Time Well Spent? A Study on How High School Students' Activities Affect Their Grades

ECON 471 Research Project

Ramsey El Lethy, Darwin Mah

04-29-2024

Introduction

It is no secret that, as students, we see a direct relationship between the time we invest in studying concepts and our exam performance. However, students' lives incorporate so much more activity than just studying. Whether finding time to care for the people in our lives or setting aside a moment for ourselves, students consistently add numerous activities to their schedules. Some of these are just too much, some seriously help, and some have no effect. Time is significant to us, so learning how we spend it can give us insights into our opportunities and shortcomings as we traverse this educational journey.

Abstract

To assess the impact of how students divide their time on their academic performance, we acquired data from two thousand individuals, all of whom are high schoolers of different genders and ages. Within the data set, each student was recorded, including whether they took part in extracurricular activities if they had a part-time job, the number of absences, and the number of hours students spent studying a week. We also recorded test scores for multiple subjects, including math, history, physics, chemistry, biology, english, and geography. Our model represented true or false records of having a part-time job or an activity as a 0 for false and a 1 for true. We performed multiple regressions, each for every individual subject. Regressing the same independent variables on the different scores. We aimed to identify the relationship between these factors and their test scores.

Table 1: Summary Statistics for Original Model

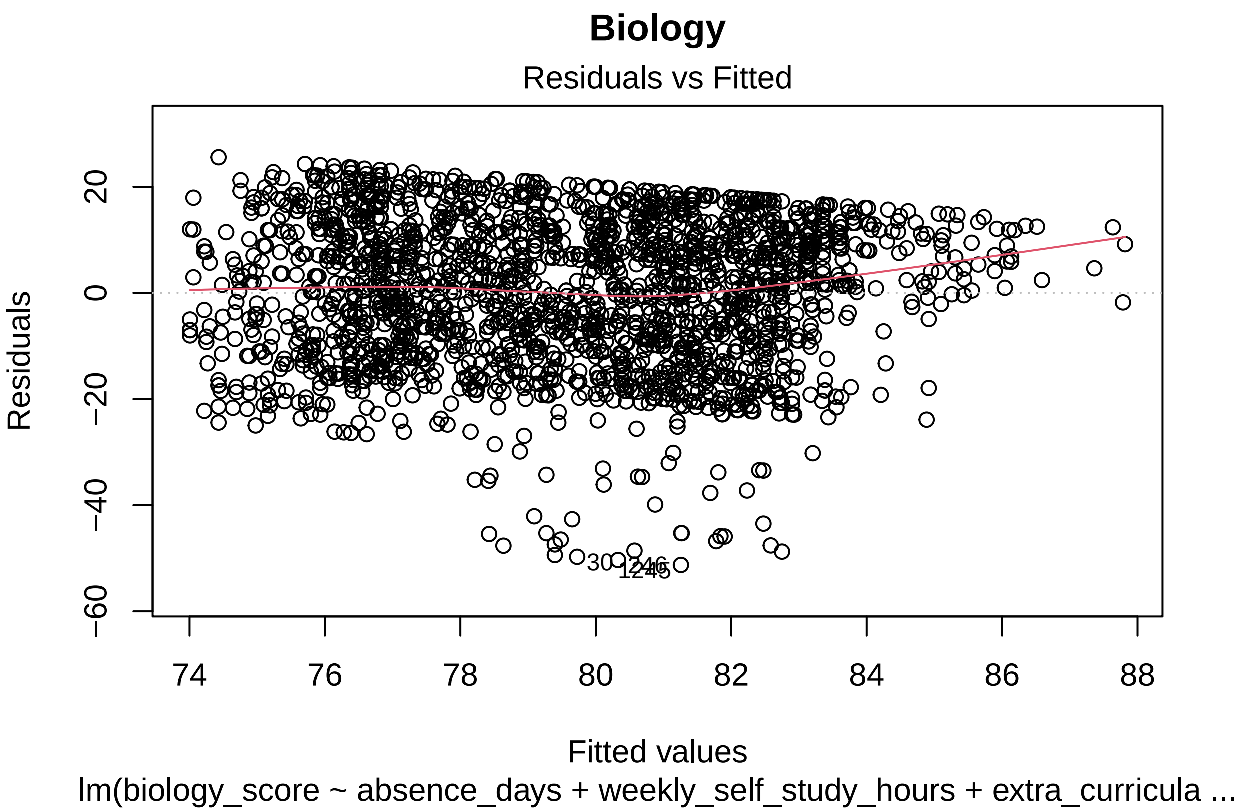
	Model	CoefficientValue	StdError	TValue	Pr	FStatistic	ResidualDF
(Intercept)...1	Math	79.8594075	0.7229505	110.4631745	0.0000000	109.47095	1995
absence_days...2	Math	-0.6090796	0.1077443	-5.6530106	0.0000000	109.47095	1995
weekly_self_study_hours...3	Math	0.3688413	0.0234189	15.7497230	0.0000000	109.47095	1995
extra_curricular...4	Math	-1.0640486	0.6663606	-1.5968060	0.1104673	109.47095	1995
has_job...5	Math	-3.2071014	0.7617484	-4.2101847	0.0000267	109.47095	1995
(Intercept)...6	Physics	79.9886727	0.7368076	108.5611325	0.0000000	27.68196	1995
absence_days...7	Physics	-0.3530683	0.1098095	-3.2152814	0.0013240	27.68196	1995
weekly_self_study_hours...8	Physics	0.1708945	0.0238678	7.1600487	0.0000000	27.68196	1995
extra_curricular...9	Physics	0.0274605	0.6791331	0.0404346	0.9677507	27.68196	1995
has_job...10	Physics	-2.5184834	0.7763492	-3.2440086	0.0011981	27.68196	1995
(Intercept)...11	History	76.5001681	0.7367620	103.8329395	0.0000000	44.38538	1995
absence_days...12	History	-0.2255766	0.1098027	-2.0543819	0.0400688	44.38538	1995
weekly_self_study_hours...13	History	0.2670831	0.0238663	11.1908000	0.0000000	44.38538	1995
extra_curricular...14	History	0.8482651	0.6790910	1.2491184	0.2117683	44.38538	1995
has_job...15	History	-1.6237800	0.7763011	-2.0916884	0.0365926	44.38538	1995
(Intercept)...16	Biology	76.4337269	0.8117978	94.1536568	0.0000000	20.58129	1995
absence_days...17	Biology	-0.2426707	0.1209856	-2.0057821	0.0450143	20.58129	1995
weekly_self_study_hours...18	Biology	0.2115550	0.0262970	8.0448407	0.0000000	20.58129	1995
extra_curricular...19	Biology	0.0507036	0.7482532	0.0677626	0.9459814	20.58129	1995
has_job...20	Biology	1.7131244	0.8553637	2.0028023	0.0453336	20.58129	1995

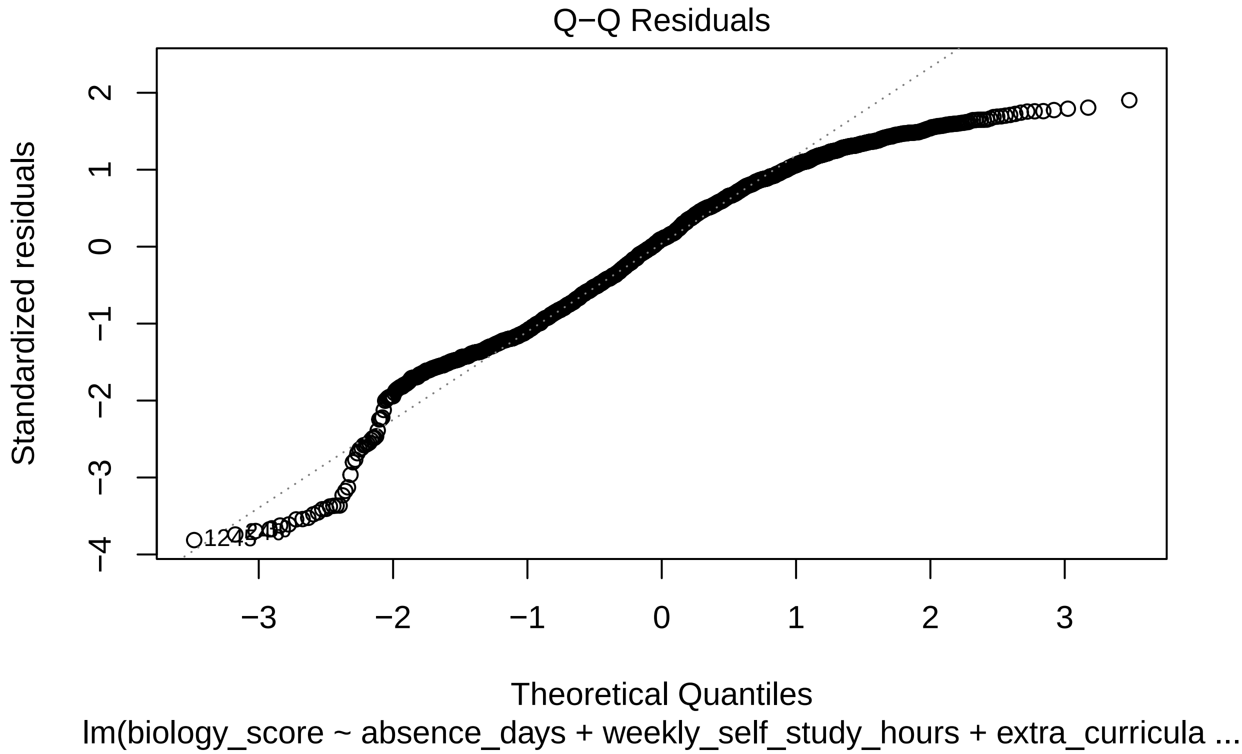
Observations and Analysis

We perform multiple regression analyses for multiple subjects to find consistent relationships between the variables across each subject. The first thing we looked at was the individual significance of each variable; we found that consistently, the only insignificant factor was whether or not students participated in extracurricular activities. In our analysis of the overall significance of the model, a rejection of the null hypothesis due to a large F statistic coincides with at least one of our parameters, in this case, extracurricular, not being able to explain student performance in any capacity. We see consistent relationships between student performance for the rest of our parameters. Table 1 below presents the values of the coefficients and respective tests. Aligning with common perceptions about studying, we see that our weekly self-study hour parameter is positive in all models. This means that an increase in time spent studying increases student performance. Looking at students' absences, we see a negative relationship between being at school and their performance. Interestingly, students with a job have decreased math, physics, and history performance. However, biology gave us conflicting results, expressing a positive relationship between having a job and performance.

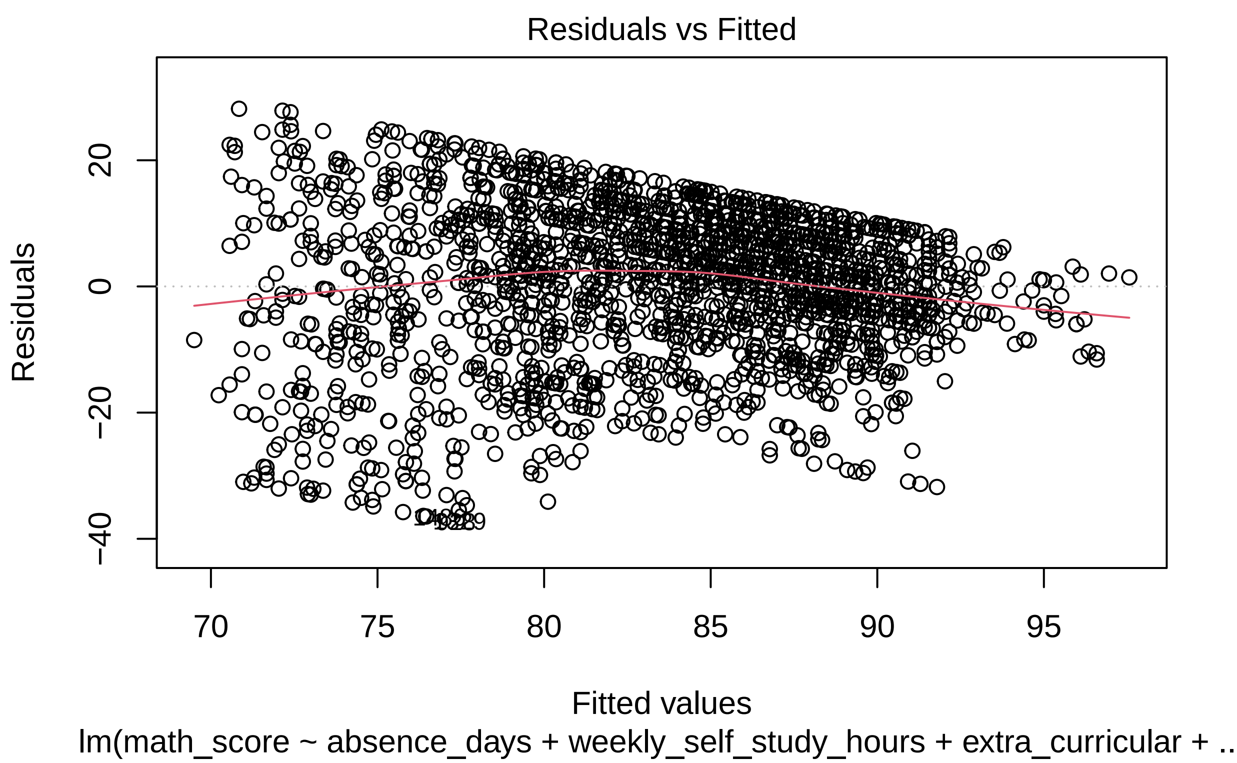
Residual Analysis

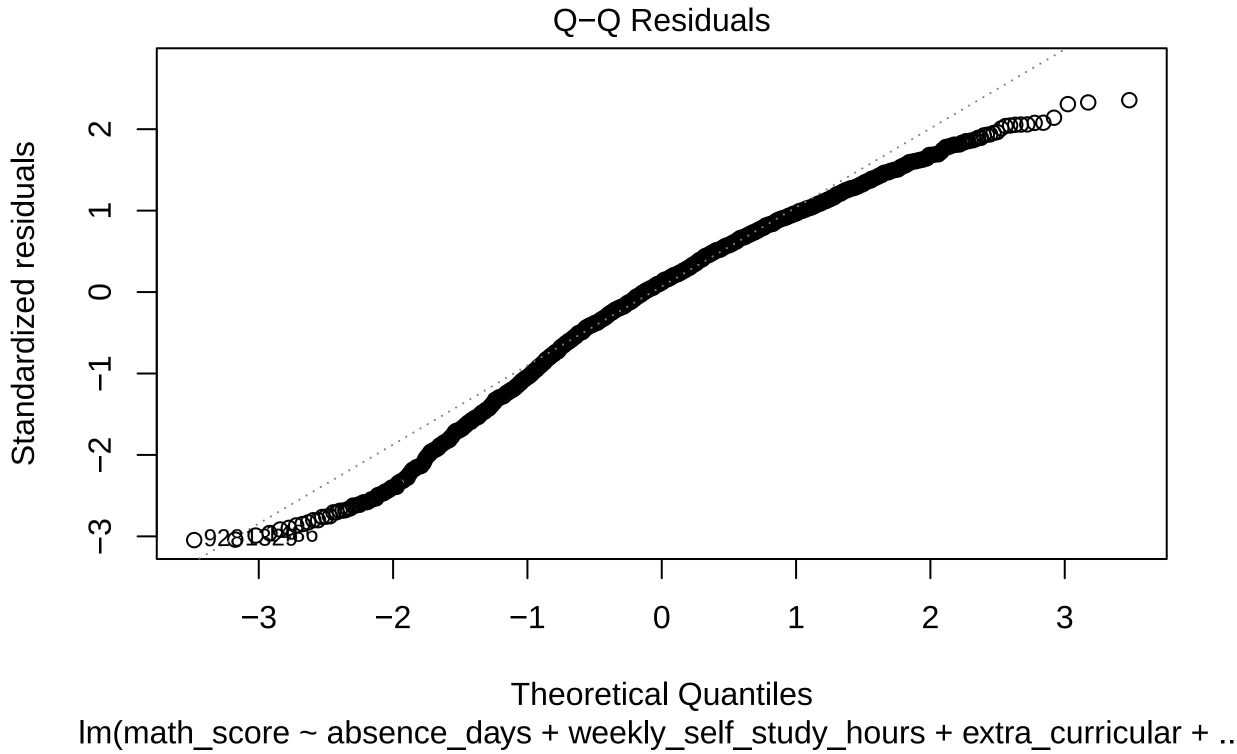
These conflicting results of students having a job between biology and the rest of the subjects led us to believe there might be some possible error in regression. We were not inclined to believe in the existence of a linear relationship between our independent variables; in other terms, multicollinearity, because in all of our subjects' regressive output, the F-test for overall significance coincided with the T-test for individual significance. Our next thought was to analyze the residuals' linearity and our distribution's normality. We generated plots for each model that can give us insights about non-normality and heteroscedasticity. Each plot grouping has been labeled according to its respective subject. Consisting of two different graphs, first, we plotted our Residuals vs. Fitted values, generating a line best fit for our residuals; we can see if there is a linear relationship between our residuals and the fitted values. For non-normality, we have plotted our standardized residuals against theoretical quantities for our standard distribution. This is represented as a Q-Q Plot, deviations from the one-to-one linear relationship might explain where skewness may exist.



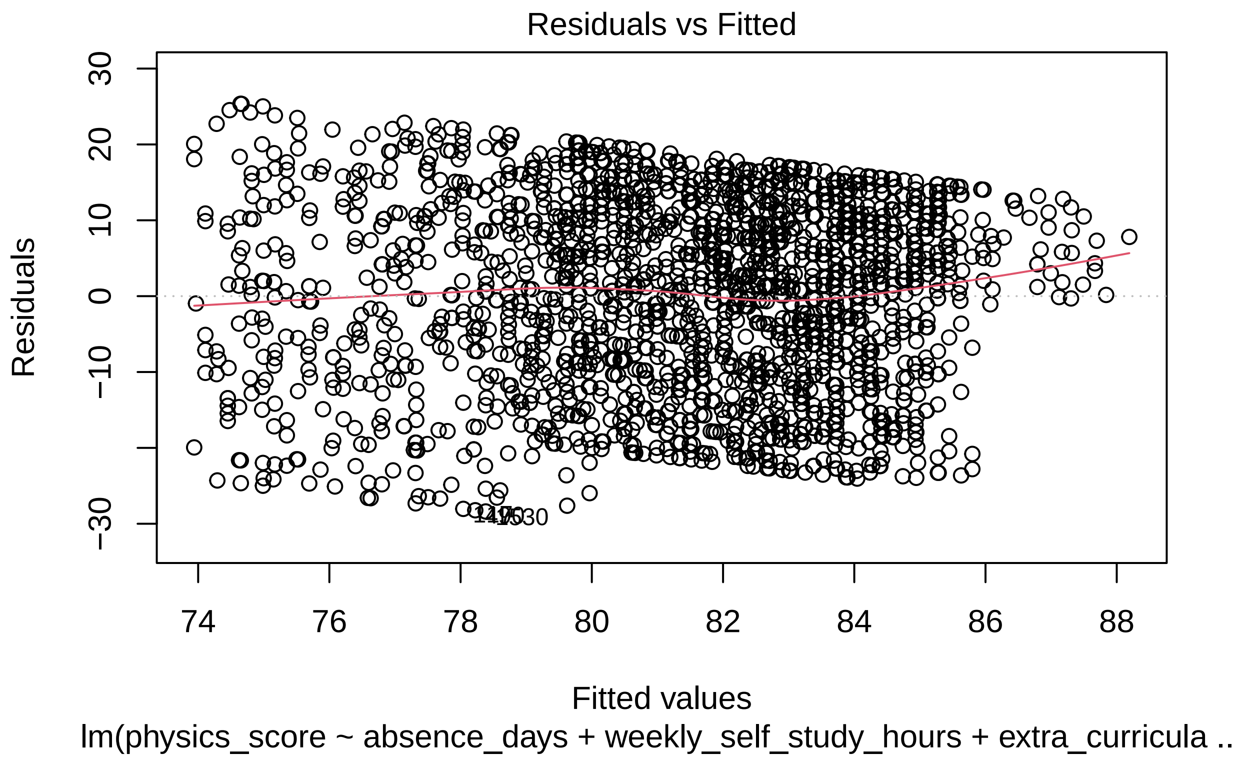

Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(biology_score ~ absence_days + weekly_self_study_hours + extra_curricula ...

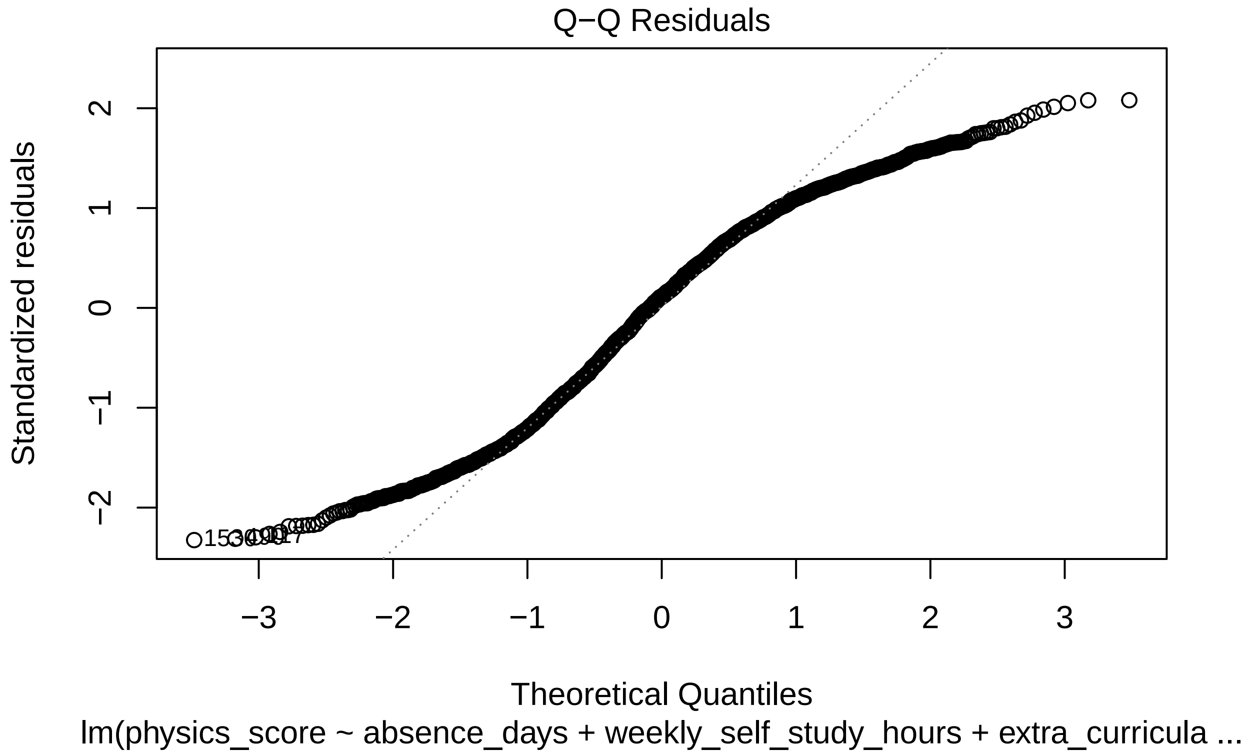
Math


Residuals vs Fitted
Residuals
Fitted values
lm(math_score ~ absence_days + weekly_self_study_hours + extra_curricular + ...

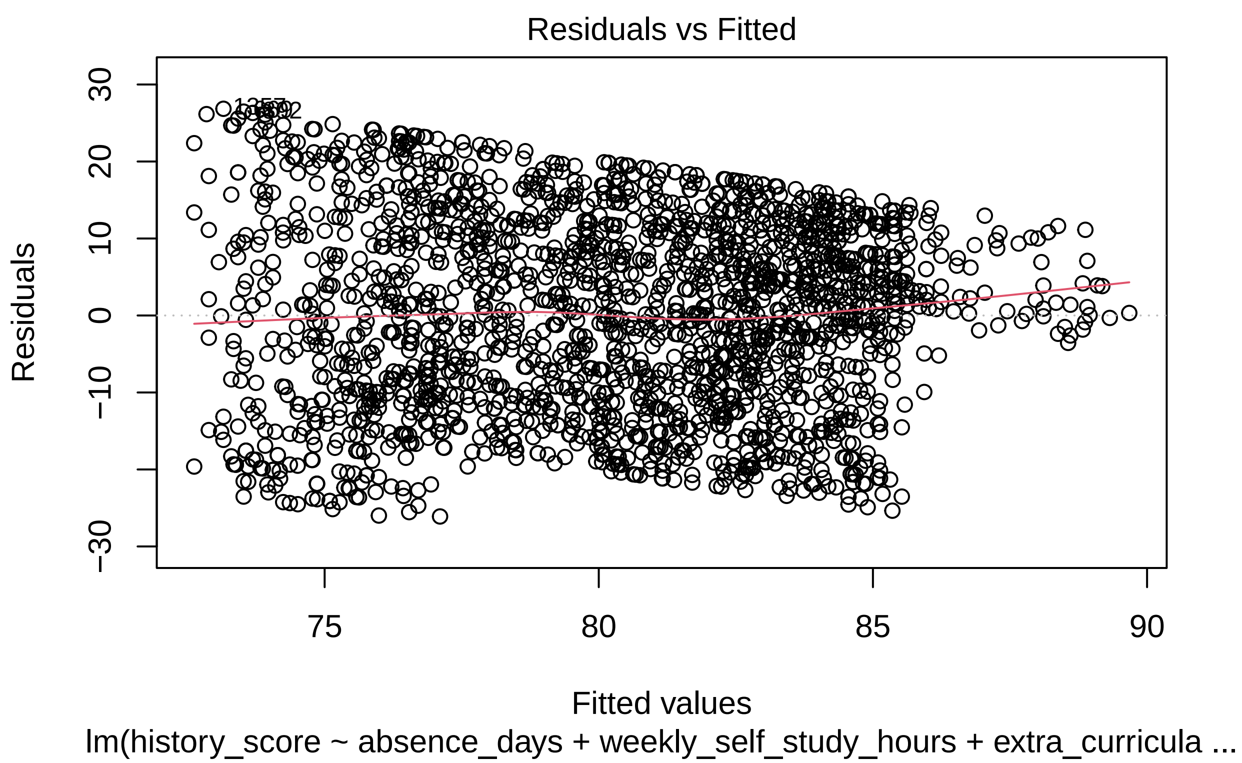

Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(math_score ~ absence_days + weekly_self_study_hours + extra_curricular + ...

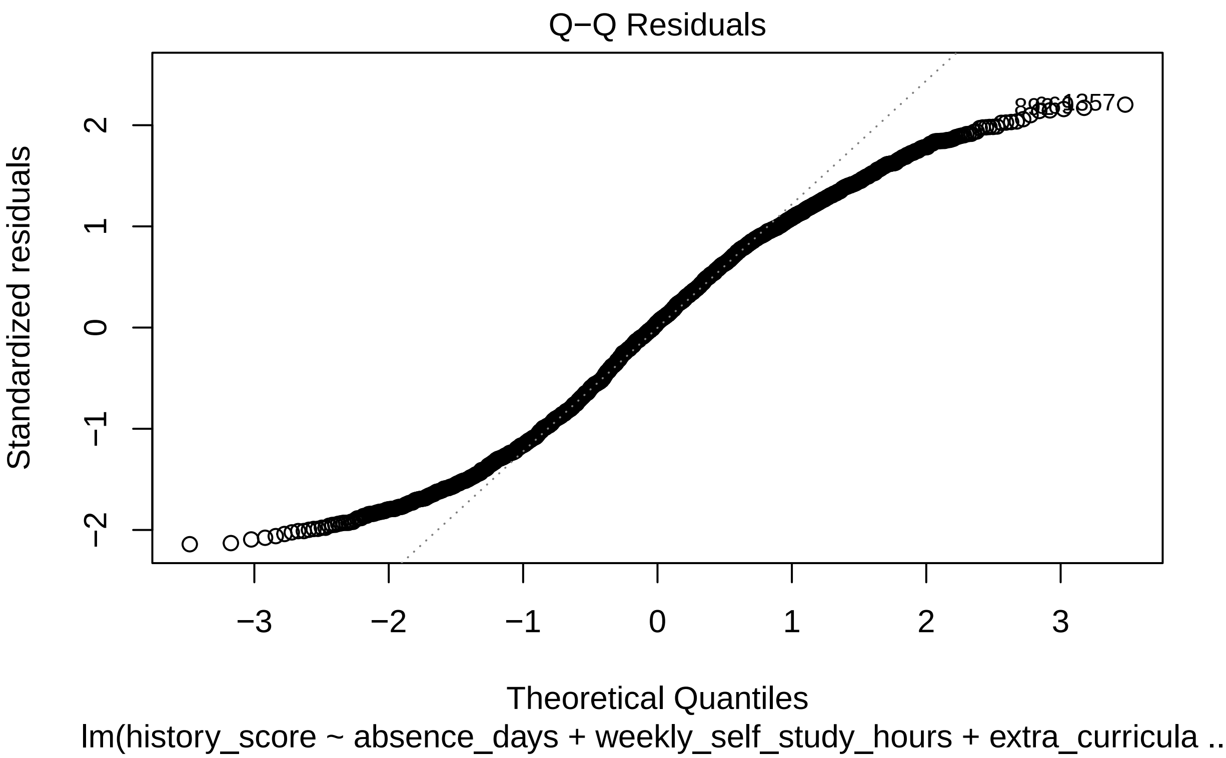
Physics


Residuals vs Fitted
Residuals
Fitted values
lm(physics_score ~ absence_days + weekly_self_study_hours + extra_curricula ...


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(physics_score ~ absence_days + weekly_self_study_hours + extra_curricula ...

History


Residuals vs Fitted
Residuals
Fitted values
lm(history_score ~ absence_days + weekly_self_study_hours + extra_curricula ...



Results of Residual Analysis

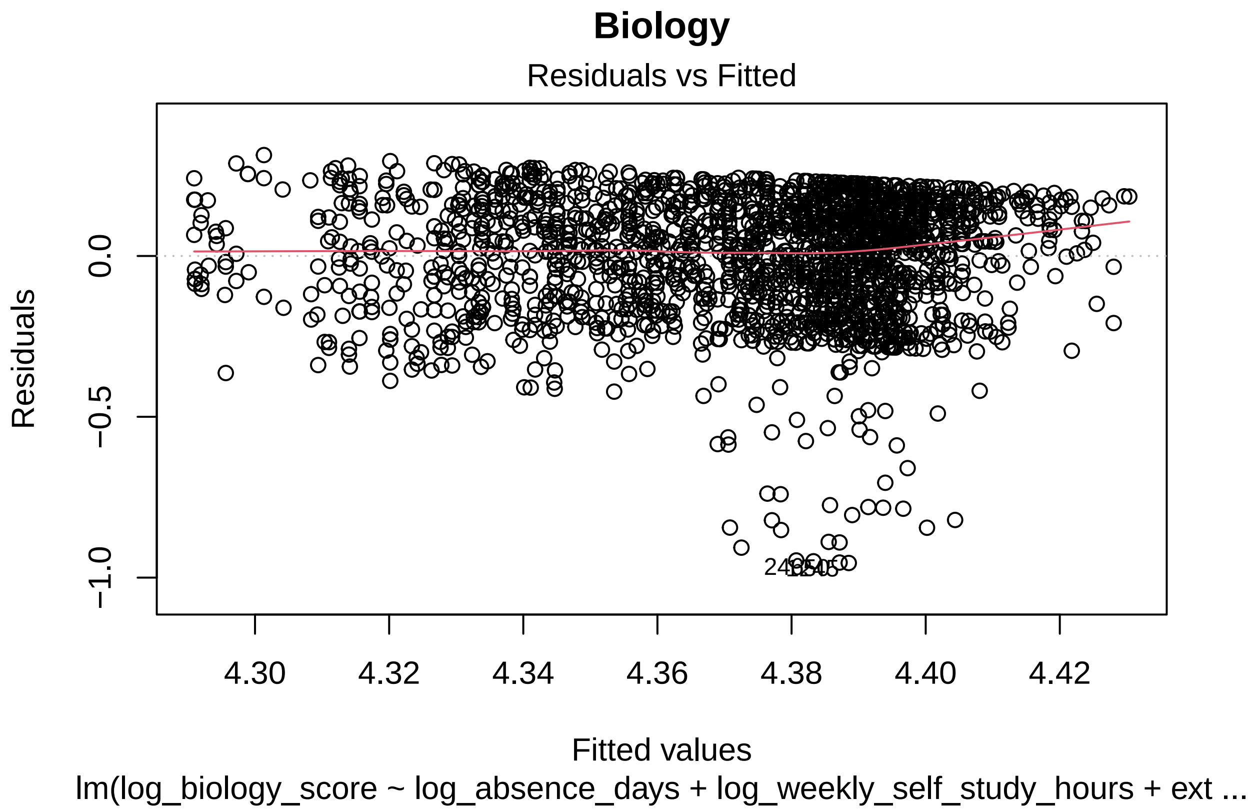
Plotting our Standardized Residuals against our fitted values consistently showed us a non-linear relationship. We would like to see our line of best fit horizontal and linear, but there seems to be some quadratic relationship between the residuals and fitted values. This tells us that our errors are not linear. Subsequently, these results put into question any statistical inference including the significance of our variables. In our Q-Q plot, we hope to have residuals hugging close to our theoretical distribution. We see a deviation in our extreme theoretical values, which leads us to believe that our model departs from normality. Skewness in our distribution violates the assumption of normality in our model, which also questions the overall significance of our model. We cannot entirely believe that students' extracurricular activities are the only insignificant parameter. We must attempt to correct this in some capacity to gain valid results.

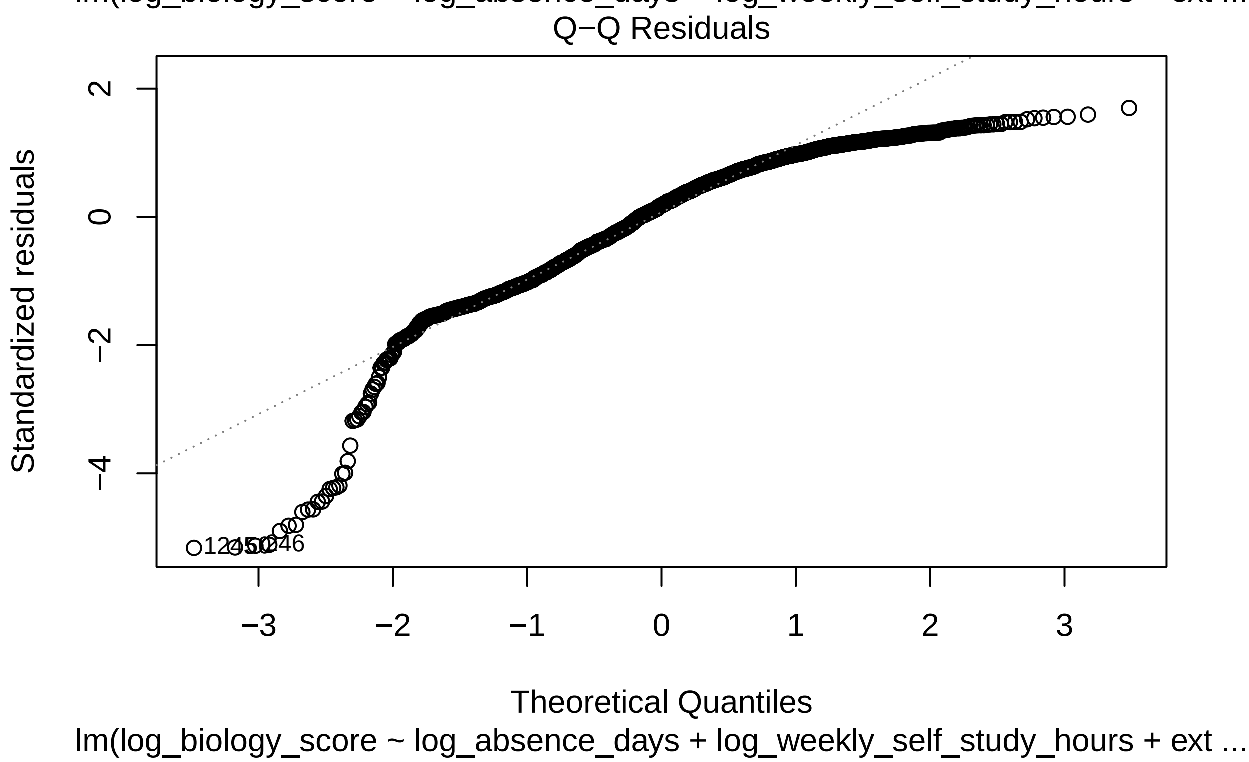
Table 2: Summary Statistics for Log-Transformed Models

	Model	CoefficientValue	StdError	TValue	Pr	FStatistic	FPValue	ModelDF	ResidualDF
(Intercept)...1	Log Math	4.2985120	0.0154558	278.1157072	0.0000000	113.99660	0	4	1995
log_absence_days...2	Log Math	-0.0250738	0.0061727	-4.0620314	0.0000505	113.99660	0	4	1995
log_weekly_self_study_hours...3	Log Math	0.0650945	0.0038971	16.7034330	0.0000000	113.99660	0	4	1995
extra_curricular...4	Log Math	-0.0172228	0.0089441	-1.9255968	0.0542967	113.99660	0	4	1995
has_job...5	Log Math	-0.0436481	0.0102973	-4.2387889	0.0000235	113.99660	0	4	1995
(Intercept)...6	Log Physics	4.3670102	0.0149847	291.4305945	0.0000000	24.24709	0	4	1995
log_absence_days...7	Log Physics	-0.0174401	0.0059846	-2.9141796	0.0036060	24.24709	0	4	1995
log_weekly_self_study_hours...8	Log Physics	0.0237094	0.0037783	6.2751633	0.0000000	24.24709	0	4	1995
extra_curricular...9	Log Physics	-0.0009574	0.0086715	-0.1104134	0.9120927	24.24709	0	4	1995
has_job...10	Log Physics	-0.0354982	0.0099834	-3.5557117	0.0003857	24.24709	0	4	1995
(Intercept)...11	Log History	4.3031768	0.0150991	284.9963931	0.0000000	40.47456	0	4	1995
log_absence_days...12	Log History	-0.0128039	0.0060302	-2.1232846	0.0338528	40.47456	0	4	1995
log_weekly_self_study_hours...13	Log History	0.0393723	0.0038071	10.3417820	0.0000000	40.47456	0	4	1995
extra_curricular...14	Log History	0.0090023	0.0087377	1.0302885	0.3029995	40.47456	0	4	1995
has_job...15	Log History	-0.0218762	0.0100596	-2.1746552	0.0297731	40.47456	0	4	1995
(Intercept)...16	Log Biology	4.3154069	0.0177777	242.7427666	0.0000000	10.98388	0	4	1995
log_absence_days...17	Log Biology	-0.0101562	0.0071000	-1.4304498	0.1527446	10.98388	0	4	1995
log_weekly_self_study_hours...18	Log Biology	0.0264967	0.0044825	5.9111262	0.0000000	10.98388	0	4	1995
extra_curricular...19	Log Biology	-0.0001242	0.0102878	-0.0120752	0.9903668	10.98388	0	4	1995
has_job...20	Log Biology	0.0199785	0.0118442	1.6867724	0.0918035	10.98388	0	4	1995

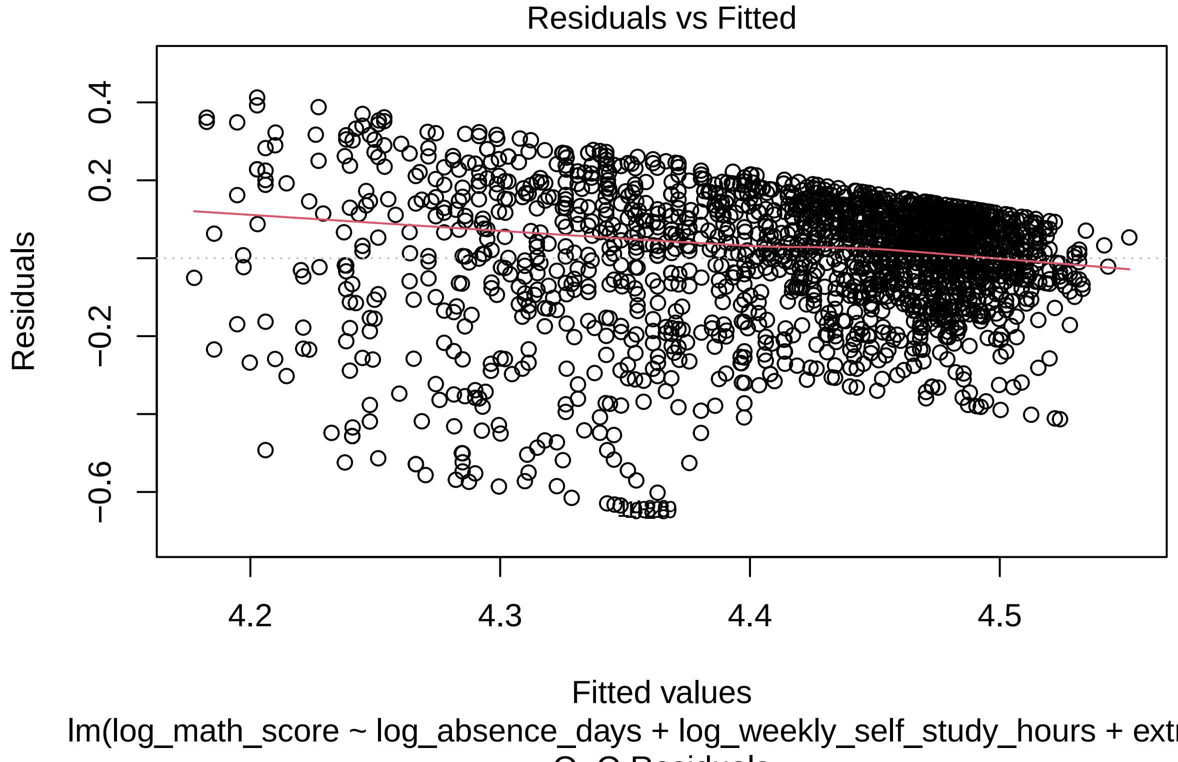
Correcting Errors in Regression

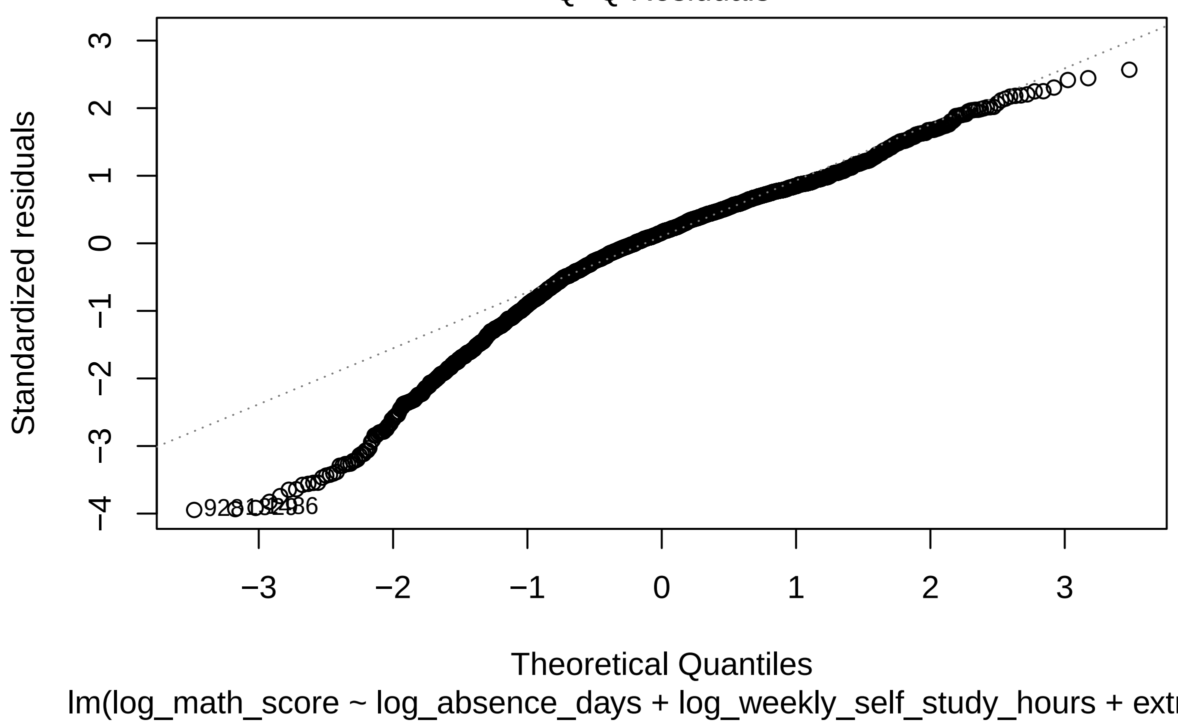
We regressed the same parameters over a log-transformed model to correct our non-linear errors. Performing a logarithmic transformation can correct for heteroscedasticity among errors in some cases. So, we decided to give it a try. Not only can it correct for non-linear errors, but it can reduce the skewness that we might be dealing with due to our model facing non-normality. If we spread out the data clumps and combine them for another regression, we might see different results that follow normality. In this case of trying to correct for non-normality and heteroscedastic errors, a logarithmic transformation became the apparent treatment of this model. Table 2, posted above, shows us the new results of our regression. If we can correct our errors in regression and still receive the same or similar results, we would have much more reason to believe in the relationships of our independent variables on student performance.


Biology
Residuals vs Fitted
Residuals
0.0
−0.5
−1.0
4.30
4.32
4.34
4.36
4.38
4.40
4.42
Fitted values
lm(log_biology_score ~ log_absence_days + log_weekly_self_study_hours + ext ...

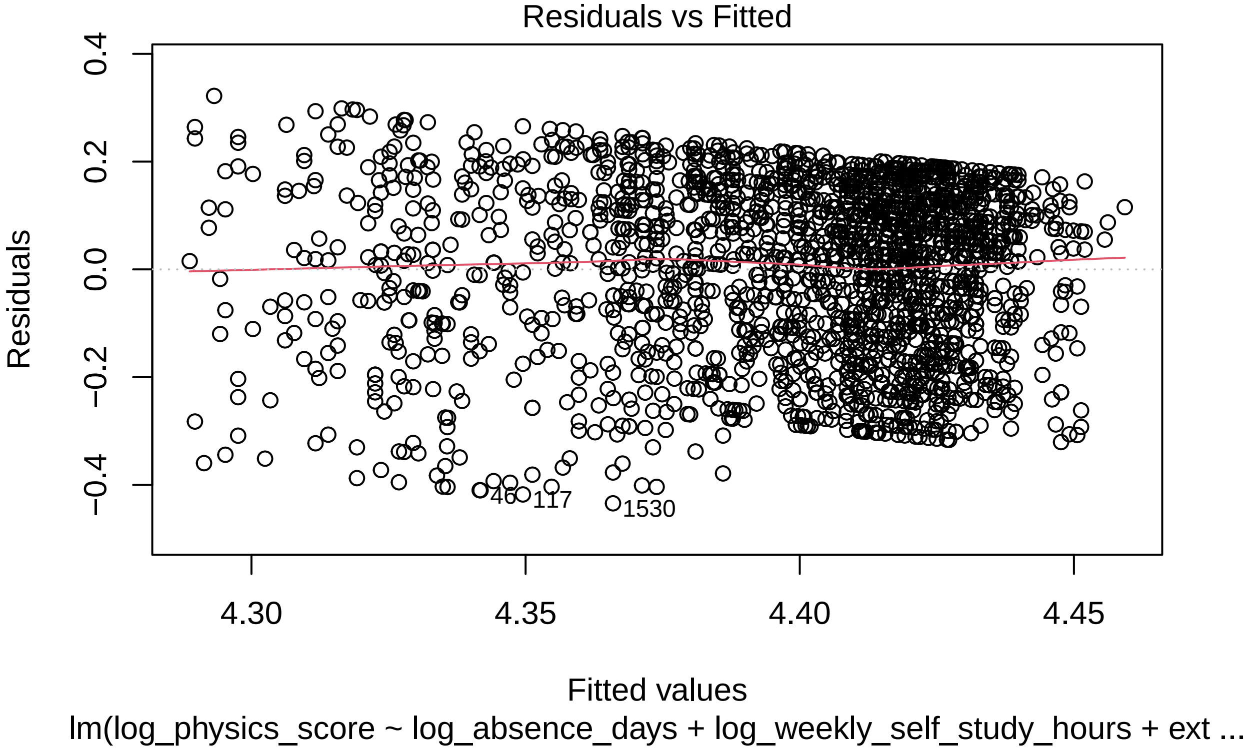

Q−Q Residuals
Standardized residuals
2
0
−2
−4
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(log_biology_score ~ log_absence_days + log_weekly_self_study_hours + ext ...

Math

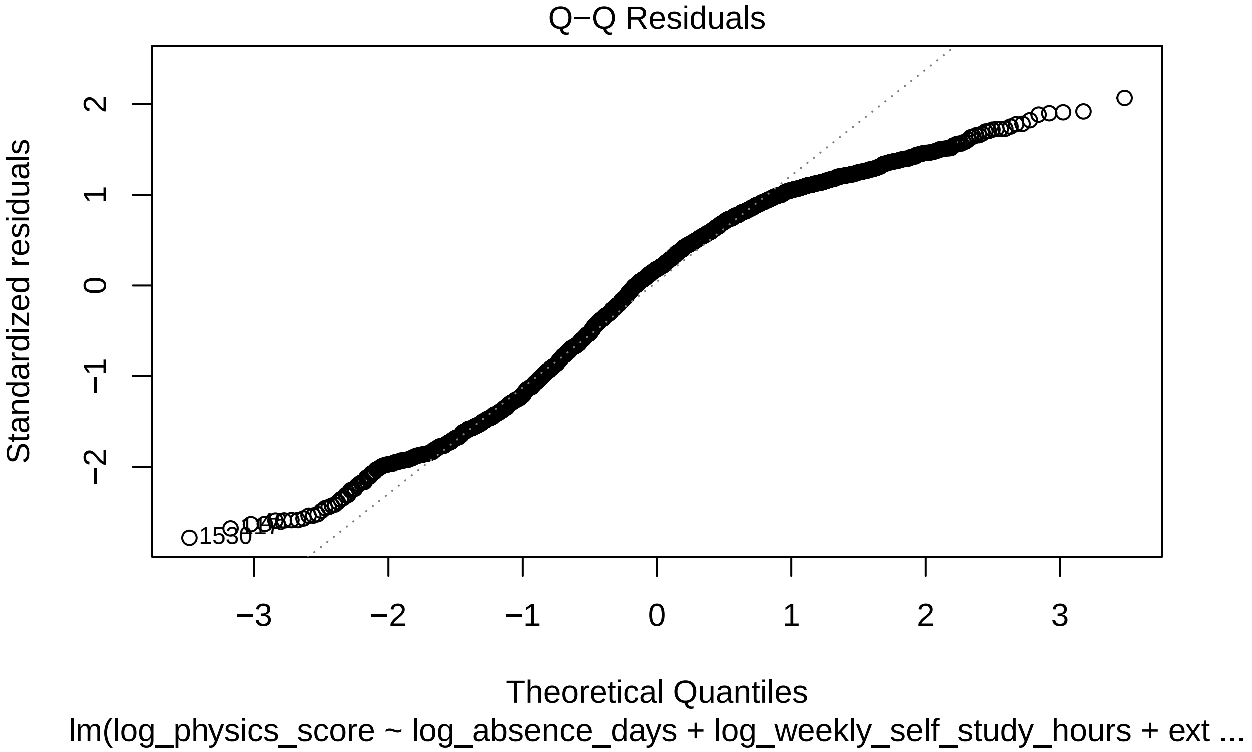

Residuals vs Fitted
Residuals
0.4
0.2
−0.2
−0.6
4.2
4.3
4.4
4.5
Fitted values
lm(log_math_score ~ log_absence_days + log_weekly_self_study_hours + extra_ ..


Q−Q Residuals
Standardized residuals
3
2
1
0
−1
−2
−3
−4
Theoretical Quantiles
lm(log_math_score ~ log_absence_days + log_weekly_self_study_hours + extra_ ..

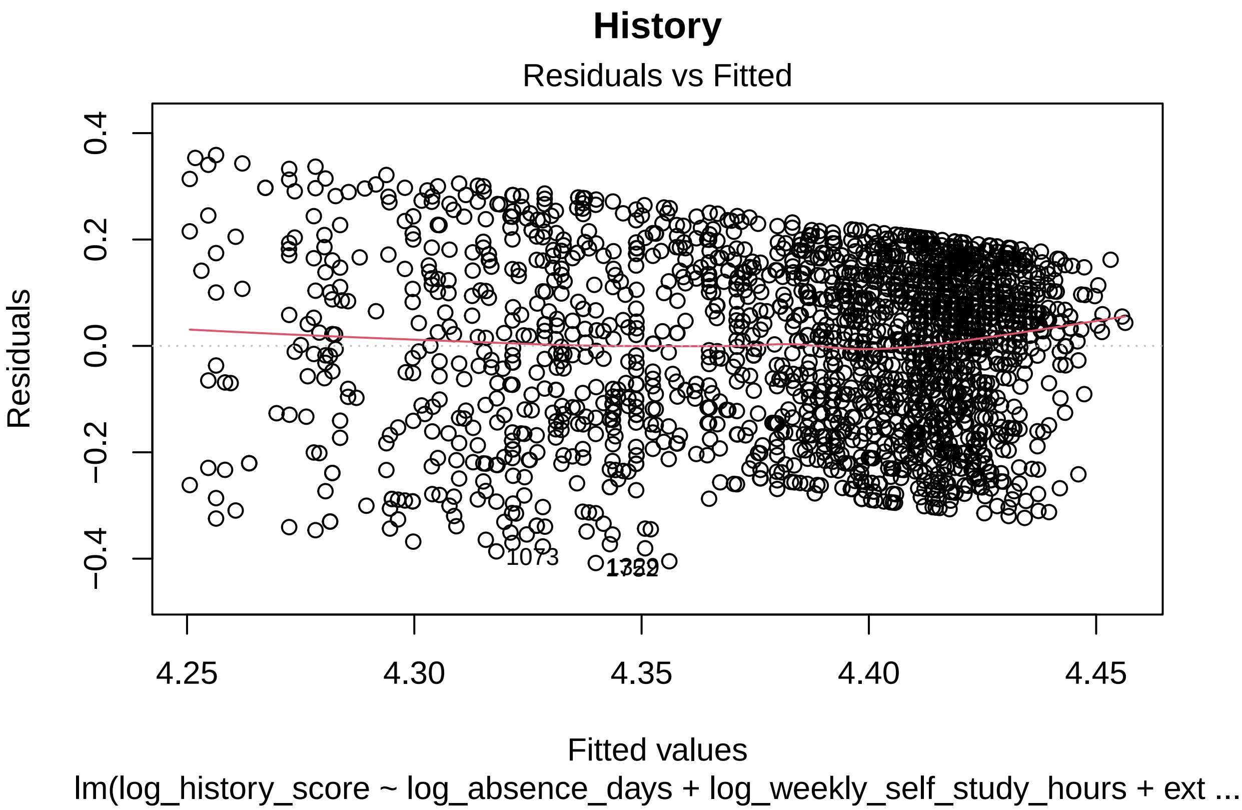
Physics

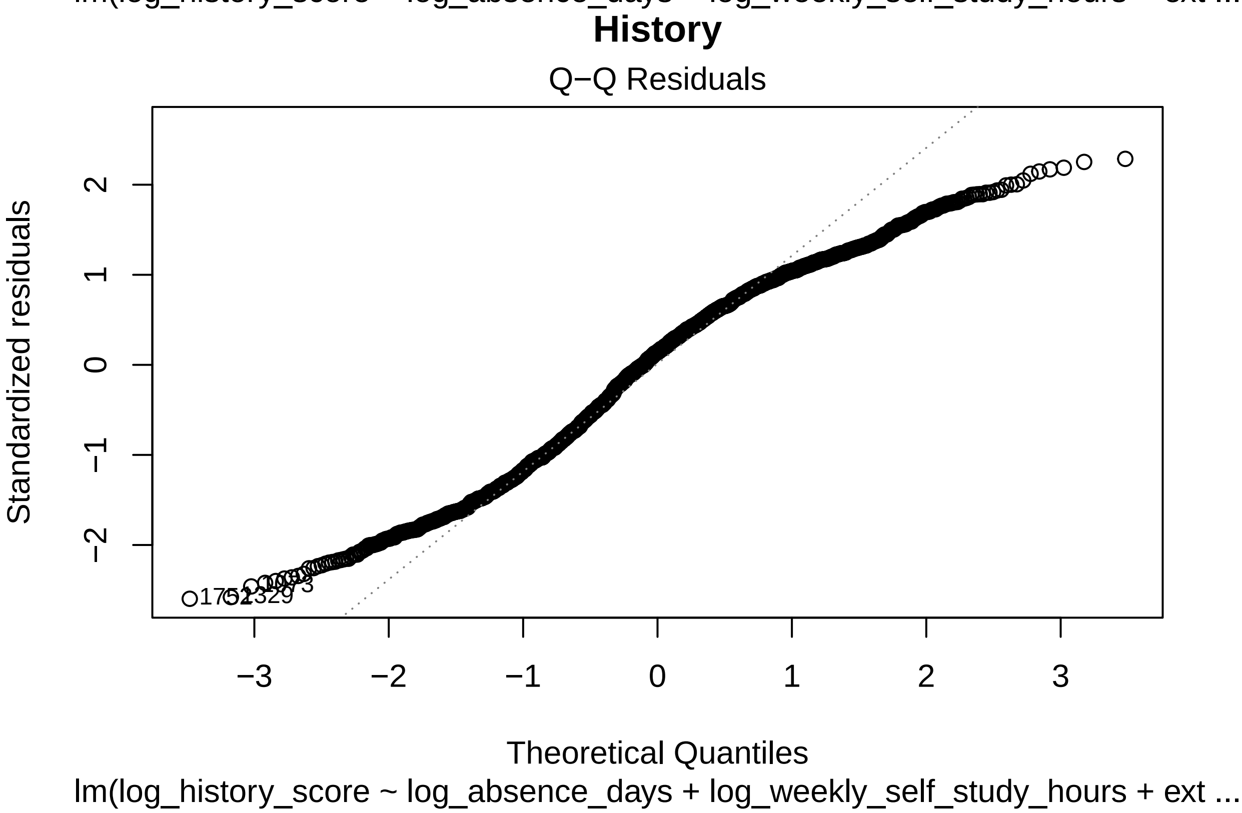


Residuals vs Fitted
Residuals
0.4
0.2
0.0
−0.2
−0.4
4.30
4.35
4.40
4.45
46
117
1530
Fitted values
lm(log_physics_score ~ log_absence_days + log_weekly_self_study_hours + ext ...



Q−Q Residuals
Standardized residuals
2
1
0
−1
−2
−3
3
1530
Theoretical Quantiles
lm(log_physics_score ~ log_absence_days + log_weekly_self_study_hours + ext ...


History
Residuals vs Fitted
Residuals
0.4
0.2
0.0
−0.2
−0.4
1073
1732
4.25
4.30
4.35
4.40
4.45
Fitted values
lm(log_history_score ~ log_absence_days + log_weekly_self_study_hours + ext ...


History
Q−Q Residuals
Standardized residuals
2
1
0
−1
−2
−3
3
Theoretical Quantiles
lm(log_history_score ~ log_absence_days + log_weekly_self_study_hours + ext ...

Results of Log Transformation

Visualizing the same residual plots and comparing our non-log-transformed residual to the log-transformed, we can consistently see more linearity among our residuals. So, heteroscedasticity is corrected to some degree. Analyzing our Q-Q Results, we see that normality is corrected on the extreme positive end, but we lose normality sooner in the negative direction. Listed in Table 2, The significance of our model aligns with our original regression for math, physics, and history, being that extracurricular activities were the only factor that was individually insignificant. Looking then at the relationships of each variable, we still see that studying is positive, absences are harmful, and having a job is also damaging. Looking at biology, we find that studying was the only significant factor in regression. This led us to believe the existence of other confounding factors that were not present in our data set, such as what kind of job students were performing, potentially students were a part of activities that were related to the subject of study, like shadowing in a lab setting for biology, etc.

Conclusion

Though our confidence in our relationships has increased as a result of transforming our data, we cannot definitively claim the relationships observed are statistically significant because we are diverging from normality, as well as our errors are slightly heteroscedastic. Aside from this, we have potentially reinforced the concept of studying and its positive attributes on student performance, as well as potentially identifying the relationship between how students spend their time outside of school and performance. Pulling quantitative and qualitative data together opened the door for us to see what life is like for students juggling plenty of activities. To further iron out and correct the errors in regression, we would like to observe more data from different attributes. Things like how much time students spend with teachers outside of class, learn about their aspirations, etc. Introducing more factors could potentially grow our understanding of these relationships, but may also invite more violations of regressive assumptions that need to be corrected.