STA108 FINAL EXAM ANALYSIS AND ALL PLOTS

Priyanshi Singh

This paper aims to identify the variables that best predict GPA based on the dataset collected and given. Using a random subset of 200 samples, I interpreted predictive models using variables such as sleep quality, cognitive ability, stress levels, and other lifestyle factors. I used a variety of methods but focused on forward selection and backward elimination to identify the most significant predictors.

I initially started by creating scatterplots and getting their correlation coefficients to assess the linear relationships between GPA and each of the given predictor variables.[Figures 1-11]. However, this was not very helpful as all of the correlations were weak, with the highest coefficient being 0.1676 for CognitionZScore. I then conducted t-tests for each variable to try to determine the potential significance of the model [Figures 12-22]. All of the variables had a p-value below the 0.05 threshold, suggesting that they all had some potential significance and relation to GPA. I then tried to look at the residual and QQ plots for each variable [Figures 23-44] to determine if a linear model was not appropriate and if any transformations were necessary. In this process, preliminary single-predictor models showed some heteroscedasticity and non-normality for some variables- ClassesMissed [Figures 23-24], PoorSleepQuality [Figures 27-28], AnxietyScore [Figures 31-32], Drinks, [Figures 39-40], and WeekdaySleep [Figures 43-44]. I then did Box-Cox transformations for these variables [Figures 45-49] to determine if there was a linear relationship after the transformation, but the multiple R-squared values, p-values, and other statistical values (the R-squared values, p-values, and others) were inconclusive and did not suggest that any one predictor variable had a significant effect.

I decided then to use another model - forward selection and backward elimination. These methods helped me iteratively evaluate each predictor based on their respective

contributions toward predicting and explaining GPA variation. Note that I had to remove the DASScore variable when running this model due to its multicollinearity likely to do its correlations with the Depression, Anxiety, and Stress Scores.

1. Forward Selection: I started off with having no predictors in the model. Then, with respect to how the forward selection process works, each variable was added stepwise based on their respective p-values and how much contribution they have to predictive accuracy for GPA. This process identified CognitionZScore, StressScore, DepressionScore, ClassesMissed, and AnxietyScore as the key predictions that helped contribute significantly to GPA [Figure 51] as seen by the addition of the star for each layer.
2. Backward Elimination: I started off with having all predictors in this model, and each predictor was removed iteratively based on which one contributed the least to the mode. This approach once again retained CognitionZScore, StressScore, DepressionScore, ClassesMissed, and AnxietyScore into its final model, reinforcing the predictive power of these variables for the model. Some other variables such as AnxietyScore and WeekdaySleep were removed towards the beginning of the process due to their minimal impact on the improvement of the adjusted R-squared values.

Diagnostic plots for the final model were conducted to assess the adjusted R-squared value [Figure 53]. We can see that the value is higher when all 6 predictors are used compared to some combination of them, and the r-adjusted squared value is 0.1561.

The final model includes CognitionZScore, CognitionZScore, StressScore, DepressionScore, ClassesMissed, and AnxietyScore, all of which help to explain 15.61% of GPA variability (with the adjusted R-squared = 0.1561). The analysis identified these 6 as significant predictors of GPA. While the model's predictive power is still sub-par highlights key factors influencing academic performance.

STA108 FINAL EXAM ALL PLOTS

Priyanshi Singh

Scatterplots of all potential predictor variables with a least square line and a correlation coefficient

Figure 1

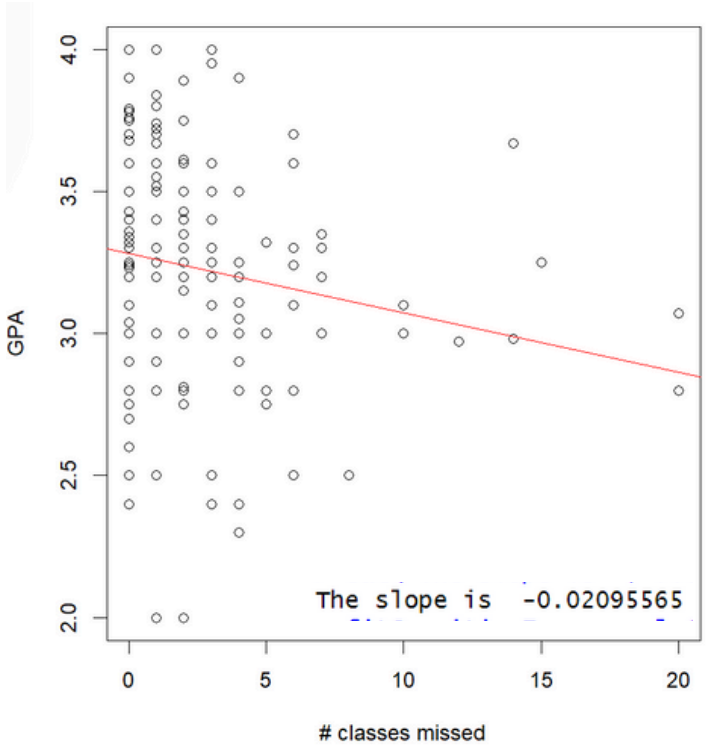


Figure 2:

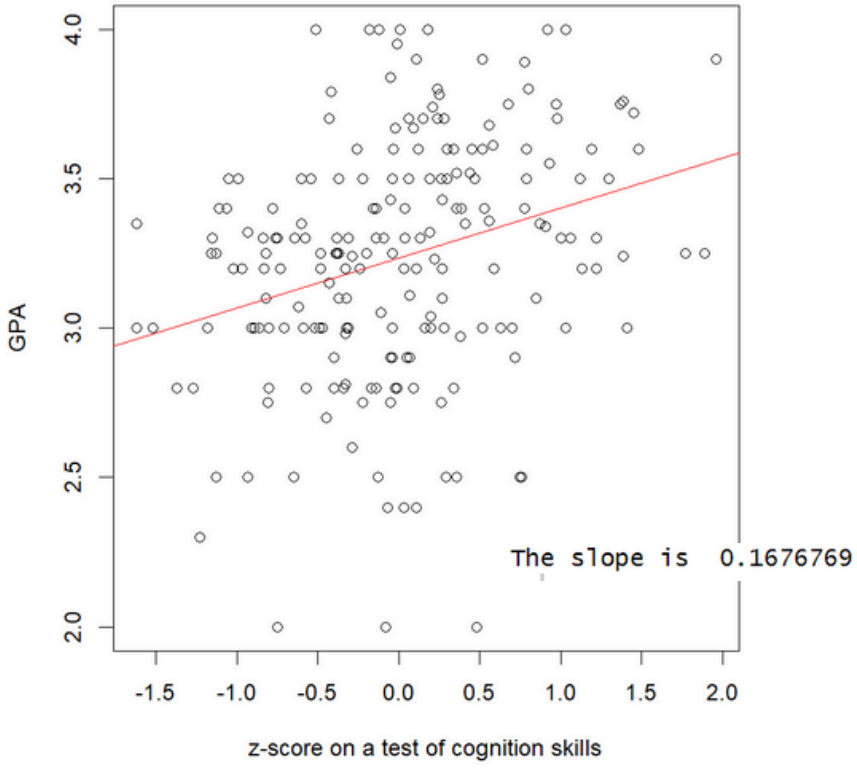


Figure 3

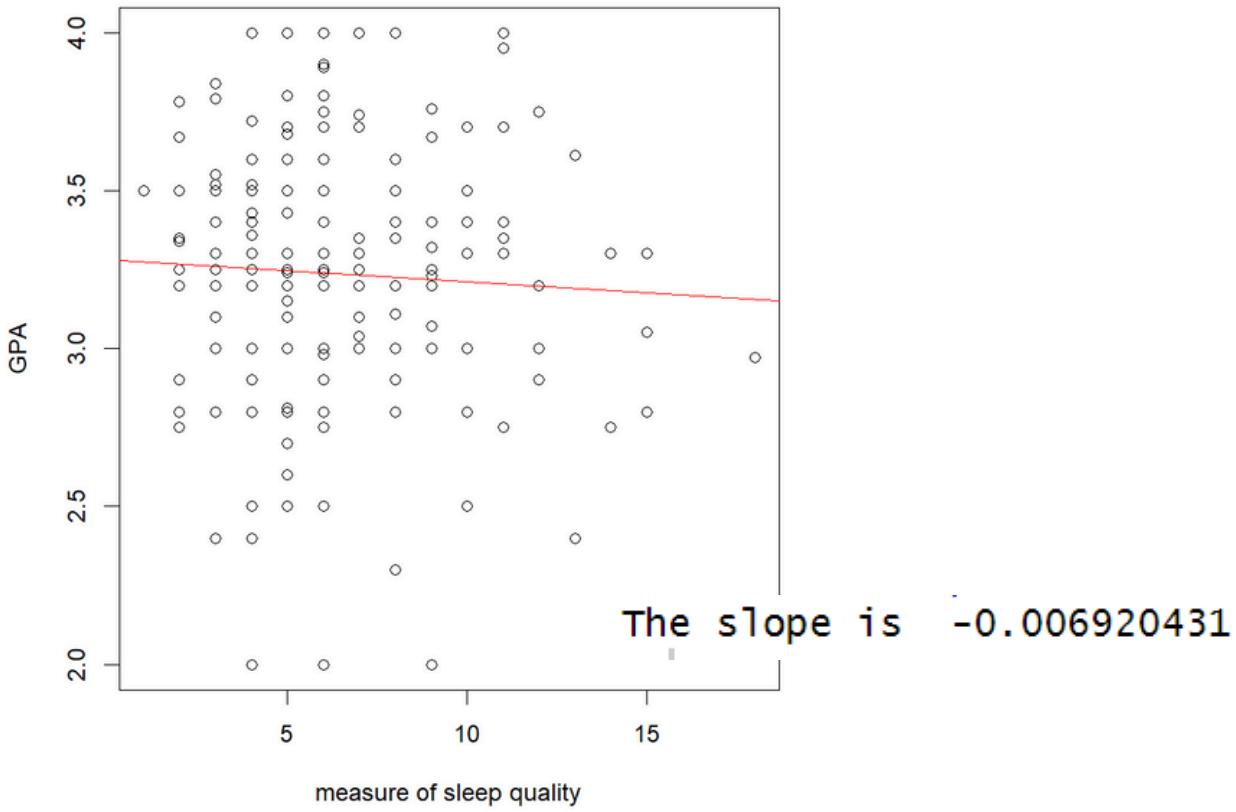


Figure 4:

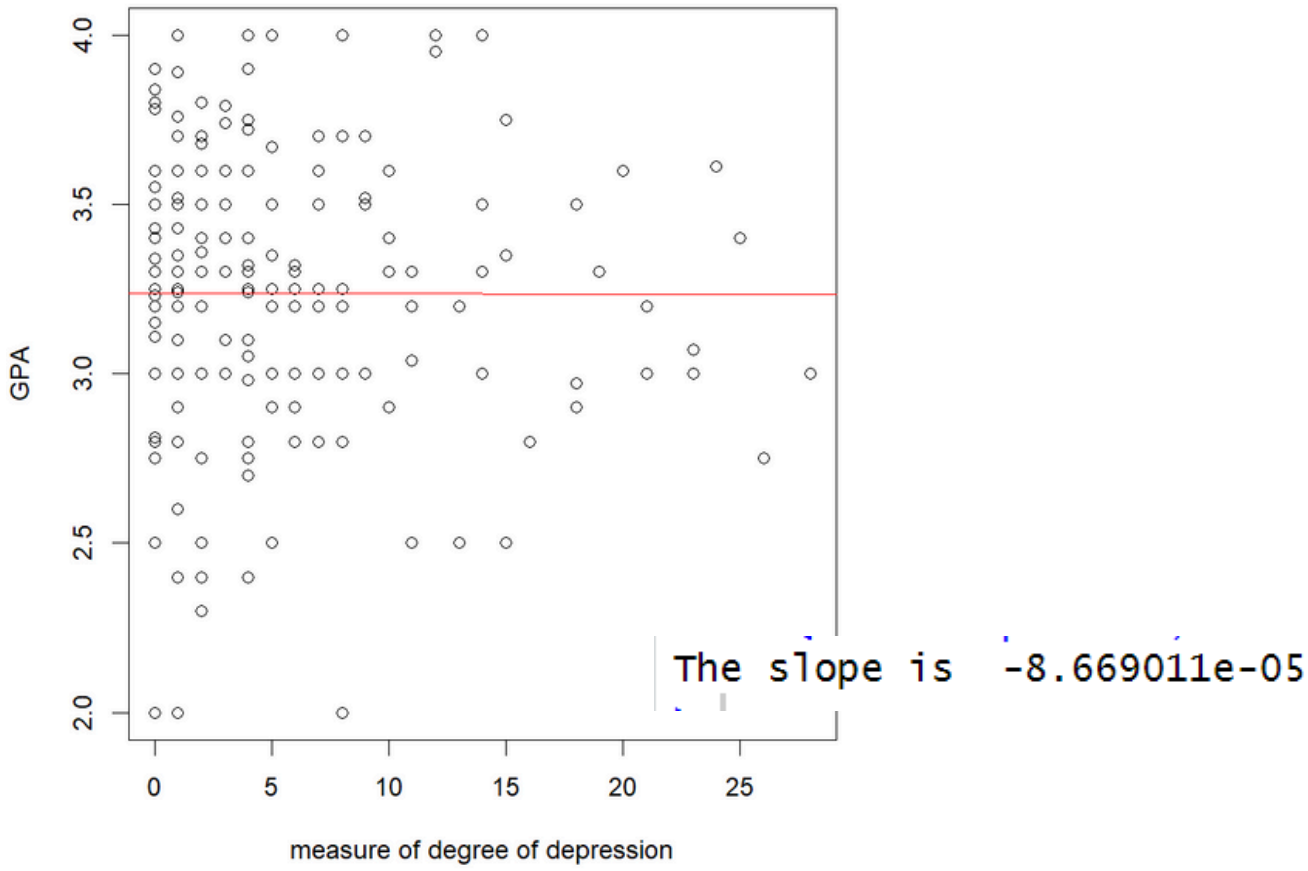


Figure 5

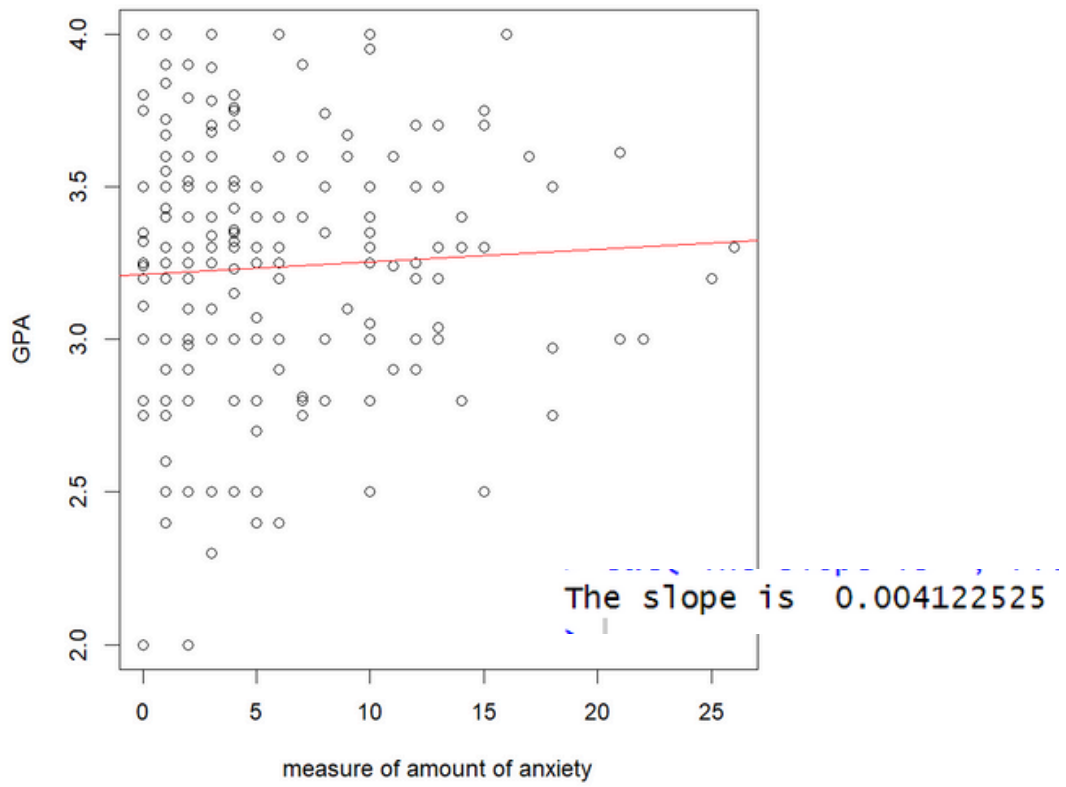


Figure 6

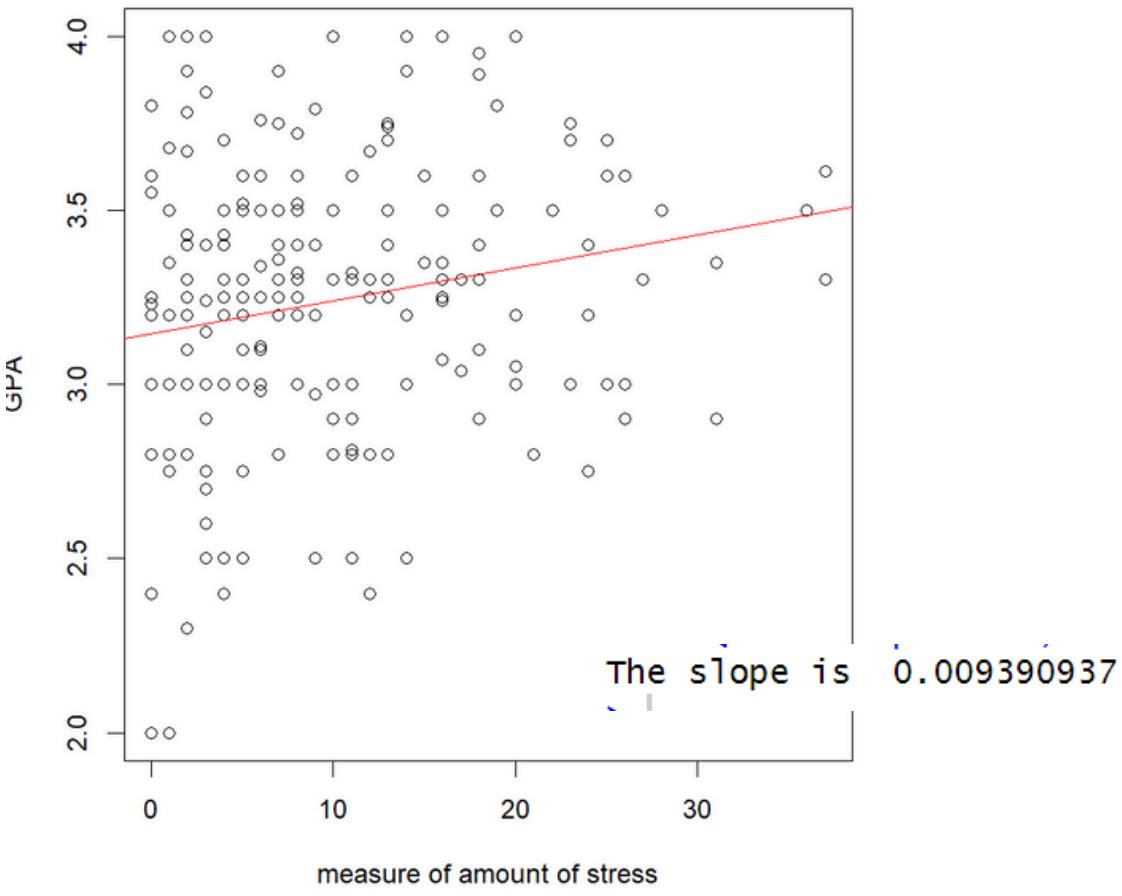


Figure 7

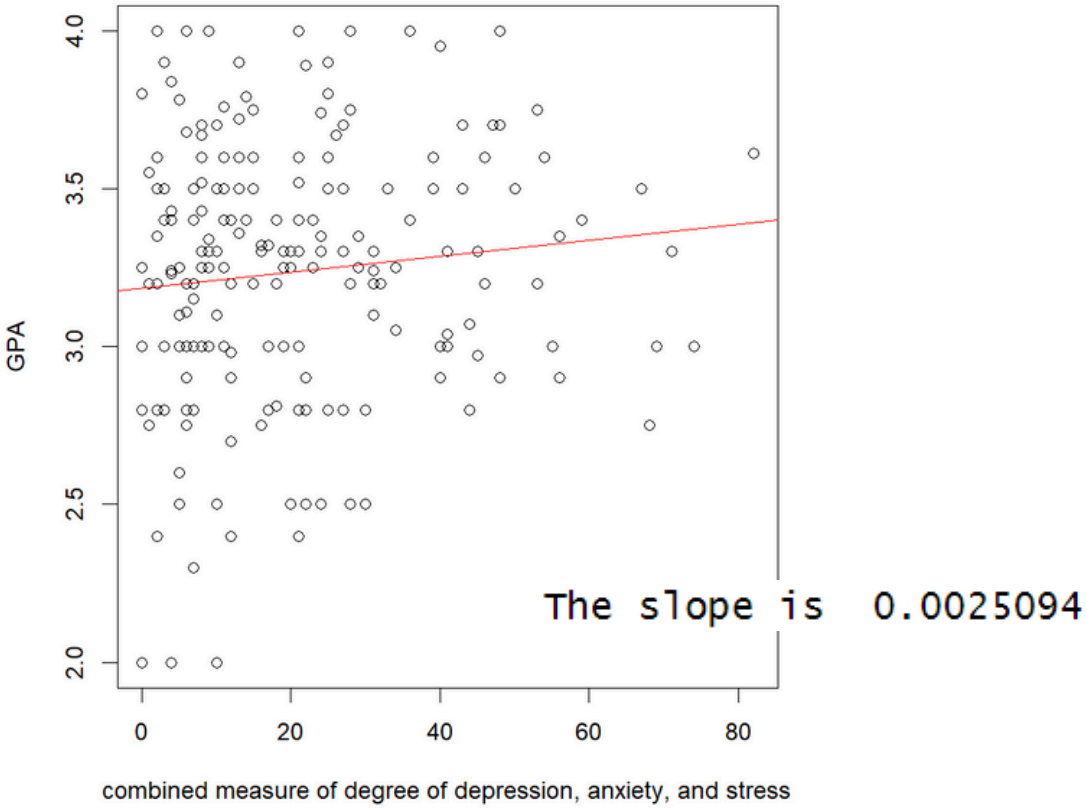


Figure 8

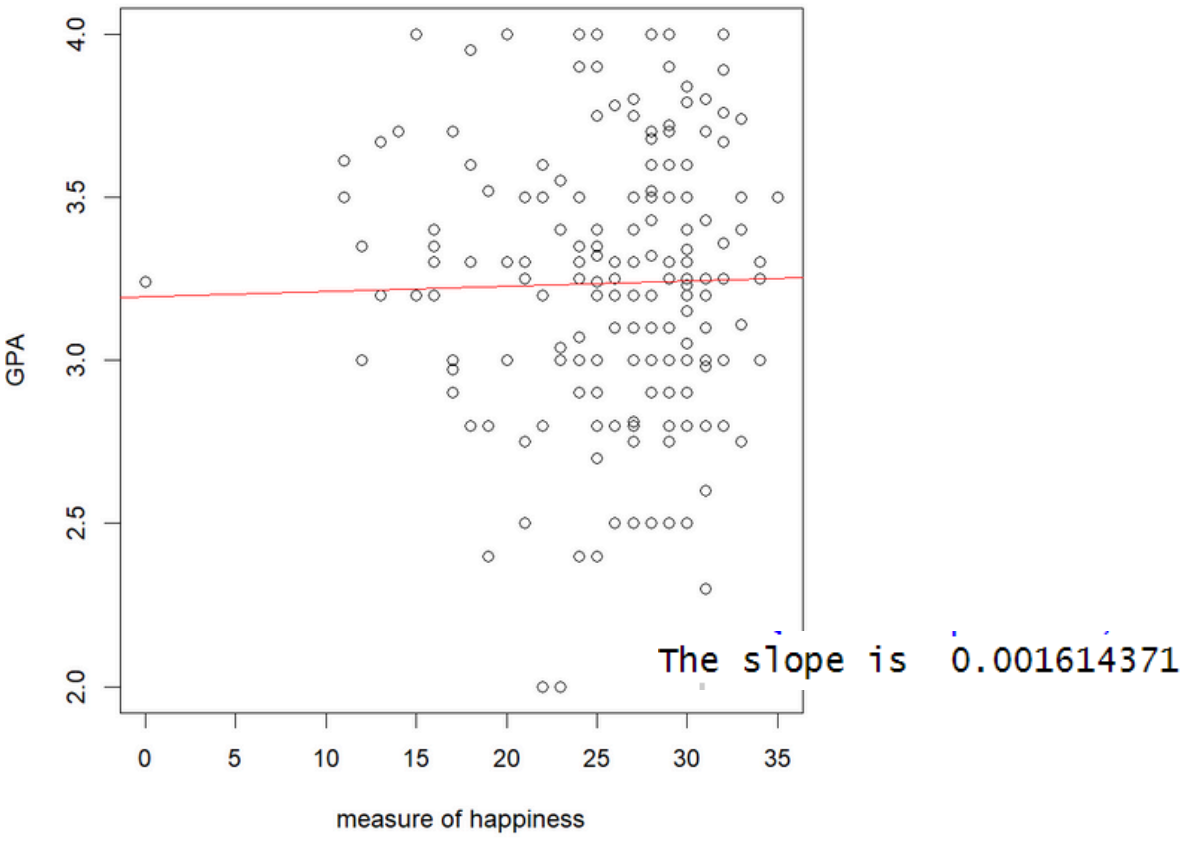


Figure 9

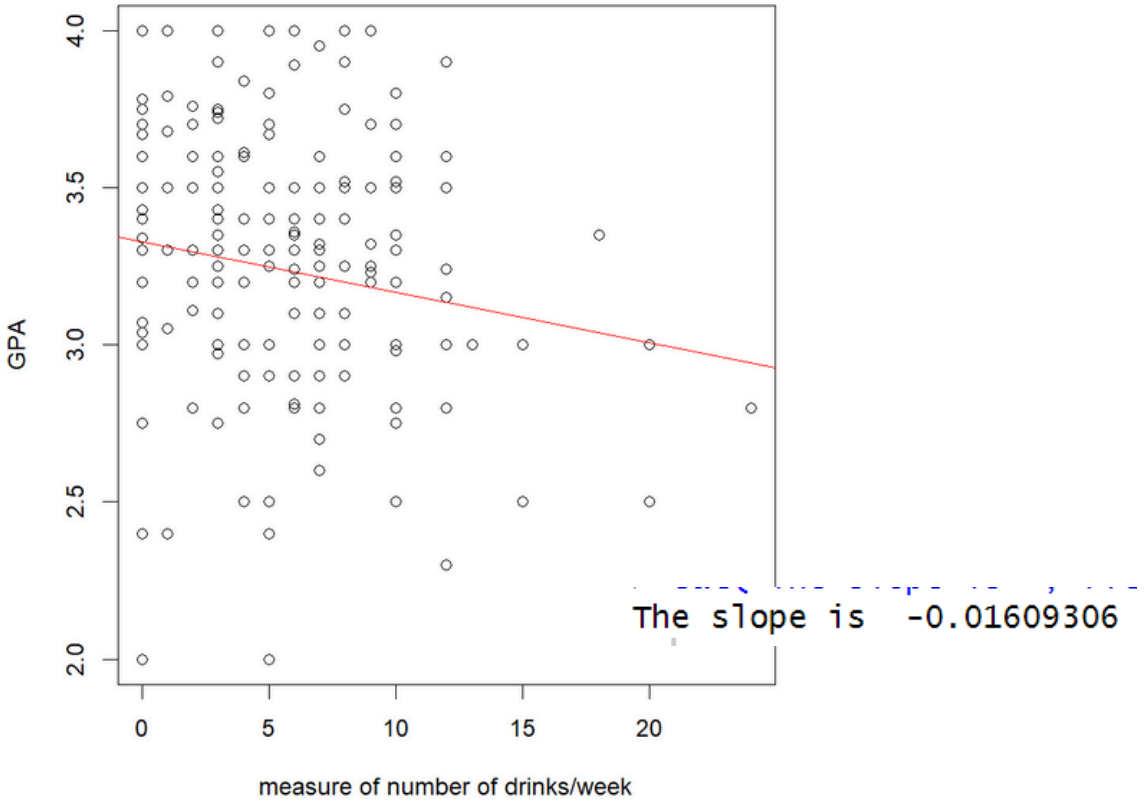


Figure 10

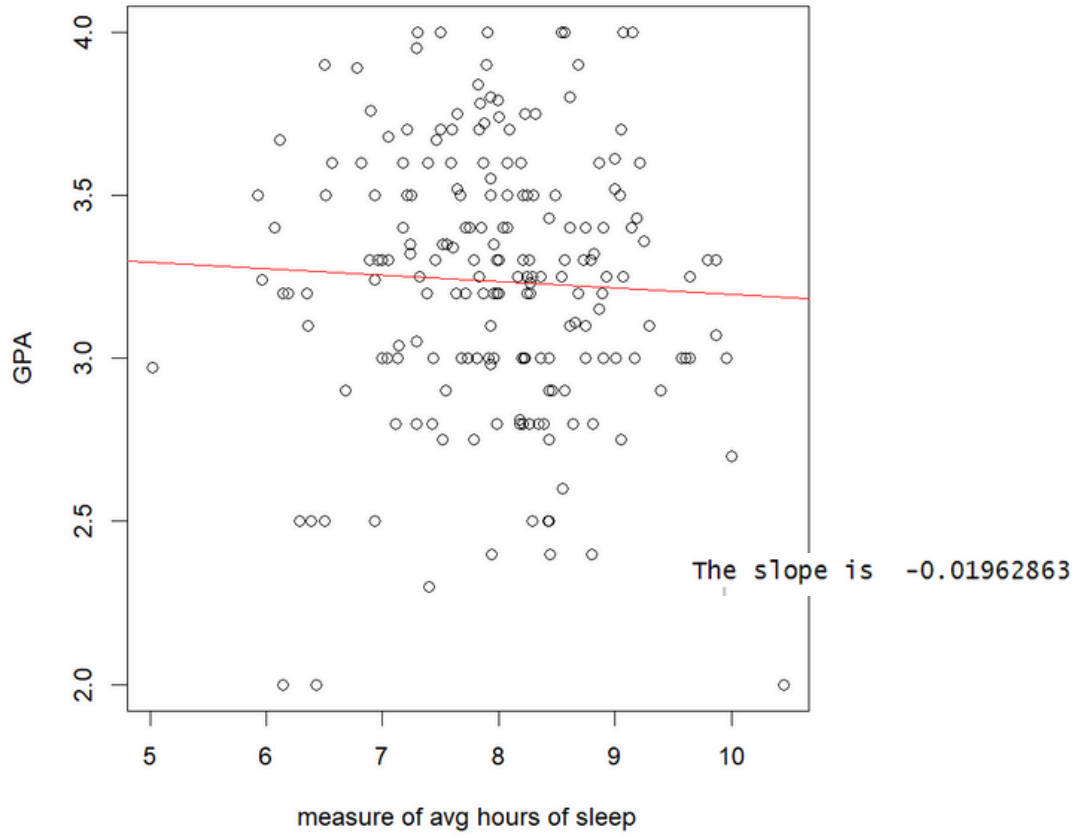
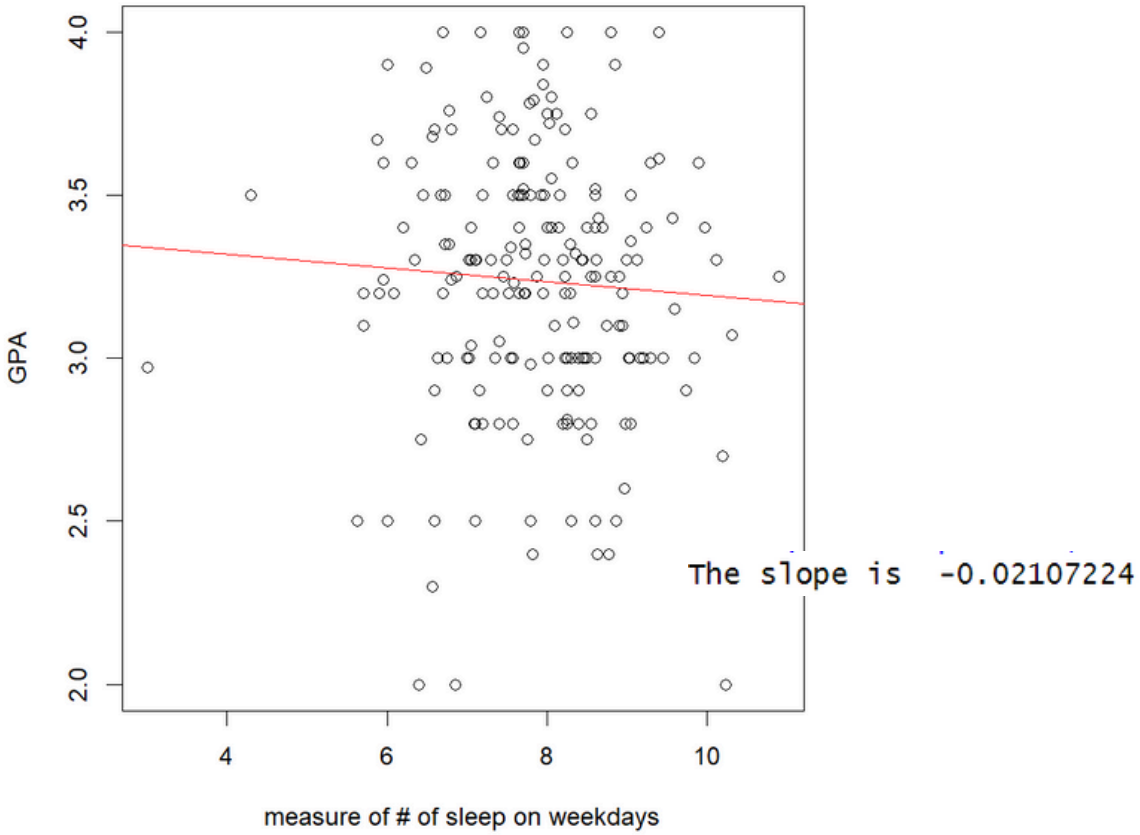


Figure 11


GPA
4.0
3.5
3.0
2.5
2.0
4
6
8
10
measure of # of sleep on weekdays
The slope is -0.02107224

Summaries/plots of t-test for each variable

Figure 12

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.28158331	0.034398219	95.399803	1.785604e-167
ClassesMissed	-0.02095565	0.008860931	-2.364949	1.899942e-02

Figure 13

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.2343322	0.02761312	117.130272	8.192750e-185
CognitionZscore	0.1676769	0.03914914	4.283029	2.873495e-05

Figure 14

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.280329169	0.068042590	48.2099401	2.313860e-111
PoorSleepQuality	-0.006920431	0.009600619	-0.7208318	4.718632e-01

Figure 15

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.236352e+00	0.038444283	84.18291030	5.61969e-157
DepressionScore	-8.669011e-05	0.004874573	-0.01778415	9.85829e-01

Figure 16

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.213514687	0.041104937	78.1783142	8.470118e-151
AnxietyScore	0.004122525	0.005397808	0.7637407	4.459309e-01

Figure 17

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.146310463	0.04361581	72.136922	4.075127e-144
StressScore	0.009390937	0.00347495	2.702467	7.480435e-03

Figure 18

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.1852603	0.044296247	71.908131	7.469826e-144
DASScore	0.0025094	0.001672068	1.500776	1.350066e-01

Figure 19

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.194031284	0.140404535	22.7487758	3.851873e-57
Happiness	0.001614371	0.005298152	0.3047046	7.609112e-01

Figure 20

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.32706721	0.048210833	69.01078	1.882581e-140
Drinks	-0.01609306	0.006868016	-2.34319	2.011182e-02

Figure 21

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.39243732	0.25530596	13.2877323	3.197164e-29
AverageSleep	-0.01962863	0.03180867	-0.6170842	5.378881e-01

Figure 22

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.40166264	0.21004055	16.1952663	4.014257e-38
WeekdaySleep	-0.02107224	0.02644851	-0.7967266	4.265641e-01

Residual Plots for fitted values vs residual plots & QQ plots

Figure 23

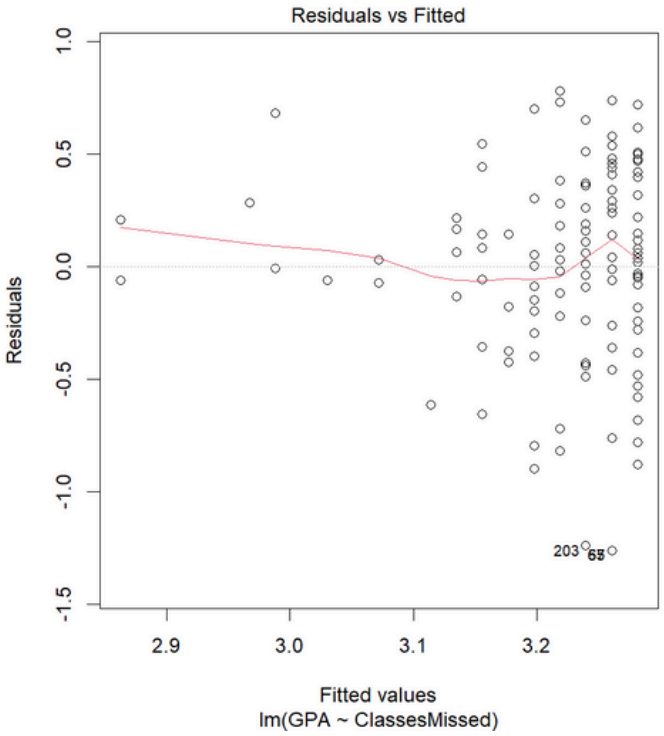


Figure 24

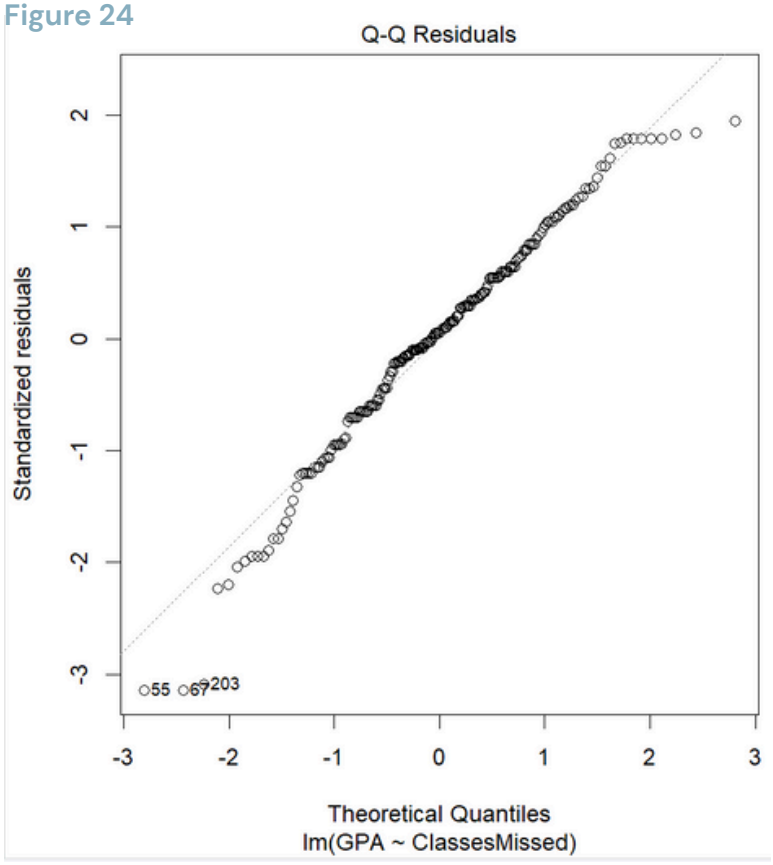


Figure 25

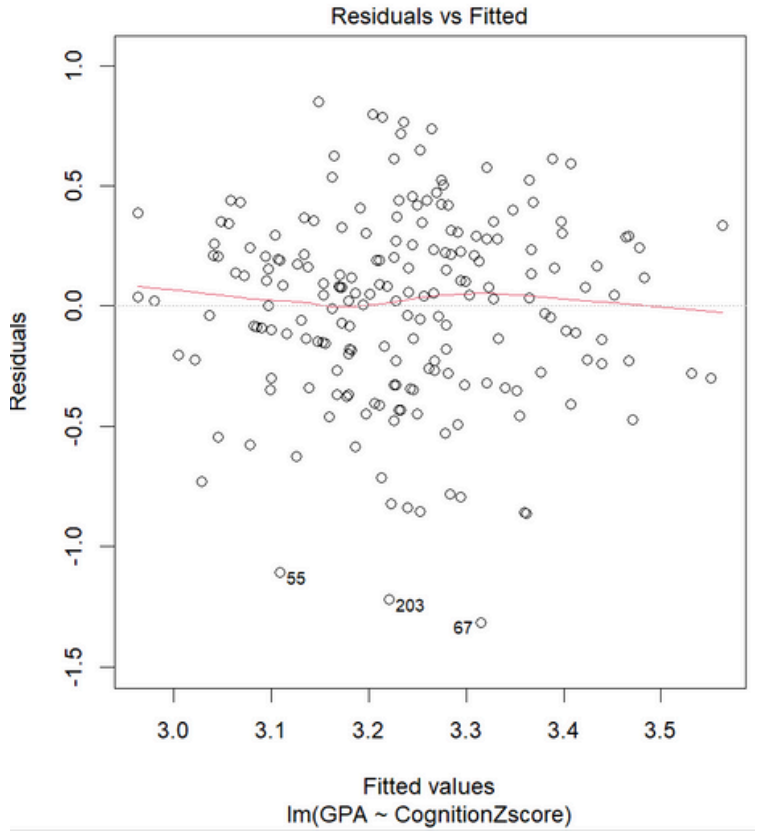


Figure 26

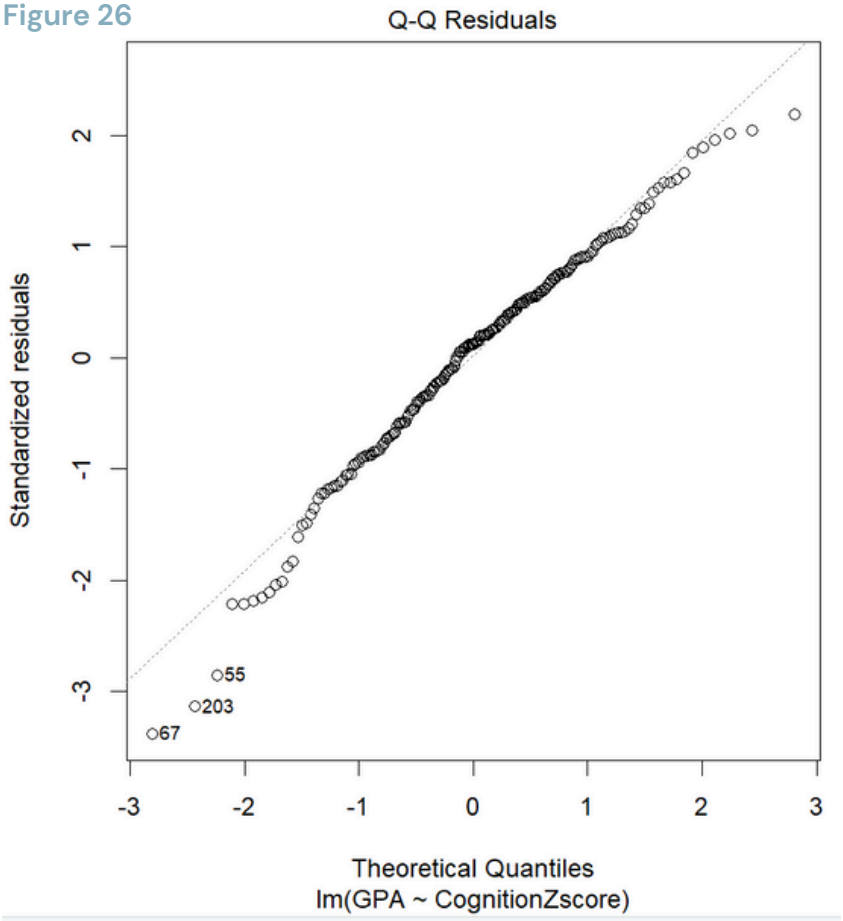


Figure 27

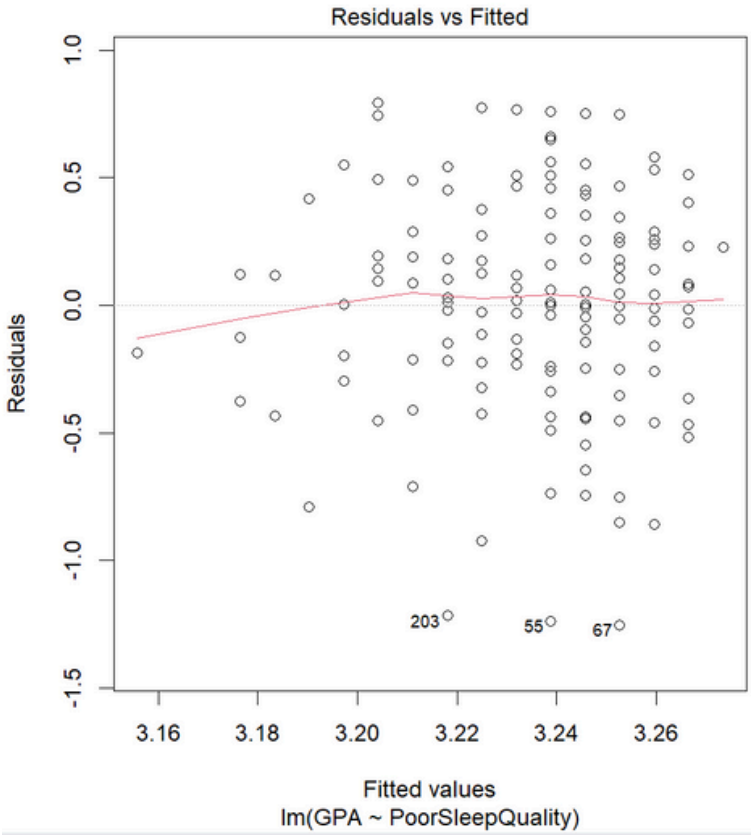


Figure 28

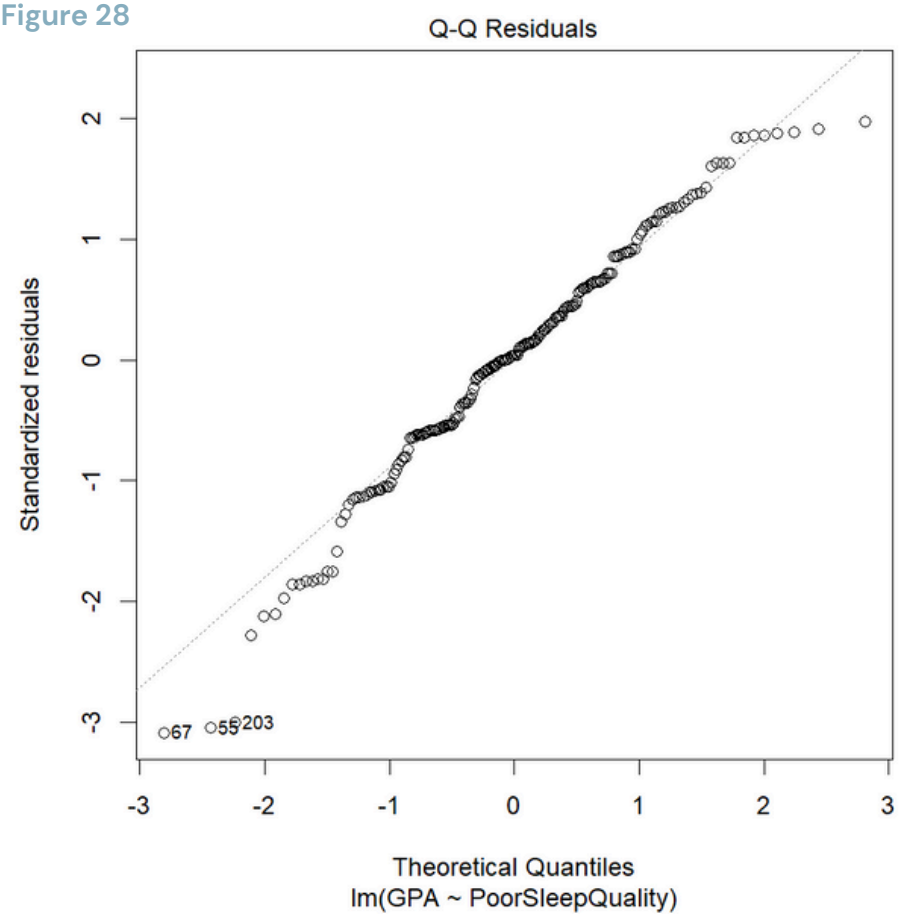


Figure 29

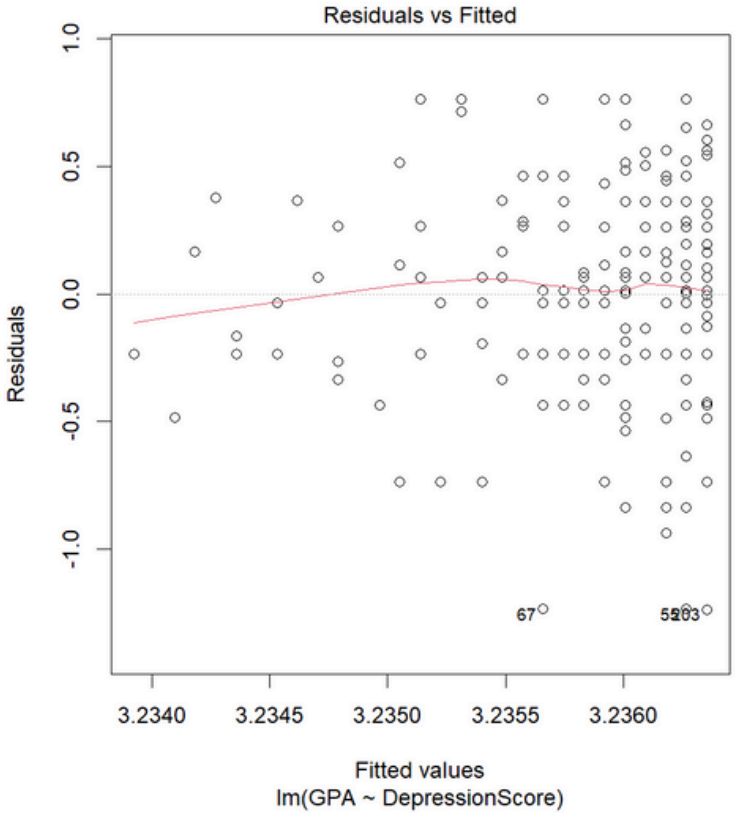


Figure 30

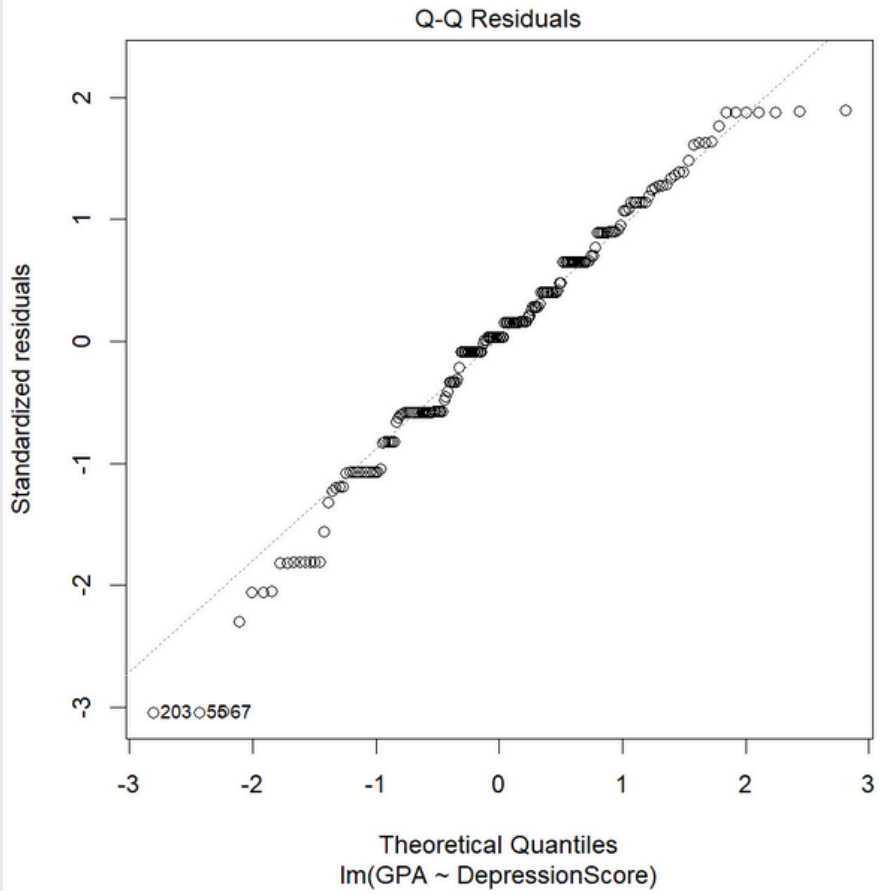


Figure 31

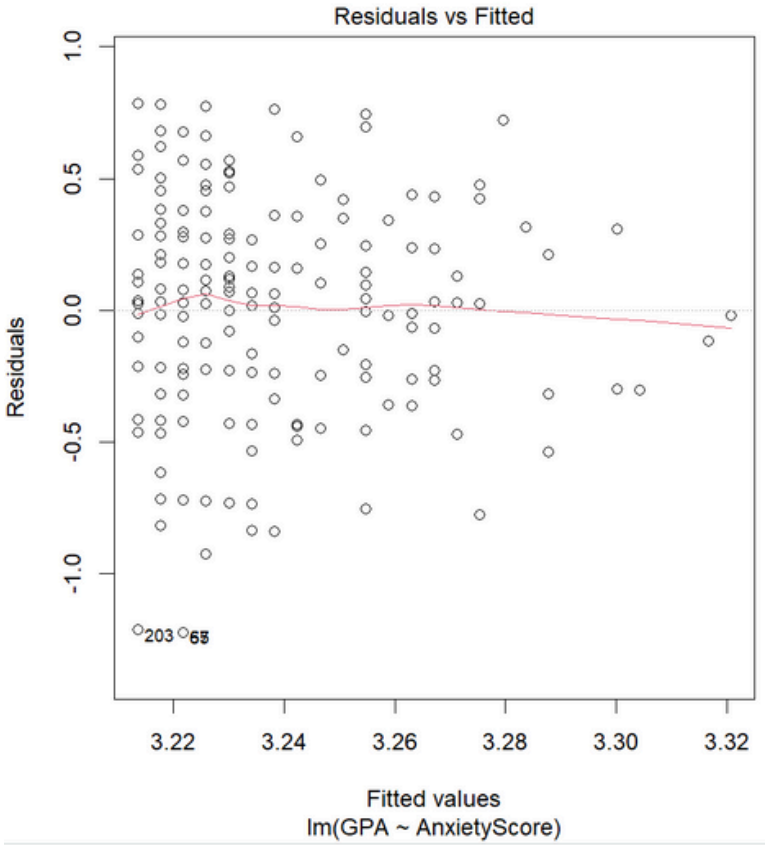


Figure 32

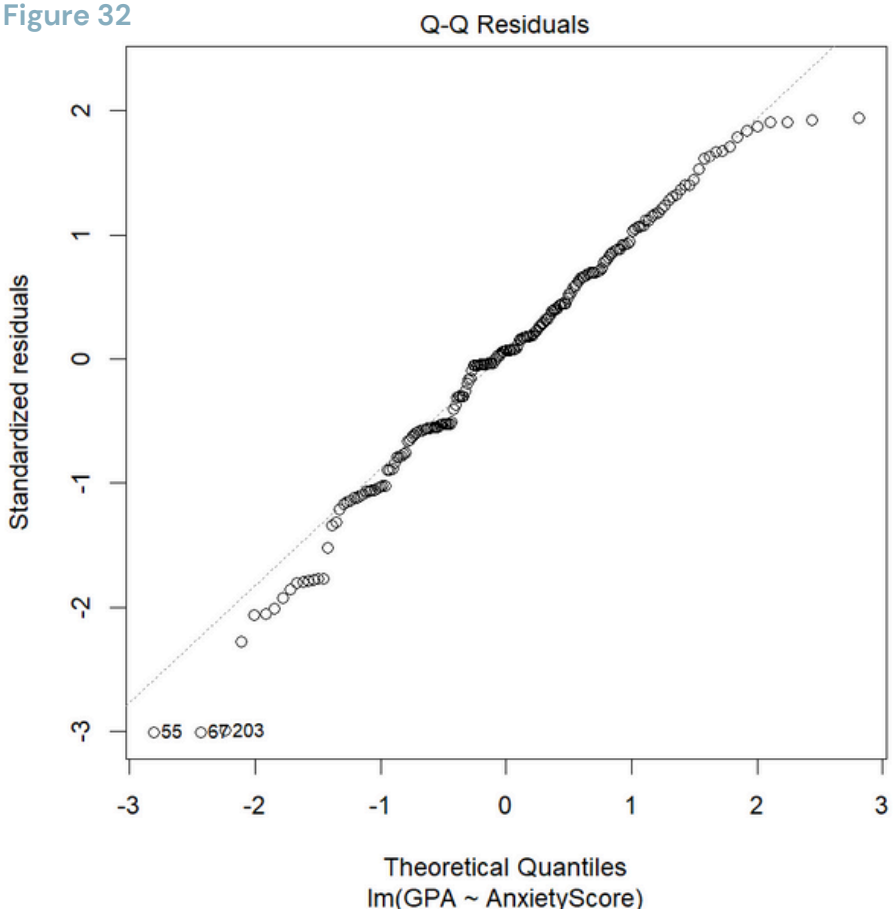


Figure 33

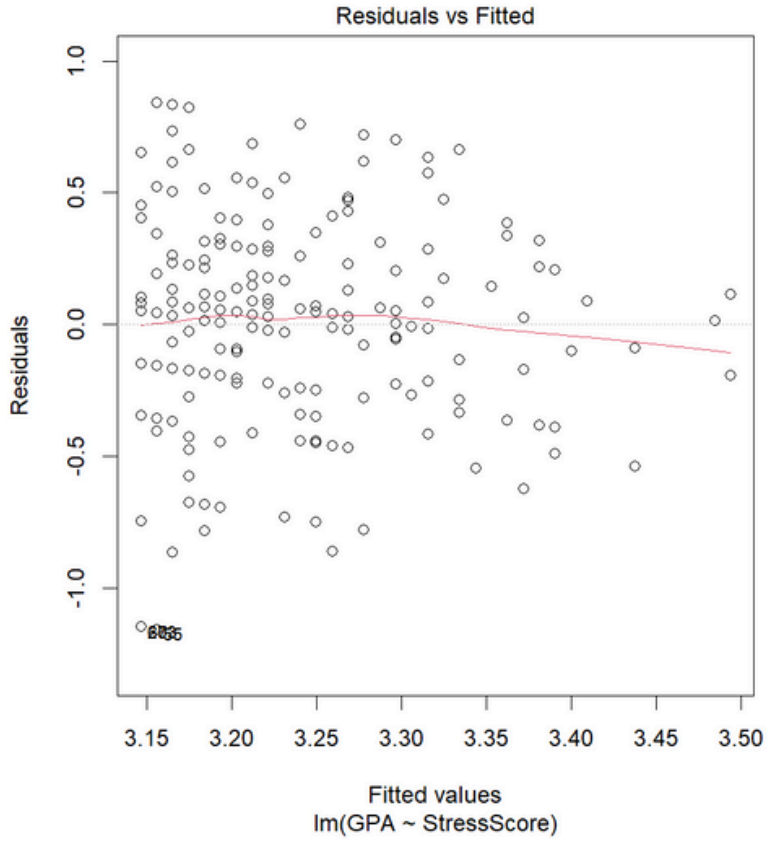


Figure 34

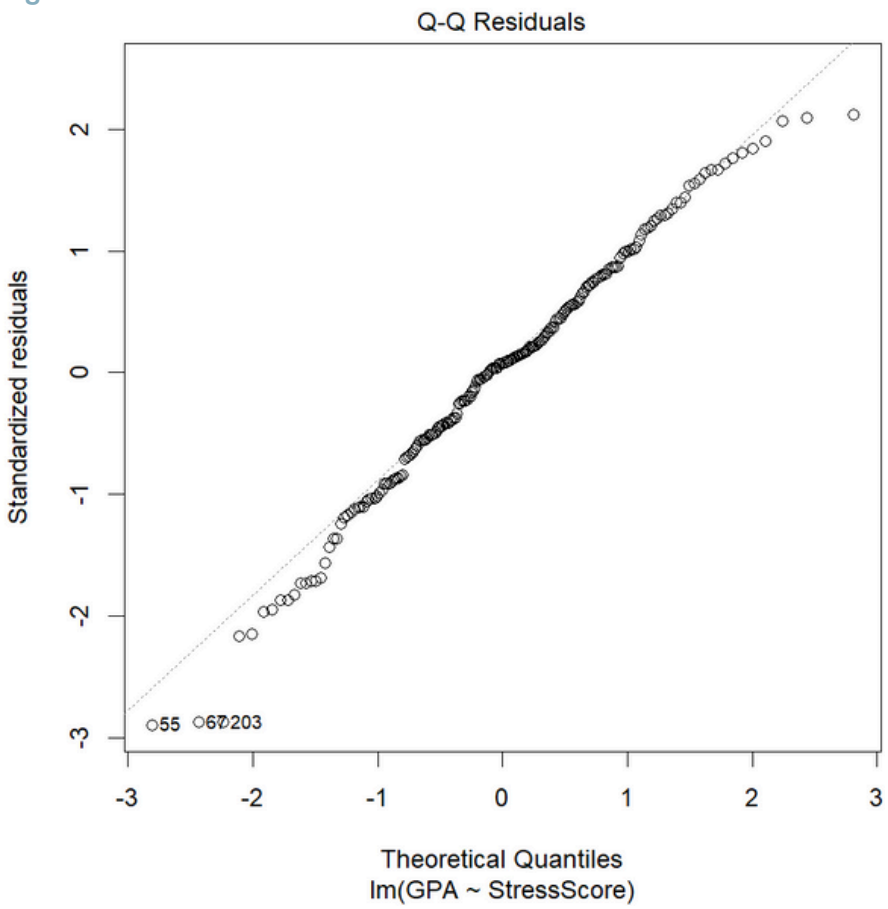


Figure 35

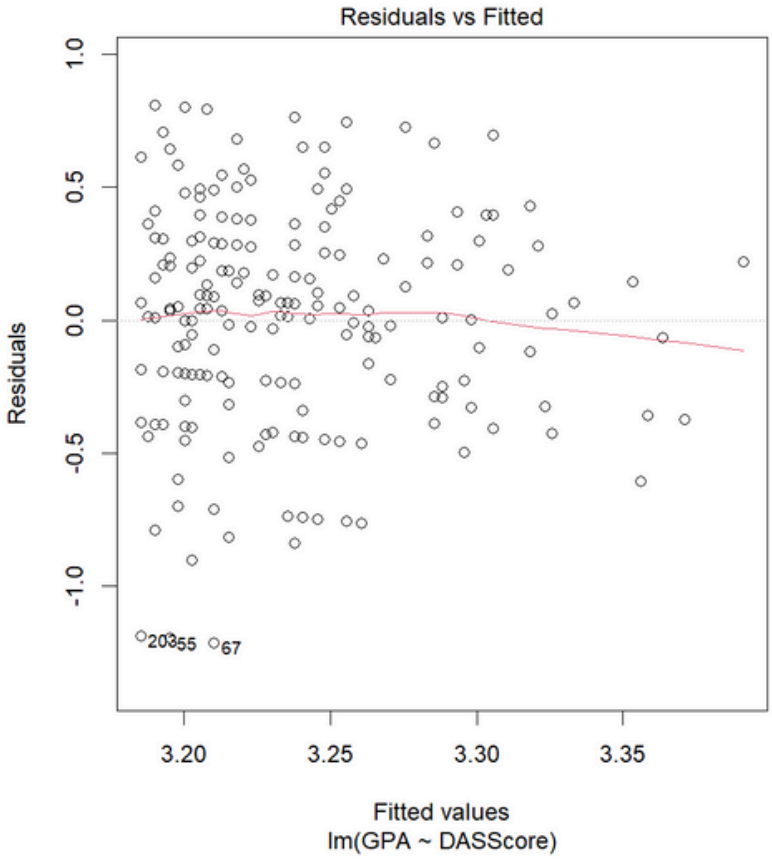


Figure 36

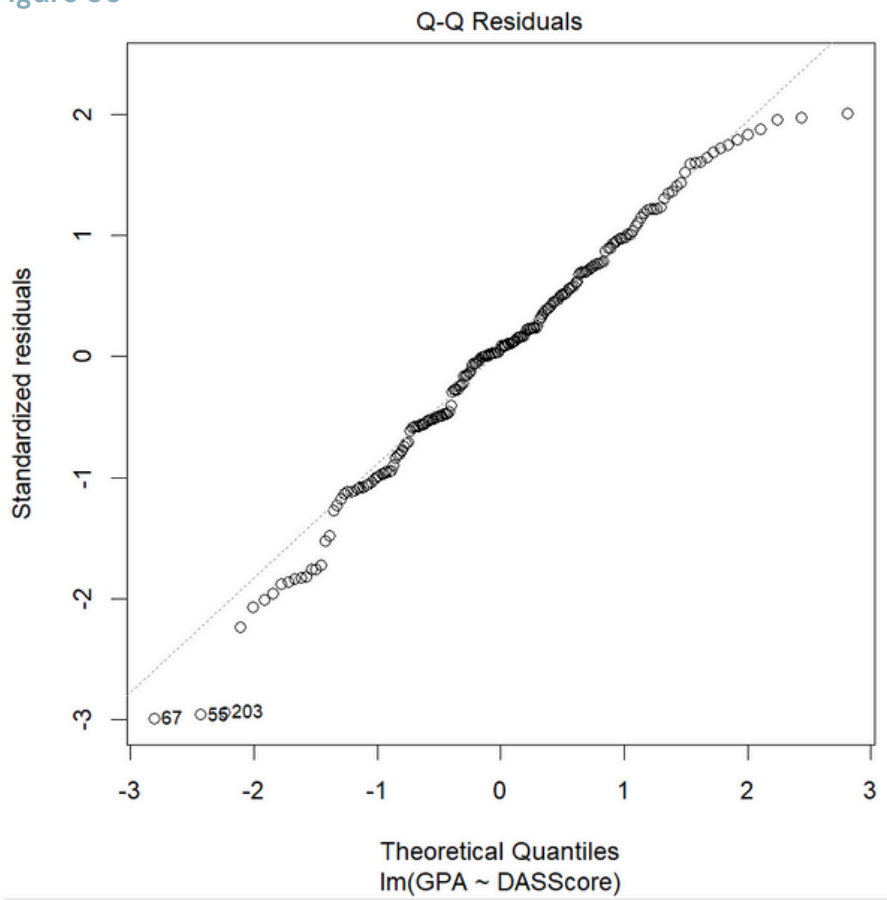


Figure 37

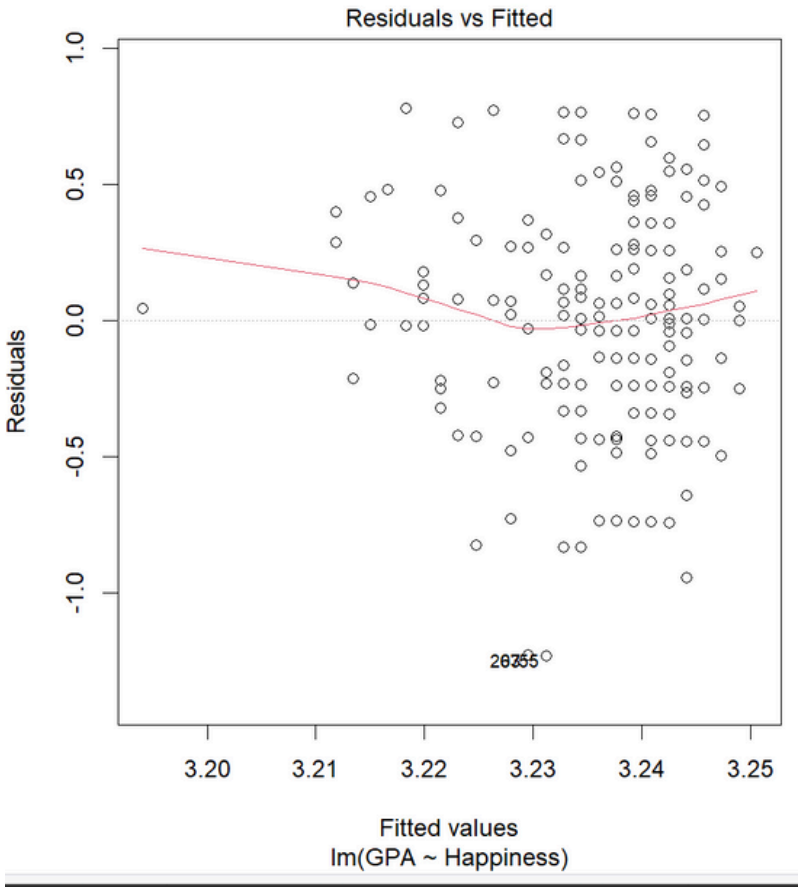


Figure 38

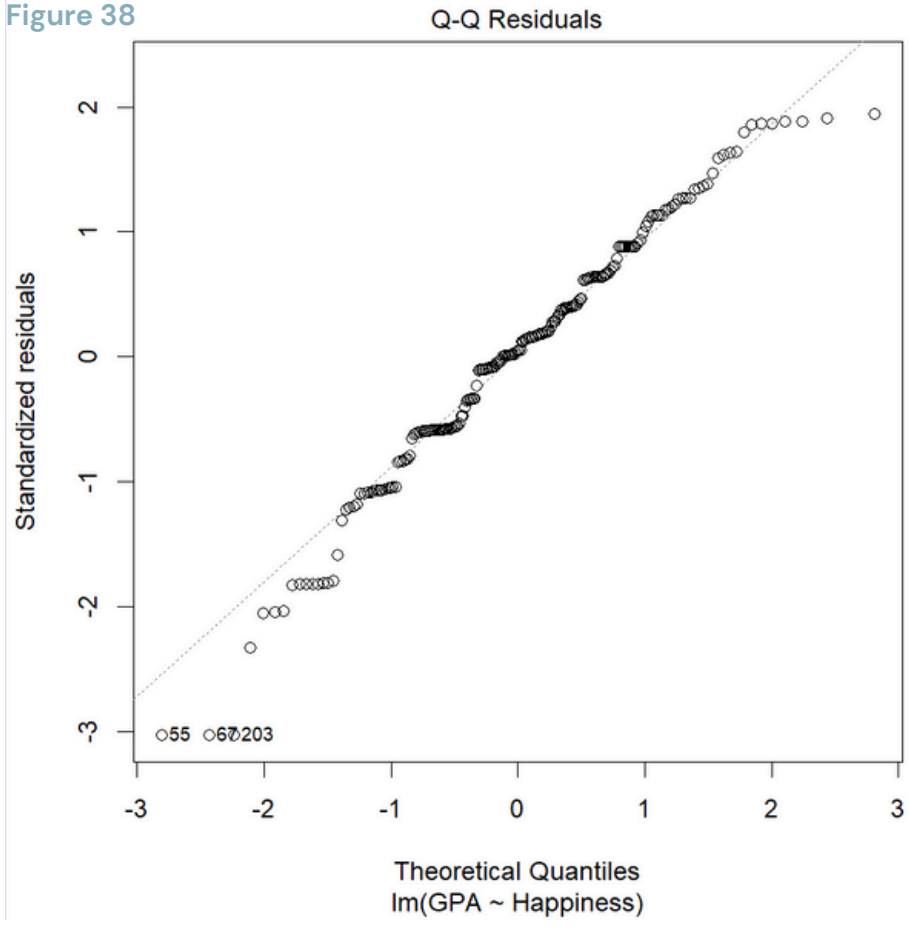


Figure 39

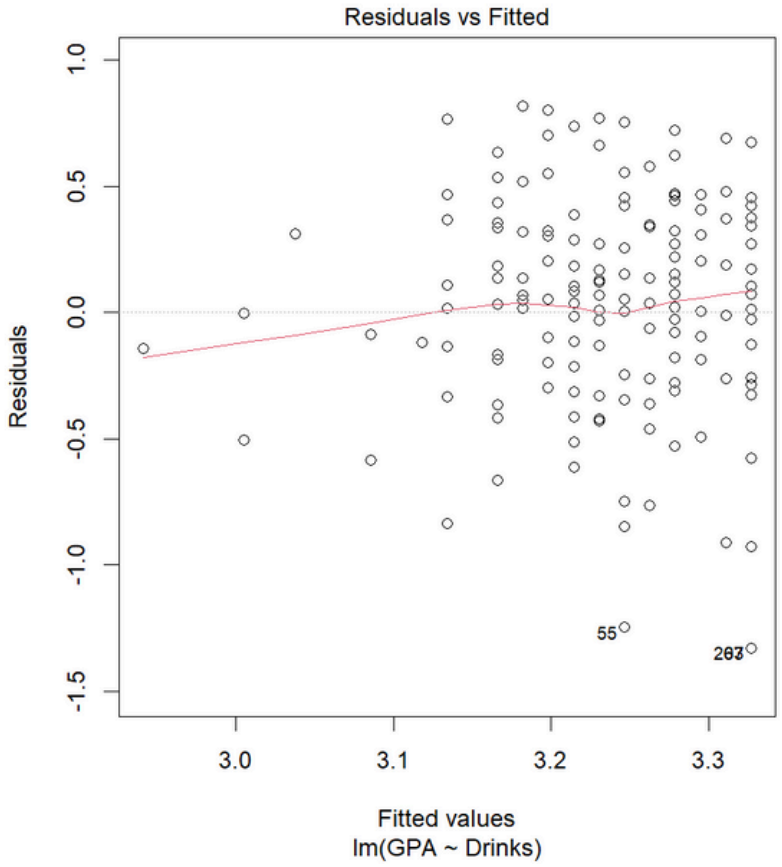


Figure 40

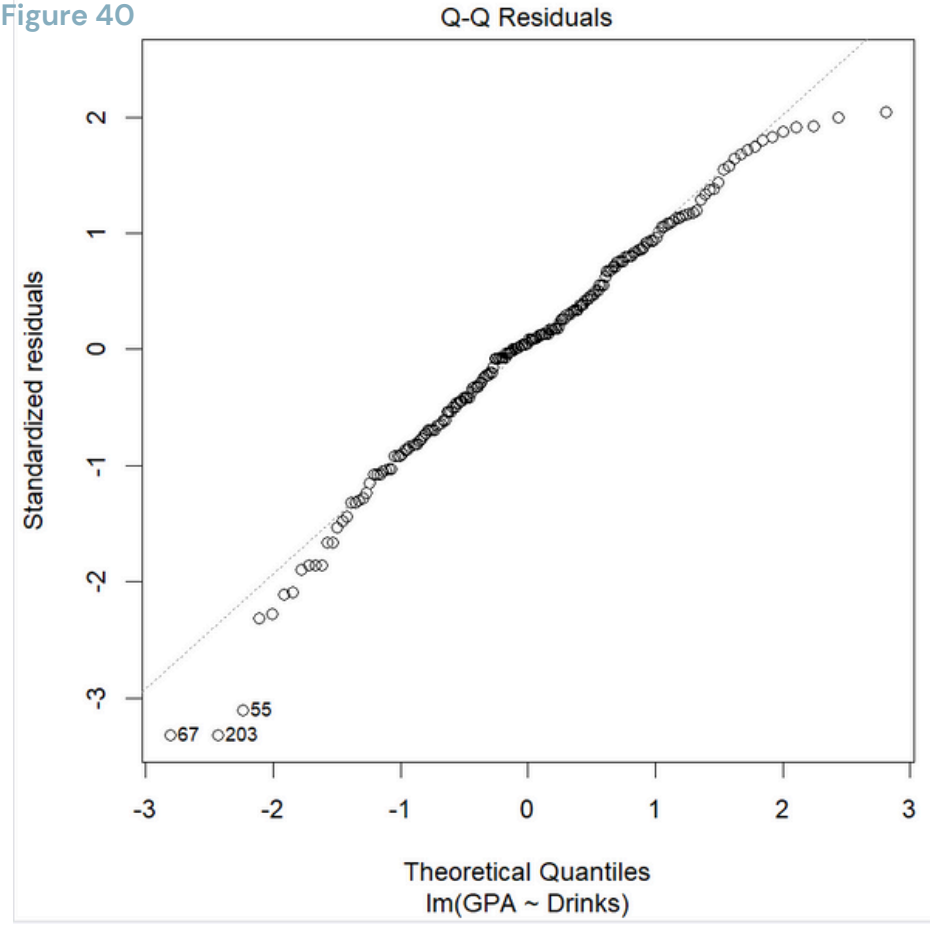


Figure 41

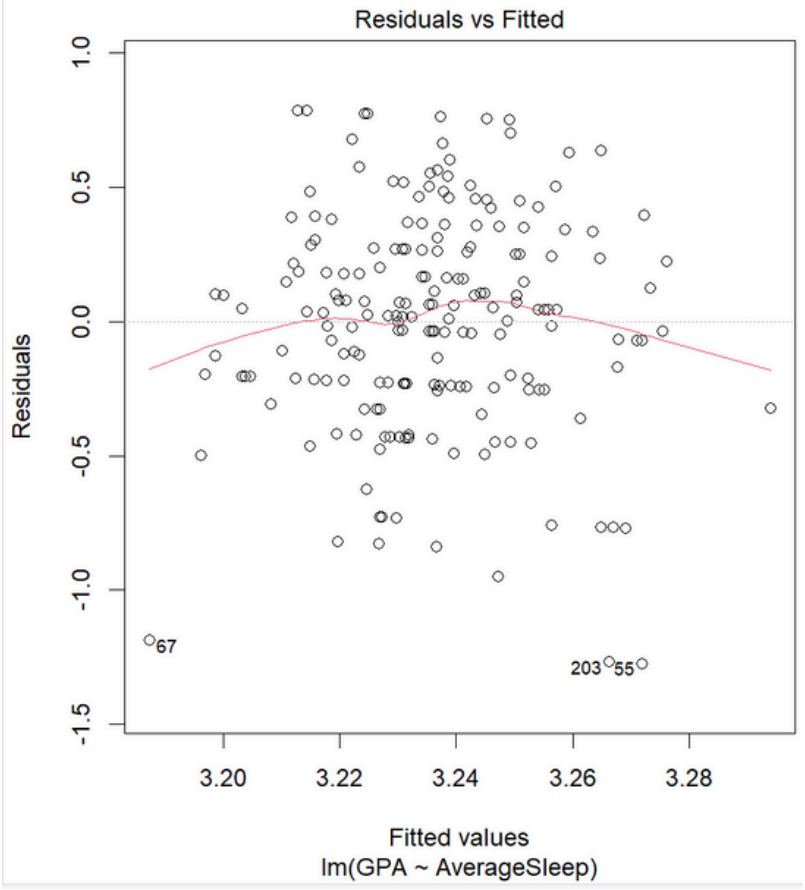


Figure 42

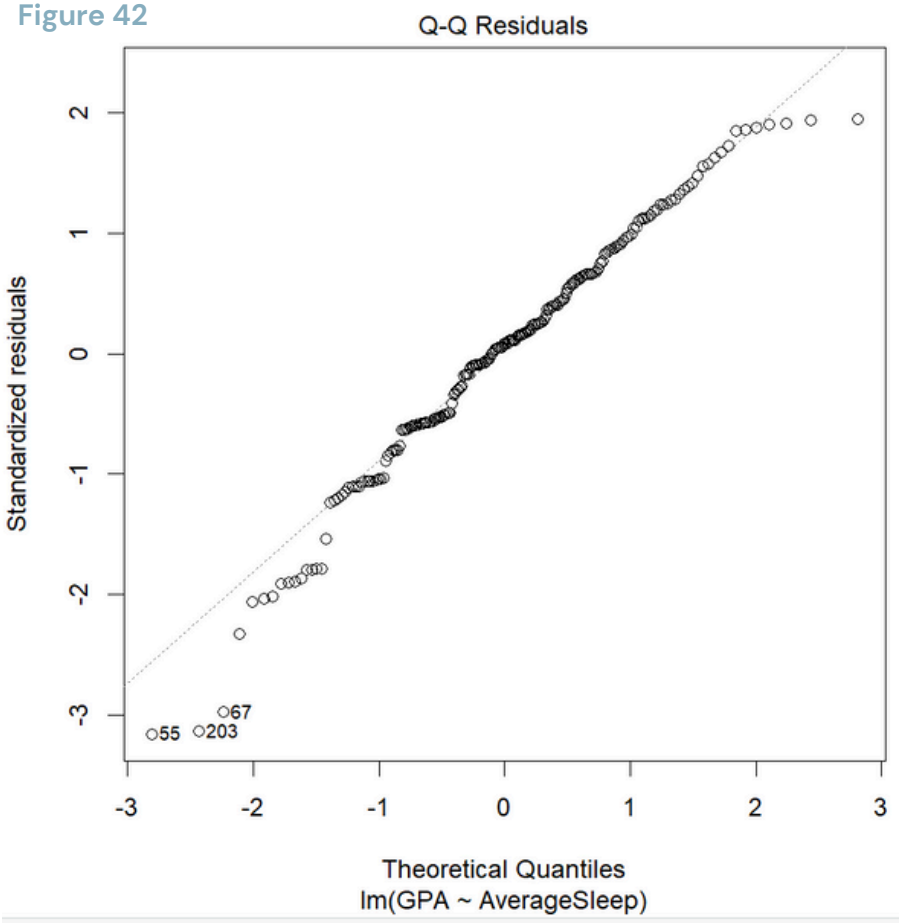


Figure 43

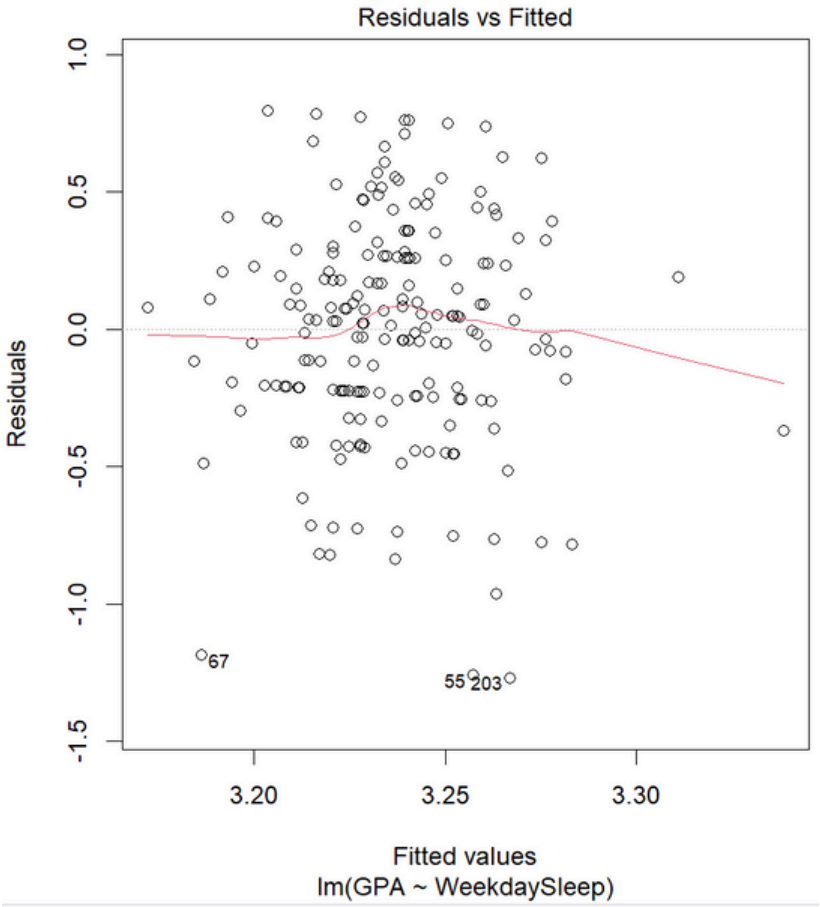
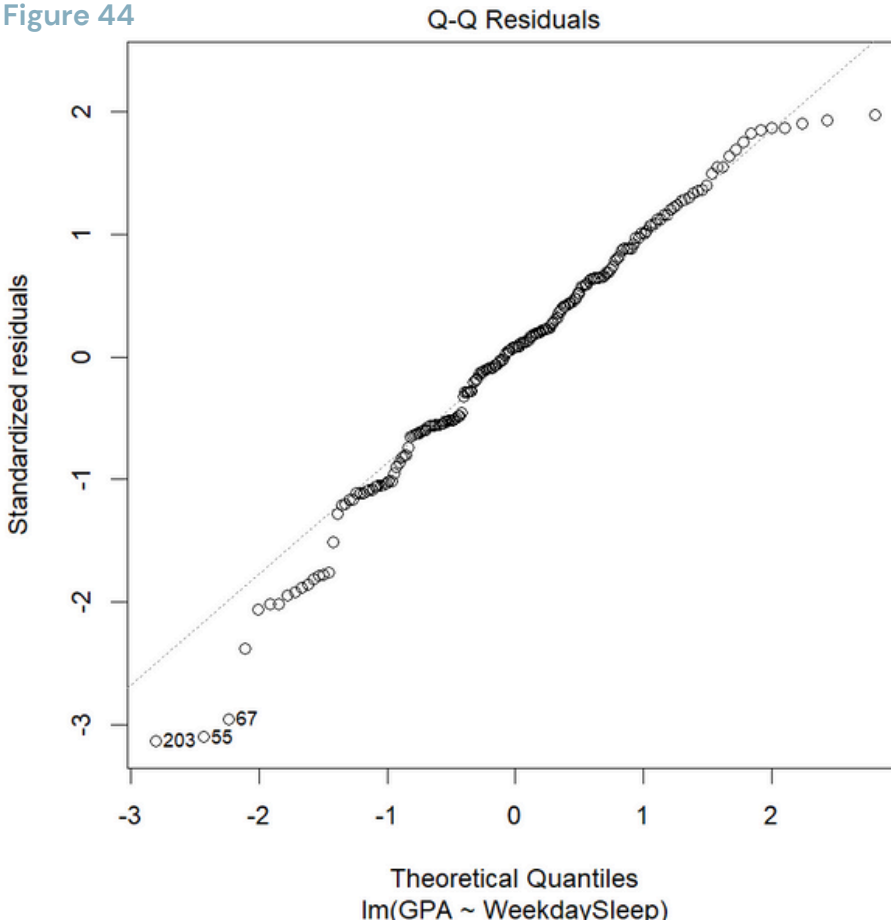
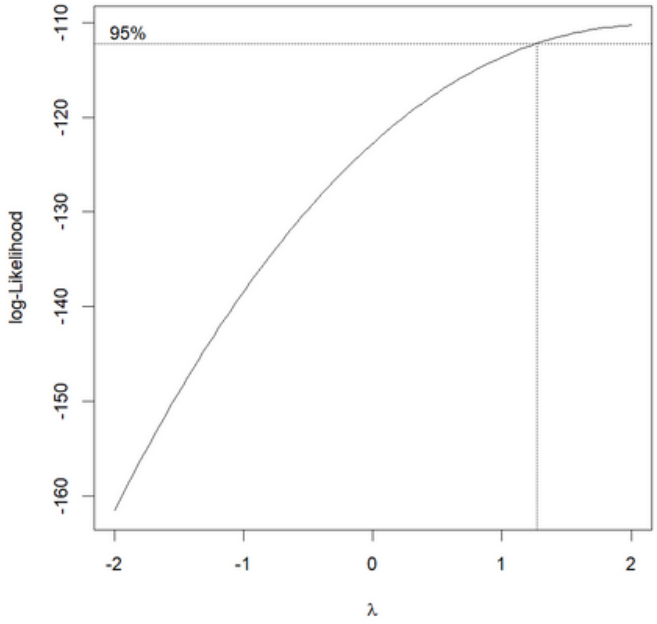


Figure 44



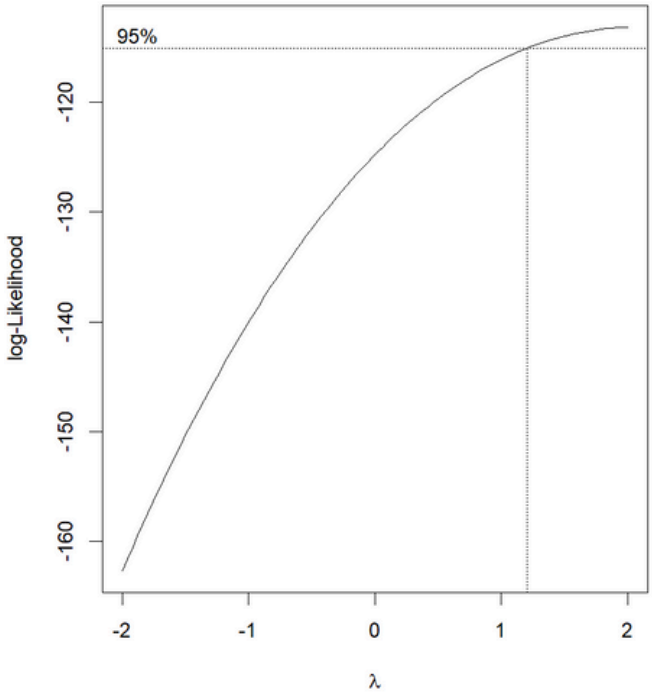
Box Cox Transformation Plots and Results

Figure 45



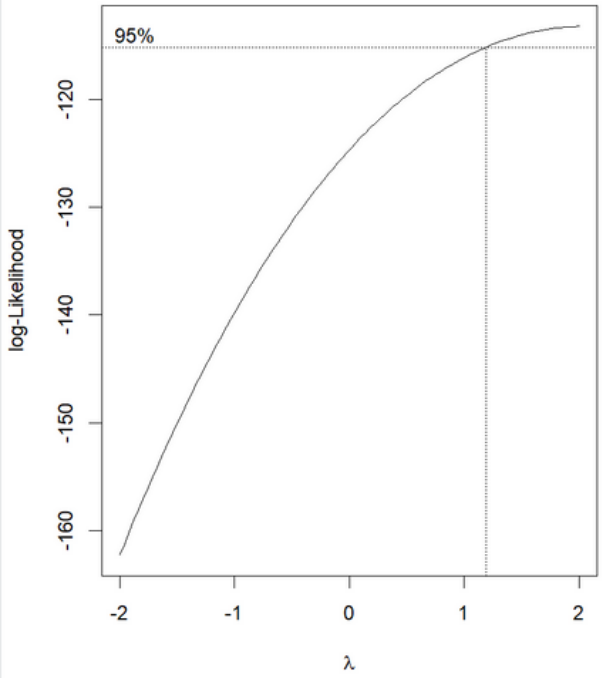

```
Residuals:
     Min       1Q   Median       3Q
-2.01437 -0.42173  0.02967  0.42867
     Max
 1.33500

Coefficients:
              Estimate Std. Error t value
(Intercept)    4.42825    0.05716  77.467
ClassesMissed -0.03547    0.01473  -2.409
              Pr(>|t|)
(Intercept)     <2e-16 ***
ClassesMissed   0.0169 *
---
Signif. codes:
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
  0.1 ‘ ’ 1

Residual standard error: 0.6689 on 198 degrees of freedom
Multiple R-squared:  0.02847,   Adjusted R-squared:  0.02356
F-statistic: 5.801 on 1 and 198 DF,  p-value: 0.01693
```

Figure 46



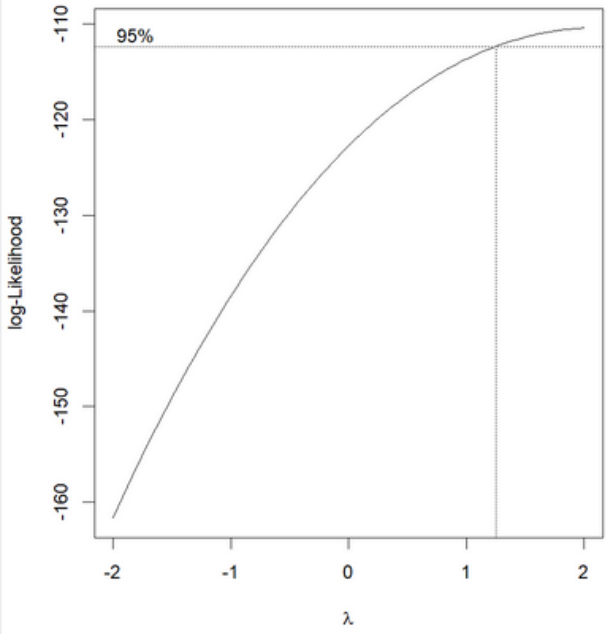

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.82834 -0.36765  0.01957  0.39173  1.22543

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.16752    0.10257  40.631   <2e-16 ***
PoorSleepQuality -0.01045    0.01447  -0.722    0.471
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6145 on 198 degrees of freedom
Multiple R-squared:  0.002625,   Adjusted R-squared:  -0.002412
F-statistic: 0.5211 on 1 and 198 DF,  p-value: 0.4712
```

Figure 47



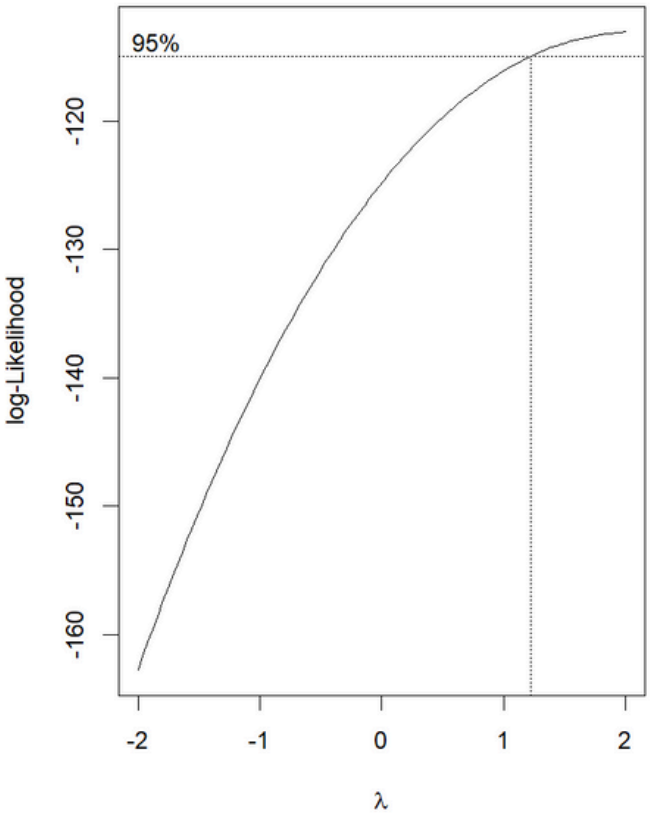

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.78231 -0.36066  0.03298  0.42291  1.21041

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.067617   0.061968  65.641   <2e-16 ***
AnxietyScore 0.006047   0.008137   0.743    0.458
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6144 on 198 degrees of freedom
Multiple R-squared:  0.002781,  Adjusted R-squared:  -0.002255
F-statistic: 0.5522 on 1 and 198 DF,  p-value: 0.4583
```

Figure 48




```
Call:
lm(formula = GPA^1.25 ~ Drinks, data = mysubset)

Residuals:
     Min       1Q   Median       3Q      Max
-2.12644 -0.42587  0.03561  0.47257  1.39653

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.50485    0.08012   56.22   <2e-16
Drinks      -0.02717    0.01141   -2.38   0.0182

(Intercept) ***
Drinks      *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6691 on 198 degrees of freedom
Multiple R-squared:  0.02782,   Adjusted R-squared:  0.02291
F-statistic: 5.666 on 1 and 198 DF,  p-value: 0.01825
```

Figure 49




```
Call:
lm(formula = GPA^1.2 ~ WeekdaySleep, data = mysubset)

Residuals:
     Min       1Q   Median       3Q      Max
-1.85080 -0.35337  0.04189  0.39852  1.22751

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.35659    0.31660  13.761   <2e-16 ***
WeekdaySleep -0.03256    0.03987  -0.817    0.415
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6142 on 198 degrees of freedom
Multiple R-squared:  0.003358,  Adjusted R-squared:  -0.001676
F-statistic: 0.6671 on 1 and 198 DF,  p-value: 0.4151
```

Multiple Regression Models + Forward Selection + Backward Elimination

Figure 50

```
Call:
lm(formula = GPA ~ ClassesMissed + CognitionZscore + PoorSleepQuality +
    DepressionScore + AnxietyScore + StressScore + DASScore +
    Happiness + Drinks + AverageSleep + WeekdaySleep, data = mysubset)

Residuals:
     Min       1Q   Median       3Q      Max
-1.23861 -0.20323  0.03394  0.23253  0.90324

Coefficients: (1 not defined because of singularities)
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.460738   0.310833  11.134  < 2e-16 ***
ClassesMissed    -0.011861   0.008892  -1.334 0.183842
CognitionZscore   0.160577   0.038472   4.174 4.56e-05 ***
PoorSleepQuality -0.002113   0.010816  -0.195 0.845311
DepressionScore  -0.012512   0.006829  -1.832 0.068490 .
AnxietyScore     -0.010931   0.007813  -1.399 0.163408
StressScore       0.021179   0.005431   3.900 0.000134 ***
DASScore                NA         NA      NA       NA
Happiness        -0.001095   0.005744  -0.191 0.848997
Drinks           -0.010682   0.006633  -1.610 0.108983
AverageSleep     -0.005770   0.072401  -0.080 0.936563
WeekdaySleep     -0.016461   0.059837  -0.275 0.783543
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3758 on 189 degrees of freedom
Multiple R-squared:  0.1909,    Adjusted R-squared:  0.1481
F-statistic: 4.458 on 10 and 189 DF,  p-value: 1.196e-05
```

Figure 51

```
    DepressionScore + AnxietyScore + StressScore + Happiness +
    Drinks + AverageSleep + WeekdaySleep, data = mysubset, method = "forward")
10 Variables  (and intercept)
                Forced in Forced out
ClassesMissed       FALSE      FALSE
CognitionZscore     FALSE      FALSE
PoorSleepQuality    FALSE      FALSE
DepressionScore     FALSE      FALSE
AnxietyScore        FALSE      FALSE
StressScore         FALSE      FALSE
Happiness           FALSE      FALSE
Drinks              FALSE      FALSE
AverageSleep        FALSE      FALSE
WeekdaySleep        FALSE      FALSE
1 subsets of each size up to 8
Selection Algorithm: forward
         ClassesMissed CognitionZscore PoorSleepQuality DepressionScore AnxietyScore StressScore
1  ( 1 ) " "           "*"             " "              " "             " "          " "
2  ( 1 ) " "           "*"             " "              " "             " "          "*"
3  ( 1 ) " "           "*"             " "              "*"             " "          "*"
4  ( 1 ) "*"           "*"             " "              "*"             " "          "*"
5  ( 1 ) "*"           "*"             " "              "*"             " "          "*"
6  ( 1 ) "*"           "*"             " "              "*"             "*"          "*"
7  ( 1 ) "*"           "*"             " "              "*"             "*"          "*"
8  ( 1 ) "*"           "*"             " "              "*"             "*"          "*"
         Happiness Drinks AverageSleep WeekdaySleep
1  ( 1 ) " "       " "    " "          " "
2  ( 1 ) " "       " "    " "          " "
3  ( 1 ) " "       " "    " "          " "
4  ( 1 ) " "       " "    " "          " "
5  ( 1 ) " "       "*"    " "          " "
6  ( 1 ) " "       "*"    " "          " "
7  ( 1 ) " "       "*"    " "          "*"
8  ( 1 ) "*"       "*"    " "          "*"
>
```

Figure 52

```
Subset selection object
Call: regsubsets.formula(GPA ~ ClassesMissed + CognitionZscore + PoorSleepQuality +
    DepressionScore + AnxietyScore + StressScore + Happiness +
    Drinks + AverageSleep + WeekdaySleep, data = mysubset, method = "backward")
10 Variables  (and intercept)
                Forced in Forced out
ClassesMissed       FALSE      FALSE
CognitionZscore     FALSE      FALSE
PoorSleepQuality    FALSE      FALSE
DepressionScore     FALSE      FALSE
AnxietyScore        FALSE      FALSE
StressScore         FALSE      FALSE
Happiness           FALSE      FALSE
Drinks              FALSE      FALSE
AverageSleep        FALSE      FALSE
WeekdaySleep        FALSE      FALSE
1 subsets of each size up to 8
Selection Algorithm: backward
         ClassesMissed CognitionZscore PoorSleepQuality DepressionScore AnxietyScore StressScore Happiness Drinks AverageSleep WeekdaySleep
1  ( 1 ) " "           "*"             " "              " "             " "           " "         " "       " "    " "          " "
2  ( 1 ) " "           "*"             " "              " "             " "           "*"         " "       " "    " "          " "
3  ( 1 ) " "           "*"             " "              "*"             " "           "*"         " "       " "    " "          " "
4  ( 1 ) " "           "*"             " "              "*"             " "           "*"         " "       "*"    " "          " "
5  ( 1 ) " "           "*"             " "              "*"             "*"           "*"         " "       "*"    " "          " "
6  ( 1 ) "*"           "*"             " "              "*"             "*"           "*"         " "       "*"    " "          " "
7  ( 1 ) "*"           "*"             " "              "*"             "*"           "*"         " "       "*"    " "          "*"
8  ( 1 ) "*"           "*"             " "              "*"             "*"           "*"         "*"       "*"    " "          "*"
```

Model Assumptions & Diagnostics

Figure 53

	CognitionZScore	StressScore	DepressionScore	classes missed	anxiety score	adjr2
1	1	0	0	0	0	0.080170048
1	0	1	0	0	0	0.030702508
1	0	0	0	1	0	0.022559636
1	0	0	0	0	1	-0.002098365
1	0	0	1	0	0	-0.005048900
2	1	1	0	0	0	0.119208260
2	1	0	0	1	0	0.093116639
2	1	0	0	0	1	0.079133956
2	1	0	1	0	0	0.075730101
2	0	1	1	0	0	0.057867298
2	0	1	0	1	0	0.056972545
2	0	1	0	0	1	0.039148713
2	0	0	0	1	1	0.023707958
2	0	0	1	1	0	0.018506923
2	0	0	1	0	1	-0.005605106
3	1	1	1	0	0	0.145256285
3	1	1	0	1	0	0.135017124
3	1	1	0	0	1	0.131027270
3	1	0	0	1	1	0.094781459
3	1	0	1	1	0	0.090076322
3	1	0	1	0	1	0.075003408
3	0	1	1	1	0	0.073549822
3	0	1	1	0	1	0.060608399
3	0	1	0	1	1	0.060498491
3	0	0	1	1	1	0.019011960
4	1	1	1	1	0	0.152791302
4	1	1	1	0	1	0.150837577
4	1	1	0	1	1	0.142214105
4	1	0	1	1	1	0.090174629
4	0	1	1	1	1	0.073875971
5	1	1	1	1	1	0.156106103