A Study of How Different Habits of Students Affect Their Academic Performance,

Jiayi Xu

Abstract

This project examines the impact of student habits on academic performance using a dataset of 1,000 students with 14 predictors. I conducted exploratory data analysis, fitted multiple linear models, and applied remedial strategies such as Box-Cox transformation to address violations of model assumptions. Through stepwise selection based on BIC, I identified a final model (Model 3) with six key predictors: study hours, social media usage, attendance percentage, sleep hours, exercise frequency, and mental health rating. To respect the natural bounds of exam scores, I clipped the model's fitted values to the range [0, 100], which improved interpretability and residual behavior without compromising performance. Model 3 explains 79.2% of the variance in the transformed exam scores and highlights the significant positive effects of study time, attendance, sleep, exercise, and mental health, while social media usage shows a negative association.

Exploratory Data Analysis

```
## [1] "Sample size: 1000"
```

```
## [1] "Number of predictors: 14"
```

The original dataset contains 1000 observations and 14 predictors. Statistics are summarized as following:

```
##       age           gender          study_hours_per_day social_media_hours
##  Min.   :17.00   Length:1000        Min.   :0.000       Min.   :0.000
##  1st Qu.:18.75   Class :character   1st Qu.:2.324       1st Qu.:1.661
##  Median :20.00   Mode  :character   Median :3.040       Median :2.217
##  Mean   :20.50                      Mean   :3.019       Mean   :2.226
##  3rd Qu.:23.00                      3rd Qu.:3.678       3rd Qu.:2.805
##  Max.   :24.00                      Max.   :6.157       Max.   :5.621
##  netflix_hours   part_time_job      attendance_percentage  sleep_hours
##  Min.   :0.000   Length:1000        Min.   : 56.00        Min.   : 4.794
##  1st Qu.:1.487   Class :character   1st Qu.: 78.00        1st Qu.: 7.055
##  Median :2.056   Mode  :character   Median : 84.40        Median : 7.639
##  Mean   :2.081                      Mean   : 84.13        Mean   : 7.646
##  3rd Qu.:2.615                      3rd Qu.: 91.03        3rd Qu.: 8.227
##  Max.   :4.826                      Max.   :100.00        Max.   :10.580
##  diet_quality       exercise_frequency parental_education_level internet_quality
##  Length:1000        Min.   :0.000      Length:1000              Length:1000
##  Class :character   1st Qu.:1.000      Class :character         Class :character
##  Mode  :character   Median :3.000      Mode  :character         Mode  :character
##                     Mean   :3.042
```

```
##                     3rd Qu.:5.000
##                     Max.   :6.000
##  mental_health_rating extracurricular_participation   exam_score
##  Min.   : 1.000       Length:1000                   Min.   : 37.14
##  1st Qu.: 3.000       Class :character              1st Qu.: 80.07
##  Median : 5.000       Mode  :character              Median : 87.02
##  Mean   : 5.438                                     Mean   : 85.31
##  3rd Qu.: 8.000                                     3rd Qu.: 92.32
##  Max.   :10.000                                     Max.   :100.00
```

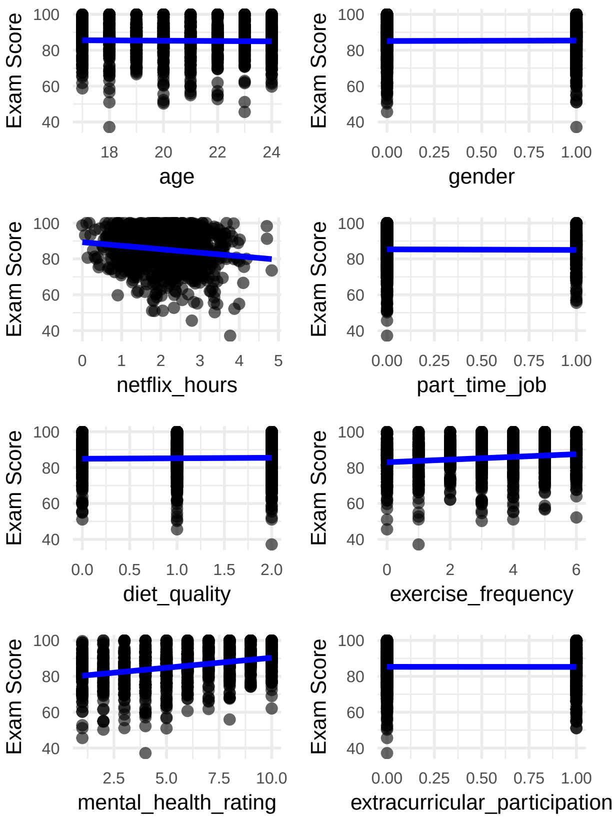
Encoding

- I find that gender, part_time_job, diet_quality, parental_education_level, internet_quality, extracurricular_participation are categorical.
- I excluded 42 observations labeled 'Other' in the gender variable because they represented less than 5% of the data and could not be reliably modeled. This was done to improve model accuracy and interpretability.
- For the variable parental_education_level, I group it into "University" vs. "Non-university", and then encode the groups as binary: University = 1, Non-university = 0.

Variable	Type	Encoding
Gender	Categorical	Female = 0, Male = 1
Part-time Job	Binary	No = 0, Yes = 1
Diet Quality	Ordinal	Poor = 0, Fair = 1, Good = 2
Parental Education Level	Binary	Non-university = 0, University = 1
Internet Quality	Ordinal	Poor = 0, Average = 1, Good = 2
Extracurricular Participation	Binary	No = 0, Yes = 1

Scatter plot

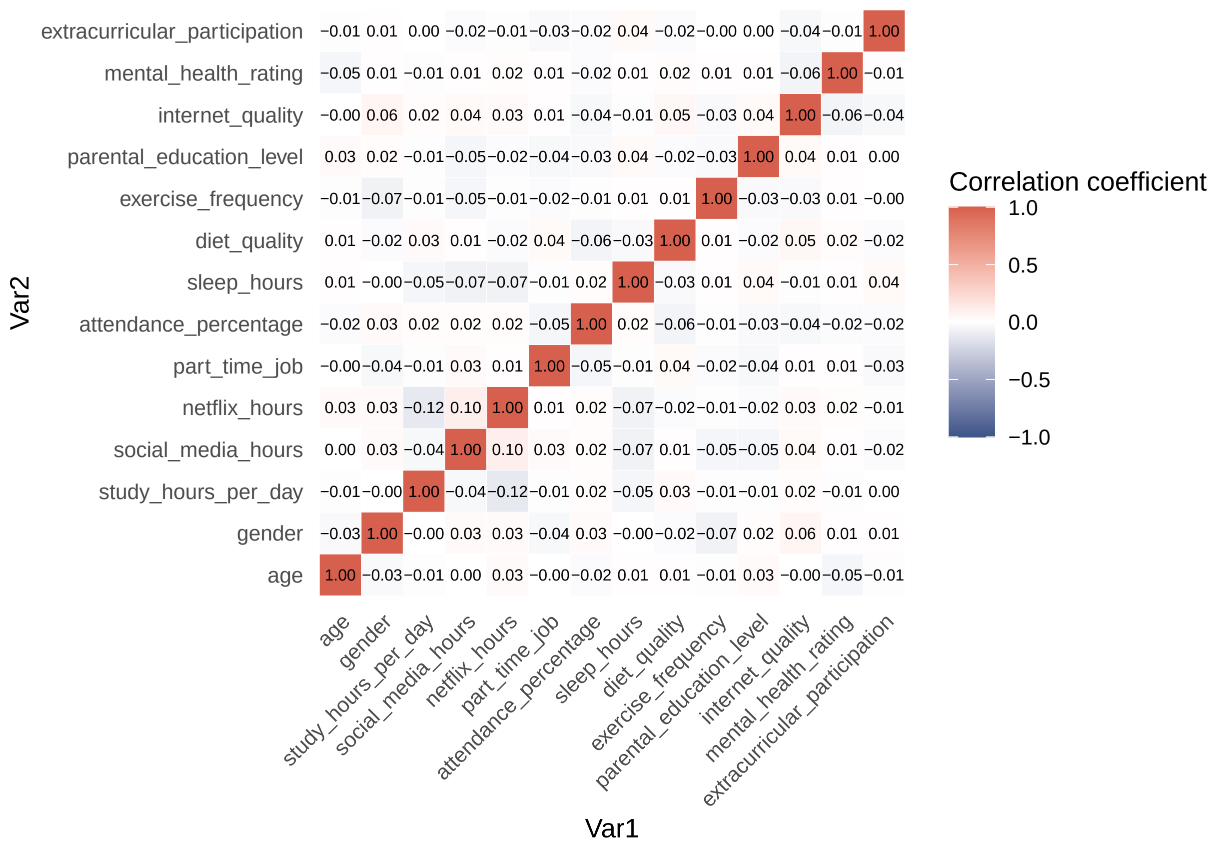
After pre-processing, I plot exam_score versus all independent variables to detect potential linear relationship.



Observations:

The scatterplot matrix reveals key insights into the relationships between individual predictors and exam scores. Among all variables, study_hours_per_day shows a strong positive linear relationship with exam scores, while mental_health_rating, sleep_hours, and attendance_percentage exhibit mild positive trends. In contrast, social_media_hours and netflix_hours show slight negative associations with exam performance. Variables such as age, gender, part_time_job, diet_quality, exercise_frequency, parental_education_level, internet_quality, and extracurricular_participation display little to no clear relationship with exam scores, as indicated by flat regression lines or low variability. These findings suggest that study habits and mental well-being are more predictive of academic performance, while many demographic or lifestyle variables may have limited explanatory power in this context.

Multicollinearity



Observations:

According to Correlation Matrix, study hours, attendance_percentage, sleep hours, exercise_frequency and mental-health rating correlate positively with exam score, while social-media and Netflix hours show negative associations.

No pair exceeds 0.8, so multicollinearity is unlikely to be a major issue.

Model Fitting

Linear model with all variables (Model 1)

I first fit a linear model with all predictors:

```
##
## Call:
## lm(formula = exam_score ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.4222  -2.7947   0.1853   3.3357  12.6204
##
```

```
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                   39.51258    2.68444  14.719  < 2e-16 ***
## age                            0.03757    0.06966   0.539  0.58981    
## gender                         0.54447    0.32217   1.690  0.09136 .  
## study_hours_per_day            7.36518    0.15506  47.499  < 2e-16 ***
## social_media_hours            -1.38401    0.19174  -7.218 1.08e-12 ***
## netflix_hours                 -0.60162    0.20505  -2.934  0.00343 ** 
## part_time_job                  0.12438    0.39211   0.317  0.75116    
## attendance_percentage          0.08050    0.01714   4.697 3.04e-06 ***
## sleep_hours                    1.57095    0.17863   8.794  < 2e-16 ***
## diet_quality                  -0.09551    0.22237  -0.429  0.66766    
## exercise_frequency             0.75605    0.07962   9.495  < 2e-16 ***
## parental_education_level      -0.06093    0.32216  -0.189  0.85004    
## internet_quality              -0.24938    0.22220  -1.122  0.26201    
## mental_health_rating           1.15254    0.05632  20.465  < 2e-16 ***
## extracurricular_participation -0.19640    0.34483  -0.570  0.56912    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.954 on 943 degrees of freedom
## Multiple R-squared:  0.7604, Adjusted R-squared:  0.7569 
## F-statistic: 213.8 on 14 and 943 DF,  p-value: < 2.2e-16
```

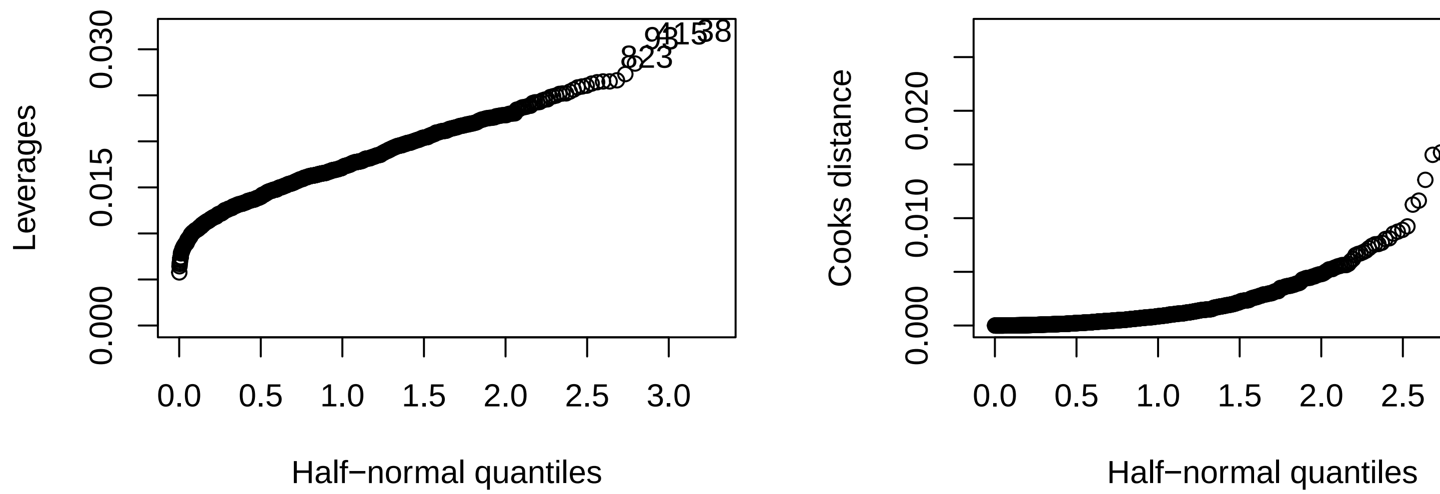
findings: It shows strong evidence that some predictors affect exam score significantly.

Diagnostics for the basic linear model

I first check unusual observations.

```
##         38        415 
## 0.03201713 0.03171924
```

```
## [1] 0.02745746
```

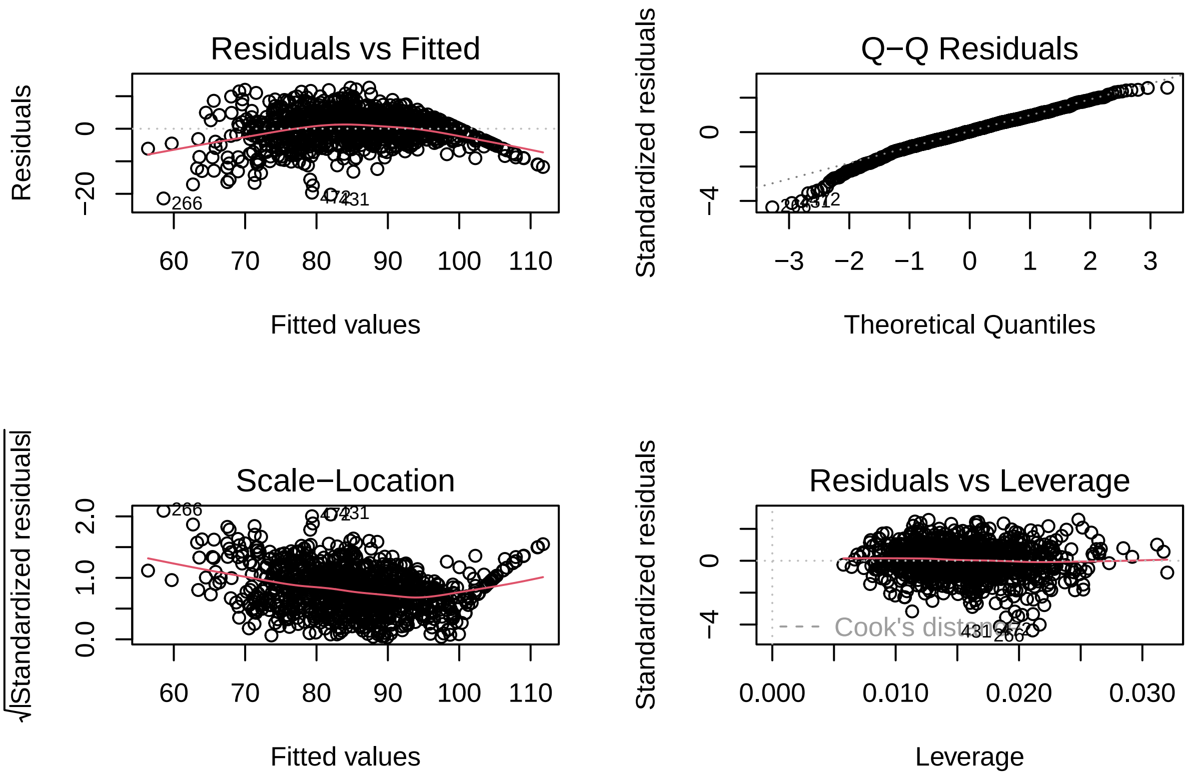


Findings: The diagnostic results for the basic linear model indicate that observations 38 and 415 have high leverage values exceeding the threshold of 0.0313, suggesting they possess unusual combinations of predictor values. However, the maximum Cook's distance is only 0.0275, which is well below common concern thresholds (e.g., 0.5 or 1), indicating that no single observation has a substantial influence on the overall regression results. Overall, the model appears to be reasonably robust, with no highly influential outliers.

```
## [1] 4.06431
```

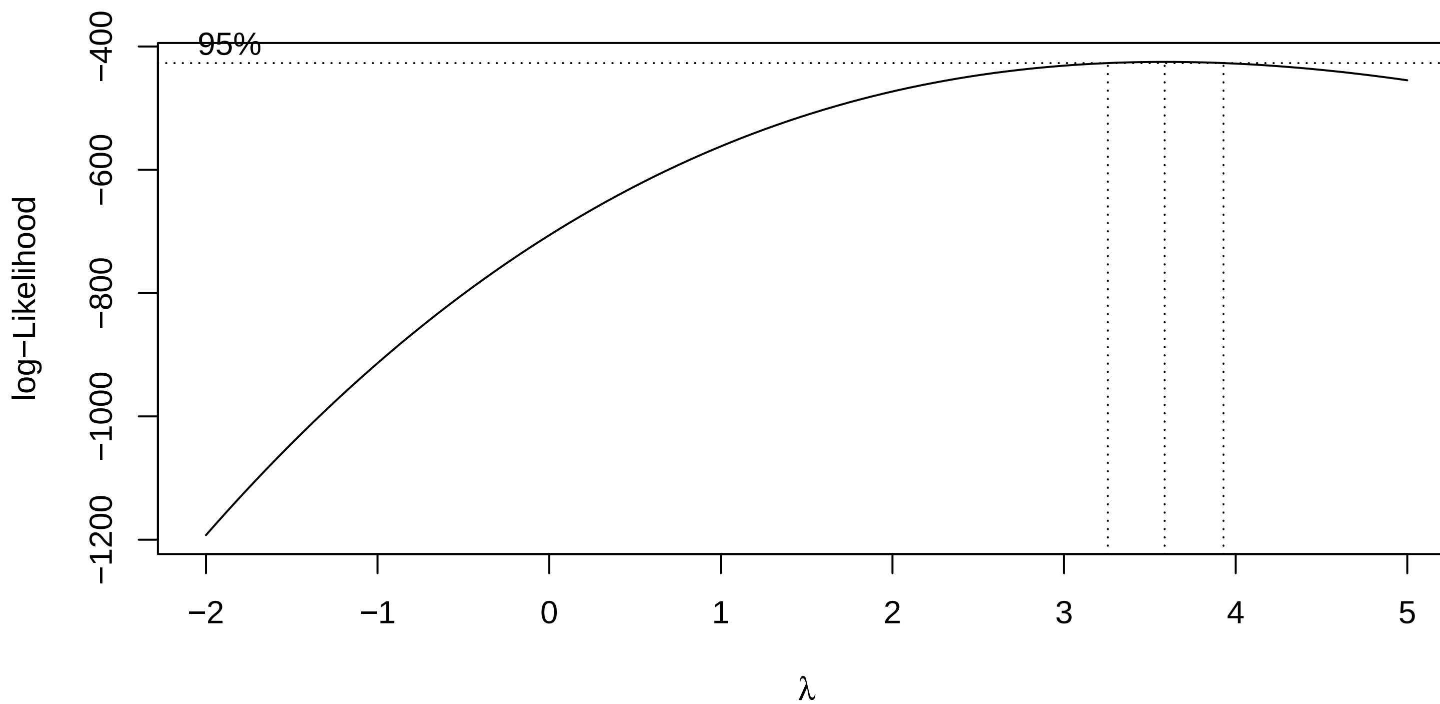
```
## 266 431
## 256 418
```

Findings: Using the Bonferroni correction to control for multiple comparisons, the cutoff for identifying outliers based on studentized residuals is approximately 4.06. Four observations (266, 431, 256, and 418) have absolute studentized residuals exceeding this threshold, indicating they are statistically significant outliers in the context of the model.



Findings: The diagnostic plots reveal several concerns with the linear model. The Residuals vs Fitted plot shows a curved pattern, indicating a violation of the linearity assumption and potential heteroscedasticity. The Normal Q-Q plot reveals deviations from the diagonal at both ends, suggesting that the residuals are not perfectly normally distributed. The Scale-Location plot further confirms unequal variance, as the spread of residuals changes across fitted values. Lastly, the Residuals vs Leverage plot identifies a few observations (e.g., 256 and 418) with moderate leverage, though none appear to be highly influential. Overall, these diagnostics suggest that the model may benefit from transformations or more flexible modeling techniques.

Remedy: Motivated by the diagnostics, I consider Box Cox transformation.




```
## [1] 3.585859
```

Findings: $\lambda_{opt} = 3.59$ gives the highest log likelihood.

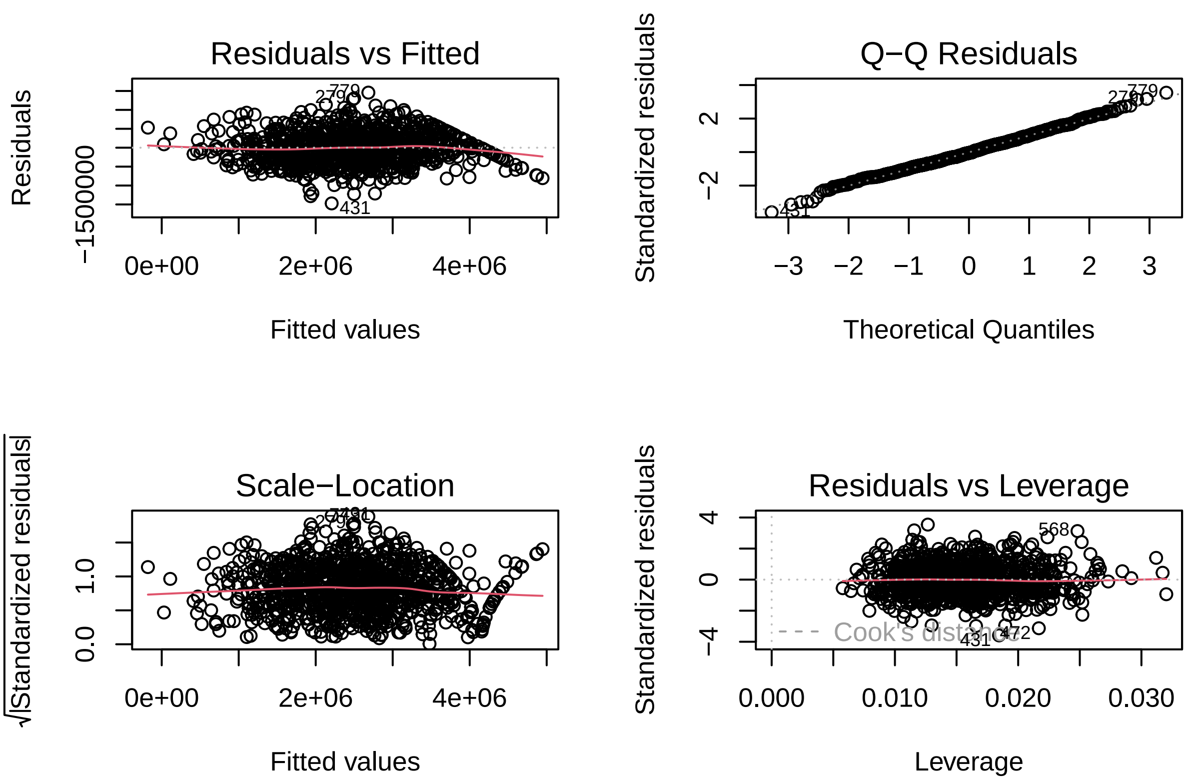
Refit the linear model using the transformed response (Model 2)

```
##
## Call:
## lm(formula = exam_score_trans ~ . - exam_score, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1472305  -281463   -21680   270870  1456871
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -1973612     224276  -8.800  < 2e-16 ***
## age                              4268       5820   0.733   0.4635
## gender                          44926      26916   1.669   0.0954 .
## study_hours_per_day            682169      12955  52.658  < 2e-16 ***
## social_media_hours            -119472      16019  -7.458 1.99e-13 ***
## netflix_hours                  -35913      17131  -2.096   0.0363 *
## part_time_job                    -359      32759  -0.011   0.9913
## attendance_percentage            8513       1432   5.944 3.90e-09 ***
## sleep_hours                    151710      14924  10.166  < 2e-16 ***
```

```
## diet_quality                         -4982      18578  -0.268   0.7886
## exercise_frequency                   71389       6652  10.731  < 2e-16 ***
## parental_education_level             -2277      26915  -0.085   0.9326
## internet_quality                    -28813      18564  -1.552   0.1210
## mental_health_rating                107124       4705  22.767  < 2e-16 ***
## extracurricular_participation        -6563      28810  -0.228   0.8199
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 413900 on 943 degrees of freedom
## Multiple R-squared:  0.7955, Adjusted R-squared:  0.7925
## F-statistic: 262.1 on 14 and 943 DF,  p-value: < 2.2e-16
```

Diagnostics for Model 2

Diagnostic plots for Model 2 are as follows:



Findings: The Box-Cox transformation with $\lambda = 3.59$ has significantly improved the model diagnostics. Linearity, normality, and constant variance assumptions are now reasonably satisfied. The model is better suited for inference and prediction in its transformed form.

```
## Loading required package: zoo

##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod2
## BP = 16.366, df = 14, p-value = 0.2916
```

Findings: There is no significant evidence of heteroskedasticity in the residuals of the transformed model (mod2).

Variable selection (Model Comparison with Model 2 Included)

I start from Model 2, and use stepwise procedure to select a subset of predictors, using BIC as the criterion.

```
##
## Call:
## lm(formula = exam_score_trans ~ study_hours_per_day + social_media_hours +
##     attendance_percentage + sleep_hours + exercise_frequency +
##     mental_health_rating, data = data)
##
## Coefficients:
##           (Intercept)    study_hours_per_day     social_media_hours
##              -2010165                 684708                -122488
## attendance_percentage            sleep_hours     exercise_frequency
##                  8603                 154028                  71049
##  mental_health_rating
##                107238
```

Findings: Using stepwise selection based on the Bayesian Information Criterion (BIC), six predictors were selected as the most informative for explaining variation in the Box-Cox transformed exam scores. These variables include study_hours_per_day, social_media_hours, attendance_percentage, sleep_hours, exercise_frequency, and mental_health_rating.

```
##
## Call:
## lm(formula = exam_score_trans ~ study_hours_per_day + social_media_hours +
##     attendance_percentage + sleep_hours + exercise_frequency +
##     mental_health_rating, data = new_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1509862  -278967   -18364   271199  1445478
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -2010165     178022 -11.292  < 2e-16 ***
## study_hours_per_day     684708      12856  53.259  < 2e-16 ***
## social_media_hours     -122488      15929  -7.689 3.68e-14 ***
## attendance_percentage     8603       1426   6.031 2.33e-09 ***
```

```
## sleep_hours             154028      14871  10.358  < 2e-16 ***
## exercise_frequency       71049       6636  10.706  < 2e-16 ***
## mental_health_rating    107238       4692  22.855  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 414300 on 951 degrees of freedom
## Multiple R-squared:  0.7934, Adjusted R-squared:  0.7921
## F-statistic: 608.7 on 6 and 951 DF,  p-value: < 2.2e-16
```

ANCOVA Model with Interactions (Model Comparison with Model 3 Included)

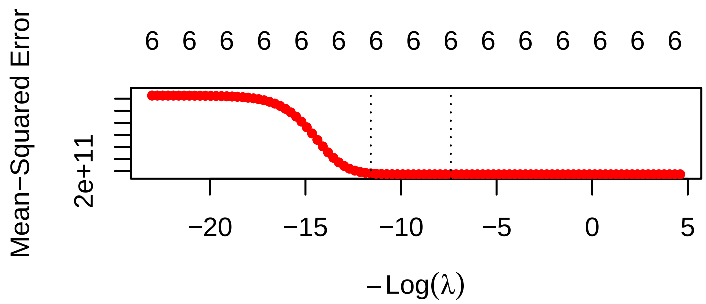
I also want to consider a model with interactions between categorical and numerical variables. An F-test that compares the additive model (Model 3) and the interaction model (Model 4) is as follows:

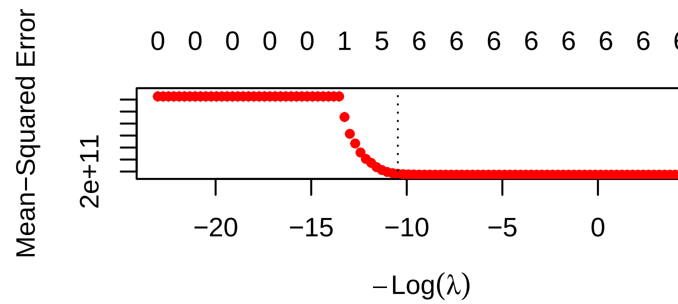
```
## Analysis of Variance Table
##
## Model 1: exam_score_trans ~ study_hours_per_day + social_media_hours +
##     attendance_percentage + sleep_hours + exercise_frequency +
##     mental_health_rating
## Model 2: exam_score_trans ~ (study_hours_per_day + social_media_hours +
##     attendance_percentage + sleep_hours) * (exercise_frequency +
##     mental_health_rating)
##   Res.Df        RSS Df  Sum of Sq     F Pr(>F)
## 1    951 1.6321e+14
## 2    943 1.6119e+14  8 2.0143e+12 1.473 0.1628
```

Findings: Since the p-value is greater than 0.05, I fail to reject the null hypothesis. This means that the inclusion of interaction terms does not significantly improve the model.

Shrinkage Methods (Ridge and Lasso Regression)

I also try some shrinkage methods including the ridge and Lasso regression. I consider the same set of variables as Model 4, and transformed exam score as the response. The following figures show the cross-validation errors of the two methods with a range of λ values.



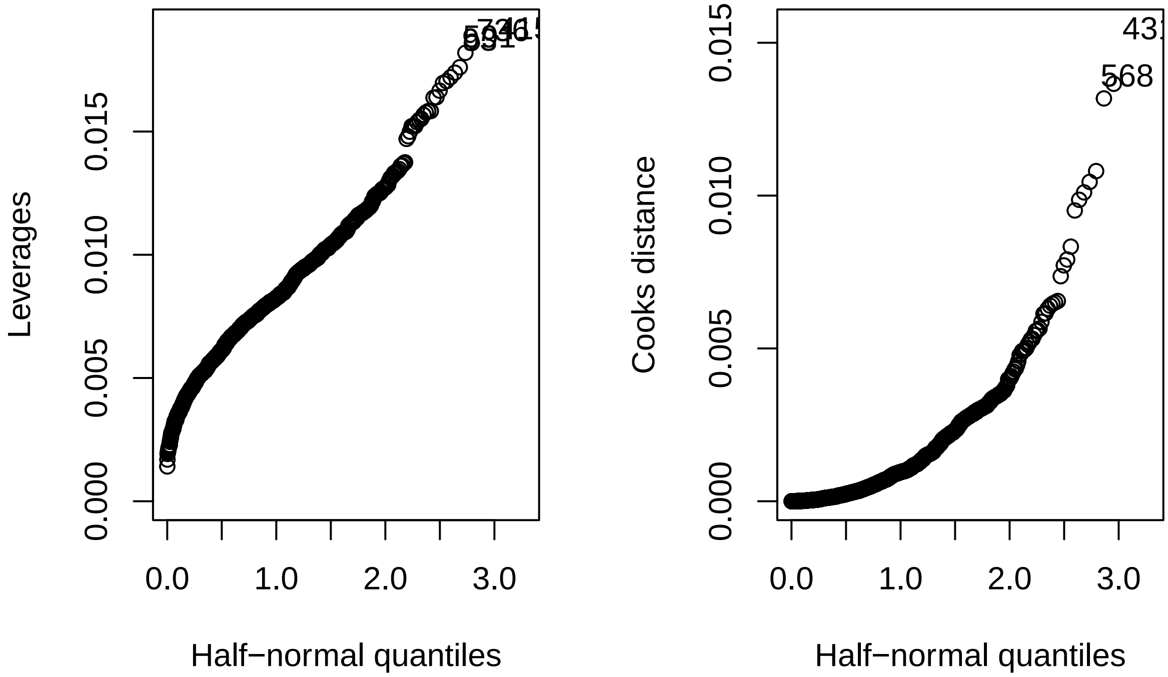


Findings: The cross-validation plots for Ridge and Lasso regression reveal that the optimal prediction accuracy is achieved at moderate-to-large values of λ. Ridge regression retains all predictors with reduced magnitude, while Lasso achieves similar error rates but also performs variable selection, shrinking some coefficients to exactly zero as λ increases.

Diagnostic for final model (Model 3)

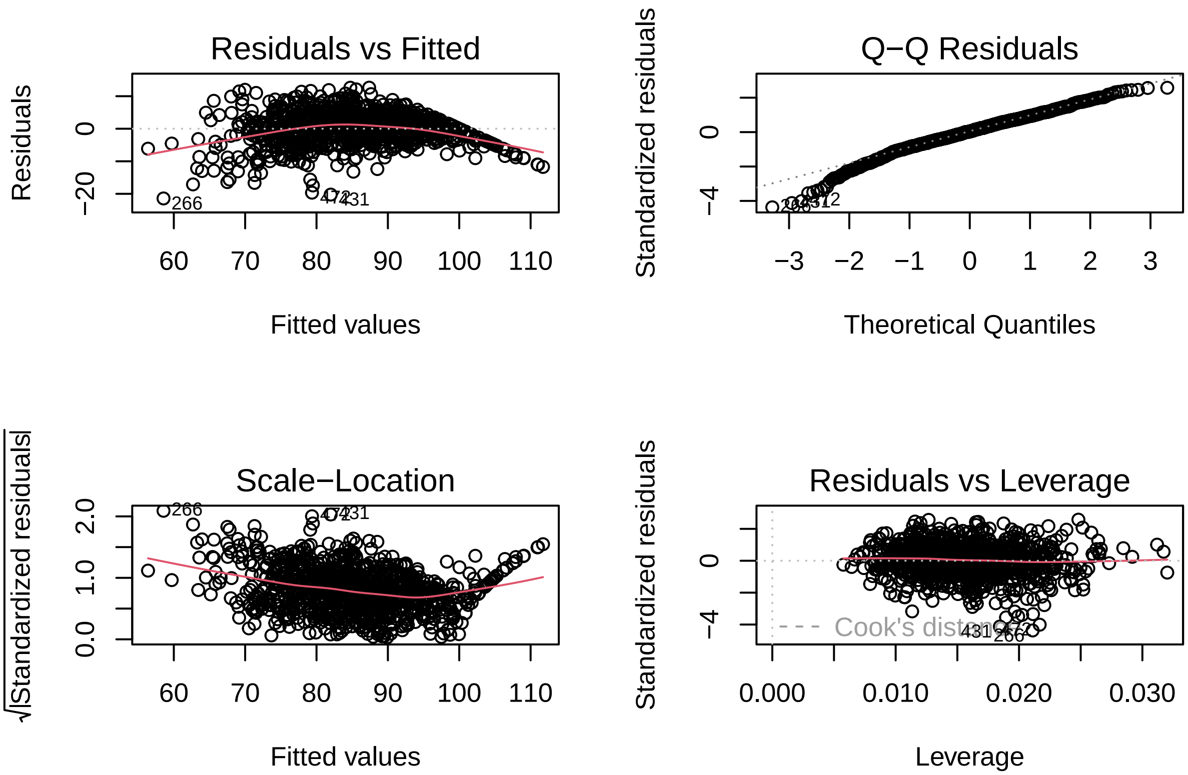
```
##          2         24         38         41         81         93        132        259
## 0.01720272 0.01547448 0.01665947 0.01704565 0.01582934 0.01872612 0.01522460 0.01470026
##        323        347        376        415        417        432        591        606
## 0.01819318 0.01567681 0.01522334 0.01918706 0.01639125 0.01577317 0.01885359 0.01637299
##        682        728        736        756        770        775        865        907
## 0.01583131 0.01738796 0.01911633 0.01551769 0.01499241 0.01538197 0.01696881 0.01760896
##        956        961        982
## 0.01521520 0.01480789 0.01859239
```

```
## [1] 0.01547163
```




```
## [1] 4.064152
```

```
## named integer(0)
```




```
##
##  RESET test
##
## data:  mod3
## RESET = 13.876, df1 = 2, df2 = 949, p-value = 1.148e-06
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod3
## BP = 6.5612, df = 6, p-value = 0.3633
```

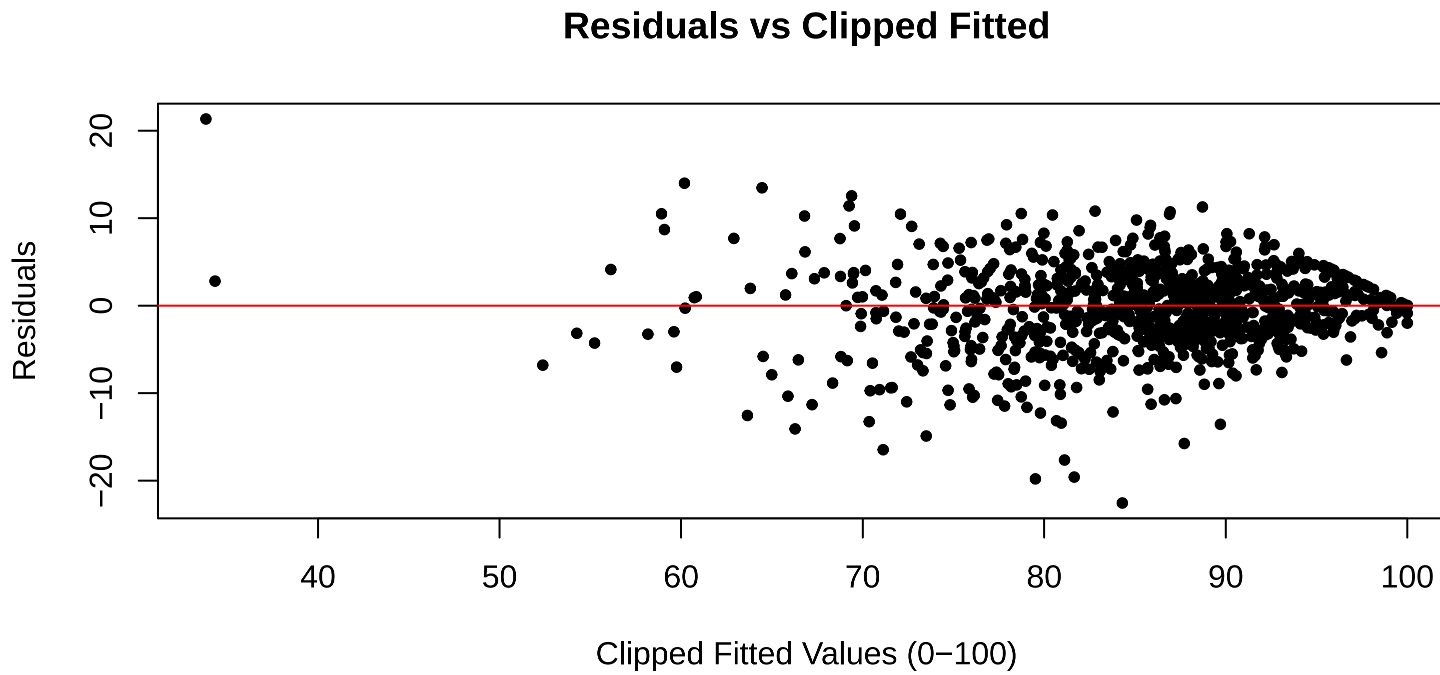
Findings: The Breusch-Pagan test suggests that the residuals of the final model exhibit homoskedasticity (p = 0.36), satisfying the assumption of constant variance.

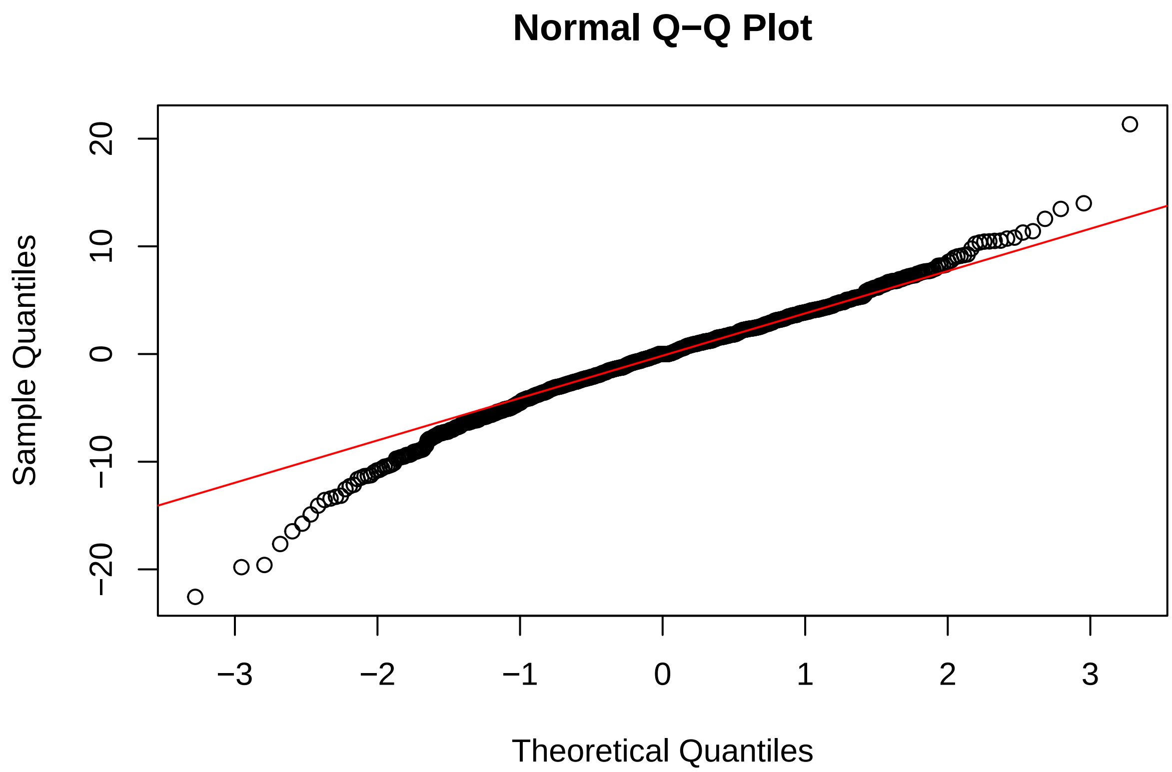
However, the RESET test indicates potential misspecification in functional form (p < 0.001). This may reflect unmodeled nonlinearities or interactions.

Clip fitted value into [0, 100]

Exam scores are naturally bounded (e.g., 0 to 100). Traditional linear regression models do not respect these bounds, often predicting unrealistic value like 110. Such predictions reduce model credibility and can mislead.

Residule plot for clipped model





Findings: The residuals vs. clipped fitted values plot shows no clear nonlinear trend across most of the prediction range, suggesting that the linearity assumption is reasonably met in the central region.

Discussion of Results and Conclusions

Summary of Findings

After fitting multiple linear models, conducting diagnostic checks, and experimenting with transformations and interaction terms, I ultimately selected Model 3 as our final model. This model was derived from the Box-Cox transformed response using a step-wise selection approach based on BIC. It includes six key predictors:

- study_hours_per_day
- social_media_hours
- attendance_percentage
- sleep_hours
- exercise_frequency
- mental_health_rating

Model 3 achieved an adjusted R^2 of 0.792, indicating strong explanatory power. Residual diagnostics showed no clear signs of heteroskedasticity (BP test p = 0.36), and the model demonstrated stability with no influential outliers. Although the RESET test indicated minor nonlinearity, the model's simplicity and interpretability justified its selection as the final model.

To further improve the model's realism and residual behavior, I clipped the fitted values to the range [0, 100], aligning them with the actual exam score boundaries. This adjustment eliminated implausible predictions and led to a clearer residual pattern, satisfying the linearity assumption.

Insights

```r
coef(mod3)
```

```
##          (Intercept)   study_hours_per_day    social_media_hours attendance_percentage
##         -2010164.944            684708.398           -122488.172              8603.112
##          sleep_hours    exercise_frequency   mental_health_rating
##           154028.449             71049.075            107237.800
```

These results suggest that habits related to focus, rest, and well-being (studying, sleeping, mental health, etc.) contribute positively to academic outcomes, while distractions such as social media use have negative effects. Importantly, the model allows me to quantify these effects in the transformed score space, offering insights into relative importance.

Habit Change	Estimated Change in Transformed Exam Score
Increase study time by 1 hour per day	+684,708
Increase sleep time by 1 hour per day	+154,02
Improve mental health rating by 1 point (1–10)	+107,238
Increase attendance by 1%	+8,603
Increase exercise frequency by 1 unit	+71,049
Increase social media use by 1 hour per day	−122,488

Note: Effects are on the Box-Cox transformed exam score scale ($\lambda = 3.59$). While exact changes in raw scores are not directly interpretable, the *direction* and *relative magnitude* of each habit's effect are valid.

Open Questions and Remaining Challenges

- The RESET test ($p < 0.001$) suggests some functional form misspecification. This remains an area for future refinement.
- Due to the Box-Cox transformation ($\lambda = 3.59$), direct interpretation of units is not intuitive. Though relative effects are valid, mapping back to the original exam score scale is non-trivial.
- Factors like motivation, teacher quality, or social economic background were not included but may also influence academic performance.