

## Code

### 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\ncolumn_means"

```

```

# 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

```

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

```
data =
  data |>
  drop_na()

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.p75, numeric.max)

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

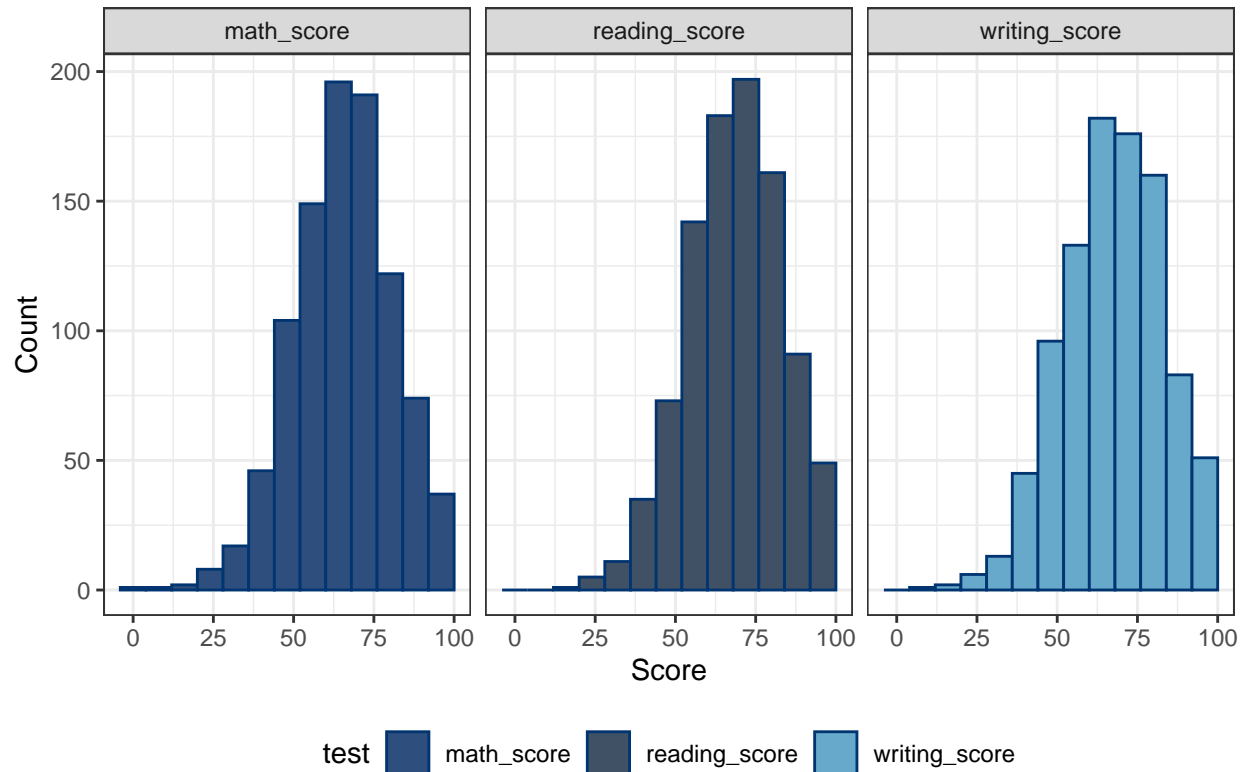
```
data_long |>
  ggplot(aes(x = score, fill = test)) +
  geom_histogram(binwidth = 8, color = "#013571") +
  labs(
    title = "Scores Distribution by Subjects",
```

```

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))

```

## Scores Distribution by Subjects



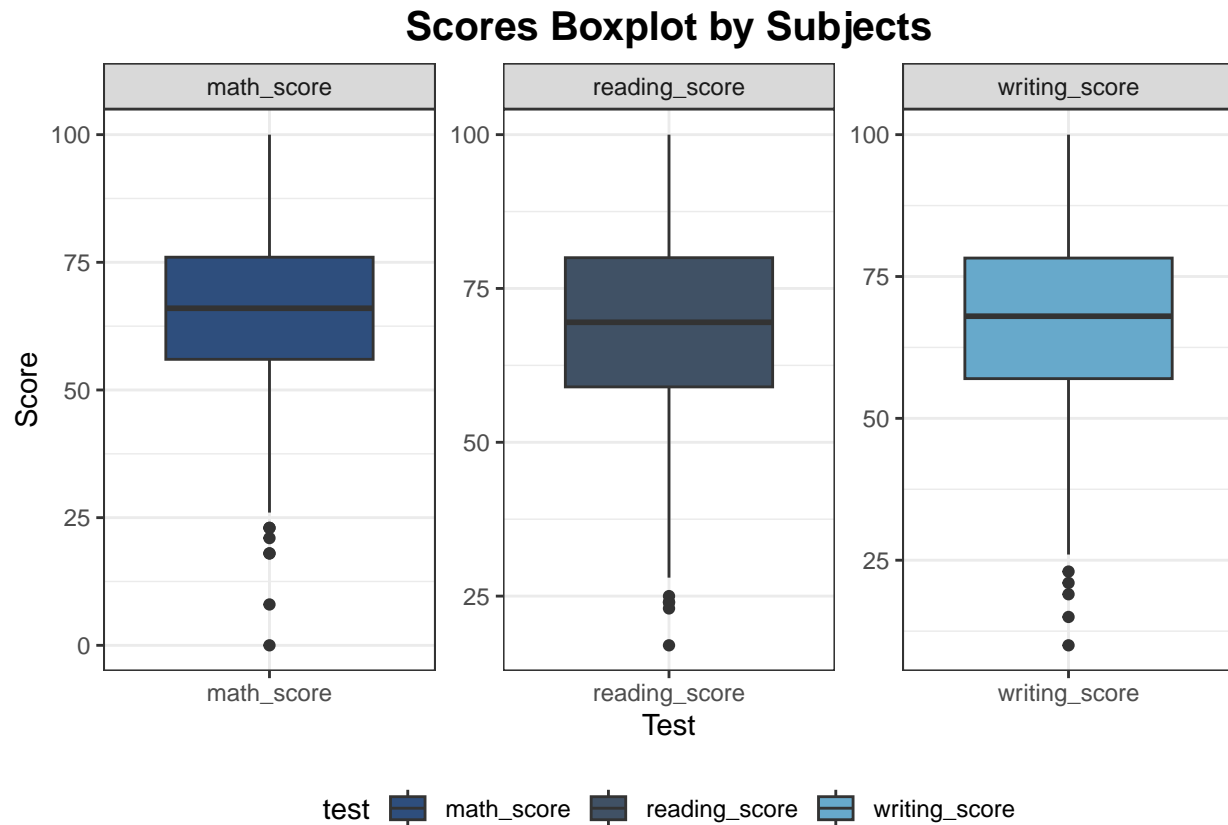
## Boxplots

```

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

```
##
## Call:
## glm(formula = math_score ~ . - reading_score - writing_score,
##      data = data)
##
## 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 **
```

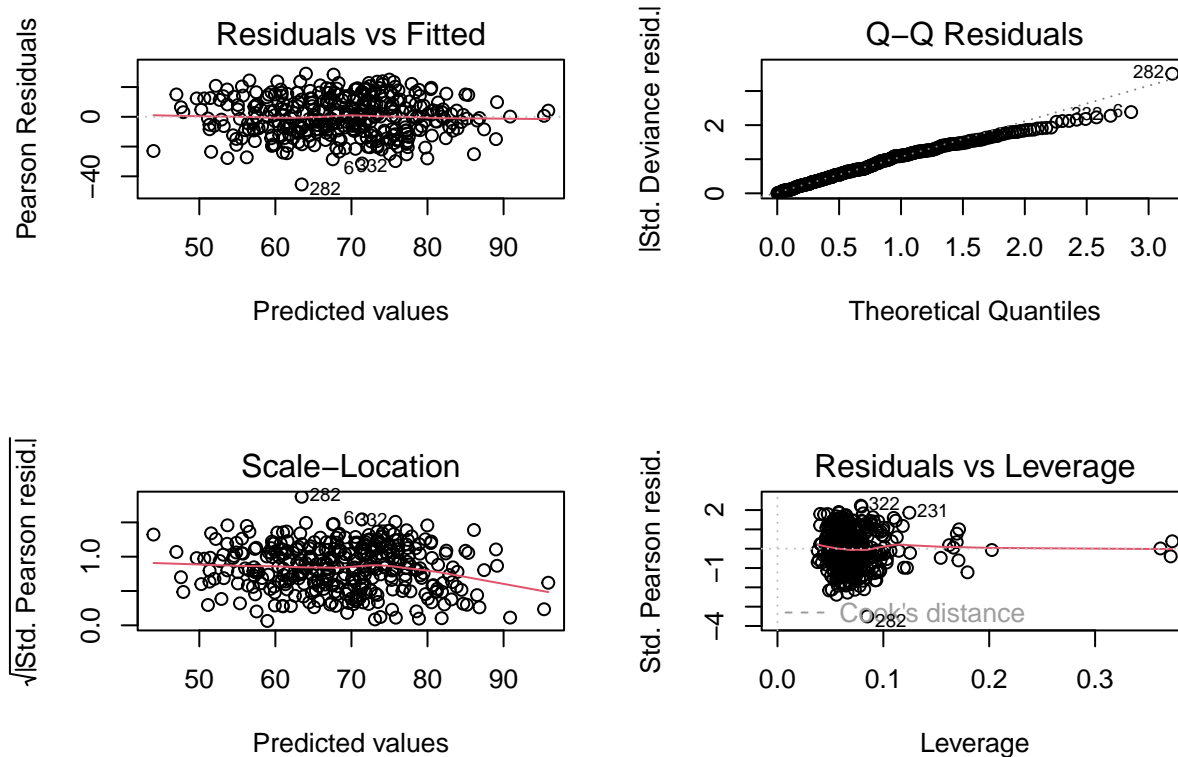
```

## 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
##
## (Dispersion parameter for gaussian family taken to be 183.6931)
##
##    Null deviance: 89074  on 353  degrees of freedom
## Residual deviance: 60068  on 327  degrees of freedom
## AIC: 2878
##
## Number of Fisher Scoring iterations: 2

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

## Warning:           :
##    186

```



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

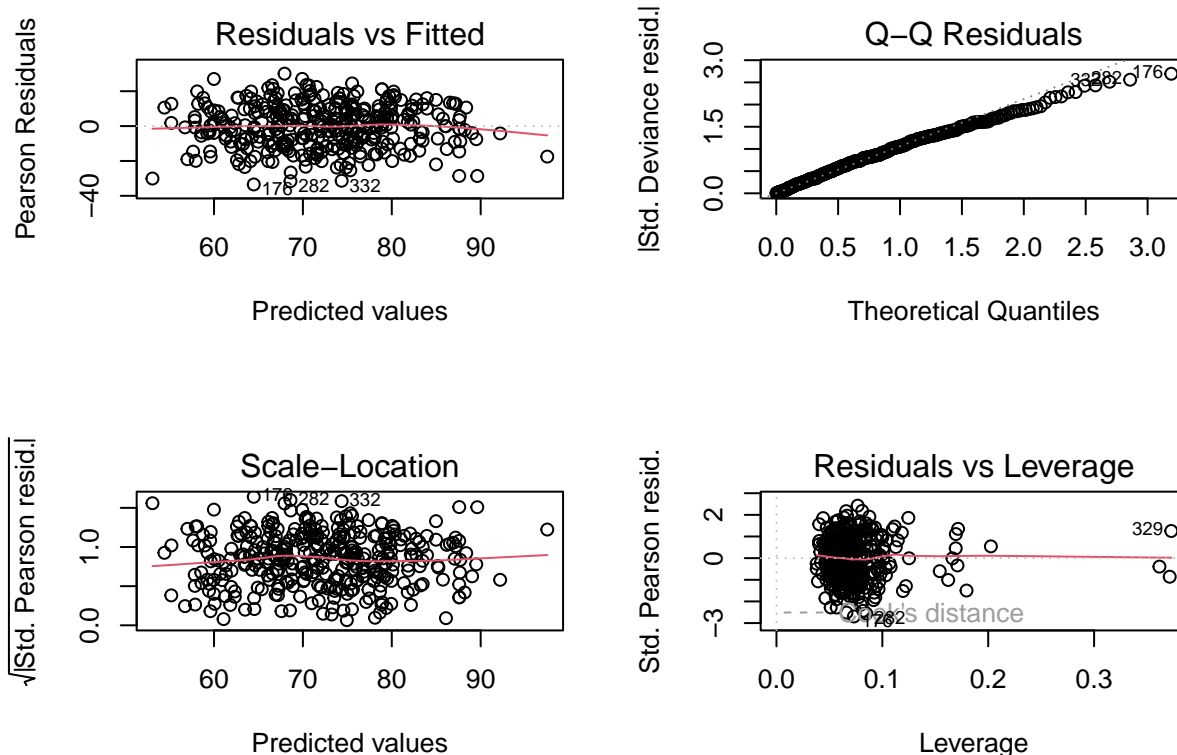
```
##
## Call:
## glm(formula = reading_score ~ . - math_score - writing_score,
##      data = data)
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    59.3627     4.7169  12.585  < 2e-16 ***
## gender1         8.2587     1.4242   5.799 1.57e-08 ***
## ethnic_group1   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
##
## (Dispersion parameter for gaussian family taken to be 166.5274)
##
## Null deviance: 77467  on 353  degrees of freedom
## Residual deviance: 54454  on 327  degrees of freedom
## AIC: 2843.3
##
## Number of Fisher Scoring iterations: 2
```

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

```
## Warning:           :
## 186
```





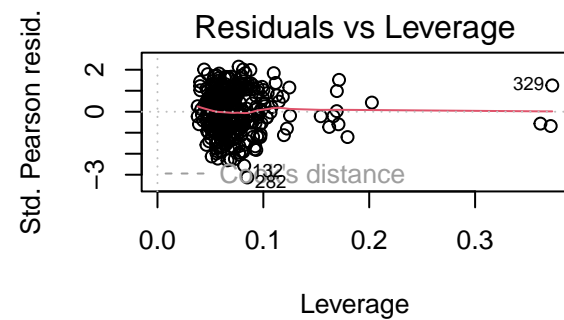
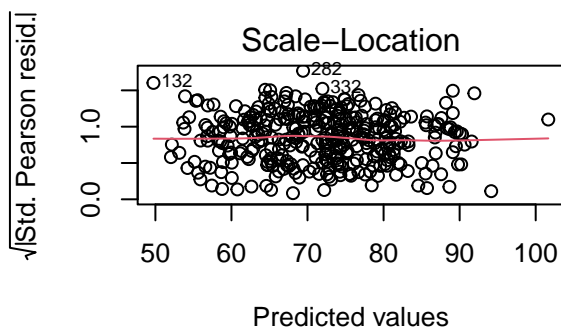
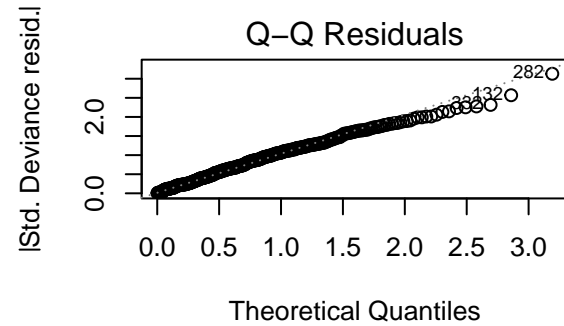
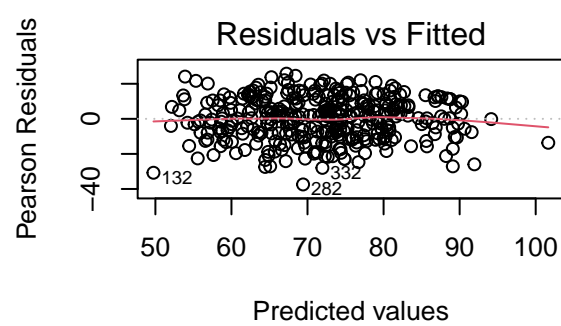
```
# Writing
model_writing_full = glm(writing_score ~ . - reading_score - math_score, data = data)
model_writing_full |> summary()
```

```
##
## Call:
## glm(formula = writing_score ~ . - reading_score - math_score,
##      data = data)
##
## 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_pre1        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
##
## (Dispersion parameter for gaussian family taken to be 156.147)
##
## Null deviance: 81858  on 353  degrees of freedom
## Residual deviance: 51060  on 327  degrees of freedom
## AIC: 2820.5
##
## Number of Fisher Scoring iterations: 2
```

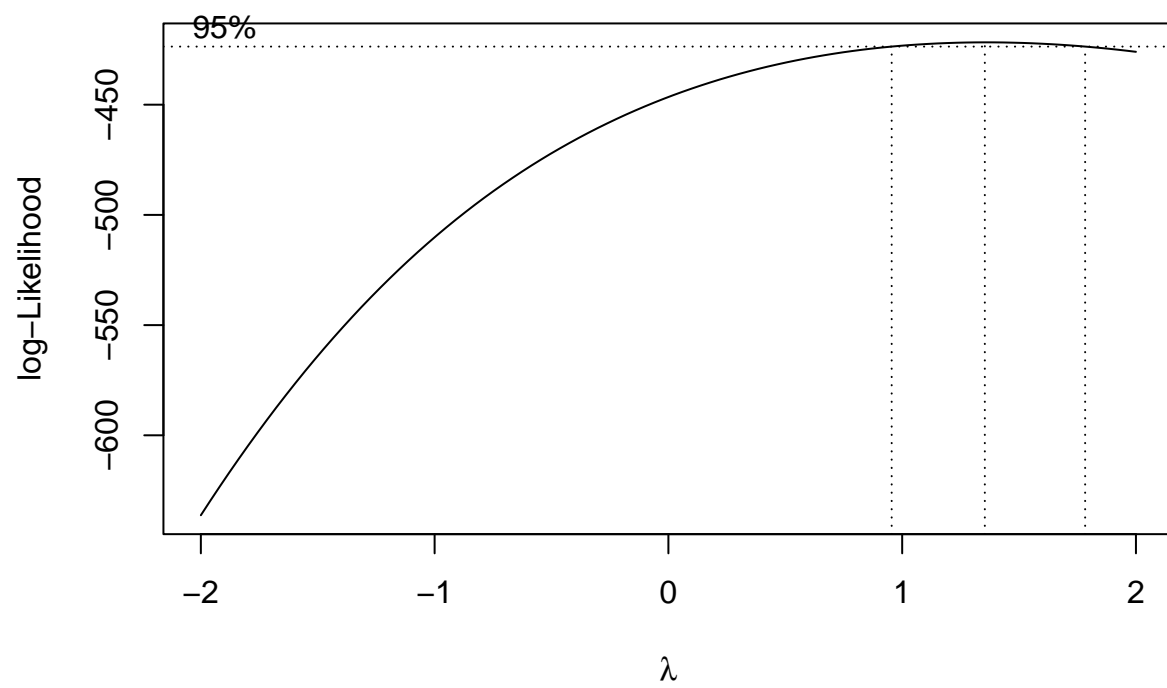
```
par(mfrow = c(2,2))
plot(model_writing_full)
```

```
## Warning:      :
## 186
```



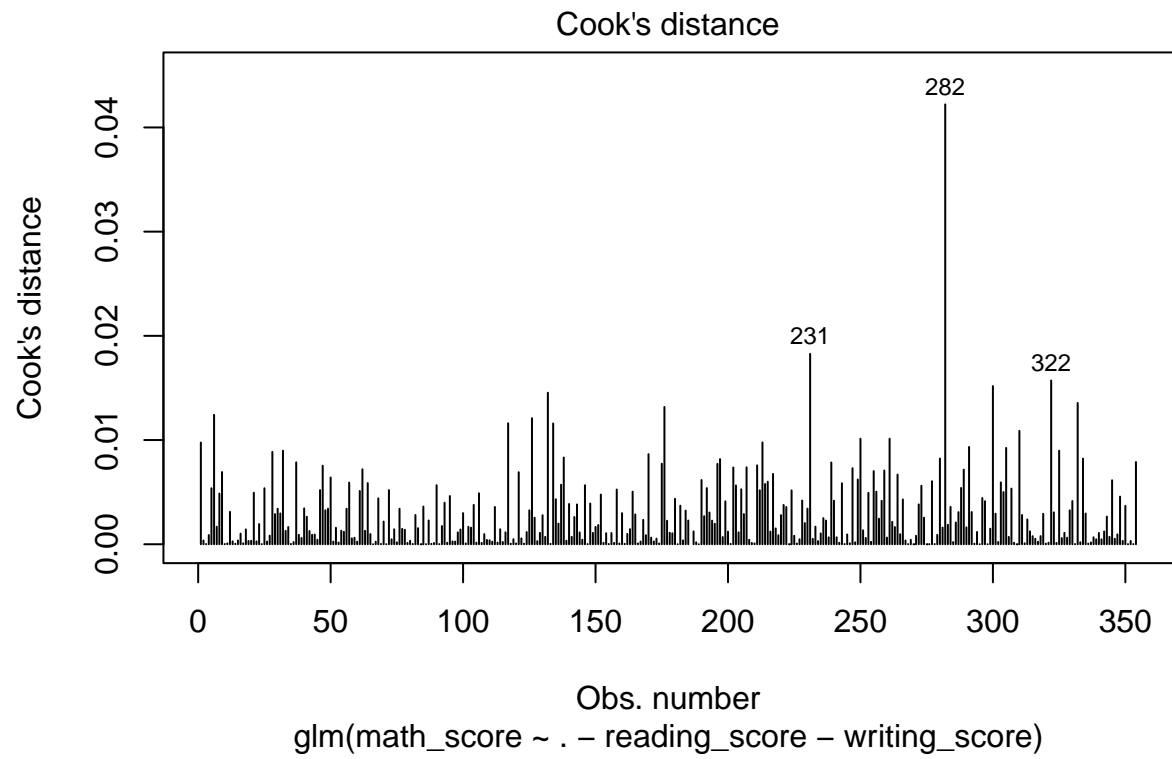
## Transformation

```
boxcox(model_reading_full)
```

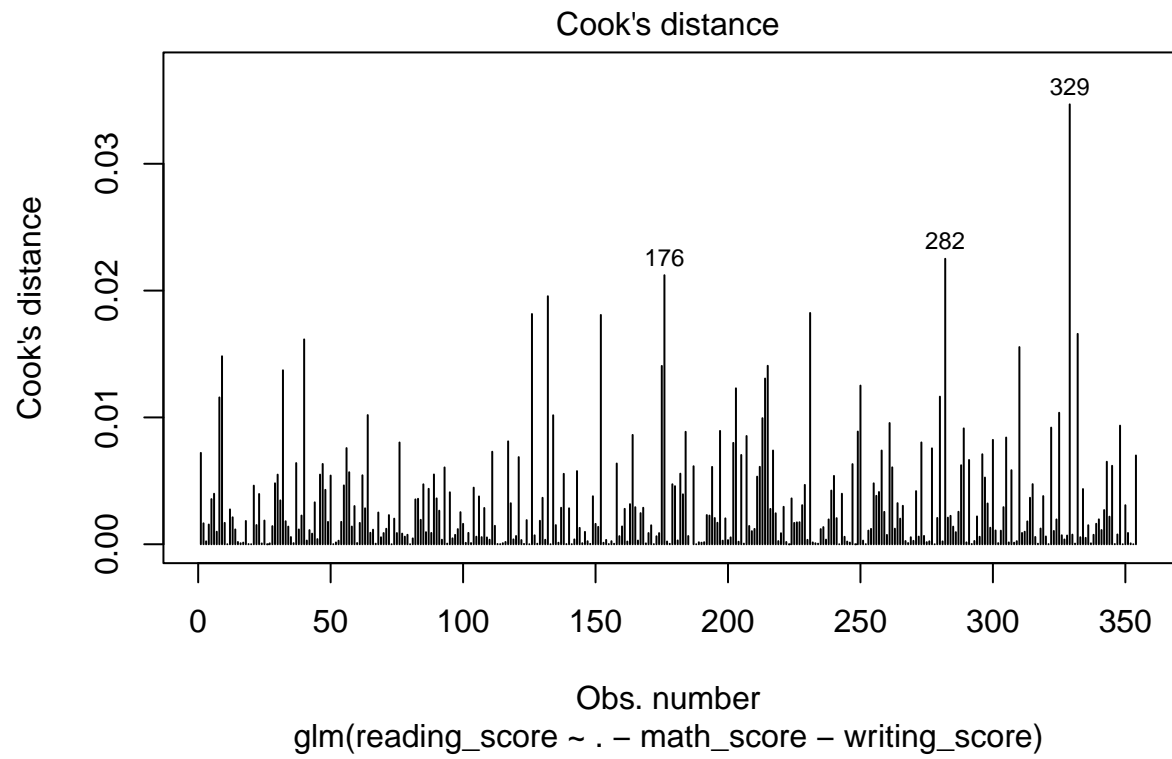


## Outlier and influence points

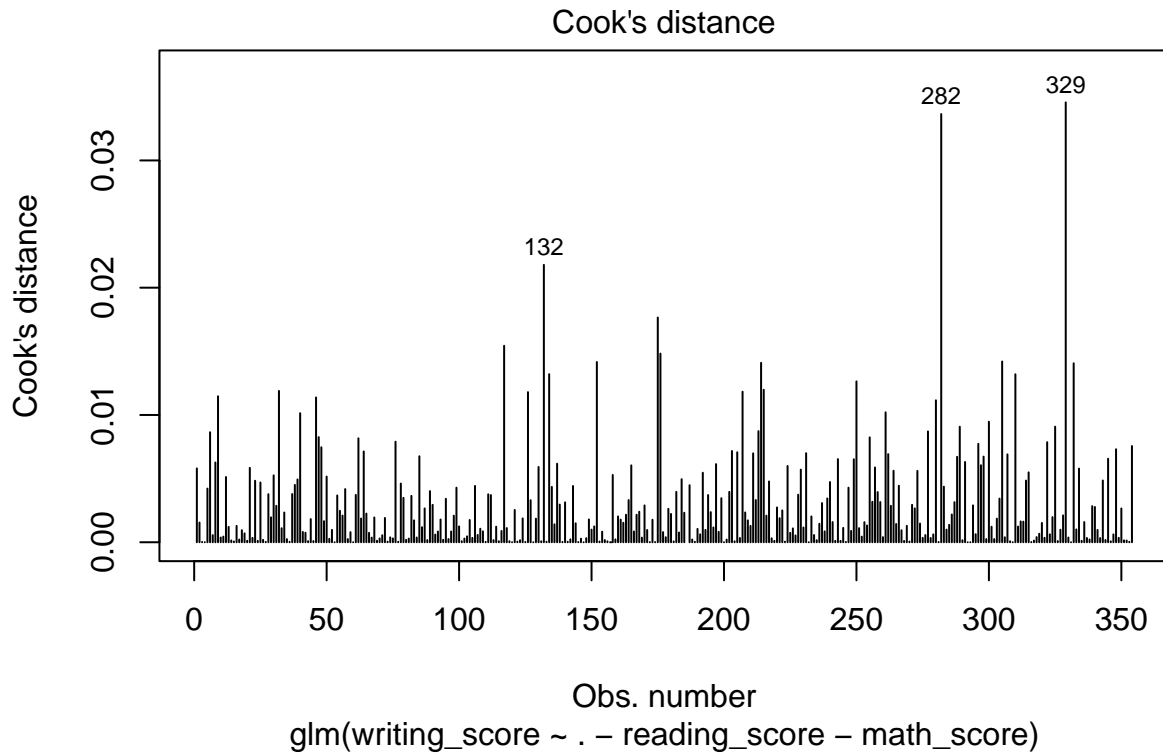
```
plot(model_math_full, which = 4)
```



```
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
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

```
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
transport_means	1.1	[1, 1.3]	0.9
wkly_study_hours	1.1	[1.1, 1.3]	0.9



## Model building for math

```
# backward model
step(model_math_full, direction='backward')

## Start:  AIC=2878.02
## 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 - writing_score
##
##           Df Deviance    AIC
## - nr_siblings      7    61456 2872.1
## - parent_educ       3    60735 2875.9
## - practice_sport    2    60412 2876.0
## - is_first_child    1    60075 2876.1
## - transport_means    1    60152 2876.5
## <none>              60068 2878.0
## - gender            1    61163 2882.4
## - parent_marital_status 3    62260 2884.7
## - wkly_study_hours  2    62582 2888.5
## - test_prep          1    62572 2890.5
## - ethnic_group       4    63860 2891.7
## - lunch_type         1    71425 2937.3
##
## Step:  AIC=2872.11
## math_score ~ gender + ethnic_group + parent_educ + lunch_type +
##   test_prep + parent_marital_status + practice_sport + is_first_child +
##   transport_means + wkly_study_hours
##
##           Df Deviance    AIC
## - parent_educ       3    62111 2869.9
## - is_first_child    1    61457 2870.1
## - practice_sport    2    61829 2870.2
## - transport_means    1    61514 2870.4
## <none>              61456 2872.1
## - gender            1    62644 2876.9
## - parent_marital_status 3    63819 2879.5
## - wkly_study_hours  2    63807 2881.4
## - test_prep          1    64028 2884.6
## - ethnic_group       4    65559 2887.0
## - lunch_type         1    73858 2935.2
##
## Step:  AIC=2869.86
## math_score ~ gender + ethnic_group + lunch_type + test_prep +
##   parent_marital_status + practice_sport + is_first_child +
##   transport_means + wkly_study_hours
##
##           Df Deviance    AIC
## - practice_sport    2    62417 2867.6
## - is_first_child    1    62113 2867.9
## - transport_means    1    62142 2868.0
```

```

## <none>                62111 2869.9
## - gender              1    63275 2874.4
## - parent_marital_status 3    64477 2877.1
## - wkly_study_hours     2    64331 2878.3
## - test_prep            1    64934 2883.6
## - ethnic_group         4    66259 2884.8
## - lunch_type           1    74436 2931.9
##
## Step: AIC=2867.6
## math_score ~ gender + ethnic_group + lunch_type + test_prep +
##   parent_marital_status + is_first_child + transport_means +
##   wkly_study_hours
##
##               Df Deviance    AIC
## - is_first_child    1    62425 2865.7
## - transport_means    1    62444 2865.8
## <none>                62417 2867.6
## - gender            1    63581 2872.1
## - parent_marital_status 3    64755 2874.6
## - wkly_study_hours   2    64625 2875.9
## - test_prep          1    65248 2881.3
## - ethnic_group       4    66529 2882.2
## - lunch_type         1    74657 2929.0
##
## Step: AIC=2865.65
## math_score ~ gender + ethnic_group + lunch_type + test_prep +
##   parent_marital_status + transport_means + wkly_study_hours
##
##               Df Deviance    AIC
## - transport_means    1    62453 2863.8
## <none>                62425 2865.7
## - gender            1    63583 2870.2
## - parent_marital_status 3    64773 2872.7
## - wkly_study_hours   2    64627 2873.9
## - test_prep          1    65251 2879.3
## - ethnic_group       4    66531 2880.2
## - lunch_type         1    74659 2927.0
##
## Step: AIC=2863.8
## math_score ~ gender + ethnic_group + lunch_type + test_prep +
##   parent_marital_status + wkly_study_hours
##
##               Df Deviance    AIC
## <none>                62453 2863.8
## - gender            1    63614 2868.3
## - parent_marital_status 3    64774 2870.7
## - wkly_study_hours   2    64646 2872.0
## - test_prep          1    65373 2878.0
## - ethnic_group       4    66550 2878.3
## - lunch_type         1    74664 2925.0
##
##
## Call: glm(formula = math_score ~ gender + ethnic_group + lunch_type +
##   test_prep + parent_marital_status + wkly_study_hours, data = data)

```

```
##
## Coefficients:
##             (Intercept)                gender1                ethnic_group1
##             67.3260                -3.7049                2.4461
##             ethnic_group2                ethnic_group3                ethnic_group4
##             0.3026                4.1687                10.1791
##             lunch_type1                test_prep1                parent_marital_status1
##             -12.3773                6.0788                -4.0821
## parent_marital_status2                parent_marital_status3                wkly_study_hours1
##             6.7982                -5.2507                5.9171
##             wkly_study_hours2
##             3.8301
##
## Degrees of Freedom: 353 Total (i.e. Null); 341 Residual
## Null Deviance: 89070
## Residual Deviance: 62450 AIC: 2864
```

```
model_math_fit_back = lm(formula = math_score ~ gender + ethnic_group + parent_educ +
  lunch_type + test_prep + parent_marital_status + practice_sport +
  is_first_child + wkly_study_hours, data = data)

summary(model_math_fit_back)
```

```
##
## Call:
## lm(formula = math_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep + parent_marital_status + practice_sport +
##     is_first_child + wkly_study_hours, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -42.641  -9.388   0.444  10.841  29.060
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    63.3058     4.0723  15.545 < 2e-16 ***
## gender1        -3.7768     1.4786  -2.554 0.011080 *
## ethnic_group1    2.0233     3.2739   0.618 0.536983
## ethnic_group2   -0.1921     3.1097  -0.062 0.950767
## ethnic_group3    3.5985     3.1572   1.140 0.255191
## ethnic_group4    9.8452     3.3254   2.961 0.003289 **
## parent_educ2     1.6680     1.7628   0.946 0.344724
## parent_educ3     3.1571     2.0672   1.527 0.127641
## parent_educ4     3.7243     2.5498   1.461 0.145058
## lunch_type1    -12.4609     1.5198  -8.199 5.22e-15 ***
## test_prep1       5.9501     1.5447   3.852 0.000140 ***
## parent_marital_status1 -4.1882     1.7844  -2.347 0.019505 *
## parent_marital_status2  7.3458     4.7089   1.560 0.119707
## parent_marital_status3 -4.9516     2.1536  -2.299 0.022104 *
## practice_sport1    3.1345     2.3452   1.337 0.182276
## practice_sport2    3.2766     2.4641   1.330 0.184500
## is_first_child1   -0.1431     1.5713  -0.091 0.927481
## wkly_study_hours1    6.1263     1.7189   3.564 0.000418 ***
## wkly_study_hours2    4.2272     2.2378   1.889 0.059752 .
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.55 on 335 degrees of freedom
## Multiple R-squared:  0.3094, Adjusted R-squared:  0.2723
## F-statistic: 8.338 on 18 and 335 DF,  p-value: < 2.2e-16

# lasso model
lambda_seq = 10 ^ seq(-3, 0, by = .1)

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

model_math_lasso = glmnet(as.matrix(data[1:11]), data$math_score, lambda = cv_object_math$lambda.min, a
coef(model_math_lasso)

## 12 x 1 sparse Matrix of class "dgCMatrix"
##
## (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
```

## Model building for reading

```
# backward model
step(model_reading_full, direction='backward')

## Start:  AIC=2843.29
## 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 Deviance    AIC
## - nr_siblings      7   55342 2835.0
```

```

## - practice_sport      2    54578 2840.1
## - transport_means     1    54476 2841.4
## - is_first_child      1    54482 2841.5
## - ethnic_group        4    55682 2843.2
## <none>                 54454 2843.3
## - parent_educ         3    56013 2847.3
## - parent_marital_status 3    56363 2849.5
## - wkly_study_hours    2    56459 2852.1
## - test_prep           1    58760 2868.2
## - lunch_type          1    59248 2871.2
## - gender              1    60054 2875.9
##
## Step: AIC=2835.02
## 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 Deviance    AIC
## - practice_sport      2    55488 2831.9
## - transport_means     1    55354 2833.1
## - is_first_child      1    55382 2833.3
## <none>                 55342 2835.0
## - ethnic_group        4    56661 2835.3
## - parent_educ         3    57024 2839.6
## - parent_marital_status 3    57267 2841.1
## - wkly_study_hours    2    57312 2843.4
## - test_prep           1    59565 2859.0
## - lunch_type          1    60780 2866.2
## - gender              1    61036 2867.7
##
## Step: AIC=2831.94
## reading_score ~ gender + ethnic_group + parent_educ + lunch_type +
##      test_prep + parent_marital_status + is_first_child + transport_means +
##      wkly_study_hours
##
##              Df Deviance    AIC
## - transport_means     1    55493 2830.0
## - is_first_child      1    55529 2830.2
## <none>                 55488 2831.9
## - ethnic_group        4    56782 2832.1
## - parent_educ         3    57143 2836.3
## - parent_marital_status 3    57391 2837.9
## - wkly_study_hours    2    57447 2840.2
## - test_prep           1    59804 2856.5
## - lunch_type          1    60909 2862.9
## - gender              1    61166 2864.4
##
## Step: AIC=2829.98
## reading_score ~ gender + ethnic_group + parent_educ + lunch_type +
##      test_prep + parent_marital_status + is_first_child + wkly_study_hours
##
##              Df Deviance    AIC
## - is_first_child      1    55533 2828.2
## <none>                 55493 2830.0

```

```

## - ethnic_group          4    56789 2830.2
## - parent_educ           3    57143 2834.3
## - parent_marital_status 3    57393 2835.9
## - wkly_study_hours      2    57452 2838.3
## - test_prep             1    59916 2855.1
## - lunch_type            1    60916 2861.0
## - gender                1    61168 2862.4
##
## Step: AIC=2828.23
## reading_score ~ gender + ethnic_group + parent_educ + lunch_type +
##      test_prep + parent_marital_status + wkly_study_hours
##
##              Df Deviance    AIC
## <none>                55533 2828.2
## - ethnic_group        4    56839 2828.5
## - parent_educ         3    57188 2832.6
## - parent_marital_status 3    57432 2834.1
## - wkly_study_hours    2    57508 2836.6
## - test_prep           1    60064 2854.0
## - lunch_type          1    60973 2859.3
## - gender              1    61177 2860.5
##
## Call: glm(formula = reading_score ~ gender + ethnic_group + parent_educ +
##      lunch_type + test_prep + parent_marital_status + wkly_study_hours,
##      data = data)
##
## Coefficients:
##      (Intercept)                gender1          ethnic_group1
##           61.6474                8.1816                1.8945
##      ethnic_group2          ethnic_group3          ethnic_group4
##           0.3778                3.3789                5.6870
##      parent_educ2          parent_educ3          parent_educ4
##           2.3964                4.6728                6.4917
##      lunch_type1          test_prep1 parent_marital_status1
##          -8.2631                7.6175          -4.5976
## parent_marital_status2 parent_marital_status3          wkly_study_hours1
##           4.1841          -4.3042                5.1565
##      wkly_study_hours2
##           1.0458
##
## Degrees of Freedom: 353 Total (i.e. Null); 338 Residual
## Null Deviance: 77470
## Residual Deviance: 55530 AIC: 2828

```

```

model_reading_back = lm(formula = reading_score ~ gender + ethnic_group + parent_educ +
      lunch_type + test_prep + parent_marital_status + is_first_child +
      transport_means + wkly_study_hours, data = data)
summary(model_reading_back)

```

```

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

```

```
##      lunch_type + test_prep + parent_marital_status + is_first_child +
##      transport_means + wkly_study_hours, data = data)
##
## Residuals:
##      Min        1Q    Median        3Q        Max
## -32.522  -9.335   0.253   9.491  29.948
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      61.0959      3.4189  17.870 < 2e-16 ***
## gender1           8.2151      1.4010   5.864 1.08e-08 ***
## ethnic_group1      1.8440      3.0962   0.596 0.55187
## ethnic_group2      0.3221      2.9318   0.110 0.91257
## ethnic_group3      3.3272      2.9801   1.116 0.26502
## ethnic_group4      5.6186      3.1503   1.784 0.07540 .
## parent_educ2       2.4730      1.6822   1.470 0.14248
## parent_educ3       4.7430      1.9674   2.411 0.01645 *
## parent_educ4       6.4579      2.4012   2.689 0.00751 **
## lunch_type1       -8.2690      1.4432  -5.730 2.24e-08 ***
## test_prep1        7.5208      1.4711   5.112 5.35e-07 ***
## parent_marital_status1 -4.5595      1.6944  -2.691 0.00748 **
## parent_marital_status2  4.3781      4.4330   0.988 0.32405
## parent_marital_status3 -4.3645      2.0421  -2.137 0.03330 *
## is_first_child1     0.7327      1.4725   0.498 0.61910
## transport_means1     0.2718      1.4551   0.187 0.85195
## wkly_study_hours1     5.1383      1.6296   3.153 0.00176 **
## wkly_study_hours2     1.0442      2.1217   0.492 0.62294
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.85 on 336 degrees of freedom
## Multiple R-squared:  0.2837, Adjusted R-squared:  0.2475
## F-statistic: 7.829 on 17 and 336 DF,  p-value: < 2.2e-16
```

```
# lasso model
lambda_seq = 10 ^ seq(-3, 0, by = .1)

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

```
## [1] 0.5011872
```

```
model_reading_lasso = glmnet(as.matrix(data[1:11]), data$reading_score, lambda = cv_object_reading$lambda.min,
                             coef(model_reading_lasso))
```

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##              s0
## (Intercept)  63.0047330
## gender       6.8714456
## ethnic_group 1.0191726
## parent_educ  1.6822432
```

```
## lunch_type -7.2445118
## test_prep 6.2890596
## parent_marital_status -0.7735146
## practice_sport .
## is_first_child .
## nr_siblings .
## transport_means .
## wkly_study_hours 0.4772919
```

```
model_reading_lasso$dev.ratio
```

```
## [1] 0.2302132
```

## Model building for writing

```
# backward model
step(model_writing_full, direction = "backward", )
```

```
## Start: AIC=2820.51
## 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 Deviance    AIC
## - nr_siblings      7   52079 2813.5
## - is_first_child    1   51064 2818.5
## - practice_sport    2   51421 2819.0
## - transport_means   1   51134 2819.0
## <none>              51060 2820.5
## - ethnic_group      4   52839 2824.6
## - parent_educ        3   53000 2827.7
## - parent_marital_status 3   53052 2828.1
## - wkly_study_hours   2   52961 2829.4
## - lunch_type         1   57235 2858.9
## - test_prep          1   57985 2863.5
## - gender             1   59341 2871.7
##
## Step: AIC=2813.5
## 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 Deviance    AIC
## - is_first_child    1   52080 2811.5
## - transport_means    1   52132 2811.9
## - practice_sport     2   52484 2812.2
## <none>              52079 2813.5
## - ethnic_group      4   53949 2818.0
## - parent_marital_status 3   54107 2821.0
```



```

## - parent_educ          3    54148 2821.3
## - wkly_study_hours     2    53910 2821.7
## - test_prep            1    58959 2855.4
## - lunch_type           1    59035 2855.9
## - gender               1    60523 2864.7
##
## Step:  AIC=2811.51
## writing_score ~ gender + ethnic_group + parent_educ + lunch_type +
##      test_prep + parent_marital_status + practice_sport + transport_means +
##      wkly_study_hours
##
##              Df Deviance    AIC
## - transport_means      1    52133 2809.9
## - practice_sport       2    52489 2810.3
## <none>                  52080 2811.5
## - ethnic_group         4    53950 2816.0
## - parent_marital_status 3    54109 2819.0
## - parent_educ          3    54149 2819.3
## - wkly_study_hours     2    53910 2819.7
## - test_prep            1    58988 2853.6
## - lunch_type           1    59035 2853.9
## - gender               1    60544 2862.8
##
## Step:  AIC=2809.87
## writing_score ~ gender + ethnic_group + parent_educ + lunch_type +
##      test_prep + parent_marital_status + practice_sport + wkly_study_hours
##
##              Df Deviance    AIC
## - practice_sport       2    52531 2808.6
## <none>                  52133 2809.9
## - ethnic_group         4    54035 2814.6
## - parent_marital_status 3    54120 2817.1
## - parent_educ          3    54175 2817.5
## - wkly_study_hours     2    53954 2818.0
## - lunch_type           1    59038 2851.9
## - test_prep            1    59324 2853.6
## - gender               1    60577 2861.0
##
## Step:  AIC=2808.56
## writing_score ~ gender + ethnic_group + parent_educ + lunch_type +
##      test_prep + parent_marital_status + wkly_study_hours
##
##              Df Deviance    AIC
## <none>                  52531 2808.6
## - ethnic_group         4    54482 2813.5
## - parent_educ          3    54457 2815.3
## - parent_marital_status 3    54494 2815.5
## - wkly_study_hours     2    54335 2816.5
## - lunch_type           1    59368 2849.9
## - test_prep            1    59741 2852.1
## - gender               1    61017 2859.6
##
##
## Call:  glm(formula = writing_score ~ gender + ethnic_group + parent_educ +

```

```
## lunch_type + test_prep + parent_marital_status + wkly_study_hours,
## data = data)
##
## Coefficients:
## (Intercept) gender1 ethnic_group1
## 58.522 10.032 2.213
## ethnic_group2 ethnic_group3 ethnic_group4
## 1.850 6.338 6.617
## parent_educ2 parent_educ3 parent_educ4
## 1.789 4.598 7.212
## lunch_type1 test_prep1 parent_marital_status1
## -9.263 9.609 -4.417
## parent_marital_status2 parent_marital_status3 wkly_study_hours1
## 4.668 -4.644 5.168
## wkly_study_hours2
## 1.893
##
## Degrees of Freedom: 353 Total (i.e. Null); 338 Residual
## Null Deviance: 81860
## Residual Deviance: 52530 AIC: 2809
```

```
model_writing_back = lm(formula = writing_score ~ gender + ethnic_group + parent_educ +
  lunch_type + test_prep + parent_marital_status + practice_sport +
  is_first_child + transport_means + wkly_study_hours, data = data)
summary(model_writing_back)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + ethnic_group + parent_educ +
## lunch_type + test_prep + parent_marital_status + practice_sport +
## is_first_child + transport_means + wkly_study_hours, data = data)
##
## Residuals:
## Min 1Q Median 3Q Max
## -35.016 -8.347 0.861 9.431 25.920
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 55.6226 3.7915 14.670 < 2e-16 ***
## gender1 10.0283 1.3627 7.359 1.45e-12 ***
## ethnic_group1 1.9857 3.0171 0.658 0.51089
## ethnic_group2 1.3766 2.8687 0.480 0.63164
## ethnic_group3 5.7836 2.9166 1.983 0.04819 *
## ethnic_group4 6.4017 3.0645 2.089 0.03747 *
## parent_educ2 1.8930 1.6347 1.158 0.24769
## parent_educ3 4.7742 1.9128 2.496 0.01305 *
## parent_educ4 7.5674 2.3506 3.219 0.00141 **
## lunch_type1 -9.3729 1.4034 -6.679 1.01e-10 ***
## test_prep1 9.5404 1.4363 6.642 1.25e-10 ***
## parent_marital_status1 -4.5162 1.6470 -2.742 0.00643 **
## parent_marital_status2 5.4329 4.3399 1.252 0.21150
## parent_marital_status3 -4.4594 1.9914 -2.239 0.02579 *
## practice_sport1 3.4669 2.1647 1.602 0.11020
## practice_sport2 3.0695 2.2715 1.351 0.17751
```

```
## is_first_child1      -0.1246      1.4489  -0.086  0.93152
## transport_means1     0.8261      1.4256   0.580  0.56263
## wkly_study_hours1    5.2430      1.5846   3.309  0.00104 **
## wkly_study_hours2    2.0645      2.0654   1.000  0.31826
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.49 on 334 degrees of freedom
## Multiple R-squared:  0.3638, Adjusted R-squared:  0.3276
## F-statistic: 10.05 on 19 and 334 DF,  p-value: < 2.2e-16
```

```
# lasso model
lambda_seq = 10 ^ seq(-3, 0, by = .1)

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

```
## [1] 0.5011872
```

```
model_writing_lasso = glmnet(as.matrix(data[1:11]), data$writing_score, lambda = cv_object_writing$lambda.min,
                             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
```