MDA9159 Data Analysis Project Report

Modelling Student Exam Performance Using Multiple Linear Regression

MDA9159 – Statistical Modelling and Data Analysis

Instructor: Dr. Guowen Huang

Team Members

Name	**Student ID**
Zhui Geng	251157401
Wen Yang	25YYYYYYY
Yuxuan Zhou	25ZZZZZZZ

Western University

December 16, 2025

1 Executive Summary

This project investigates which factors drive students' exam performance and how accurately we can predict individual exam scores using demographic, behavioural, and school-level characteristics. The dataset contains 6,607 students and includes study behaviour (`Hours_Studied`, `Attendance`, `Previous_Scores`, `Tutoring_Sessions`), family background (`Family_Income`, `Parental_Education_Level`, `Parental_Involvement`), school environment (`Teacher_Quality`, `School_Type`, `Peer_Influence`, `Access_to_Resources`) and health-related variables (`Sleep_Hours`, `Physical_Activity`, `Learning_Disabilities`), with `Exam_Score` as the continuous outcome.

We first fit a multiple linear regression (MLR) with all main effects, using a 70/30 train–test split. Categorical predictors were encoded with dummy variables via `model.matrix`. We then extended the analysis with: (i) an OLS model with selected two-way interactions chosen through stepwise AIC, (ii) Ridge regression and Lasso regression on all main effects, and (iii) Lasso with all two-way interactions, all tuned using cross-validated `glmnet`. A comparison table summarizes $RMSE$ and adjusted R^2 between models.

Across all approaches, predictive performance is high and very similar. The best out-of-sample $RMSE$ is around ≈ 1.82, and adjusted R^2 lies between 0.70–0.71, indicating that roughly 70% of the variation in exam scores is explained by observed features. Attendance, hours studied, previous academic performance, tutoring sessions, internet access, and a supportive school/family environment all show strong positive associations with exam scores, while limited resources, low parental involvement, low motivation, and learning disabilities are associated with lower scores. In terms of a final model, Lasso on main effects offers a good balance of interpretability, parsimony, and predictive accuracy, with performance effectively matching more complex interaction models.

2 Data Description and Exploratory Analysis

2.1 Variable descriptions

The dataset contains 6,607 observations. The key variables are:

- **Exam_Score**: target variable; continuous exam score.
- **Study behaviour**
  - *Hours_Studied* (1–44 hours).
  - *Attendance* (% of classes attended).

- *Previous_Scores* (prior exam performance).
    - *Tutoring_Sessions* (0–8 sessions per week).

- **Lifestyle & health**
    - *Sleep_Hours* (4–10 hours/night).
    - *Physical_Activity* (0–6 sessions/week).
    - *Learning_Disabilities* (Yes/No).

- **Family background**
    - *Family_Income* (Low/Medium/High).
    - *Parental_Involvement* (Low/Medium/High).
    - *Parental_Education_Level* (High School/College/Postgraduate).

- **School environment**
    - *Teacher_Quality* (Low/Medium/High).
    - *School_Type* (Public/Private).
    - *Access_to_Resources* (Low/Medium/High).
    - *Peer_Influence* (Negative/Neutral/Positive).

- **Other**
    - *Internet_Access* (Yes/No), *Extracurricular_Activities* (Yes/No), *Distance_from_Home* (Near/Moderate/Far), *Gender* (Male/Female).

2.2 Summary statistics and simple tables

Table 1: Summary statistics for numeric variables ($n = 6607$).

Variable	Mean (SD)	Median [Q1, Q3]	Range
`Exam_Score`	67.24 (3.89)	67 [65, 69]	55–101
`Hours_Studied`	19.98 (5.99)	20 [16, 24]	1–44
`Attendance`	79.98 (11.55)	80 [70, 90]	60–100
`Previous_Scores`	75.07 (14.40)	75 [63, 88]	50–100
`Tutoring_Sessions`	1.49 (1.23)	1 [1, 2]	0–8
`Sleep_Hours`	7.03 (1.47)	7 [6, 8]	4–10
`Physical_Activity`	2.97 (1.03)	3 [2, 4]	0–6

Table 2: Distribution of key ordinal predictors ($n = 6607$).

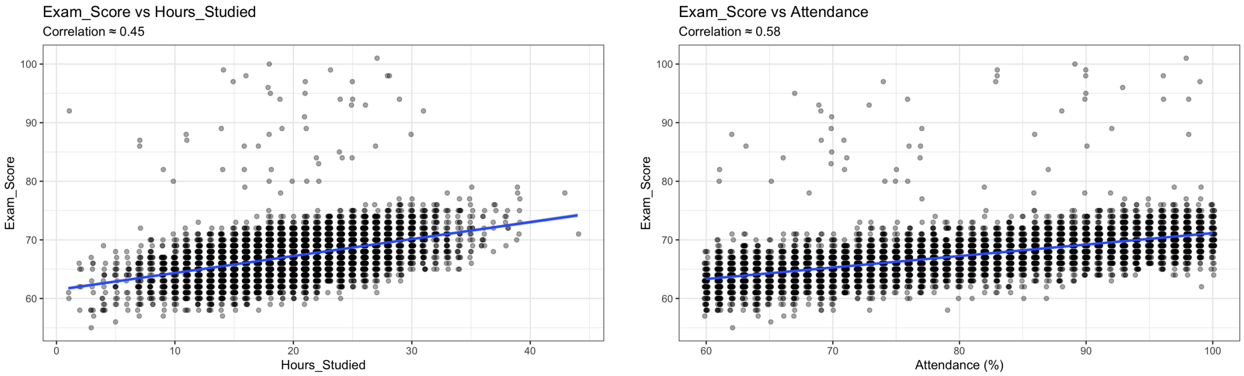
Variable	Low	Medium	High
`Parental_Involvement`	20.2% (1337)	50.9% (3362)	28.9% (1908)
`Access_to_Resources`	19.9% (1313)	50.2% (3319)	29.9% (1975)
`Motivation_Level`	29.3% (1937)	50.7% (3351)	20.0% (1319)

Table 3: Mean `Exam_Score` by selected groups.

Factor	Contrast (A vs B)	Mean(A)	Mean(B)	Diff (A–B)
`Parental_Involvement`	High vs Low	68.09	66.36	+1.73
`Access_to_Resources`	High vs Low	68.09	66.20	+1.89
`Motivation_Level`	High vs Low	67.70	66.75	+0.95
`Teacher_Quality`	High vs Low	67.68	66.75	+0.92
`Family_Income`	High vs Low	67.84	66.85	+0.99
`Peer_Influence`	Positive vs Negative	67.62	66.56	+1.06
`Learning_Disabilities`	Yes vs No	66.27	67.35	–1.08
`Gender`	Female vs Male	67.24	67.23	+0.02

2.3 Exploratory plots

Exploratory visualisations were generated in R using `ggplot2` (scatterplots and boxplots); the key patterns are summarised below:



(a) `Exam_Score` vs `Hours_Studied` (positive trend).

(b) `Exam_Score` vs `Attendance` (positive trend).

Figure 1: Scatterplots with fitted linear trend lines showing positive associations between `Exam_Score` and key study behaviours.

- **Study behaviour (scatterplots).** `Exam_Score` increases with both `Hours_Studied` and `Attendance`, showing approximately linear trends. The correlations are $r \approx 0.45$ (Hours_Studied) and $r \approx 0.58$ (Attendance). In bivariate linear regressions, the slope is about ≈ 0.29 points per additional study hour (i.e., ≈ 2.9 points per 10 hours) and ≈ 0.20 points per 1% attendance (i.e., ≈ 2.0 points per +10% attendance). While the trend is clear, there is still noticeable spread at each x-value, indicating other factors also contribute to performance.

- **Monotonic shifts in ordinal factors (boxplots).** Boxplots of `Exam_Score` by `Parental_Involvement`, `Access_to_Resources`, `Motivation_Level`, and `Teacher_Quality` show broadly monotonic patterns from Low → Medium → High. Mean score differences between High and Low are approximately +1.73 (Parental involvement), +1.89 (Access to resources), +0.95 (Motivation), and +0.92 (Teacher quality), consistent with an overall $\approx$ 1–2 point upward shift in the distribution under more supportive conditions.

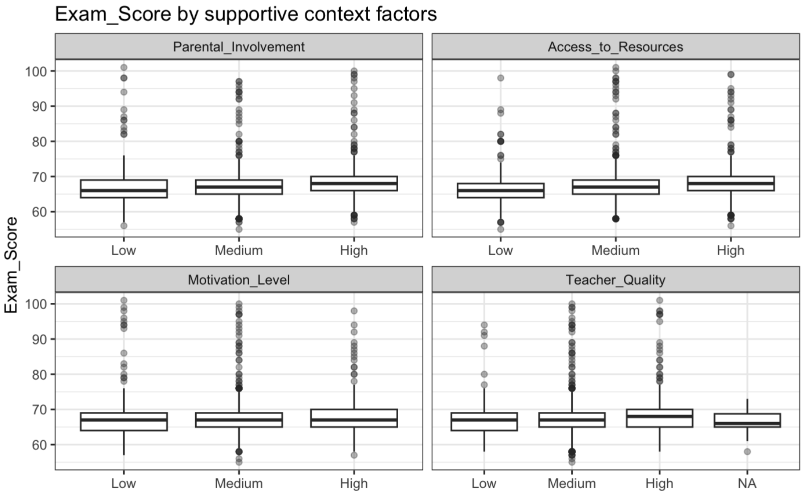


Figure 2: Boxplots of `Exam_Score` by ordinal supportive factors (`Parental_Involvement`, `Access_to_Resources`, `Motivation_Level`, `Teacher_Quality`).

- **Lower performance groups (boxplots).** Students with `Learning_Disabilities` = Yes have a lower mean score by about 1.08 points compared with No. For `Peer_Influence`, the mean score under Negative is about 1.06 points lower than under Positive, suggesting negative peer environments are associated with reduced exam performance.

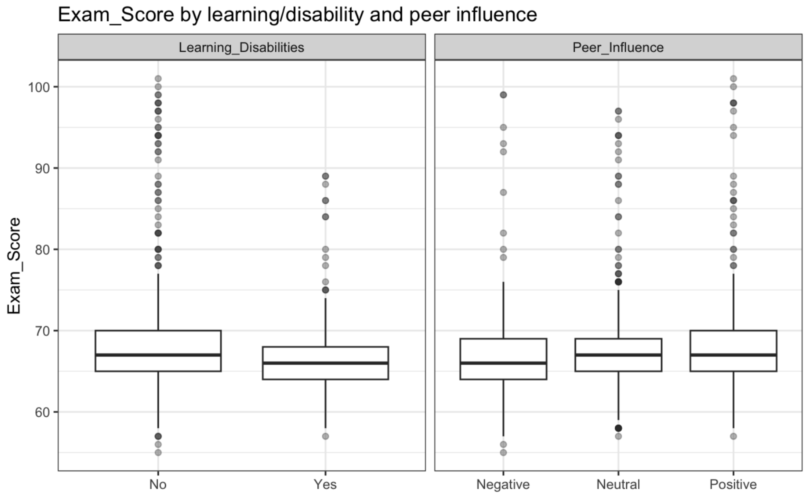


Figure 3: Boxplots of `Exam_Score` by `Learning_Disabilities` and `Peer_Influence`.

2.4 Discussion of patterns and potential issues

The EDA suggests that study effort and attendance are among the strongest individual predictors of `Exam_Score`, with additional (smaller) contributions from prior performance, tutoring, and psychosocial factors (e.g., motivation, parental involvement, and peer influence). Within the observed ranges, relationships between numeric predictors and `Exam_Score` appear approximately linear. Because the exploratory plots are primarily marginal views, interaction effects (e.g., `Hours_Studied` × `Attendance` or interactions involving `School_Type`) are not directly identified from EDA alone, but are conceptually plausible and will be formally tested in later models.

Missingness appears limited overall; however, a small amount of missing data is present in a few categorical predictors (e.g., `Teacher_Quality` contains some NA values). In the OLS models, missing values are handled by complete-case analysis (listwise deletion), so results are based on available observations. Because many predictors capture related constructs (e.g., parental involvement, access to resources, and family income), multicollinearity is also possible. These considerations motivate examining model diagnostics and comparing OLS with penalized regression approaches (ridge and lasso) to improve stability.

3 Model Building and Justification

3.1 Base Model Construction:

We began by splitting the data into a 70% training set and a 30% test set (with a fixed seed) so that model choices are made on the training data while performance is judged on unseen data. This matches the project expectation that model building should be justified and validated, not only fit on the same data used to train the model.

As the baseline, we fit an ordinary multiple linear regression (MLR) using all predictors as main effects:

$$target \sim .$$

implemented as $lm(target \sim ., data = train\_dat)$. This baseline is a transparent starting point for interpretation and for later comparison against more flexible/regularized models. On the training data, the baseline achieved Adjusted $R^2 \approx 0.509$

For penalized regression (Ridge/Lasso), we explicitly used model.matrix(...) to build design matrices (and remove the intercept column). This approach also ensures that if any predictors are stored as categorical/factor variables, they will be handled consistently through dummy encoding.

We used $RMSE$ (train and test) as the primary predictive metric because it measures average prediction error in the same unit as the response and heavily penalizes large errors, which is appropriate for comparing predictive accuracy across candidate models. The summary table reports both Train/Test $RMSE$ and Adjusted R^2 to balance predictive performance with explanatory fit.

3.2 Model Selection Process:

Our model selection compared candidates using a logical process (AIC-based selection and regularization), while explicitly considering multicellularity and nonlinearity/interactions. We evaluated five model structures: baseline OLS, stepwise with interactions, Ridge, Lasso, and Lasso with full two-way interactions.

(1) Regularization to address multicollinearity (Ridge and Lasso main effects).

Because predictors can be correlated, we fit Ridge ($\alpha = 0$) and Lasso ($\alpha = 1$) using cross-validation to select λ (with cv.glmnet). Ridge shrinks coefficients smoothly (helpful when correlated predictors share signal), while Lasso can additionally set some coefficients exactly to zero (feature selection). In our results:

- Ridge selected: $\lambda_{min} = 0.0469$ with Test RMSE $= 0.3531$.
- Lasso selected: $\lambda_{min} = 0.0054$ with Test RMSE $= 0.3536$.

These were very close to the baseline OLS test $RMSE$ (0.353), suggesting that with main effects only, shrinkage did not materially change generalization error, but it still provides a principled safeguard against unstable estimates when predictors are correlated.

(2) Capturing nonlinearity via interactions (Stepwise AIC). To address potential nonlinearity and effect modification, we expanded the candidate space to include all two-way interactions using an upper scope of $(.)^2$, and applied bidirectional stepwise selection based on AIC. This produced a model with main effects plus a subset of interactions and improved predictive accuracy: Adjusted $R^2 = 0.6604$ and Test $RMSE$ $= 0.3263$. This jump indicates that interactions contain meaningful structure that the main-effects-only models cannot capture.

(3) High-dimensional interactions with control for overfitting (Lasso with full $(.)^2$).

Stepwise selection can improve fit, but when many interaction terms are available, the model space becomes large and the risk of overfitting increases. To keep the flexibility of $(.)^2$ while controlling complexity, we fit Lasso on the full set of main effects + all two-way interactions, letting the penalty automatically shrink weak terms.

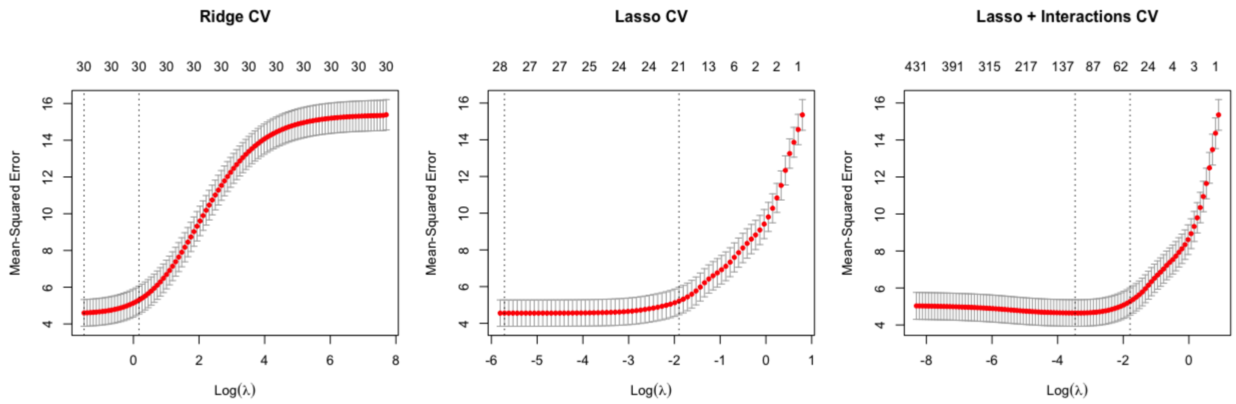


Figure 4: Cross-validation curves for Ridge, Lasso, and Lasso+Interactions.

3.3 Final model choice and justification:

We selected Lasso with second-order interactions as the final model because it achieved the best held-out test $RMSE$ among all candidates (0.325) while still using regularization to control complexity in an interaction-rich design.

This choice directly satisfies the rubric requirement for a clear justification of the final model and demonstrates that the main gain in predictive performance comes from modeling interaction effects, with Lasso providing a structured way to keep only the most useful signal.

4 Model Interpretation

4.1 Interpretation of regression coefficients

We interpret the coefficients from the baseline multiple linear regression (MLR) fit on the training set. In this model, each coefficient represents the expected change in `Exam_Score` for a one-unit increase in the predictor, holding other variables constant. For categorical variables, each coefficient is interpreted relative to a reference category.

Overall, the strongest positive academic-behaviour drivers are `Attendance` and `Hours_Studied`. The coefficient for `Attendance` is about 0.201, meaning that a one-unit increase in attendance is associated with average around a 0.20 point increase in `Exam_Score` after controlling for other variables. Similarly, `Hours_Studied` has an estimated coefficient of 0.290, suggesting that additional study time is associated with higher exam scores. We also see positive effects from `Tutoring_Sessions` (estimate about 0.512) and `Previous_Scores` (estimate about 0.051), showing that extra academic support and prior performance are strongly related to the current exam result.

Several support and environment variables show meaningful negative associations when the support level is low. For example, `Access_to_Resources = Low` is associated with a drop of about 2.12 points, and `Parental_Involvement = Low` is associated with a drop of about 2.04 points, compared with their reference groups. These are relatively large effects given the scale of `Exam_Score`. The model also suggests that `Motivation_Level = Low` and `Family_Income = Low` are associated with lower exam scores.

Table 4: Key coefficient estimates from the baseline MLR model (training set).

Variable	Estimate	Std. Error	t value	Pr(> $\lvert t\rvert$)
`(Intercept)`	41.16692864	0.633239003	65.01010	0.000000e+00
`Hours_Studied`	0.29004395	0.005216571	55.60049	0.000000e+00
`Attendance`	0.20053685	0.002729594	73.46765	0.000000e+00
`Access_to_ResourcesLow`	-2.12034115	0.090413933	-23.45149	5.710101e-115
`Previous_Scores`	0.05118284	0.002192241	23.34726	5.066308e-114
`Parental_InvolvementLow`	-2.03637620	0.091047705	-22.36604	2.927146e-105
`Tutoring_Sessions`	0.51236041	0.025481090	20.10748	2.966784e-86
`Parental_InvolvementMedium`	-1.16540167	0.073104373	-15.94161	9.937939e-56
`Access_to_ResourcesMedium`	-1.05505657	0.072018077	-14.64989	1.570775e-47
`Family_IncomeLow`	-1.17637518	0.086365430	-13.62090	1.902372e-41
`Motivation_LevelLow`	-1.04738649	0.090468178	-11.57740	1.420284e-30
`Peer_InfluencePositive`	0.95935090	0.085041250	11.28101	3.923481e-29

4.2 Interpretation of R^2 and adjusted R^2

The baseline MLR has $R^2 \approx 0.708$ and adjusted $R^2 \approx 0.706$. This means that about 70% of the variation in `Exam_Score` can be explained by the predictors included in the model. The adjusted R^2 is very close to R^2, which suggests the model fit is not only coming from adding many predictors, but that the predictors overall provide real explanatory power.

4.3 Interpretation of confidence intervals vs prediction intervals

To explain uncertainty clearly, we computed both confidence intervals (CI) and prediction intervals (PI) for two example student profiles (a "high-support" case and a "low-support" case). A 95% CI describes uncertainty for the mean predicted score (average outcome for similar students), while a 95% PI describes uncertainty for a single individual student's score, so the PI is wider.

For the high-support example, the predicted score is 72.39, with a 95% CI of [72.14, 72.63], but the 95% PI is [68.21, 76.56]. For the low-support example, the predicted score is 60.40, with a 95% CI of [60.06, 60.75], and a 95% PI of [56.22, 64.58]. This matches the key idea: CI is narrow (mean effect), PI is wider (individual variability).

Table 5: Predicted `Exam_Score` for two example profiles with 95% confidence intervals (CI) and 95% prediction intervals (PI).

Profile	Predicted	CI_Lower	CI_Upper	PI_Lower	PI_Upper
High-support example	72.39	72.14	72.63	68.21	76.56
Low-support example	60.40	60.06	60.75	56.22	64.58

5 Model Diagnostics & Validation

5.1 Residual plots

We first checked the standard residual diagnostics from the baseline MLR. The residuals versus fitted plot shows most residuals centered around zero, but there are several observations with extremely large positive residuals (around 20–30), meaning the model strongly under-predicts exam scores for a small number of students. The Q–Q plot also shows a clear deviation in the upper tail, consistent with these extreme positive residuals and suggesting that residuals are not perfectly normal. The scale-location plot does not show a dramatic funnel shape, but it suggests mild changes in spread, so we also checked constant variance more formally below.

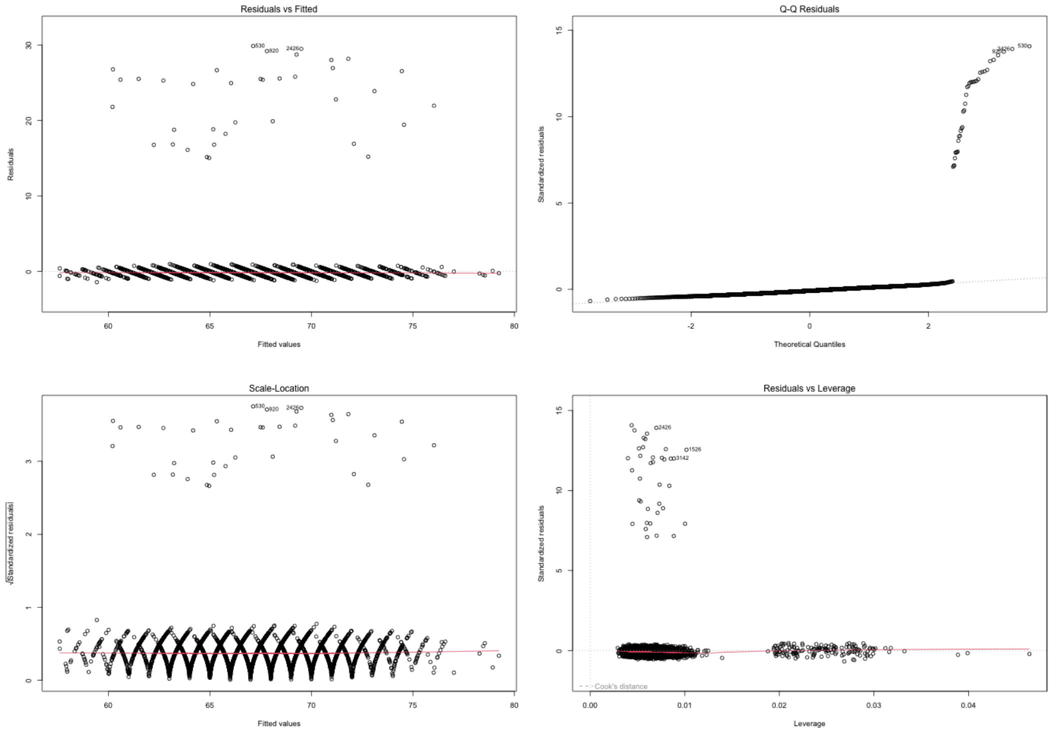


Figure 5: Standard diagnostic plots for the baseline MLR model.

5.2 Outlier and influence analysis

Next, we assessed influential observations using Cook's distance. The Cook's distance plot shows a small set of observations standing out above the usual threshold line, and the top influential points have Cook's D values around 0.03–0.05 (largest about 0.052). These are noticeable, but still far below extreme values such as 1. This suggests that we do have influential cases worth checking, but the model is not dominated by a single observation.

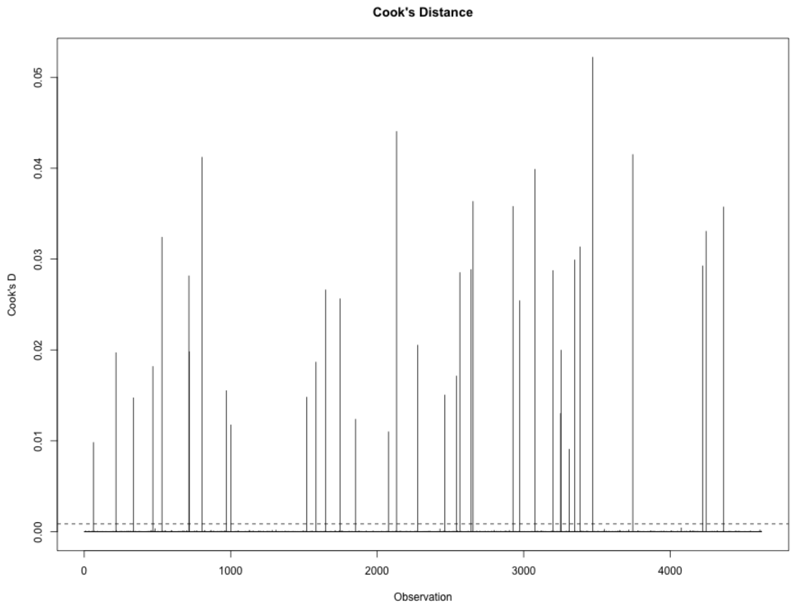


Figure 6: Cook's distance for the baseline MLR model.

Table 6: Top 10 observations by Cook's distance.

obs_id	Observation	Cook's D
1526	3470	0.0522
2426	2132	0.0440
3142	3745	0.0415
4780	805	0.0412
4532	3076	0.0399
2905	2654	0.0363
6394	2927	0.0358
2596	4365	0.0357
3580	4245	0.0331
95	531	0.0324

5.3 Linearity and constant variance assumptions

To check linearity, we used component plus residual (partial residual) plots. Overall, these plots suggest that for many predictors the linear structure is reasonable, but the presence of extreme observations indicates that some patterns (especially the highest scores) may not be fully captured by a simple linear model.

For constant variance, we used the studentized Breusch–Pagan test, which gives p-value = 0.235. Since this p-value is not small, we do not have strong evidence of heteroscedasticity in this model. In other words, variance may not be perfectly constant visually, but the formal test does not indicate a serious violation.

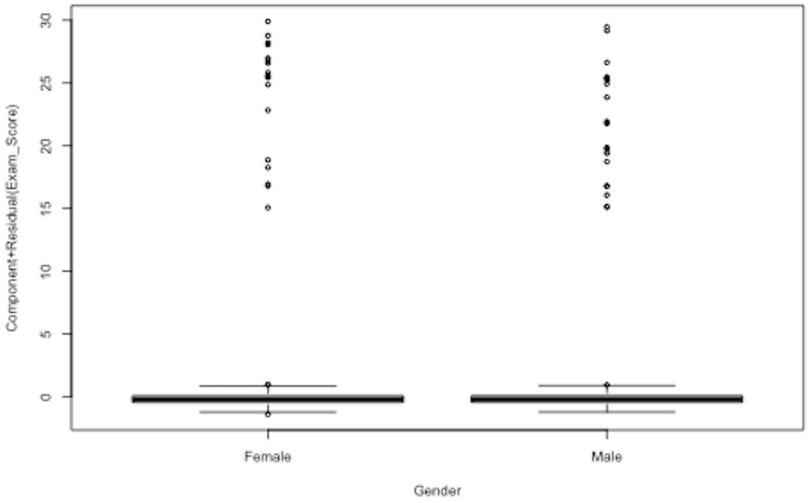


Figure 7: Component plus residual plots for checking linearity (baseline MLR).

5.4 Final assessment of model adequacy

Finally, we evaluated predictive performance on the held-out test set. The baseline MLR achieves test RMSE ≈ 1.818, which indicates good average predictive accuracy in the same unit as `Exam_Score`. The predicted versus actual plot shows that most points follow the 45-degree line closely for the main range of exam scores, supporting overall model adequacy. However, there are several high-score observations where the model under-predicts substantially (actual scores in the high 80s/90s but predicted around the mid-to-high 60s). This matches what we saw in the residual plots: the model performs well for most students, but it struggles with a small number of extreme cases.

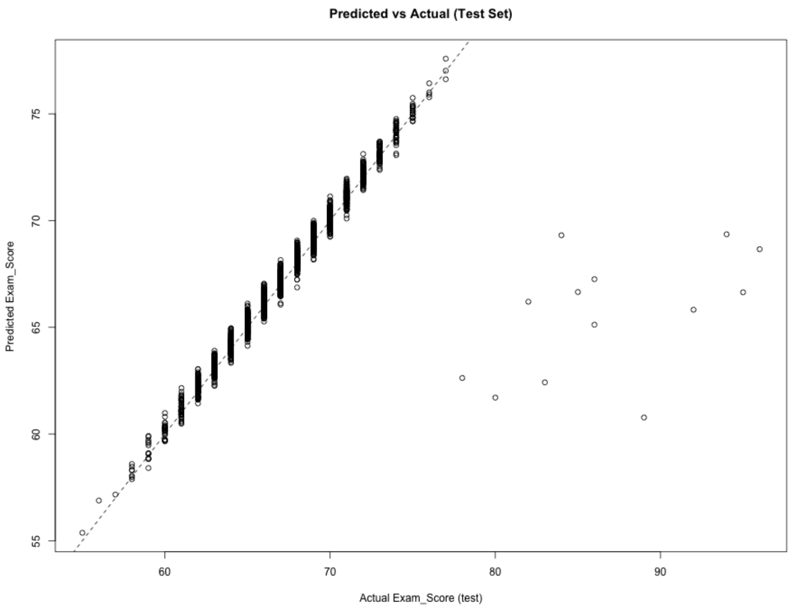


Figure 8: Predicted vs actual `Exam_Score` on the test set

6 Conclusion & Communication

6.1 Clear summary of findings

Using a multiple linear regression framework, we find that students' exam performance is strongly associated with both academic behaviours and support/resource factors. In the baseline MLR, `Attendance` and `Hours_Studied` show large positive associations with `Exam_Score`, and `Tutoring_Sessions` and `Previous_Scores` also contribute positively. In contrast, limited support variables such as low access to resources, low parental involvement, and low motivation are associated with lower exam scores, even after controlling for the other factors. Overall, the model explains about 70% of the variation in exam scores ($R^2 \approx 0.708$, adjusted $R^2 \approx 0.706$).

6.2 Limitations and next steps

There are several limitations to communicate clearly. First, this is an observational dataset, so the regression coefficients represent associations, not guaranteed causal effects. Second, diagnostics show a small number of extreme observations that create strong right-tail deviations in the Q–Q plot and large residuals, meaning the model may not capture the mechanisms behind unusually high scores. Third, although the Breusch–Pagan test does not show strong evidence of heteroscedasticity, some visual patterns suggest that model assumptions are not perfect for all cases.

As next steps, we would (1) inspect the most influential observations (Table 6) to check for unusual profiles or potential data quality issues, (2) consider robust regression or alternative modelling approaches to reduce sensitivity to extreme cases, and (3) explore whether a small number of carefully chosen nonlinear terms or interactions can improve performance for the high-score group, while still keeping the model explainable.

7 R Code Quality and Reproducibility

All code and supplementary material can be found: `https://github.com/HelloWorld-0711/STAT-MODELING`

All analyses are fully reproducible from a single R Markdown file (`student-perform2.Rmd`) and one input dataset (`StudentPerformanceFactors.csv`). A fixed random seed is set (`set.seed(42)`), followed by a clearly documented 70/30 train–test split. Model fitting is deterministic, using `lm` for the baseline OLS model, `step` for AIC-based stepwise selection of interaction terms, and cross-validated `cv.glmnet` for ridge/lasso models.

The code follows a logical workflow (base OLS → ridge/lasso main effects → stepwise interactions → lasso with all interactions → final comparison table), with intermediate objects stored under clear names (e.g., `rmse_*`, `adjR2_*`, `model_summary`).