# MATH 189 Project 4

Group 29

March 2021

## 1 Introduction

For many people, entering the universities will be the last stop transiting them from academy to society, as students learn necessary academic knowledge but start to accumulate social experiences. However, almost a third of students in universities leave without earning a degree and on average students in the universities spend longer time on accomplishing their academic requirements. Hence, it is crucial for people to explore what kinds of activities will influence students' academic achievements, how will the activities influence the academic achievements, and prove the credibility of the finds. Since the health is almost the basic for pursuing people's life goals including their academic goals, the main question will be "How Health Behaviors Relate to Academic Performance via Affect?". In order to decompose the big question into specific parts, we focus on exploring how sleep and physical activities relate to students' academic performance and form up two main parts followed by five sub-questions in this project. The first main part is based on between-person level followed with four questions. The first question will be exploring whether better sleep quality or higher physical activity predicts better academic achievements. Then, the second part will be focused on whether average affects which may be positive or negative mediates the relation between average sleep quality and average learning goal achievement. Similarly, the third one will be whether average affects which may be positive or negative mediates the relation between average physical activity and average learning goal achievement. After exploring how these two variables influence students' studying goals, we will finish this part by utilizing data to prove or disprove the potential negative or positive influences imposed on the learning goals and we will apply the datasets illustrating students who on average report better learning goal achievement are more likely to pass their examinations. The second part will focus on within-person level with one exploring question including three aspects which are explorations on studying connections between higher physical sleep quality or higher physical goals and learning gals on the day-to-day level and proof for the relations with corresponding dataset.

## 2 Data

The dataset we use in this project is HealthBehavAcadPerfAffect.tab.tsv. which is provided by the professor Jelena Bradic. This dataset contains 13 columns and 72 rows, which shows 67392 observations. The 13 columns refer to 13 variables which are Day (Survey day), Sex (Participants' sex), Age (Participants' age), Sem (Semester: Number of semesters studied), SQ (Sleep quality: 1 (very bad) to 4 (very good)), PhysAct (Physical activity: Number of minutes engaged in mild, moderate and strenuous exercise weighted by metabolic equivalents and then summed to produce a total daily leisure activity score), PA (Positive affect 1 (not at all) to 7 (extremely)), NA (Negative affect 1 (not at all) to 7 (extremely)), LGA (Learning goal achievement 0 (not at all) to 4 (completely)), Exam (Examination success 0 (fail) to 6 (highest grade)), HSG (High school grades 1 (lowest grade) to 6 (highest grade)), and BDI (Beck Depression Inventory 1 (not) 2 (mild to moderate) 3 (clinically relevant symptoms)). Among these variables described above, the numerical continuous variables are Sem, Day while the numerical discrete variable is Age. The numerical discrete variables are SQ, PhysAct, PA, NA, LGA, Exam, HSG, and BDI and the regular categorical variable includes Sex. This dataset focuses on the results of 32 day including 14 days of test preparation and 18 days of examinations period. Meanwhile, even though the dataset originally collects data from 82 students, but since 5 did not take examinations and 5 did not report grades, the final size will be 72.

# 3 Background

Improving students' performance and learning experiences in universities and colleges is key to ensure students with more desirable future and to fulfill students' expectations. It is students all over the world undergo difficulties successfully graduating from college or graduating on expected time, which increases educational cost and harms students' future. We intend to find health factors, such as sleep and physical activities, that could affect students' academic performance.

In 2013, several Finnish scholars, led by SYVÄOJA, studied the relationships between objectively measured and self-reported physical activity, sedentary behavior, and academic performance in Finnish children. The academic achievement score was calculated according to students' GPA that students provided. Activities were measured through questionnaire.The authors suggested that "this is the first study to examine the associations of objectively measured and self-reported physical activity on teacher-rated academic achievement". They concluded that self reported increase in physical activity would lead to increase in higher levels of academic performance. They argued that the reason behind this was because that physical activity positively influenced children's executive functions and working memory. They also found that "It may be that some of the most active children spend time in physical activities at the expense of time devoted to homework". However, the study found that objectively observed physical activities did not affect academic performance.

# 4 Investigation

## 4.1 Between-person level

### 4.1.1 Q1

After setting the directory and importing the data into RStudio, we first perform data cleaning to make sure each used datapoint is valid. Then we construct linear regression to see if there is a statistical relationship between the factors.

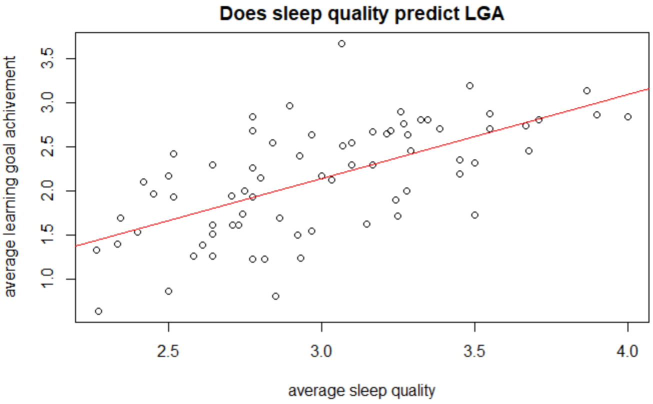
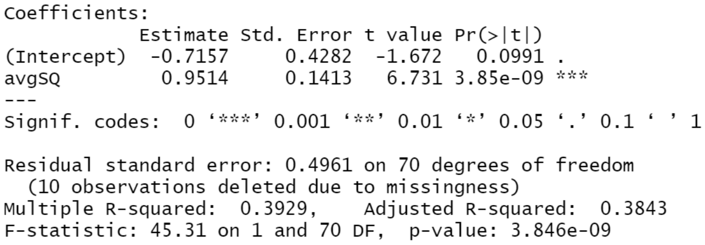


Figure 1. We first see the scatter plot along with the linear regression line of average sleep quality on average learning goal achievement.

In this graph, we can see that the datapoints range from around 2 to 4 in sleep quality and 0 to 3.5 in LGA. We observe that there is a possible positive relationship between the two variables.



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.7157     0.4282  -1.672   0.0991 .
avgSQ         0.9514     0.1413   6.731 3.85e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4961 on 70 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.3929,    Adjusted R-squared:  0.3843
F-statistic: 45.31 on 1 and 70 DF,  p-value: 3.846e-09
```

Figure 2. We perform a rigorous linear regression.

We can see that the least square regression predicts that for each increase of 1 in average sleep quality, 0.9514 unit of increase in LGA is estimated by the relationship. By the theory of T-test, p-value is 3.95*exp(-9). We say that we are 99.9% confident that it is statistically significant that there is a positive relationship between average sleeping quality and average LGA. We also observed that the R-squared is 0.3929. Then the R score is 0.6268. Then we say that there is a moderate, linear positive relationship between average sleeping quality and average learning goal achievement. We can then say that better average sleeping quality predicts better average LGA.

Then we look at the relationship between average physical activity and learning goal achievement.

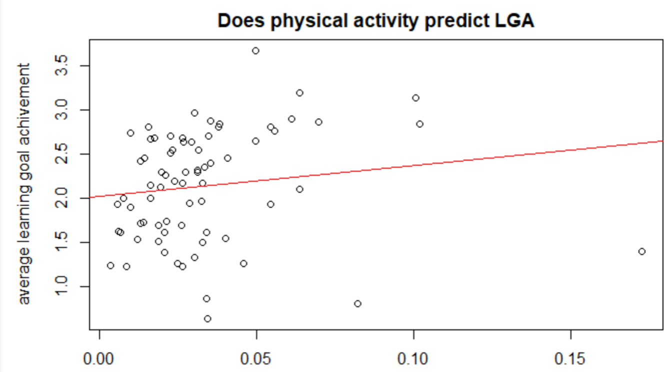
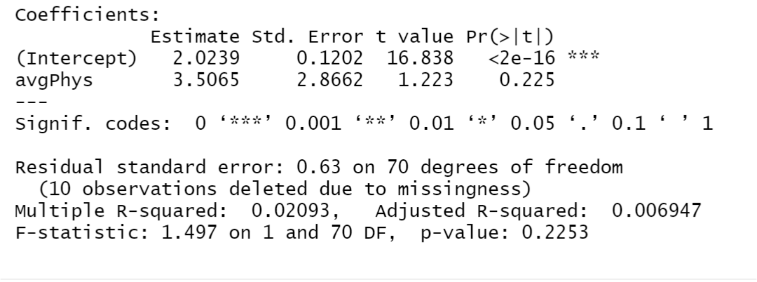


Figure 3. We then look at the scatter plot along with the linear regression line of average physical activity on average learning goal achievement.

In this graph, we can see that the datapoints mostly accumulate around 0 to 0.075 in physical activity (times 0.0001 for better comparison in following sub analysis) and 0 to 3.5 in LGA. We observe that there is possibly a shallow, positive relationship between the two variables.



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.0239     0.1202  16.838   <2e-16 ***
avgPhys       3.5065     2.8662   1.223    0.225    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.63 on 70 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.02093,	Adjusted R-squared:  0.006947 
F-statistic: 1.497 on 1 and 70 DF,  p-value: 0.2253
```

Figure 4.linear regression summary for physical activity on LGA

From the analysis above, we conclude that for each 1 unit of average physical activity increases, average learning goal achievement will increase by 3.5065*0.0001 (for easier comparison later). However, we observe that the p-value is 0.2225, which does not statistically support that there is a positive relationship at a 90% confidence level. Hence, even though there is somewhat a relationship we may observe graphically, rigorous statistics does not support this observation. Then we can't say that higher physical activity level predicts better learning goal achievement.

### 4.1.2 Q2

In this question, we want to investigate whether positive or negative affect mediates the relation between average sleep quality and average learning goal achievement.

(Positive effect) In order to perform a mediate analysis, we first investigate the relationship between average sleeping quality and average positive affect.

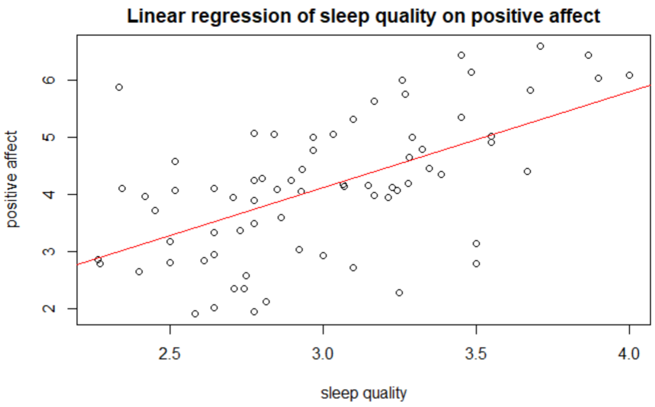
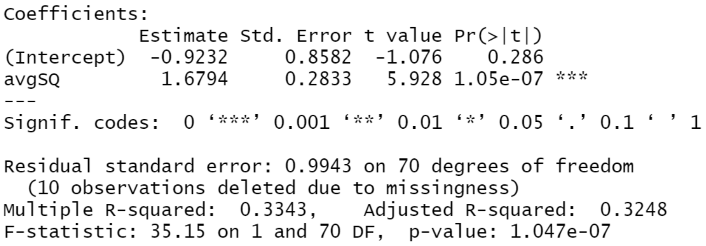


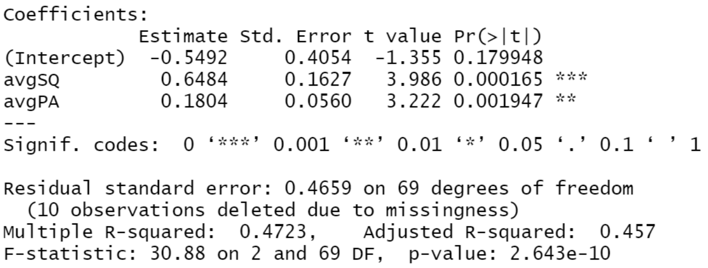
Figure 5. We can observe from this scatter plot that there is a pretty strong positive linear relationship between the two variables. We then investigate this relationship using rigorous test.



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -0.9232     0.8582  -1.076    0.286    
avgSQ         1.6794     0.2833   5.928 1.05e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9943 on 70 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.3343,    Adjusted R-squared:  0.3248 
F-statistic: 35.15 on 1 and 70 DF,  p-value: 1.047e-07
```

Figure 6.linear regression summary of sleep quality on positive effect

In this test, we observe that when average sleep quality increases by 1 unit, average positive affect will increase by 1.6794 units. The p-value is 1.05e-07, which implies that it is statistically significant at 99.9% level that average sleep quality will positively affect average positive affect.



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -0.5492     0.4054  -1.355 0.179948    
avgSQ         0.6484     0.1627   3.986 0.000165 ***
avgPA         0.1804     0.0560   3.222 0.001947 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4659 on 69 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.4723,    Adjusted R-squared:  0.457 
F-statistic: 30.88 on 2 and 69 DF,  p-value: 2.643e-10
```

Figure 7.linear regression summary of sleep quality on positive effect

In this test, we perform a regression of average sleep quality and positive affect on average learning goal achievement. We observe that when average sleep quality increases by 1unit, average LGA will increase by 0.6484 units. The p-value is 0.000165, which implies that it is statistically significant at 99.9% level that average sleep quality will positively affect average LGA. We observe that when average positive affect increases by 1unit, average LGA will increase by 0.1804 units. The p-value is 0.001947, which implies that it is statistically significant at 99% level that average positive affect will positively affect average LGA.

We then assume variables already omitted since linear model in R automatically do that. Hence, we don't take that into consideration. In addition, we can assume there's no causal effect because there is no such obvious logic. We then have the indirect effect = 1.6794*0.1804 = 0.3029638. Total effect = 0.9514. Then by the mediate formula: total effect = indirect effect + direct effect, we can calculate that direct effect = 0.6484, which is exactly what we have in the test. Then we perform Sobel test: T = 0.3029638 / sqrt((b*exp(2))*(SE(a)*exp(2)) + (a*exp(2))*(SE(b)*exp(2))) = 0.3029638 / 0.1070359 = 2.830488. Since absolute value of 2.830488 ¿ 1.96, we reject the null hypothesis that indirect effect equals to 0. Therefore, since direct effect is not zero statistically, average positive affect partially mediates the relationship.

To finish the analysis, we find that the 95% confidence interval for indirect effect: 0.3029638 +/- 0.1070359*1.96 = (0.09317344, 0.5127542). Also, all four coefficients for total effect, direct effect, and indirect effect (two coefficients) are statistically significant at a 95% level.

Negative effect In order to perform a mediate analysis, we first investigate the relationship between average sleeping quality and average negative affect.

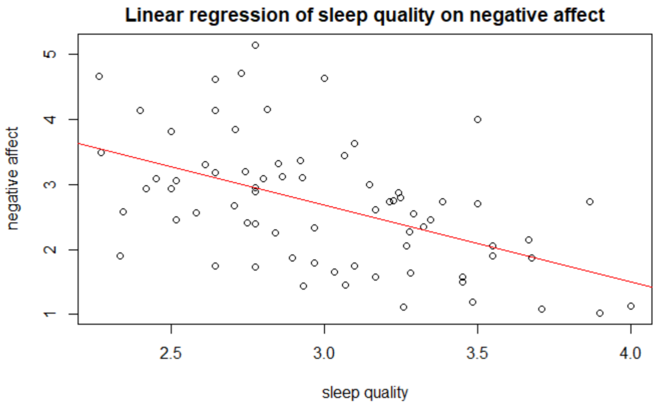
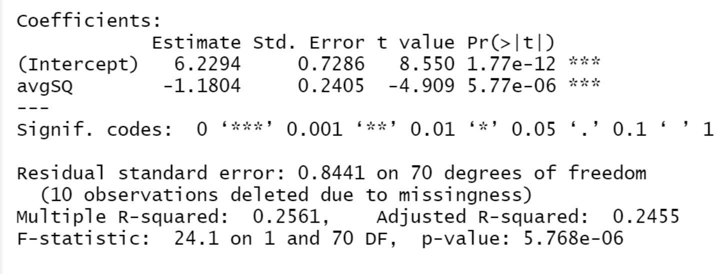
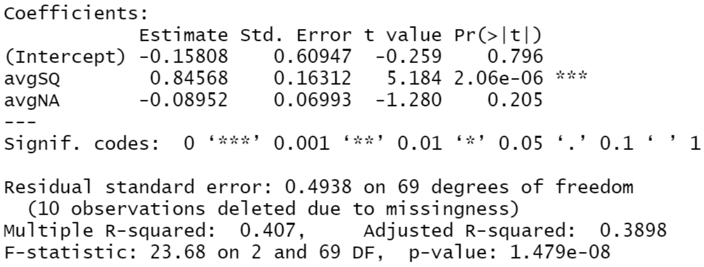


Figure 8.We can observe from this scatter plot that there is a moderately strong negative linear relationship between the two variables. We then investigate this relationship using rigorous test.



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   6.2294     0.7286   8.550 1.77e-12 ***
avgSQ        -1.1804     0.2405  -4.909 5.77e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8441 on 70 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.2561,	Adjusted R-squared:  0.2455 
F-statistic:  24.1 on 1 and 70 DF,  p-value: 5.768e-06
```

Figure 9. In this test, we observe that when average sleep quality increases by 1unit, average negative affect will decrease by 1.1804 units. The p-value is 5.77e-06, which implies that it is statistically significant at a 99.9% level that average sleep quality will negatively affect average negative affect.



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.15808    0.60947  -0.259    0.796    
avgSQ        0.84568    0.16312   5.184 2.06e-06 ***
avgNA       -0.08952    0.06993  -1.280    0.205    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4938 on 69 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.407,	Adjusted R-squared:  0.3898 
F-statistic: 23.68 on 2 and 69 DF,  p-value: 1.479e-08
```

Figure 10. In this test, we perform a regression of average sleep quality and negative affect on average learning goal achievement.

We observe that when average sleep quality increases by 1unit, average LGA will increase by 0.84568 units. The p-value is 2.06e-06, which implies that it is statistically significant at a 99.9% level that average sleep quality will positively affect average LGA. We also observe that when average negative affect increases by 1unit, average LGA will decrease by 0.08952 units. The p-value is 0.205, which implies that it is not statistically significant that average negative affect will negatively affect average LGA.

We then assume variables already omitted since linear model in R automatically do that. Hence, we don't take that into consideration. In addition, we can assume there's no causal effect because there is no such obvious logic. We then have the indirect effect = (-1.1804) * (-0.08952) = 0.1056694. Total effect = 0.9514. Then by the mediate formula: total effect = indirect effect + direct effect, we can calculate that direct effect = 0.8457306, which is exactly what we have in the test. Then we perform Sobel test: T =

0.1056694 / sqrt((b*exp(2))*(SE(a)*exp(2)) + (a*exp(2))*(SE(b)*exp(2))) = 0.1056694/ 0.08530686= 1.238698. Since absolute value of 1.238698 ¡ 1.96, we fail to reject the null hypothesis that indirect effect equals to 0. Therefore, negative effect statistically doesn't mediate the relationship. However, the direct effect is not zero.

To finish the analysis, we find that the 95% confidence interval for indirect effect: 0.1056694+/-0.8457306*1.96 = (-1.551963, 1.763301). The interval includes 0. Also, all four coefficients for total effect, direct effect, and indirect effect (two coefficients) are statistically significant at a 95% level except for the one between average negative effect and average learning goal achievement.

### 4.1.3 Q3

In this question, we want to investigate whether positive or negative affect mediates the relation between average physical activity and average learning goal achievement.

Positive effect In order to perform a mediate analysis, we first investigate the relationship between average physical activity and average positive affect.

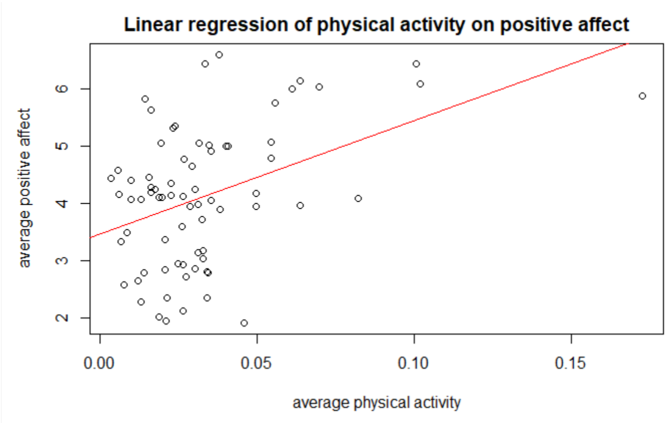
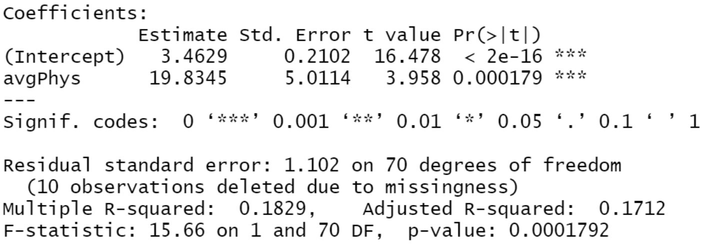
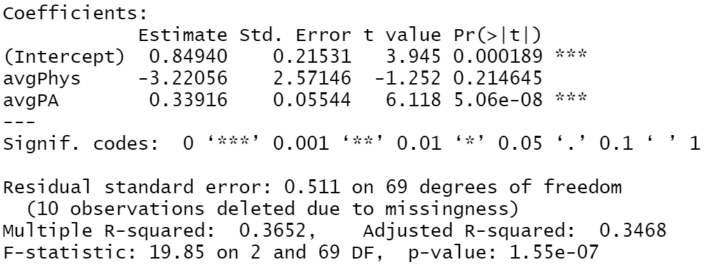


Figure 11. We can observe from this scatter plot that there is a relatively strong, positive linear relationship between the two variables. We then investigate this relationship using rigorous test.



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   3.4629     0.2102  16.478  < 2e-16 ***
avgPhys      19.8345     5.0114   3.958 0.000179 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.102 on 70 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.1829,	Adjusted R-squared:  0.1712 
F-statistic: 15.66 on 1 and 70 DF,  p-value: 0.0001792
```

Figure 12. In this test, we observe that when average physical activity increases by 1unit, average positive affect will increase by 19.8345 units. The p-value is 0.000179, which implies that it is statistically significant at a 99.9% level that average physical activity will positively affect average positive affect.



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.84940    0.21531   3.945 0.000189 ***
avgPhys     -3.22056    2.57146  -1.252 0.214645    
avgPA        0.33916    0.05544   6.118 5.06e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.511 on 69 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.3652,	Adjusted R-squared:  0.3468 
F-statistic: 19.85 on 2 and 69 DF,  p-value: 1.55e-07
```

Figure 13. In this test, we perform a regression of average physical activity and positive affect on average learning goal achievement.

We observe that when average physical activity increases by 1unit, average LGA will decrease by 3.222056 units. The p-value is 0.214645, which implies that it is not statistically significant that average physical activity will negatively affect average LGA. We also observe that when average positive affect increases by 1unit, average LGA will increase by 0.33916 units. The p-value is 5.06e-08, which implies that it is statistically significant at a 99.9% level that average positive affect will positively affect average LGA.

We then assume variables already omitted since linear model in R automatically do that. Hence, we don't take that into consideration. In addition, we can assume there's no causal effect because there is no such obvious logic. We then have the indirect effect = 19.8345*0.33919 = 6.727664. Total effect = 3.5065. Then by the mediate formula: total effect = indirect effect + direct effect, we can calculate that direct effect = -3.221164, which is exactly what we have in the test. Then we perform Sobel test: T = 6.727664 / sqrt((b*exp(2))*(SE(a)*exp(2)) + (a*exp(2))*(SE(b)*exp(2))) = 6.727664 / 2.024362= 3.32335. Since absolute value of 3.32335 ¿ 1.96, we reject the null hypothesis that indirect effect equals to 0. Therefore, since direct effect is zero statistically (-3.22056), average positive effect completely mediates the relationship.

To finish the analysis, we find that the 95% confidence interval for indirect effect: 6.727664 +/- 2.024362*1.96 = (2.759914, 10.69541). Also, two of the coefficients for indirect effect are statistically significant at a 95% level, while the ones for total effect and direct effect are not significant.

Negative effect In order to perform a mediate analysis, we first investigate the relationship between average physical activity and average negative affect.

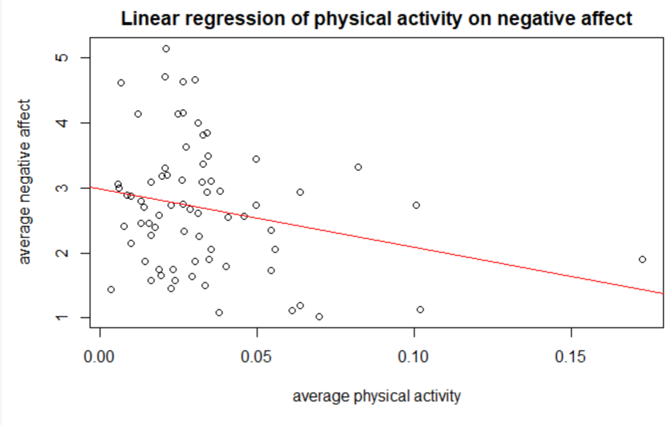
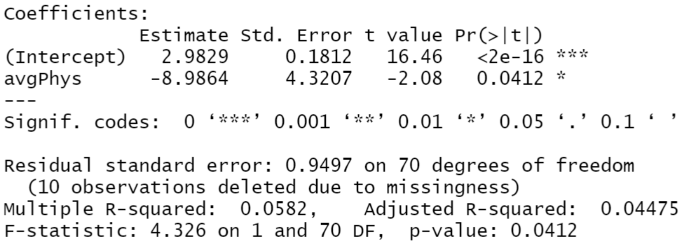
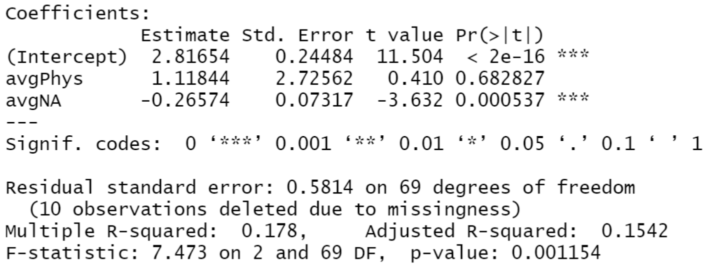


Figure 14. We can observe from this scatter plot that there is a moderately negative linear relationship between the two variables. We then investigate this relationship using rigorous test.



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.9829     0.1812   16.46   <2e-16 ***
avgPhys      -8.9864     4.3207   -2.08   0.0412 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9497 on 70 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.0582,	Adjusted R-squared:  0.04475 
F-statistic: 4.326 on 1 and 70 DF,  p-value: 0.0412
```

Figure 15. In this test, we observe that when average physical activity increases by 1unit, average negative affect will decrease by 8.9864 units. The p-value is 0.0412, which implies that it is statistically significant at a 95% level that average physical activity will positively affect average positive affect.



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.81654    0.24484  11.504  < 2e-16 ***
avgPhys      1.11844    2.72562   0.410 0.682827
avgNA       -0.26574    0.07317  -3.632 0.000537 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5814 on 69 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.178,	Adjusted R-squared:  0.1542
F-statistic: 7.473 on 2 and 69 DF,  p-value: 0.001154
```

Figure 16. In this test, we perform a regression of average physical activity and negative effect on average learning goal achievement.

We observe that when average physical activity increases by 1unit, average LGA will increase by 1.11844 units. The p-value is 0.682827, which implies that it is not statistically significant that average physical activity will negatively affect average LGA. We also observe that when average positive affect increases by 1unit, average LGA will decrease by 0.26574 units. The p-value is 0.000537, which implies that it is statistically significant at a 99.9% level that average negative affect will negatively affect average LGA.

We then assume variables already omitted since linear model in R automatically do that. Hence, we don't take that into consideration. In addition, we can assume there's no causal effect because there is no such obvious logic. We then have the indirect effect = (-8.9864) * (-0.26574) = 2.388046. Total effect = 3.5065. Then by the mediate formula: total effect = indirect effect + direct effect, we can calculate that direct effect = 1.118454, which is exactly what we have in the test. Then we perform Sobel test: T = 2.388046 / sqrt((b*exp(2))*(SE(a)*exp(2)) + (a*exp(2))*(SE(b)*exp(2))) = 2.388046 / 1.323131 = 1.804845. Since absolute value of 1.804845 ¡ 1.96, we fail to reject the null hypothesis that indirect effect equals to 0. Therefore, negative effect doesn't mediate the relationship statistically.

To finish the analysis, we find that the 95% confidence interval for indirect effect: 2.388046 +/- 1.323131 *1.96 = (-0.2052908, 4.981383), which includes 0. Also, two of the coefficients for indirect effect are statistically significant, while the ones for total effect and direct effect are not significant.

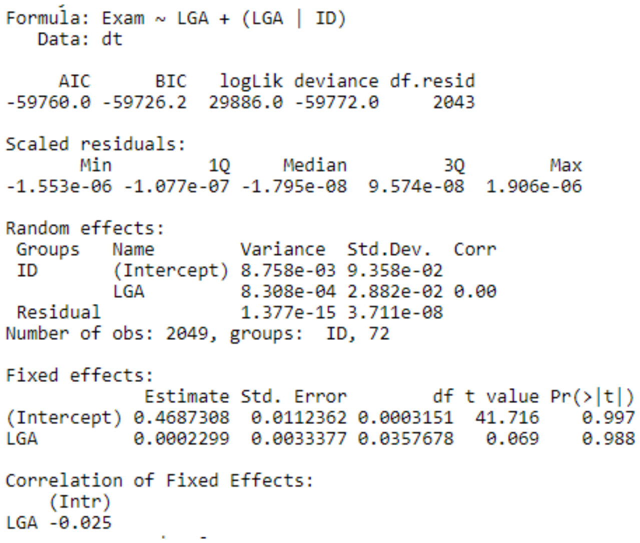
In the correlated model, we assume the random effect is also correlated to the fixed effect. To get a clear relationship between LGA and Exam result, we need to take the random effect of different subject on LGA into consideration.

#### 4.1.4 Q4

In this section, we intend to discuss if a better learning goal achievement lead to a better grade in the exam? We will examine this using both Mixed Linear Model and Generalized Estimate Equation.

**Mixed Linear Model** We will try fit two different types of Linear Mixed Model, one with correlated random effects, and another with uncorrelated random effects.

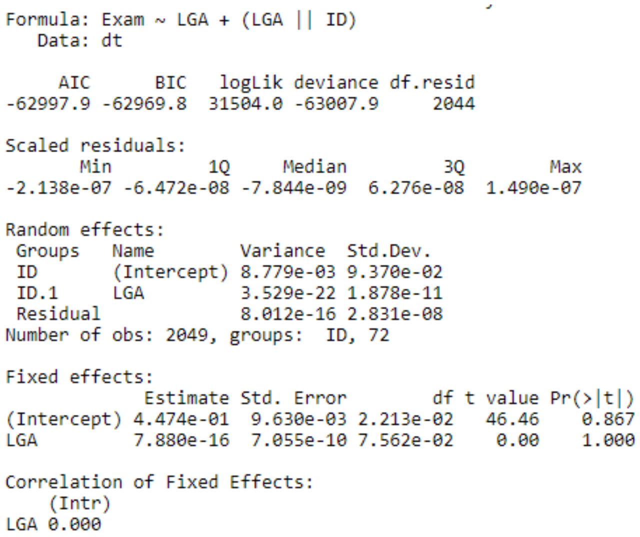
In the correlated model, we assume the random effects is correlated to the fixed effect, which means the difference appeared in learning goal achievement is correlated to the difference among individuals.



```
Formula: Exam ~ LGA + (LGA | ID)
   Data: dt

     AIC      BIC   logLik deviance df.resid
-59760.0 -59726.2  29886.0 -59772.0     2043

Scaled residuals:
       Min         1Q     Median         3Q        Max
-1.553e-06 -1.077e-07 -1.795e-08  9.574e-08  1.906e-06

Random effects:
 Groups   Name        Variance  Std.Dev.  Corr
 ID       (Intercept) 8.758e-03 9.358e-02
          LGA         8.308e-04 2.882e-02 0.00
 Residual             1.377e-15 3.711e-08
Number of obs: 2049, groups:  ID, 72

Fixed effects:
             Estimate Std. Error        df t value Pr(>|t|)
(Intercept) 0.4687308  0.0112362 0.0003151  41.716    0.997
LGA         0.0002299  0.0033377 0.0357678   0.069    0.988

Correlation of Fixed Effects:
    (Intr)
LGA -0.025
```

Variance for the random effect is very small, and correlation for the fixed effects is -0.025, which indicates that the Exam result will decrease as LGA increases.

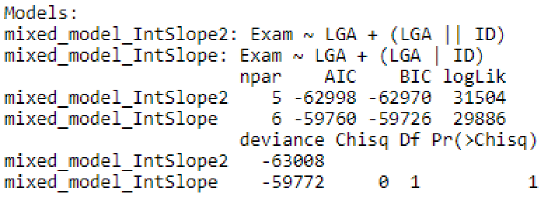
Then we apply uncorrelated model, where we assume the random effect - the difference in subject - is uncorrelated to the fixed effect, which means the correlation between within-group and between-group is 0.



```
Formula: Exam ~ LGA + (LGA || ID)
   Data: dt

     AIC      BIC   logLik deviance df.resid
-62997.9 -62969.8  31504.0 -63007.9     2044

Scaled residuals:
       Min         1Q     Median         3Q        Max
-2.138e-07 -6.472e-08 -7.844e-09  6.276e-08  1.490e-07

Random effects:
 Groups   Name        Variance  Std.Dev.
 ID       (Intercept) 8.779e-03 9.370e-02
 ID.1     LGA         3.529e-22 1.878e-11
 Residual             8.012e-16 2.831e-08
Number of obs: 2049, groups:  ID, 72

Fixed effects:
             Estimate Std. Error        df t value Pr(>|t|)
(Intercept) 4.474e-01  9.630e-03 2.213e-02   46.46    0.867
LGA         7.880e-16  7.055e-10 7.562e-02    0.00    1.000

Correlation of Fixed Effects:
    (Intr)
LGA 0.000
```

The correlation for the fixed effects is 0, which means under this condition LGA is completely uncorrelated to the Exam result.

Finally, we use ANOVA test to compare these two models, and the resulted Chisquare of the test is 0, which corresponds to a pvalue of 1, which indicates the correlated model is a better fit compared to the uncorrelated model. Hence, the effect of LGA on Exam result is correlated to the difference of LGA between subjects.

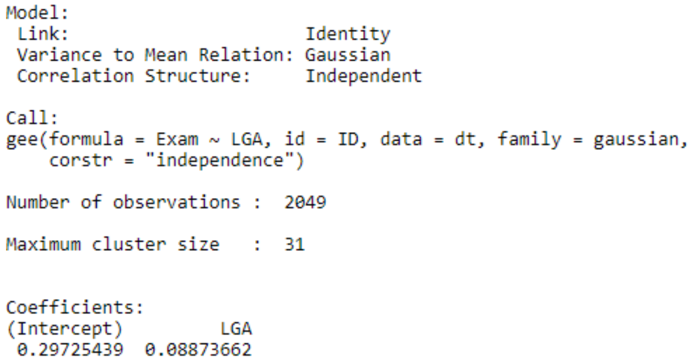


```
Models:
mixed_model_IntSlope2: Exam ~ LGA + (LGA || ID)
mixed_model_IntSlope: Exam ~ LGA + (LGA | ID)
                      npar    AIC    BIC logLik
mixed_model_IntSlope2    5 -62998 -62970  31504
mixed_model_IntSlope     6 -59760 -59726  29886
                      deviance Chisq Df Pr(>Chisq)
mixed_model_IntSlope2   -63008
mixed_model_IntSlope    -59772     0  1          1
```

Resulted Chisquare is 0

In the uncorrelated model, the LGA has a correlation around -0.025 with Exam result, which means the more learning goals achieved, the worse the grade, which is very counter intuitive. I believe linear mixed model is not a good choice here, since the explanatory variable only has four choice and the type of the outcome is binary. Therefore, the nature of the data type induces a lot of error on the result of the correlation, and we can't fully trust the result here.

**Generalized Estimation Equation** Generalized Estimation Equation(GEE) can apply to all different types of outcome, including continuous, count, and binary outcome, and this will fix the problem caused by the last section.



```
Model:
 Link:                      Identity
 Variance to Mean Relation: Gaussian
 Correlation Structure:     Independent

Call:
gee(formula = Exam ~ LGA, id = ID, data = dt, family = gaussian,
    corstr = "independence")

Number of observations :  2049

Maximum cluster size   :  31


Coefficients:
(Intercept)         LGA
 0.29725439  0.08873662
```

The Generalized Estimation Equation resulted in a coefficient of 0.089 for LGA

The result produced by the GEE is much more accurate, and coefficient for LGA (0.089) indicates that the exam result will be better as more learning goal is achieved. This coefficient looks very small, and it's probably due to the nature of the binary outcome, where exam result can only take 0 or 1. Therefore, this small coefficient doesn't mean the impact of the LGA is weak.

To conclude, we will adopt the result of the Generalized Estimation Equation, and we can prove that there is a positive correlation between LGA and Exam result. However, I believe the result of the this analysis is not entirely accurate, since the data of the exam is only divided into pass and no pass, so we can't differentiate the grade A, B, C or D. Therefore, the result can't clear indicate the relationship between learning goal achievement and exam grade.

## 4.2 Within-person level

### 4.2.1 Q5

To study the connections between higher sleep quality and learning goals achievement on a day-to-day level, we apply the mixed effect model. We plot the regression model for each person and intend to see the relationships between sleep quality and learning goals achievement.

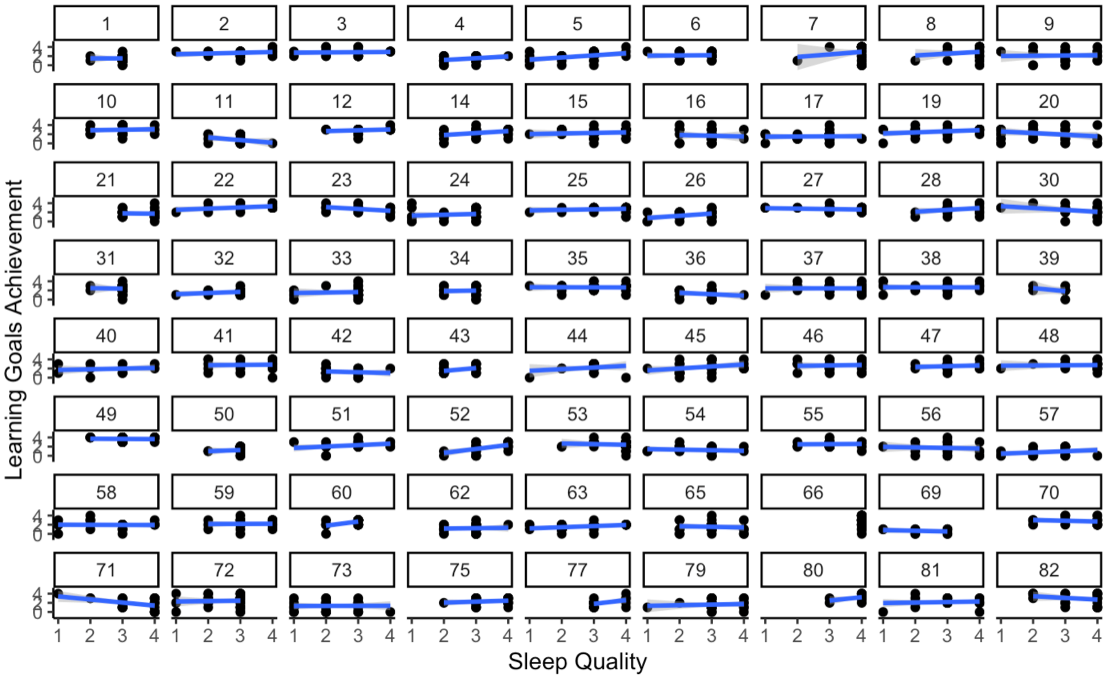
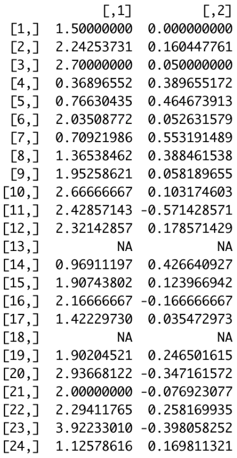


Figure 17. The plot of sleep quality and learning goals achievement on a day-to-day level.

From the regression model, We can't observe a systematic relationship between the subject's sleep quality and learning goals achievement of each of 72 people, so in aim to further estimate their relationships, we print out their coefficients of intercepts and slope as shown below.



```
             [,1]         [,2]
 [1,]  1.50000000  0.000000000
 [2,]  2.24253731  0.160447761
 [3,]  2.70000000  0.050000000
 [4,]  0.36896552  0.389655172
 [5,]  0.76630435  0.464673913
 [6,]  2.03508772  0.052631579
 [7,]  0.70921986  0.553191489
 [8,]  1.36538462  0.388461538
 [9,]  1.95258621  0.058189655
[10,]  2.66666667  0.103174603
[11,]  2.42857143 -0.571428571
[12,]  2.32142857  0.178571429
[13,]          NA           NA
[14,]  0.96911197  0.426640927
[15,]  1.90743802  0.123966942
[16,]  2.16666667 -0.166666667
[17,]  1.42229730  0.035472973
[18,]          NA           NA
[19,]  1.90204521  0.246501615
[20,]  2.93668122 -0.347161572
[21,]  2.00000000 -0.076923077
[22,]  2.29411765  0.258169935
[23,]  3.92233010 -0.398058252
[24,]  1.12578616  0.169811321
```

Figure 18. Result of the coefficients of sleep quality and learning goals achievement (partial). The first column is the intercept while the second column is the slope estimates accorsing to regression.

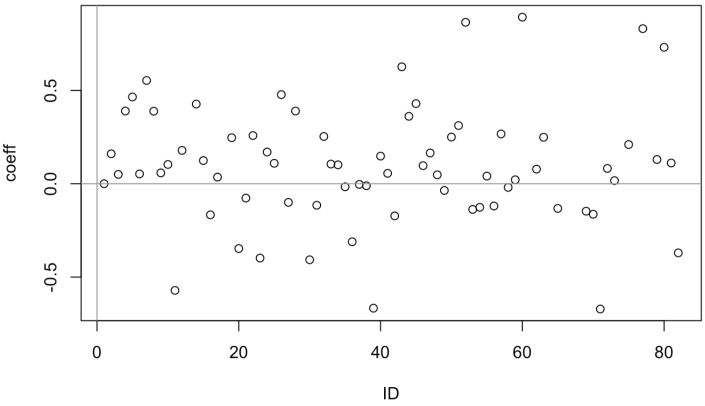


Figure 19. Scatter plot of the coefficients of sleep quality and learning goals achievement

According to both the coefficients and the figure, we find that although most of the estimates are positive, there are still negative estimates of slope, so we can hardly conclude that higher sleep quality will lead to higher learning goals achievement.

### 4.2.2 Q6

In this question, we wanted to investigate the connections between higher physical and learning goals achievement on a day-to-day level. Thus, we first took a close look of the relationship pattern between these two variables for each individual.

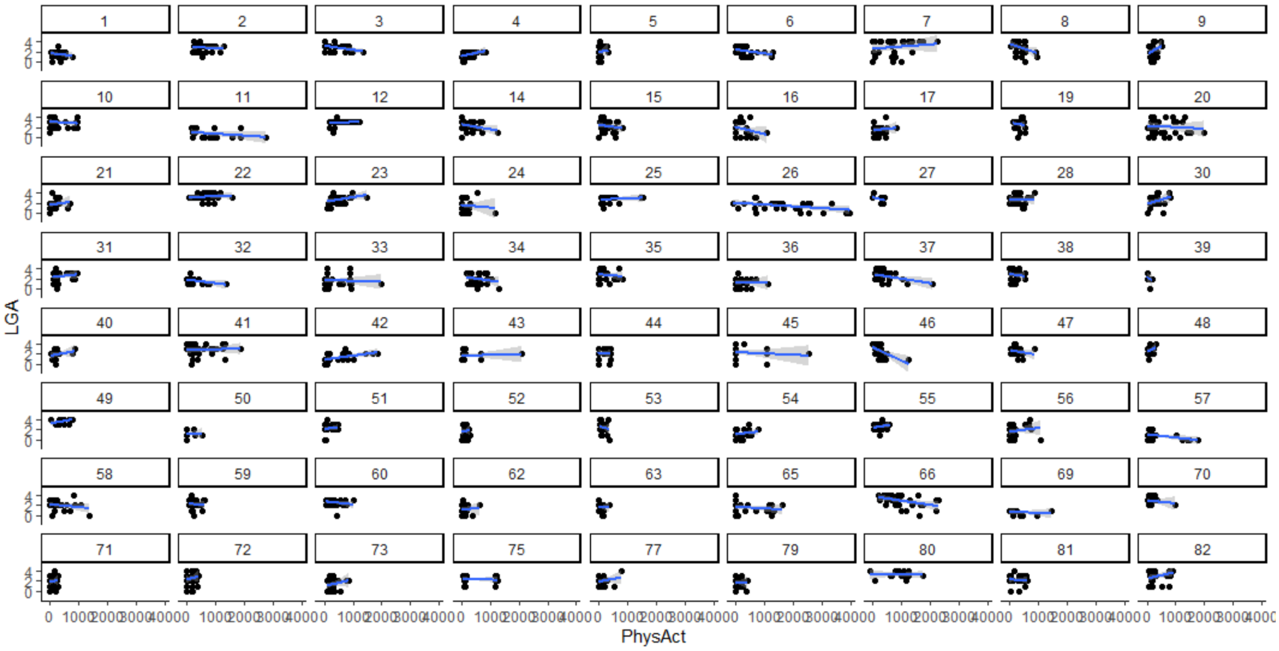
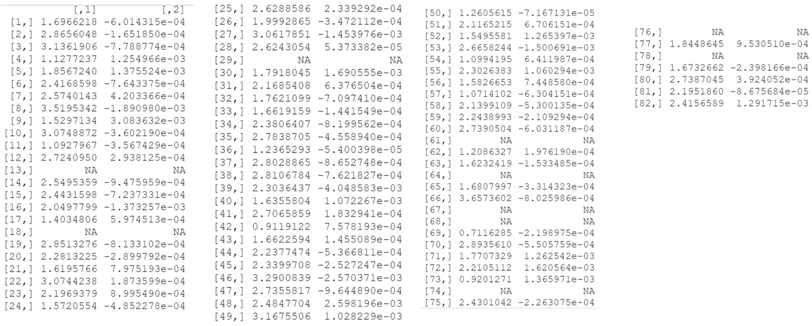


Figure 16. Linear relationship graphs between physical activity and learning goals achievement for each subject

As the graph above showed, all 71 people had very different linear relationship between higher physical and learning goals achievement. We could not find the common pattern directly through the graphs. Individual 26, for example, had a long descending linear liner, whereas individual 21 had a short ascending linear line. Therefore, we did calculation for the coefficient of these two variables for each individual using linear regression model.



```
              [,1]          [,2]
 [1,] 1.6966218 -6.014315e-04
 [2,] 2.8656048 -1.651850e-04
 [3,] 3.1361906 -7.788774e-04
 [4,] 1.1277237  1.254966e-03
 [5,] 1.8567240  1.375524e-03
 [6,] 2.4168598 -7.643375e-04
 [7,] 2.5740143  4.203366e-04
 [8,] 3.5195342 -1.890980e-03
 [9,] 1.5297134  3.083632e-03
[10,] 3.0748872 -3.602190e-04
[11,] 1.0927967 -3.567429e-04
[12,] 2.7240950  2.938125e-04
[13,]        NA            NA
[14,] 2.5495359 -9.475959e-04
[15,] 2.4431598 -7.237331e-04
[16,] 2.0497799 -1.373257e-03
[17,] 1.4034806  5.974513e-04
[18,]        NA            NA
[19,] 2.8513276 -8.133102e-04
[20,] 2.2813225 -2.899792e-04
[21,] 1.6195766  7.975193e-04
[22,] 3.0744238  1.873599e-04
[23,] 2.1969379  8.995490e-04
[24,] 1.5720554 -4.852278e-04
```

```
[25,] 2.6288586  2.339292e-04
[26,] 1.9992865 -3.472112e-04
[27,] 3.0617851 -1.453976e-03
[28,] 2.6243054  5.373382e-05
[29,]        NA            NA
[30,] 1.7918045  1.690555e-03
[31,] 2.1685408  6.376504e-04
[32,] 1.7621099 -7.097410e-04
[33,] 1.6619159 -1.441549e-04
[34,] 2.3806407 -8.199562e-04
[35,] 2.7838705 -4.558940e-04
[36,] 1.2365293 -5.400398e-05
[37,] 2.8028865 -8.652748e-04
[38,] 2.8106784 -7.621827e-04
[39,] 2.3036437 -4.048583e-03
[40,] 1.6355804  1.072267e-03
[41,] 2.7065859  1.832941e-04
[42,] 0.9119122  7.578193e-04
[43,] 1.6622594  1.455089e-04
[44,] 2.2377474 -5.366811e-04
[45,] 2.3399708 -2.527247e-04
[46,] 3.2900839 -2.570371e-03
[47,] 2.7355817 -9.644890e-04
[48,] 2.4847704  2.598196e-03
[49,] 3.1675506  1.028229e-03
```

```
[50,] 1.2605615 -7.167131e-05
[51,] 2.1165215  6.706151e-04
[52,] 1.5495581  1.265397e-03
[53,] 2.6658244 -1.500691e-03
[54,] 1.0994195  6.411987e-04
[55,] 2.3026383  1.060294e-03
[56,] 1.5826653  7.448580e-04
[57,] 1.0714102 -6.304151e-04
[58,] 2.1399109 -5.300135e-04
[59,] 2.2438993 -2.109294e-04
[60,] 2.7390504 -6.031187e-04
[61,]        NA            NA
[62,] 1.2086327  1.976190e-04
[63,] 1.6232419 -1.533485e-04
[64,]        NA            NA
[65,] 1.6807997 -3.314323e-04
[66,] 3.6573602 -8.025986e-04
[67,]        NA            NA
[68,]        NA            NA
[69,] 0.7116285 -2.198975e-04
[70,] 2.8935610 -5.505759e-04
[71,] 1.7707329  1.262542e-03
[72,] 2.2105112  1.620564e-03
[73,] 0.9201271  1.365971e-03
[74,]        NA            NA
[75,] 2.4301042 -2.263075e-04
```

```
[76,]        NA            NA
[77,] 1.8448645  9.530510e-04
[78,]        NA            NA
[79,] 1.6732662 -2.398166e-04
[80,] 2.7387045  3.924052e-04
[81,] 2.1951860 -8.675684e-05
[82,] 2.4156589  1.291715e-03
```

Figure 17. Result of the coefficients of physical activity and learning goals achievement (partial)

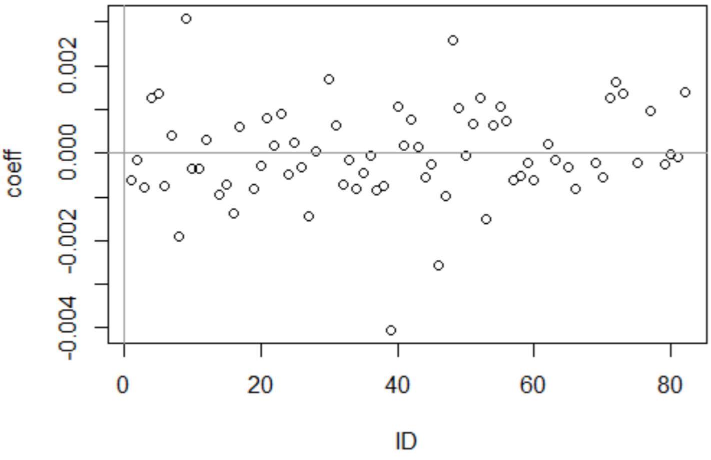


Figure 18. Scatter plot of the coefficients of physical activity and learning goals achievement

From the graph above, we could see that the coefficients of the linear regression model lies on both sides of y=0 almost uniformly. No obvious trend was discovered at this point. Thus, we found almost no connection between physical activity and learning goals achievement

#### 4.2.3 Q7

In this question, we need to find if daily positive/negative affect mediates the relation between sleep quality and learning goal achievement as well as the relation between physical activity and learning goal achievement. On other words, we test if when students have higher sleep quality, the PA (Positive Affect) is higher which predicts better LGA (Learning Goal Achievement)

We first graph the relations between sleeping quality and daily positive affect for each of our students. By observing the graph directly, it is hard to find relations between the two variables. Check figure 19.

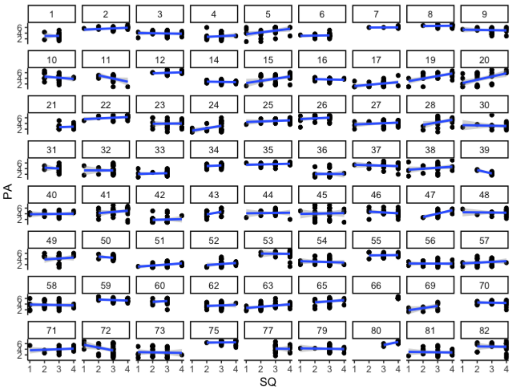


Figure 19. The plot of relations between sleep qualities and daily positive affect

Then we take a look at the graph of the slopes of each linear regression line. We obtain an average slope of 0.13654 for each r. It means that in average Positive Affect would increase 0.13654 if Sleeping Quality increase by 1 for each student.

We repeat this process to the relations between Physical Activities and Daily Positive Affect. We get an average slope of 0.00098. It means that in average Positive Affect would increase 0.00098 if Sleeping Quality increase by 1 for each student. By similar approach, we found that when Positive Affect = increase by 1,learning goal achievement will increase in average by 0.16551 for each student.

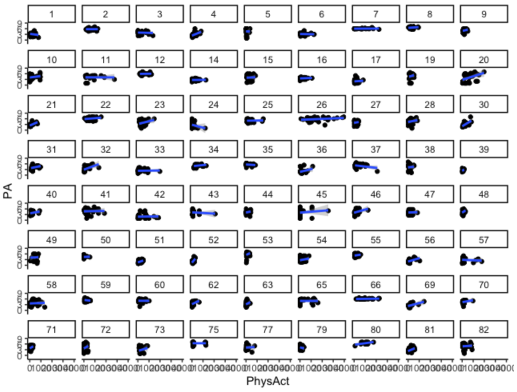


Figure 20. The plot of relations between physical activities and daily positive affect

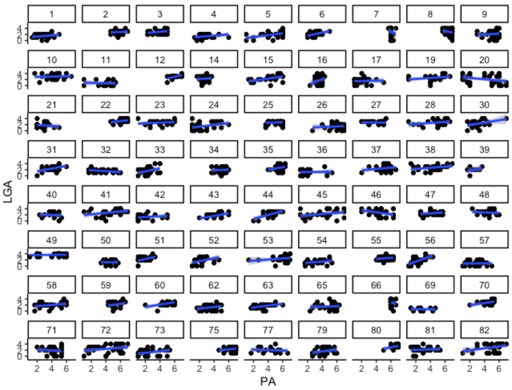


Figure 21. The plot of relations between daily positive affect and learning goal achievement

As a result, we have the indirect effect of sleeping quality within individuals by 0.13654*0.16551 = 0.02260. We have the indirect effect of physical activities within individuals by 0.00098*0.16551 = 0.00016, which is relatively small.

# 5 Conclusion

In the investigation section, we found following results of between-person level and within-person level through methods of mixed effect model, between-subjects model and generalized estimation equation. They worked well for longitudinal data.

For the between-person level, we cannot conclude that higher physical activity level predicts better learning goal achievement as the p-value do not provide statistical support. According to the Sobel test and statistically values, the average negative and positive affects mediate the relation between average sleep quality and average learning goal achievement, but the average positive affect. We got that negative effect doesn't mediate the relationship between physical activity and LGA. Through the use of ANOVA and Generalized Estimation Equation, we prove that there is a positive correlation between LGA and Exam result.

From the within-person perspective, the connections between higher sleep quality and learning goals achievement are limited. Moreover, the connections between higher physical and learning goals achievement are not apparent as well. Then, daily positive/negative affect mediates the relation between sleep quality and learning goal achievement as well as the relation between physical activity and learning goal achievement is approved by taking the average of each slope of the individuals' relation graph. By comparing individual student's sleeping quality/physical activities and daily positive affect, we could conclude that both would increase daily positive affect of students. By comparing and computing the relations between daily positive affect and the learning goal achievement, we found that both sleeping quality and physical activities have positive indirect affect on learning goal achievement. Even though the effect might by insignificant, daily positive affect does mediate between physical activity and learning goal achievement as well as between sleeping quality and learning goal achievement

# 6 Theory

## 6.1 Mean

Mean equals to the average value of data in your sample or population. Such value can always be basically categorized sample mean and population mean, and the sample mean is always used to estimate the situation of the whole population. The sample mean can be an unbaised estimator as it can be the least squares estimator according to the Gauss Markov theorem.

## 6.2 Standard Error

The estimator for standard error are defined as the following:

$$S\hat{E}(\bar{x}) = \frac{\sqrt{\bar{x}(1-\bar{x})}}{\sqrt{n-1}} \frac{\sqrt{N-n}}{\sqrt{N}} \tag{1}$$

## 6.3 (Multivariate) Linear Regression

Multivariate linear regression is used to estimate the association between multiple explanatory variables and a response variable. The model comes in the following format:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \tag{2}$$

Several important assumptions are made regarding to this model:

- The response variable is linearly dependent on all of covariates $x_1, ..., x_p$.
- All of the observations $y_1, ..., y_n$ are independent.
- $x_j$ is uncorrelated to the error $\epsilon$.
- The error$\epsilon$ is normally distributed.
- The covariates $x_1, ..., x_p$ are not highly correlated.
- The mean and variance of $\epsilon_i$ are the same for i.

## 6.4 Confidence Interval

This is a theory used to cover a range of possible values varying from the estimated mean. Using this theory, we can uncover an interval that we are highly confident that the population mean is within it. This is an estimating approach to predict true mean from sample.

$$samplemean + / - standarderror * t - score \tag{3}$$

## 6.5 Mixed Effects Model

### 6.5.1 Within-subjects Model

$$Y_{ij} = \beta_{i,0} + \beta_{i,1} X_{ij} + \epsilon_{ij} \tag{4}$$

where $Y_{ij}$ denote the outcome for subject i in the j-th measurement, $X_i j$ denote the covariates for subject i in the j-th meastuement, and $\epsilon_{ij} \sim N(0, \sigma^2)$ are the error terms.

For within-subjects model we assume that $\beta_{i,0}$, $\beta_{i,1}$ are parameters (fixed constants) for individual i. Therefore, we have 2N unknown parameters. Specifically, we are assuming the following models:

$$\begin{gathered} \mathrm{Y}_{1j} = \beta_{1,0} + \beta_{1,1} X_{1j} + \epsilon_{1j} \\ \mathrm{Y}_{2j} = \beta_{2,0} + \beta_{2,1} X_{2j} + \epsilon_{2j} \\ ... \\ \mathrm{Y}_{nj} = \beta_{n,0} + \beta_{n,1} X_{1j} + \epsilon_{nj}. \end{gathered}$$

### 6.5.2 Between-subjects Model

For the Between-subjects Model for Mixed Effect Model, it follows the equations below:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + b_{i,0} + b_{i,1} X_{ij} + \epsilon_{ij}$$

where Yij indicates the outcome for subject i in the j-th measurement, Xij is the covariates for subject i in the jth measure, and

$$\begin{pmatrix} b_{i,0} \\ b_{i,1} \end{pmatrix} \sim N\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} D_{00} & D_{01} \\ D_{10} & D_{11} \end{pmatrix} \right)$$

and given

$$\epsilon_{ij} \sim_{i.i.d} N(0, \sigma^2).$$

### 6.6 Generalized Estimation Equation

Generalized Estimation Equation tries to estimate the parameters of the generalized linear model where is the correlation between outcomes might be unknown.It focuses on estimating the average response in the given population. The following is the equation used for estimation:

$$U(\beta) = \sum_{i=1}^{N} \frac{\partial \mu_i}{\partial \beta} V_i^{-1} \{Y_i - \mu_i(\beta)\}$$

where

$$\mu_{ij}$$

is the mean for subject i at time j

$$\beta$$

is the regression parameter, and

$$i$$

is the regression structure used.

### 6.7 ANOVA test

ANOVA test is a statistical model attempts to compare means of different groups. Based on the total law of variance, it partitioned the variance according to its source, and it generalizes the t-test further beyond two means.

### 6.8 Sobel Test

H$_0$ : $ab = 0(wherea, bareindirecteffects)$

$$T = \frac{\hat{a}\hat{b}}{\sqrt{(\hat{b} * exp(2))(SE(\hat{a}) * exp(x) + (\hat{a} * exp(2))(SE(\hat{b}) * exp(x))}} \tag{5}$$

When the test statistic is larger than 1.96, the hypothesis will be rejected. This is a very conservative test.

## 7 Appendix

### 7.1 Contribution

1. Shuyang Zhang: Did Question 6 and Conclusion
2. Zetong Lai: Question 3 and introduction
3. Yibei Cai: Question 4 and review of the code
4. Yuan Lin: Question 7 and Background
5. Junqian Liu: Question 5 and data
6. Yiteng Lu: Question 1 and 2: first two analysis at the between-person level and review the report.

### 7.2 Citation

1. SYVÄOJA, HEIDI J.1,2; KANTOMAA, MARKO T.1,3; AHONEN, TIMO2; HAKONEN, HARTO1; KANKAANPÄÄ, ANNA1; TAMMELIN, TUIJA H.1 Physical Activity, Sedentary Behavior, and Academic Performance in Finnish Children, Medicine Science in Sports Exercise: November 2013 - Volume 45 - Issue 11 - p 2098-2104 doi: 10.1249/MSS.0b013e318296d7b8

2. "Generalized Estimating Equation." Wikipedia, Wikimedia Foundation, 9 Nov. 2020, `en.wikipedia.org/wiki/Generalized_estimating_equation`.

3. "Analysis of Variance." Wikipedia, Wikimedia Foundation, 21 Feb. 2021, `en.wikipedia.org/wiki/Analysis_of_variance.`

4. lab9-Yuyao.nb.html