Bayesian Linear Regression Based Determination of Relevant Correlates of Exam Performance

Carlton J. Berthold

Department of Cognitive Science, Rensselaer Polytechnic Institute

PSYC 4960: Bayesian Data Analysis

Dr. Stefan T. Radev

December 11, 2024

Introduction

Standardized testing and cumulative exam scores remain important predictors of academic achievement and future socioeconomic success and significant research has been conducted on biopsychosocial correlates of test scores. Notably, most previous research has focused on a small number of predictors and has not attempted to quantify the degree to which certain predictors impact test outcomes or identify relationships between predictors. The present paper aims to analyze a synthetic dataset containing a number of potential correlates of exam scores and report which predictors are most relevant to student performance.

A significant correlation has repeatedly been observed between standardized test scores and socioeconomic status (Dixon-Roman et al., 2013; Sacket et al., 2009). Prior research suggests that there does not appear to be a significant effect of parental involvement on exam scores (Curry, 2008). Gender differences seem to exist for traditional multiple-choice tests with males generally performing better (Ben-Shakhar & Sinai, 1991; Bolger & Kellaghan, 1990), but not for performance assessments (Klein et al., 1997). A number of predictors included in the synthetic dataset (access to resources, motivation level, peer influence, and distance from home) are not clearly defined, and it is difficult to assess the potential effects of these variables based on prior literature.

It is hypothesized that the most impactful predictor of exam performance will be family income as an operationalization of socioeconomic status. Usually, school districts are funded through local taxes, meaning wealthier areas tend to have better funded schools and can hire and develop better quality teachers. Family income is thus expected to correlate positively with tutoring sessions and teacher quality, with these two variables

potentially mediating any relationship between family income and exam scores. It is likely that whether a student attends a public or a private school is also related to family income, with a higher degree of private school students expected to come from higher income families. Additionally, there is likely a correlation between parental education level and family income that. Other significant predictors are expected to be hours studied, attendance, access to resources, sleep hours, previous scores, tutoring sessions, teacher quality, learning disabilities, and gender.

Methods

Data Collection

A synthetic dataset containing potential correlates of exam performance was obtained from https://www.kaggle.com/datasets/lainguyn123/student-performance-factors/data. The present paper investigated the variables: hours studied, attendance, parental involvement, access to resources, extracurricular activities, sleep hours, previous scores, motivation level, internet access, tutoring sessions, family income, teacher quality, school type, peer influence, physical activity, learning disabilities, parental education level, distance from home, gender, and exam score. No data points were missing, and it was not necessary to impute missing data. The dataset was analyzed using R 4.3.2 with the brms, BayesianMediationA, car, dplyr, ggplot2, loo, Metrics, mice, and recipes packages and the RStudio IDE version 2024.09.1+394.

Data Preprocessing

The dataset (n = 6607) was split into a training set (n = 5000) and a test set (n = 1607) by simple data frame slicing with the first 5000 observations selected as the training set. A

preprocessing pipeline was established to prepare the test and training sets for analysis in a standardized manner while assuring independence during imputation. For both sets, imputation of missing values was handled using Multivariate Imputation by Chained Equations (MICE) using the mice package. Any NA values remaining after MICE imputation were dropped with na.omit().

Classical dummy coding was conducted for both sets using the step_dummy() function and the recipes package to generate separate dummy coded data frames. The BayesianMediationA package used to conduct the exploratory mediation analysis requires manual dummy coding of categorical variables when used as either outcome or mediator covariates, necessitating this extra step.

Exploratory Analysis of Potential Mediating Effects of Family Income

The BayesianMediationA package was used to conduct a mediation analysis across four potential mediators for the relationship between family income and exam score. BayesianMediationA is designed for biostatistical applications and natively accommodates categorical exposure variables making it ideal for this analysis (Yu & Li, 2022). Exam score was used as the outcome variable while family income was used as the exposure variable. School type, teacher quality, tutoring sessions, and access to resources were chosen as potential mediators. All other predictors were considered as outcome covariates, while parental education level was considered as a possible covariate for the selected mediators. Categorical variables selected as outcome and mediator covariates were replaced with classically dummy coded binary variables as described in the ‘Data Preprocessing’ section. Default priors were set over all parameters to reflect the limited information available in the literature.

Regression Modeling of Potential Exam Score Correlates

Three linear regression models were constructed using brms. All three models treated exam scores as the outcome variable, with varying predictors and specified priors. In all cases, the data was assumed to be independent and exchangeable without time or spatial components. The linearity and homoscedasticity of error terms assumptions were assessed by generating residuals versus fitted values plots. Normality of errors was checked with residual Q-Q plots. For all three models, convergence was assessed using the Rhat and ESS metrics calculated by the brms summary function. Predictive performance was assessed for all models by comparing model predictions using test set data with actual exam scores from the test set. Model fit was visualized and assessed using the brms pp_check function. Model comparison was carried out using leave one out cross validation via the loo package and ELPD LOO to determine the best fit model and examine the effects of varying priors, predictors, and model structures on predictive performance. ELPD WAIC was also computed for all models.

Model 1: Complete Pooling with All Predictors (Naïve Fit)

A full pooling, linear regression model with a normal likelihood was set up according to the following joint model specification:

$$\beta_j \sim N(0, 5) \text{ for } j = 0, 1, \ldots, 19 \quad (1)$$

$$\sigma \sim Gamma(1, 1), \quad (2)$$

$$y_n \sim N(\beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_{19} x_{n19}, \sigma^2) \text{ for } n = 1, \ldots, 6607. \quad (3)$$

As in the preliminary analyses, a Gaussian prior with a mean of 0 and standard deviation of 5 was placed over the intercept and model coefficients, while a gamma prior with shape and rate parameters set to 1 was placed over the standard deviation.

The outcome variable for the regression was exam score with hours studied, attendance, parental involvement, access to resources, extracurricular activities, sleep hours, previous scores, motivation level, internet access, tutoring sessions, family income, teacher quality, school type, peer influence, physical activity, learning disabilities, parental education level, distance from home, and gender as predictors.

Model 2: Complete Pooling with Hypothesized Relevant Predictors (Hypothesized Fit)

A simplified full pooling, linear regression model with a normal likelihood was constructed per the following joint model specification:

$$\beta_j \sim N(0, 5) \text{ for } j = 0, 1, \ldots, 10 \quad (1)$$

$$\sigma \sim Gamma(1, 1), \quad (2)$$

$$y_n \sim N(\beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_{10} x_{n10}, \sigma^2) \text{ for } n = 1, \ldots, 6607. \quad (3)$$

The same priors were used for the simplified model as for the model with all predictors used.

The second regression retained exam score as the outcome variable but reduced the predictors to just those hypothesized to be potentially significant (family income, hours studied, attendance, access to resources, sleep hours, previous scores, tutoring sessions, teacher quality, learning disabilities, and gender) to simplify the model and limit overfitting.

Model 3: Complete Pooling with Automatic Relevance Determination for Predictors (R2D2 Fit)

A second full pooling, linear regression model on the full set of predictors was specified as follows:

$$\beta_0 \sim N(0, 5) \quad (1)$$

$$\beta_j \sim R2D2(0.8, 10, 0.6) \text{ for } j = 0, 1, \dots, 19 \quad (2)$$

$$\sigma \sim Gamma(1, 1), \quad (3)$$

$$y_n \sim N(\beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_{19} x_{n19}, \sigma^2) \text{ for } n = 1, \dots, 6607. \quad (4)$$

The only difference between model 3 and model 1 is the use of an R2D2 prior for automatic relevance determination, intended to reduce overfitting and select relevant predictor variables from the 19 total possible predictors. As in model 1 the full set of predictors were used in the regression.

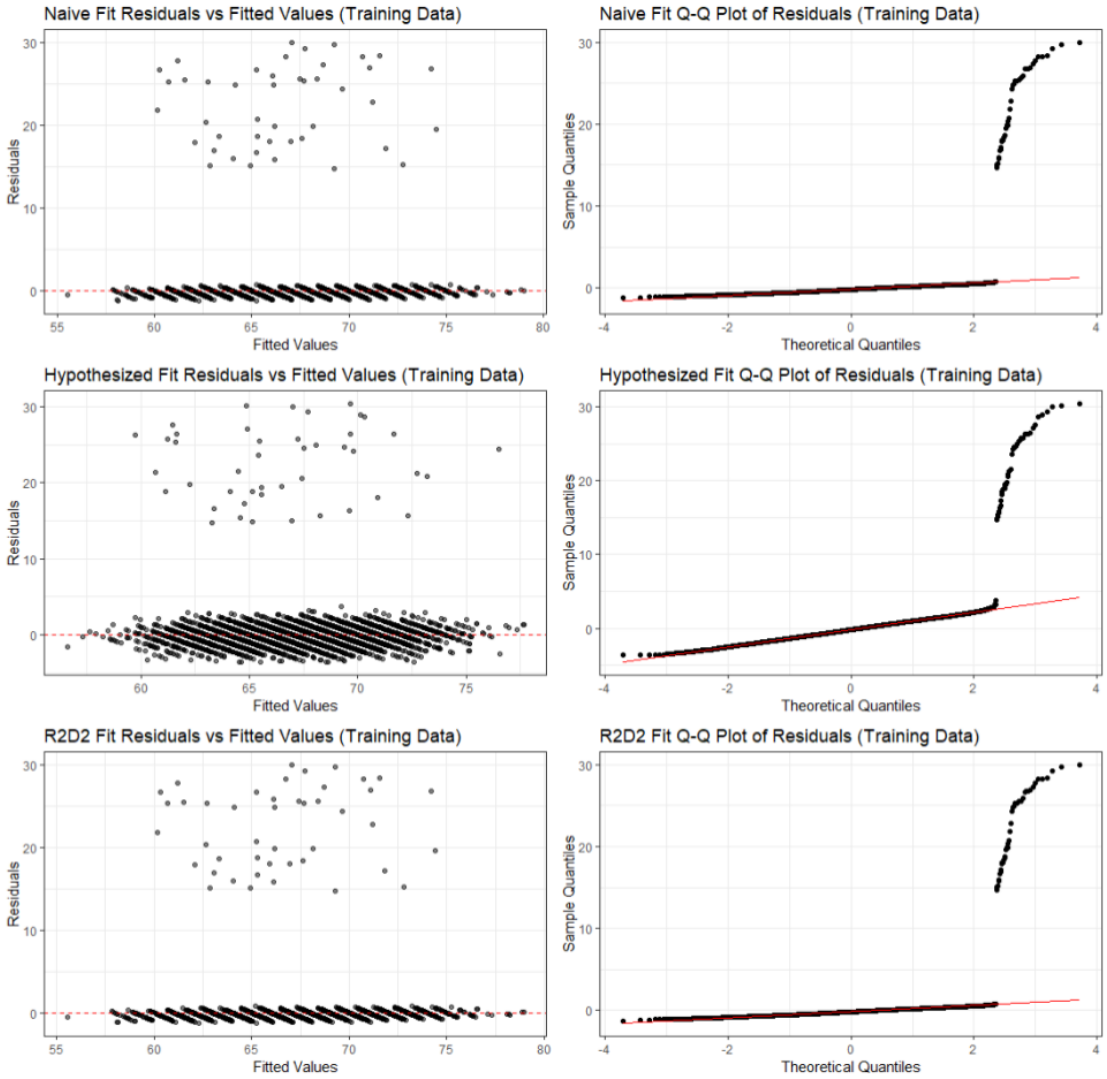
Results

Validation of Assumptions

For all models, residuals versus fitted data plots show fairly random clustering or bouncing of residuals around the 0-line for most fitted values suggesting that the linearity assumption is reasonable. Additionally, for all cases the residuals form a largely rectangular band around the 0-line indicating homoscedasticity of error terms, although this assumption may be slightly weaker for the hypothesized fit. The presence of some relatively large residuals suggests there may be a limited number of outliers. All three Q-Q plots show a sharp departure from the theoretical expectation so there is likely a violation of the normality of errors assumption. Notably the error terms may instead be bimodal.

Figure 1

Residuals vs. Fitted Values Plots & Q-Q Plots of Residuals



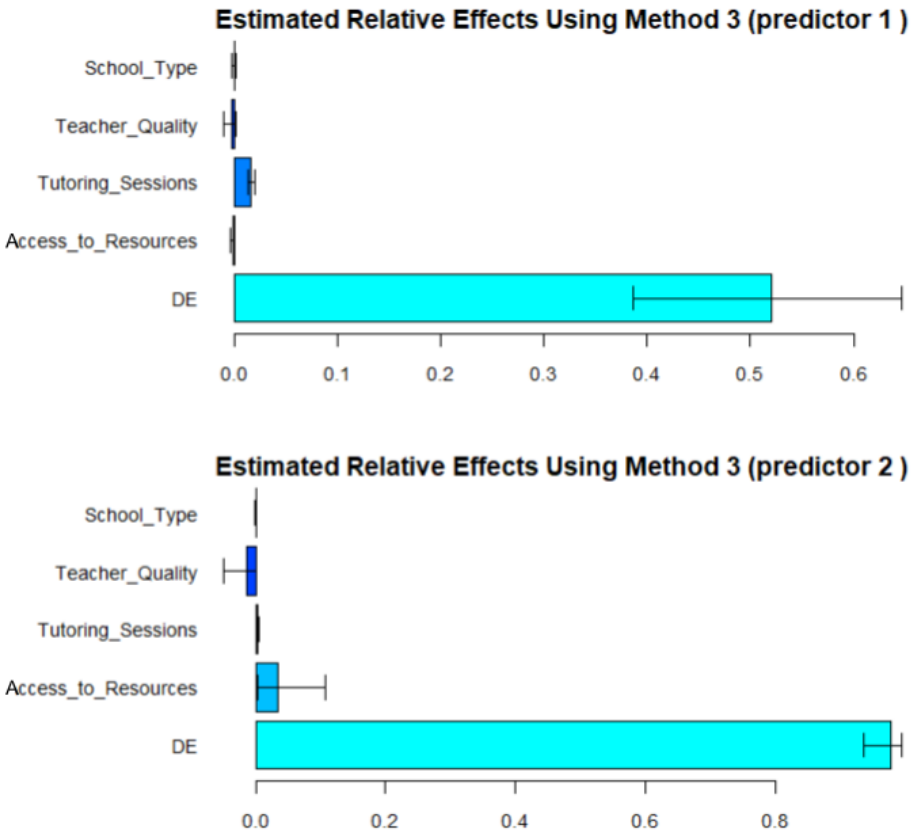
Mediation Analysis

The mediation analysis does not suggest the presence of a mediating interaction between family income and any of the hypothesized mediator variables. No potential mediator

featured a 95% CI not including 0 for any level of family income, suggesting there is not a significant difference in either the low or medium income conditions from the high income condition.

Figure 2.

Mediation Analysis of Family Income on School Type, Teacher Quality, Tutoring Sessions, & Access to Resources



Note. Predictor 1 represents the low family income condition while predictor 2 represents medium family income. High family income is the reference level.

Regression Model Convergence

All models converged with all Rhats within ± 0.00015 of 1 and all ESS metrics greater than 2000. Visual inspections of the trace plots for all models indicate good mixing.

Test Set Validation & Assessment of Model Predictive Performance

Table 1

Model Predictive Performance Metrics & Comparison

Fit	Naïve	Hypothesized	R2D2
MAE	0.4100389	0.9917534	0.4113888
RMSE	1.474457	1.833741	1.476871
R^2	0.8486399	0.7646785	0.84824

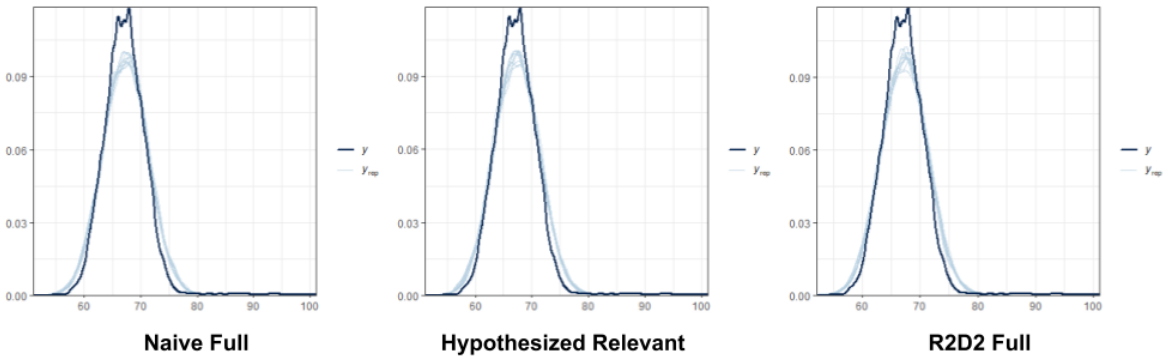
Analysis of the mean absolute error (MAE), root mean square error (RMSE) and R^2 values comparing model predictions on test set values and actual test set exam score observations suggest similar out of sample predictive performance for the naïve and R2D2 models, with a significant drop in predictive performance for the hypothesized fit. R^2 goodness of fit is similarly slightly lower for the hypothesized fit than for the other two fits.

Posterior Predictive Checks & Comparison of Model Fits

Posterior predictive checks for all three models suggest similar generative performance with all models producing data that closely matches the observed values. There are no clear differences or deviations between the graphical PP check plots for any of the assessed models.

Figure 3

Posterior Predictive Checks of Fitted Models



Comparison of ELPD LOO and ELPD WAIC values across all models shows essentially no difference in model fit between the naïve and R2D2 models with notably worse fit for the hypothesized relevant model. Between the naïve and R2D2 models, the standard error terms of the ELPD LOO and ELPD WAIC estimates are not sufficiently different for the minor observed difference in LOO and WAIC criterion values to be accepted as likely. For the hypothesized relevant fit, both ELPD LOO and ELPD WAIC values and therefore differences are significantly higher than the standard error difference, and it is clear that the hypothesized model has inferior fit compared to the other two models. Criterion values are summarized in Table 2.

Table 2

ELPD LOO & ELPD WAIC Model Comparison

Fit	Naïve	Hypothesized	R2D2
ELPD LOO	-10774.86	-11260.2	-10775.16

SE ELPD LOO	384.1423	309.7044	383.9809
ELPD LOO Dif.	0.000000	-485.338529	-0.2943159
SE ELPD LOO Dif.	0.000000	81.1922838	0.7352044
ELPD WAIC	-10775.04	-11259.95	-10775.15
SE ELPD WAIC	384.179	309.6536	383.9804
ELPD WAIC Dif.	0.000000	-484.91	-0.11
SE ELPD WAIC Dif.	0.000000	74.5254	0.1986

Regression Results

Positive peer influence, access to the internet, and being close to home saw the greatest improvement in exam scores. Neutral peer influence (relative to negative peer influence), extracurricular participation, tutoring sessions, highly educated (postgraduate) parents, moderate distance from home, hours studied, attendance, and physical activity all predicted moderately increased exam scores. Previous test scores, the type of school (public or private), hours of sleep, and gender had very small coefficients with 95% credible intervals including zero. Medium quality teachers, parents with only a high school diploma, moderate motivation, and moderate family income all predicted moderately depressed exam scores. Finally, low to medium access to resources, low to medium parental involvement, low family income, low motivation, low quality teachers (relative to high quality teachers) and learning disabilities all strongly reduced exam performance.

Table 3

Naive Prior Full Predictors Regression Results

Effect	Estimate	Est.Error	l-95% CI	u-95% CI
Intercept	42.12912	0.41953	41.30418	42.94373
Peer_InfluencePositive	1.03810	0.08704	0.86694	1.21014
Distance_from_HomeNear	0.89151	0.10870	0.68028	1.10440
Internet_AccessYes	0.86449	0.12115	0.62861	1.10036
Peer_InfluenceNeutral	0.55598	0.08710	0.38503	0.72690
Extracurricular_ActivitiesYes	0.52969	0.06531	0.40042	0.65708
Tutoring_Sessions	0.50700	0.02581	0.45676	0.55762
Parental_Education_LevelPostgraduate	0.47381	0.09243	0.29246	0.65478
Distance_from_HomeModerate	0.39798	0.11595	0.17043	0.62462
Hours_Studied	0.29315	0.00539	0.28263	0.30364
Attendance	0.19681	0.00278	0.19141	0.20227
Physical_Activity	0.16870	0.03122	0.10766	0.22927
Previous_Scores	0.04818	0.00225	0.04376	0.05264
School_TypePublic	0.00950	0.06921	-0.12648	0.14435
Sleep_Hours	0.00043	0.02168	-0.04252	0.04270
GenderMale	-0.03094	0.06518	-0.15797	0.09605
Teacher_QualityMedium	-0.47126	0.07269	-0.61329	-0.32799
Parental_Education_LevelHighSchool	-0.50984	0.07430	-0.65505	-0.36346
Motivation_LevelMedium	-0.52990	0.08582	-0.69698	-0.35941
Family_IncomeMedium	-0.53809	0.08918	-0.71345	-0.36464
Learning_DisabilitiesYes	-0.82006	0.10432	-1.02449	-0.61448

Teacher_QualityLow	-1.00402	0.11496	-1.22977	-0.77790
Family_IncomeLow	-1.01581	0.08918	-1.18942	-0.84146
Access_to_ResourcesMedium	-1.02362	0.07397	-1.16836	-0.87747
Parental_InvolvementMedium	-1.06362	0.07455	-1.20985	-0.91641
Motivation_LevelLow	-1.10244	0.09385	-1.28671	-0.91783
Parental_InvolvementLow	-1.98077	0.09353	-2.16366	-1.79720
Access_to_ResourcesLow	-2.09227	0.09269	-2.27361	-1.91063

Table 4

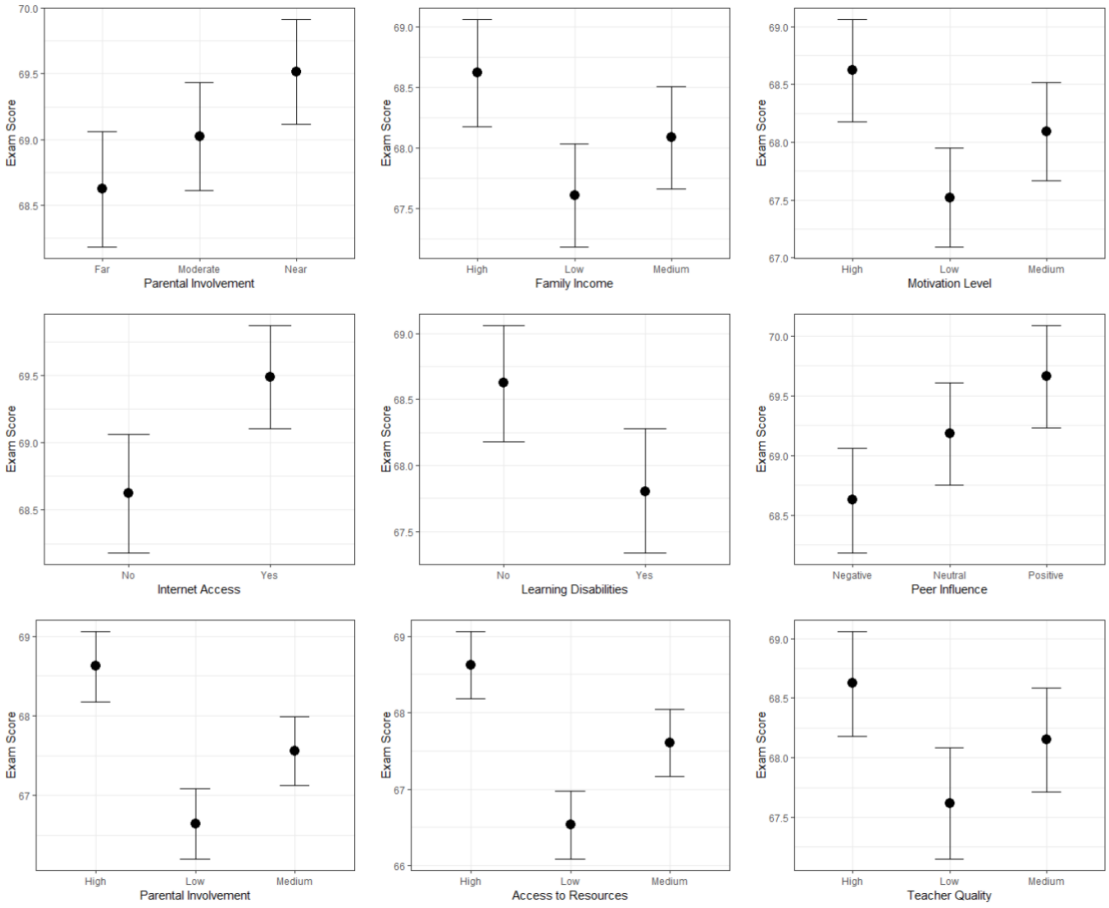
R2D2 Prior Full Predictors Regression Results

Effect	Estimate	Est.Error	l-95% CI	u-95% CI
Intercept	42.15729	0.413969	41.34619	42.96524
Peer_InfluencePositive	1.018999	0.087336	0.847494	1.190095
Distance_from_HomeNear	0.853758	0.110571	0.635924	1.070033
Internet_AccessYes	0.840993	0.124312	0.597892	1.084254
Peer_InfluenceNeutral	0.534876	0.087451	0.363617	0.705637
Extracurricular_ActivitiesYes	0.522625	0.065743	0.394266	0.651685
Tutoring_Sessions	0.505948	0.026026	0.454647	0.556684
Parental_Education_LevelPostgraduate	0.465368	0.092832	0.283093	0.648035
Distance_from_HomeModerate	0.356787	0.119217	0.12168	0.589611
Hours_Studied	0.293012	0.005391	0.282528	0.303532

Attendance	0.196782	0.002781	0.191307	0.202197
Physical_Activity	0.165028	0.031441	0.102471	0.22661
Previous_Scores	0.048171	0.002238	0.043748	0.052559
School_TypePublic	0.007143	0.061891	-0.11622	0.133623
Sleep_Hours	0.000169	0.020222	-0.04065	0.040791
GenderMale	-0.02473	0.059102	-0.147	0.089014
Teacher_QualityMedium	-0.45708	0.072789	-0.60085	-0.31346
Parental_Education_LevelHighSchool	-0.50609	0.074996	-0.65337	-0.3593
Motivation_LevelMedium	-0.51038	0.08527	-0.67696	-0.34301
Family_IncomeMedium	-0.51537	0.088986	-0.69043	-0.34203
Learning_DisabilitiesYes	-0.80354	0.103699	-1.00772	-0.60068
Teacher_QualityLow	-0.97672	0.115841	-1.20223	-0.75083
Family_IncomeLow	-0.99466	0.089403	-1.16947	-0.82173
Access_to_ResourcesMedium	-1.00955	0.074088	-1.15473	-0.86506
Parental_InvolvementMedium	-1.05179	0.074798	-1.19931	-0.90656
Motivation_LevelLow	-1.08185	0.093469	-1.26355	-0.89866
Parental_InvolvementLow	-1.96472	0.093139	-2.1476	-1.78256
Access_to_ResourcesLow	-2.07537	0.093213	-2.25774	-1.89354

Figure 4

Conditional Effects of Key Variables



Note. Error bars represent 95% credible intervals.

Discussion

Most surprisingly, the preliminary analysis revealed no relationship between family income and any of the other predictors hypothesized to be potential mediators due to the role of socioeconomic status in determining access to resources like tutoring, high quality teachers, and private schooling. This may be a limitation of the synthetic dataset in that

these real-world relationships are not captured in the artificial generative process responsible for creating the dataset used in the present paper. Regardless, these analyses suggest no need for a hierarchical regression and indicate that a flat model is acceptable for further analysis.

As expected, low access to resources, low family income, low teacher quality, and learning disabilities are associated with significantly lower exam scores. However, very few of the other hypothesized factors were strong predictors of exam performance. It is clear that excluding several key predictors such as parental involvement and motivation from the hypothesized relevant fit harmed both the predictive and generative performance of that model, and the results of this analysis do not support the original hypothesis. It may be reasonable to fit another model using only the relevant predictors and exclude those predictors found to have negligible effects from future analyses.

There are a number of potential avenues for future research. One interesting possibility is to investigate parental factors and assess for potential relationships between parental education, parental involvement, and student motivation. This is especially important as low motivation was identified as a key predictor of low exam scores, and understanding the dynamics behind student motivation may provide an avenue to address motivation issues in educational settings.

Positive peer support was identified as the top predictor of high exam scores, and extracurricular participation had a positive, though diminished, effect as well. It may be worthwhile to investigate social correlates of exam performance and determine what features of peer relationships and a student’s social space promote high academic achievement.

References

Ben-Shakhar, G., & Sinai, Y. (1991). Gender Differences in Multiple-Choice Tests: The Role of Differential Guessing Tendencies. *Journal of Educational Measurement*, *28*(1), 23–35. doi:10.1111/j.1745-3984.1991.tb00341.x

Bolger, N., & Kellaghan, T. (1990). Method of Measurement and Gender Differences in Scholastic Achievement. *Journal of Educational Measurement*, *27*(2), 165–174. doi:10.1111/j.1745-3984.1990.tb00740.x

Curry, A. H. (2008). Increasing student test scores: A study of if parent involvement, initiated by NCLB, affects student test scores [ProQuest Information & Learning]. In *Dissertation Abstracts International Section A: Humanities and Social Sciences* (Vol. 68, Issue 9–A, p. 3676).

Dixon-Roman, E., Everson, H., & Mcardle, J. (2013). Race, Poverty and SAT Scores: Modeling the Influences of Family Income on Black and White High School Students' SAT Performance. *Teachers College Record*, *115*. doi:10.1177/016146811311500406

Klein, S. P., Jovanovic, J., Stecher, B. M., McCaffrey, D., Shavelson, R. J., Haertel, E., Solano-Flores, G., & Comfort, K. (1997). Gender and Racial/Ethnic Differences on Performance Assessments in Science. Educational Evaluation and Policy Analysis, 19(2), 83-97. https://doi-org.libproxy.rpi.edu/10.3102/01623737019002083

Sackett, P. R., Kuncel, N. R., Arneson, J. J., Cooper, S. R., & Waters, S. D. (2009). Does socioeconomic status explain the relationship between admissions tests and post-secondary academic performance?. *Psychological bulletin*, *135*(1), 1–22. https://doi.org/10.1037/a0013978

Yu, Q. and Li, B., 2022. Statistical Methods for Mediation, Confounding and Moderation Analysis Using R and SAS. *Chapman and Hall/CRC.* ISBN 9780367365479. Link