P8130 Final Report (Project 1)

Huanyu Chen (hc3451) Xiaoting Tang (xt2288) Yifei Liu (yl5508)
Longyu Zhang (lz2951)

2023-12-16

Read and Clean Data

```
data =
  read_csv("./data.csv") |>
  janitor::clean_names() |>
  mutate(
    gender = factor(case_when(
      gender == "male" ~ 0,
      gender == "female" ~ 1,
      )),
    ethnic_group = factor(case_when(
      ethnic_group == "group A" ~ 0,
      ethnic_group == "group B" ~ 1,
      ethnic_group == "group C" ~ 2,
      ethnic_group == "group D" ~ 3,
      ethnic_group == "group E" ~ 4,
      )),
    parent_educ = factor(case_when(
      parent_educ == "some highschool" ~ 0,
      parent_educ == "some college" ~ 1,
      parent_educ == "associate's degree" ~ 2,
      parent_educ == "bachelor's degree" ~ 3,
      parent_educ == "master's degree" ~ 4,
      )),
    lunch_type = factor(case_when(
      lunch_type == "standard" ~ 0,
      lunch_type == "free/reduced" ~ 1,
      )),
    test_prep = factor(case_when(
      test_prep == "none" ~ 0,
      test_prep == "completed" ~ 1,
      )),
    parent_marital_status = factor(case_when(
      parent_marital_status == "married" ~ 0,
      parent_marital_status == "single" ~ 1,
      parent_marital_status == "widowed" ~ 2,
      parent_marital_status == "divorced" ~ 3,
      )),
```

```
    practice_sport = factor(case_when(
      practice_sport == "never" ~ 0,
      practice_sport == "sometimes" ~ 1,
      practice_sport == "regularly" ~ 2,
      )),
    is_first_child = factor(case_when(
      is_first_child == "no" ~ 0,
      is_first_child == "yes" ~ 1,
      )),
    transport_means = factor(case_when(
      transport_means == "school_bus" ~ 0,
      transport_means == "private" ~ 1,
      )),
    wkly_study_hours = factor(case_when(
      wkly_study_hours == "< 5" ~ 0,
      wkly_study_hours == "10-May" ~ 1,
      wkly_study_hours == "> 10" ~ 2,
      ))
    ) |>
  mutate(nr_siblings = factor(nr_siblings))
```

```
## Rows: 948 Columns: 14
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (10): Gender, EthnicGroup, ParentEduc, LunchType, TestPrep, ParentMarita...
## dbl  (4): NrSiblings, MathScore, ReadingScore, WritingScore
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
'
# Deal with NA -- Calculate the column mean (round to integer) and plug it into NA cell
column_means <- round(colMeans(data, na.rm = TRUE), digits = 0)
for (col in names(data)) {
  data[[col]][is.na(data[[col]])] <- column_means[col]
  }

head(data)
'
```

```
## [1] "\n# Deal with NA -- Calculate the column mean (round to integer) and plug it into NA cell\ncolu
```

```
# Another data set for EDA
data_long <- data |>
  pivot_longer(cols = c(math_score, reading_score, writing_score),
               names_to = "test", values_to = "score")
```

Summary

```r
sum_data_fct =
  data |>
  dplyr::select(1:11) |>
  skimr::skim() |>
  dplyr::select(skim_variable, n_missing, complete_rate, factor.n_unique, factor.top_counts)

colnames(sum_data_fct) = c("Variable", "Missing", "Complete Rate", "Unique", "Top Counts")

knitr::kable(x = sum_data_fct, caption = "Categorical Variables pre-analysis", digits = 1)
```

Table 1: Categorical Variables pre-analysis

Variable	Missing	Complete Rate	Unique	Top Counts
gender	0	1.0	2	1: 488, 0: 460
ethnic_group	59	0.9	5	2: 277, 3: 237, 1: 171, 4: 124
parent_educ	392	0.6	4	1: 199, 2: 198, 3: 104, 4: 55
lunch_type	0	1.0	2	0: 617, 1: 331
test_prep	55	0.9	2	0: 571, 1: 322
parent_marital_status	49	0.9	4	0: 516, 1: 213, 3: 146, 2: 24
practice_sport	16	1.0	3	1: 477, 2: 343, 0: 112
is_first_child	30	1.0	2	1: 604, 0: 314
nr_siblings	46	1.0	8	1: 245, 2: 213, 3: 198, 0: 101
transport_means	102	0.9	2	0: 509, 1: 337
wkly_study_hours	37	1.0	3	1: 508, 0: 253, 2: 150

```r
data =
  data |>
  drop_na()
```

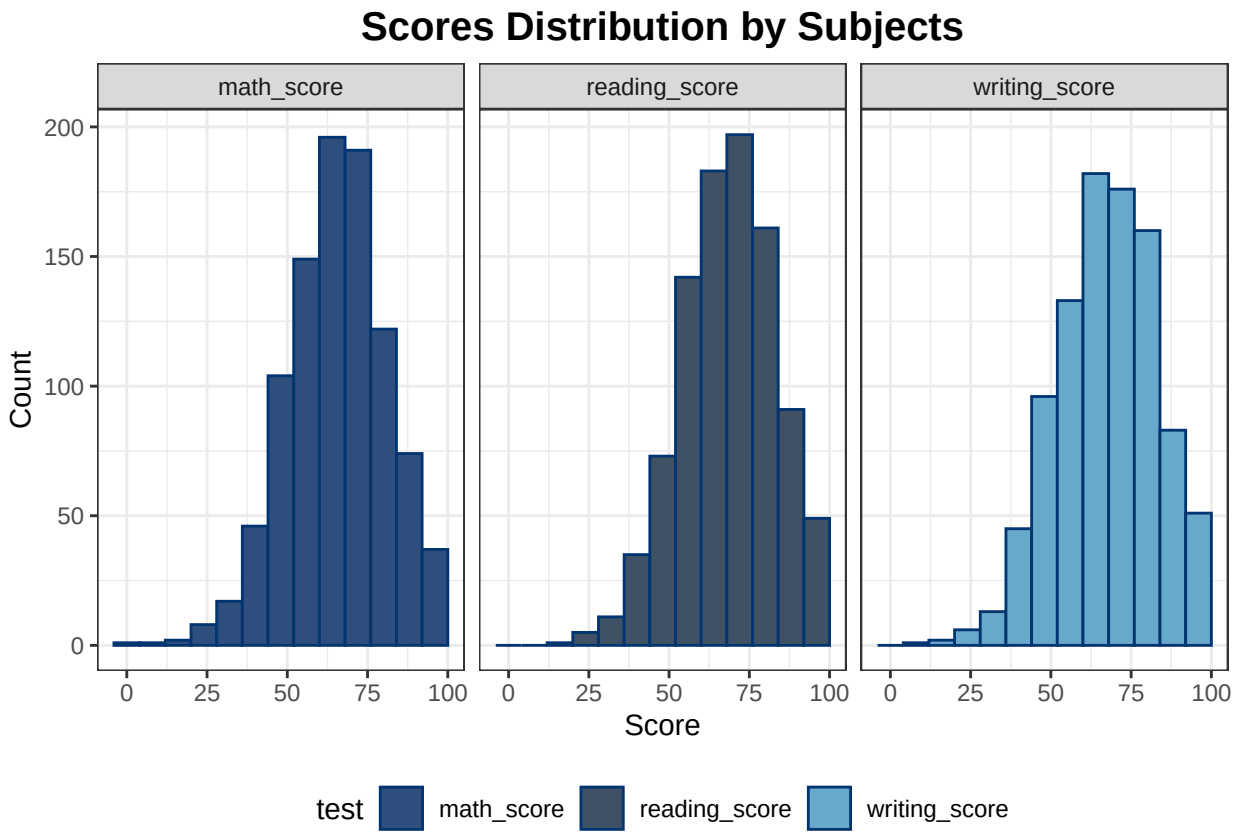
```r
sum_data_score =
  data |>
  dplyr::select(12:14) |>
  skimr::skim() |>
  dplyr::select(skim_variable, numeric.mean, numeric.sd, numeric.p0, numeric.p25, numeric.p50, numeric.

colnames(sum_data_score) = c("Variable", "Mean", "SD", "Min", "Q1", "Median", "Q3", "Max")

knitr::kable(x = sum_data_score, caption = "Continuous Variables pre-analysis", digits = 1)
```

Table 2: Continuous Variables pre-analysis

Variable	Mean	SD	Min	Q1	Median	Q3	Max
math_score	68.7	15.9	18	57	69.0	81	100
reading_score	72.3	14.8	23	61	73.0	84	100
writing_score	72.0	15.2	19	62	72.5	84	100

Histograms

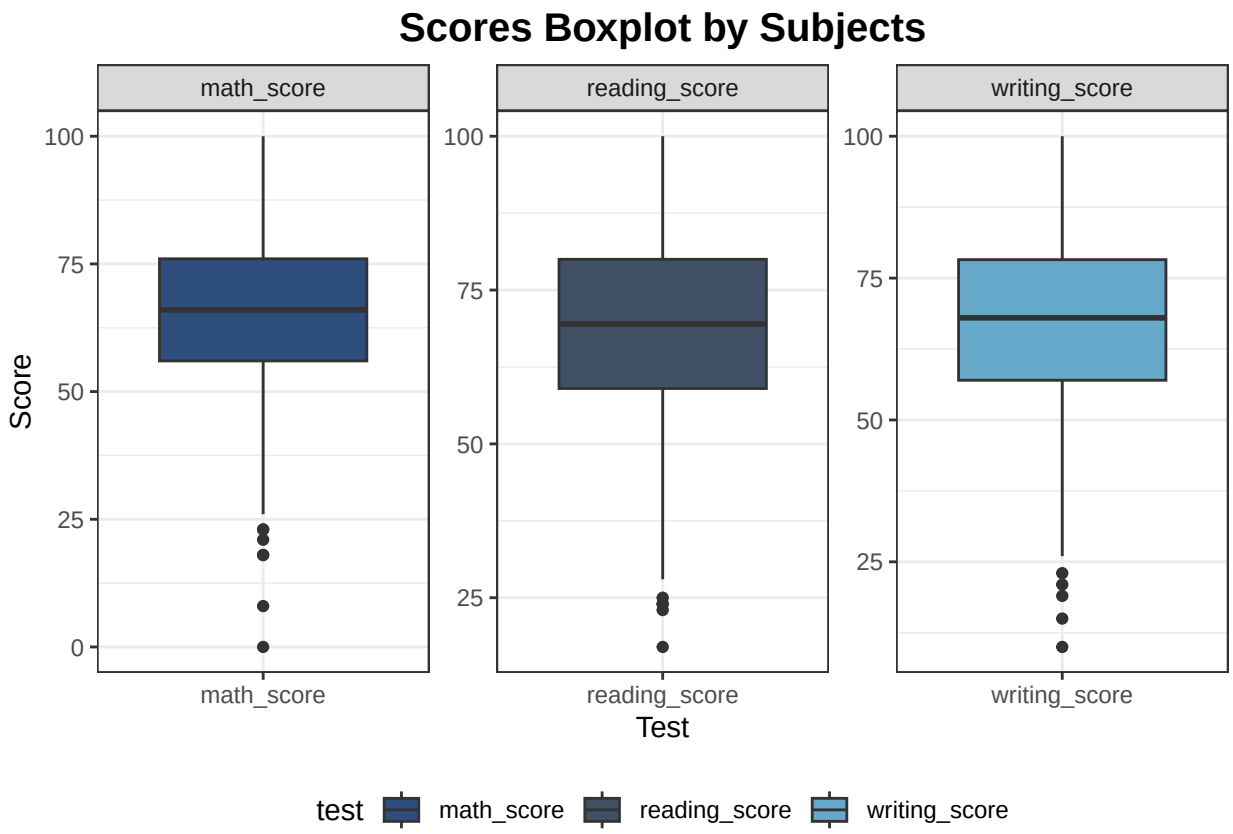
```
data_long |>
  ggplot(aes(x = score, fill = test)) +
  geom_histogram(binwidth = 8, color = "#013571") +
  labs(
    title = "Scores Distribution by Subjects",
    x = "Score",
    y = "Count"
    ) +
  scale_fill_manual(values = c("#2E4E7D", "#405165", "#67A9CB")) +
  facet_grid(~ test) +
  theme_bw() +
  theme(legend.position = "bottom") +
  theme(plot.title = element_text(size = 15, face = "bold", hjust = 0.5))
```



Boxplots

```
data_long |>
  ggplot(aes(x = test, y = score, fill = test)) +
  geom_boxplot() +
  labs(
    title = "Scores Boxplot by Subjects",
```

```
    x = "Test",
    y = "Score"
    ) +
  facet_wrap(~ test, scales = "free") +
  scale_fill_manual(values = c("#2E4E7D", "#405165", "#67A9CB")) +
  theme_bw() +
  theme(legend.position = "bottom") +
  theme(plot.title = element_text(size = 15, face = "bold", hjust = 0.5))
```



Diagnostics

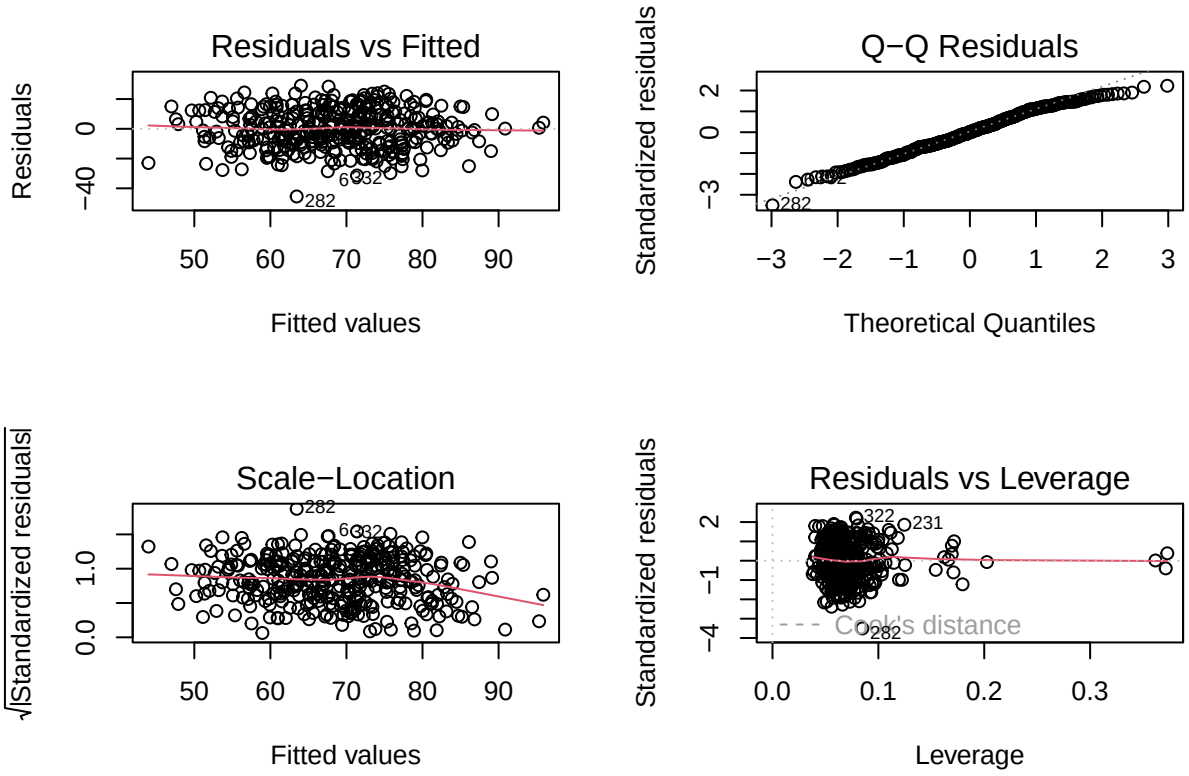
```
# Math
model_math_full = lm(math_score ~ . - reading_score - writing_score, data = data)
model_math_full |> summary()
```

```
##
## Call:
## lm(formula = math_score ~ . - reading_score - writing_score,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.458  -8.961   0.089   9.800  28.981
```

```
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               62.3523     4.9540  12.586  < 2e-16 ***
## gender1                   -3.6522     1.4958  -2.442 0.015150 *
## ethnic_group1              1.8120     3.2790   0.553 0.580912
## ethnic_group2             -1.1247     3.1319  -0.359 0.719748
## ethnic_group3              3.0342     3.1826   0.953 0.341109
## ethnic_group4              8.7423     3.3555   2.605 0.009598 **
## parent_educ2               1.8031     1.7975   1.003 0.316545
## parent_educ3               3.1775     2.0927   1.518 0.129886
## parent_educ4               4.0051     2.5782   1.553 0.121282
## lunch_type1              -12.1275     1.5423  -7.863 5.49e-14 ***
## test_prep1                 5.7990     1.5706   3.692 0.000260 ***
## parent_marital_status1    -4.2006     1.8079  -2.323 0.020770 *
## parent_marital_status2     7.0930     4.7226   1.502 0.134083
## parent_marital_status3    -4.8362     2.1726  -2.226 0.026694 *
## practice_sport1            3.0566     2.3818   1.283 0.200295
## practice_sport2            3.2296     2.4896   1.297 0.195466
## is_first_child1           -0.3254     1.6378  -0.199 0.842638
## nr_siblings1              -0.1780     2.7665  -0.064 0.948739
## nr_siblings2              -1.1446     2.8721  -0.399 0.690507
## nr_siblings3               3.1546     2.8049   1.125 0.261548
## nr_siblings4               2.8587     3.3920   0.843 0.399963
## nr_siblings5               2.4937     3.9289   0.635 0.526071
## nr_siblings6              14.5158    13.9723   1.039 0.299617
## nr_siblings7               9.5593     8.3433   1.146 0.252735
## transport_means1           1.0585     1.5640   0.677 0.499003
## wkly_study_hours1          6.4822     1.7525   3.699 0.000254 ***
## wkly_study_hours2          4.2523     2.2536   1.887 0.060065 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.55 on 327 degrees of freedom
## Multiple R-squared:  0.3256, Adjusted R-squared:  0.272
## F-statistic: 6.073 on 26 and 327 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(model_math_full)
```

```
## Warning: not plotting observations with leverage one:
##   186
```



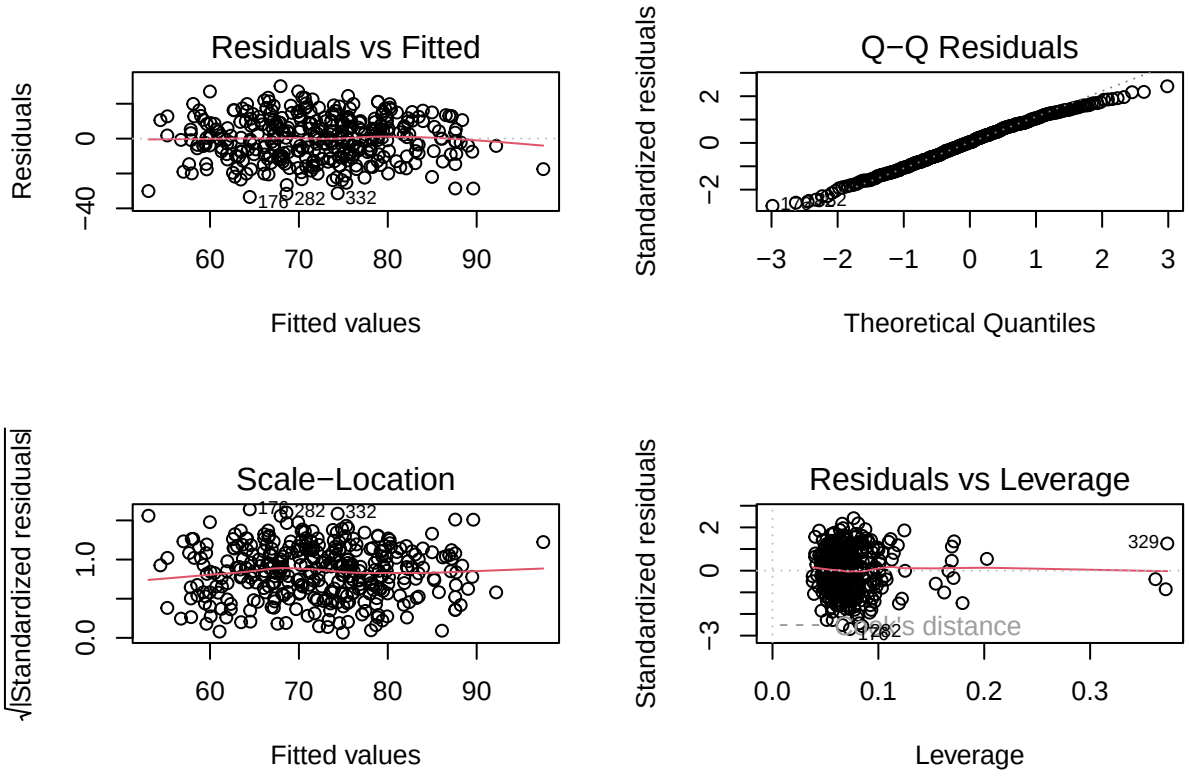

```
# Reading
model_reading_full = lm(reading_score ~ . - math_score - writing_score, data = data)
model_reading_full |> summary()
```

```
##
## Call:
## lm(formula = reading_score ~ . - math_score - writing_score,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.470  -8.942   0.403   9.553  30.063
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               59.3627     4.7169  12.585  < 2e-16 ***
## gender1                    8.2587     1.4242   5.799 1.57e-08 ***
## ethnic_group1              1.4533     3.1220   0.466  0.64188
## ethnic_group2             -0.5044     2.9819  -0.169  0.86578
## ethnic_group3              2.8080     3.0302   0.927  0.35479
## ethnic_group4              4.7359     3.1949   1.482  0.13921
## parent_educ2               2.6502     1.7114   1.549  0.12246
## parent_educ3               4.5816     1.9925   2.299  0.02211 *
## parent_educ4               6.4240     2.4548   2.617  0.00929 **
## lunch_type1               -7.8783     1.4685  -5.365 1.54e-07 ***
## test_prep1                 7.6036     1.4954   5.085 6.21e-07 ***
## parent_marital_status1    -4.6412     1.7214  -2.696  0.00738 **
## parent_marital_status2     4.6364     4.4966   1.031  0.30325
## parent_marital_status3    -4.2660     2.0686  -2.062  0.03997 *
```

```
## practice_sport1          1.9156     2.2678   0.845  0.39890
## practice_sport2          1.2989     2.3705   0.548  0.58408
## is_first_child1          0.6384     1.5594   0.409  0.68252
## nr_siblings1             0.4794     2.6341   0.182  0.85569
## nr_siblings2            -1.4869     2.7347  -0.544  0.58700
## nr_siblings3             1.8958     2.6706   0.710  0.47830
## nr_siblings4             2.3345     3.2296   0.723  0.47028
## nr_siblings5            -1.4797     3.7408  -0.396  0.69269
## nr_siblings6            11.7473    13.3034   0.883  0.37787
## nr_siblings7             7.7275     7.9439   0.973  0.33139
## transport_means1         0.5365     1.4891   0.360  0.71890
## wkly_study_hours1        5.3310     1.6686   3.195  0.00154 **
## wkly_study_hours2        1.1401     2.1458   0.531  0.59557
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.9 on 327 degrees of freedom
## Multiple R-squared:  0.2971, Adjusted R-squared:  0.2412
## F-statistic: 5.315 on 26 and 327 DF,  p-value: 6.451e-14
```

```r
par(mfrow = c(2,2))
plot(model_reading_full)
```

```
## Warning: not plotting observations with leverage one:
##   186
```



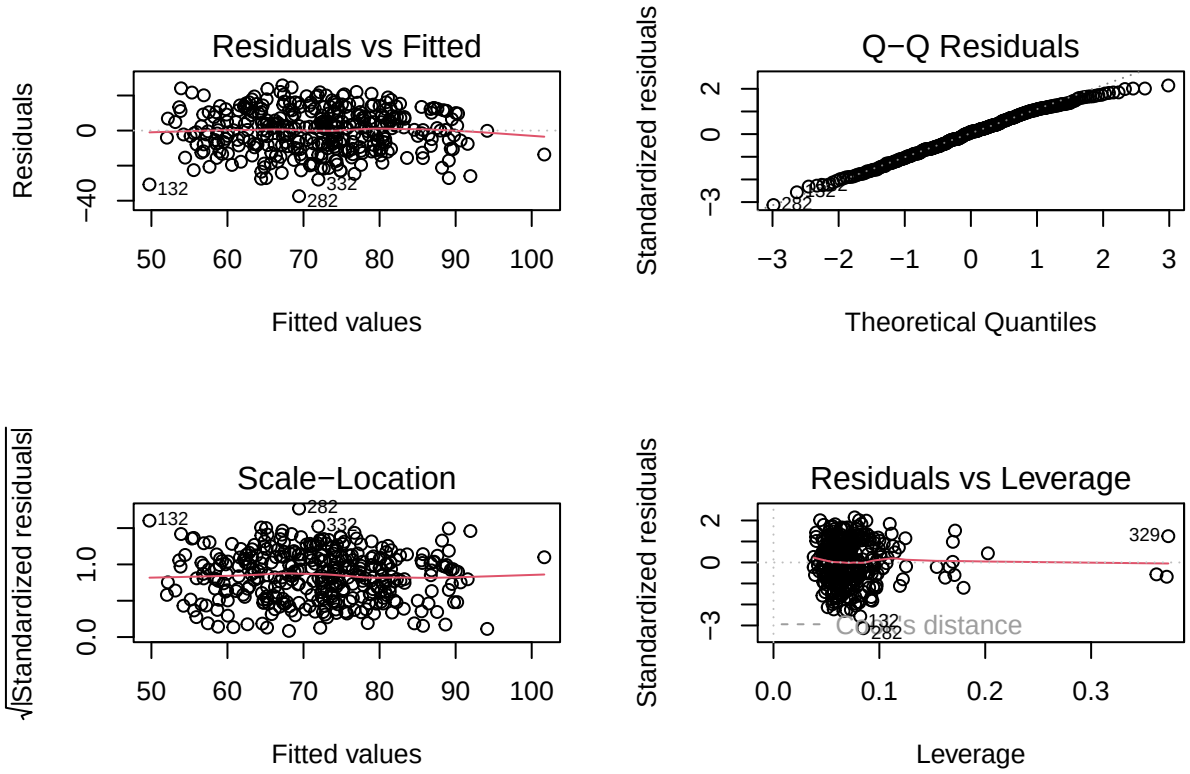
```r
# Writing
model_writing_full = lm(writing_score ~ . - reading_score - math_score, data = data)
model_writing_full |> summary()
```

```
##
## Call:
## lm(formula = writing_score ~ . - reading_score - math_score,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.416  -8.131   1.123   9.165  25.765
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              55.1871     4.5675  12.083  < 2e-16 ***
## gender1                  10.0433     1.3791   7.283 2.46e-12 ***
## ethnic_group1             1.7982     3.0232   0.595 0.552382
## ethnic_group2             0.7708     2.8875   0.267 0.789684
## ethnic_group3             5.5577     2.9343   1.894 0.059101 .
## ethnic_group4             5.5666     3.0937   1.799 0.072893 .
## parent_educ2              2.0224     1.6572   1.220 0.223203
## parent_educ3              4.5673     1.9294   2.367 0.018507 *
## parent_educ4              7.5525     2.3771   3.177 0.001629 **
## lunch_type1              -8.9424     1.4220  -6.289 1.03e-09 ***
## test_prep1                9.6428     1.4480   6.659 1.16e-10 ***
## parent_marital_status1   -4.5781     1.6669  -2.747 0.006356 **
## parent_marital_status2    5.2451     4.3542   1.205 0.229221
## parent_marital_status3   -4.4305     2.0031  -2.212 0.027669 *
## practice_sport1           3.3011     2.1960   1.503 0.133746
## practice_sport2           3.0186     2.2954   1.315 0.189415
## is_first_child1          -0.2525     1.5100  -0.167 0.867295
## nr_siblings1              0.3186     2.5507   0.125 0.900665
## nr_siblings2             -1.2993     2.6481  -0.491 0.624008
## nr_siblings3              2.2515     2.5860   0.871 0.384594
## nr_siblings4              2.9536     3.1273   0.944 0.345630
## nr_siblings5             -0.5419     3.6224  -0.150 0.881167
## nr_siblings6             14.3830    12.8821   1.117 0.265024
## nr_siblings7              8.0232     7.6923   1.043 0.297708
## transport_means1          0.9938     1.4420   0.689 0.491208
## wkly_study_hours1         5.4344     1.6157   3.363 0.000861 ***
## wkly_study_hours2         2.0335     2.0778   0.979 0.328454
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.5 on 327 degrees of freedom
## Multiple R-squared:  0.3762, Adjusted R-squared:  0.3266
## F-statistic: 7.586 on 26 and 327 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2,2))
plot(model_writing_full)
```

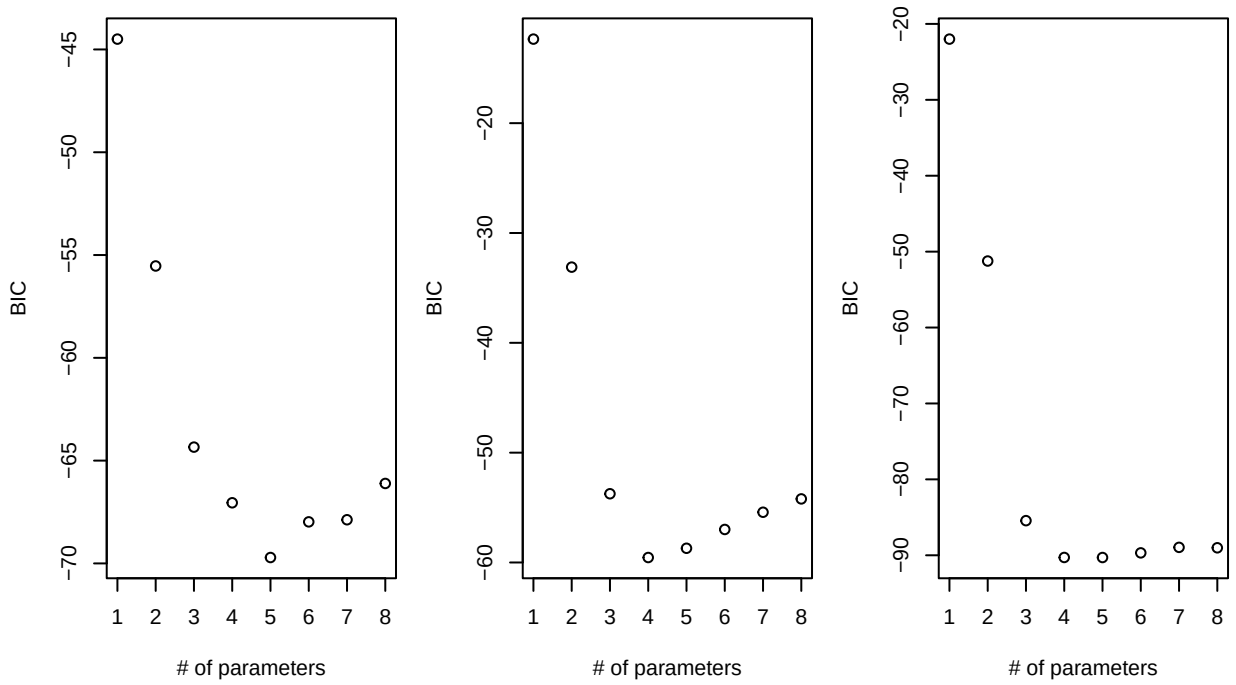
```
## Warning: not plotting observations with leverage one:
```

```
##    186
```



criterion-based Procedures

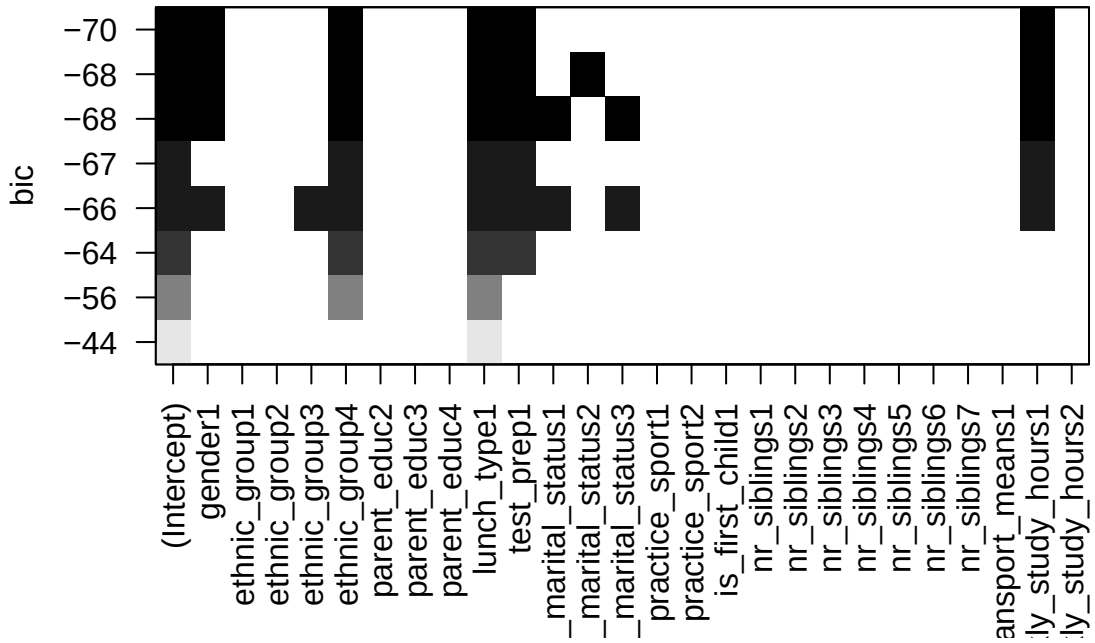
```
math_c = regsubsets(math_score ~ . - reading_score - writing_score, data = data)
res_math =
  math_c |>
  summary()

reading_c = regsubsets(reading_score ~ . - math_score - writing_score, data = data)
res_reading =
  reading_c |>
  summary()

writing_c = regsubsets(writing_score ~ . - math_score - reading_score, data = data)
res_writing =
  writing_c |>
  summary()

par(mfrow = c(1, 3), mar = c(8, 4, 4, 1))
plot(1:8, res_math$bic, xlab = "# of parameters", ylab = "BIC")
plot(1:8, res_reading$bic, xlab = "# of parameters", ylab = "BIC")
plot(1:8, res_writing$bic, xlab = "# of parameters", ylab = "BIC")
```




```
par(mfrow = c(1, 1))
res_math$outmat[5,]
```

```
##                gender1          ethnic_group1          ethnic_group2
##                    "*"                    " "                    " "
##          ethnic_group3          ethnic_group4           parent_educ2
##                    " "                    "*"                    " "
##           parent_educ3           parent_educ4            lunch_type1
##                    " "                    " "                    "*"
##             test_prep1 parent_marital_status1 parent_marital_status2
##                    "*"                    " "                    " "
## parent_marital_status3        practice_sport1        practice_sport2
##                    " "                    " "                    " "
##        is_first_child1           nr_siblings1           nr_siblings2
##                    " "                    " "                    " "
##           nr_siblings3           nr_siblings4           nr_siblings5
##                    " "                    " "                    " "
##           nr_siblings6           nr_siblings7       transport_means1
##                    " "                    " "                    " "
##      wkly_study_hours1      wkly_study_hours2
##                    "*"                    " "
```

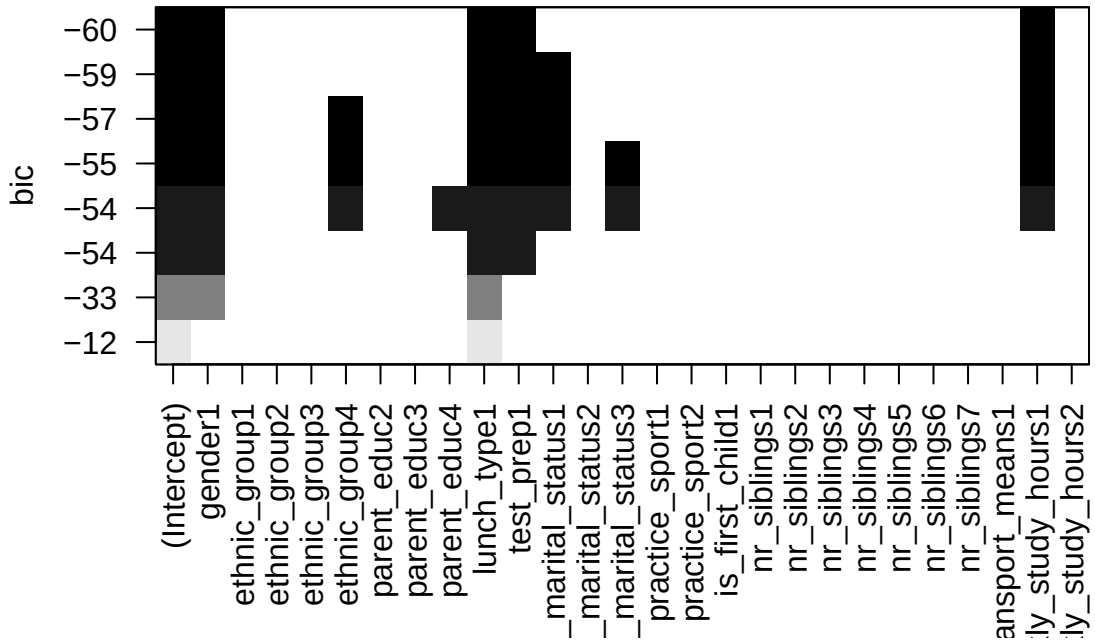
```
plot(math_c, scale = "bic")
```




```
res_reading$outmat[5,]
```

```
##                gender1          ethnic_group1          ethnic_group2
##                    "*"                    " "                    " "
##          ethnic_group3          ethnic_group4           parent_educ2
##                    " "                    " "                    " "
##           parent_educ3           parent_educ4            lunch_type1
##                    " "                    " "                    "*"
##             test_prep1 parent_marital_status1 parent_marital_status2
##                    "*"                    "*"                    " "
## parent_marital_status3        practice_sport1        practice_sport2
##                    " "                    " "                    " "
##        is_first_child1           nr_siblings1           nr_siblings2
##                    " "                    " "                    " "
##           nr_siblings3           nr_siblings4           nr_siblings5
##                    " "                    " "                    " "
##           nr_siblings6           nr_siblings7       transport_means1
##                    " "                    " "                    " "
##      wkly_study_hours1      wkly_study_hours2
##                    "*"                    " "
```

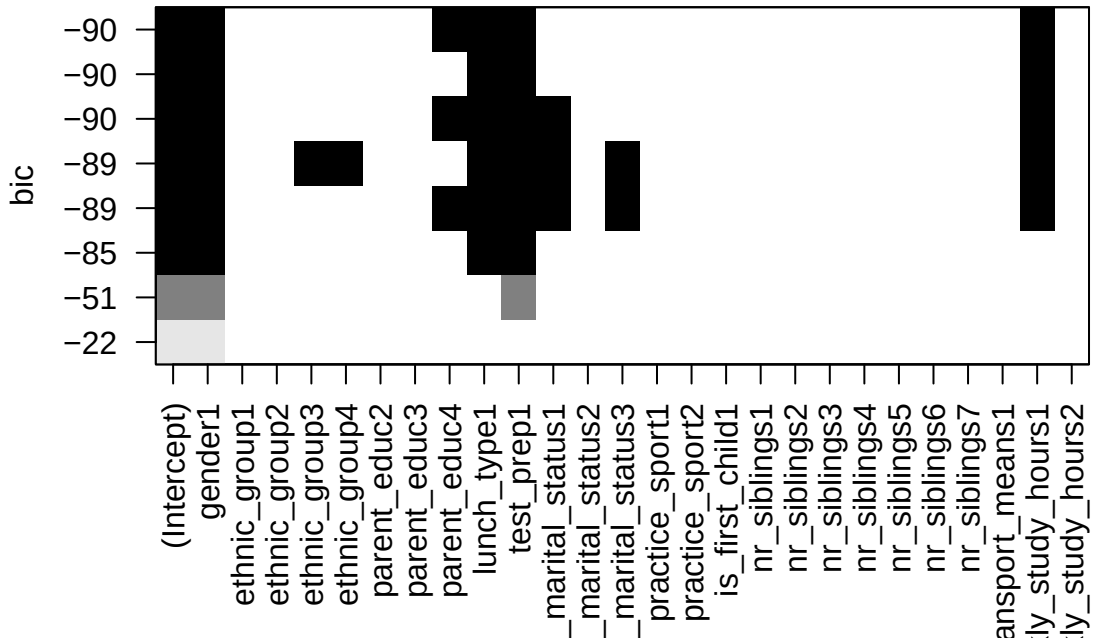
```
plot(reading_c, scale = "bic")
```




```
res_writing$outmat[5,]
```

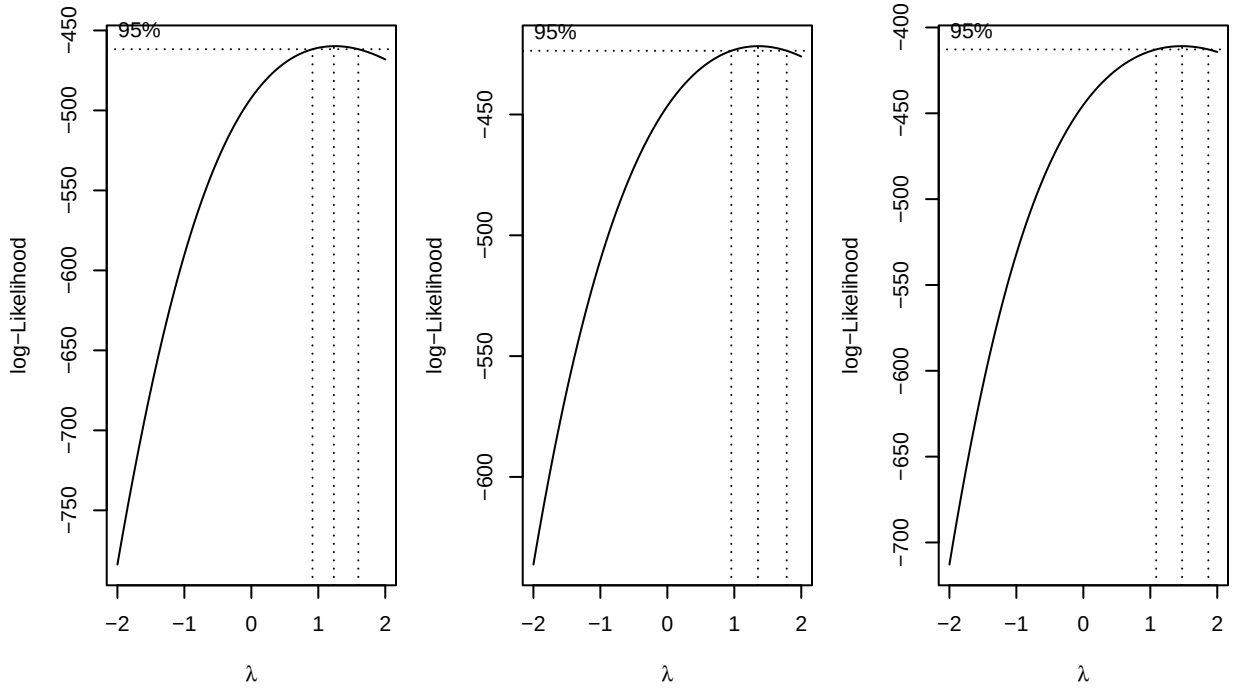
```
##                gender1          ethnic_group1          ethnic_group2
##                    "*"                    " "                    " "
##          ethnic_group3          ethnic_group4           parent_educ2
##                    " "                    " "                    " "
##           parent_educ3           parent_educ4            lunch_type1
##                    " "                    "*"                    "*"
##             test_prep1 parent_marital_status1 parent_marital_status2
##                    "*"                    " "                    " "
## parent_marital_status3        practice_sport1        practice_sport2
##                    " "                    " "                    " "
##        is_first_child1           nr_siblings1           nr_siblings2
##                    " "                    " "                    " "
##           nr_siblings3           nr_siblings4           nr_siblings5
##                    " "                    " "                    " "
##           nr_siblings6           nr_siblings7       transport_means1
##                    " "                    " "                    " "
##      wkly_study_hours1      wkly_study_hours2
##                    "*"                    " "
```

```
plot(writing_c, scale = "bic")
```



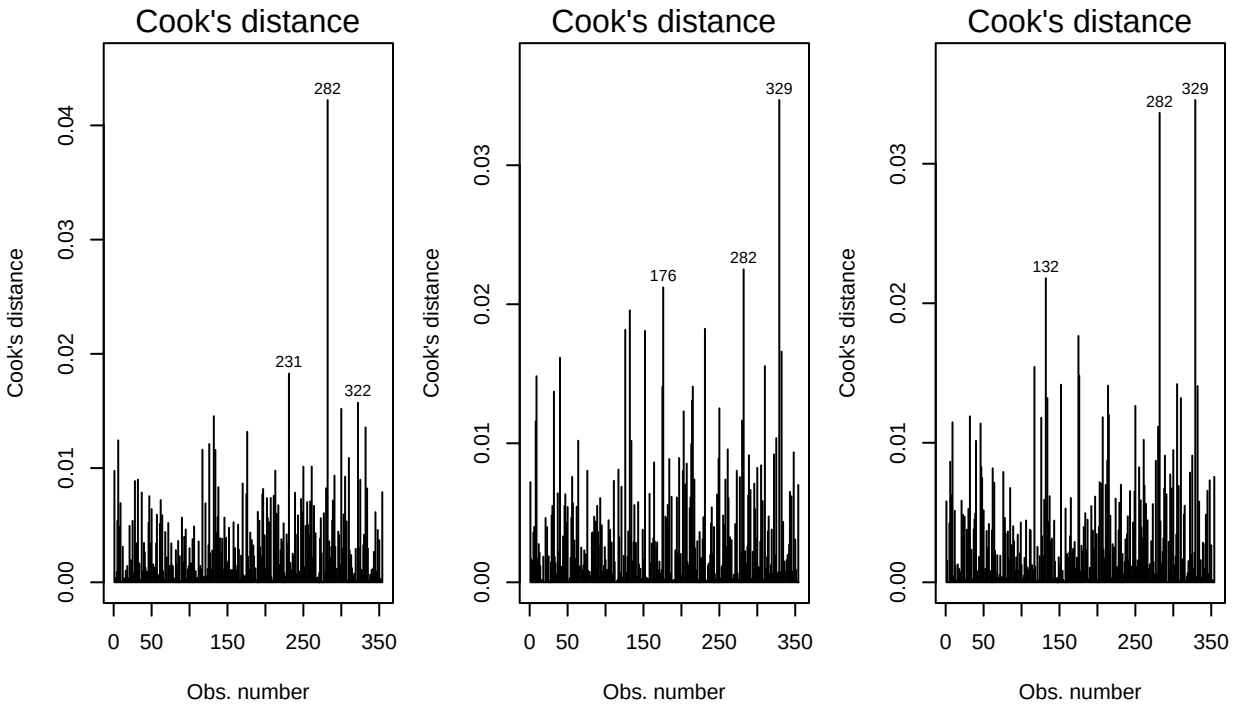
Transformation

```
par(mfrow = c(1, 3), mar = c(8, 4, 4, 1))
boxcox(model_math_full)
boxcox(model_reading_full)
boxcox(model_writing_full)
```



Outlier and influence points

```
par(mfrow = c(1, 3), mar = c(8, 4, 4, 1))
plot(model_math_full, which = 4)
plot(model_reading_full, which = 4)
plot(model_writing_full, which = 4)
```



Multicollinearity

```
# check VIF
vif_math =
  performance::check_collinearity(model_math_full) |>
  as_tibble() |>
  mutate(VIF_CI = str_c("[", round(VIF_CI_low, 1), ", ", round(VIF_CI_high, 1), "]")) |>
  dplyr::select(Term, VIF, VIF_CI, Tolerance)
knitr::kable(x = vif_math, caption = "VIF for Math Score", digits = 1)
```

Table 3: VIF for Math Score

Term	VIF	VIF_CI	Tolerance
gender	1.1	[1, 1.4]	0.9
ethnic_group	1.2	[1.1, 1.4]	0.8
parent_educ	1.2	[1.1, 1.4]	0.8
lunch_type	1.1	[1, 1.4]	1.0
test_prep	1.1	[1, 1.3]	0.9

Term	VIF	VIF_CI	Tolerance
parent_marital_status	1.2	[1.1, 1.4]	0.9
practice_sport	1.2	[1.1, 1.4]	0.9
is_first_child	1.2	[1.1, 1.3]	0.9
nr_siblings	1.5	[1.4, 1.8]	0.6
transport_means	1.1	[1, 1.3]	0.9
wkly_study_hours	1.1	[1.1, 1.3]	0.9

```
vif_reading =
  performance::check_collinearity(model_reading_full) |>
  as_tibble() |>
  mutate(VIF_CI = str_c("[", round(VIF_CI_low, 1), ", ", round(VIF_CI_high, 1), "]")) |>
  dplyr::select(Term, VIF, VIF_CI, Tolerance)
knitr::kable(x = vif_reading, caption = "VIF for Reading Score", digits = 1)
```

Table 4: VIF for Reading Score

Term	VIF	VIF_CI	Tolerance
gender	1.1	[1, 1.4]	0.9
ethnic_group	1.2	[1.1, 1.4]	0.8
parent_educ	1.2	[1.1, 1.4]	0.8
lunch_type	1.1	[1, 1.4]	1.0
test_prep	1.1	[1, 1.3]	0.9
parent_marital_status	1.2	[1.1, 1.4]	0.9
practice_sport	1.2	[1.1, 1.4]	0.9
is_first_child	1.2	[1.1, 1.3]	0.9
nr_siblings	1.5	[1.4, 1.8]	0.6
transport_means	1.1	[1, 1.3]	0.9
wkly_study_hours	1.1	[1.1, 1.3]	0.9

```
vif_writing =
  performance::check_collinearity(model_writing_full) |>
  as_tibble() |>
  mutate(VIF_CI = str_c("[", round(VIF_CI_low, 1), ", ", round(VIF_CI_high, 1), "]")) |>
  dplyr::select(Term, VIF, VIF_CI, Tolerance)
knitr::kable(x = vif_writing, caption = "VIF for Reading Score", digits = 1)
```

Table 5: VIF for Reading Score

Term	VIF	VIF_CI	Tolerance
gender	1.1	[1, 1.4]	0.9
ethnic_group	1.2	[1.1, 1.4]	0.8
parent_educ	1.2	[1.1, 1.4]	0.8
lunch_type	1.1	[1, 1.4]	1.0
test_prep	1.1	[1, 1.3]	0.9
parent_marital_status	1.2	[1.1, 1.4]	0.9
practice_sport	1.2	[1.1, 1.4]	0.9
is_first_child	1.2	[1.1, 1.3]	0.9
nr_siblings	1.5	[1.4, 1.8]	0.6

Term	VIF	VIF_CI	Tolerance
transport_means	1.1	[1, 1.3]	0.9
wkly_study_hours	1.1	[1.1, 1.3]	0.9

Model Selections

Stepwise Regressions

```
# math
math_sr = step(model_math_full, direction = 'both', trace = FALSE)

res_math_sr = math_sr |>
  summary()
res_math_sr
```

```
##
## Call:
## lm(formula = math_score ~ gender + ethnic_group + lunch_type +
##     test_prep + parent_marital_status + wkly_study_hours, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.985  -9.397   0.110  10.638  30.842
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               67.3260     3.2947  20.435  < 2e-16 ***
## gender1                   -3.7049     1.4716  -2.518 0.012274 *
## ethnic_group1              2.4461     3.2493   0.753 0.452084
## ethnic_group2              0.3026     3.0827   0.098 0.921866
## ethnic_group3              4.1687     3.1287   1.332 0.183624
## ethnic_group4             10.1791     3.3078   3.077 0.002258 **
## lunch_type1              -12.3773     1.5158  -8.166 6.28e-15 ***
## test_prep1                 6.0788     1.5224   3.993 8.00e-05 ***
## parent_marital_status1    -4.0821     1.7731  -2.302 0.021925 *
## parent_marital_status2     6.7982     4.6451   1.464 0.144250
## parent_marital_status3    -5.2507     2.1346  -2.460 0.014398 *
## wkly_study_hours1          5.9171     1.7108   3.459 0.000612 ***
## wkly_study_hours2          3.8301     2.2148   1.729 0.084647 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.53 on 341 degrees of freedom
## Multiple R-squared:  0.2989, Adjusted R-squared:  0.2742
## F-statistic: 12.11 on 12 and 341 DF,  p-value: < 2.2e-16
```

```
tb_math_sr = res_math_sr |>
  broom::tidy() |>
  filter(term != "(Intercept)") |>
```

```
  dplyr::select(term, estimate, p.value)
colnames(tb_math_sr) = c("Term", "Estimate", "P Value")
knitr::kable(x = tb_math_sr, caption = "Math Scores Models by Stepwise Regression", digits = 1)
```

Table 6: Math Scores Models by Stepwise Regression

Term	Estimate	P Value
gender1	-3.7	0.0
ethnic_group1	2.4	0.5
ethnic_group2	0.3	0.9
ethnic_group3	4.2	0.2
ethnic_group4	10.2	0.0
lunch_type1	-12.4	0.0
test_prep1	6.1	0.0
parent_marital_status1	-4.1	0.0
parent_marital_status2	6.8	0.1
parent_marital_status3	-5.3	0.0
wkly_study_hours1	5.9	0.0
wkly_study_hours2	3.8	0.1

```
# reading
rea_sr = step(model_reading_full, direction = 'both')
```

```
## Start:  AIC=1836.68
## 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 + math_score +
##     writing_score) - math_score - writing_score
##
##                         Df Sum of Sq   RSS    AIC
## - nr_siblings            7     887.9 55342 1828.4
## - practice_sport         2     123.8 54578 1833.5
## - transport_means        1      21.6 54476 1834.8
## - is_first_child         1      27.9 54482 1834.9
## - ethnic_group           4    1227.5 55682 1836.6
## <none>                               54454 1836.7
## - parent_educ            3    1558.4 56013 1840.7
## - parent_marital_status  3    1908.7 56363 1842.9
## - wkly_study_hours       2    2004.0 56459 1845.5
## - test_prep              1    4305.6 58760 1861.6
## - lunch_type             1    4793.1 59248 1864.5
## - gender                 1    5599.8 60054 1869.3
##
## Step:  AIC=1828.41
## reading_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + parent_marital_status + practice_sport + is_first_child +
##     transport_means + wkly_study_hours
##
##                         Df Sum of Sq   RSS    AIC
## - practice_sport         2     145.3 55488 1825.3
## - transport_means        1      11.4 55354 1826.5
```

```
## - is_first_child         1      40.1 55382 1826.7
## <none>                                55342 1828.4
## - ethnic_group           4    1318.8 56661 1828.7
## - parent_educ            3    1681.4 57024 1833.0
## - parent_marital_status  3    1924.1 57267 1834.5
## + nr_siblings            7     887.9 54454 1836.7
## - wkly_study_hours       2    1969.7 57312 1836.8
## - test_prep              1    4222.4 59565 1852.4
## - lunch_type             1    5437.8 60780 1859.6
## - gender                 1    5693.8 61036 1861.1
##
## Step:  AIC=1825.34
## reading_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + parent_marital_status + is_first_child + transport_means +
##     wkly_study_hours
##
##                         Df Sum of Sq   RSS    AIC
## - transport_means        1       5.8 55493 1823.4
## - is_first_child         1      40.9 55529 1823.6
## <none>                                55488 1825.3
## - ethnic_group           4    1294.8 56782 1825.5
## + practice_sport         2     145.3 55342 1828.4
## - parent_educ            3    1654.8 57143 1829.7
## - parent_marital_status  3    1902.9 57391 1831.3
## + nr_siblings            7     909.4 54578 1833.5
## - wkly_study_hours       2    1959.0 57447 1833.6
## - test_prep              1    4316.3 59804 1849.8
## - lunch_type             1    5421.7 60909 1856.3
## - gender                 1    5678.5 61166 1857.8
##
## Step:  AIC=1823.37
## reading_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + parent_marital_status + is_first_child + wkly_study_hours
##
##                         Df Sum of Sq   RSS    AIC
## - is_first_child         1      39.4 55533 1821.6
## <none>                                55493 1823.4
## - ethnic_group           4    1295.8 56789 1823.5
## + transport_means        1       5.8 55488 1825.3
## + practice_sport         2     139.7 55354 1826.5
## - parent_educ            3    1649.4 57143 1827.7
## - parent_marital_status  3    1899.1 57393 1829.3
## + nr_siblings            7     901.8 54592 1831.6
## - wkly_study_hours       2    1958.5 57452 1831.7
## - test_prep              1    4422.7 59916 1848.5
## - lunch_type             1    5422.5 60916 1854.4
## - gender                 1    5674.9 61168 1855.8
##
## Step:  AIC=1821.62
## reading_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + parent_marital_status + wkly_study_hours
##
##                         Df Sum of Sq   RSS    AIC
## <none>                                55533 1821.6
```

```
## - ethnic_group           4    1305.9 56839 1821.8
## + is_first_child         1      39.4 55493 1823.4
## + transport_means        1       4.3 55529 1823.6
## + practice_sport         2     140.5 55392 1824.7
## - parent_educ            3    1654.8 57188 1826.0
## - parent_marital_status  3    1899.5 57432 1827.5
## + nr_siblings            7     917.5 54615 1829.7
## - wkly_study_hours       2    1974.9 57508 1830.0
## - test_prep              1    4531.6 60064 1847.4
## - lunch_type             1    5440.2 60973 1852.7
## - gender                 1    5644.2 61177 1853.9
```

```
res_rea_sr = rea_sr |>
  summary()
res_rea_sr
```

```
##
## Call:
## lm(formula = reading_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep + parent_marital_status + wkly_study_hours,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.083  -9.288   0.232   9.530  30.265
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              61.6474     3.2387  19.035  < 2e-16 ***
## gender1                   8.1816     1.3959   5.861 1.09e-08 ***
## ethnic_group1             1.8945     3.0850   0.614  0.53956
## ethnic_group2             0.3778     2.9216   0.129  0.89720
## ethnic_group3             3.3789     2.9670   1.139  0.25559
## ethnic_group4             5.6870     3.1395   1.811  0.07096 .
## parent_educ2              2.3964     1.6650   1.439  0.15101
## parent_educ3              4.6728     1.9527   2.393  0.01726 *
## parent_educ4              6.4917     2.3912   2.715  0.00697 **
## lunch_type1              -8.2631     1.4360  -5.754 1.95e-08 ***
## test_prep1                7.6175     1.4505   5.252 2.67e-07 ***
## parent_marital_status1   -4.5976     1.6839  -2.730  0.00666 **
## parent_marital_status2    4.1841     4.4058   0.950  0.34296
## parent_marital_status3   -4.3042     2.0287  -2.122  0.03460 *
## wkly_study_hours1         5.1565     1.6242   3.175  0.00164 **
## wkly_study_hours2         1.0458     2.1132   0.495  0.62102
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.82 on 338 degrees of freedom
## Multiple R-squared:  0.2831, Adjusted R-squared:  0.2513
## F-statistic:   8.9 on 15 and 338 DF,  p-value: < 2.2e-16
```

```
tb_rea_sr = res_rea_sr |>
  broom::tidy() |>
```

```
  filter(term != "(Intercept)") |>
  dplyr::select(term, estimate, p.value)
colnames(tb_rea_sr) = c("Term", "Estimate", "P Value")
knitr::kable(x = tb_rea_sr, caption = "Reading Scores Models by Stepwise Regression", digits = 1)
```

Table 7: Reading Scores Models by Stepwise Regression

Term	Estimate	P Value
gender1	8.2	0.0
ethnic_group1	1.9	0.5
ethnic_group2	0.4	0.9
ethnic_group3	3.4	0.3
ethnic_group4	5.7	0.1
parent_educ2	2.4	0.2
parent_educ3	4.7	0.0
parent_educ4	6.5	0.0
lunch_type1	-8.3	0.0
test_prep1	7.6	0.0
parent_marital_status1	-4.6	0.0
parent_marital_status2	4.2	0.3
parent_marital_status3	-4.3	0.0
wkly_study_hours1	5.2	0.0
wkly_study_hours2	1.0	0.6

```
# writing
wri_sr = step(model_writing_full, direction = 'both')
```

```
## Start:  AIC=1813.9
## 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 + math_score +
##     reading_score) - reading_score - math_score
##
##                         Df Sum of Sq   RSS    AIC
## - nr_siblings            7    1019.1 52079 1806.9
## - is_first_child         1       4.4 51064 1811.9
## - practice_sport         2     361.2 51421 1812.4
## - transport_means        1      74.2 51134 1812.4
## <none>                               51060 1813.9
## - ethnic_group           4    1779.1 52839 1818.0
## - parent_educ            3    1940.3 53000 1821.1
## - parent_marital_status  3    1991.7 53052 1821.4
## - wkly_study_hours       2    1901.4 52961 1822.8
## - lunch_type             1    6175.3 57235 1852.3
## - test_prep              1    6924.6 57985 1856.9
## - gender                 1    8281.3 59341 1865.1
##
## Step:  AIC=1806.89
## writing_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + parent_marital_status + practice_sport + is_first_child +
##     transport_means + wkly_study_hours
```

```
##
##                         Df Sum of Sq   RSS    AIC
## - is_first_child         1       1.2 52080 1804.9
## - transport_means        1      52.4 52132 1805.2
## - practice_sport         2     404.8 52484 1805.6
## <none>                               52079 1806.9
## - ethnic_group           4    1870.2 53949 1811.4
## + nr_siblings            7    1019.1 51060 1813.9
## - parent_marital_status  3    2027.5 54107 1814.4
## - parent_educ            3    2069.1 54148 1814.7
## - wkly_study_hours       2    1830.3 53910 1815.1
## - test_prep              1    6879.5 58959 1848.8
## - lunch_type             1    6955.3 59035 1849.3
## - gender                 1    8444.0 60523 1858.1
##
## Step:  AIC=1804.9
## writing_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + parent_marital_status + practice_sport + transport_means +
##     wkly_study_hours
##
##                         Df Sum of Sq   RSS    AIC
## - transport_means        1      53.0 52133 1803.3
## - practice_sport         2     408.3 52489 1803.7
## <none>                               52080 1804.9
## + is_first_child         1       1.2 52079 1806.9
## - ethnic_group           4    1869.4 53950 1809.4
## + nr_siblings            7    1015.9 51064 1811.9
## - parent_marital_status  3    2028.9 54109 1812.4
## - parent_educ            3    2068.7 54149 1812.7
## - wkly_study_hours       2    1829.2 53910 1813.1
## - test_prep              1    6907.4 58988 1847.0
## - lunch_type             1    6954.4 59035 1847.3
## - gender                 1    8463.2 60544 1856.2
##
## Step:  AIC=1803.26
## writing_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + parent_marital_status + practice_sport + wkly_study_hours
##
##                         Df Sum of Sq   RSS    AIC
## - practice_sport         2     397.6 52531 1802.0
## <none>                               52133 1803.3
## + transport_means        1      53.0 52080 1804.9
## + is_first_child         1       1.8 52132 1805.2
## - ethnic_group           4    1901.9 54035 1808.0
## + nr_siblings            7     992.6 51141 1810.5
## - parent_marital_status  3    1986.8 54120 1810.5
## - parent_educ            3    2041.4 54175 1810.9
## - wkly_study_hours       2    1821.0 53954 1811.4
## - lunch_type             1    6905.0 59038 1845.3
## - test_prep              1    7190.9 59324 1847.0
## - gender                 1    8443.2 60577 1854.4
##
## Step:  AIC=1801.95
## writing_score ~ gender + ethnic_group + parent_educ + lunch_type +
```

```
##     test_prep + parent_marital_status + wkly_study_hours
##
##                          Df Sum of Sq   RSS    AIC
## <none>                                 52531 1802.0
## + practice_sport          2     397.6 52133 1803.3
## + transport_means         1      42.3 52489 1803.7
## + is_first_child          1       6.3 52525 1803.9
## - ethnic_group            4    1950.7 54482 1806.9
## - parent_educ             3    1925.8 54457 1808.7
## + nr_siblings             7    1035.0 51496 1808.9
## - parent_marital_status   3    1962.6 54494 1808.9
## - wkly_study_hours        2    1804.0 54335 1809.9
## - lunch_type              1    6837.1 59368 1843.3
## - test_prep               1    7210.3 59741 1845.5
## - gender                  1    8486.0 61017 1853.0
```

```
res_wri_sr = wri_sr |>
  summary()
res_wri_sr
```

```
##
## Call:
## lm(formula = writing_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep + parent_marital_status + wkly_study_hours,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.350  -8.531   0.962   9.592  25.874
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)               58.522      3.150  18.579  < 2e-16 ***
## gender1                   10.032      1.358   7.389 1.17e-12 ***
## ethnic_group1              2.213      3.000   0.738  0.46124
## ethnic_group2              1.850      2.842   0.651  0.51544
## ethnic_group3              6.338      2.886   2.196  0.02874 *
## ethnic_group4              6.617      3.053   2.167  0.03094 *
## parent_educ2               1.789      1.619   1.105  0.27005
## parent_educ3               4.598      1.899   2.421  0.01599 *
## parent_educ4               7.212      2.326   3.101  0.00209 **
## lunch_type1               -9.263      1.397  -6.633 1.31e-10 ***
## test_prep1                 9.609      1.411   6.811 4.44e-11 ***
## parent_marital_status1    -4.417      1.638  -2.697  0.00734 **
## parent_marital_status2     4.668      4.285   1.089  0.27678
## parent_marital_status3    -4.644      1.973  -2.353  0.01917 *
## wkly_study_hours1          5.168      1.580   3.271  0.00118 **
## wkly_study_hours2          1.893      2.055   0.921  0.35769
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.47 on 338 degrees of freedom
## Multiple R-squared:  0.3583, Adjusted R-squared:  0.3298
## F-statistic: 12.58 on 15 and 338 DF,  p-value: < 2.2e-16
```

```r
tb_wri_sr = res_wri_sr |>
  broom::tidy() |>
  filter(term != "(Intercept)") |>
  dplyr::select(term, estimate, p.value)
colnames(tb_wri_sr) = c("Term", "Estimate", "P Value")
knitr::kable(x = tb_wri_sr, caption = "Writing Scores Models by Stepwise Regression", digits = 1)
```

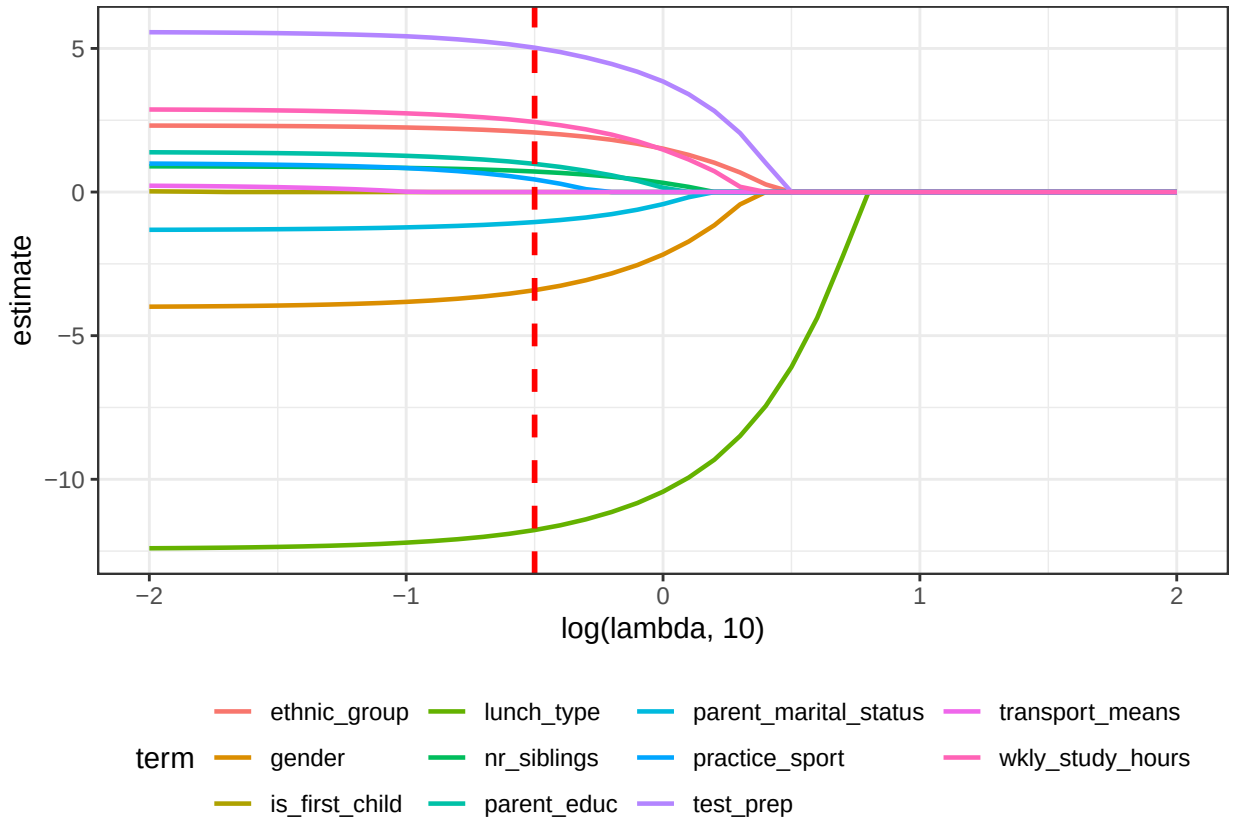
Table 8: Writing Scores Models by Stepwise Regression

Term	Estimate	P Value
gender1	10.0	0.0
ethnic_group1	2.2	0.5
ethnic_group2	1.9	0.5
ethnic_group3	6.3	0.0
ethnic_group4	6.6	0.0
parent_educ2	1.8	0.3
parent_educ3	4.6	0.0
parent_educ4	7.2	0.0
lunch_type1	-9.3	0.0
test_prep1	9.6	0.0
parent_marital_status1	-4.4	0.0
parent_marital_status2	4.7	0.3
parent_marital_status3	-4.6	0.0
wkly_study_hours1	5.2	0.0
wkly_study_hours2	1.9	0.4

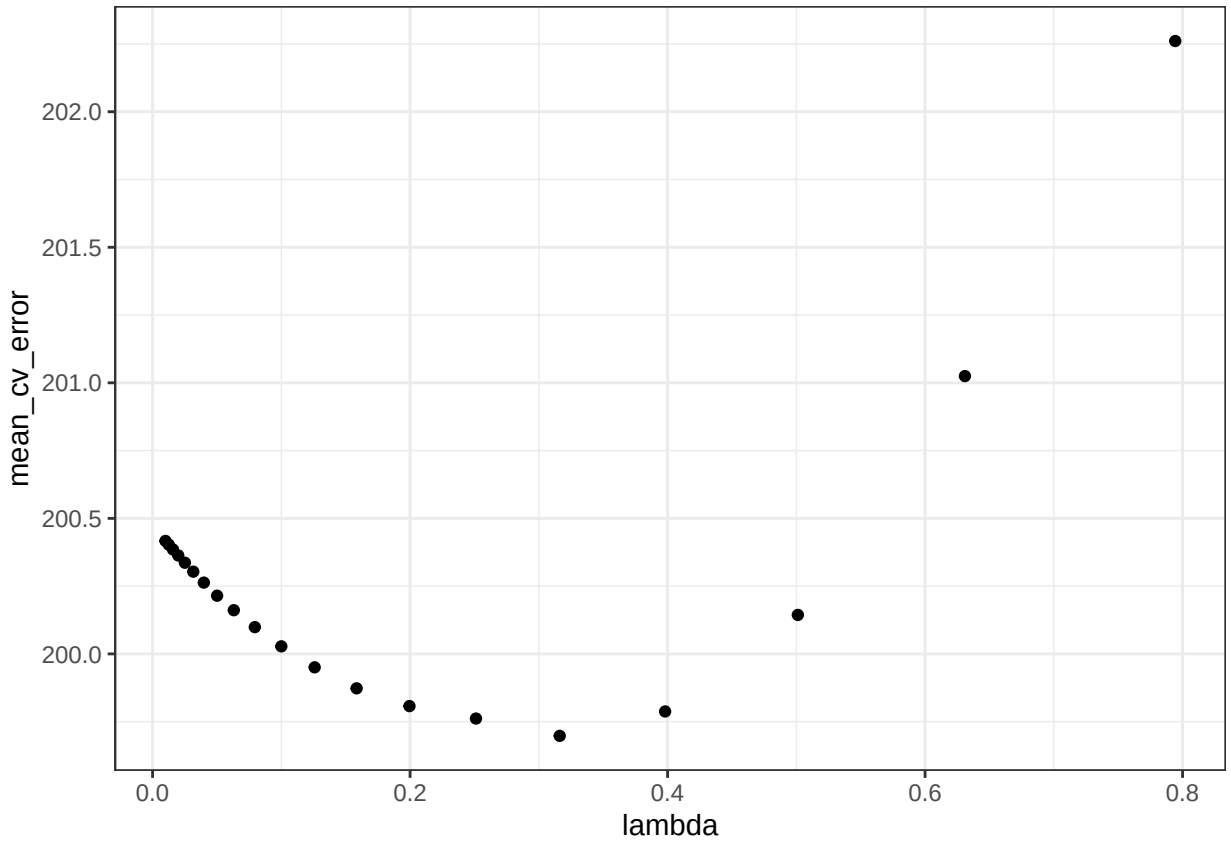
Lasso Models

```r
lambda_seq = 10 ^ seq(-2, 2, by = .1)

#math
cv_object_math =
  cv.glmnet(as.matrix(data[1:11]), data$math_score,
                          lambda = lambda_seq,
                          nfolds = 5)

opt_lambda_math = cv_object_math$lambda.min

#variables contraction
glmnet(as.matrix(data[1:11]), data$math_score, lambda = lambda_seq) |>
  broom::tidy() |>
  dplyr::select(term, lambda, estimate) |>
  complete(term, lambda, fill = list(estimate = 0) ) |>
  filter(term != "(Intercept)") |>
  ggplot(aes(x = log(lambda, 10), y = estimate, group = term, color = term)) +
  geom_path(size = 0.8) +
  geom_vline(xintercept = log(opt_lambda_math, 10), color = "red", linetype = "dashed", size = 1) +
  theme_bw() +
  theme(legend.position = "bottom")
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```




```r
tb_la_math = tibble(
  lambda = cv_object_math$lambda,
  mean_cv_error = cv_object_math$cvm) |>
  filter(lambda < 1)

#choosing optimal lambda
tb_la_math |>
  ggplot(aes(x = lambda, y = mean_cv_error)) +
  geom_point() +
  theme_bw()
```




```
#math result
model_math_lasso = glmnet(as.matrix(data[1:11]), data$math_score, lambda = opt_lambda_math, alpha = 1)
coef(model_math_lasso)
```

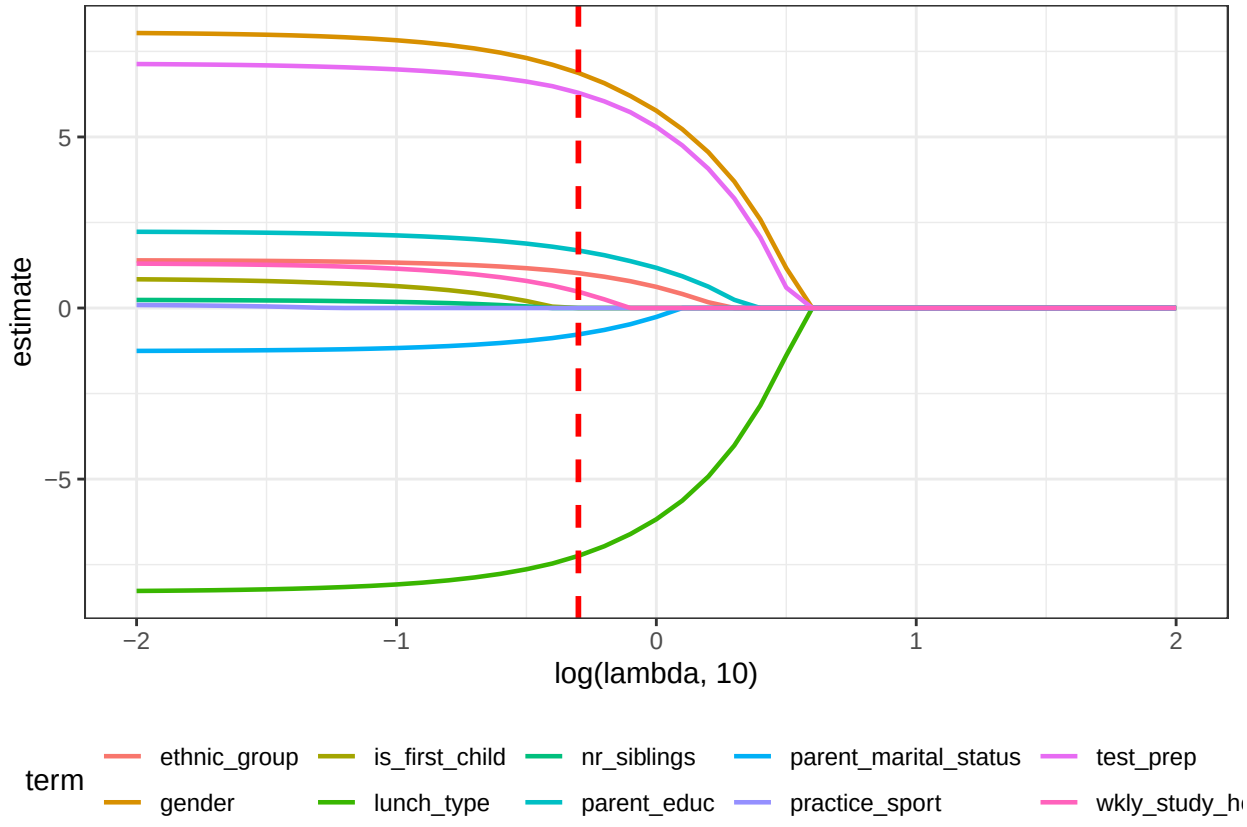
```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                               s0
## (Intercept)           62.7158706
## gender                -3.4172517
## ethnic_group           2.0740949
## parent_educ            0.9804808
## lunch_type           -11.7678104
## test_prep              5.0255504
## parent_marital_status -1.0446103
## practice_sport         0.4391390
## is_first_child         .
## nr_siblings            0.7146589
## transport_means        .
## wkly_study_hours       2.4395500
```

```
model_math_lasso$dev.ratio
```

```
## [1] 0.2622201
```

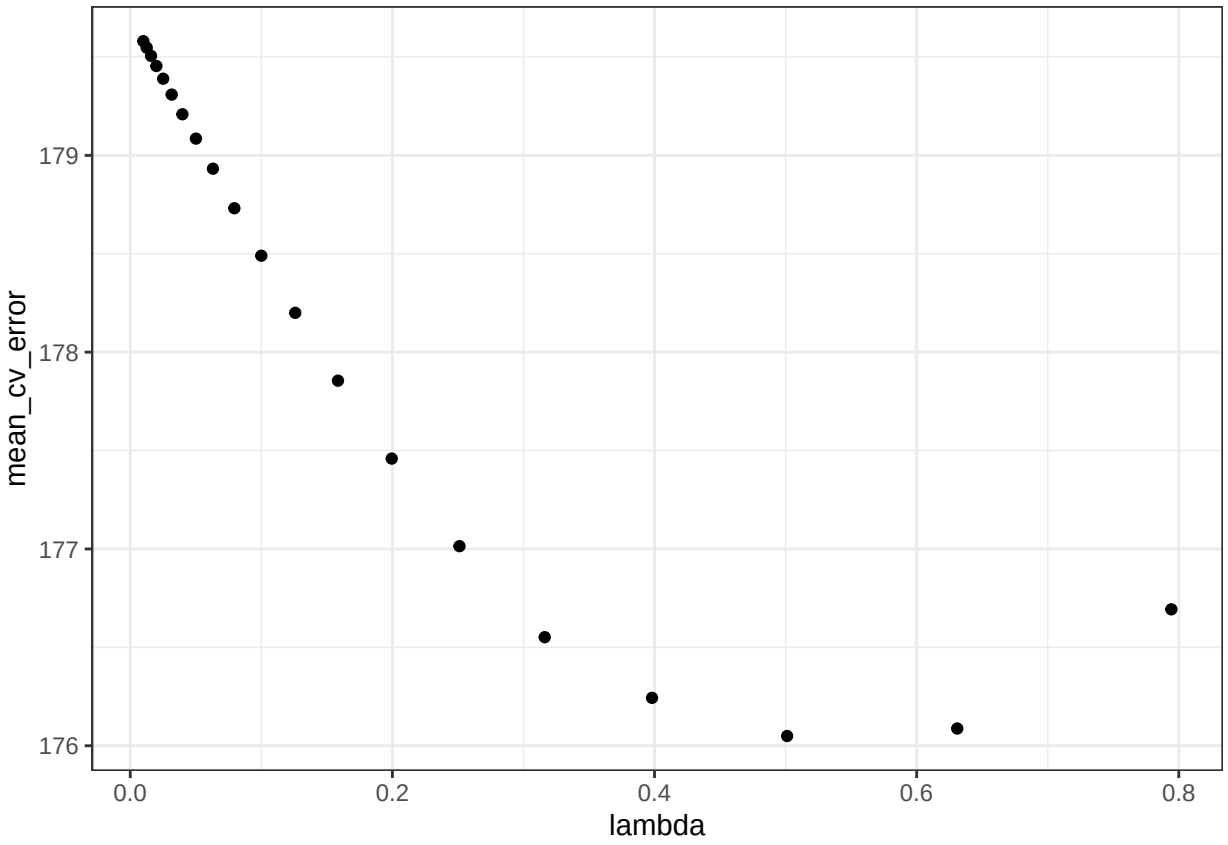
```
#reading
cv_object_reading =
```

```
  cv.glmnet(as.matrix(data[1:11]), data$reading_score,
                          lambda = lambda_seq,
                          nfolds = 5)

opt_lambda_reading = cv_object_reading$lambda.min

#variables contraction
glmnet(as.matrix(data[1:11]), data$reading_score, lambda = lambda_seq) |>
  broom::tidy() |>
  dplyr::select(term, lambda, estimate) |>
  complete(term, lambda, fill = list(estimate = 0) ) |>
  filter(term != "(Intercept)") |>
  ggplot(aes(x = log(lambda, 10), y = estimate, group = term, color = term)) +
  geom_path(size = 0.8) +
  geom_vline(xintercept = log(opt_lambda_reading, 10), color = "red", linetype = "dashed", size = 1) +
  theme_bw() +
  theme(legend.position = "bottom")
```




```
tb_la_reading = tibble(
  lambda = cv_object_reading$lambda,
  mean_cv_error = cv_object_reading$cvm) |>
  filter(lambda < 1)

#choosing optimal lambda
tb_la_reading |>
  ggplot(aes(x = lambda, y = mean_cv_error)) +
```

```
  geom_point() +
  theme_bw()
```



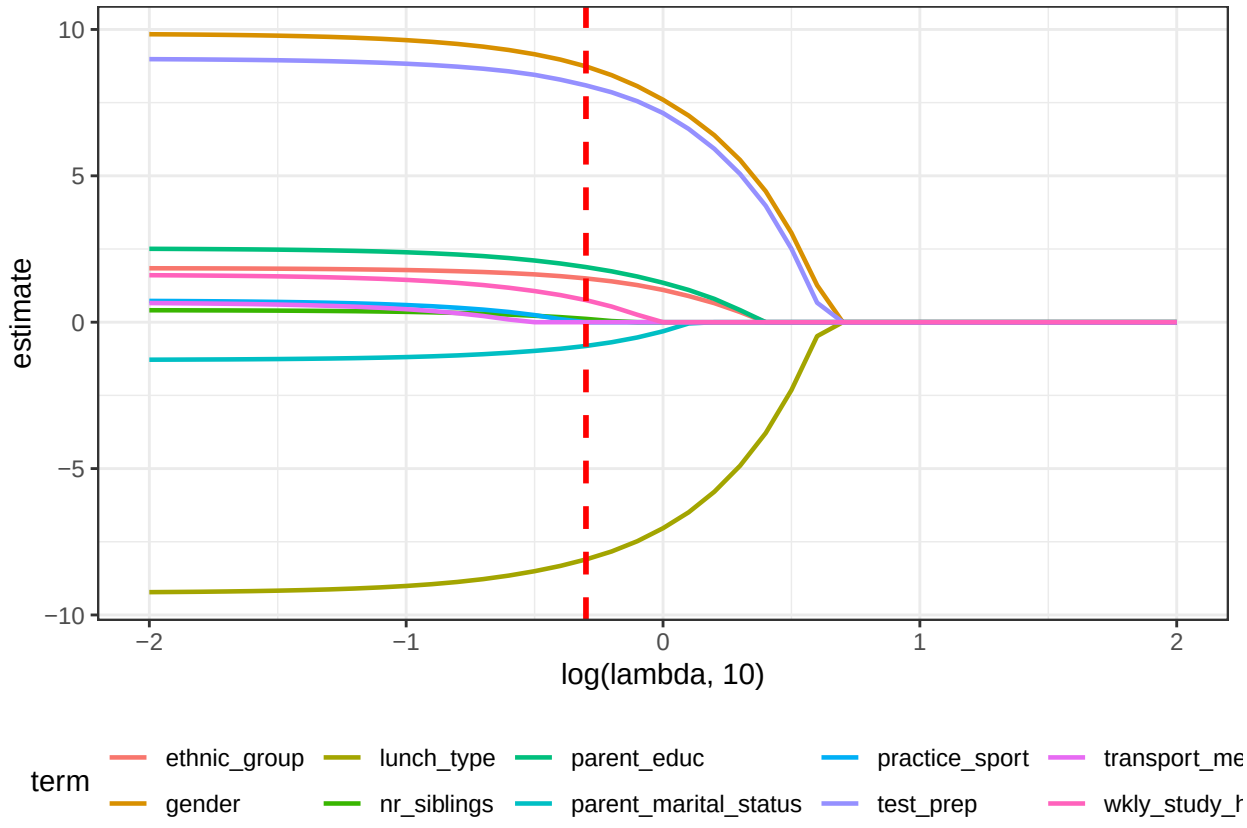

```
#reading result
model_reading_lasso = glmnet(as.matrix(data[1:11]), data$reading_score, lambda = opt_lambda_math, alpha
coef(model_reading_lasso)
```

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                                  s0
## (Intercept)           61.64060428
## gender                 7.30735439
## ethnic_group           1.16352168
## parent_educ            1.88265624
## lunch_type            -7.63626242
## test_prep              6.61888731
## parent_marital_status -0.95887140
## practice_sport         .
## is_first_child         0.20510870
## nr_siblings            0.04774974
## transport_means        .
## wkly_study_hours       0.79132994
```

```
model_reading_lasso$dev.ratio
```

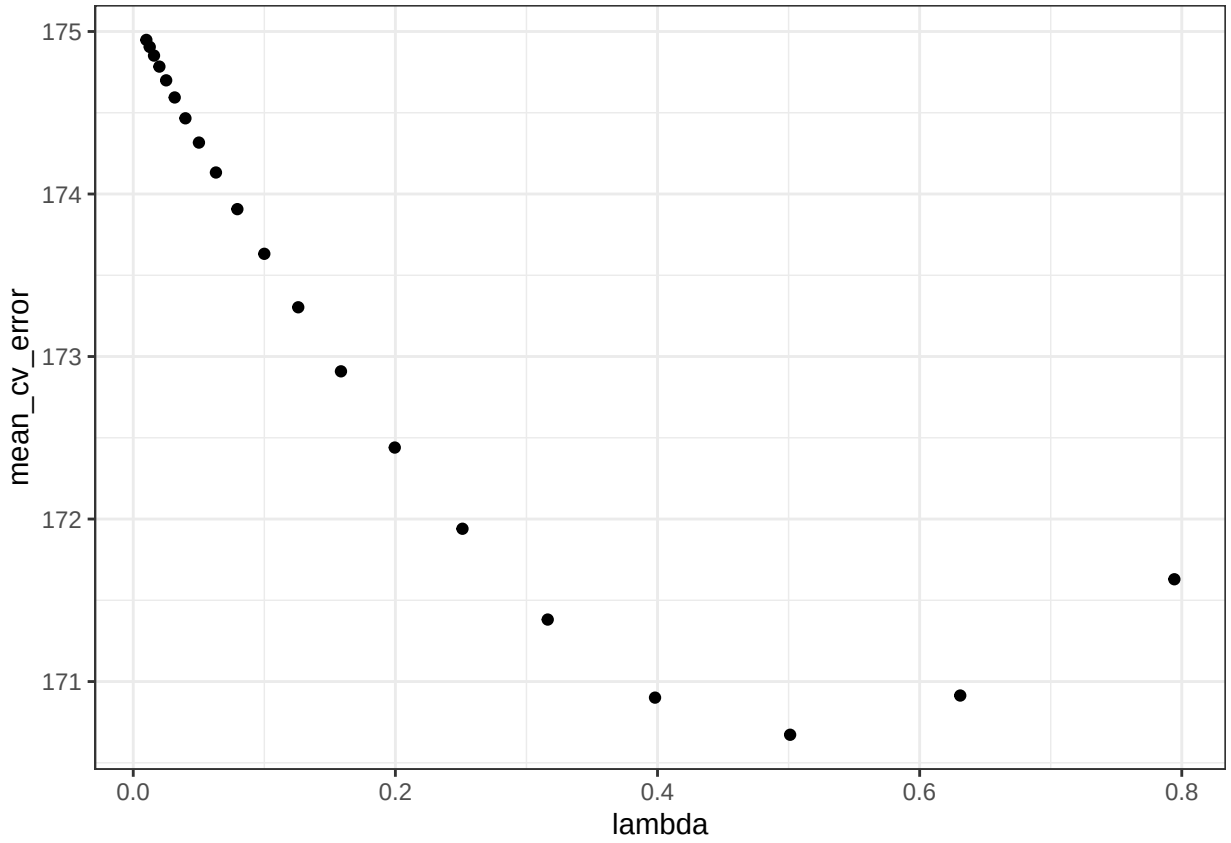
```
## [1] 0.2357325
```

```r
#writing
cv_object_writing =
  cv.glmnet(as.matrix(data[1:11]), data$writing_score,
                          lambda = lambda_seq,
                          nfolds = 5)

opt_lambda_writing = cv_object_writing$lambda.min

#variables contraction
glmnet(as.matrix(data[1:11]), data$writing_score, lambda = lambda_seq) |>
  broom::tidy() |>
  dplyr::select(term, lambda, estimate) |>
  complete(term, lambda, fill = list(estimate = 0) ) |>
  filter(term != "(Intercept)") |>
  ggplot(aes(x = log(lambda, 10), y = estimate, group = term, color = term)) +
  geom_path(size = 0.8) +
  geom_vline(xintercept = log(opt_lambda_writing, 10), color = "red", linetype = "dashed", size = 1) +
  theme_bw() +
  theme(legend.position = "bottom")
```




```r
tb_la_writing = tibble(
  lambda = cv_object_writing$lambda,
  mean_cv_error = cv_object_writing$cvm) |>
  filter(lambda < 1)

#choosing optimal lambda
tb_la_writing |>
```

```
  ggplot(aes(x = lambda, y = mean_cv_error)) +
  geom_point() +
  theme_bw()
```




```
#writing result
model_writing_lasso = glmnet(as.matrix(data[1:11]), data$writing_score, lambda = opt_lambda_writing, al
coef(model_writing_lasso)
```

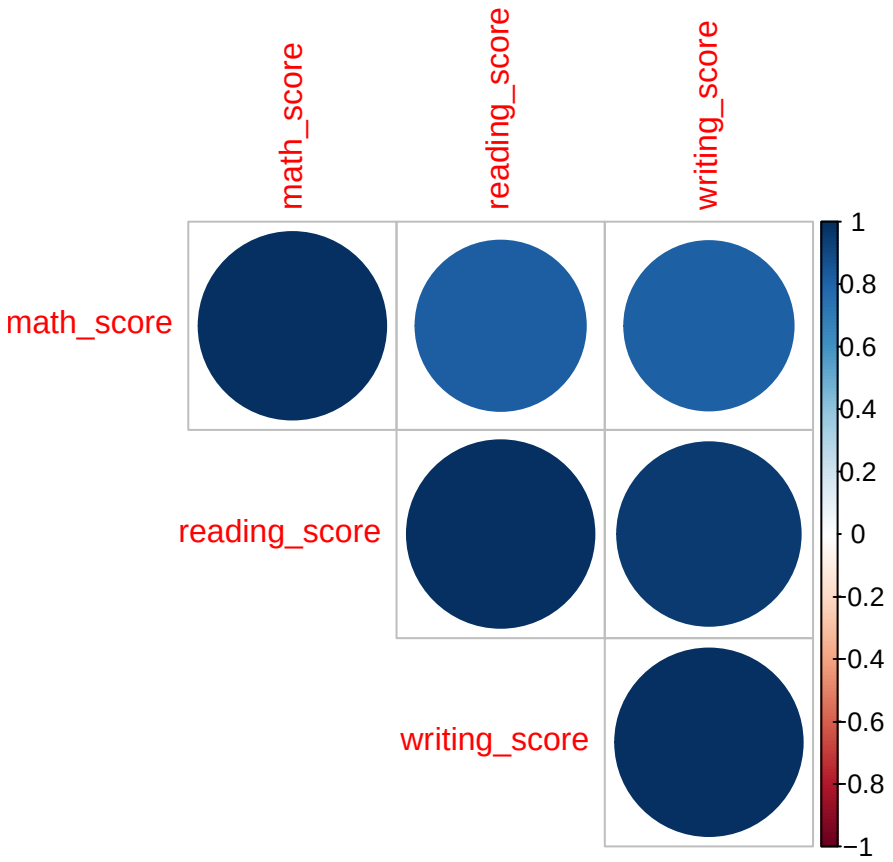
```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                                s0
## (Intercept)           59.3844759
## gender                 8.7384396
## ethnic_group           1.4955961
## parent_educ            1.8826016
## lunch_type            -8.1037819
## test_prep              8.0886240
## parent_marital_status -0.8123378
## practice_sport         .
## is_first_child         .
## nr_siblings            0.1133873
## transport_means        .
## wkly_study_hours       0.7539334
```

```
model_writing_lasso$dev.ratio
```

```
## [1] 0.3119987
```

adding one score variable

```
#correlation plot for scores
corrplot::corrplot(cor(data[12:14]), type = "upper")
```




```
# Math
math_reading = lm(math_score ~ . - writing_score, data = data) |>
  step(direction = "both")
```

```
## Start:  AIC=1306.61
## 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 + reading_score +
##     writing_score) - writing_score
##
##                          Df Sum of Sq   RSS    AIC
## - parent_marital_status   3       100 12214 1303.5
## - parent_educ             3       128 12241 1304.3
## - nr_siblings             7       418 12532 1304.6
## - transport_means         1        23 12137 1305.3
## - practice_sport          2       124 12238 1306.2
## - is_first_child          1        58 12172 1306.3
## <none>                               12114 1306.6
## - test_prep               1       123 12237 1308.2
## - wkly_study_hours        2       368 12482 1313.2
```

```
## - ethnic_group           4       927 13040 1324.7
## - lunch_type             1      1591 13705 1348.3
## - gender                 1      9679 21793 1512.5
## - reading_score          1     47954 60068 1871.4
##
## Step:  AIC=1303.53
## math_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + practice_sport + is_first_child + nr_siblings +
##     transport_means + wkly_study_hours + reading_score
##
##                         Df Sum of Sq   RSS    AIC
## - parent_educ            3       131 12345 1301.3
## - transport_means        1        18 12232 1302.0
## - nr_siblings            7       446 12660 1302.2
## - practice_sport         2       116 12330 1302.9
## <none>                               12214 1303.5
## - is_first_child         1        78 12292 1303.8
## - test_prep              1       122 12336 1305.0
## + parent_marital_status  3       100 12114 1306.6
## - wkly_study_hours       2       369 12583 1310.1
## - ethnic_group           4       907 13121 1320.9
## - lunch_type             1      1553 13767 1343.9
## - gender                 1      9860 22074 1511.0
## - reading_score          1     50046 62260 1878.1
##
## Step:  AIC=1301.3
## math_score ~ gender + ethnic_group + lunch_type + test_prep +
##     practice_sport + is_first_child + nr_siblings + transport_means +
##     wkly_study_hours + reading_score
##
##                         Df Sum of Sq   RSS    AIC
## - transport_means        1        24 12369 1300.0
## <none>                               12345 1301.3
## - nr_siblings            7       503 12848 1301.4
## - practice_sport         2       147 12492 1301.5
## - is_first_child         1        80 12425 1301.6
## - test_prep              1       126 12470 1302.9
## + parent_educ            3       131 12214 1303.5
## + parent_marital_status  3       103 12241 1304.3
## - wkly_study_hours       2       416 12760 1309.0
## - ethnic_group           4       918 13262 1318.7
## - lunch_type             1      1622 13966 1343.0
## - gender                 1      9838 22183 1506.8
## - reading_score          1     50580 62925 1875.9
##
## Step:  AIC=1299.99
## math_score ~ gender + ethnic_group + lunch_type + test_prep +
##     practice_sport + is_first_child + nr_siblings + wkly_study_hours +
##     reading_score
##
##                         Df Sum of Sq   RSS    AIC
## - nr_siblings            7       497 12866 1299.9
## <none>                               12369 1300.0
## - practice_sport         2       153 12522 1300.3
```

```
## - is_first_child         1        84 12453 1300.4
## - test_prep              1       116 12485 1301.3
## + transport_means        1        24 12345 1301.3
## + parent_educ            3       137 12232 1302.0
## + parent_marital_status  3        97 12272 1303.2
## - wkly_study_hours       2       405 12774 1307.4
## - ethnic_group           4       907 13276 1317.0
## - lunch_type             1      1603 13972 1341.1
## - gender                 1      9833 22202 1505.1
## - reading_score          1     50565 62934 1873.9
##
## Step:  AIC=1299.94
## math_score ~ gender + ethnic_group + lunch_type + test_prep +
##     practice_sport + is_first_child + wkly_study_hours + reading_score
##
##                         Df Sum of Sq   RSS    AIC
## - is_first_child         1        70 12936 1299.8
## <none>                                12866 1299.9
## + nr_siblings            7       497 12369 1300.0
## - practice_sport         2       159 13025 1300.3
## - test_prep              1        88 12954 1300.4
## + parent_educ            3       192 12674 1300.6
## + transport_means        1        18 12848 1301.4
## + parent_marital_status  3       128 12738 1302.4
## - wkly_study_hours       2       436 13302 1307.7
## - ethnic_group           4       957 13823 1317.3
## - lunch_type             1      1631 14497 1340.2
## - gender                 1     10310 23176 1506.3
## - reading_score          1     51613 64479 1868.5
##
## Step:  AIC=1299.85
## math_score ~ gender + ethnic_group + lunch_type + test_prep +
##     practice_sport + wkly_study_hours + reading_score
##
##                         Df Sum of Sq   RSS    AIC
## <none>                                12936 1299.8
## + is_first_child         1        70 12866 1299.9
## + nr_siblings            7       483 12453 1300.4
## - test_prep              1        97 13033 1300.5
## + parent_educ            3       194 12742 1300.5
## - practice_sport         2       195 13131 1301.2
## + transport_means        1        20 12916 1301.3
## + parent_marital_status  3       146 12790 1301.8
## - wkly_study_hours       2       427 13363 1307.3
## - ethnic_group           4       940 13875 1316.7
## - lunch_type             1      1627 14563 1339.8
## - gender                 1     10255 23191 1504.5
## - reading_score          1     51552 64488 1866.5
```

```
math_reading |> summary()
```

```
##
## Call:
## lm(formula = math_score ~ gender + ethnic_group + lunch_type +
```

```
##     test_prep + practice_sport + wkly_study_hours + reading_score,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.0114  -4.4424   0.3468   3.9324  16.4718
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.68933    2.28581   2.489 0.013288 *
## gender1           -11.48241    0.69837 -16.442  < 2e-16 ***
## ethnic_group1       0.57151    1.48058   0.386 0.699733
## ethnic_group2      -0.17064    1.40777  -0.121 0.903593
## ethnic_group3       0.73640    1.43230   0.514 0.607488
## ethnic_group4       4.63904    1.51388   3.064 0.002355 **
## lunch_type1        -4.72335    0.72113  -6.550 2.12e-10 ***
## test_prep1         -1.15663    0.72328  -1.599 0.110716
## practice_sport1     1.41324    1.05323   1.342 0.180550
## practice_sport2     2.39702    1.09146   2.196 0.028753 *
## wkly_study_hours1   1.43080    0.78754   1.817 0.070127 .
## wkly_study_hours2   3.37591    1.00648   3.354 0.000886 ***
## reading_score       0.93521    0.02537  36.864  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.159 on 341 degrees of freedom
## Multiple R-squared:  0.8548, Adjusted R-squared:  0.8497
## F-statistic: 167.3 on 12 and 341 DF,  p-value: < 2.2e-16
```

```
math_writing = lm(math_score ~ . - reading_score, data = data) |>
  step(direction = "both")
```

```
## Start:  AIC=1235.37
## 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 + reading_score +
##     writing_score) - reading_score
##
##                          Df Sum of Sq   RSS    AIC
## - nr_siblings             7       238 10143 1229.8
## - parent_marital_status   3        48  9954 1231.1
## - practice_sport          2        14  9920 1231.9
## - is_first_child          1         0  9906 1233.4
## - transport_means         1         0  9906 1233.4
## <none>                                 9905 1235.4
## - wkly_study_hours        2       183 10088 1237.8
## - parent_educ             3       376 10282 1242.6
## - lunch_type              1       734 10639 1258.7
## - test_prep               1       926 10832 1265.0
## - ethnic_group            4      1304 11209 1271.1
## - gender                  1     13079 22985 1531.3
## - writing_score           1     50162 60068 1871.4
##
## Step:  AIC=1229.77
```

```
## math_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + parent_marital_status + practice_sport + is_first_child +
##     transport_means + wkly_study_hours + writing_score
##
##                         Df Sum of Sq   RSS    AIC
## - parent_marital_status  3        58 10202 1225.8
## - practice_sport         2        13 10156 1226.2
## - is_first_child         1         0 10144 1227.8
## - transport_means        1         0 10144 1227.8
## <none>                               10143 1229.8
## - wkly_study_hours       2       184 10328 1232.2
## + nr_siblings            7       238  9905 1235.4
## - parent_educ            3       470 10614 1239.8
## - lunch_type             1       721 10864 1252.1
## - test_prep              1       883 11027 1257.3
## - ethnic_group           4      1306 11449 1264.6
## - gender                 1     13592 23736 1528.7
## - writing_score          1     51313 61456 1865.5
##
## Step:  AIC=1225.8
## math_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + practice_sport + is_first_child + transport_means +
##     wkly_study_hours + writing_score
##
##                         Df Sum of Sq   RSS    AIC
## - practice_sport         2        12 10214 1222.2
## - transport_means        1         0 10202 1223.8
## - is_first_child         1         1 10202 1223.8
## <none>                               10202 1225.8
## - wkly_study_hours       2       182 10384 1228.1
## + parent_marital_status  3        58 10143 1229.8
## + nr_siblings            7       248  9954 1231.1
## - parent_educ            3       476 10678 1235.9
## - lunch_type             1       701 10902 1247.3
## - test_prep              1       884 11086 1253.2
## - ethnic_group           4      1289 11490 1259.9
## - gender                 1     13897 24098 1528.1
## - writing_score          1     53617 63819 1872.9
##
## Step:  AIC=1222.22
## math_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + is_first_child + transport_means + wkly_study_hours +
##     writing_score
##
##                         Df Sum of Sq   RSS    AIC
## - transport_means        1         0 10214 1220.2
## - is_first_child         1         1 10215 1220.3
## <none>                               10214 1222.2
## - wkly_study_hours       2       189 10403 1224.7
## + practice_sport         2        12 10202 1225.8
## + parent_marital_status  3        58 10156 1226.2
## + nr_siblings            7       246  9967 1227.6
## - parent_educ            3       493 10706 1232.9
## - lunch_type             1       701 10915 1243.7
```

```
## - test_prep              1       909 11123 1250.4
## - ethnic_group           4      1338 11551 1257.8
## - gender                 1     13891 24105 1524.2
## - writing_score          1     53941 64155 1870.7
##
## Step:  AIC=1220.22
## math_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + is_first_child + wkly_study_hours + writing_score
##
##                         Df Sum of Sq   RSS    AIC
## - is_first_child         1         1 10215 1218.3
## <none>                               10214 1220.2
## + transport_means        1         0 10214 1222.2
## - wkly_study_hours       2       189 10403 1222.7
## + practice_sport         2        12 10202 1223.8
## + parent_marital_status  3        57 10157 1224.2
## + nr_siblings            7       246  9968 1225.6
## - parent_educ            3       493 10707 1230.9
## - lunch_type             1       703 10917 1241.8
## - test_prep              1       917 11131 1248.7
## - ethnic_group           4      1342 11556 1255.9
## - gender                 1     13894 24108 1522.2
## - writing_score          1     53949 64163 1868.8
##
## Step:  AIC=1218.27
## math_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + wkly_study_hours + writing_score
##
##                         Df Sum of Sq   RSS    AIC
## <none>                               10215 1218.3
## + is_first_child         1         1 10214 1220.2
## + transport_means        1         0 10215 1220.3
## - wkly_study_hours       2       188 10403 1220.7
## + practice_sport         2        13 10202 1221.8
## + parent_marital_status  3        58 10157 1222.2
## + nr_siblings            7       244  9972 1223.7
## - parent_educ            3       497 10712 1229.1
## - lunch_type             1       702 10917 1239.8
## - test_prep              1       930 11145 1247.1
## - ethnic_group           4      1341 11556 1253.9
## - gender                 1     13907 24122 1520.4
## - writing_score          1     53963 64179 1866.8
```

```
math_writing |> summary()
```

```
##
## Call:
## lm(formula = math_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep + wkly_study_hours + writing_score,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9895  -3.6039   0.0824   3.4587  14.3526
```

```
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         7.52091    1.89952   3.959 9.15e-05 ***
## gender1           -13.70836    0.63717 -21.514  < 2e-16 ***
## ethnic_group1       0.13982    1.31690   0.106 0.915510
## ethnic_group2      -1.62573    1.24646  -1.304 0.193019
## ethnic_group3      -2.24157    1.27532  -1.758 0.079706 .
## ethnic_group4       3.48045    1.35134   2.576 0.010430 *
## parent_educ2       -0.11677    0.71185  -0.164 0.869805
## parent_educ3       -1.43687    0.84147  -1.708 0.088627 .
## parent_educ4       -3.83877    1.03267  -3.717 0.000235 ***
## lunch_type1        -3.14880    0.65149  -4.833 2.04e-06 ***
## test_prep1         -3.66767    0.65934  -5.563 5.39e-08 ***
## wkly_study_hours1   0.98708    0.70387   1.402 0.161723
## wkly_study_hours2   2.25503    0.90230   2.499 0.012918 *
## writing_score       0.99512    0.02348  42.380  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.481 on 340 degrees of freedom
## Multiple R-squared:  0.8853, Adjusted R-squared:  0.8809
## F-statistic: 201.9 on 13 and 340 DF,  p-value: < 2.2e-16
```

```
# Reading
reading_math = lm(reading_score ~ . - writing_score, data = data) |>
  step(direction = "both")
```

```
## Start:  AIC=1271.88
## 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 + math_score +
##     writing_score) - writing_score
##
##                         Df Sum of Sq   RSS    AIC
## - nr_siblings            7       303 11285 1267.5
## - parent_marital_status  3        75 11057 1268.3
## - transport_means        1        10 10992 1270.2
## - practice_sport         2        75 11057 1270.3
## - is_first_child         1        57 11039 1271.7
## <none>                               10982 1271.9
## - parent_educ            3       308 11290 1275.7
## - ethnic_group           4       389 11371 1276.2
## - wkly_study_hours       2       279 11261 1276.8
## - lunch_type             1       386 11368 1282.1
## - test_prep              1       510 11492 1286.0
## - gender                 1     10416 21398 1506.0
## - math_score             1     43473 54454 1836.7
##
## Step:  AIC=1267.53
## reading_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + parent_marital_status + practice_sport + is_first_child +
##     transport_means + wkly_study_hours + math_score
##
```

```
##                          Df Sum of Sq   RSS    AIC
## - parent_marital_status  3        75 11360 1263.9
## - practice_sport         2        69 11355 1265.7
## - transport_means        1         9 11295 1265.8
## - is_first_child         1        51 11336 1267.1
## <none>                               11285 1267.5
## - ethnic_group           4       385 11670 1271.4
## + nr_siblings            7       303 10982 1271.9
## - wkly_study_hours       2       313 11598 1273.2
## - parent_educ            3       386 11672 1273.5
## - lunch_type             1       351 11637 1276.4
## - test_prep              1       466 11752 1279.9
## - gender                 1     10743 22028 1502.3
## - math_score             1     44057 55342 1828.4
##
## Step:  AIC=1263.87
## reading_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + practice_sport + is_first_child + transport_means +
##     wkly_study_hours + math_score
##
##                          Df Sum of Sq   RSS    AIC
## - practice_sport         2        66 11426 1261.9
## - transport_means        1        11 11371 1262.2
## <none>                               11360 1263.9
## - is_first_child         1        66 11426 1263.9
## + parent_marital_status  3        75 11285 1267.5
## - ethnic_group           4       396 11756 1268.0
## + nr_siblings            7       303 11057 1268.3
## - wkly_study_hours       2       294 11654 1268.9
## - parent_educ            3       376 11736 1269.4
## - lunch_type             1       349 11709 1272.6
## - test_prep              1       452 11813 1275.7
## - gender                 1     10678 22038 1496.5
## - math_score             1     45906 57267 1834.5
##
## Step:  AIC=1261.92
## reading_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + is_first_child + transport_means + wkly_study_hours +
##     math_score
##
##                          Df Sum of Sq   RSS    AIC
## - transport_means        1        15 11441 1260.4
## <none>                               11426 1261.9
## - is_first_child         1        89 11514 1262.7
## + practice_sport         2        66 11360 1263.9
## + parent_marital_status  3        71 11355 1265.7
## - ethnic_group           4       405 11831 1266.2
## + nr_siblings            7       299 11127 1266.5
## - wkly_study_hours       2       300 11726 1267.1
## - parent_educ            3       419 11845 1268.7
## - lunch_type             1       334 11759 1270.1
## - test_prep              1       490 11916 1274.8
## - gender                 1     10627 22053 1492.7
## - math_score             1     45965 57391 1831.3
```

```
##
## Step:  AIC=1260.37
## reading_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + is_first_child + wkly_study_hours + math_score
##
##                         Df Sum of Sq   RSS    AIC
## <none>                               11441 1260.4
## - is_first_child         1        93 11533 1261.2
## + transport_means        1        15 11426 1261.9
## + practice_sport         2        70 11371 1262.2
## + parent_marital_status  3        72 11368 1264.1
## - ethnic_group           4       401 11841 1264.6
## + nr_siblings            7       297 11144 1265.1
## - wkly_study_hours       2       295 11735 1265.4
## - parent_educ            3       430 11870 1267.4
## - lunch_type             1       326 11766 1268.3
## - test_prep              1       477 11917 1272.8
## - gender                 1     10632 22073 1491.0
## - math_score             1     45952 57393 1829.3
```

```
reading_math |> summary()
```

```
##
## Call:
## lm(formula = reading_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep + is_first_child + wkly_study_hours +
##     math_score, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.9739  -3.9387   0.0553   3.9475  17.0165
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.044602   2.097935   2.405 0.016728 *
## gender1           11.324635   0.638017  17.750  < 2e-16 ***
## ethnic_group1     -0.151881   1.396437  -0.109 0.913455
## ethnic_group2      0.219983   1.320008   0.167 0.867743
## ethnic_group3     -0.068102   1.345327  -0.051 0.959657
## ethnic_group4     -2.916552   1.442569  -2.022 0.043985 *
## parent_educ2       1.009829   0.755142   1.337 0.182031
## parent_educ3       2.105538   0.887662   2.372 0.018250 *
## parent_educ4       3.521963   1.084545   3.247 0.001281 **
## lunch_type1        2.206613   0.710249   3.107 0.002051 **
## test_prep1         2.528751   0.672906   3.758 0.000202 ***
## is_first_child1    1.093717   0.660559   1.656 0.098699 .
## wkly_study_hours1 -0.008659   0.748514  -0.012 0.990777
## wkly_study_hours2 -2.402879   0.960663  -2.501 0.012846 *
## math_score         0.846270   0.022934  36.900  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.809 on 339 degrees of freedom
## Multiple R-squared:  0.8523, Adjusted R-squared:  0.8462
```

```
## F-statistic: 139.7 on 14 and 339 DF,  p-value: < 2.2e-16
```

```r
reading_writing = lm(reading_score ~ . - math_score, data = data) |>
  step(direction = "both")
```

```
## Start:  AIC=1017.78
## 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 + math_score +
##     writing_score) - math_score
##
##                          Df Sum of Sq   RSS    AIC
## - nr_siblings             7        34  5391 1006.0
## - parent_marital_status   3         4  5361 1012.0
## - wkly_study_hours        2        37  5394 1016.2
## - transport_means         1        14  5372 1016.7
## - parent_educ             3        77  5434 1016.8
## <none>                                 5357 1017.8
## - practice_sport          2        82  5439 1019.1
## - is_first_child          1        54  5411 1019.3
## - lunch_type              1        55  5412 1019.4
## - ethnic_group            4       265  5622 1026.8
## - gender                  1       179  5536 1027.4
## - test_prep               1       225  5582 1030.3
## - writing_score           1     49097 54454 1836.7
##
## Step:  AIC=1005.99
## reading_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + parent_marital_status + practice_sport + is_first_child +
##     transport_means + wkly_study_hours + writing_score
##
##                          Df Sum of Sq   RSS    AIC
## - parent_marital_status   3         5  5396 1000.3
## - parent_educ             3        67  5458 1004.4
## - wkly_study_hours        2        42  5433 1004.8
## - transport_means         1        14  5405 1004.9
## <none>                                 5391 1006.0
## - practice_sport          2        85  5475 1007.5
## - is_first_child          1        54  5445 1007.5
## - lunch_type              1        56  5446 1007.6
## - ethnic_group            4       272  5663 1015.4
## - gender                  1       182  5573 1015.7
## + nr_siblings             7        34  5357 1017.8
## - test_prep               1       233  5624 1019.0
## - writing_score           1     49952 55342 1828.4
##
## Step:  AIC=1000.34
## reading_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + practice_sport + is_first_child + transport_means +
##     wkly_study_hours + writing_score
##
##                          Df Sum of Sq   RSS     AIC
## - parent_educ             3        67  5464  998.75
## - wkly_study_hours        2        41  5437  999.03
```

```
## - transport_means        1        14  5410  999.23
## <none>                                5396 1000.34
## - practice_sport         2        85  5481 1001.88
## - lunch_type             1        54  5451 1001.90
## - is_first_child         1        59  5456 1002.22
## + parent_marital_status  3         5  5391 1005.99
## - ethnic_group           4       275  5671 1009.96
## - gender                 1       187  5583 1010.42
## + nr_siblings            7        35  5361 1012.03
## - test_prep              1       236  5633 1013.53
## - writing_score          1     51870 57267 1834.51
##
## Step:  AIC=998.75
## reading_score ~ gender + ethnic_group + lunch_type + test_prep +
##     practice_sport + is_first_child + transport_means + wkly_study_hours +
##     writing_score
##
##                         Df Sum of Sq   RSS     AIC
## - wkly_study_hours       2        34  5497  996.93
## - transport_means        1        19  5482  997.95
## <none>                                5464  998.75
## - practice_sport         2        72  5536  999.39
## - lunch_type             1        49  5512  999.89
## - is_first_child         1        51  5515 1000.06
## + parent_educ            3        67  5396 1000.34
## + parent_marital_status  3         5  5458 1004.40
## - gender                 1       192  5655 1008.95
## - ethnic_group           4       301  5765 1009.74
## - test_prep              1       214  5677 1010.34
## + nr_siblings            7        25  5439 1011.12
## - writing_score          1     53435 58899 1838.46
##
## Step:  AIC=996.93
## reading_score ~ gender + ethnic_group + lunch_type + test_prep +
##     practice_sport + is_first_child + transport_means + writing_score
##
##                         Df Sum of Sq   RSS     AIC
## - transport_means        1        16  5514  995.97
## <none>                                5497  996.93
## - practice_sport         2        73  5571  997.61
## - lunch_type             1        48  5546  998.02
## - is_first_child         1        51  5548  998.19
## + wkly_study_hours       2        34  5464  998.75
## + parent_educ            3        60  5437  999.03
## + parent_marital_status  3         4  5494 1002.68
## - gender                 1       185  5682 1006.64
## - ethnic_group           4       298  5795 1007.59
## + nr_siblings            7        31  5467 1008.95
## - test_prep              1       240  5737 1010.04
## - writing_score          1     55486 60983 1846.77
##
## Step:  AIC=995.97
## reading_score ~ gender + ethnic_group + lunch_type + test_prep +
##     practice_sport + is_first_child + writing_score
```

```
##
##                          Df Sum of Sq   RSS     AIC
## <none>                                  5514  995.97
## - practice_sport          2        74  5588  996.69
## - lunch_type              1        45  5559  996.85
## + transport_means         1        16  5497  996.93
## - is_first_child          1        53  5566  997.33
## + parent_educ             3        65  5448  997.76
## + wkly_study_hours        2        31  5482  997.95
## + parent_marital_status   3         4  5510 1001.74
## - gender                  1       184  5698 1005.62
## - ethnic_group            4       323  5836 1008.09
## + nr_siblings             7        29  5485 1008.13
## - test_prep               1       256  5770 1010.06
## - writing_score           1     55481 60995 1844.83
```

```
reading_writing |> summary()
```

```
##
## Call:
## lm(formula = reading_score ~ gender + ethnic_group + lunch_type +
##     test_prep + practice_sport + is_first_child + writing_score,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.4698  -2.9709   0.1727   2.8592   9.4855
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.05123    1.46329   3.452 0.000626 ***
## gender1         -1.57013    0.46429  -3.382 0.000803 ***
## ethnic_group1   -0.19079    0.96363  -0.198 0.843166
## ethnic_group2   -1.25887    0.91422  -1.377 0.169419
## ethnic_group3   -2.74423    0.93633  -2.931 0.003608 **
## ethnic_group4   -0.69216    0.98825  -0.700 0.484163
## lunch_type1      0.79483    0.47574   1.671 0.095691 .
## test_prep1      -1.91374    0.47994  -3.987 8.17e-05 ***
## practice_sport1 -1.20783    0.68759  -1.757 0.079881 .
## practice_sport2 -1.53333    0.71810  -2.135 0.033451 *
## is_first_child1  0.83168    0.46063   1.806 0.071869 .
## writing_score    0.97757    0.01666  58.663  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.015 on 342 degrees of freedom
## Multiple R-squared:  0.9288, Adjusted R-squared:  0.9265
## F-statistic: 405.7 on 11 and 342 DF,  p-value: < 2.2e-16
```

```
# Writing
writing_math = lm(writing_score ~ . - reading_score, data = data) |>
  step(direction = "both")
```

```
## Start:  AIC=1177.85
```

```
## 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 + math_score +
##     reading_score) - reading_score
##
##                         Df Sum of Sq   RSS    AIC
## - nr_siblings            7       176  8596 1171.2
## - parent_marital_status  3        62  8482 1174.4
## - practice_sport         2        23  8443 1174.8
## - is_first_child         1         0  8420 1175.8
## - transport_means        1         1  8421 1175.9
## <none>                               8420 1177.8
## - wkly_study_hours       2       117  8537 1178.8
## - lunch_type             1       106  8526 1180.3
## - parent_educ            3       552  8972 1194.3
## - ethnic_group           4       855  9275 1204.1
## - test_prep              1      1618 10038 1238.1
## - gender                 1     13880 22300 1520.6
## - math_score             1     42640 51060 1813.9
##
## Step:  AIC=1171.17
## writing_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + parent_marital_status + practice_sport + is_first_child +
##     transport_means + wkly_study_hours + math_score
##
##                         Df Sum of Sq   RSS    AIC
## - parent_marital_status  3        53  8649 1167.4
## - practice_sport         2        25  8621 1168.2
## - is_first_child         1         0  8596 1169.2
## - transport_means        1         1  8597 1169.2
## <none>                               8596 1171.2
## - wkly_study_hours       2       130  8726 1172.5
## - lunch_type             1        88  8684 1172.8
## + nr_siblings            7       176  8420 1177.8
## - parent_educ            3       657  9253 1191.2
## - ethnic_group           4       826  9422 1195.7
## - test_prep              1      1558 10154 1228.1
## - gender                 1     14336 22932 1516.5
## - math_score             1     43483 52079 1806.9
##
## Step:  AIC=1167.36
## writing_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + practice_sport + is_first_child + transport_means +
##     wkly_study_hours + math_score
##
##                         Df Sum of Sq   RSS    AIC
## - practice_sport         2        24  8673 1164.3
## - is_first_child         1         0  8649 1165.4
## - transport_means        1         0  8649 1165.4
## <none>                               8649 1167.4
## - wkly_study_hours       2       120  8769 1168.2
## - lunch_type             1        90  8739 1169.0
## + parent_marital_status  3        53  8596 1171.2
## + nr_siblings            7       168  8482 1174.4
```

```
## - parent_educ            3       646  9295 1186.8
## - ethnic_group           4       842  9491 1192.2
## - test_prep              1      1536 10186 1223.2
## - gender                 1     14285 22934 1510.6
## - math_score             1     45458 54107 1814.4
##
## Step:  AIC=1164.35
## writing_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + is_first_child + transport_means + wkly_study_hours +
##     math_score
##
##                         Df Sum of Sq   RSS    AIC
## - transport_means        1         0  8673 1162.3
## - is_first_child         1         0  8674 1162.4
## <none>                               8673 1164.3
## - wkly_study_hours       2       125  8798 1165.4
## - lunch_type             1        92  8766 1166.1
## + practice_sport         2        24  8649 1167.4
## + parent_marital_status  3        52  8621 1168.2
## + nr_siblings            7       169  8504 1171.4
## - parent_educ            3       650  9323 1183.9
## - ethnic_group           4       896  9570 1191.2
## - test_prep              1      1568 10242 1221.2
## - gender                 1     14301 22974 1507.2
## - math_score             1     45807 54480 1812.8
##
## Step:  AIC=1162.35
## writing_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + is_first_child + wkly_study_hours + math_score
##
##                         Df Sum of Sq   RSS    AIC
## - is_first_child         1         0  8674 1160.4
## <none>                               8673 1162.3
## - wkly_study_hours       2       125  8799 1163.4
## - lunch_type             1        93  8766 1164.1
## + transport_means        1         0  8673 1164.3
## + practice_sport         2        24  8649 1165.4
## + parent_marital_status  3        52  8621 1166.2
## + nr_siblings            7       169  8504 1169.4
## - parent_educ            3       650  9323 1181.9
## - ethnic_group           4       903  9577 1189.4
## - test_prep              1      1591 10264 1220.0
## - gender                 1     14301 22975 1505.2
## - math_score             1     45813 54486 1810.9
##
## Step:  AIC=1160.36
## writing_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + wkly_study_hours + math_score
##
##                         Df Sum of Sq   RSS    AIC
## <none>                               8674 1160.4
## - wkly_study_hours       2       125  8799 1161.4
## - lunch_type             1        93  8766 1162.1
## + is_first_child         1         0  8673 1162.3
```

```
## + transport_means        1         0  8674 1162.4
## + practice_sport         2        24  8650 1163.4
## + parent_marital_status  3        53  8621 1164.2
## + nr_siblings            7       169  8505 1167.4
## - parent_educ            3       652  9326 1180.0
## - ethnic_group           4       904  9577 1187.4
## - test_prep              1      1606 10280 1218.5
## - gender                 1     14332 23005 1503.7
## - math_score             1     45820 54494 1808.9
```

```
writing_math |> summary()
```

```
##
## Call:
## lm(formula = writing_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep + wkly_study_hours + math_score, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7862  -3.3414  -0.0339   3.3581  12.6292
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.63866    1.78450   1.479   0.1402
## gender1           13.13078    0.55400  23.702  < 2e-16 ***
## ethnic_group1      0.30538    1.21338   0.252   0.8014
## ethnic_group2      1.78262    1.14737   1.554   0.1212
## ethnic_group3      2.94237    1.16965   2.516   0.0123 *
## ethnic_group4     -1.86655    1.25322  -1.489   0.1373
## parent_educ2       0.34846    0.65570   0.531   0.5955
## parent_educ3       1.99752    0.77113   2.590   0.0100 **
## parent_educ4       4.31411    0.94209   4.579 6.56e-06 ***
## lunch_type1        1.17610    0.61731   1.905   0.0576 .
## test_prep1         4.62520    0.58292   7.935 3.09e-14 ***
## wkly_study_hours1  0.04936    0.65046   0.076   0.9396
## wkly_study_hours2 -1.52614    0.83494  -1.828   0.0684 .
## math_score         0.84495    0.01994  42.380  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.051 on 340 degrees of freedom
## Multiple R-squared:  0.894,  Adjusted R-squared:   0.89
## F-statistic: 220.7 on 13 and 340 DF,  p-value: < 2.2e-16
```

```
writing_reading = lm(writing_score ~ . - math_score, data = data) |>
  step(direction = "both")
```

```
## Start:  AIC=994.99
## 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 + math_score +
##     reading_score) - math_score
##
```

```
##                          Df Sum of Sq   RSS     AIC
## - nr_siblings             7        50  5073  984.48
## - parent_marital_status   3        23  5046  990.60
## - wkly_study_hours        2        36  5060  993.55
## - transport_means         1        19  5042  994.32
## <none>                                 5023  994.99
## - is_first_child          1        48  5071  996.38
## - parent_educ             3       119  5142  997.27
## - practice_sport          2       101  5124  998.06
## - lunch_type              1       205  5228 1007.14
## - ethnic_group            4       312  5335 1008.30
## - gender                  1       447  5470 1023.16
## - test_prep               1       485  5508 1025.64
## - reading_score           1     46037 51060 1813.90
##
## Step:  AIC=984.48
## writing_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + parent_marital_status + practice_sport + is_first_child +
##     transport_means + wkly_study_hours + reading_score
##
##                          Df Sum of Sq   RSS     AIC
## - parent_marital_status   3        26  5098  980.25
## - wkly_study_hours        2        38  5111  983.10
## - transport_means         1        17  5090  983.67
## <none>                                 5073  984.48
## - is_first_child          1        48  5121  985.79
## - parent_educ             3       110  5183  986.08
## - practice_sport          2       106  5179  987.81
## + nr_siblings             7        50  5023  994.99
## - lunch_type              1       217  5290  997.31
## - ethnic_group            4       319  5392  998.08
## - gender                  1       453  5526 1012.75
## - test_prep               1       494  5567 1015.37
## - reading_score           1     47006 52079 1806.89
##
## Step:  AIC=980.25
## writing_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + practice_sport + is_first_child + transport_means +
##     wkly_study_hours + reading_score
##
##                          Df Sum of Sq   RSS     AIC
## - wkly_study_hours        2        40  5139  979.03
## - transport_means         1        14  5112  979.20
## <none>                                 5098  980.25
## - parent_educ             3       108  5206  981.67
## - is_first_child          1        53  5151  981.91
## - practice_sport          2       105  5203  983.46
## + parent_marital_status   3        26  5073  984.48
## + nr_siblings             7        52  5046  990.60
## - lunch_type              1       209  5308  992.49
## - ethnic_group            4       322  5421  993.95
## - gender                  1       442  5540 1007.67
## - test_prep               1       491  5590 1010.82
## - reading_score           1     49008 54107 1814.41
```

```
##
## Step:  AIC=979.03
## writing_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + practice_sport + is_first_child + transport_means +
##     reading_score
##
##                          Df Sum of Sq   RSS     AIC
## - transport_means         1        11  5150  977.79
## <none>                                 5139  979.03
## - parent_educ             3        95  5234  979.51
## + wkly_study_hours        2        40  5098  980.25
## - is_first_child          1        50  5189  980.49
## - practice_sport          2       103  5241  982.04
## + parent_marital_status   3        28  5111  983.10
## + nr_siblings             7        53  5085  989.34
## - lunch_type              1       201  5340  990.62
## - ethnic_group            4       320  5459  992.44
## - gender                  1       421  5559 1004.89
## - test_prep               1       518  5656 1011.02
## - reading_score           1     51051 56189 1823.78
##
## Step:  AIC=977.79
## writing_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + practice_sport + is_first_child + reading_score
##
##                          Df Sum of Sq   RSS     AIC
## <none>                                 5150  977.79
## - parent_educ             3        98  5248  978.49
## + transport_means         1        11  5139  979.03
## + wkly_study_hours        2        37  5112  979.20
## - is_first_child          1        52  5202  979.37
## - practice_sport          2       103  5253  980.81
## + parent_marital_status   3        25  5125  982.10
## + nr_siblings             7        52  5098  988.21
## - lunch_type              1       196  5346  989.03
## - ethnic_group            4       340  5490  992.42
## - gender                  1       420  5569 1003.54
## - test_prep               1       546  5695 1011.45
## - reading_score           1     51045 56195 1821.82
```

```
writing_reading |> summary()
```

```
##
## Call:
## lm(formula = writing_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep + practice_sport + is_first_child +
##     reading_score, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.960 -2.728 -0.169  2.615  9.475
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)      0.92461    1.45587   0.635 0.525799
## gender1          2.31862    0.44099   5.258 2.59e-07 ***
## ethnic_group1    0.44110    0.93737   0.471 0.638253
## ethnic_group2    1.36322    0.88761   1.536 0.125512
## ethnic_group3    3.04721    0.90559   3.365 0.000853 ***
## ethnic_group4    1.30540    0.95925   1.361 0.174464
## parent_educ2    -0.46713    0.50706  -0.921 0.357577
## parent_educ3     0.25030    0.59714   0.419 0.675365
## parent_educ4     1.34702    0.73597   1.830 0.068089 .
## lunch_type1     -1.64123    0.45661  -3.594 0.000373 ***
## test_prep1       2.73895    0.45695   5.994 5.24e-09 ***
## practice_sport1  1.47778    0.66828   2.211 0.027681 *
## practice_sport2  1.81448    0.70192   2.585 0.010154 *
## is_first_child1 -0.83255    0.44865  -1.856 0.064366 .
## reading_score    0.92657    0.01598  57.968  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.897 on 339 degrees of freedom
## Multiple R-squared:  0.9371, Adjusted R-squared:  0.9345
## F-statistic: 360.7 on 14 and 339 DF,  p-value: < 2.2e-16
```

```
# check VIF
vif_math_reading =
  performance::check_collinearity(math_reading) |>
  as_tibble() |>
  mutate(VIF_CI = str_c("[", round(VIF_CI_low, 1), ", ", round(VIF_CI_high, 1), "]")) |>
  dplyr::select(Term, VIF, VIF_CI, Tolerance)
knitr::kable(x = vif_math_reading, caption = "VIF for Math Score", digits = 1)
```

Table 9: VIF for Math Score

Term	VIF	VIF_CI	Tolerance
gender	1.1	[1, 1.3]	0.9
ethnic_group	1.1	[1, 1.3]	0.9
lunch_type	1.1	[1, 1.3]	0.9
test_prep	1.1	[1, 1.3]	0.9
practice_sport	1.0	[1, 1.5]	1.0
wkly_study_hours	1.1	[1, 1.3]	0.9
reading_score	1.3	[1.2, 1.5]	0.8

```
vif_math_writing =
  performance::check_collinearity(math_writing) |>
  as_tibble() |>
  mutate(VIF_CI = str_c("[", round(VIF_CI_low, 1), ", ", round(VIF_CI_high, 1), "]")) |>
  dplyr::select(Term, VIF, VIF_CI, Tolerance)
knitr::kable(x = vif_math_writing, caption = "VIF for Math Score", digits = 1)
```

Table 10: VIF for Math Score

Term	VIF	VIF_CI	Tolerance
gender	1.2	[1.1, 1.4]	0.8
ethnic_group	1.1	[1, 1.3]	0.9
parent_educ	1.1	[1, 1.3]	0.9
lunch_type	1.1	[1.1, 1.4]	0.9
test_prep	1.2	[1.1, 1.4]	0.8
wkly_study_hours	1.1	[1, 1.3]	0.9
writing_score	1.5	[1.3, 1.7]	0.7

```
vif_reading_math =
  performance::check_collinearity(reading_math) |>
  as_tibble() |>
  mutate(VIF_CI = str_c("[", round(VIF_CI_low, 1), ", ", round(VIF_CI_high, 1), "]")) |>
  dplyr::select(Term, VIF, VIF_CI, Tolerance)
knitr::kable(x = vif_reading_math, caption = "VIF for Reading Score", digits = 1)
```

Table 11: VIF for Reading Score

Term	VIF	VIF_CI	Tolerance
gender	1.1	[1, 1.4]	1.0
ethnic_group	1.1	[1.1, 1.4]	0.9
parent_educ	1.1	[1, 1.3]	0.9
lunch_type	1.2	[1.1, 1.4]	0.8
test_prep	1.1	[1, 1.3]	0.9
is_first_child	1.0	[1, 3.7]	1.0
wkly_study_hours	1.1	[1, 1.3]	0.9
math_score	1.4	[1.2, 1.6]	0.7

```
vif_reading_writing =
  performance::check_collinearity(reading_writing) |>
  as_tibble() |>
  mutate(VIF_CI = str_c("[", round(VIF_CI_low, 1), ", ", round(VIF_CI_high, 1), "]")) |>
  dplyr::select(Term, VIF, VIF_CI, Tolerance)
knitr::kable(x = vif_reading_writing, caption = "VIF for Reading Score", digits = 1)
```

Table 12: VIF for Reading Score

Term	VIF	VIF_CI	Tolerance
gender	1.2	[1.1, 1.4]	0.9
ethnic_group	1.1	[1, 1.3]	0.9
lunch_type	1.1	[1.1, 1.3]	0.9
test_prep	1.2	[1.1, 1.4]	0.9
practice_sport	1.1	[1, 1.4]	0.9
is_first_child	1.0	[1, 1.6]	1.0
writing_score	1.4	[1.3, 1.6]	0.7

```
vif_writing_math =
  performance::check_collinearity(writing_math) |>
  as_tibble() |>
  mutate(VIF_CI = str_c("[", round(VIF_CI_low, 1), ", ", round(VIF_CI_high, 1), "]")) |>
  dplyr::select(Term, VIF, VIF_CI, Tolerance)
knitr::kable(x = vif_writing_math, caption = "VIF for Reading Score", digits = 1)
```

Table 13: VIF for Reading Score

Term	VIF	VIF_CI	Tolerance
gender	1.0	[1, 1.5]	1.0
ethnic_group	1.1	[1.1, 1.3]	0.9
parent_educ	1.1	[1, 1.4]	0.9
lunch_type	1.2	[1.1, 1.4]	0.8
test_prep	1.1	[1, 1.3]	0.9
wkly_study_hours	1.1	[1, 1.3]	0.9
math_score	1.4	[1.2, 1.6]	0.7

```
vif_writing_reading =
  performance::check_collinearity(writing_reading) |>
  as_tibble() |>
  mutate(VIF_CI = str_c("[", round(VIF_CI_low, 1), ", ", round(VIF_CI_high, 1), "]")) |>
  dplyr::select(Term, VIF, VIF_CI, Tolerance)
knitr::kable(x = vif_writing_reading, caption = "VIF for Reading Score", digits = 1)
```

Table 14: VIF for Reading Score

Term	VIF	VIF_CI	Tolerance
gender	1.1	[1, 1.3]	0.9
ethnic_group	1.1	[1, 1.3]	0.9
parent_educ	1.1	[1, 1.3]	0.9
lunch_type	1.1	[1, 1.3]	0.9
test_prep	1.1	[1, 1.3]	0.9
practice_sport	1.1	[1, 1.3]	0.9
is_first_child	1.0	[1, 1.5]	1.0
reading_score	1.3	[1.2, 1.5]	0.8