.11361673
rmeq	0.10346857	-0.13908724	-0.131266996	-0.07664575	1.00000000	0.08404108	-0.029704605	-0.154778194	0.20952575	-0.069234478	0.207378168	0.11611291
female	0.02161144	0.01104358	0.116071971	0.08781532	0.08404108	1.00000000	0.040826157	-0.078934007	0.03645431	0.060107137	0.031065612	0.02963784
minority	-0.02799238	0.01383746	0.006686886	0.06786729	-0.02970460	0.04082616	1.000000000	0.056111825	-0.06007822	-0.060672377	-0.061646372	-0.04715386
student	0.06354809	0.03346686	0.030248743	0.06052679	-0.15477819	-0.07893401	0.056111825	1.000000000	-0.59981292	-0.005224733	-0.585146500	0.01406539
age	-0.06631162	-0.03138730	-0.016096455	-0.04379889	0.20952575	0.03645431	-0.060078215	-0.599812919	1.00000000	-0.023691437	0.991833442	-0.03417686
optimalsleep	0.24276535	0.02223076	0.022373514	0.07352682	-0.06923448	0.06010714	-0.060672377	-0.005224733	-0.02369144	1.000000000	-0.028530689	0.34917468
lnage	-0.07002609	-0.03109821	-0.009219781	-0.04284457	0.20737817	0.03106561	-0.061646372	-0.585146500	0.99183344	-0.028530689	1.000000000	-0.03771075
lwsleep	0.60322239	-0.12457072	-0.228582132	-0.11361673	0.11611291	0.02963784	-0.047153864	0.014065389	-0.03417686	0.349174682	-0.037710751	1.00000000

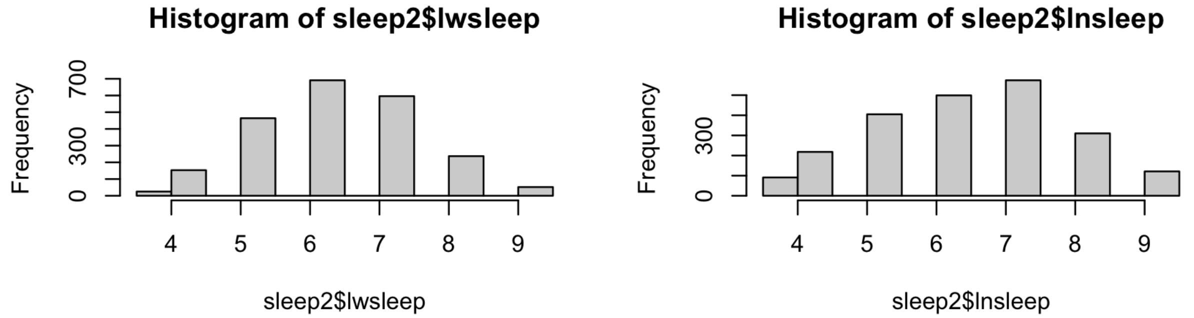
Box Plot:

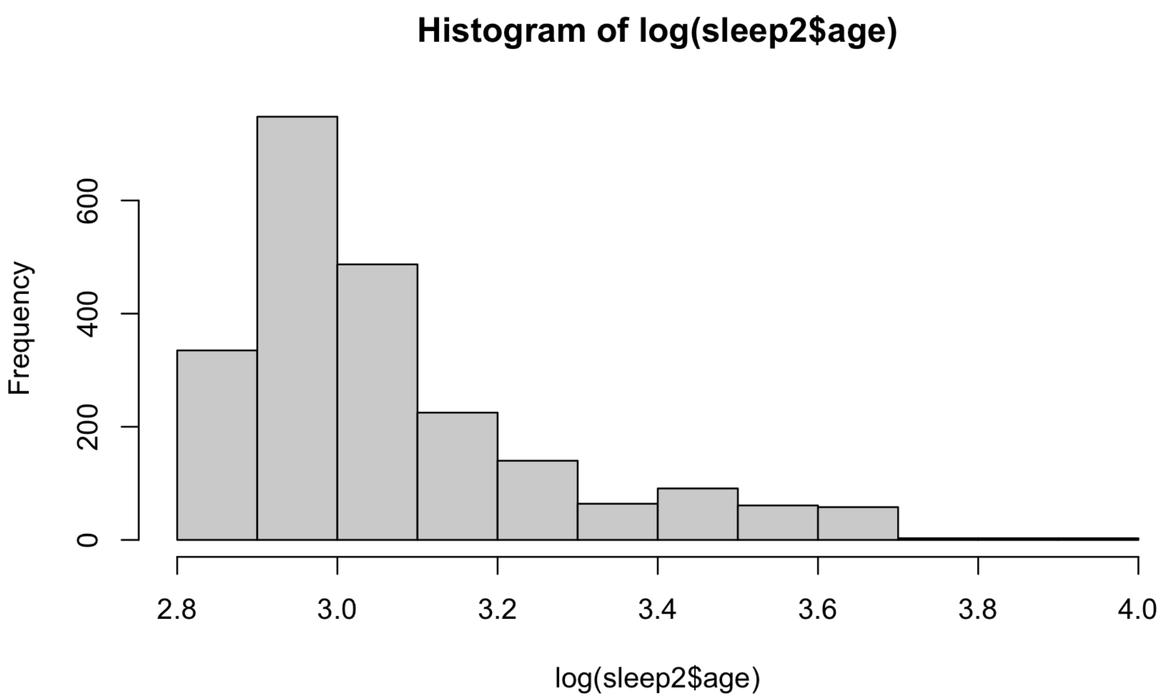



lwsleep
9
8
7
6
5
4
0
1
student


lwsleep
9
8
7
6
5
4
0
1
female

Histograms of Quantitative Variables:





Structure of Dataset:

```
tibble [2,218 × 14] (S3: tbl_df/tbl/data.frame)
 $ lwsleep     : num [1:2218] 6.5 6.5 9.5 6.5 7.5 6.5 7.5 5.5 8.5 7.5 ...
 $ lnsleep     : num [1:2218] 7.5 7.5 8.5 5.5 7.5 6.5 6.5 6.5 3.5 9.5 ...
 $ depression  : num [1:2218] 1 0 0 0 5 0 0 5 3 2 ...
 $ anxiety     : num [1:2218] 4 9 4 0 16 8 5 14 11 1 ...
 $ epworth     : num [1:2218] 7 7 3 7 11 6 10 4 2 4 ...
 $ rmeq        : num [1:2218] 14 5 16 12 10 14 15 8 4 19 ...
 $ chronotype  : chr [1:2218] "Intermediate-type" "Strong Evening-type" "Intermediate-type" "Intermediate-type" ...
 $ female      : num [1:2218] 0 0 0 0 0 0 1 1 1 1 ...
 $ minority    : num [1:2218] 0 0 0 0 0 0 0 0 0 0 ...
 $ student     : num [1:2218] 1 1 1 1 1 1 1 1 0 1 ...
 $ age         : num [1:2218] 19 21 24 19 23 28 19 19 34 22 ...
 $ optimalsleep: num [1:2218] 9 7.5 6 8 8 7.5 8.5 8 14 8.5 ...
 $ sdisorder   : num [1:2218] 0 0 0 0 0 0 0 0 1 0 ...
 $ lnage       : num [1:2218] 2.94 3.04 3.18 2.94 3.14 ...
 - attr(*, "na.action")= 'omit' Named int [1:992] 2219 2220 2221 2222 2223 2224 2225 2226 2227 2228 ...
  ..- attr(*, "names")= chr [1:992] "2219" "2220" "2221" "2222" ...
```

Summary of Dataset:

```
    lwsleep         lnsleep      depression       anxiety         epworth          rmeq        chronotype
 Min.   :3.500   Min.   :3.5   Min.   :0.000   Min.   : 0.000   Min.   : 0.00   Min.   : 4.0   Length:2218
 1st Qu.:5.500   1st Qu.:5.5   1st Qu.:0.000   1st Qu.: 3.000   1st Qu.: 5.00   1st Qu.: 9.0   Class :character
 Median :6.500   Median :6.5   Median :1.000   Median : 6.000   Median : 8.00   Median :12.0   Mode  :character
 Mean   :6.672   Mean   :6.7   Mean   :1.318   Mean   : 6.469   Mean   : 8.05   Mean   :12.2
 3rd Qu.:7.500   3rd Qu.:7.5   3rd Qu.:2.000   3rd Qu.: 9.000   3rd Qu.:10.00   3rd Qu.:15.0
 Max.   :9.500   Max.   :9.5   Max.   :6.000   Max.   :21.000   Max.   :23.00   Max.   :24.0
     female          minority          student            age         optimalsleep      sdisorder           lnage
 Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :18.00   Min.   : 0.600   Min.   :0.00000   Min.   :2.890
 1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:1.0000   1st Qu.:19.00   1st Qu.: 7.500   1st Qu.:0.00000   1st Qu.:2.944
 Median :1.0000   Median :0.00000   Median :1.0000   Median :21.00   Median : 8.000   Median :0.00000   Median :3.045
 Mean   :0.6596   Mean   :0.06132   Mean   :0.8963   Mean   :22.25   Mean   : 8.048   Mean   :0.02976   Mean   :3.082
 3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:23.00   3rd Qu.: 9.000   3rd Qu.:0.00000   3rd Qu.:3.135
 Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :53.00   Max.   :14.000   Max.   :1.00000   Max.   :3.970
```

Model 1 Summary:

```
Call:
lm(formula = lwsleep ~ lnsleep + depression + anxiety + epworth +
    rmeq + female + minority + student + optimalsleep + sdisorder +
    lnage, data = sleep2)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5145 -0.5889  0.0094  0.5462  4.0247

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.248273   0.466818   4.816 1.56e-06 ***
lnsleep       0.413942   0.013924  29.729  < 2e-16 ***
depression    0.020311   0.016885   1.203 0.229141
anxiety      -0.035983   0.005560  -6.472 1.19e-10 ***
epworth      -0.020066   0.005507  -3.644 0.000275 ***
rmeq          0.018254   0.005228   3.491 0.000490 ***
female        0.052128   0.041916   1.244 0.213766
minority     -0.071488   0.081470  -0.877 0.380326
student      -0.031782   0.079011  -0.402 0.687544
optimalsleep  0.261998   0.018734  13.985  < 2e-16 ***
sdisorder    -0.165143   0.115149  -1.434 0.151666
lnage        -0.100590   0.126390  -0.796 0.426191
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9131 on 2206 degrees of freedom
Multiple R-squared:  0.4336,    Adjusted R-squared:  0.4308
F-statistic: 153.5 on 11 and 2206 DF,  p-value: < 2.2e-16
```

Model 2 Summary:

```
Call:
lm(formula = lwsleep ~ lnsleep + anxiety + rmeq + epworth + optimalsleep,
    data = sleep2)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.6245 -0.5913  0.0090  0.5448  4.0384 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1.898199   0.177264  10.708  < 2e-16 ***
lnsleep       0.416327   0.013821  30.123  < 2e-16 ***
anxiety      -0.031852   0.004432  -7.187    9e-13 ***
rmeq          0.018404   0.005056   3.640 0.000279 ***
epworth      -0.019350   0.005461  -3.543 0.000403 ***
optimalsleep  0.263626   0.018647  14.138  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9132 on 2212 degrees of freedom
Multiple R-squared:  0.432,	Adjusted R-squared:  0.4307 
F-statistic: 336.5 on 5 and 2212 DF,  p-value: < 2.2e-16
```

Model 3 Summary:

```
Call:
lm(formula = lwsleep ~ lnsleep + optimalsleep + anxiety + rmeq +
    epworth + sdisorder, data = sleep2)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4601 -0.5854  0.0107  0.5458  4.0337

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.907586   0.177338  10.757  < 2e-16 ***
lnsleep       0.415439   0.013831  30.037  < 2e-16 ***
optimalsleep  0.264357   0.018649  14.175  < 2e-16 ***
anxiety      -0.031431   0.004440  -7.079 1.94e-12 ***
rmeq          0.017931   0.005065   3.540 0.000409 ***
epworth      -0.019511   0.005461  -3.573 0.000361 ***
sdisorder    -0.166849   0.114885  -1.452 0.146557
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9129 on 2211 degrees of freedom
Multiple R-squared:  0.4326,    Adjusted R-squared:  0.431
F-statistic: 280.9 on 6 and 2211 DF,  p-value: < 2.2e-16
```

Model 4 Summary:

```
Call:
lm(formula = lwsleep ~ lnsleep + anxiety + epworth + rmeq + optimalsleep +
    sdisorder, data = sleep2)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4601 -0.5854  0.0107  0.5458  4.0337

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.907586   0.177338  10.757  < 2e-16 ***
lnsleep       0.415439   0.013831  30.037  < 2e-16 ***
anxiety      -0.031431   0.004440  -7.079 1.94e-12 ***
epworth      -0.019511   0.005461  -3.573 0.000361 ***
rmeq          0.017931   0.005065   3.540 0.000409 ***
optimalsleep  0.264357   0.018649  14.175  < 2e-16 ***
sdisorder    -0.166849   0.114885  -1.452 0.146557
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9129 on 2211 degrees of freedom
Multiple R-squared:  0.4326,    Adjusted R-squared:  0.431
F-statistic: 280.9 on 6 and 2211 DF,  p-value: < 2.2e-16
```

Model 5 Summary:

```
Call:
lm(formula = lwsleep ~ lnsleep + depression + anxiety + epworth +
    rmeq + female + minority + student + optimalsleep + sdisorder +
    lnage + depression * anxiety, data = sleep2)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4491 -0.5854  0.0111  0.5517  3.9311

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.329580   0.466421   4.995 6.36e-07 ***
lnsleep             0.412971   0.013896  29.719  < 2e-16 ***
depression         -0.055070   0.028357  -1.942 0.052257 .
anxiety            -0.050160   0.007013  -7.153 1.15e-12 ***
epworth            -0.020144   0.005495  -3.666 0.000252 ***
rmeq                0.017908   0.005218   3.432 0.000610 ***
female              0.053790   0.041825   1.286 0.198557
minority           -0.075084   0.081295  -0.924 0.355794
student            -0.023892   0.078870  -0.303 0.761972
optimalsleep        0.262267   0.018692  14.031  < 2e-16 ***
sdisorder          -0.175993   0.114938  -1.531 0.125862
lnage              -0.098728   0.126108  -0.783 0.433781
depression:anxiety  0.008385   0.002537   3.305 0.000966 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9111 on 2205 degrees of freedom
Multiple R-squared:  0.4364,    Adjusted R-squared:  0.4334
F-statistic: 142.3 on 12 and 2205 DF,  p-value: < 2.2e-16
```

Model 6 Summary:

```
Call:
lm(formula = lwsleep ~ lnsleep + optimalsleep + anxiety + rmeq +
    epworth + depression + anxiety * depression, data = sleep2)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5913 -0.5846  0.0149  0.5539  3.9026

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.987869   0.179133  11.097  < 2e-16 ***
lnsleep             0.415234   0.013797  30.097  < 2e-16 ***
optimalsleep        0.263917   0.018607  14.184  < 2e-16 ***
anxiety            -0.049242   0.006957  -7.078 1.95e-12 ***
rmeq                0.018448   0.005060   3.646 0.000273 ***
epworth            -0.019698   0.005457  -3.610 0.000313 ***
depression         -0.055630   0.028329  -1.964 0.049691 *
anxiety:depression  0.008210   0.002535   3.239 0.001217 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9112 on 2210 degrees of freedom
Multiple R-squared:  0.435,     Adjusted R-squared:  0.4332
F-statistic: 243.1 on 7 and 2210 DF,  p-value: < 2.2e-16
```

Table for Model Comparison:

Model	R^2	Adjusted R^2	s	F-statistic
Model 1	0.4336	0.4308	0.9131	153.5
Model 2	0.432	0.4307	0.9132	336.5
Model 3	0.4326	0.431	0.9129	280.9
Model 4	0.4326	0.431	0.9129	280.9
Model 5	0.4364	0.4334	0.9111	142.3
Model 6 *	0.435	0.4332	0.9112	243.1

Anova Table for Model 1 and Model 2:

```
Analysis of Variance Table

Model 1: lwsleep ~ lnsleep + depression + anxiety + epworth + rmeq + female +
    minority + student + optimalsleep + sdisorder + lnage
Model 2: lwsleep ~ lnsleep + anxiety + rmeq + epworth + optimalsleep
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   2206 1839.3
2   2212 1844.5 -6   -5.1995 1.0393 0.3976
```

Anova Table for Model 5 and Model 6:

```
Analysis of Variance Table

Model 1: lwsleep ~ lnsleep + anxiety + rmeq + epworth + optimalsleep
Model 2: lwsleep ~ lnsleep + optimalsleep + anxiety + rmeq + epworth +
    depression + anxiety * depression
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1   2212 1844.5
2   2210 1834.8  2    9.6936 5.8378 0.00296 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Anova Table for Model 2 and Model 6:

```
Analysis of Variance Table

Model 1: lwsleep ~ lnsleep + anxiety + rmeq + epworth + optimalsleep
Model 2: lwsleep ~ lnsleep + optimalsleep + anxiety + rmeq + epworth +
    depression + anxiety * depression
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1   2212 1844.5
2   2210 1834.8  2    9.6936 5.8378 0.00296 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Anova Table for Interaction Term:

```
Analysis of Variance Table

Model 1: lwsleep ~ anxiety + depression
Model 2: lwsleep ~ anxiety + depression + anxiety * depression
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1   2215 3076.7
2   2214 3064.1  1    12.548 9.0668 0.002632 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

T-test for Anxiety:

````
```{r}
n <- 2218
k <- 11
alpha <- 0.01
qt(1-(alpha/2), n-k-1)
```
````

```
[1] 2.57806
```

References

Dickinson, David. Online Sleep Survey Data. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2016-12-13. https://doi.org/10.3886/E100375V1; Survey completed by respondents: https://docs.google.com/document/d/1KoQqhyzSpy0TJ8qs14VO2s0QfI7HL1nD/edit?usp=sharing&ouid=104495783261358379335&rtpof=true&sd=true