603 Quant Final

Research Question

How does cumulative GPA and sleep time predict term GPA in first year university students?

Sleep is an essential component for cognitive processes such as problem-solving, memory, and attention. Previous literature has established that sleeping is necessary to strengthen learning and memory through the conversion of short-term information into long-term (Hershner, 2020). Most notably, lack of sleep has been extensively studied in adcademic settings and students and have been found to be associated with impairment domains like attention, learning, and decision-making (Chen & Chen, 2019b, Hysing et al., 2016). Therefore, sleep factors may negatively impact their performance in school. In the last few decades, scholars have investigated the relationships and factors associated with academic performance and sleep. For example, in 2013, Biswas found that students who reported inadequate sleep (specificed as 6 or less hours of sleep) had significantly lower academic performance, as measure by their GPA, than their peers who slept the recommended 7-9 hours. Many studies further emphasize the importance to investigate the associations behind sleep with students and expand on the current pool of sleep knowledge.

Biswas, A. E. (2013). Whose Code of Conduct Matters Most? Examining the Link Between Academic Integrity and Student Development. Journal of College and Character, 14(3). https://doi.org/10.1515/jcc-2013-0034

Hershner, S. (2020). Sleep and academic performance: measuring the impact of sleep. Current Opinion in Behavioral Sciences, 33(2352-1546), 51–56. https://doi.org/10.1016/j.cobeha.2019.11.009.

Hysing, M., Harvey, A. G., Linton, S. J., Askeland, K. G., & Sivertsen, B. (2016). Sleep and academic performance in later adolescence: results from a large population-based study. Journal of Sleep Research, 25(3), 318–324. https://doi.org/10.1111/jsr.12373

Chen, W.-L., & Chen, J.-H. (2019b). Consequences of inadequate sleep during the college years: Sleep deprivation, grade point average, and college graduation. Preventive Medicine, 124(124), 23–28. https://doi.org/10.1016/j.ypmed.2019.04.017

Hypothesis

Based on the previous literature, the role of sleep plays a vital role in the life of a student as well as humans in general. A lack of knowledge is present in te literature for determining the short and long term outcomes of academic performance to provide comprehensive insights. Therefore, this analysis will undertake looking at how the term GPA of first year university students is impacted by the total sleep time they have at night and as suggested by previous literature, incorporate student's cumulative GPA to gain a better understanding into the dynamics behind sleep and effecting academic performanc.

H1: Shorter total sleep time and lower cumulative GPA are associated with lower term GPA

```r
library(dplyr)
```

```

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
library(tidyverse)
```

```
Warning: package 'tidyverse' was built under R version 4.4.2
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v forcats   1.0.0     v readr     2.1.5
v ggplot2   3.5.1     v stringr   1.5.1
v lubridate 1.9.3     v tibble    3.2.1
v purrr     1.0.2     v tidyr     1.3.1
```

```
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
library(ggplot2)
```

```r
# Loading data
sleep <- read_csv("cmu-sleep.csv")
```

```
Rows: 634 Columns: 15
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (1): cohort
dbl (14): subject_id, study, demo_race, demo_gender, demo_firstgen, bedtime_...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
sleep
```

```
# A tibble: 634 x 15
   subject_id study cohort demo_race demo_gender demo_firstgen bedtime_mssd
        <dbl> <dbl> <chr>      <dbl>       <dbl>         <dbl>        <dbl>
 1        185     5 lac1           1           1             0       0.117
 2        158     5 lac1           0           1             0       0.142
 3        209     5 lac1           1           1             0       1.53
 4        102     5 lac1           0           1             1       0.130
 5        174     5 lac1           1           1             0       0.130
 6        184     5 lac1           1           1             0       0.209
 7        255     5 lac1           1           1             0       0.675
 8        265     5 lac1           1           1             0       0.130
 9        343     5 lac1           1           0             0       1.48
10        137     5 lac1           1           1             0       0.0850
# i 624 more rows
# i 8 more variables: TotalSleepTime <dbl>, midpoint_sleep <dbl>,
#   frac_nights_with_data <dbl>, daytime_sleep <dbl>, cum_gpa <dbl>,
#   term_gpa <dbl>, term_units <dbl>, Zterm_units_ZofZ <dbl>
```

```r
# Inspecting Data
summary(sleep)
```

```
   subject_id          study          cohort            demo_race    
 Min.   :   1.0   Min.   :1.000   Length:634         Min.   :0.000  
 1st Qu.: 178.0   1st Qu.:2.000   Class :character   1st Qu.:1.000  
 Median :  358.5   Median :3.000   Mode  :character   Median :1.000  
 Mean   :13005.9   Mean   :3.181                      Mean   :0.812  
 3rd Qu.:  592.8   3rd Qu.:4.000                      3rd Qu.:1.000  
 Max.   :99978.0   Max.   :5.000                      Max.   :1.000  
                                                      NA's   :1      
  demo_gender     demo_firstgen     bedtime_mssd       TotalSleepTime 
 Min.   :0.0000   Min.   :0.0000   Min.   : 0.004505   Min.   :194.8  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 0.074694   1st Qu.:366.9  
 Median :1.0000   Median :0.0000   Median : 0.135007   Median :400.4  
 Mean   :0.5832   Mean   :0.1667   Mean   : 0.451688   Mean   :397.3  
 3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.: 0.291698   3rd Qu.:430.1  
 Max.   :1.0000   Max.   :2.0000   Max.   :20.849225   Max.   :587.7  
 NA's   :3        NA's   :4                                           
 midpoint_sleep  frac_nights_with_data daytime_sleep        cum_gpa     
 Min.   :247.1   Min.   :0.2143        Min.   :  2.269   Min.   :1.210  
 1st Qu.:345.2   1st Qu.:0.8214        1st Qu.: 23.098   1st Qu.:3.232  
 Median :388.2   Median :0.9322        Median : 34.982   Median :3.558  
 Mean   :398.7   Mean   :0.8674        Mean   : 41.164   Mean   :3.466  
 3rd Qu.:437.7   3rd Qu.:1.0000        3rd Qu.: 51.249   3rd Qu.:3.790  
 Max.   :724.7   Max.   :1.0000        Max.   :292.304   Max.   :4.000  
                                                                        
    term_gpa       term_units    Zterm_units_ZofZ 
 Min.   :0.350   Min.   : 5.00   Min.   :-3.98252  
 1st Qu.:3.233   1st Qu.:15.00   1st Qu.:-0.55104  
 Median :3.556   Median :17.00   Median : 0.04121  
 Mean   :3.450   Mean   :29.39   Mean   : 0.00000  
 3rd Qu.:3.810   3rd Qu.:48.00   3rd Qu.: 0.56027  
 Max.   :4.000   Max.   :73.00   Max.   : 4.05529  
                 NA's   :147     NA's   :147       
```

```r
# Inspecting Data
str(sleep)
```

```
spc_tbl_ [634 x 15] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ subject_id           : num [1:634] 185 158 209 102 174 184 255 265 343 137 ...
 $ study                : num [1:634] 5 5 5 5 5 5 5 5 5 5 ...
 $ cohort               : chr [1:634] "lac1" "lac1" "lac1" "lac1" ...
 $ demo_race            : num [1:634] 1 0 1 0 1 1 1 1 1 1 ...
 $ demo_gender          : num [1:634] 1 1 1 1 1 1 1 1 0 1 ...
 $ demo_firstgen        : num [1:634] 0 0 0 1 0 0 0 0 0 0 ...
 $ bedtime_mssd         : num [1:634] 0.117 0.142 1.529 0.13 0.13 ...
 $ TotalSleepTime       : num [1:634] 432 392 344 393 423 ...
 $ midpoint_sleep       : num [1:634] 459 364 561 416 369 ...
 $ frac_nights_with_data: num [1:634] 0.862 1 0.793 1 0.655 ...
 $ daytime_sleep        : num [1:634] 24.2 13.1 15 54.6 10.5 ...
 $ cum_gpa              : num [1:634] 3 3.66 3.57 3.61 3.21 3.2 3.4 3.86 3.79 3.53 ...
 $ term_gpa             : num [1:634] 3.38 2.6 3.07 3.56 4 3.36 3.19 3.28 3.5 2.55 ...
 $ term_units           : num [1:634] 73 64 63 61 61 60 60 60 60 59 ...
 $ Zterm_units_ZofZ     : num [1:634] 4.06 2.48 2.31 1.96 1.96 ...
 - attr(*, "spec")=
  .. cols(
  ..   subject_id = col_double(),
  ..   study = col_double(),
  ..   cohort = col_character(),
  ..   demo_race = col_double(),
  ..   demo_gender = col_double(),
  ..   demo_firstgen = col_double(),
  ..   bedtime_mssd = col_double(),
  ..   TotalSleepTime = col_double(),
  ..   midpoint_sleep = col_double(),
  ..   frac_nights_with_data = col_double(),
  ..   daytime_sleep = col_double(),
  ..   cum_gpa = col_double(),
  ..   term_gpa = col_double(),
  ..   term_units = col_double(),
  ..   Zterm_units_ZofZ = col_double()
  .. )
 - attr(*, "problems")=<externalptr>
```

Descriptive Statistics

The data comes from the Carnegie Mellon University (CMU), Department of Statistics & Data Science Data Repository. I am choosing to use the "Nightly sleep time and GPA in first years" dataset. The data was collected by its original collectors (Crewswell et al., 2023). Their sample consisted of 634 first year college students participants which was collected via the CMU registar.The important variables of interest for my hypothesis are bedtime_mssd, TotalSleepTime, cum_gpa, and term_gpa.

The dimensions of the data is 634 observations and 15 variables.

```
glimpse(sleep)
```

```
Rows: 634
Columns: 15
```

```
$ subject_id           <dbl> 185, 158, 209, 102, 174, 184, 255, 265, 343, 137~
$ study                <dbl> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, ~
$ cohort               <chr> "lac1", "lac1", "lac1", "lac1", "lac1", "lac1", ~
$ demo_race            <dbl> 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
$ demo_gender          <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, ~
$ demo_firstgen        <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, ~
$ bedtime_mssd         <dbl> 0.11672695, 0.14168084, 1.52928949, 0.13014845, ~
$ TotalSleepTime       <dbl> 432.2000, 391.9310, 344.3043, 392.6207, 423.4211~
$ midpoint_sleep       <dbl> 458.6600, 364.4655, 560.8913, 416.4828, 368.7632~
$ frac_nights_with_data <dbl> 0.8620690, 1.0000000, 0.7931034, 1.0000000, 0.65~
$ daytime_sleep        <dbl> 24.160000, 13.137931, 14.956522, 54.551724, 10.5~
$ cum_gpa              <dbl> 3.00, 3.66, 3.57, 3.61, 3.21, 3.20, 3.40, 3.86, ~
$ term_gpa             <dbl> 3.38, 2.60, 3.07, 3.56, 4.00, 3.36, 3.19, 3.28, ~
$ term_units           <dbl> 73, 64, 63, 61, 61, 60, 60, 60, 60, 59, 59, 58, ~
$ Zterm_units_ZofZ     <dbl> 4.0552949, 2.4825341, 2.3077829, 1.9582805, 1.95~
```

Below is a exploration of each variable:

TotalSleepTime

```
mean_TotalSleepTime <- mean(sleep$TotalSleepTime)
mean_TotalSleepTime
```

```
[1] 397.3239
```

The mean for total sleep time is 397.324 minutes.

```
sd_TotalSleepTime <- sd(sleep$TotalSleepTime)
sd_TotalSleepTime
```

```
[1] 50.85673
```

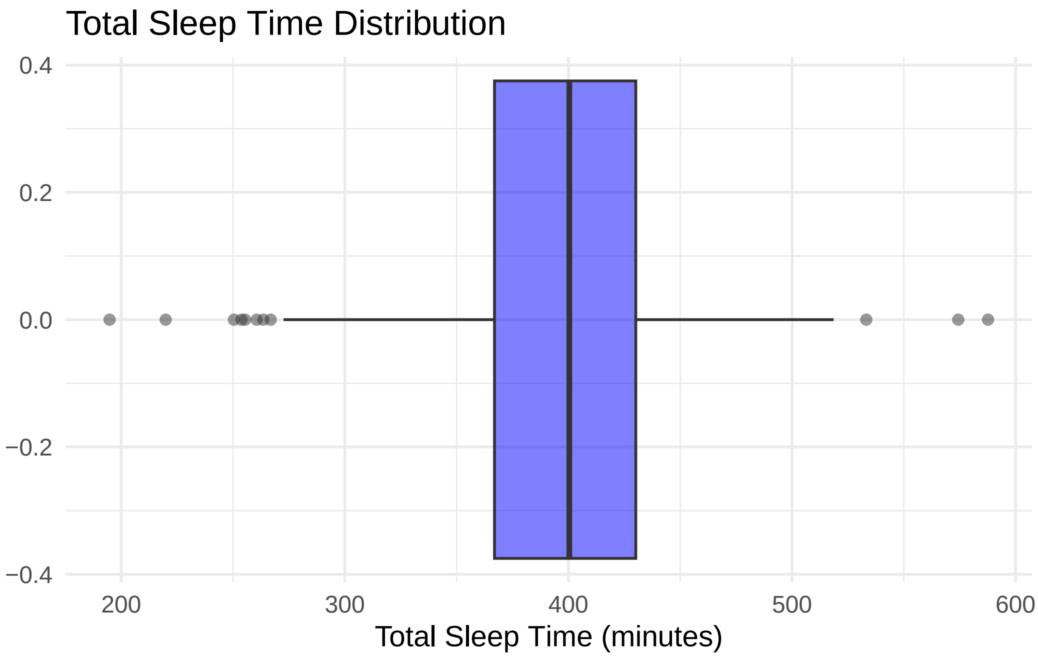
The standard deviation for total sleep time is 50.857 minutes.

```
iqr_TotalSleepTime <- IQR(sleep$TotalSleepTime)
iqr_TotalSleepTime
```

```
[1] 63.18451
```

The interquartile range for total sleep time is 63.18 minutes.

```
ggplot(sleep, aes(x = TotalSleepTime)) +
  geom_boxplot(fill = "blue", alpha = 0.5) +
  labs(title = "Total Sleep Time Distribution", x = "Total Sleep Time (minutes)") +
  theme_minimal()
```



cum_gpa

```
mean_cum_gpa <- mean(sleep$cum_gpa)
mean_cum_gpa
```

```
[1] 3.465596
```

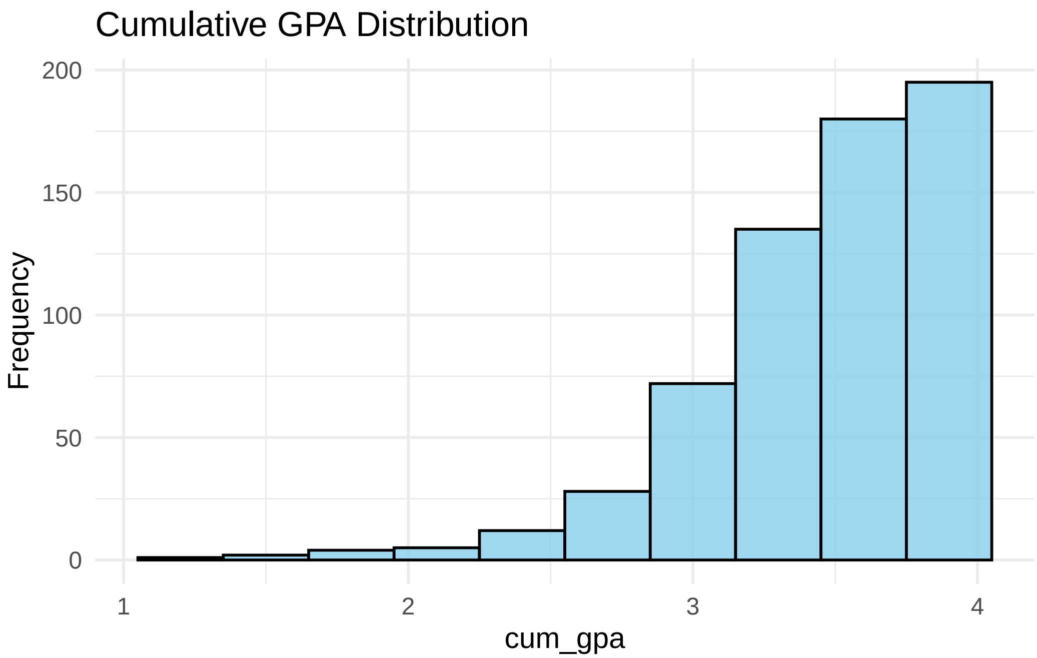
The mean cumulative GPA was 3.466.

```
sd_cum_gpa <- sd(sleep$cum_gpa)
sd_cum_gpa
```

```
[1] 0.4375772
```

The standard deviation for cumulative GPA was 0.438.

```
ggplot(sleep, aes(x = cum_gpa)) +
  geom_histogram(binwidth = 0.3, fill = "skyblue", color = "black", alpha = 0.8) +
  labs(title = "Cumulative GPA Distribution",
       y = "Frequency") +
  theme_minimal()
```



#term_gpa

```
mean_term_gpa <- mean(sleep$term_gpa)
mean_term_gpa
```

```
[1] 3.449598
```

The mean term gpa was 3.450.

```
sd_term_gpa <- sd(sleep$term_gpa)
sd_term_gpa
```

```
[1] 0.5004669
```

The standard deviation for term gpa was 0.500.

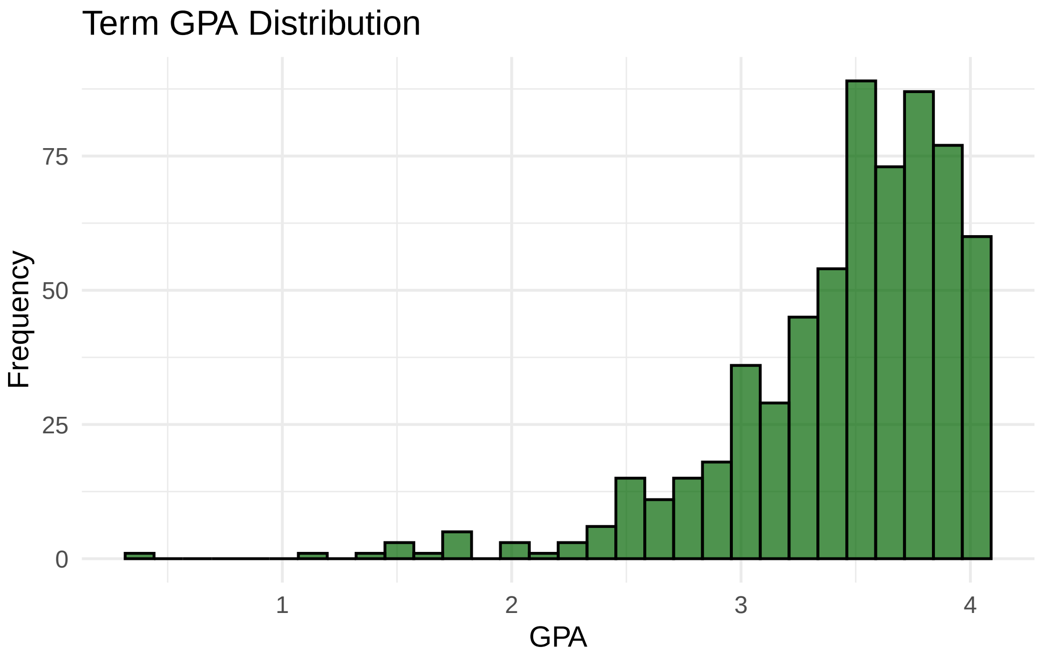
```
quantile_term_gpa <- quantile(sleep$term_gpa)
quantile_term_gpa
```

```
      0%      25%      50%      75%     100% 
0.350000 3.233333 3.555667 3.810000 4.000000 
```

```
ggplot(sleep, aes(x = term_gpa)) +
  geom_histogram(bindwidth = 0.5, fill = "darkgreen", color = "black", alpha = 0.7) +
  labs(title = "Term GPA Distribution",
       x = "GPA",
       y = "Frequency") +
  theme_minimal()
```

```
Warning in geom_histogram(bindwidth = 0.5, fill = "darkgreen", color = "black",
: Ignoring unknown parameters: `bindwidth`
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Model Fitting/Hypothesis Testing

The outcome variable is term GPA. The explanatory/predictor variables are total sleep time, and cum_GPA (With control variable `bedtime_mssd`)

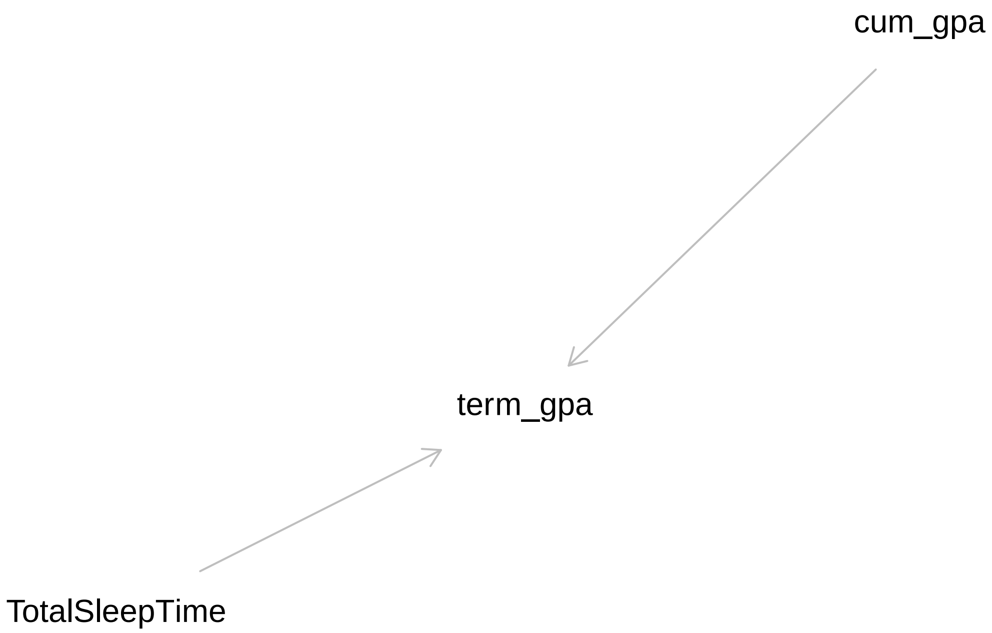
```
# DAG based on proposed hypothesis
library(dagitty)
```

```
Warning: package 'dagitty' was built under R version 4.4.1
```

```
dag <- dagitty('
    TotalSleepTime -> term_gpa
    cum_gpa -> term_gpa
  ')
```

```
plot(dag)
```

```
Plot coordinates for graph not supplied! Generating coordinates, see ?coordinates for how to
```



```r
# Fitting Model for Predictors

model_mr1 <- lm(term_gpa ~ TotalSleepTime + cum_gpa, data = sleep)

summary(model_mr1)
```

```
Call:
lm(formula = term_gpa ~ TotalSleepTime + cum_gpa, data = sleep)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.77214 -0.15072  0.06582  0.21925  1.02308 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    0.459234   0.160277   2.865  0.00431 ** 
TotalSleepTime 0.001308   0.000299   4.373 1.43e-05 ***
cum_gpa        0.712962   0.034752  20.516  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3803 on 631 degrees of freedom
Multiple R-squared:  0.4245,    Adjusted R-squared:  0.4227 
F-statistic: 232.7 on 2 and 631 DF,  p-value: < 2.2e-16
```

Interpretations:

The coefficient of my explanatory variables were 0.001208 for TotalSleepTime and 0.712962 for cum_gpa. This means that for every additional minute of sleep, the sample's participants term GPAs are predicted to increase by 0.001208 (when holding cum_gpa constant). Likewise, the coefficient value for cumulative GPA suggests that for every 1 additional unit increase in cumulative GPA, the students' term GPA are predicted to increase by 0.712962 (when holding TotalSleepTime constant).

```r
# Changing p-value of TotalSleepTime to scientific notation
num = 1.43e-05
print("Modified Num: ")
```

```
[1] "Modified Num: "
```

```
print(num)
```

```
[1] 1.43e-05
```

H0: mu = 397.324 Ha: mu != 397.324

```
# CI for TotalSleepTime

n_TST <- length(sleep$TotalSleepTime)

mean_TotalSleepTime
```

```
[1] 397.3239
```

```
se_TST <- sd_TotalSleepTime / sqrt(length(sleep$TotalSleepTime))
se_TST
```

```
[1] 2.019779
```

```
t_TST <- qt(0.975, df = length(sleep$TotalSleepTime) - 1)
t_TST
```

```
[1] 1.963719
```

```
ci_TST <- c(mean_TotalSleepTime - t_TST * se_TST, mean_TotalSleepTime + t_TST * se_TST)
ci_TST
```

```
[1] 393.3576 401.2902
```

The 95% confidence interval for TotalSleepTime is [393.4, 401.3]. Therfore, we fail to reject the null hypothesis. This implies that there is no significant evidence to suggest that the mean of total sleep time for the sample's participants deviates from 397.324 minutes at 0.05 significance level.

On the other hand, the p-value for TotalSleepTime is 1.43e-05, which indicates a predictor significantly predicting the term GPA.

Ho: mu = 3.466 Ha: mu != 3.466

```
# CI for cum_gpa
mean_cum_gpa
```

```
[1] 3.465596
```

```
se_cgpa <- sd_cum_gpa / sqrt(length(sleep$cum_gpa))
se_cgpa
```

```
[1] 0.01737841
```

```
t_cgpa <- qt(0.975, df = length(sleep$cum_gpa))
t_cgpa
```

```
[1] 1.963713
```

```
ci_cgpa <- c(mean_cum_gpa - t_cgpa * se_cgpa, mean_cum_gpa + t_cgpa * se_cgpa)
ci_cgpa
```

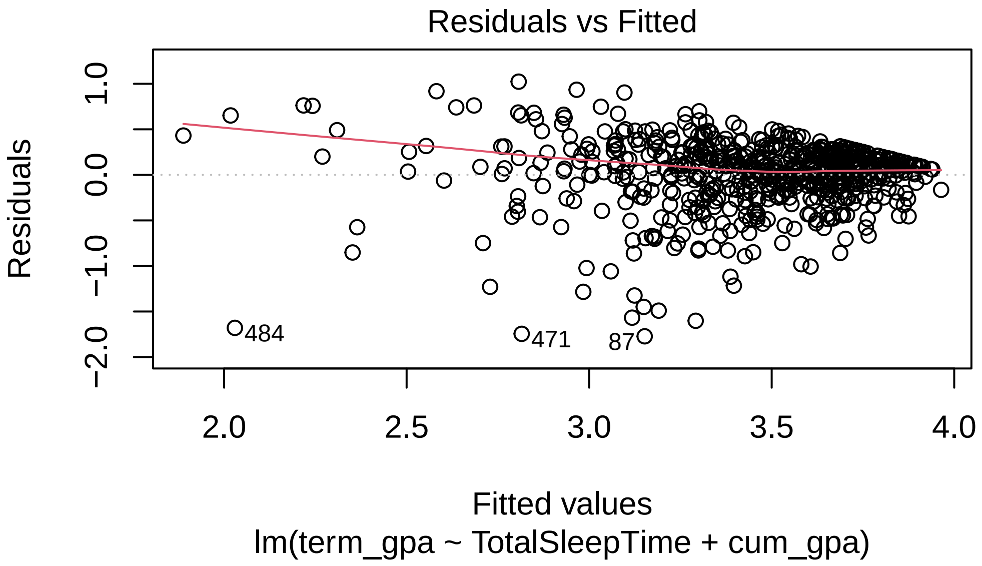
```
[1] 3.431470 3.499722
```

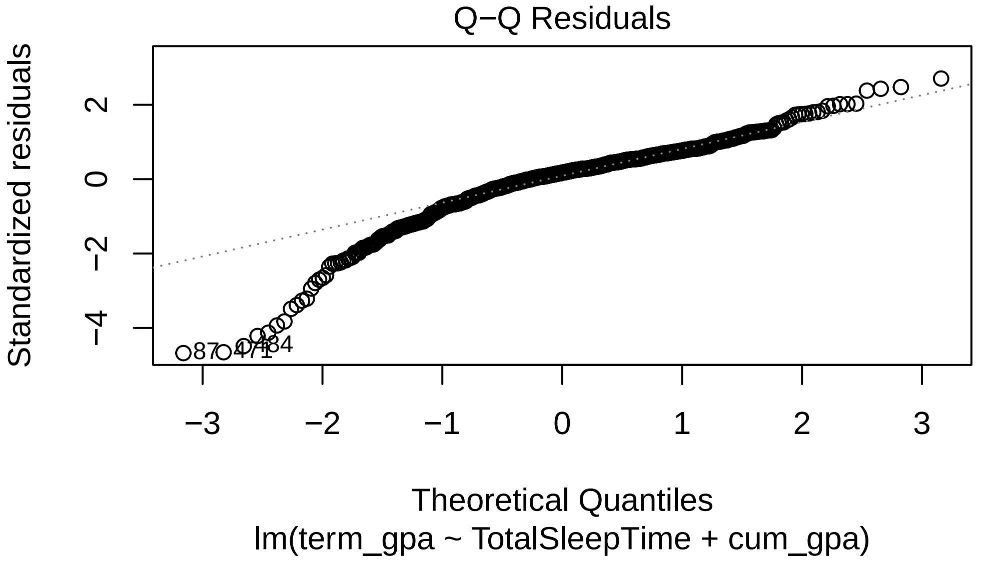
The 95% confidence interval for cum_gpa is [3.431, 3.500]. The hypothesized mu value is within the confidence interval. Hence, we fail to reject the null hypothesis.

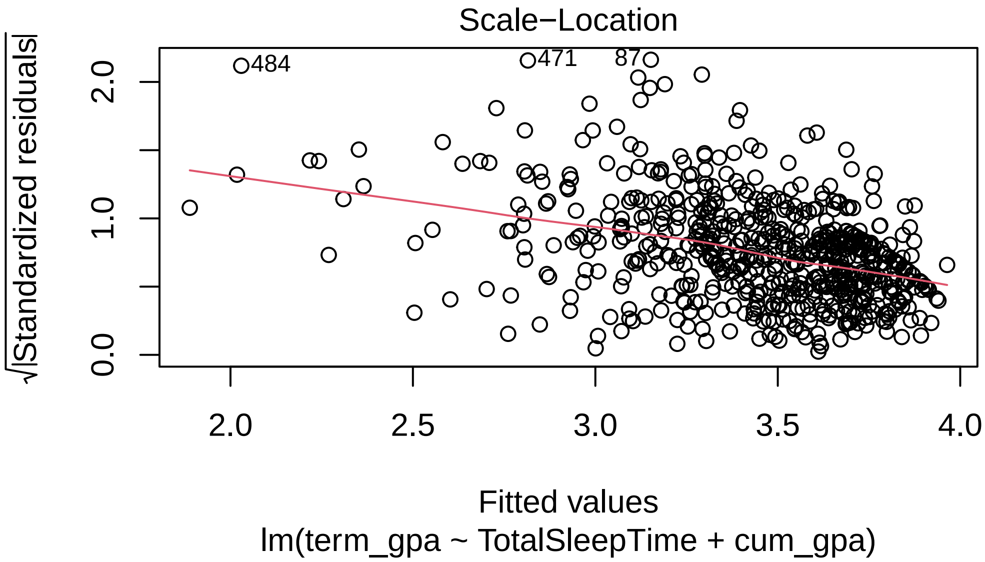
Similar to TotalSleepTime, the p-value for cum_gpa (< 2e-16) implies that cumulative GPA in the sample has a significant effect on the outcome variable term GPA.

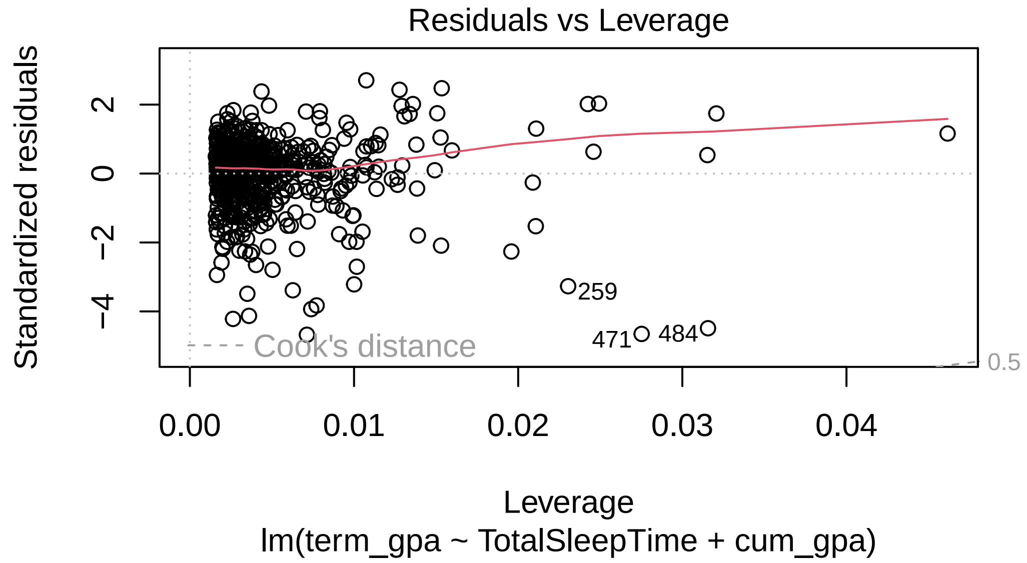
Diagnostics and Model Evaluation

```
plot(model_mr1)
```


Residuals vs Fitted
Residuals
1.0
0.0
−1.0
−2.0
2.0
2.5
3.0
3.5
4.0
484
471
87
Fitted values
lm(term_gpa ~ TotalSleepTime + cum_gpa)


Q−Q Residuals
Standardized residuals
2
0
−2
−4
−3
−2
−1
0
1
2
3
87
471
484
Theoretical Quantiles
lm(term_gpa ~ TotalSleepTime + cum_gpa)





The Residuals vs Fitted plot suggests that the model is not adequately capturing the variability in the data. The funnel-like shape and curved red line indicates that that the prediction of the model may not be fitting the data. Next, I will look at the Q-Q Residuals plot. In terms of linearity, the deviation from the line may imply that the data from the sample is not normally distributed. Furthermore, the slope of the line is relatively steeper than the line for the theoretical distribution. This may suggest that this sample data has higher variance. While not determinant, there appear to be outliers that may indictae the presence of sampling and/or measurement error. The Scale Location plot tests that assumptions of homoskedasticity. Based on the plot, the spread of the points suggests against the assumption of constant variance. The points are not evenly spread and are not consistent throughout the theoretical line. Given this information, I think I need to consider alternative modeling techniques and/or transformations. Lastly, the Residuals vs Leverage plot. There are a few values that deviate far from zero and have high leverage values. These points may have exerted a strong influence on the regression model above. It seems that the values that have both large positive standardized residuals and high leverage have provided a strong influence on the lm model.

Adjusted R-Squared Value: 0.4227 Mutliple R-Squared Value: 0.4245

An Adjusted R-Squared value of 0.4227 and a Multiple R-Squared value of 0.4245 suggests that the provided predictors may not over fit the model and the inclusion of the given variables are reasonable. In other words, the estimate of the variance explained for the model for the population and the varaince explained for the data is very similar.

Discussion

The results of the linear regression analysis suggest that the variables TotalSleepTime and cum_gpa are predictors of term_gpa. The findings of these studies have practical implications for students and faculty of institutional institutes. Adequtae rest could contribute, even modestly, to improving academic performance. However, cum_gpa seemed to be a stronger predictor of academic success (measured by GPA). Which hints at the importance of consistent high academic performance over the long-run.Some limitations of the analysis are that this analysis assumes the presence of a linear relationship between the predictor variables TotalSleepTime and cum_gpa and the outcome variable, term_GPA. Exploring non-linear relationships with these predictors could offer deeper insights.

Adjustments for 604 Assignment 3

```r
sleep_std <- sleep |>
  mutate(across(c(TotalSleepTime, cum_gpa, term_gpa), ~scale(.)[,1]))
```

```r
model_std <- lm(term_gpa ~ TotalSleepTime + cum_gpa, data = sleep_std)
summary(model_std)
```

```
Call:
```

```
lm(formula = term_gpa ~ TotalSleepTime + cum_gpa, data = sleep_std)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5410 -0.3012  0.1315  0.4381  2.0442

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.208e-16  3.018e-02   0.000        1
TotalSleepTime 1.329e-01  3.038e-02   4.373 1.43e-05 ***
cum_gpa        6.234e-01  3.038e-02  20.516  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7598 on 631 degrees of freedom
Multiple R-squared:  0.4245,    Adjusted R-squared:  0.4227
F-statistic: 232.7 on 2 and 631 DF,  p-value: < 2.2e-16
```