math_prediction

Firstly, we were interested in identifying any covariates are associated with each other or not. There are ten categorical potential predictors for which we first tested for chi-squared test for each pair of two variables. However, there wasn't enough evidence suggesting association between any two categorical variables.

Table 1: Chi-Squared Test: Top 2 results (NS)

statistic	p.value	group
15.206	0.0553	ethnic_group:wkly_study_hours
9.385	0.0947	parent_educ:transport_means

The remaining continuous variable, `nr_siblings`, that measure number of siblings was tested against all the categorical variables to test for number of sibling differs between different categorical variables. Number of siblings were different between students who were the first child versus those who were not, also between students with different number of weekly study hours. We needed to account for both `is_first_child` and `wkly_study_hours`, if number of siblings were to be included the model.

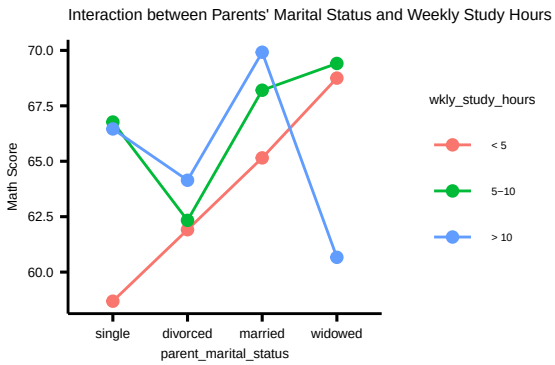
Table 2: ANOVA: Number of Siblings v/s Other Covariates (<0.05)

term	df	statistic	p.value
is_first_child	1	16.366	5.68e-05
wkly_study_hours	2	3.024	4.91e-02

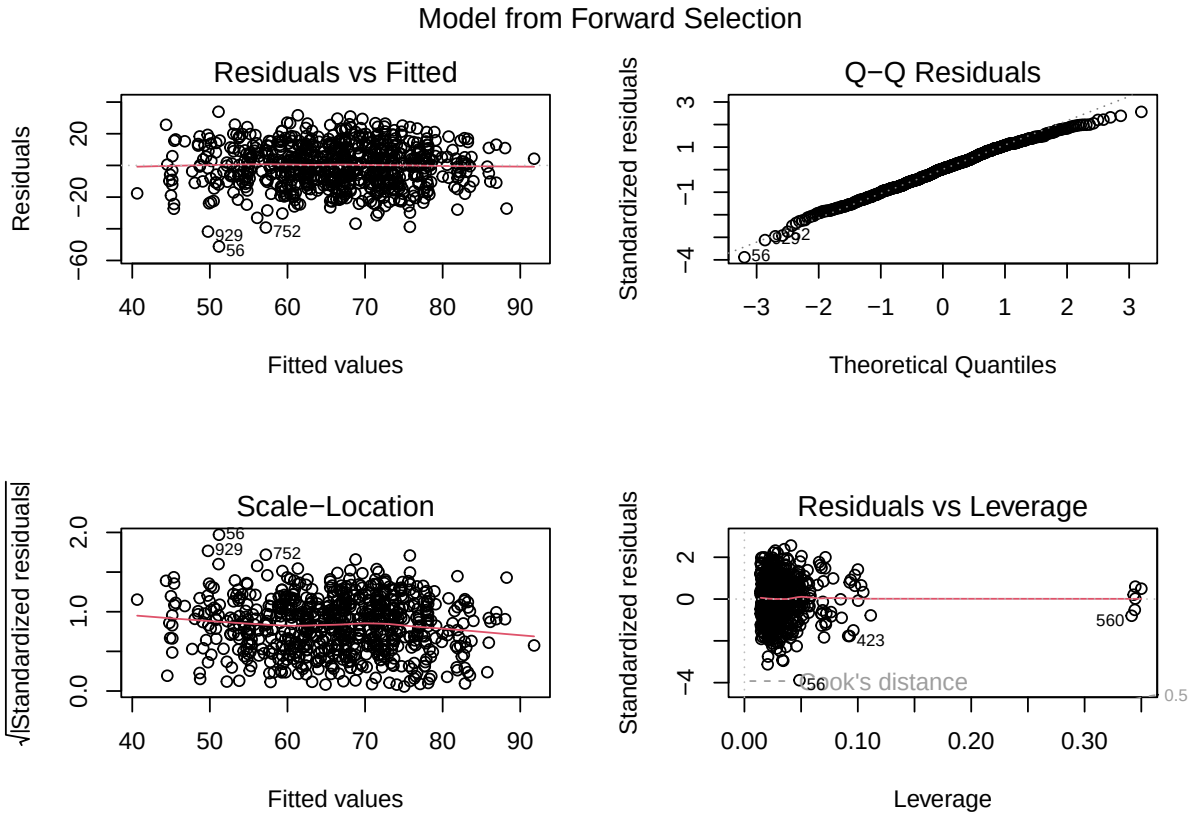
To model for math, reading and writing scores of students, we utilized forward selection and step-wise regression, along with test based procedures and LASSO to find the "best fitting" models. Among the shortlisted models, we ran 10-fold cross validation to finalize on the better performing model.

Math Score Prediction

With forward selection, we obtained seven significant predictors. Subsequently,we visualized math score between pairs of categorical variables. Among the pairs where the mean values score are associated within sub-categories, we tested for interaction, and found that weekly study hours and parental marital status has significant combined effect.

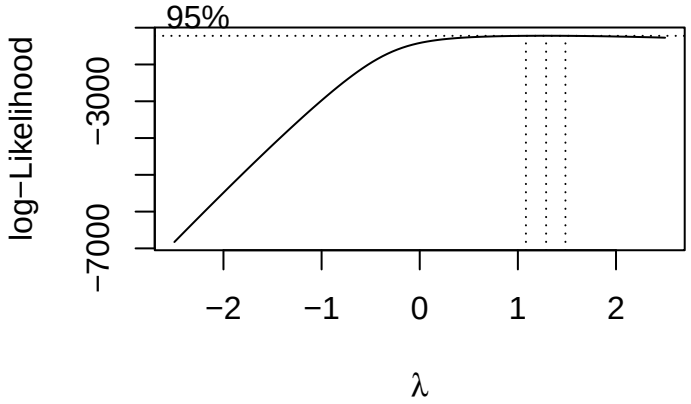


The diagnostic plots and test (included in corresponding R code) suggests that the model follows homoscedasticity, mean of 0 for residuals, and there are no outliers or any influential points. However, there is slight deviation from normality (as shown by the formal test below).

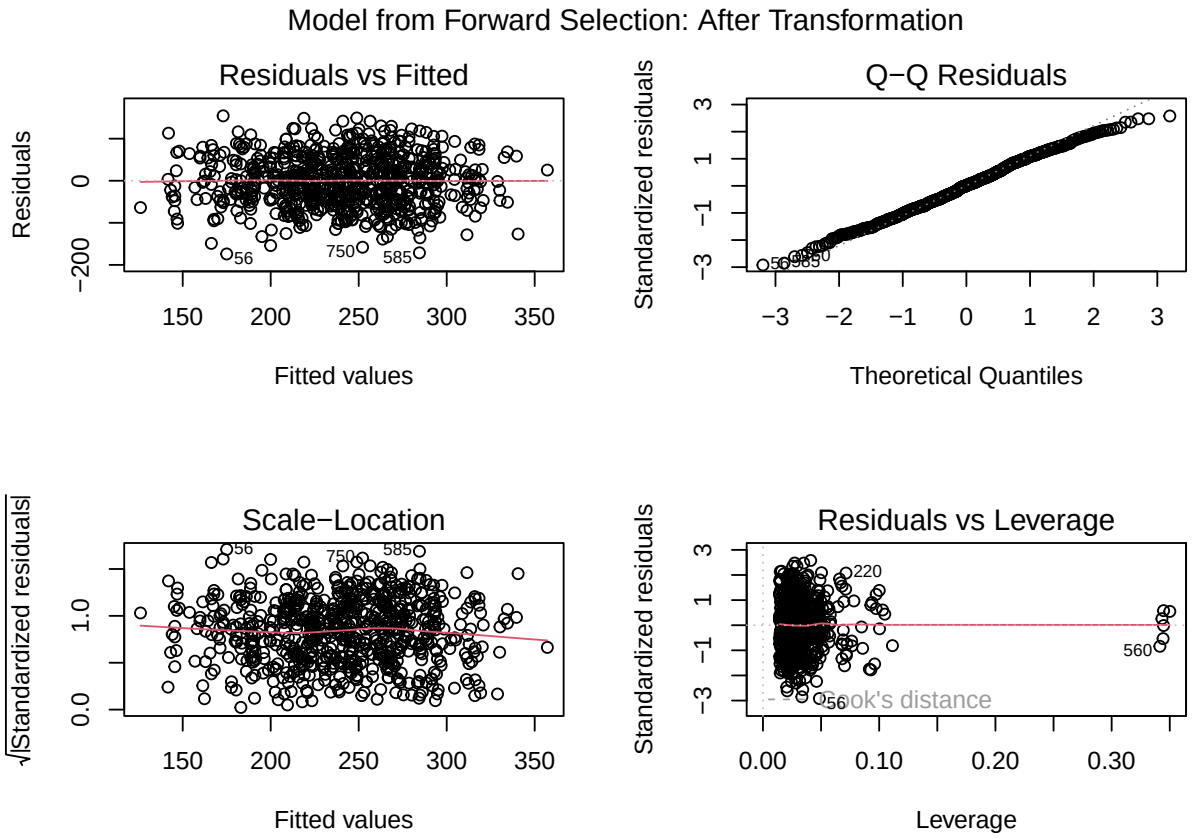


statistic	p.value	method
0.9944663	0.010165	Shapiro-Wilk normality test

We ran boxcox transformation to find the optimal transformation needed for residual for math score to follow the assumptions.



Upon transformation of $(Math\ Score + 1)^{1.3}$, the residuals follow the normality (Shapiro-Wilk test, p-value = 0.055), homoscedascity and mean 0 assumptions.



statistic	p.value	method
0.9958945	0.0558016	Shapiro-Wilk normality test

The final model is as below:

Table 5: Forward Selection: Coefficients

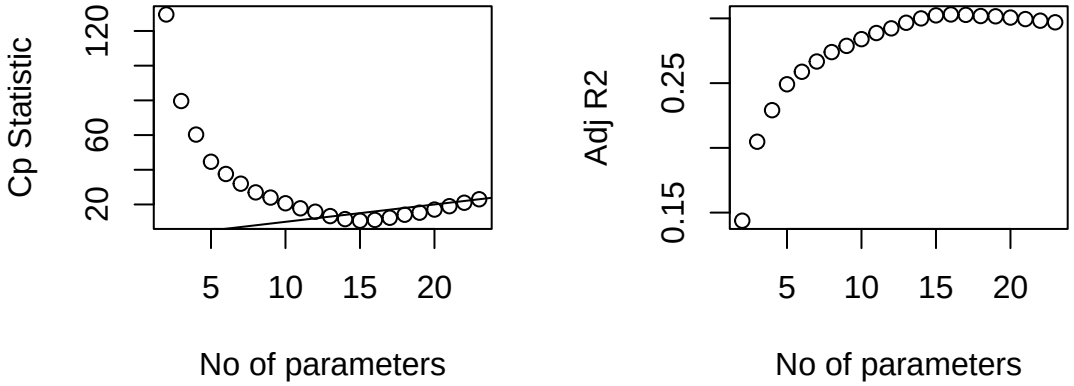
term	estimate	std.error	statistic	p.value
(Intercept)	205.59	13.33	15.43	2.19e-46
lunch_typefree/reduced	-54.83	4.83	-11.35	1.67e-27
test_prepnone	-27.81	4.83	-5.76	1.26e-08
gendermale	22.21	4.65	4.77	2.20e-06
ethnic_groupgroup B	3.07	9.36	0.33	7.43e-01
ethnic_groupgroup C	4.74	8.82	0.54	5.91e-01
ethnic_groupgroup D	22.26	8.93	2.49	1.29e-02
ethnic_groupgroup E	47.96	9.84	4.87	1.35e-06
parent_educassociate's degree	20.34	7.33	2.78	5.65e-03
parent_educbachelor's degree	24.75	8.55	2.90	3.90e-03
parent_educhigh school	-3.26	7.41	-0.44	6.60e-01
parent_educmaster's degree	31.81	10.63	2.99	2.87e-03
parent_educsome college	19.11	7.35	2.60	9.49e-03
parent_marital_statusdivorced	18.72	14.28	1.31	1.90e-01
parent_marital_statusmarried	35.40	10.47	3.38	7.58e-04
parent_marital_statuswidowed	64.31	36.63	1.76	7.96e-02

term	estimate	std.error	statistic	p.value
wkly_study_hours5-10	34.02	10.95	3.11	1.98e-03
wkly_study_hours> 10	47.31	14.42	3.28	1.09e-03
parent_marital_statusdivorced:wkly_study_hours5-10	-31.65	17.32	-1.83	6.81e-02
parent_marital_statusmarried:wkly_study_hours5-10	-25.90	12.98	-1.99	4.65e-02
parent_marital_statuswidowed:wkly_study_hours5-10	-37.22	41.33	-0.90	3.68e-01
parent_marital_statusdivorced:wkly_study_hours> 10	-63.80	23.29	-2.74	6.31e-03
parent_marital_statusmarried:wkly_study_hours> 10	-38.25	17.07	-2.24	2.54e-02
parent_marital_statuswidowed:wkly_study_hours> 10	-76.96	52.09	-1.48	1.40e-01

Table 6: Forward Selection: Model Summary

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.31	0.29	61	14	3.2e-43	23	-3964.719	7979	8094	2593221	695	719

Next, we applied step-wise regression method to select a model. The initial model had five covariates (weekly study hours, test preparation, lunch type, gender and ethnic group) that can potentially be related to academic performance. New covariates were added, insignificant covariates were dropped in the following steps to obtain a model with all significant predictors. Both forward and Stepwise regression gives us the same model.



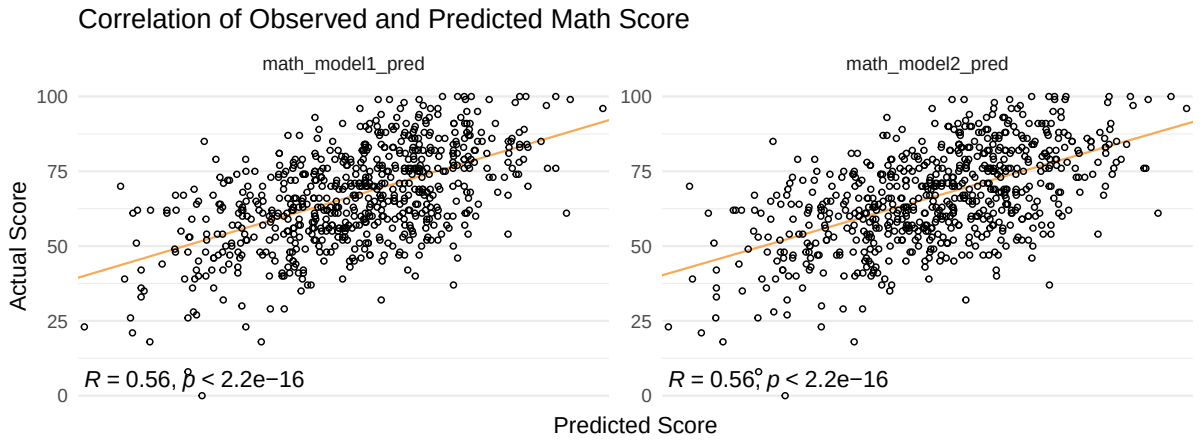
The models were further investigated using test based criteria. Based on the Mallow's Cp and adjusted R^2, an optimal model for math score will should 15-17 main effect parameters. This aligns with our model selected from forward and backward model selection. LASSO suggests taking number of siblings into account. So, we decided to cross-validate and compare performances between model with and without number of sibling as a modifier.

term	step	estimate	lambda	dev.ratio
(Intercept)	1	65.269	0.05	0.004
nr_siblings	1	0.658	0.05	0.004

Among the two, the model without number of sibling has lower root mean square error, higher R^2 has similar predictive ability as the more complex model. Therefore, it is the better fitting model.

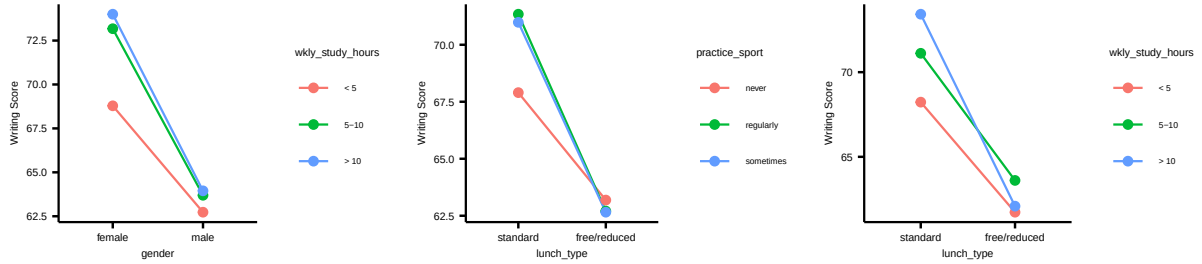
Table 8: Performance matrices of the 2 Models in Predicting (Math Score + 1)^1.3

model_id	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
w/o no. of sibling	62.188	0.283	51.084	3.817	0.060	3.556
w/ no. of sibling	62.479	0.277	51.456	4.739	0.085	3.737

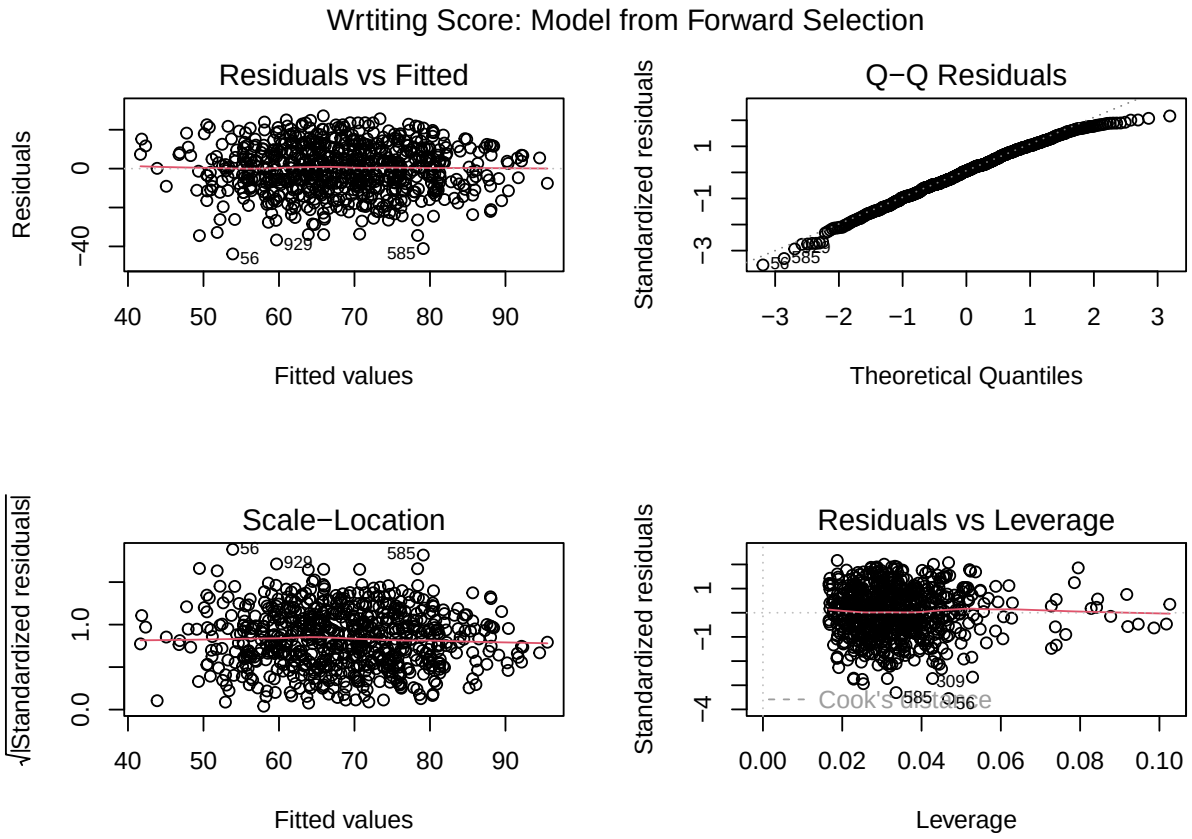


Writing Score Prediction

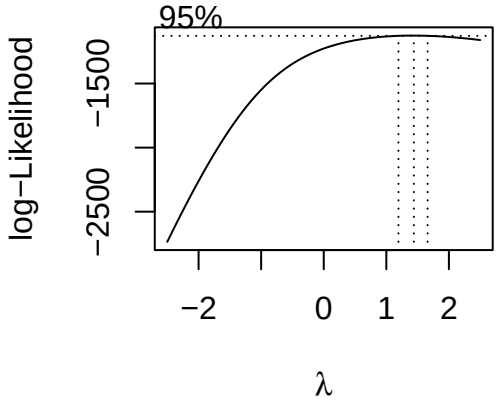
Using forward selection, we obtained seven significant covariates to predict writing score, namely test preparation, gender, lunch type, parent education, ethnic group, parent marital status and sports practice status. We visualized writing score between pairs of categorical variables. Among the pairs where the mean values score are associated within sub-categories, we tested for interaction, and found that signigicant interaction coefficient for `gender:wkly_study_hours` and `lunch_type:wkly_study_hours`. Similarly, we the same model was selected for upon step-wise regression selection.



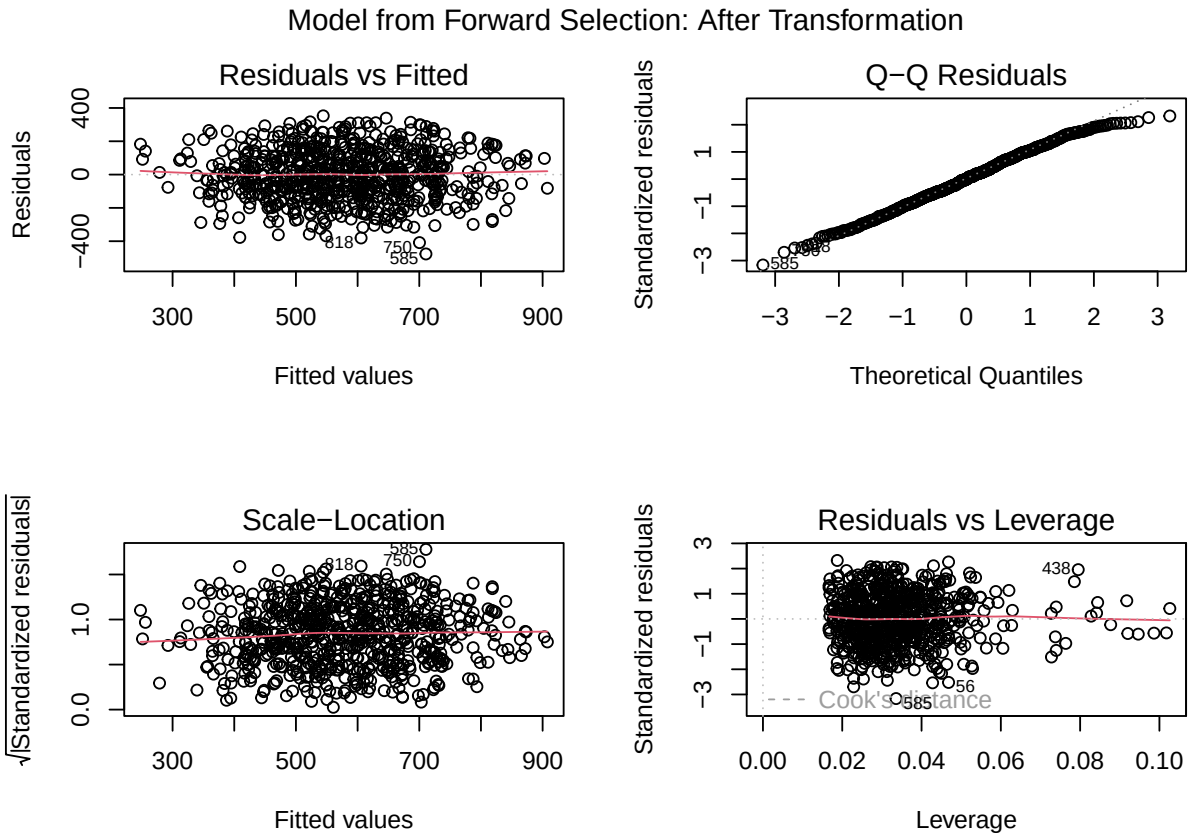
Similar to math model, the residuals were followed homoscedascticity, mean zero assumptions, but deviated from normality assumptions. Therefore, box-cox transformation was used to identify the optimal transformation, and model was re-fitted using $\sqrt{(Writing\ Score)^3}$.



statistic	p.value	method
0.9897292	7.61e-05	Shapiro-Wilk normality test



Upon transformation of $\sqrt{(Writing\ Score)^3}$, the normality, homoscedascity and mean 0 assumptions for residuals are met.



The final model for writing score is:

Table 10: Forward Selection: Coefficients

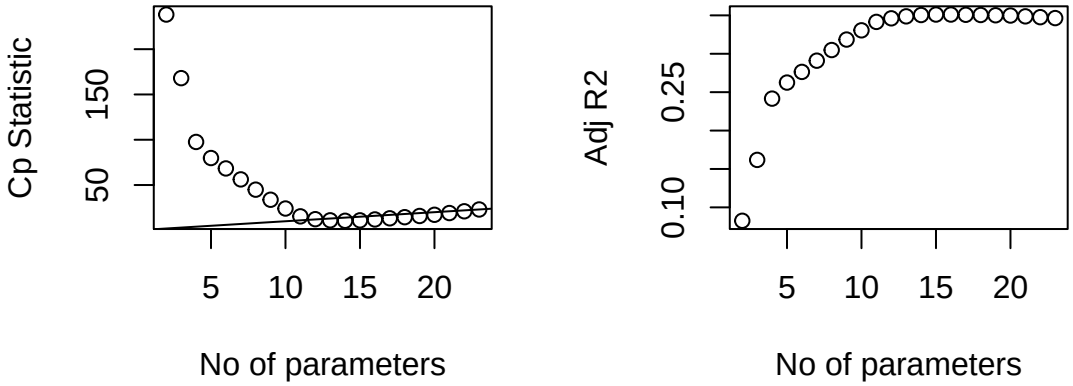
term	estimate	std.error	statistic	p.value
(Intercept)	568.09	35.27	16.11	1.00e-49
test_prepnone	-117.48	12.25	-9.59	1.61e-20
gendermale	-68.47	22.35	-3.06	2.27e-03
lunch_typefree/reduced	-97.08	22.94	-4.23	2.64e-05
parent_educassociate's degree	68.07	18.45	3.69	2.43e-04
parent_educbachelor's degree	99.07	21.40	4.63	4.39e-06
parent_educhigh school	-7.95	18.67	-0.43	6.71e-01
parent_educmaster's degree	151.74	27.12	5.60	3.19e-08
parent_educsome college	64.77	18.67	3.47	5.56e-04
ethnic_groupgroup B	-11.32	23.63	-0.48	6.32e-01
ethnic_groupgroup C	7.60	22.30	0.34	7.33e-01
ethnic_groupgroup D	65.57	22.66	2.89	3.93e-03
ethnic_groupgroup E	64.00	24.99	2.56	1.07e-02
parent_marital_statusdivorced	-25.76	18.76	-1.37	1.70e-01
parent_marital_statusmarried	46.44	14.25	3.26	1.18e-03
parent_marital_statuswidowed	63.62	39.53	1.61	1.08e-01
practice_sportregularly	44.60	19.25	2.32	2.08e-02
practice_sportsometimes	38.46	18.61	2.07	3.92e-02
wkly_study_hours5-10	58.65	21.93	2.68	7.65e-03
wkly_study_hours> 10	80.58	29.73	2.71	6.88e-03
gendermale:wkly_study_hours5-10	-63.27	27.40	-2.31	2.13e-02
gendermale:wkly_study_hours> 10	-67.54	36.09	-1.87	6.17e-02

term	estimate	std.error	statistic	p.value
lunch_typefree/reduced:wkly_study_hours5-10	-6.36	28.40	-0.22	8.23e-01
lunch_typefree/reduced:wkly_study_hours> 10	-77.34	37.21	-2.08	3.80e-02

Table 11: Forward Selection: Model Summary

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.38	0.36	150	18	0	23	-4542.314	9135	9249	16022165	682	706

We applied step-wise regression method to select a model with the same initial five covariates as used for prediction model for math. New covariates were added, insignificant covariates were dropped in the following steps to obtain a model with all significant predictors. Both forward and Stepwise regression gives us the same model.



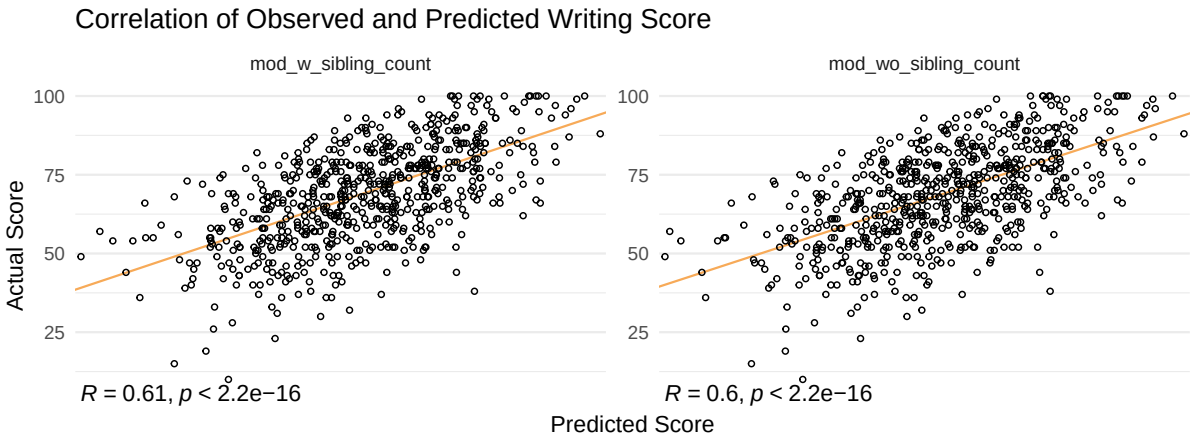
The models were further investigated using test based criteria. Based on the Mallow's Cp and adjusted R^2, an optimal model for math score will should 13-18 main effect parameters. This aligns with our model selected from forward and backward model selection. LASSO suggests taking number of siblings into account. So, we decided to cross-validate and compare performances between model with and without number of sibling as a modifier.

term	step	estimate	lambda	dev.ratio
(Intercept)	1	565.413	0.501	0.005
nr_siblings	1	8.268	0.501	0.005

Among the two, the model without number of sibling has lower root mean square error, higher R^2 has similar predictive ability as the more complex model. Therefore, it is the better fitting model.

Table 13: Performance matrices of the 2 Models in Predicting (Writing Score)^1.5

model_id	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
w/o no. of sibling	157.769	0.330	128.940	10.013	0.097	8.118
w/ no. of sibling	158.340	0.325	129.863	10.698	0.063	9.883


Correlation of Observed and Predicted Writing Score
mod_w_sibling_count
mod_wo_sibling_count
Actual Score
100
75
50
25
R = 0.61, p < 2.2e−16
R = 0.6, p < 2.2e−16
Predicted Score