Paris Dauphine University
MIDO departement

Project :Linear Models and their generalization :

Factors Influencing Students' Exam Scores : A Linear Regression Analysis

Teacher :
Ms. Katia MEZIANI
Mr. Henri PANJO

Author :
Antonin Durousseau
Thaïs Forest

2025-2026

Contents

1 Introduction

1.1 Problematic

Exam grades are the main criterion used by universities to rank students, admit them, or in some cases dismiss them. As universities worldwide report growing disparities in student achievement and increasing concerns about engagement in lectures, exam performance has become a major topic in current educational debates. Attendance and academic background are now frequently cited as key determinants of student success.

In this context, this project investigates the relationship between exam grades and several explanatory variables using linear regression models. The objective is to identify and interpret the main factors influencing exam performance.

All graphics and numerical results, as well as the corresponding code, are to be found at the following github depository:

https://github.com/AntoninDurousseau/Studies/tree/main/Project/GLM_Project

AI use statement,Artificial intelligence tools were used for translation purposes (ChatGPT) and for limited debugging support, notably in the subsampling code related to stability results (30 repetitions per fraction) and the construction of df_bins_both (Claude.ai). All interpretations, analyses, and written commentary were produced and independently verified by the authors.

1.2 Dataset and Variables

The *project* dataset is a table of size 5000 x 16, containing information on 5000 students (the variable Id here stands for the student's Id number which is not relevant to our prediction). Fifteen variables were measured for each student. These variables are presented in the following table, along with their relevance to the exam score variable.

It appears that the variables can be grouped into three categories: variables directly related to the student's academic work, variables associated with quality of life(lifestyle-related variables), and social variables. Several interaction effects may be of interest, as the impact of certain academic or lifestyle variables on exam performance may depend on other student characteristics or contextual factors.

Furthermore, several variables provide overlapping information and should not be included simultaneously in the regression model. In particular, age and agecat, as well as attend pct and attend pct cat, represent the same underlying concepts in continuous and categorical forms and would introduce redundancy and multicollinearity if included together. Similarly, sleep related variables may be correlated and should be handled with caution, with model selection guided by exploratory analysis and interpretability considerations.

Variables and their relevance to exam performance

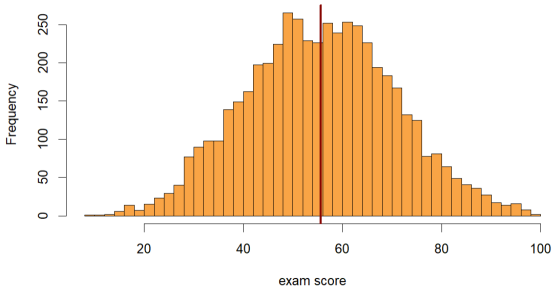
Variable	Relevance
Y	Score obtained at the exam; this is the response variable that the analysis aims to explain and predict.
Age	Student's age in years; it may reflect academic maturity or atypical educational trajectories affecting performance.
Agecat	Categorized version of age; allows capturing potential non-linear effects of age on exam performance.
Sexe	Student's gender; may be associated with performance differences related to social or educational factors.
School type	Type of school (public or private); may reflect differences in educational resources and academic support.
Parent educ	Indicator of family cultural capital and academic support at home, often correlated with student achievement.
Study hrs	Weekly study hours; a direct measure of academic effort, typically positively associated with exam scores.
Sleep hrs	Sleep duration; adequate sleep is essential for concentration, memory, and learning efficiency.
Sleep qual	Sleep quality; complements sleep duration by capturing how restorative sleep is, which impacts cognitive performance.
Attend pct	Class attendance rate; reflects exposure to course material and is often strongly related to exam results.
Attend pct cat	Categorized attendance rate; allows the analysis of threshold effects between low and high attendance.
Web access	Internet access; facilitates access to online learning resources and independent study.
Trav time	Time spent traveling between home and school; long commutes may increase fatigue and reduce study time.
Extra act	Participation in extracurricular activities; may have positive effects (motivation, time management) or negative ones (reduced study time).
Study method	Primary study method; different learning strategies may vary in effectiveness for exam performance.

1.3 Project Goals

1. Identify and estimate a final linear regression model explaining exam scores based on the available predictors.
2. Provide a rigorous and transparent variable selection procedure, with all model selection criteria explicitly stated and justified.
3. Interpretability and predictive performance are systematically compared in order to derive a parsimonious, interpretable, and reproducible model.

2 Exploratory Data Analysis

2.1 Exam score



Distribution of Y (Exam score)

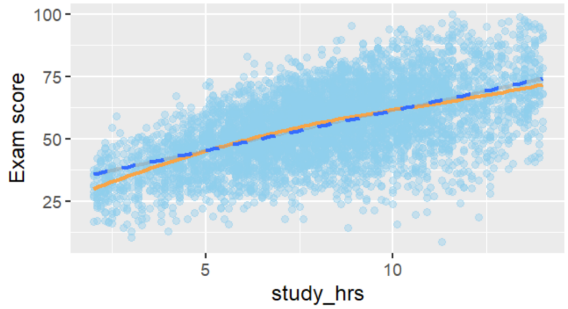
Min	1st Qu.	Median	Mean	3rd Qu.	Max
8.90	45.10	55.60	55.62	65.90	99.80

The distribution of exam scores appears symmetric, with a moderate spread without extreme values. All exam scores fall within the expected range $[0; 100]$, therefore, values such as 8.9 are plausible and do not correspond to aberrant observations. The overall shape of the distribution suggests that a Gaussian linear model is appropriate for modeling the exam score variable.

2.2 Student's academic work variables

In this section, we focus on the real work provides by the student.

2.2.1 Weekly study hours



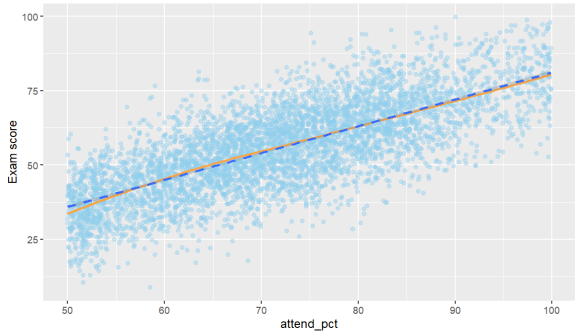
Exam score vs Study hours

Min	1st Qu.	Median	Mean	3rd Qu.	Max
2	6.3	8.2	8.187	10.1	14

From the figure, we observe that as the number of study hours increases, the exam score increases on average. Studying more is therefore associated with better performance. The LOESS curve (orange) closely follows the linear line (dashed) over almost the entire range, indicating an approximately linear relationship. However, a slight curvature remains at the extremes: at low values, the LOESS curve is slightly lower, and beyond 10-11 hours it tends to level off slightly. This suggests that the first hours of study yield substantial gains, whereas beyond a certain threshold, each additional hour results in smaller marginal improvements.

The points show considerable vertical dispersion: for a given number of study hours, exam scores vary widely. Thus, study time alone does not fully explain performance. Finally, the dispersion appears to increase slightly with the number of study hours, so the assumption of homoskedasticity should be examined carefully.

2.2.2 School attendance and School attendance categories

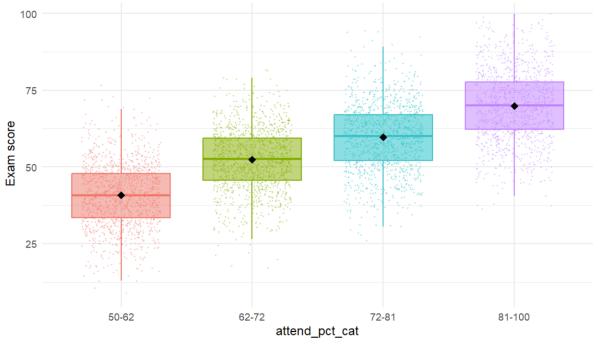


Exam score vs Attendance

Min	1st Qu.	Median	Mean	3rd Qu.	Max
50	62.20	71.70	71.77	80.80	99.90

We observe that as the attendance rate increases, exam scores increase markedly. Moreover, the LOESS curve and the linear regression line almost overlap, suggesting a largely linear relationship. Vertical dispersion remains substantial, and the variance appears to increase slightly with the attendance rate, indicating possible heteroskedasticity (very light).

Thus, attendance rate is more than a simple behavioral measure: it also reflects the student's engagement, exposure to instructional content, and consistency in study habits, which helps explain the increase in exam performance.



Exam score vs Attendance cat

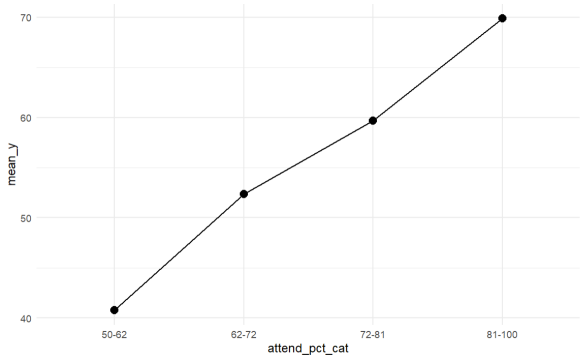
	Category	Count	Proportion (%)
1	50–62	1264	25.3
2	62–72	1241	24.8
3	72–81	1250	25.0
4	81–100	1245	24.9

We observe that each category corresponds to a higher performance level. Both medians and means increase sharply from one category to the next. The difference between the extreme categories is very large, on the order of 25 to 30 points. The 50-62 and 81-100 groups overlap slightly.

Within each group, there is substantial variability; however, the dispersion is similar across groups, and the differences between groups are much larger than the within-group variance. This indicates that the groups have been appropriately defined.

Therefore, if a more precise model is desired, the continuous attendance variable should be preferred, as it captures more detailed information. However, if the focus is on interpretability, the categorical variable is more than sufficient, since the groups are clearly differentiated.

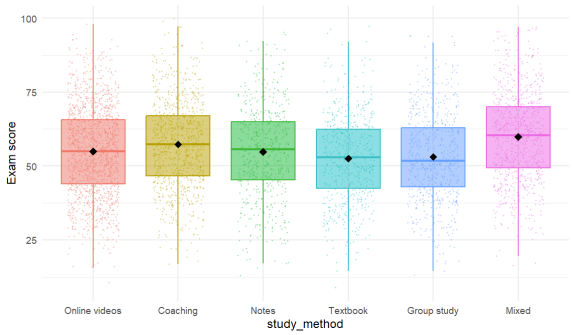
To underline monotone trends in y across ordered levels we can see the. The graph was created by grouping observations according to attendance rate categories and computing the mean exam score for each category. These means were then plotted as points connected by a line to visualize how average exam performance varies with attendance level.



Average Exam Score by Attendance Category

For other ordinal factor, the lector can see the same graphic in the text code.

2.2.3 Study method



Exam score vs Study methods

	Category	Count	Proportion (%)
1	Online videos	1498	30.0
2	Coaching	860	17.2
3	Notes	569	11.4
4	Textbook	594	11.9
5	Group study	664	13.3
6	Mixed	815	16.3

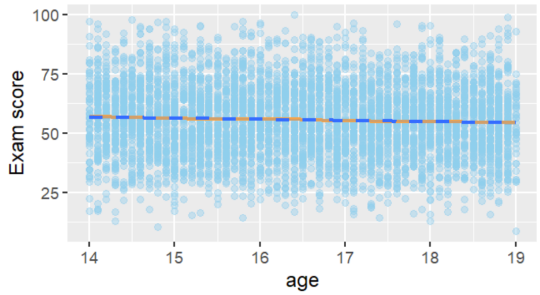
We observe that, on average, the highest scores are associated with a mixed learning method. Coaching and online videos are slightly above the overall mean, while the other methods are slightly below it. This suggests that the type of learning method has an effect on exam performance.

However, the boxplots overlap substantially, and for each method there are both very high and very low scores. The learning method therefore appears to be a complementary rather than a primary factor. Mixing learning resources may help better accommodate different students' learning styles and enhance retention. Similarly, group study can sometimes be more passive, which could explain slightly lower performance.

2.3 Social factors

In this section, we focus on social determination, sociological factors.

2.3.1 Age and Age categories



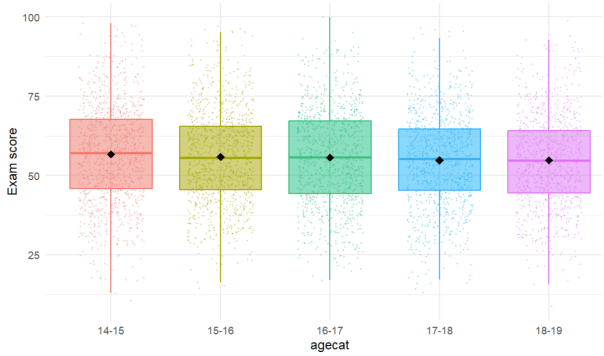
Exam score vs Age

Min	1st Qu.	Median	Mean	3rd Qu.	Max
14	15.3	16.5	16.5	17.8	19

The linear regression line is almost horizontal, and the LOESS curve is also flat. Therefore, there is no noticeable linear effect of age on exam scores. Moreover, for each age, scores display wide dispersion: within-age variability is much larger than between-age variability.

This suggests that the age range may be too narrow to observe meaningful differences in cognitive abilities or structural educational differences. Regardless of age, academic level appears to be the main determinant of performance.

Thus, it is appropriate to focus on the categorical variable to assess whether any group stands out from the others.



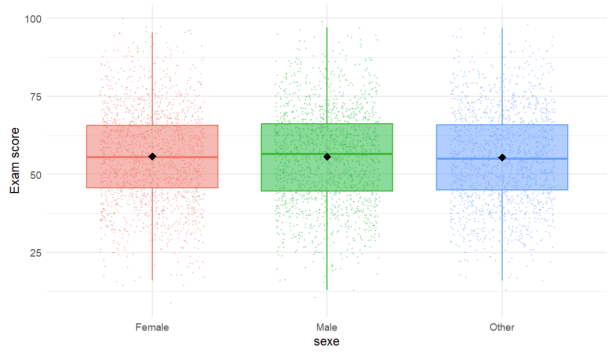
Exam score vs age cat

	Category	Count	Proportion (%)
1	14–15	1048	21.0
2	15–16	1027	20.5
3	16–17	983	19.7
4	17–18	983	19.7
5	18–19	959	19.2

The medians of exam scores are almost identical across all age categories. The means are also very similar, around 55-58, so no age category clearly stands out. The interquartile range is comparable across groups, and extreme values are of similar magnitude. Thus, within-group variability is both high and homogeneous.

We can therefore conclude that age, even when categorized, does not explain exam performance. Although it is not a key variable, it may still capture certain structural differences when combined with other variables.

2.3.2 Gender(Sexe)



Exam score vs Gender

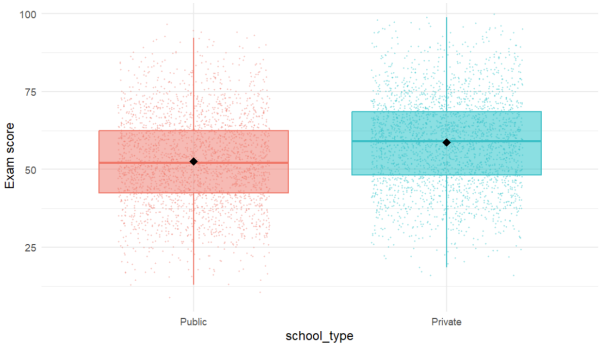
	Category	Count	Proportion (%)
1	Female	1730	34.6
2	Male	1550	31.0
3	Other	1720	34.4

We observe that the medians and means are almost identical across groups, as are the interquar-

tile ranges and whiskers. Therefore, gender does not appear to explain differences in exam scores. Variability is substantial within each group.

Thus, differences in exam performance do not seem to be attributable to gender.

2.3.3 School type



Exam score vs School type

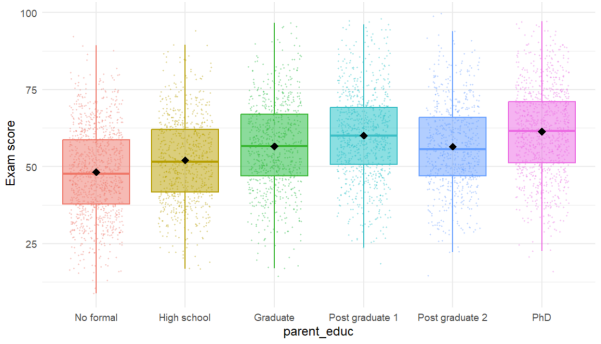
	Category	Count	Proportion (%)
1	Public	2514	50.3
2	Private	2486	49.7

Private schools exhibit higher median and mean exam scores than public schools, with a difference of about 6 to 8 points. The score distribution for private schools is generally shifted upward, meaning that high scores are more frequent in private institutions.

However, there is substantial dispersion in scores within each school type. Thus, school type is not deterministic of exam performance. Instead, it likely reflects several underlying factors: private schools often select students more strongly, tend to have greater and better-utilized educational resources, and students in private schools often benefit from a more structured and closely supervised learning environment, sometimes with less autonomy.

Therefore, there is an association between school type and exam performance, but not necessarily a causal relationship.

2.3.4 Parent education



Exam score vs parent education

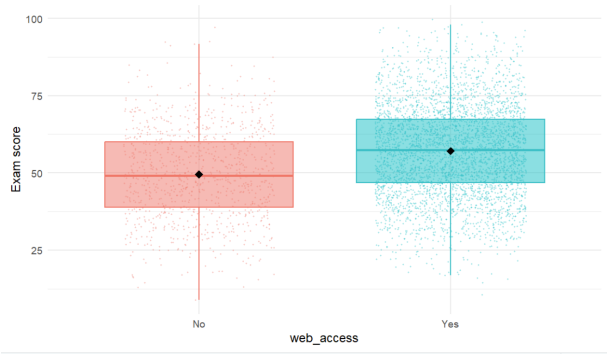
	Category	Count	Proportion (%)
1	No formal	1071	21.4
2	High school	736	14.7
3	Graduate	736	14.7
4	Post graduate 1	823	16.5
5	Post graduate 2	675	13.5
6	PhD	959	19.2

We observe a steady increase in exam scores as parental education level rises. Both medians and means increase at each level, and the gap between No formal education and PhD is on the order of 15 to 20 points. This effect appears stronger than that of school type or study method.

The distributions are progressively shifted upward, meaning that high scores are more frequent at higher levels of parental education. Parental education thus reflects cultural capital, the availability of academic support, academic expectations, and a supportive educational environment, all of which influence students' motivation, access to resources, and study habits.

Once again, this represents an association rather than a causal relationship. Parental education emerges as one of the strongest social predictors of academic performance. Moreover, the clear ordering of the categories is highly informative, making this variable important to retain in the analysis.

2.3.5 Web access



Exam score vs Web access

	Category	Count	Proportion (%)
1	No	986	19.7
2	Yes	4014	80.3

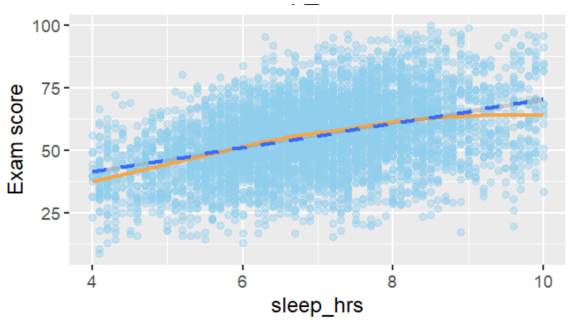
Students with access to the internet have higher median and mean exam scores, with a difference of about 8 to 10 points. The distribution for the Yes group is generally shifted upward, making high scores more frequent among students with internet access.

The dispersion indicates that most of the highest scores are concentrated in the Yes group, while low scores are more frequent in the No group. Internet access therefore appears to be a discriminating factor. Indeed, it enables access to information, educational resources, and digital tools, and is often associated with socio-economic status.

2.4 Lifestyle-related variables

In this section, we focus on student behavior, and lifestyle.

2.4.1 Sleep hours



Exam score vs Sleep hours

Min	1st Qu.	Median	Mean	3rd Qu.	Max
4	6.1	6.9	6.939	7.8	10

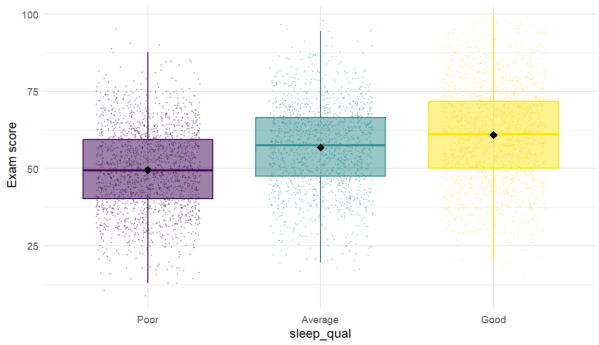
We observe that as the number of sleep hours increases, exam scores increase on average. The slope is positive but less pronounced than for study hours. The linear regression line and the LOESS curve are close in the central range, between 5 and 8 hours of sleep, but they diverge at the extremes.

Between 4 hours and 7-8 hours of sleep, exam scores increase, while beyond 8 hours the LOESS curve flattens or even stagnates. This indicates diminishing returns to sleep: sleeping too little is harmful, sleeping a "sufficient" amount is beneficial, but sleeping more provides almost no additional gains. The optimal range therefore appears to be around 7-8 hours of sleep.

This supports the idea that fatigue leads to a decline in cognitive performance. The effect of sleep is thus non-linear.

To account for this non-linearity, several approaches can be considered: including a quadratic term to capture performance losses at low or high sleep durations, categorizing the variable to improve interpretability, or interacting sleep with study hours. Indeed, sleep not only enhances learning itself but also the efficiency of studying. One could even test whether studying a lot without sufficient sleep is counterproductive.

2.4.2 Sleep quality



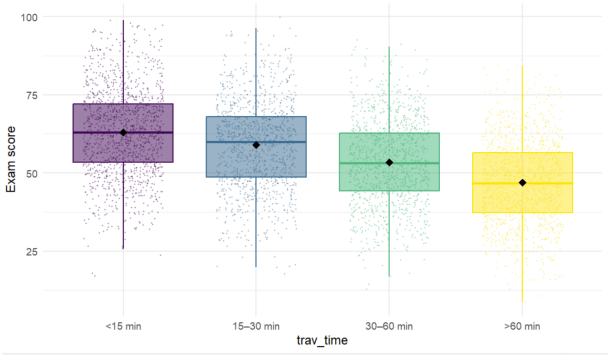
Exam score vs Sleep quality

	Category	Count	Proportion (%)
1	Poor	1811	36.2
2	Average	1442	28.8
3	Good	1747	34.9

We observe that exam scores increase as sleep quality improves. Medians and means are clearly differentiated, with a gap of about 12 to 15 points between the Poor and Good categories. The center of the distributions is noticeably shifted upward, and high scores are more frequent when sleep quality is good.

Thus, sleep quality reflects regularity and restorative sleep, which are conducive to better concentration and more effective cognitive processing. It would be interesting to relate this variable to sleep duration: indeed, two individuals sleeping 7 hours may exhibit different performance levels depending on the quality of their sleep.

2.4.3 Travel time



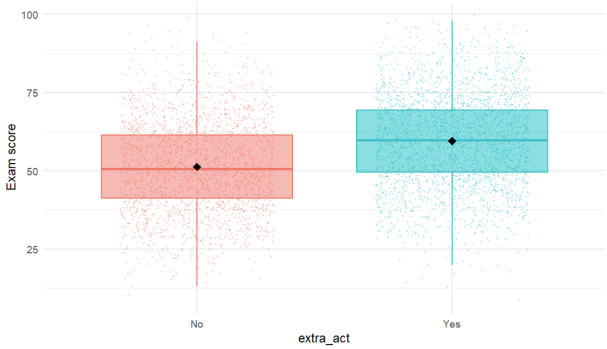
Exam Score vs Travel time

	Category	Count	Proportion (%)
1	$<$ 15 min	1376	27.5
2	15–30 min	1131	22.6
3	30–60 min	1185	23.7
4	$>$ 60 min	1308	26.2

We observe a monotonic decrease in exam scores as commuting time increases. Means and medians are clearly differentiated, with a difference of about 14 to 16 points between the extreme categories. The distributions are progressively shifted downward, and longer commutes are associated with a higher concentration of low scores.

Thus, longer commuting times increase fatigue and stress, reduce the time available for studying, and may also disrupt sleep, both in terms of duration and quality.

2.4.4 Extracurricular activities



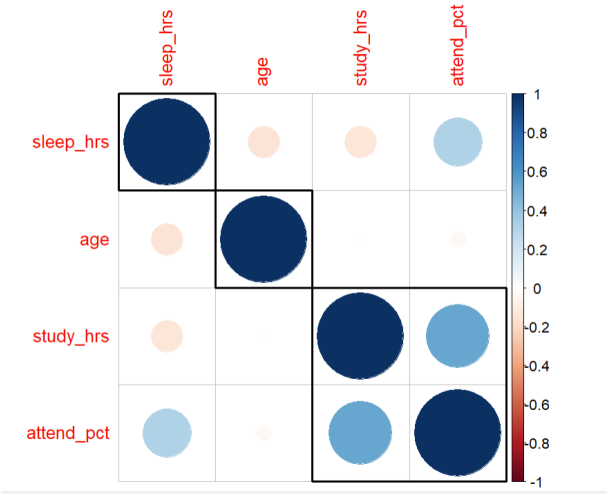
Exam score vs Extracurricular Act

	Category	Count	Proportion (%)
1	No	2370	47.4
2	Yes	2630	52.6

Students who engage in activities outside of class have higher median and mean exam scores, with an increase of about 8 to 10 points. High scores are more frequent in the Yes group, while low scores are more frequent in the No group.

Thus, participating in extracurricular activities may promote better time management, increased motivation, improved psychological well-being, and social interactions that contribute to overall balance and academic success. Once again, this reflects an association rather than a causal relationship: more motivated students are also more likely to engage in such activities.

2.5 Correlation matrix



Correlation Matrix

Between attendance and study hours, the correlation is strong and positive: students with higher attendance tend to study more. Therefore, potential multicollinearity should be considered in the model.

Attendance and sleep hours show a moderate positive correlation, suggesting that more regular attendance is associated with slightly better or more regular sleep. Sleep hours and study hours display a weak negative correlation, indicating that more study time is associated with slightly less sleep, although the effect is small.

Age and the other variables exhibit nearly zero correlations. There are no extreme correlations or signs of critical multicollinearity at this stage.

2.6 Relevant Interactions

Following the variable by variable analysis, we turn to the examination of potentially plausible interactions and discuss their theoretical relevance to the study. These interactions will subsequently be tested in the linear regression analysis.

To illustrate the relevance of some interaction for ordinal variable, we present some graphical representation modeling the outcome variable as a function of the two explanatory variables (Y, Z). When two variables do not interact, the effect of X on Y remains the same regardless of the level of Z. Graphically, this implies that the curves have the same slope and are therefore parallel.

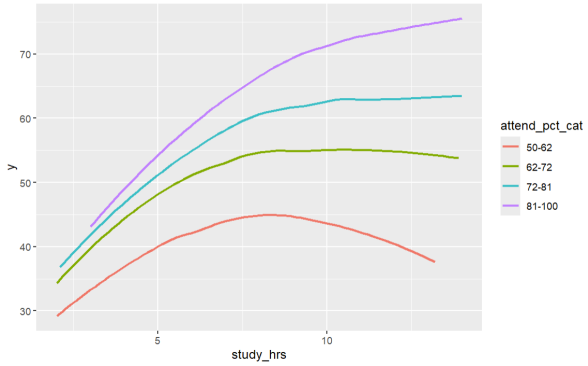
In contrast, when an interaction effect is present, the effect of X on Y depends on the level of Z. Graphically, this is reflected by differences in slopes: the curves are no longer parallel and may even intersect.

If the variable is nominal we examine the differences between groups.

2.6.1 Attendance rate categories X Study hours

We hypothesize that studying for long hours without regular attendance is less effective, as attendance enhances the productivity of study time. In this regard, attendance may moderate

the relationship between study time and academic performance, such that the positive effect of study time is stronger for students who attend classes regularly.

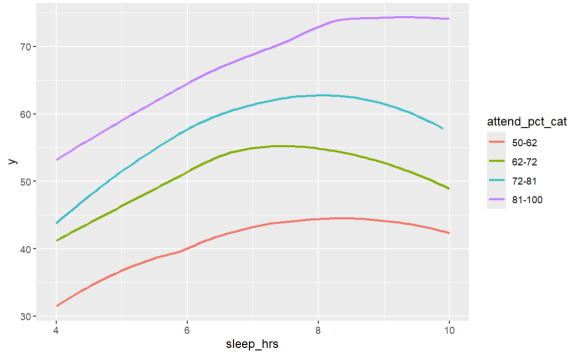


Exam score vs study hours stratified by attendance

The curves are not parallel, indicating the presence of an interaction effect. For low levels of attendance (50-62), an optimum is observed around 7-8 hours of study, beyond which the score begins to decline. In contrast, for high levels of attendance (81-100), the score continues to increase, with no clear optimum within the observed range.

These patterns indicate that the effect of study hours clearly depends on the level of attendance. The graph suggests that increased study time is particularly beneficial when attendance is high, whereas for students with low attendance, excessive study hours may be less effective or even counterproductive. Attendance therefore appears to condition the effectiveness of study time.

2.6.2 Attendance rate X Sleep hours



Exam score vs sleep hours stratified by attendance

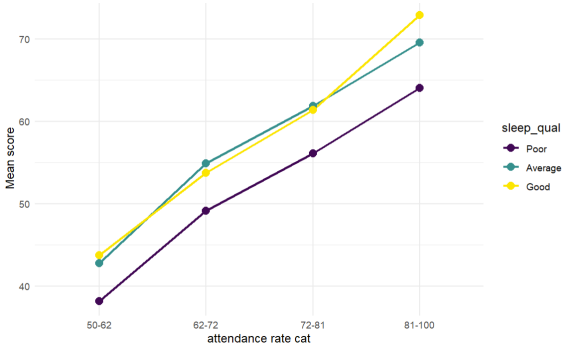
The curves are not parallel, which is the key result. For low attendance levels (50-62), the performance gains associated with increased sleep are limited. In contrast, for high attendance levels (81-100), the positive effect of sleep is much more pronounced. This indicates that the effect of sleep duration depends on the level of attendance, and that the optimal amount of sleep varies accordingly.

For students with low attendance, performance reaches a maximum around 7.5-8 hours of sleep and quickly plateaus. For students with high attendance, the plateau occurs at a higher level, with benefits remaining visible up to approximately 8.5-9 hours of sleep. Overall, the graph suggests that good sleep amplifies the effect of strong attendance. Sleeping better alone is

not sufficient when attendance is low; rather, attendance makes sleep more productive. Sleep therefore appears as a facilitating factor rather than an autonomous driver of performance.

In sum, organization and recovery jointly contribute to academic performance.

2.6.3 Attendance rate X Sleep quality



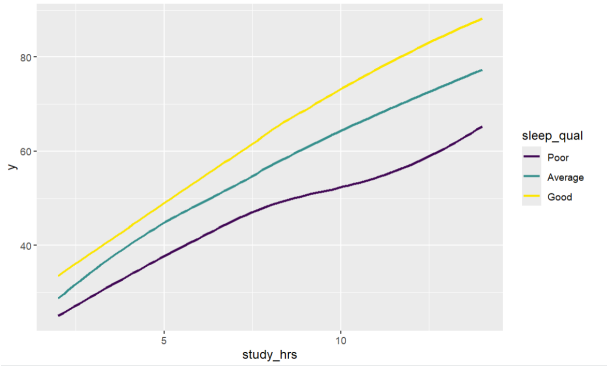
Exam score vs attendance rate cat stratified by sleep quality

The gap between the lines widens as attendance increases. At low attendance levels (50-62%), the difference in performance between poor and good sleep quality is approximately six points, whereas at very high attendance levels (81-100%), this gap increases to about nine points. The lines are therefore not perfectly parallel, with a steeper slope for good sleep quality and a flatter slope for poor sleep quality, which constitutes a clear signature of a positive interaction. Overall, the graph conveys a coherent pattern: while attendance improves performance for all students, it is substantially more rewarding when sleep quality is high.

Conversely, even strong attendance does not fully compensate for poor sleep quality. In sum, organization and recovery jointly drive maximal academic performance.

2.6.4 Study hours X Sleep quality

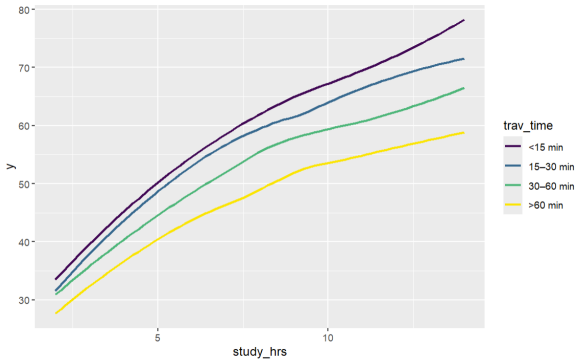
We hypothesize that studying for long hours while experiencing poor sleep quality may be counterproductive, as sleep quality moderates the effectiveness of study time. In this context, the positive effect of study hours on performance is expected to be weaker when sleep quality is low and stronger when sleep quality is high.



Exam score vs study hours stratified by sleep quality

The curves are nearly parallel, with the gaps between poor, average, and good sleep quality remaining approximately constant. The slopes are very similar across groups. These graphical patterns suggest a minim interaction effect. This indicates that increased study time improves performance regardless of sleep quality, while better sleep quality raises performance in an additive manner without altering the effect of study time.

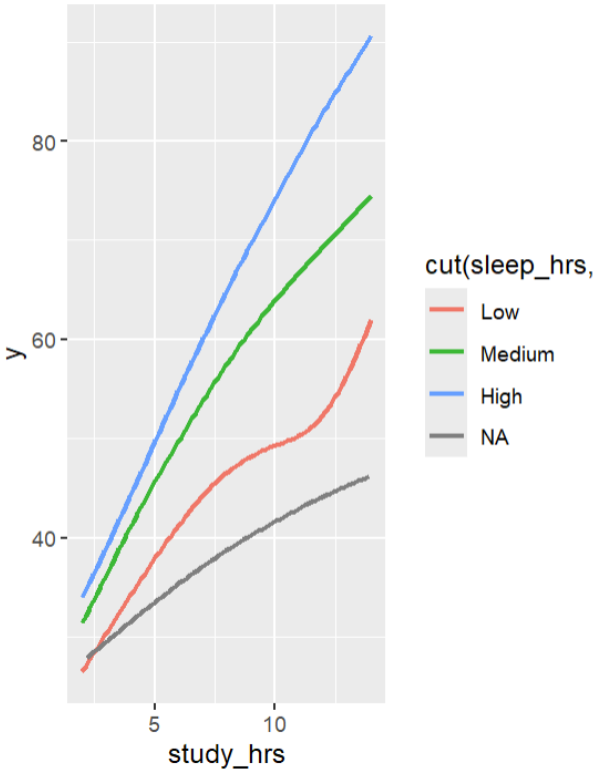
2.6.5 Study hours X Travel time



Exam score vs study hours stratified by travel time

The curves are not parallel, which is the key result. The slope is steeper for short commutes and flatter for long commutes, indicating that the effectiveness of study time depends on commute duration. In practical terms, an additional two hours of study yield substantial gains for students with commutes shorter than 15 minutes, but much more limited gains for those with commutes exceeding 60 minutes. Overall, the graph conveys an intuitive pattern: while studying more benefits all students, fatigue, time loss, and stress associated with long commutes reduce the marginal effectiveness of academic effort.

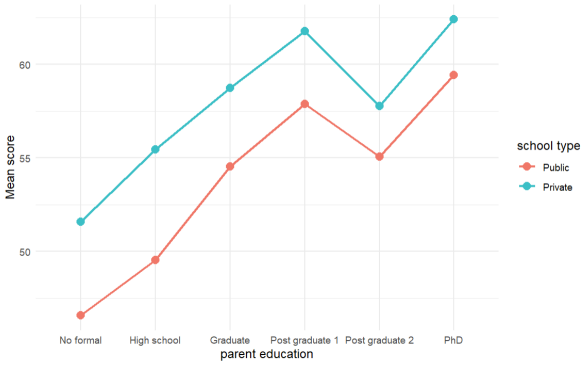
2.6.6 Study hours X Sleep hours



y vs study hours stratified by sleep hours cut

The curves are not parallel, which is the key finding. The slope is steeper for high sleep quality and flatter for low sleep quality, and the gap between the curves widens as study hours increase. This indicates that the marginal effect of study time depends on sleep quality. Studying more is highly productive when sleep quality is good but much less effective when sleep quality is poor. Overall, the graph shows that sleep is not merely an additive factor; it conditions the productivity of study time. In other words, studying while fatigued yields low returns, whereas studying while well rested leads to substantially higher returns.

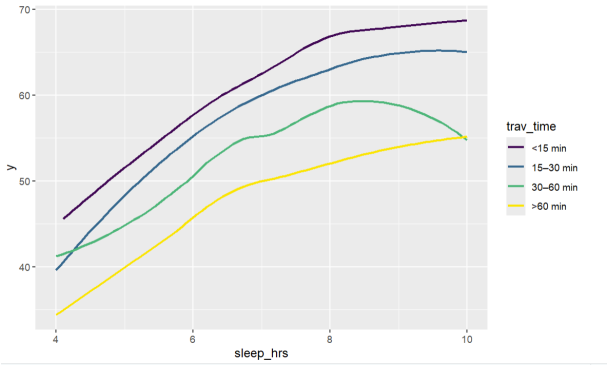
2.6.7 Parental education X School type



mean score vs parent education stratified by school type

The two curves display very similar shapes, and the gap between public and private schools remains nearly constant. This suggests a weak interaction in a strict statistical sense.

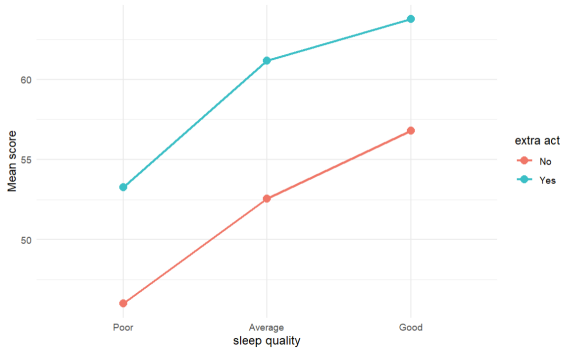
2.6.8 Travel time X Sleep hours



y vs sleep hours stratified by travel time

The curves are not parallel and exhibit distinct shapes depending on commute time. For short commutes (less than 15 minutes), performance increases sharply with sleep duration and quickly reaches a plateau, suggesting that adequate sleep is particularly beneficial when commute time is short. For intermediate commutes (30-60 minutes), an optimal sleep duration appears around 8-9 hours, followed by a slight decline. In contrast, for long commutes (over 60 minutes), the effect of sleep is weaker and nearly linear. Overall, these patterns indicate that the effect of sleep duration on academic performance depends on commute time, consistent with the presence of an interaction effect.

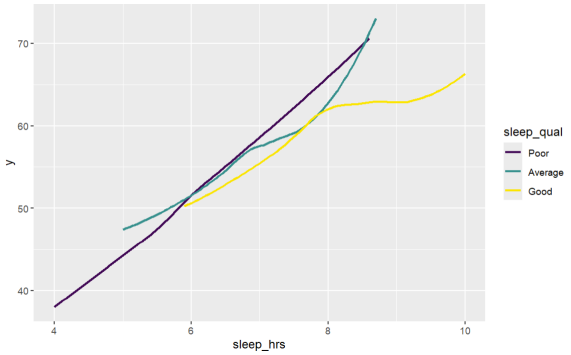
2.6.9 Extra act X Sleep quality



Y mean vs sleep quality stratified by extra act

The two lines are nearly parallel, indicating that the increase in performance associated with better sleep quality is similar for students with and without extracurricular activities. The gap between the two groups varies only slightly across sleep quality levels. These graphical patterns do not suggest a strong interaction effect. This implies that improved sleep quality enhances academic performance regardless of participation in extracurricular activities, while extracurricular involvement provides an additive benefit without altering the effect of sleep.

2.6.10 Sleep quality X sleep hours



Y vs sleep hours stratified by sleep quality

The curves are neither strictly parallel nor strictly ordered. At low levels of sleep duration (approximately 4-5 hours), the differences across sleep quality categories are pronounced. Around 6-7 hours, the curves converge and may even intersect, while at higher levels of sleep (8 hours or more), they diverge again, with a distinct curvature for the Good sleep quality category. These patterns indicate that the effect of sleep duration depends on sleep quality, both in terms of slope and functional form. Overall, the graph suggests that sleeping more is beneficial for all students; however, the optimal amount of sleep and the marginal gain associated with an additional hour vary by sleep quality. In particular, good sleep quality appears to reduce marginal returns at high sleep levels, whereas poor sleep quality increases the importance of obtaining additional sleep.

2.6.11 Attendance X Study hours X Sleep hours or quality

The effectiveness of study time depends on class attendance, and this effectiveness is itself conditioned by sleep hours/quality. Studying for long hours without regular attendance tends to be inefficient, just as excessive studying under poor sleep conditions may lead to fatigue and reduced effectiveness. In contrast, high study intensity combined with strong attendance and good sleep quality is expected to produce the maximal effect on academic performance.

2.6.12 Travel Time X Sleep Quality X Study hours

A long commute particularly penalizes students who experience poor sleep quality and study for long hours, reflecting the effects of cumulative fatigue. However, including three-way interactions is generally not advisable, even with a large sample size of 5,000 observations. Such specifications substantially increase the number of parameters, leading to a loss of parsimony and making the effects more difficult to interpret clearly and meaningfully.

2.7 Final Choices

After examining each variable in detail, we can summarize the results in the following table, which provides an overview of the relevance of the variables and the potentially interesting interactions to be tested.

Variables and choice for the future model

Variable	Decision
Age	Student's age in years; For reasons of interpretability, we choose the age category variable.
Agecat	Categorized version of age; It does not appear to provide substantial information regarding exam performance, and we therefore tend to remove it from the model (this will of course be formally verified using a Type I analysis test).
Sexe	Student's gender; same as age cat.
School type	seems to be included.
Parent educ	seems to be included.
Study hrs	seems to be included and add a ^{2}.
Sleep hrs	seems to be included and add a ^{2}.
Sleep qual	seems to be included.
Attend pct	Class attendance rate; For reasons of interpretability, we choose the attend pct category variable.
Attend pct cat	seems to be included.
Web access	seems to be included.
Trav time	seems to be included.
Extra act	seems to be included.
Study method	seems to be included.

So we can remove from our data Age and Attend pct to avoid any confusion and replication of a variable.

3 Regression Models

The data were randomly split into training (80%) and testing (20%) sets using a fixed random seed for reproducibility.

Based on the analyses conducted during the exploratory data analysis (EDA), there does not appear to be any measurement error or data transcription error. However, several outliers are present. These may reflect adversarial errors intended to distort the analysis, or they may represent meaningful observations that reveal specific phenomena not captured by the model followed by the majority of the data.

To find the best model we base are selection on some criterion as AIC, BIC, nested model test, RMSE and C_p of mallow's.

Reminder AIC and BIC :

The residual sum of squares associated with model m is defined as

$$\mathrm{RSS}(m) = \sum_{i=1}^{n} \left(y_i - \hat{y}_i^{(m)}\right)^2,$$

where y_i denotes the observed value and $\hat{y}_i^{(m)}$ the value predicted by model m.

$$\mathrm{AIC}(m) = n \log\Big(\tfrac{\mathrm{RSS}(m)}{n}\Big) + 2m$$

The AIC criterion is used to identify a good predictive model, as it applies a relatively mild penalty for model complexity. Smaller is better, the difference between the smaller and other AIC is ΔAIC if $\Delta AIC \in [0; 2]$ models are equivalent, if $\Delta AIC \in [2; 6]$ models have moderate difference, if $\Delta AIC > 10$ models tested is really not good.

$$\mathrm{BIC}(m) = n \log\Big(\tfrac{\mathrm{RSS}(m)}{n}\Big) + \log(n)\, m$$

The BIC criterion is more focused on identifying the true underlying model and the key determinants of exam performance. It applies a stronger penalty for model complexity.

Reminder Test for nested model : Let $\mathcal{M}_0$ be a reduced model and $\mathcal{M}_1$ a full model such that $\mathcal{M}_0 \subset \mathcal{M}_1$. Let p_0 and p_1 be the numbers of parameters (including the intercept), with $p_1 > p_0$. Let n be the sample size.

classical ANOVA F-test For the normal linear model $Y \sim \mathcal{N}(X\beta, \sigma^2 I)$, nested-model comparison

is often expressed via sums of squares:

$$F = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(n - p_1)} \quad \sim \quad F_{p_1 - p_0,\, n - p_1} \quad \text{under } H_0.$$

Here $RSS_m = \sum_{i=1}^{n} (y_i - \hat{y}_{i,m})^2$.

Reminder RMSE : A common RMS measure of fit is the root mean squared error:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

In linear regression, the residual standard error (an estimator of σ) is

$$\widehat{\sigma} = \sqrt{\frac{RSS}{n-p}},$$

and is also sometimes referred to informally as an RMS residual measure. Smaller RMS/RMSE indicates better in-sample fit, but it does not by itself penalize model complexity.

Reminder C_p : Mallows' C_p compares a candidate model with p parameters to an estimate of the noise level (often taken from the full model with p_{full} parameters):

$$C_p = \frac{RSS_p}{\widehat{\sigma}^2} - (n - 2p),$$

where $\widehat{\sigma}^2$ is typically estimated from the largest (or "full") model:

$$\widehat{\sigma}^2 = \frac{RSS_{\text{full}}}{n - p_{\text{full}}}.$$

Rule of thumb: prefer models with small C_p, and especially those with $C_p \approx p$, which suggests low bias without excessive variance.

For normal linear regression, minimizing C_p is closely related to minimizing AIC: both trade off fit (via RSS) and complexity (via p), differing mainly by constants and scaling.

3.1 First Variable Selection

To find the most relevant predictors, we apply stepwise selection using AIC and BIC (Akaike Information Criterion). This process helps to identify a simpler model that still explains the data well.

Using the AIC criterion, we obtain a model that includes all variables except age and gender, which is consistent with the findings from the exploratory data analysis (EDA).

According to the BIC criterion, we obtain an almost identical model; however, due to its more restrictive nature, parental education is not retained. But do not eject the variable because we want to explore the interaction parental education and school type.

We also conducted analyses of variance using both Type I and Type II approaches. The Type II analysis yields the same model as the one selected by the AIC criterion. In contrast, under the Type I approach, age appears to be statistically significant (because it's the first variable). We therefore performed a nested model test, which led us to conclude that the model selected by the AIC criterion remains preferable at this stage.

Test for nested model result

Model	p-value
modboth/modanova	0.6251931
modboth/modfull	0.8458208

Model comparison using AIC

Model	Degrees of freedom	AIC
modboth	25	28326.94
modanova	26	28328.70
modfull	28	28332.12

Since the objective of this report is a balance between prediction of exam scores and identification of the most important determinants of performance, we choose to retain the variables selected by the AIC criterion. Consequently, the variables age and gender are excluded from the model (as we predicted in the EDA).

3.2 Test for $sleep - hour^2$ and $study - hour^2$

As observed during exploratory data analysis (EDA), the relationships between sleep hours and exam performance and study hours and exam performance do not appear to be linear. To avoid any heteroskedasticity problem on the residuals, We therefore include a squared term for variables in order to capture this curvature, as the optimal level of sleep/study lies in the middle of the observed range. However, including both sleep hours/study hours and its squared term may induce strong multicollinearity. To address this issue, sleep hours/study hours are centered prior to squaring.

Model	$\boldsymbol{R^2_{\text{adj}}}$	$\boldsymbol{C_p}$	**AIC**
mod1 (with $sleep^2$)	0.71098	24.75	28125.44
mod1bis (with $sleep$)	0.70209	145.98	28245.63
mod_1 (with $study^2$)	0.71098	24.75	28125.44
mod_1bis (with $study$)	0.70464	110.95	28211.27
modfull	–	–	28128.68

model comparison with R^2_{adj}, C_p and AIC

Comparison of nested models (ANOVA tests)

Model comparison	p-value
mod1 / mod1bis	3.17×10^{-28}
mod_1 / mod_1bis	9.59×10^{-21}

Once the models are estimated, we assess whether the inclusion of the squared sleep term and the study squares term is statistically justified.Among the selection criteria considered, Mallows' Cp support keeping this term, despite of its particularly strong penalization of model complexity. In addition, the p-value from the nested model test is highly significant. Consequently, we retain the squared sleep-hours and study hours terms in the final model.

3.3 Model With Interactions, interaction selection with anova test

We will test the interactions selected during the EDA. We proceed step by step, estimating models incrementally, group by group. We create a few models that can be tested by Nested fisher test. We begin by adding a single interaction term and comparing them based on which yields the lowest p-value. We then add interactions sequentially in ascending order of p-values, while simultaneously performing nested model tests to assess whether each additional interaction

significantly improves the model. As we can see the interaction Parental education x School type is always possible and we start put it in every models (we can check at the end if it's relevant or not).

We start with the interactions involving study hours. The theoretically relevant interactions that can be jointly included in a model are:

- Attendance rate categories X Study hours
- Study hours X Sleep quality
- Study hours X Travel time
- Study hours X Sleep hours
- Parental education X School type

For the best model involving Study hours's interaction we select 2 models :

ModI : With the interaction study hours X sleep hours and parent education X school type and all the other variables.

ModII : With the interaction Study hours X sleep quality, study hours X attendance and parent education X school type and all the other variables.

We repeat the operation for the interactions involving Sleep Quality. The theoretically relevant interactions that can be jointly included in a model are:

- Attendance rate X Sleep quality
- Study hours X Sleep quality
- Extra act X Sleep quality
- Sleep quality X sleep hours
- Parental education X School type

For the best model involving sleep quality's interaction we select model : ModIII : with the

interaction sleep quality X study hours, sleep quality X attendance and parent education X school type.And all the other variables.

The operation for the interactions involving Sleep hours. The theoretically relevant interactions that can be jointly included in a model are:

- Attendance rate X Sleep hours
- Study hours X Sleep hours
- Sleep quality X sleep hours
- Parental education X School type

For the best model involving sleep hours's interaction we select model :

ModIV : with the interaction sleep hours X attendance, sleep hours X study hours, school type X parent education. And all the other variables.

We repeat the operation for the interactions involving Attendance pct cat. The theoretically relevant interactions that can be jointly included in a model are:

- Attendance rate categories X Study hours
- Attendance rate X Sleep hours
- Attendance rate X Sleep quality
- Parental education X School type

For the best model involving attendance's interaction we select the model :

ModV : with attendance X sleep hours, Sleep quality X study hours and Parent education X school type. And all the other variables.

The operation for the interactions involving Travel time. The theoretically relevant interactions that can be jointly included in a model are:

- Study hours X Travel time
- Travel time X Sleep hours
- Extra act X Sleep quality
- Attendance rate X Sleep quality
- Parental education X School type

For the best model involving travel time's interaction we select model :

Mod VI : with travel time X study hours,Travel time X sleep hours, sleep quality X attendance, Parent education X school type, and all the other variables.

Then we can think of the model with 3 interactions because theoretically is really interesting.

ModVII : with the triple interaction Attendance X Sleep hours X study hours

Furthermore with all the models tested we can denoted wich interactions add the most from the AIC model. and we can add some models that we can compare with other criterion because nested model is not efficient. We can limit our number of interactions at 4 to avoid over-adjustment.

ModVIII : With the interaction Study hours X sleep hours, sleep hours X attendance, study hours X attendance, parent education X school type

3.4 Final Selection with diverse criterion

Then we have at least 12 models for further exploration. The tables show all the criterion by interesting order use to chose the final model.

Model	AIC
modVII	27780.45
modVIII	27783.70
modX	27787.69
modXI	27794.59
modIX	27797.63
modIV	27809.83
modI	27814.08
modXII	27816.68
modII	27844.18
modIII	27882.08
modV	27886.84
modVI	28065.94

Model	BIC
modI	28040.66
modVIII	28048.05
modXII	28068.44
modIV	28074.18
modX	28077.21
modXI	28077.82
modVII	28082.56
modIX	28087.16
modII	28108.53
modIII	28146.43
modV	28151.19
modVI	28380.64

Model	R^2_{adj}
modVII	0.7362
modVIII	0.7356
modX	0.7356
modXI	0.7351
modIX	0.7350
modIV	0.7339
modXII	0.7333
modI	0.7332
modII	0.7316
modIII	0.7291
modV	0.7287
modVI	0.7168

Model	C_p	p
modVII	40.58	47
modVIII	43.65	41
modX	47.68	45
modXI	54.50	44
modIX	57.54	45
modIV	69.62	41
modI	73.85	35
modXII	76.45	39
modII	104.01	41
modIII	142.30	41
modV	147.13	41
modVI	332.31	49

Across the 12 candidate linear models, the model selection criteria convey a clear and consistent message: predictive performance reaches a plateau quickly, while model complexity continues to increase. Using AIC, which prioritizes goodness of fit with a moderate penalty for complexity, modVII is the best model. However, the gain over the next best alternative is modest: modVIII is only $\Delta AIC = 3.25$higher, whereas the remaining models are substantially worse (modXI $\Delta AIC \approx 7.2$; modX ΔAIC ≈ 14; modIX ΔAIC ≈ 17), implying limited support under AIC for those specifications.

In contrast, the BIC imposes a stronger penalty on complexity and therefore tends to favor more parsimonious models, particularly with large samples such as $n = 5000$. Under BIC, modI is selected as the best model, and the gap to more complex models is meaningful (for instance, modVIII is +7.39). Importantly, the adjusted coefficient of determination confirms that additional complexity produces only marginal improvements: the best value is $R^2_{\text{adj}} = 0.7362$ (modVII), but modVIII is essentially identical ($R^2_{\text{adj}} = 0.7356$), and even the simpler modI remains close ($R^2_{\text{adj}} = 0.7332$). This pattern indicates strong diminishing returns: sizable increases in the number of parameters translate into only about a 0.3% point improvement in adjusted explained variance.

Finally, Mallows' C_p supports modVIII as a balanced compromise because C_p is relatively close to the parameter count ($C_p = 43.65$ and $p = 41$), whereas more highly parameterized models do not clearly justify their added complexity.

Overall, the evidence suggests retaining only the models that are not dominated by others: modVIII as the strongest overall trade-off between fit and complexity (near best AIC and R^2_{adj} with favorable C_p), and modVII only if the primary objective is maximizing in sample fit despite substantially higher complexity.

Finally we keep the model VII and VIII for the RMSE. we have :

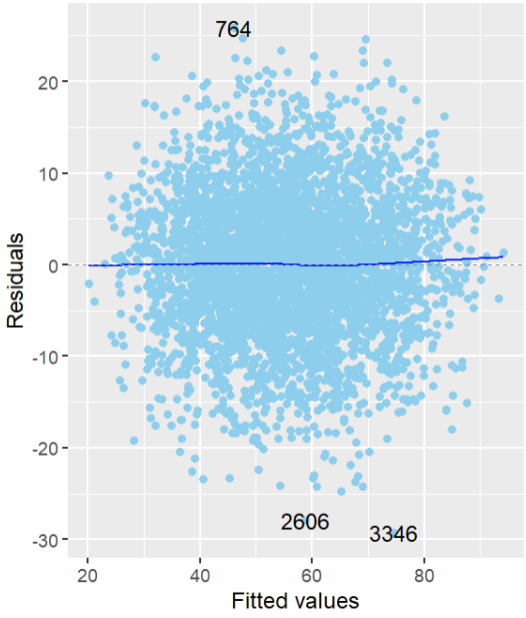
	modVII	modVIII
Train	7.703056	7.717755
Test	7.718212	7.705898

Based on the RMSE results, modVIII is the preferred model because it delivers the best out of sample predictive accuracy. On the test set, modVIII achieves the lowest RMSE (7.7059), improving upon modVII (7.7182). In addition, the train and test RMSE values are very close for all three models, suggesting no strong evidence of overfitting and good generalization overall.

Given that model selection should prioritize performance on unseen data, these results support choosing modVIII as the final model for prediction, while modI would only be favored if interpretability and parsimony were the dominant objectives.

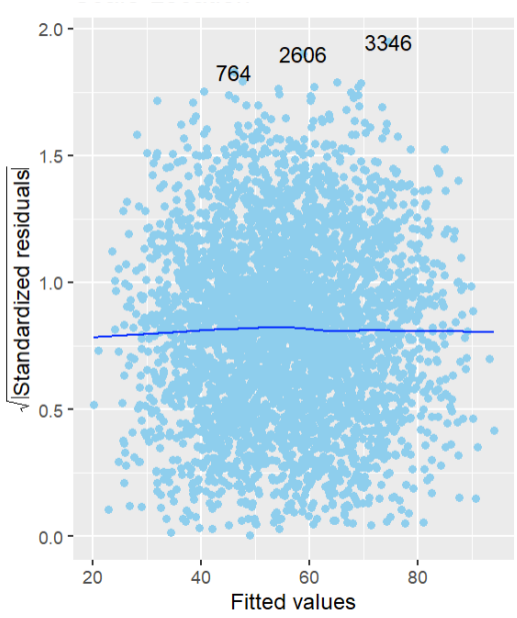
4 Diagnostics, assumptions, and robustness

4.1 Assumptions



Residuals vs Fitted

The residuals are centered around zero, and without any particular structure indicating that the conditional mean is generally well captured.

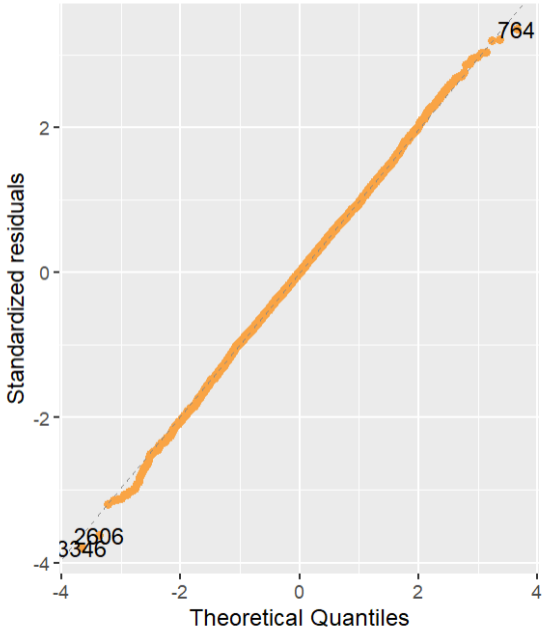


Scale-location

The residuals are homoskedastic due to scatter plot centered, aligned around 1, and without any structure. And the Breusch-Pagan test conclude the same.

The Durbin-Watson Test give the residuals uncorrelated ($pvalue = 0.13$).

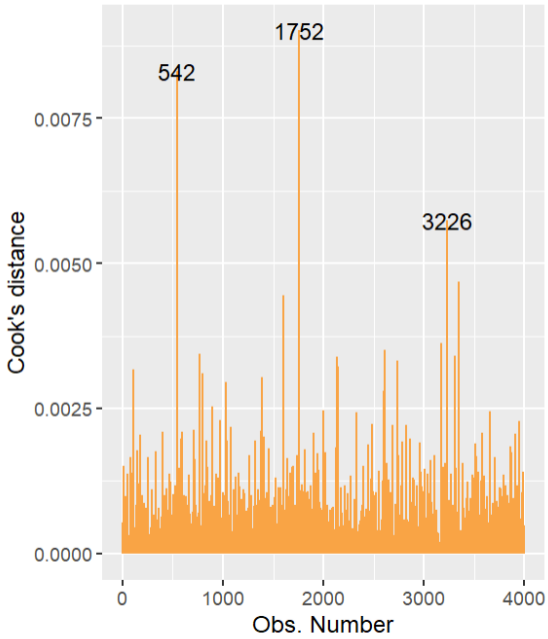
And finally, the points fall approximately along a straight line, then the data are approximately normally distributed.



Normal Q-Q plot

4.2 Outliers detection

Cook's distance quantifies how strongly each observation influences the fitted regression coefficients by combining the size of its residual and its leverage. In the plot, the vast majority of observations have Cook's distances very close to zero, which indicates that most data points have negligible impact on the model fit and the estimated coefficients are not being driven by many individual cases.



Cook's distance

With n around 4,000, the common screening cutoff $D_i > 4/n = 0.001$ flags a few observations (Cook's distance $= 0.006 - 0.01$) as worth checking. However, these values are still far below 1, so no single point is dominating the regression; overall the model appears broadly stable. Such influential cases can be driven by large residuals, high leverage, or both.

The recommended follow-up is to inspect those records for possible data issues, examine leverage and studentized residuals, and run a sensitivity analysis by refitting the model without them. If estimates change noticeably, conclusions may be sensitive and robust methods could be considered; if changes are small, the results are not overly dependent on a handful of observations.

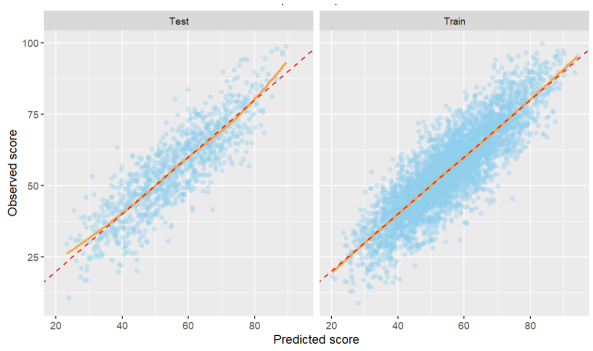
Cook's distance check indicates that the three most influential observations do not materially drive the conclusions of modVIII. After dropping just these three cases, the fitted model remains essentially unchanged: the coefficient estimates shift only slightly (often in the third decimal place), the pattern of statistical significance is nearly identical, and the overall fit changes only marginally. For example, the residual standard error decreases from $7.758 to 7.734$, while R^2 increases from 0.7383 to 0.7392 and adjusted R^2 from 0.7356 to0.7366. This small improvement is consistent with the idea that the removed points were somewhat influential (they slightly worsened fit when included), but the magnitude of the change is too small to suggest instability or serious leverage problems in a dataset of this size.

Overall, this sensitivity analysis supports the interpretation that, although a few points have higher Cook's distance, the model is not unduly sensitive to them.

In reporting, you can state that modVIII appears robust to influential case diagnostics, and that the main findings persist under the reduced-sample refit.

4.3 Efficiency of the model

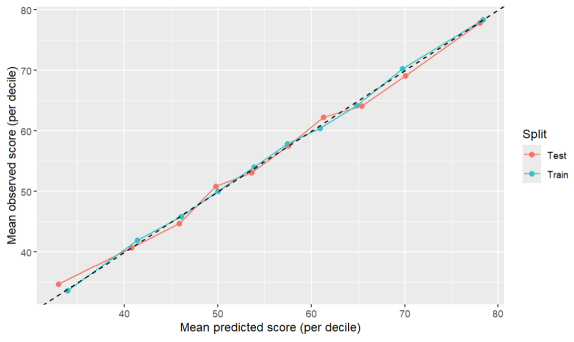
We performed a calibration check by plotting observed scores against the model's predicted scores on the train and test sets to assess whether predictions align with outcomes on average (whether points follow the 45-degree line).



Observed vs predicted

The calibration plot (observed vs predicted) indicates that modVIII is well calibrated on both the training and test sets. The LOESS curve closely tracks the 45-degree identity line across most of the prediction range, suggesting that predicted scores align well with observed outcomes on average. Train and test panels show very similar patterns, providing evidence of good generalization and limited overfitting. Minor departures appear at the extremes, consistent with mild regression to the mean behavior: the model slightly underpredicts the lowest outcomes and slightly overpredicts the highest outcomes, but these deviations remain small relative to the overall fit.

We also conducted a binned calibration analysis by grouping observations into deciles of predicted score and comparing the mean observed score to the mean predicted score within each decile, which provides a clearer view of calibration in the aggregate.



Binned calibration:Train vs Test

The decile binned calibration plot shows strong agreement between mean predicted and mean observed scores on both the training and test sets. Across nearly all deciles, the points lie close to the 45-degree identity line, indicating that the model is well calibrated in the aggregate: groups of students with similar predicted outcomes also have similar observed average outcomes. The train and test curves closely overlap, suggesting good generalization and limited overfitting.

Minor deviations appear mainly at the extremes (slight underprediction for the lowest decile and slight overprediction for the highest deciles), consistent with mild regression to the mean behavior, but these departures are small relative to the overall calibration accuracy.

We evaluate predictive performance on the held-out test set using four complementary metrics and compare them to a simple baseline. The baseline model predicts a constant value for every student, equal to the mean of y in the training set; it provides a minimal benchmark that any useful model should beat. We report Mean Squared Error (MSE), defined as

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

which penalizes large errors more heavily due to squaring. We also report Root Mean Squared Error (RMSE),

$$\text{RMSE} = \sqrt{\text{MSE}},$$

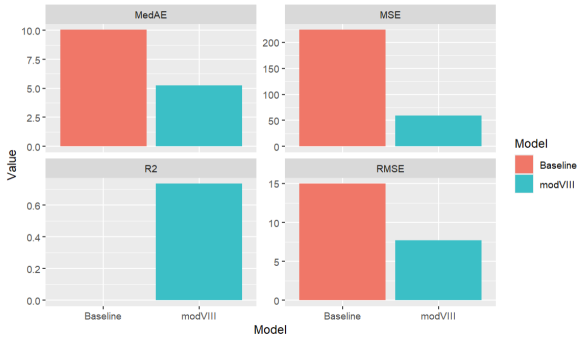
which is on the same scale as the outcome and is easier to interpret in outcome units. To summarize a robust notion of typical error, we use the Median Absolute Error (MedAE),

$$\text{MedAE} = \text{median}\left(|y_i - \hat{y}_i|\right),$$

which is less sensitive to outliers than MSE/RMSE. Finally, we compute out-of-sample R^2 on the test set,

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

which measures the fraction of variance explained on unseen data (higher is better, whereas lower values of MSE, RMSE, and MedAE indicate better predictive accuracy).



Predictive performance on Test: Baseline vs modVIII

Based on the Median Absolute Error (MedAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) Model VIII performs substantially better than the baseline model. It reduces the typical prediction error (MedAE), limits large errors (MSE), and lowers the average prediction error on the scale of y (RMSE).

In terms of out of sample R^2, where higher values indicate better performance, the baseline model yields a value close to zero, which is expected since predicting the mean explains virtually no variation in the test set. In contrast, Model VIII achieves a very high R^2, capturing a large share of the predictive signal.

Overall, this plot demonstrates that Model VIII provides a substantial improvement over a naïve baseline strategy (predicting the mean) across all evaluation metrics, with markedly lower prediction errors and significantly higher explanatory power on unseen data.

4.4 Stability of the model

We conducted a subsampling based stability analysis to evaluate how robust the final model modVIII is to changes in training sample size. The training data were randomly subsampled at five fractions $(20\%, 40\%, 60\%, 80\%, and 100\%)$, and for each fraction the procedure was repeated 30 times using sampling without replacement. In each repetition, modVIII was refit on the reduced training set, predictions were generated on the same held out test set, and predictive performance was recorded using MSE, RMSE, MedAE, and out of sample R^2.

The results were then summarized by the mean and standard deviation of each metric across repetitions for each fraction, providing a clear assessment of both average accuracy and the variability (stability) of performance as the amount of training data decreases.

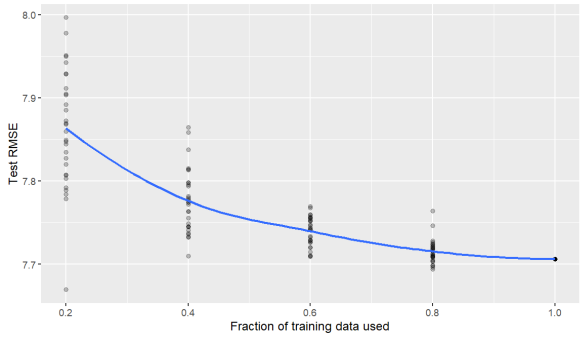
Subsampling stability results (30 repetitions per fraction). Values are mean ± standard deviation on the test set.

Training fraction	MSE	RMSE	MedAE	R^2
0.2	61.835 ± 1.107	7.863 ± 0.071	5.370 ± 0.103	0.724 ± 0.005
0.4	60.469 ± 0.589	7.776 ± 0.038	5.303 ± 0.099	0.731 ± 0.003
0.6	59.902 ± 0.273	7.740 ± 0.018	5.294 ± 0.075	0.733 ± 0.001
0.8	59.528 ± 0.222	7.715 ± 0.014	5.271 ± 0.036	0.735 ± 0.001
1.0	59.381 ± 0.000	7.706 ± 0.000	5.233 ± 0.000	0.735 ± 0.000

The subsampling results show that modVIII's predictive performance improves steadily as more training data are used, while the variability of that performance across random subsamples decreases markedly. When only 20% of the training set is used, test performance is weaker and less stable (mean $RMSE = 7.86$ with a larger standard deviation, and mean $R^2 = 0.724$), indicating that parameter estimates (especially in an interaction rich specification) are more sensitive to which observations are included.

As the training fraction increases to $40 - 60\%$, errors decline and dispersion shrinks, and by $80\tilde{}100\%$ the model becomes highly stable (mean $RMSE = 7.72 - 7.71$ and mean $R^2 = 0.735$ with very small standard deviations). The pattern also indicates diminishing returns: beyond roughly $60\tilde{}80\%$ of the training data, additional observations yield only modest improvements in RMSE/MSE and R^2.

Overall, this analysis supports that modVIII generalizes reliably when trained on a sufficiently large sample and that its test performance is not driven by a particular random subset of the data.



Stability check: Test RMSE vs training sample fraction

When only 20% of the training data are used, the test RMSE is higher and noticeably more dispersed, indicating greater instability. This reflects the fact that the model must estimate a relatively large number of coefficients with limited information.

As the training sample size increases (from 40%to 60%and then to 80%), the test RMSE decreases and its variability is substantially reduced. Beyond approximately $60 - 80\%$, the curve flattens and additional gains become marginal, indicating the presence of a genuine performance plateau.

The monotonic decline in test RMSE, together with the tightening dispersion, suggests that the model generalizes better when estimated on larger samples. Importantly, the model does not appear to be overfitted to a particular subsample; instead, its performance stabilizes as the training size increases, which is characteristic of a model capturing a genuine underlying signal. However, because the specification is relatively rich it requires a sufficiently large sample to be estimated reliably, explaining the higher variance observed at smaller training sizes.

Finally, the fact that even when using 100%of the training data the test RMSE remains around 7.71 suggests that the model's performance is robust and reproducible.

Using the training-mean baseline as a benchmark, the model achieves large and consistent improvements in test performance across all training fractions. The baseline RMSE on the test set is approximately 14.98, whereas refitting modVIII on only 20% of the training data yields a mean test RMSE of 7.86, corresponding to an absolute improvement of 7.12 RMSE points (a 47.51% reduction relative to baseline). As the training fraction increases, performance improves slightly and stabilizes: with 100%of the training data, the mean test RMSE is 7.71, an improvement of 7.27 points (a 48.56%reduction). Overall, these results show that modVIII substantially outperforms the naive baseline even when trained on smaller subsamples, with only modest additional gains as the training set approaches full size.

4.5 A example of two invented Student

We constructed a scenario-based interpretation of the final model by defining two realistic student profiles a lower performance (call lower support) profile and a higher performance (call higher support) profile and computing predicted mean scores with 95% confidence intervals.

For the continuous predictors expressed in centered units, the lower-support profile was set to below average levels ($study_c = -1$, $sleep_c = -1$) and included the corresponding quadratic terms ($study_c^2$ and $sleep_c^2$), while the higher support profile was set to above-average levels ($study_c = 1.5$, $sleep_c = 1$), again including the quadratic terms.

The categorical covariates were chosen to reflect contrasting environments: the lower-support profile assumes lower attendance ($62 - 72\%$), the longest travel time category, relatively poorer sleep quality, no extracurricular activities, no web access, a less favorable study method (Notes), lower parental education (High school), and a public school setting. In contrast, the higher performance profile assumes high attendance ($81 - 100\%$), the shortest travel time category, higher sleep quality, participation in extracurricular activities, web access, a more favorable study method (Mixed), higher parental education (Post graduate 2), and a private school setting.

Using these two sets of inputs, the model generates predicted mean outcomes that illustrate the practical difference in expected performance between the two scenarios.

The model predicts a substantial difference in expected performance between the two scenarios. For Profile A (lower performance), the predicted mean score is 45.50 with a 95% confidence interval of $[44.09, 46.91]$. For Profile B (higher performance), the predicted mean score is 77.87 with a 95%confidence interval of $[76.55, 79.19]$. The intervals are relatively narrow (about ± 1.4 points around each mean), indicating that the expected scores for these profiles are estimated with high precision.

Moreover, the predicted means are far apart and the confidence intervals do not meaningfully overlap, implying a large and clearly distinguishable difference in average predicted outcomes between the lower performance and higherperformance student profiles.

Profile	Predicted mean	CI_low (95%)	CI_high (95%)
Profile A (lower support)	45.50	44.09	46.91
Profile B (higher support)	77.87	76.55	79.19

Contrast	Predicted difference	CI_low (95%)	CI_high (95%)
Profile B – Profile A	22.44	20.13	24.75

The implied difference between profiles is 22.44 points in favor of Profile B ($95\% CI : [20.13, 24.75]$), indicating a large and precisely estimated improvement when moving from the lower support to the higher support scenario.

5 Conclusion

5.1 Mathematic formalism and estimation table

Finlay, after all the model selection phase, we decided to use what we called the model modVIII defined as follow : for all $i \in \{1, ..., 5000\}$

$$\begin{aligned} y_i = b_0 &+ \sum_{k=1}^{4} b_k X_i^{(k)} + \sum_{\substack{k=1,2 \\ \ell=3,4}} b_{k\ell}\, X_i^{(k)} X_i^{(\ell)} \\ &+ \alpha^\top A_i + \sum_{k=3}^{4} \alpha_k^\top \left(A_i \cdot X_i^{(k)}\right) + \beta^\top B_i + \gamma^\top C_i + \delta^\top D_i \\ &+ e^\top E_i + \eta^\top F_i + \kappa^\top (E_i \odot F_i) + \theta^\top G_i + \tau^\top H_i + \varepsilon_i, \end{aligned} \tag{1}$$

Where $\epsilon_i \sim^{i.i.d.} \mathcal{N}(0,1)$ and

$$E_i \odot F_i \;=\; \begin{pmatrix} E_{i1}F_{i1} \\ E_{i1}F_{i2} \\ E_{i2}F_{i1} \\ E_{i2}F_{i2} \\ E_{i3}F_{i1} \\ E_{i3}F_{i2} \\ E_{i4}F_{i1} \\ E_{i4}F_{i2} \\ E_{i5}F_{i1} \\ E_{i5}F_{i2} \\ E_{i6}F_{i1} \\ E_{i6}F_{i2} \end{pmatrix} \in \mathbb{R}^{12}.$$

The variable use in this equation are given by :

Quantitative variables of our model

Variables	Reference
$X^{(1)} \in \mathbb{R}^{5000}$	$sleep\_c = sleep\_hours - mean\_sleep$, where $mean\_sleep$ is the mean of the sleep hours on the train set
$X^{(2)} \in \mathbb{R}^{5000}$	$sleep\_c2 = (sleep\_c)^2$.
$X^{(3)} \in \mathbb{R}^{5000}$	$study\_c = study\_hours - mean\_study$, where $mean\_study$ is the mean of the study hours on the train set
$X^{(4)} \in \mathbb{R}^{5000}$	$study\_c2 = (study\_c)^2$.

Qualitative variables of our model

Variables	Reference
$A \in \mathbb{R}^{4\times5000}$	attend_pct_cat
$B \in \mathbb{R}^{4\times5000}$	trav_time
$C \in \mathbb{R}^{3\times5000}$	sleep_qual
$D \in \mathbb{R}^{2\times5000}$	extra_act
$E \in \mathbb{R}^{6\times5000}$	parent_educ
$F \in \mathbb{R}^{2\times5000}$	school_type
$G \in \mathbb{R}^{2\times5000}$	web_access
$H \in \mathbb{R}^{6\times5000}$	study_method
$K \in \mathbb{R}^{12\times5000}$	interaction between parent_educ and school_type

Coefficients of our model

coefficient	Reference
$b_0 \in \mathbb{R}$	the intercept
$b_k \in \mathbb{R}$ **for** $k \in \{1,2,3,4\}$	represent the influence of each qualitative variables
$b_{kl} \in \mathbb{R}$ **for** $k \in \{1,2\}$ **and** $l \in \{3,4\}$	represent the iteration between $X^{(k)}$ and $X^{(l)}$
$\alpha^T \in \mathbb{R}^4$	represent the influence of attend_pct_cat
$\beta^T \in \mathbb{R}^4$	represent the influence of trav_time
$\gamma^T \in \mathbb{R}^3$	represent the influence of sleep_qual
$\delta^T \in \mathbb{R}^2$	represent the influence of extra_act
$e^T \in \mathbb{R}^6$	represent the influence of parent_educ
$\eta^T \in \mathbb{R}^2$	represent the influence of school_type
$\theta^T \in \mathbb{R}^2$	represent the influence of web_access
$\tau^T \in \mathbb{R}^6$	represent the influence of study_method
$\kappa \in \mathbb{R}^{12}$	represent the interaction between parent_educ and school_type

5.2 Interpretation

Ultimately, the final model highlights a limited number of key variables and interactions. First, age and gender do not appear to be sufficiently strong explanatory factors to be retained in the model. Whether a student is male or female, younger or older, these characteristics do not play a decisive role in determining final exam performance once other factors are taken into account.

One of the most significant findings of the model is the interaction between study hours and sleep duration. Study time is inherently linked to sleep patterns, as the amount of time devoted to studying necessarily affects sleep rhythms, which in turn influence cognitive clarity, concentration, and the ability to organize one's work effectively. Adequate sleep is essential not only for maintaining mental alertness but also for sustaining a productive and structured approach to academic work.

This result is consistent with findings from the academic literature. In 2016, researchers at McGill University demonstrated that a good night's sleep is associated with improved perfor-

mance in mathematics and language-related subjects. This research led to the implementation of a sleep education program that successfully improved both student wellbeing and academic outcomes. Similarly, a 2018 study conducted in Morocco among 308 medical students in Fez found that more than half of the participants reported poor sleep quality, which was significantly associated with lower academic performance, difficulties in time management, and ineffective study strategies.

More broadly, numerous studies have documented the strong relationship between sleep, memory consolidation, and the formation of neural networks, including research conducted by CNRS. These findings strongly support the mechanisms captured by our model, which effectively predicts declines in exam performance as a function of the interaction between insufficient sleep and intensive study time. Overall, the model's results align closely with established evidence, reinforcing the central role of sleep as a key facilitator of academic success rather than a purely additive factor.

The two remaining interactions are less compelling from a medical or psychological perspective but nonetheless present meaningful sociological relevance. In particular, the interaction between attendance and study hours appears to reflect the fact that students who attend classes more regularly are not only more directly exposed to course content but are also embedded in an environment that encourages academic engagement.

This interpretation is supported by a study conducted in 2024 by Farges and Monso, which shows that boarding students are more likely to succeed academically. Such students are effectively required to attend classes regularly and, more importantly, are immersed in a collective learning environment composed of peers pursuing similar studies. This environment fosters mutual support, academic exchanges, and shared study practices, all of which can positively influence academic outcomes.

From this perspective, the interaction between attendance and study hours, while seemingly intuitive at first glance, reveals deeper sociological mechanisms that may play a decisive role in shaping academic performance. These mechanisms help explain the weight of this interaction in our model and highlight the importance of social and institutional contexts in supporting student success.

Finally, the interaction between parental education level and school type highlights important social mechanisms shaping academic performance. Access to higher education tends to condition parents' expectations regarding their children's academic success. As the perceived quality of public education declines, parents who have attained higher levels of education—and often, concomitantly, a higher socio-economic status—are more likely to enroll their children in private schools.

This dynamic contributes to the reinforcement of educational inequalities. More educated parents are not only better equipped to support their children academically, but also more able to finance private schooling or supplementary tutoring. In contrast, parents with fewer academic and economic resources often have limited alternatives and are more likely to rely on public schools, sometimes without fully grasping the structural disadvantages their children may face.

Moreover, differences in academic outcomes between private and public schools are not solely attributable to student composition. They also reflect disparities in instructional pace, individualized support, and institutional expectations. Private schools tend to provide closer supervision and exert stronger social and academic pressure, factors that can significantly influence student performance. In contemporary educational contexts, academic excellence is often more socially valued and encouraged in private schools than in public ones, further amplifying performance gaps.

Taken together, these mechanisms help explain the significance of the interaction between parental education and school type in our model, underscoring the role of social stratification and institutional context in shaping educational outcomes.

Finally, the remaining variables also exhibit meaningful effects on exam performance. However, it is the combination of all these factors that allows us to obtain a robust and effective model, capable of delivering strong predictive performance on both medium sized and large samples.

5.3 Limits

Throughout this study, we faced limitations due to the absence of certain key pieces of information that could have enabled a more targeted and precise analysis. In particular, no data were available on the students' geographic origin or on the specific educational systems of their schools. As a result, the interpretations remain relatively general and cannot be fully aligned with the academic frameworks or institutional rules of specific countries.

Indeed, all the criteria examined such as the role of private versus public schools, study hours, class participation, or access to the internet are highly context dependent and vary substantially across countries and time periods. Without information on the national or institutional context in which the study was conducted, the estimated effects of these variables may differ considerably from what would be observed in another setting.

This constitutes the main limitation of the study, as it restricts the analysis to a surface level exploration of the topic rather than allowing for context specific conclusions. Nevertheless, certain patterns in the data such as the strong link between school type and parental education, or the prevalence of particular study methods suggest that the sample may correspond to a specific region of the world. While such inferences remain speculative, they highlight the importance of contextual information for interpreting educational outcomes accurately.

5.4 References

Articles used in the conclusion :

https://opiq.qc.ca/wp-content/uploads/2016/10/Meilleur_sommeil.pdf

https://www.ih2ef.gouv.fr/frequenter-linternat-lentree-du-lycee-t-il-un-impact-sur-la-reuss

https://www.insb.cnrs.fr/fr/cnrsinfo/larchitecture-du-sommeil-la-cle-pour-une-memoire-optim

https://www.sciencedirect.com/science/article/pii/S039876202300250X