

biostats_final_combined

Miao Fu

2023-12-07

Descriptive summary statistics for all variables

Two table with summary information on the descriptive statistics of all variables are listed below. The frequency and percentage of each categories in each categorical variable is listed out. For each numeric variable, the table includes values of mean, median, standard deviation, minimum, maximum, Q1 and Q3 values.

Categorical Variables

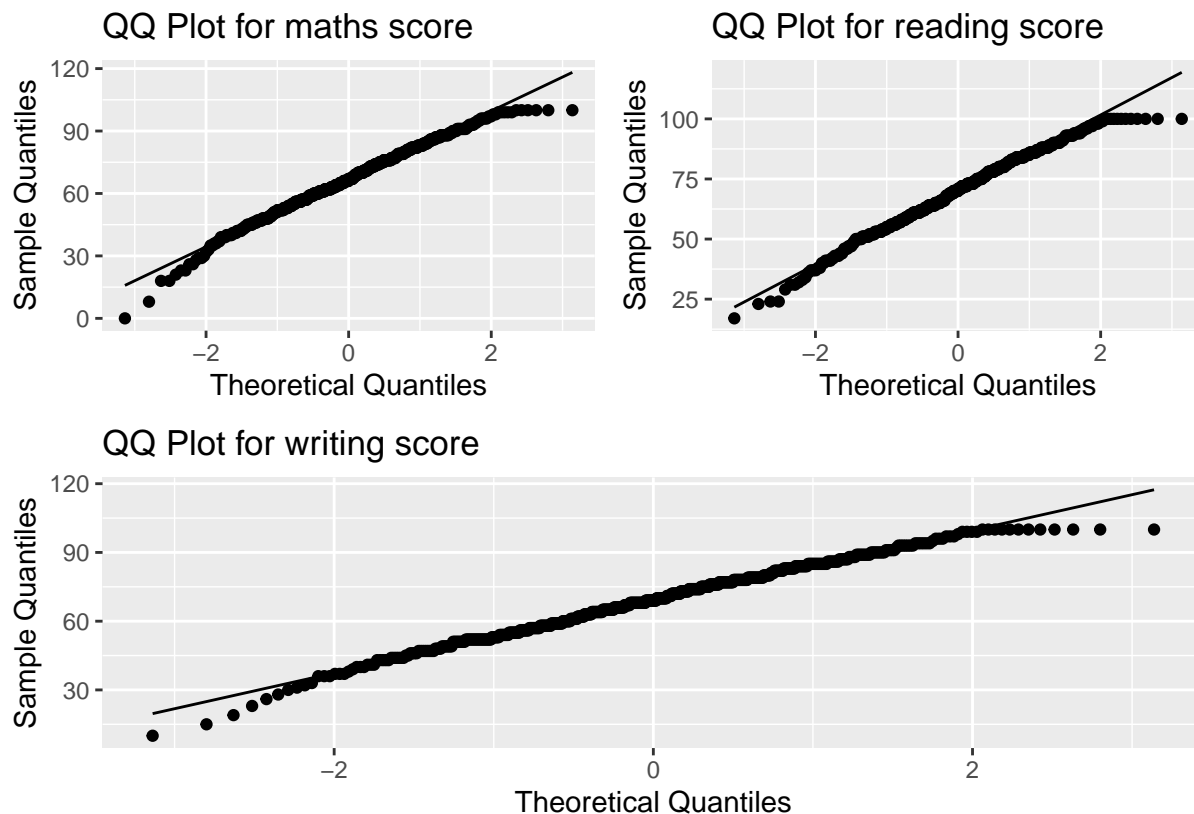
| variable | category | count | percent |
|-----------------------|--------------------|-------|-----------|
| gender | female | 315 | 53.662692 |
| gender | male | 272 | 46.337308 |
| ethnic_group | group A | 50 | 8.517888 |
| ethnic_group | group B | 123 | 20.954003 |
| ethnic_group | group C | 174 | 29.642249 |
| ethnic_group | group D | 155 | 26.405451 |
| ethnic_group | group E | 85 | 14.480409 |
| parent_educ | associate's degree | 128 | 21.805792 |
| parent_educ | bachelor's degree | 71 | 12.095400 |
| parent_educ | high school | 122 | 20.783646 |
| parent_educ | master's degree | 39 | 6.643952 |
| parent_educ | some college | 116 | 19.761499 |
| parent_educ | some high school | 111 | 18.909710 |
| lunch_type | free/reduced | 206 | 35.093697 |
| lunch_type | standard | 381 | 64.906303 |
| test_prep | completed | 208 | 35.434412 |
| test_prep | none | 379 | 64.565588 |
| parent_marital_status | divorced | 92 | 15.672913 |
| parent_marital_status | married | 343 | 58.432709 |
| parent_marital_status | single | 137 | 23.339012 |
| parent_marital_status | widowed | 15 | 2.555366 |
| practice_sport | never | 68 | 11.584327 |
| practice_sport | regularly | 218 | 37.137990 |
| practice_sport | sometimes | 301 | 51.277683 |
| is_first_child | no | 192 | 32.708688 |
| is_first_child | yes | 395 | 67.291312 |
| transport_means | private | 229 | 39.011925 |
| transport_means | school_bus | 358 | 60.988075 |
| wkly_study_hours | < 5 | 154 | 26.235094 |
| wkly_study_hours | > 10 | 104 | 17.717206 |
| wkly_study_hours | 5-10 | 329 | 56.047700 |

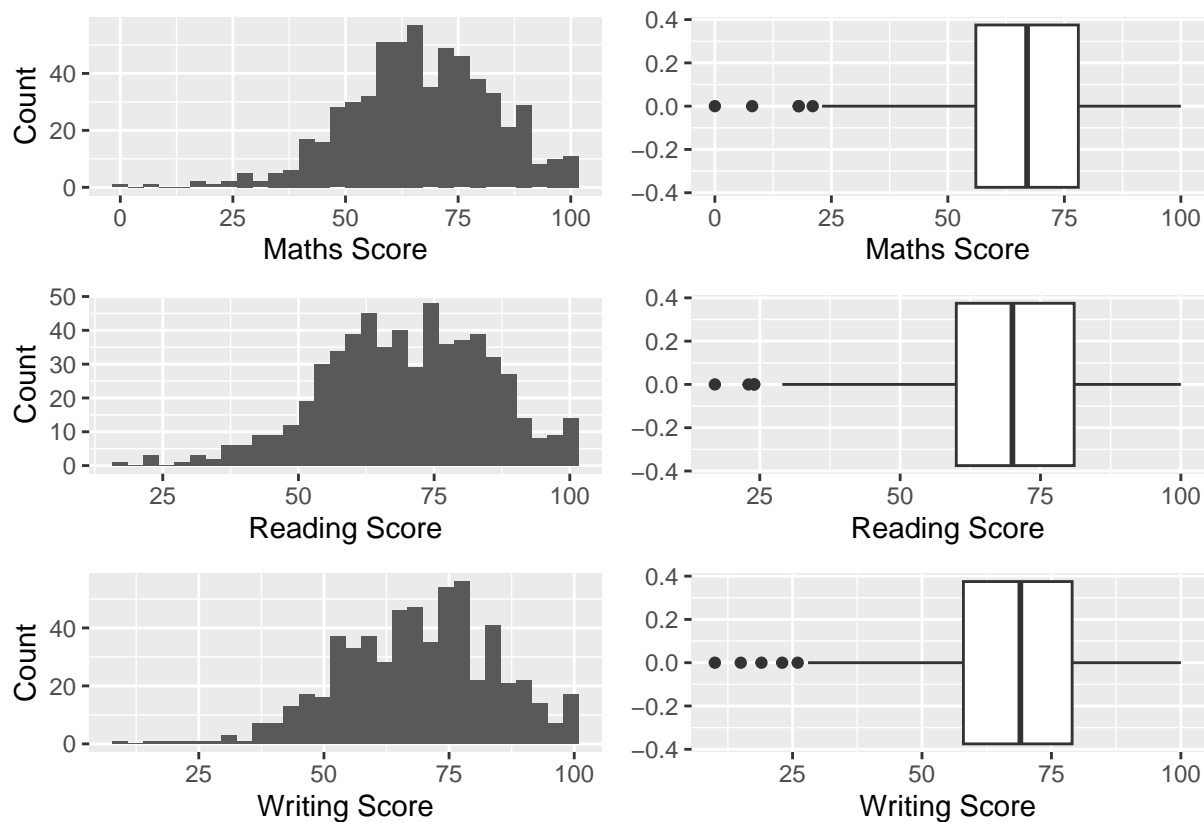
Numeric Variables

| variable | mean | median | sd | minimum | maximum | q1 | q3 |
|---------------|-----------|--------|-----------|---------|---------|----|----|
| nr_siblings | 2.139693 | 2 | 1.481712 | 0 | 7 | 1 | 3 |
| math_score | 66.676320 | 67 | 16.113744 | 0 | 100 | 56 | 78 |
| reading_score | 69.846678 | 70 | 15.166662 | 17 | 100 | 60 | 81 |
| writing_score | 68.901192 | 69 | 15.550000 | 10 | 100 | 58 | 79 |

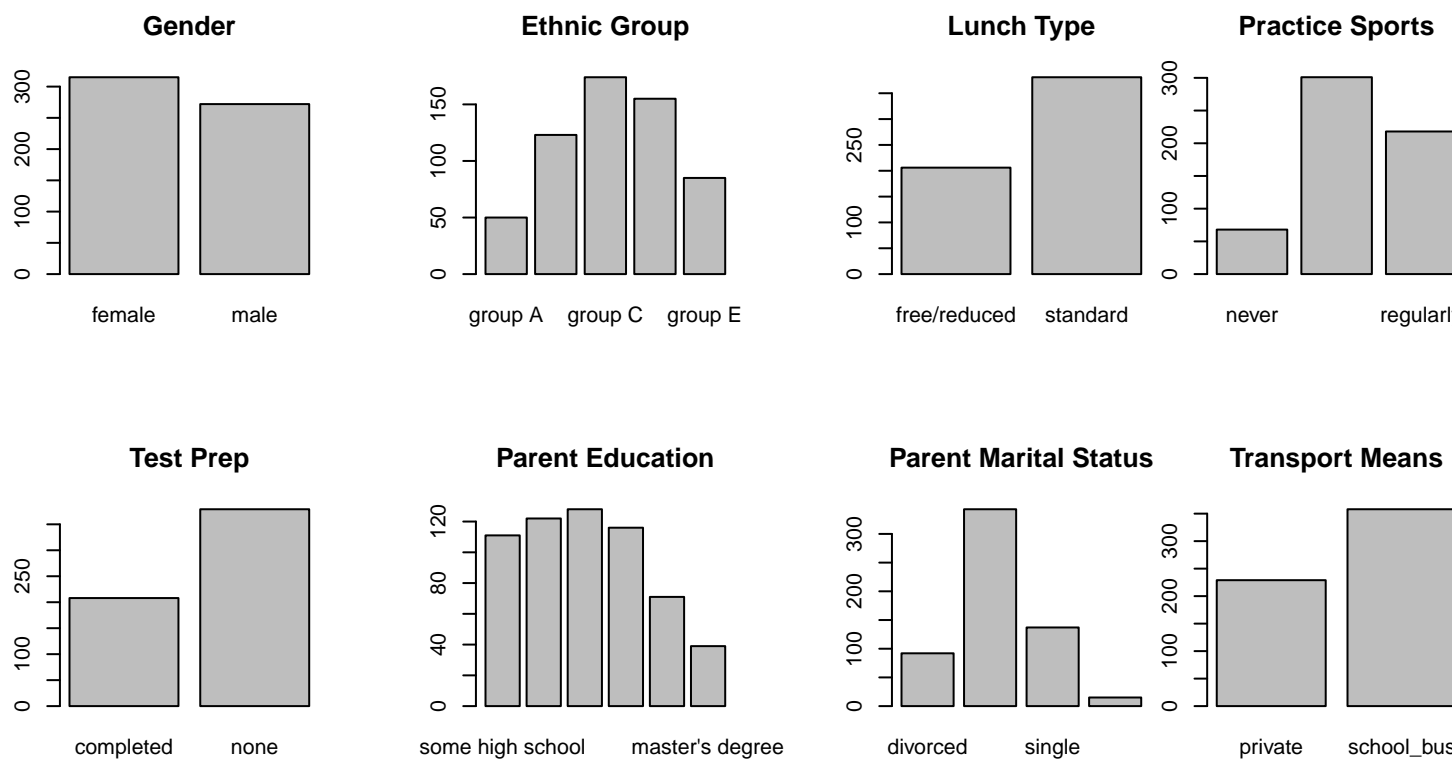
Exploration of outcome variables

The outcome of this study includes the following variables: maths scores, reading scores, and writing scores. QQplots of the outcome variables are created to explore the distribution of each score. QQplot compares the quantiles of the data against the quantiles of a normal distribution. According the plots, majority of the data points of all three scores follow the straight qqline, which indicates they follow the normal distribution. However, there are some deviations from the line on the two ends of the distribution, which indicates the distributions might have heavier tails than normal distribution. To further explore the distribution of outcomes, histograms and boxplots for the scores were incorporated. They are slightly left-tailed but mainly normally distributed. Thus, transformations are tested in the later section to further investigate the variables.





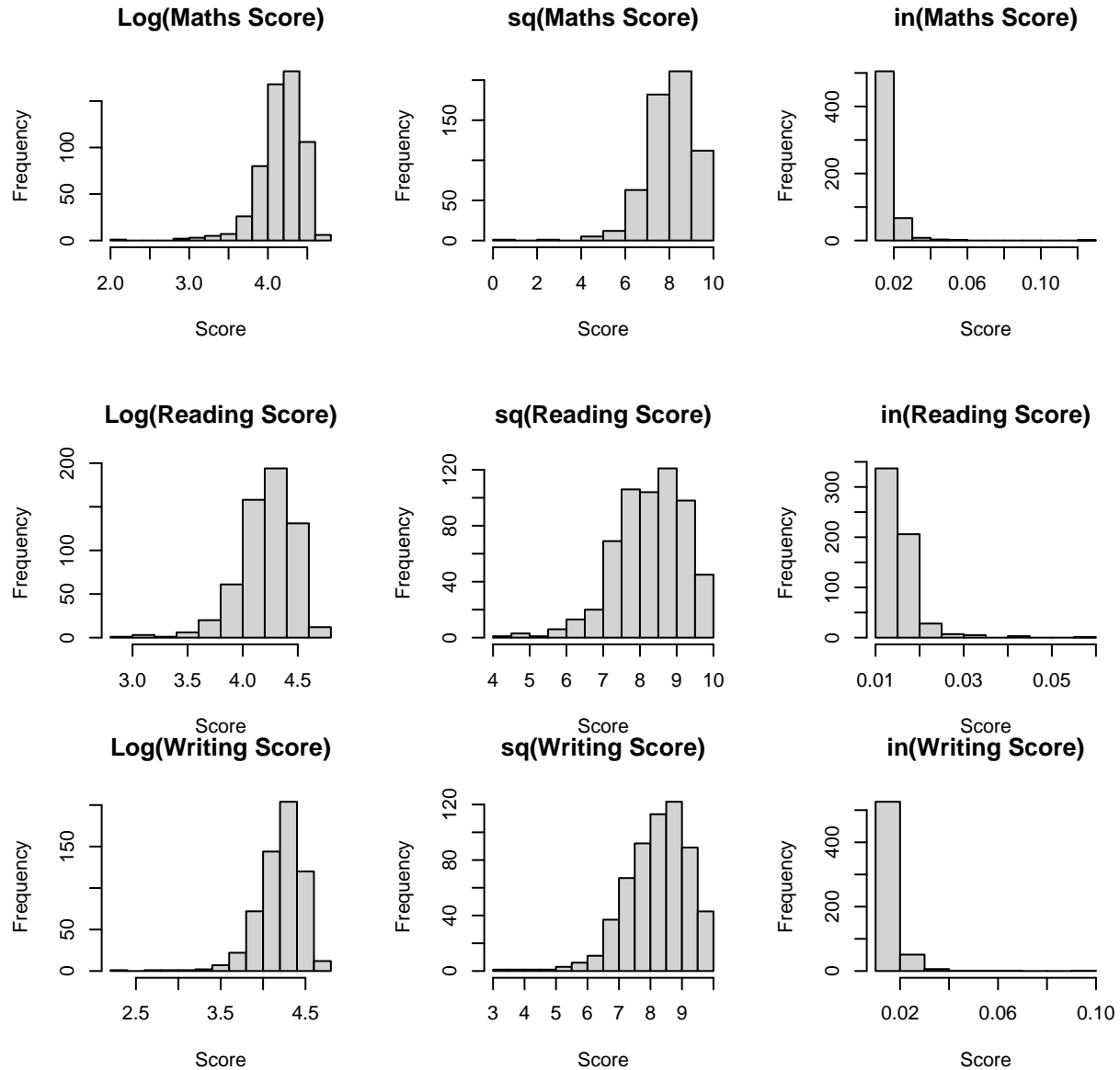
Exploration of predictor variables



From the distribution bar graphs, it was decided that no transformation is needed for any predictor variables.

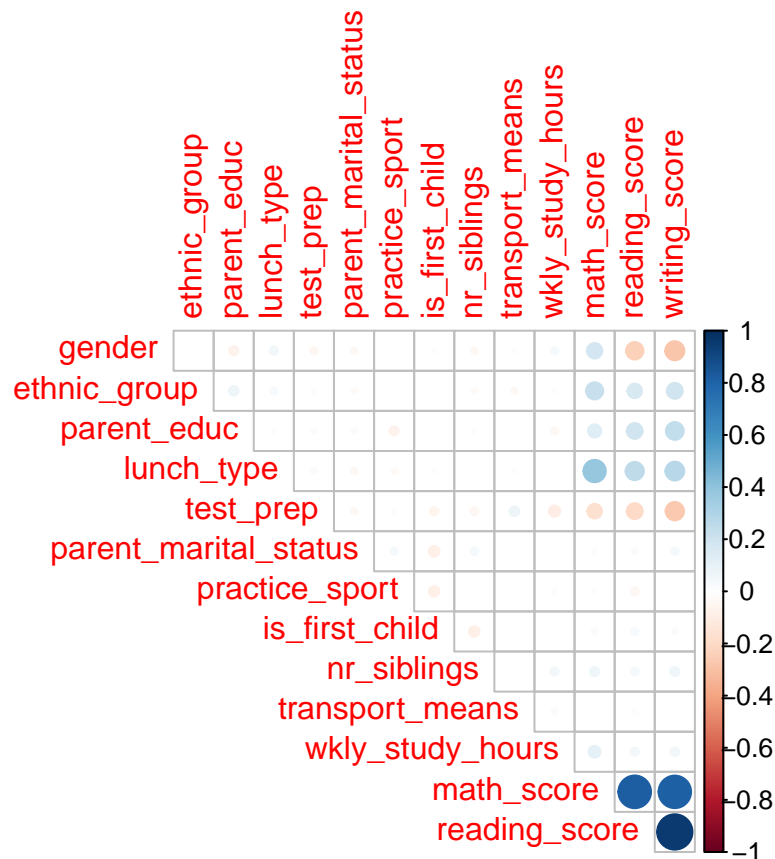
Admittedly, the distributions might not be good references for transformation because the variables are all categorical.

Potential transformations for outcome variables



Since the distribution of the three scores showed slightly left-tailed, three types of transformations were tested: 1) Natural logarithm 2) Square Root 3) Inverse. The resulting plots are plotted in histograms shown above. There is no apparent improvement on the distribution of the outcome through the three transformations. Thus, the original outcome data were chosen to be used in following statistical modeling steps.

Pairwise relationships



By plotting our the pairwise correlation between variables, there is apparent linearity among the three scores. Other correlation coefficients are relatively small, indicating weak linear relationship between the variables.

MLR `lm()`

MLR - Math

```
##
## Call:
## lm(formula = math_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep + parent_marital_status + practice_sport +
##     is_first_child + nr_siblings + transport_means + wkly_study_hours,
##     data = df_transformed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -48.916  -9.265   0.725  10.104  33.013
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    44.1006     3.6704  12.015  < 2e-16 ***
## gendermale       5.0855     1.1386   4.467  9.61e-06 ***
## ethnic_groupgroup B   -0.1788     2.3136  -0.077  0.93841
## ethnic_groupgroup C   -0.2089     2.2149  -0.094  0.92489
```

```
## ethnic_groupgroup D          3.6247      2.2286      1.626      0.10441
## ethnic_groupgroup E         11.1752      2.4434      4.574      5.90e-06 ***
## parent_educhigh school      -0.3235      1.8015     -0.180      0.85757
## parent_educassociate's degree 4.9058      1.7728      2.767      0.00584 **
## parent_educsome college      3.1933      1.8163      1.758      0.07927 .
## parent_educbachelor's degree 6.6652      2.0763      3.210      0.00140 **
## parent_educmaster's degree   6.8096      2.5417      2.679      0.00760 **
## lunch_typestandard          12.3539      1.1771     10.495     < 2e-16 ***
## test_preptime               -4.7717      1.2007     -3.974     7.99e-05 ***
## parent_marital_statusmarried 5.4805      1.6170      3.389      0.00075 ***
## parent_marital_statussingle 2.1682      1.8454      1.175      0.24053
## parent_marital_statuswidowed 7.7944      3.8119      2.045      0.04134 *
## practice_sportsometimes      1.5255      1.8439      0.827      0.40838
## practice_sportregularly      1.6701      1.9046      0.877      0.38092
## is_first_childyes           1.1303      1.2125      0.932      0.35162
## nr_siblings                 0.7403      0.3844      1.926      0.05461 .
## transport_meansschool_bus    -0.4319      1.1629     -0.371      0.71050
## wkly_study_hours5-10         3.5394      1.3429      2.636      0.00863 **
## wkly_study_hours> 10         3.0384      1.7540      1.732      0.08378 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.52 on 564 degrees of freedom
## Multiple R-squared:  0.3221, Adjusted R-squared:  0.2956
## F-statistic: 12.18 on 22 and 564 DF,  p-value: < 2.2e-16
```

Coefficients and Significance Levels:

- Intercept (44.1006): The expected value of math_score when all other predictors are at their reference level or zero.
- gendermale (5.0855, $p < 0.001$): Being male is associated with an average increase of 5.0855 points in math_score compared to females, holding all else constant. This is statistically significant.
- ethnic_group: Only ethnic_groupgroup E (11.1752, $p < 0.001$) is significant, suggesting students in this group score higher in math compared to the reference group.
- parent_educ: The associate's degree (4.9058, $p = 0.00584$), bachelor's degree (6.6652, $p = 0.00140$), and master's degree (6.8096, $p = 0.00760$) are significant and associated with higher math scores compared to the reference category.
- lunch_typestandard (12.3539, $p < 0.001$): Students with standard lunch type score significantly higher.
- test_preptime (-4.7717, $p < 0.001$): Not participating in test preparation is associated with lower math scores.
- parent_marital_status: Married (5.4805, $p = 0.00075$) and Widowed (7.7944, $p = 0.04134$) are associated with higher scores.
- practice_sport: Not significant.
- is_first_childyes: Not significant.
- nr_siblings (0.7403, $p = 0.05461$): A borderline significant positive association with math scores.
- transport_meansschool_bus: Not significant.
- wkly_study_hours: Studying 5-10 hours (3.5394, $p = 0.00863$) shows a significant positive effect.

Residuals:

The spread of residuals suggests the errors are somewhat symmetrically distributed around the predicted values, which is a good sign for linear regression assumptions.

Model Fit:

Residual Standard Error (13.52): Indicates the average difference between the observed values and the values predicted by the model.

R-squared:

- Multiple R-squared (0.3221): About 32.21% of the variability in math_score is explained by the model.
- Adjusted R-squared (0.2956): Adjusts the R-squared for the number of predictors, a better measure for models with multiple predictors.

Statistic & p-value

F-statistic (12.18) and p-value ($< 2.2e-16$): The model is statistically significant, meaning it performs better than a model with no predictors.

MLR - reading

```
##
## Call:
## lm(formula = reading_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep + parent_marital_status + practice_sport +
##     is_first_child + nr_siblings + transport_means + wkly_study_hours,
##     data = df_transformed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -41.754  -8.793   0.635   9.118  30.513
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      60.8028     3.5826  16.972 < 2e-16 ***
## gendermale       -7.6725     1.1114  -6.904 1.37e-11 ***
## ethnic_groupgroup B    -1.4287     2.2582  -0.633 0.527220
## ethnic_groupgroup C    -0.8558     2.1619  -0.396 0.692355
## ethnic_groupgroup D     2.5663     2.1753   1.180 0.238600
## ethnic_groupgroup E     5.9165     2.3850   2.481 0.013402 *
## parent_educhigh school  -0.5785     1.7584  -0.329 0.742303
## parent_educassociate's degree  4.7948     1.7305   2.771 0.005776 **
## parent_educsome college   2.4082     1.7729   1.358 0.174896
## parent_educbachelor's degree  7.3496     2.0266   3.627 0.000313 ***
## parent_educmaster's degree   8.7149     2.4809   3.513 0.000479 ***
## lunch_typestandard    8.4374     1.1489   7.344 7.31e-13 ***
## test_preptime       -6.2822     1.1720  -5.360 1.21e-07 ***
## parent_marital_statusmarried  5.2439     1.5783   3.322 0.000950 ***
## parent_marital_statussingle  1.9235     1.8013   1.068 0.286046
## parent_marital_statuswidowed  5.5863     3.7208   1.501 0.133813
## practice_sportsometimes    0.6757     1.7998   0.375 0.707488
## practice_sportregularly   -0.6843     1.8590  -0.368 0.712923
## is_first_childyes        1.3046     1.1835   1.102 0.270780
## nr_siblings           0.3882     0.3752   1.035 0.301309
## transport_meansschool_bus   0.2841     1.1351   0.250 0.802472
## wkly_study_hours5-10      2.6835     1.3108   2.047 0.041104 *
## wkly_study_hours> 10      1.0970     1.7121   0.641 0.521971
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.2 on 564 degrees of freedom
## Multiple R-squared:  0.2709, Adjusted R-squared:  0.2425
## F-statistic: 9.527 on 22 and 564 DF,  p-value: < 2.2e-16
```

Coefficients and Significance Levels:

- **Intercept (60.8028)**: The expected value of `reading_score` when all other predictors are at their reference level or zero.
- **gendermale (-7.6725, p < 0.001)**: Being male is associated with an average decrease of 7.6725 points in `reading_score` compared to females, holding all else constant. This is statistically significant.
- **ethnic_group**: Only `ethnic_groupgroup E` (5.9165, p = 0.013402) is significant, suggesting students in this group score higher in reading compared to the reference group.
- **parent_educ**: The associate's degree (4.7948, p = 0.005776), bachelor's degree (7.3496, p = 0.000313), and master's degree (8.7149, p = 0.000479) are significant and associated with higher reading scores compared to the reference category.
- **lunch_typedstandard (8.4374, p < 0.001)**: Students with standard lunch type score significantly higher.
- **test_preptime (-6.2822, p < 0.001)**: Not participating in test preparation is associated with lower reading scores.
- **parent_marital_statusmarried (5.2439, p = 0.000950)**: Children of married parents score higher.
- **practice_sport**: Not significant.
- **is_first_childyes**: Not significant.
- **nr_siblings (0.3882, p = 0.301309)**: No significant association with reading scores.
- **transport_meansschool_bus**: Not significant.
- **wkly_study_hours**: Studying 5-10 hours (2.6835, p = 0.041104) shows a significant positive effect.

Residuals:

The spread of residuals suggests the errors are somewhat symmetrically distributed around the predicted values, which is a good sign for linear regression assumptions.

Model Fit:

- **Residual Standard Error (13.2)**: Indicates the average difference between the observed values and the values predicted by the model.

R-squared:

- **Multiple R-squared (0.2709)**: About 27.09% of the variability in `reading_score` is explained by the model.
- **Adjusted R-squared (0.2425)**: Adjusts the R-squared for the number of predictors, a better measure for models with multiple predictors.

Statistic & p-value

- **F-statistic (9.527)** and **p-value (< 2.2e-16)**: The model is statistically significant, meaning it performs better than a model with no predictors.

MLR - writing

```
##
## Call:
## lm(formula = writing_score ~ gender + ethnic_group + parent_educ +
```



```
## lunch_type + test_prep + parent_marital_status + practice_sport +
## is_first_child + nr_siblings + transport_means + wkly_study_hours,
## data = df_transformed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -46.922  -8.043   1.071   8.811  26.214
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    57.808758   3.432409   16.842 < 2e-16 ***
## gendermale     -9.268845   1.064760   -8.705 < 2e-16 ***
## ethnic_groupgroup B  -1.372239   2.163560   -0.634 0.526175
## ethnic_groupgroup C    0.005008   2.071256    0.002 0.998072
## ethnic_groupgroup D    5.010576   2.084123    2.404 0.016531 *
## ethnic_groupgroup E    6.018419   2.284980    2.634 0.008673 **
## parent_educhigh school -0.230994   1.684700   -0.137 0.890990
## parent_educassociate's degree 6.130783   1.657904    3.698 0.000239 ***
## parent_educsome college  4.338798   1.698536    2.554 0.010898 *
## parent_educbachelor's degree 9.217680   1.941668    4.747 2.62e-06 ***
## parent_educmaster's degree 11.712279   2.376896    4.928 1.10e-06 ***
## lunch_typerstandard  9.390698   1.100772    8.531 < 2e-16 ***
## test_preprnone     -8.754351   1.122889   -7.796 3.09e-14 ***
## parent_marital_statusmarried 5.246610   1.512157    3.470 0.000561 ***
## parent_marital_statussingle 2.144248   1.725778    1.242 0.214575
## parent_marital_statuswidowed 6.877832   3.564779    1.929 0.054184 .
## practice_sportsometimes 1.674659   1.724312    0.971 0.331863
## practice_sportregularly 1.606102   1.781092    0.902 0.367574
## is_first_childyes    1.045414   1.133850    0.922 0.356921
## nr_siblings        0.546033   0.359485    1.519 0.129340
## transport_meansschool_bus 0.240107   1.087508    0.221 0.825338
## wkly_study_hours5-10  2.802323   1.255870    2.231 0.026048 *
## wkly_study_hours> 10  1.188892   1.640324    0.725 0.468881
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.65 on 564 degrees of freedom
## Multiple R-squared:  0.3634, Adjusted R-squared:  0.3385
## F-statistic: 14.63 on 22 and 564 DF, p-value: < 2.2e-16
```

Coefficients and Significance Levels:

- **Intercept (57.808758):** The expected value of `writing_score` when all other predictors are at their reference level or zero.
- **gendermale (-9.268845, $p < 0.001$):** Being male is associated with an average decrease of 9.268845 points in `writing_score` compared to females, holding all else constant. This is statistically significant.
- **ethnic_group:** `ethnic_groupgroup D` (5.010576, $p = 0.016531$) and `ethnic_groupgroup E` (6.018419, $p = 0.008673$) are significant, suggesting students in these groups score higher in writing compared to the reference group.
- **parent_educ:** `associate's degree` (6.130783, $p = 0.000239$), `some college` (4.338798, $p = 0.010898$), `bachelor's degree` (9.217680, $p = 2.62e-06$), and `master's degree` (11.712279, $p = 1.10e-06$) are significant and associated with higher writing scores compared to the reference category.
- **lunch_typerstandard (9.390698, $p < 0.001$):** Students with standard lunch type score significantly higher.

- **test_prepnone (-8.754351, p < 0.001)**: Not participating in test preparation is associated with lower writing scores.
- **parent_marital_statusmarried (5.246610, p = 0.000561)**: Children of married parents score higher.
- **practice_sport**: Not significant.
- **is_first_childyes**: Not significant.
- **nr_siblings (0.546033, p = 0.129340)**: No significant association with writing scores.
- **transport_meansschool_bus**: Not significant.
- **wkly_study_hours**: Studying 5-10 hours (2.802323, p = 0.026048) shows a significant positive effect.

Residuals:

The spread of residuals suggests the errors are somewhat symmetrically distributed around the predicted values, which is a good sign for linear regression assumptions.

Model Fit:

- **Residual Standard Error (12.65)**: Indicates the average difference between the observed values and the values predicted by the model.

R-squared:

- **Multiple R-squared (0.3634)**: About 36.34% of the variability in `writing_score` is explained by the model.
- **Adjusted R-squared (0.3385)**: Adjusts the R-squared for the number of predictors, a better measure for models with multiple predictors.

Statistic & p-value

- **F-statistic (14.63) and p-value (< 2.2e-16)**: The model is statistically significant, meaning it performs better than a model with no predictors.

Cleaned datasets

Step-wise + criteria-based: stepAIC()

Math Score

```
##
## Call:
## lm(formula = math_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep + nr_siblings + wkly_study_hours,
##     data = math_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -53.440  -8.894   0.776  10.134  32.889
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    28.0713     4.2756   6.565 1.15e-10 ***
## gender           5.3017     1.1486   4.616 4.83e-06 ***
## ethnic_group     2.7439     0.4896   5.605 3.23e-08 ***
## parent_educ      1.5210     0.3826   3.976 7.90e-05 ***
## lunch_type     12.5737     1.1964  10.510 < 2e-16 ***
## test_prep      -5.2926     1.1989  -4.414 1.21e-05 ***
```

```
## nr_siblings      0.6927      0.3860      1.795      0.0732 .
## wkly_study_hours 2.0825      0.8723      2.387      0.0173 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.8 on 579 degrees of freedom
## Multiple R-squared:  0.2758, Adjusted R-squared:  0.2671
## F-statistic: 31.5 on 7 and 579 DF,  p-value: < 2.2e-16
```

The step-wise-AIC model predicting math score contains gender, ethnic group, parent education level, lunch type, test prep, number of siblings, and weekly study hours. The p-values for gender, ethnic group, parent education level, lunch type, test prep, and weekly study hours were all < 0.05 and are therefore significant. Number of siblings was the only variable whose p-value > 0.05 . The overall p-value of the model < 0.05 as well.

Reading Score

```
##
## Call:
## lm(formula = reading_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep, data = reading_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -44.354  -8.959   0.802   9.901  32.216
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   66.7121     3.6485  18.285 < 2e-16 ***
## gender        -7.5066     1.1139  -6.739 3.84e-11 ***
## ethnic_group    1.7930     0.4753   3.773 0.000178 ***
## parent_educ    1.7606     0.3713   4.742 2.66e-06 ***
## lunch_type     8.6667     1.1618   7.459 3.18e-13 ***
## test_prep     -6.8289     1.1580  -5.897 6.28e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.4 on 581 degrees of freedom
## Multiple R-squared:  0.2264, Adjusted R-squared:  0.2197
## F-statistic: 34.01 on 5 and 581 DF,  p-value: < 2.2e-16
```

The step-wise-AIC model predicting reading score contains gender, ethnic group, parent education level, lunch type, and test prep. The p-values for all of these variables were < 0.05 and are therefore significant. The overall p-value of the model < 0.05 as well.

Writing Score

```
##
## Call:
## lm(formula = writing_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep + wkly_study_hours, data = writing_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -49.917  -8.391   0.613   9.143  29.293
##
## Coefficients:
```

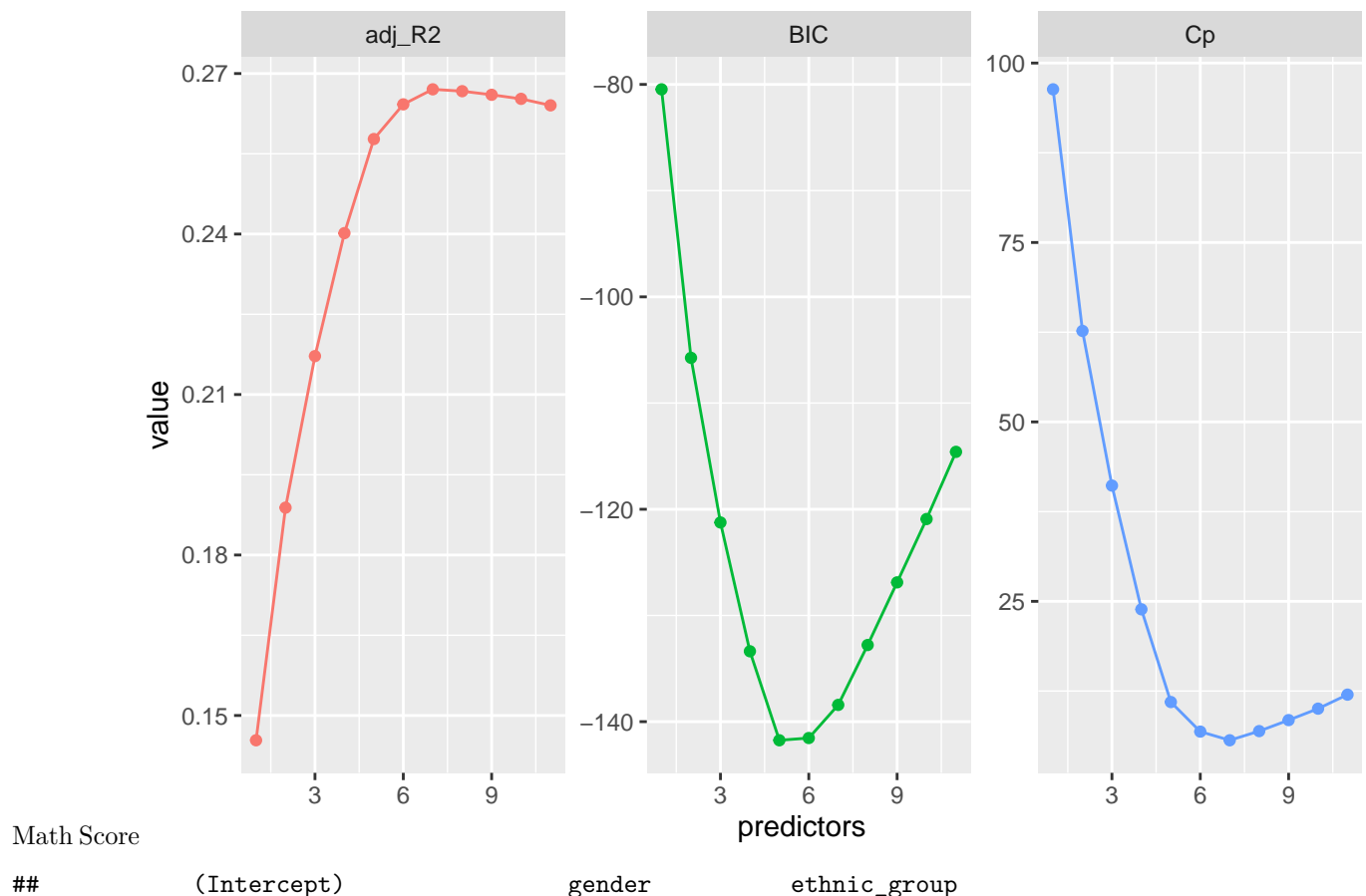
```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    65.3125     3.8874  16.801 < 2e-16 ***
## gender         -9.1792     1.0698  -8.581 < 2e-16 ***
## ethnic_group    2.1684     0.4562   4.753 2.53e-06 ***
## parent_educ     2.3242     0.3566   6.519 1.54e-10 ***
## lunch_type      9.4976     1.1151   8.517 < 2e-16 ***
## test_prep     -9.0360     1.1163  -8.094 3.40e-15 ***
## wkly_study_hours 1.1762     0.8121   1.448  0.148
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.86 on 580 degrees of freedom
## Multiple R-squared:  0.3233, Adjusted R-squared:  0.3163
## F-statistic: 46.18 on 6 and 580 DF,  p-value: < 2.2e-16
```

The step-wise-AIC model predicting writing score contains gender, ethnic group, parent education level, lunch type, test prep, and weekly study hours. The p-values for gender, ethnic group, parent education level, lunch type, and test prep were all < 0.05 and are therefore significant. Weekly study hours was the only variable whose p-value > 0.05 . The overall p-value of the model < 0.05 as well.

The writing score's step-AIC model seemed to have lowest residual standard error out of all three scores' models. It is also interesting to note that the adjusted R^2 values for all three models only differed slightly from their R^2 counterparts by about -0.01 to -0.02.

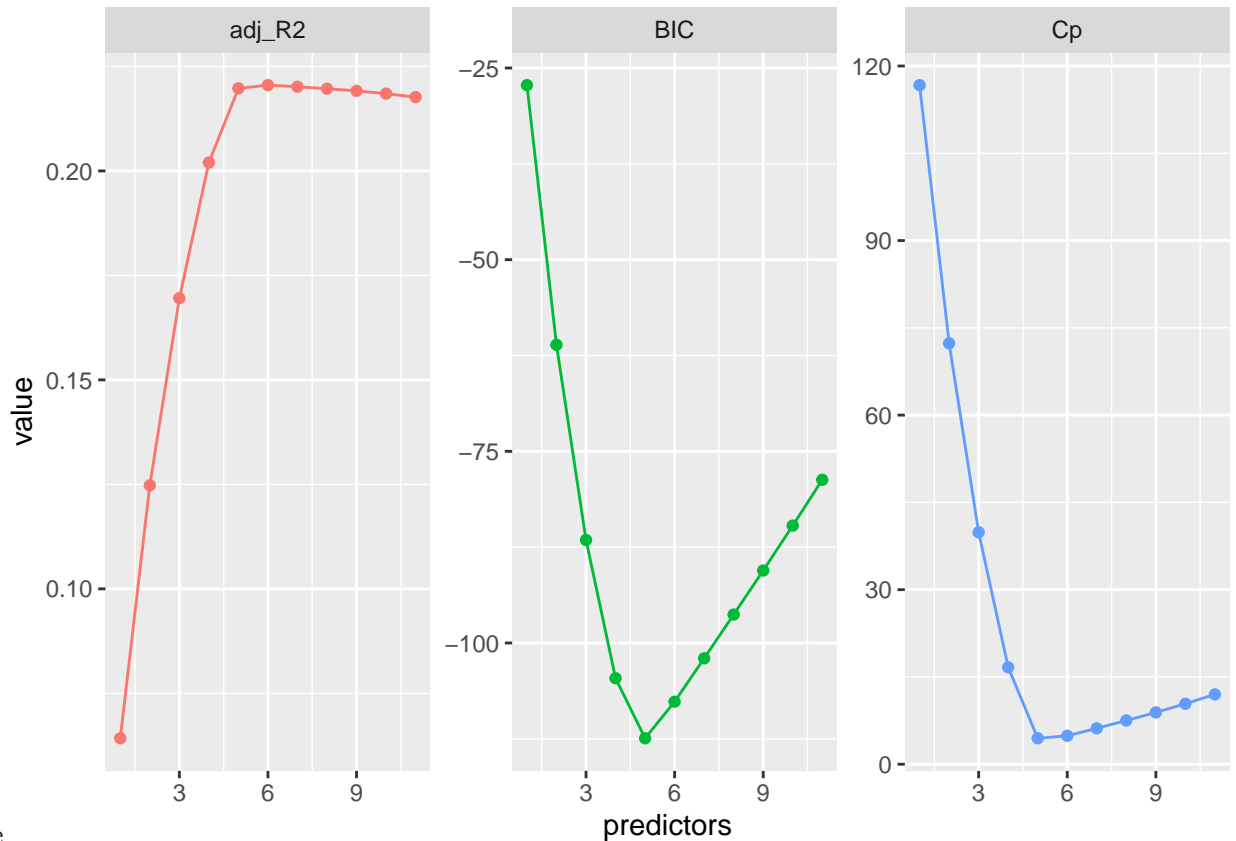
Criteria-based approach - Adjusted R^2 , Cp, BIC

(Note: BIC has a larger penalty, leading to less predictors present within the model.)



| | | | |
|----|-----------------------|-----------------|------------------|
| ## | TRUE | TRUE | TRUE |
| ## | parent_educ | lunch_type | test_prep |
| ## | TRUE | TRUE | TRUE |
| ## | parent_marital_status | practice_sport | is_first_child |
| ## | FALSE | FALSE | FALSE |
| ## | nr_siblings | transport_means | wkly_study_hours |
| ## | TRUE | FALSE | TRUE |

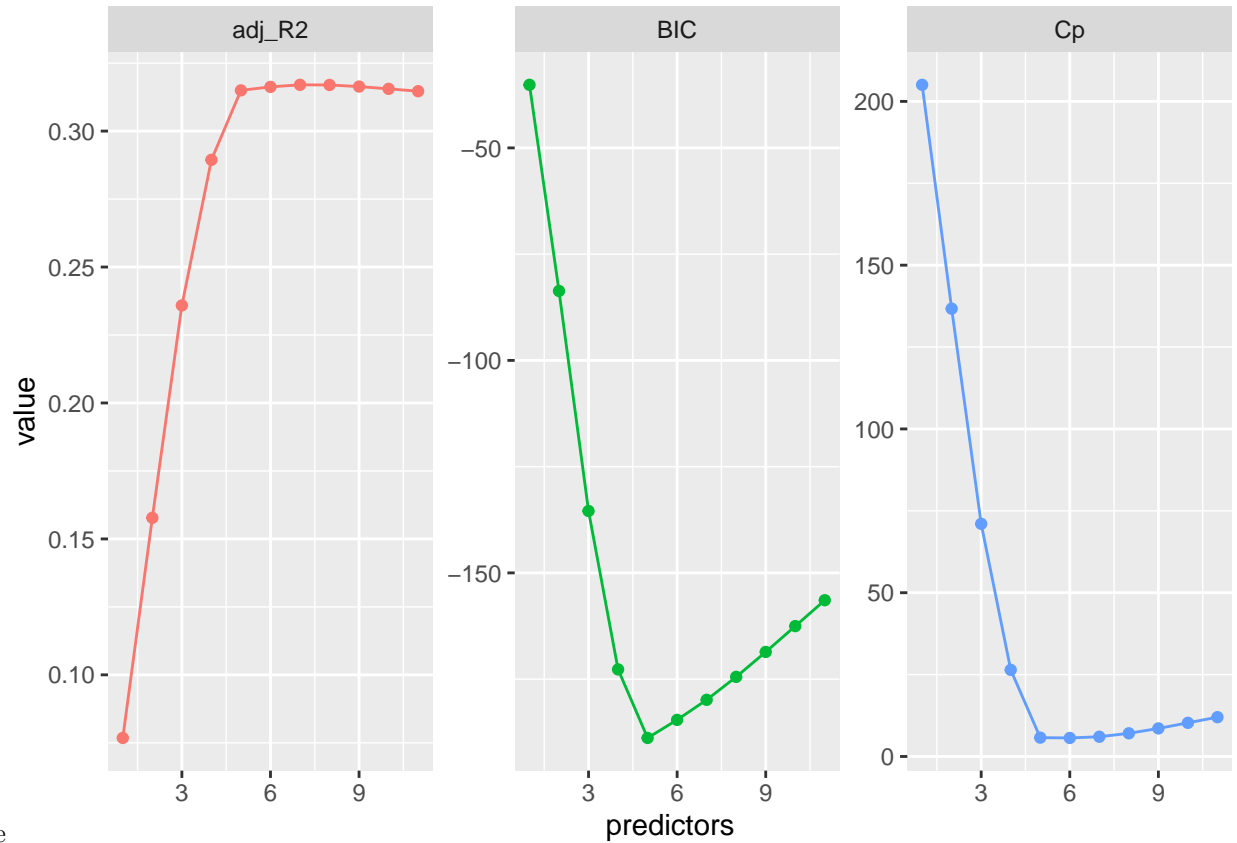
To predict math score, the adjusted R^2 statistic and C_p shows that a 7-variable model is optimal. The BIC shows the predictors selected are:



Reading Score

| | | | |
|----|-----------------------|-----------------|------------------|
| ## | (Intercept) | gender | ethnic_group |
| ## | TRUE | TRUE | TRUE |
| ## | parent_educ | lunch_type | test_prep |
| ## | TRUE | TRUE | TRUE |
| ## | parent_marital_status | practice_sport | is_first_child |
| ## | FALSE | TRUE | FALSE |
| ## | nr_siblings | transport_means | wkly_study_hours |
| ## | FALSE | FALSE | TRUE |

To predict reading score, the adjusted R^2 statistic and C_p shows that the predictors selected are:



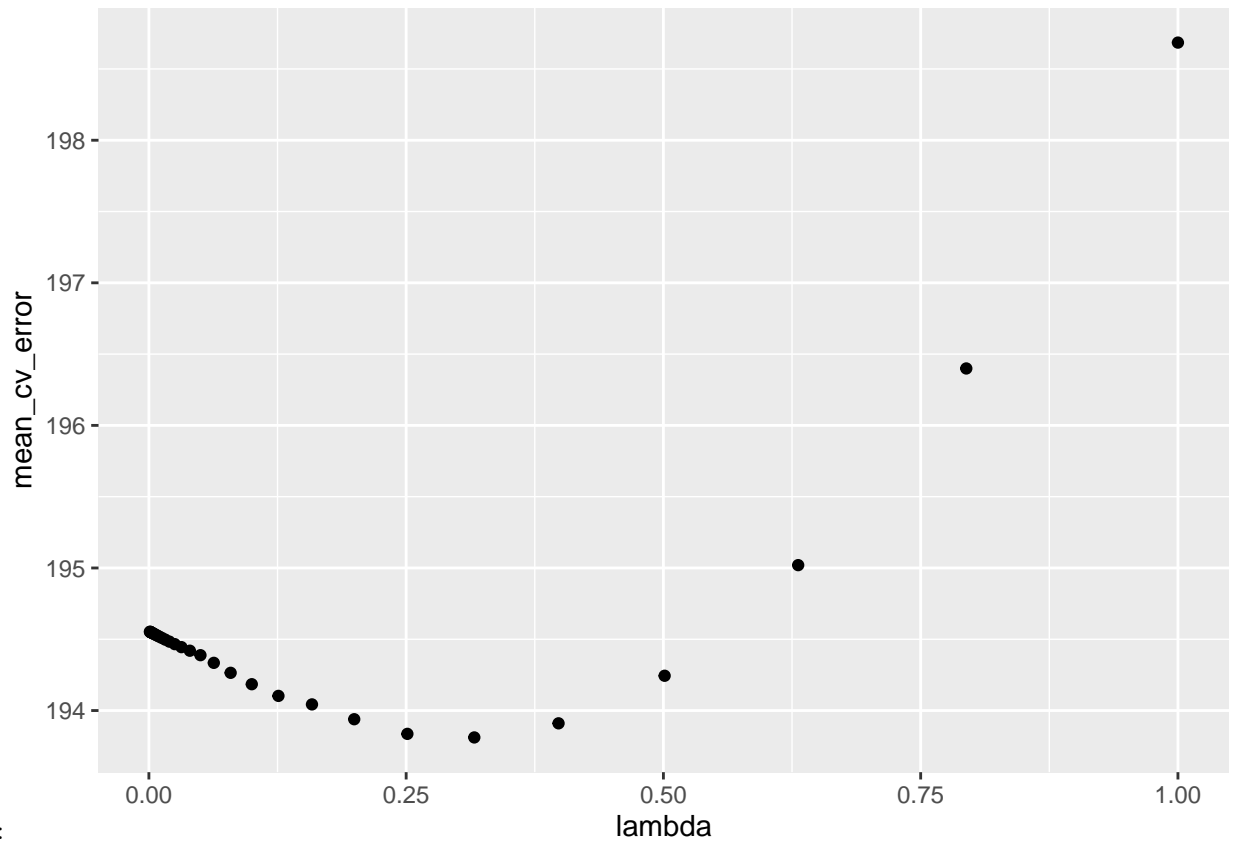
Writing Score

| | | | |
|----|-----------------------|-----------------|------------------|
| ## | (Intercept) | gender | ethnic_group |
| ## | TRUE | TRUE | TRUE |
| ## | parent_educ | lunch_type | test_prep |
| ## | TRUE | TRUE | TRUE |
| ## | parent_marital_status | practice_sport | is_first_child |
| ## | FALSE | FALSE | FALSE |
| ## | nr_siblings | transport_means | wkly_study_hours |
| ## | TRUE | FALSE | TRUE |

To predict writing score, the adjusted R^2 statistic shows that a 7 or 8-variable model is optimal. The C_p suggests a 7-variable model as well.

The predictors selected are:

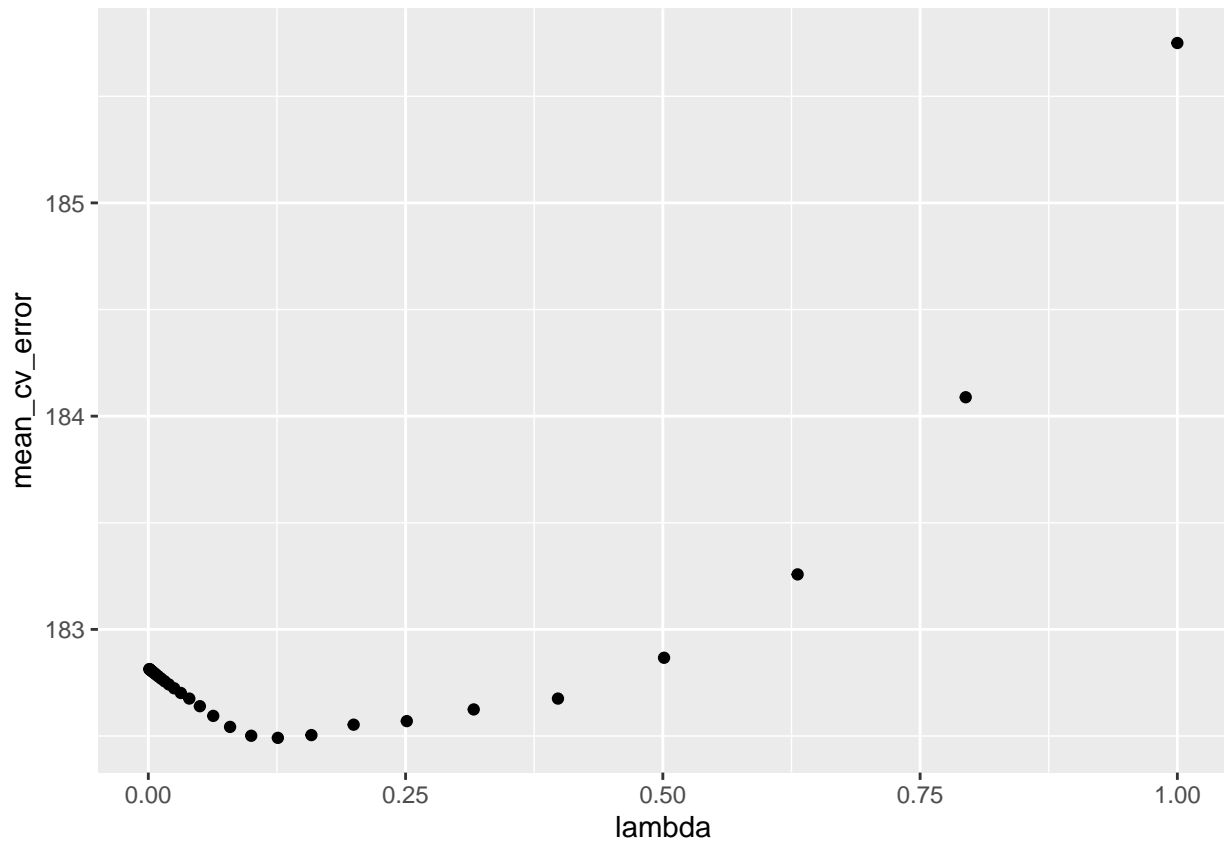
LASSO approach -



Math score:

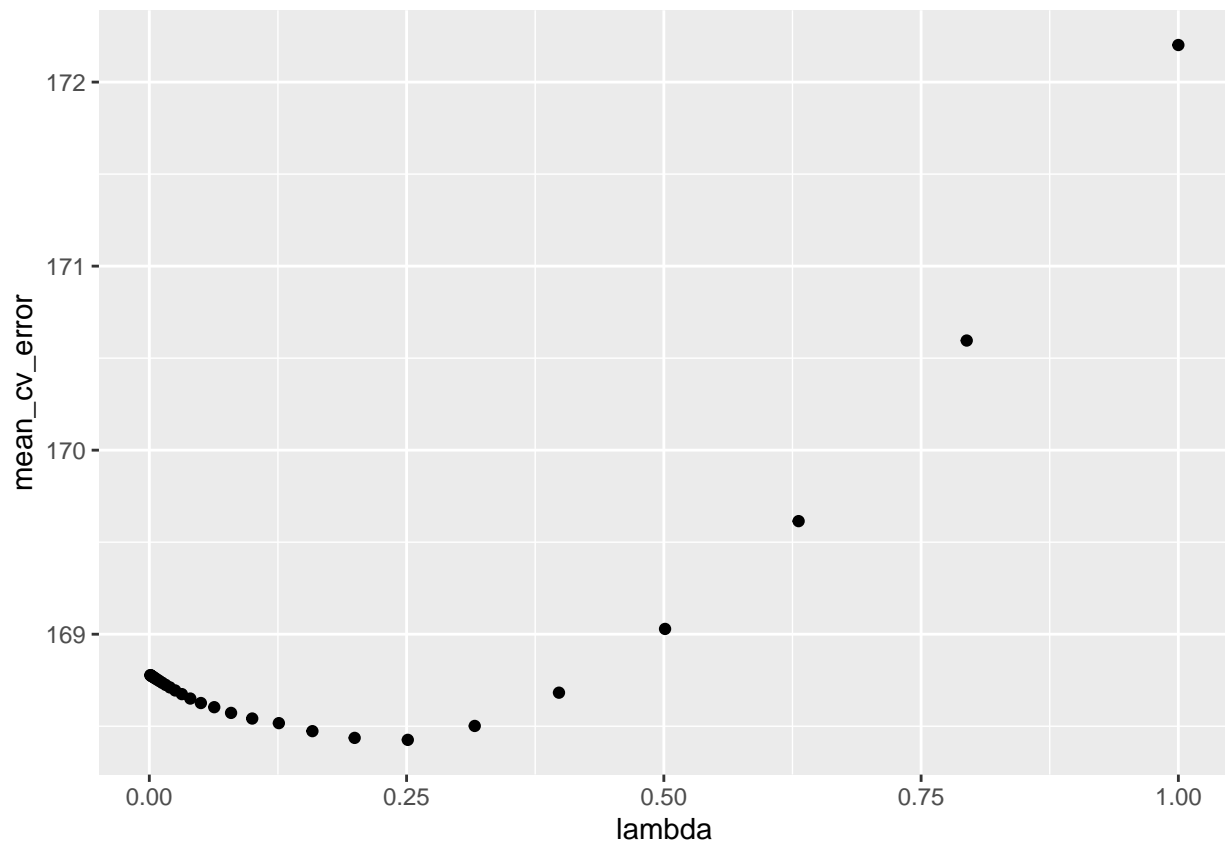
```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##              s0
## (Intercept)  30.9514949
## gender       4.6893056
## ethnic_group  2.5085903
## parent_educ   1.2932847
## lunch_type   11.9603880
## test_prep    -4.7061459
## parent_marital_status 0.2406793
## practice_sport 0.1053101
## is_first_child .
## nr_siblings   0.4775542
## transport_means .
## wkly_study_hours 1.6725295
```

Reading score:



```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##              s0
## (Intercept)  63.0168462
## gender      -7.2578814
## ethnic_group  1.7148165
## parent_educ  1.6792539
## lunch_type   8.3560449
## test_prep   -6.4054407
## parent_marital_status 0.3113991
## practice_sport -0.5218650
## is_first_child  0.6325478
## nr_siblings   0.2364662
## transport_means 0.6090418
## wkly_study_hours 0.8425385
```

Writing score:



```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##              s0
## (Intercept)  64.24045111
## gender       -8.58684295
## ethnic_group  2.00086017
## parent_educ   2.16675381
## lunch_type    8.95045903
## test_prep    -8.47239858
## parent_marital_status 0.40572033
## practice_sport 0.04910284
## is_first_child .
## nr_siblings   0.29894240
## transport_means 0.22598323
## wkly_study_hours 0.76604805
```