Code

Read and Clean Data

```
data <- read.csv("./data.csv") |>
  janitor::clean_names() |>
  mutate(
    gender = case_when(
      gender == "male" ~ 0,
      gender == "female" ~ 1,
      ),
    ethnic_group = 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 = 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 = case_when(
      lunch_type == "standard" ~ 0,
      lunch_type == "free/reduced" ~ 1,
      ),
    test_prep = case_when(
      test_prep == "none" ~ 0,
      test_prep == "completed" ~ 1,
      ),
    parent_marital_status = case_when(
      parent_marital_status == "married" ~ 0,
      parent_marital_status == "single" ~ 1,
      parent_marital_status == "widowed" ~ 2,
      parent_marital_status == "divorced" ~ 3,
      ),
    practice_sport = case_when(
      practice_sport == "never" ~ 0,
      practice_sport == "sometimes" ~ 1,
      practice_sport == "regularly" ~ 2,
      ),
    is_first_child = case_when(
      is_first_child == "no" ~ 0,
```

```
      is_first_child == "yes" ~ 1,
      ),
    transport_means = case_when(
      transport_means == "school_bus" ~ 0,
      transport_means == "private" ~ 1,
      ),
    wkly_study_hours = case_when(
      wkly_study_hours == "< 5" ~ 0,
      wkly_study_hours == "10-May" ~ 1,
      wkly_study_hours == "> 10" ~ 2,
      )
    )


# 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)
```

```
##   gender ethnic_group parent_educ lunch_type test_prep parent_marital_status
## 1      1           2           3          0         0                      0
## 2      1           2           1          0         0                      0
## 3      1           1           4          0         0                      1
## 4      0           0           2          1         0                      0
## 5      0           2           1          0         0                      0
## 6      1           1           2          0         0                      0
##   practice_sport is_first_child nr_siblings transport_means wkly_study_hours
## 1              2              1           3               0                0
## 2              1              1           0               0                1
## 3              1              1           4               0                0
## 4              0              0           1               0                1
## 5              1              1           0               0                1
## 6              2              1           1               0                1
##   math_score reading_score writing_score
## 1         71            71            74
## 2         69            90            88
## 3         87            93            91
## 4         45            56            42
## 5         76            78            75
## 6         73            84            79
```

```
# 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

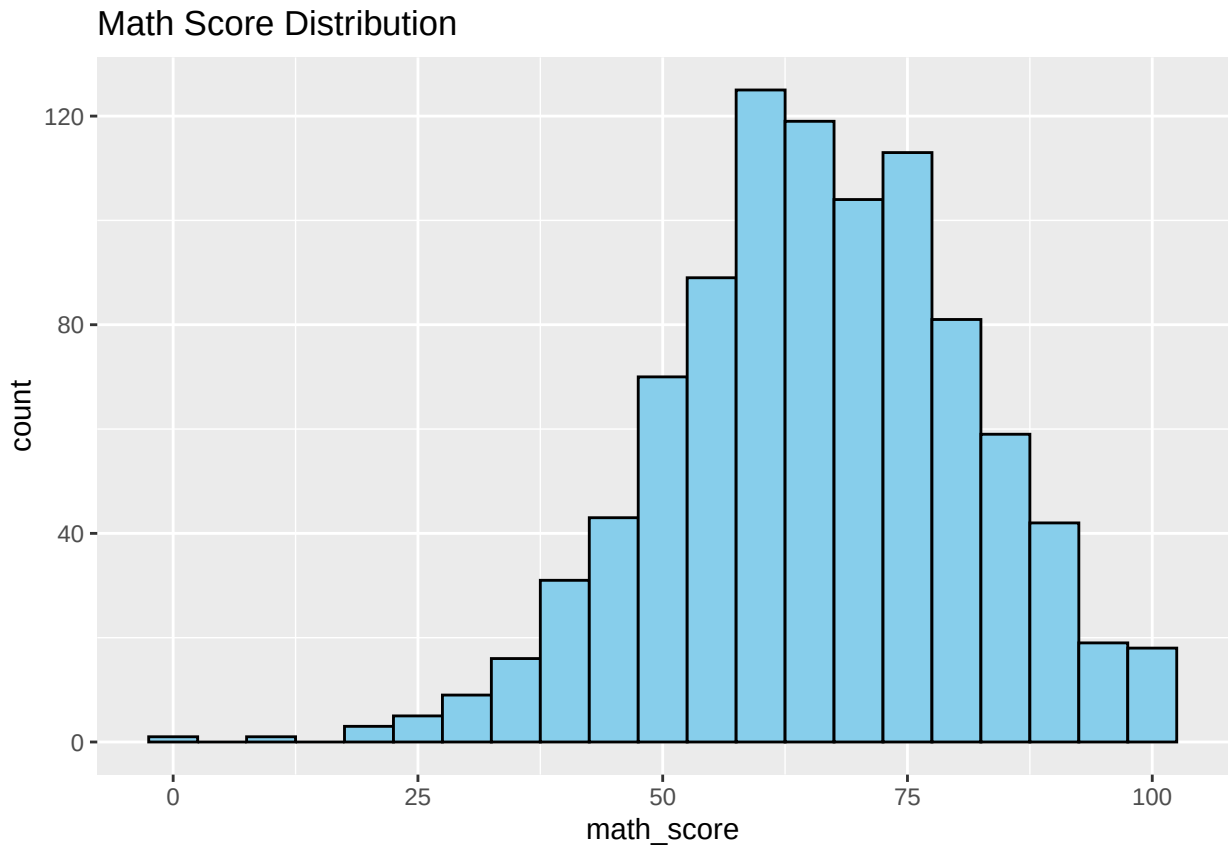
```
continuous_vars <- data[, c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14)]
summary_df <- data.frame(
Min = sapply(continuous_vars, min, na.rm = TRUE),
Q1 = sapply(continuous_vars, function(x) quantile(x, probs = 0.25, na.rm = TRUE)), Median = sapply(cont
Mean = sapply(continuous_vars, mean, na.rm = TRUE),
Q3 = sapply(continuous_vars, function(x) quantile(x, probs = 0.75, na.rm = TRUE)), Max = sapply(continu
)
kable(summary_df, caption = "Summary Statistics of Data", digits = 1)
```

Table 1: Summary Statistics of Data

	Min	Q1	Median	Mean	Q3	Max
gender	0	0	1.0	0.5	1.0	1
ethnic_group	0	1	2.0	2.2	3.0	4
parent_educ	1	2	2.0	2.0	2.0	4
lunch_type	0	0	0.0	0.3	1.0	1
test_prep	0	0	0.0	0.3	1.0	1
parent_marital_status	0	0	0.0	0.8	1.0	3
practice_sport	0	1	1.0	1.2	2.0	2
is_first_child	0	0	1.0	0.7	1.0	1
nr_siblings	0	1	2.0	2.1	3.0	7
transport_means	0	0	0.0	0.4	1.0	1
wkly_study_hours	0	0	1.0	0.9	1.0	2
math_score	0	56	66.0	66.0	76.0	100
reading_score	17	59	69.5	68.8	80.0	100
writing_score	10	57	68.0	67.9	78.2	100

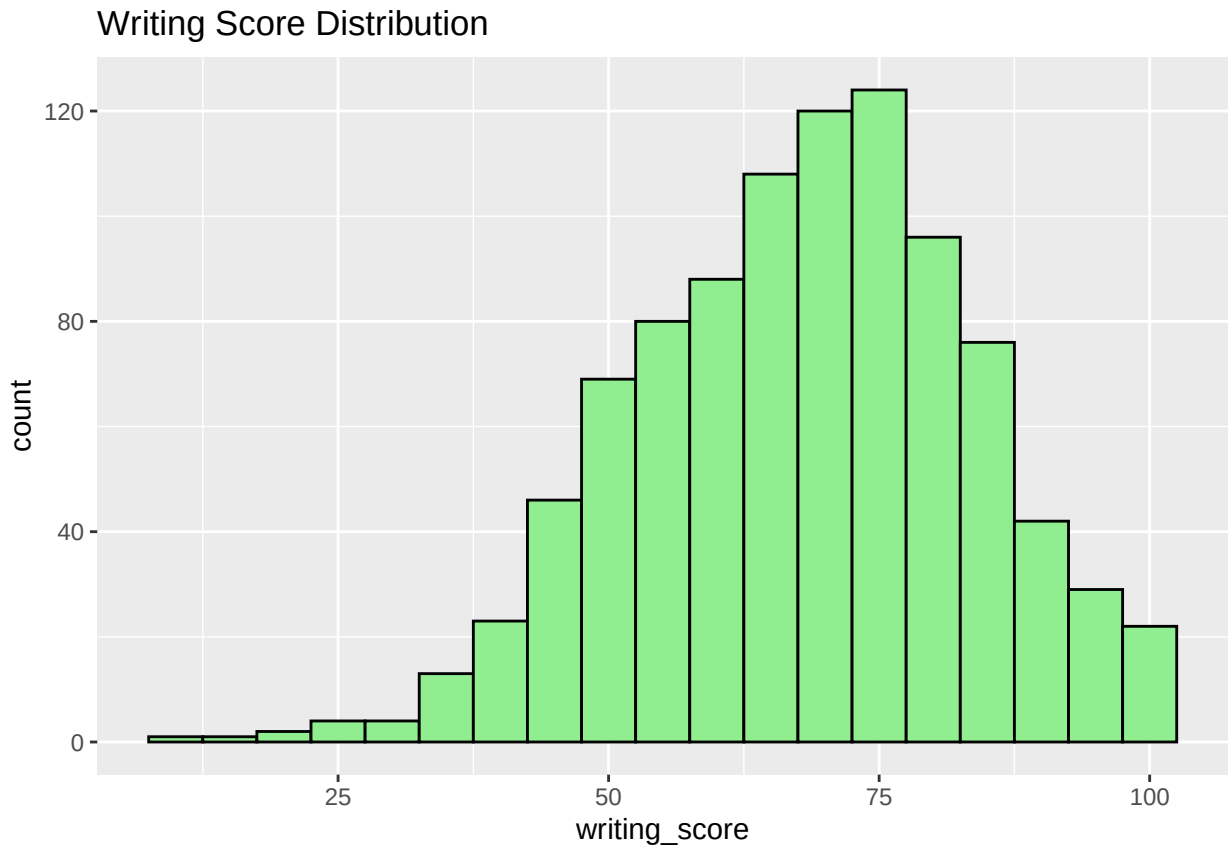
Histograms

```
ggplot(data, aes(x = math_score)) +
  geom_histogram(binwidth = 5, fill = "skyblue", color = "black") +
  labs(title = "Math Score Distribution")
```



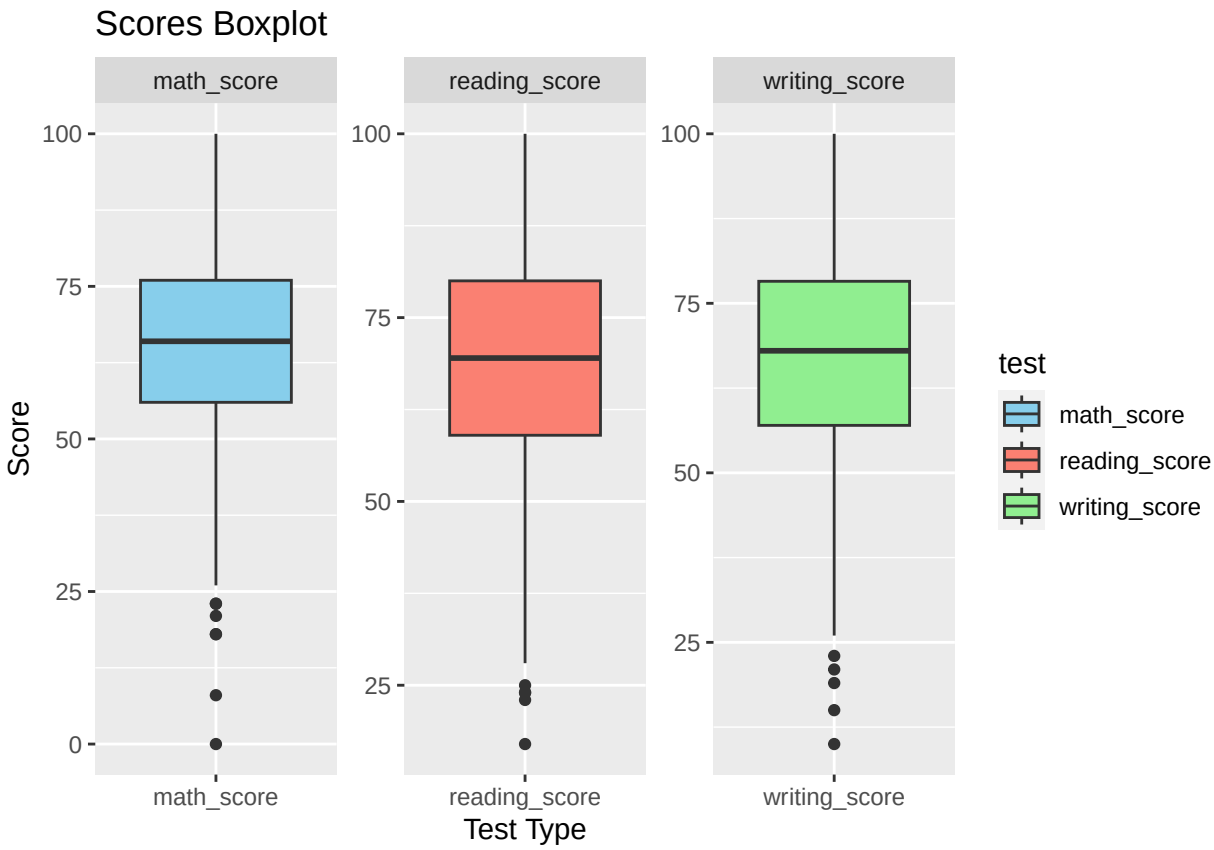

```
ggplot(data, aes(x = reading_score)) +
  geom_histogram(binwidth = 5, fill = "salmon", color = "black") +
  labs(title = "Reading Score Distribution")
```


```
ggplot(data, aes(x = writing_score)) +
  geom_histogram(binwidth = 5, fill = "lightgreen", color = "black") +
  labs(title = "Writing Score Distribution")
```



Boxplots

```
ggplot(data_long, aes(x = test, y = score, fill = test)) +
  geom_boxplot() +
  labs(title = "Scores Boxplot", x = "Test Type", y = "Score") +
  facet_wrap(~ test, scales = "free") +
  scale_fill_manual(values = c("skyblue", "salmon", "lightgreen"))
```



Diagnostics

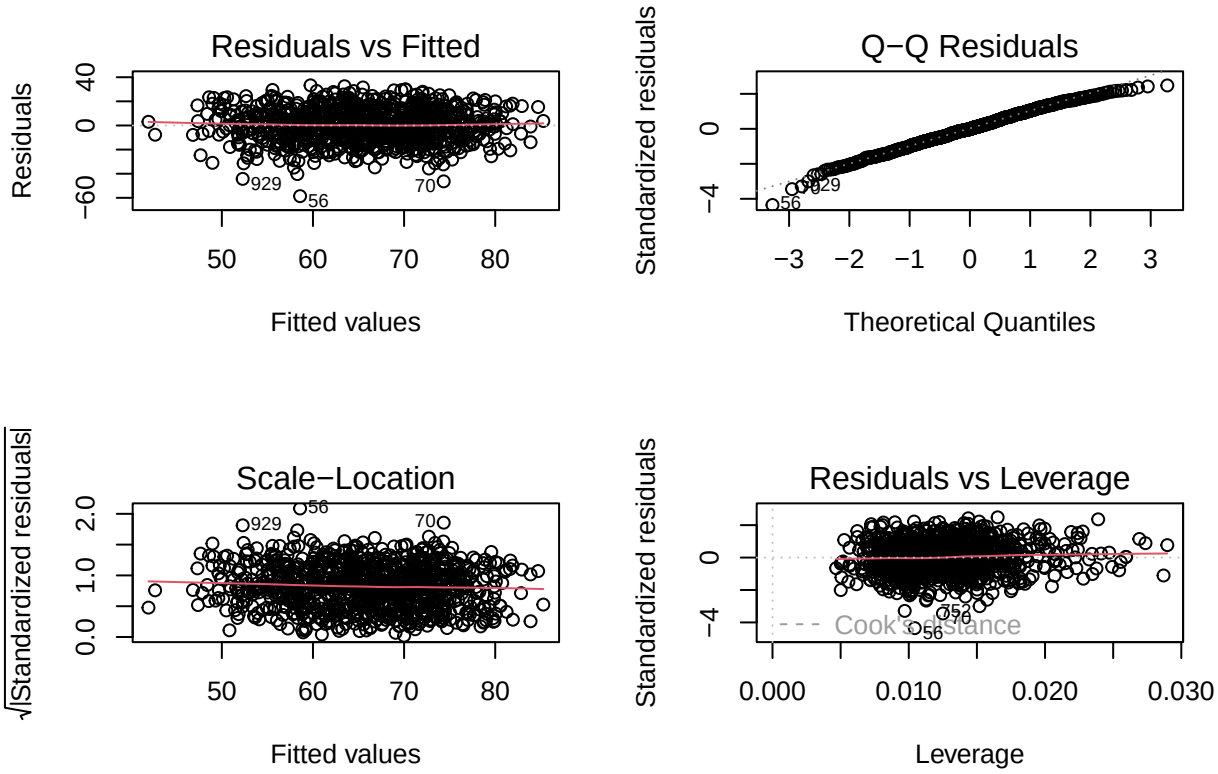
```
# Math
model_math_full = lm(math_score ~ .-reading_score -writing_score, data = data)

summary(model_math_full)
```

```
##
## Call:
## lm(formula = math_score ~ . - reading_score - writing_score,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.585  -8.814  -0.184   9.527  33.267
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           57.9983     2.2317  25.988  < 2e-16 ***
## gender                -5.0460     0.8814  -5.725 1.39e-08 ***
## ethnic_group           2.5850     0.3923   6.590 7.33e-11 ***
## parent_educ            1.1943     0.5965   2.002 0.045532 *
## lunch_type           -11.0862     0.9263 -11.969  < 2e-16 ***
## test_prep              5.6420     0.9328   6.048 2.11e-09 ***
```

```
## parent_marital_status  -1.4248     0.4133  -3.447 0.000591 ***
## practice_sport          1.3910     0.6807   2.044 0.041282 *
## is_first_child          2.5152     0.9467   2.657 0.008019 **
## nr_siblings             0.2883     0.3081   0.936 0.349645
## transport_means        -0.9564     0.9188  -1.041 0.298143
## wkly_study_hours        2.2115     0.6903   3.203 0.001404 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.53 on 936 degrees of freedom
## Multiple R-squared:   0.25,  Adjusted R-squared:  0.2411
## F-statistic: 28.36 on 11 and 936 DF,  p-value: < 2.2e-16
```

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



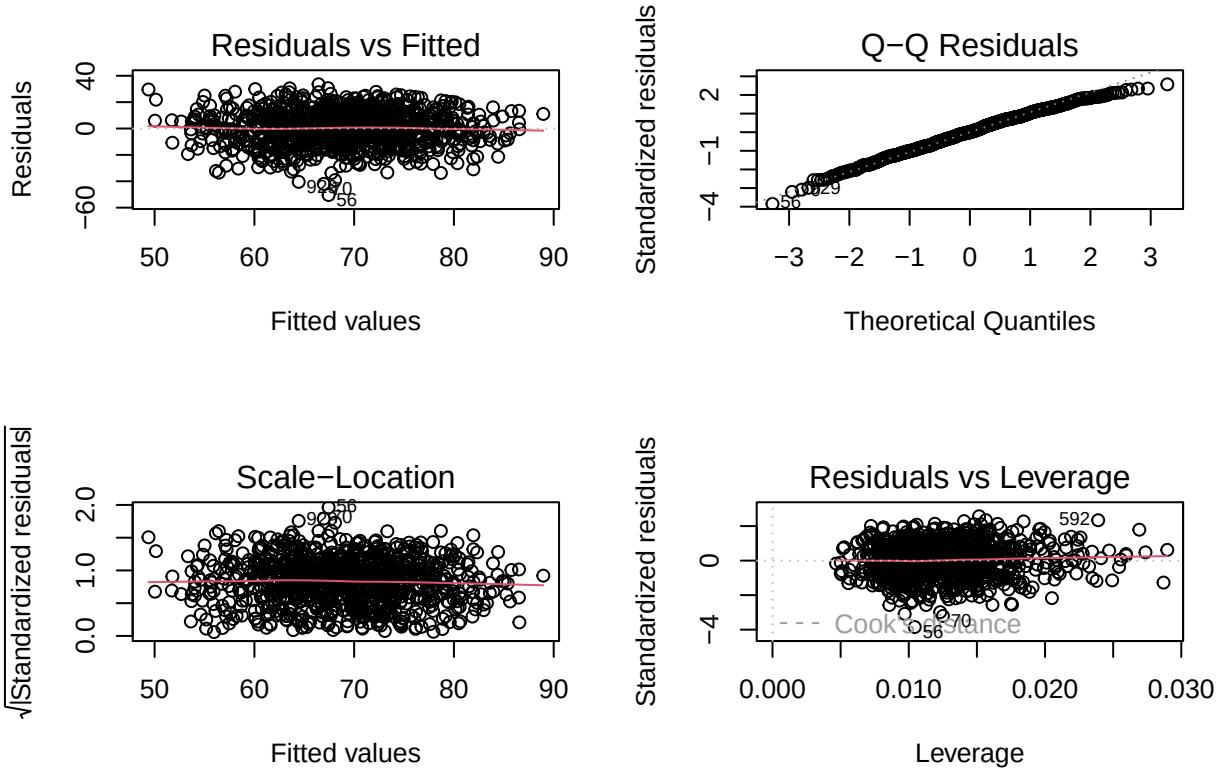

```
# Reading
model_reading_full = lm(reading_score ~ .-math_score -writing_score, data = data)

summary(model_reading_full)
```

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

```
## -50.440  -9.377   0.127   9.441  33.568
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             56.802514   2.172261  26.149  < 2e-16 ***
## gender                   7.292818   0.857973   8.500  < 2e-16 ***
## ethnic_group             1.628603   0.381808   4.266 2.20e-05 ***
## parent_educ              1.928543   0.580575   3.322 0.000929 ***
## lunch_type              -7.373578   0.901610  -8.178 9.33e-16 ***
## test_prep                6.974917   0.907968   7.682 3.95e-14 ***
## parent_marital_status -1.470414   0.402282  -3.655 0.000271 ***
## practice_sport        -0.044698   0.662570  -0.067 0.946229
## is_first_child           2.510168   0.921443   2.724 0.006566 **
## nr_siblings           -0.009534   0.299855  -0.032 0.974641
## transport_means       -1.566068   0.894307  -1.751 0.080247 .
## wkly_study_hours         1.341146   0.671953   1.996 0.046235 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.17 on 936 degrees of freedom
## Multiple R-squared:  0.2175, Adjusted R-squared:  0.2083
## F-statistic: 23.65 on 11 and 936 DF,  p-value: < 2.2e-16
```

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

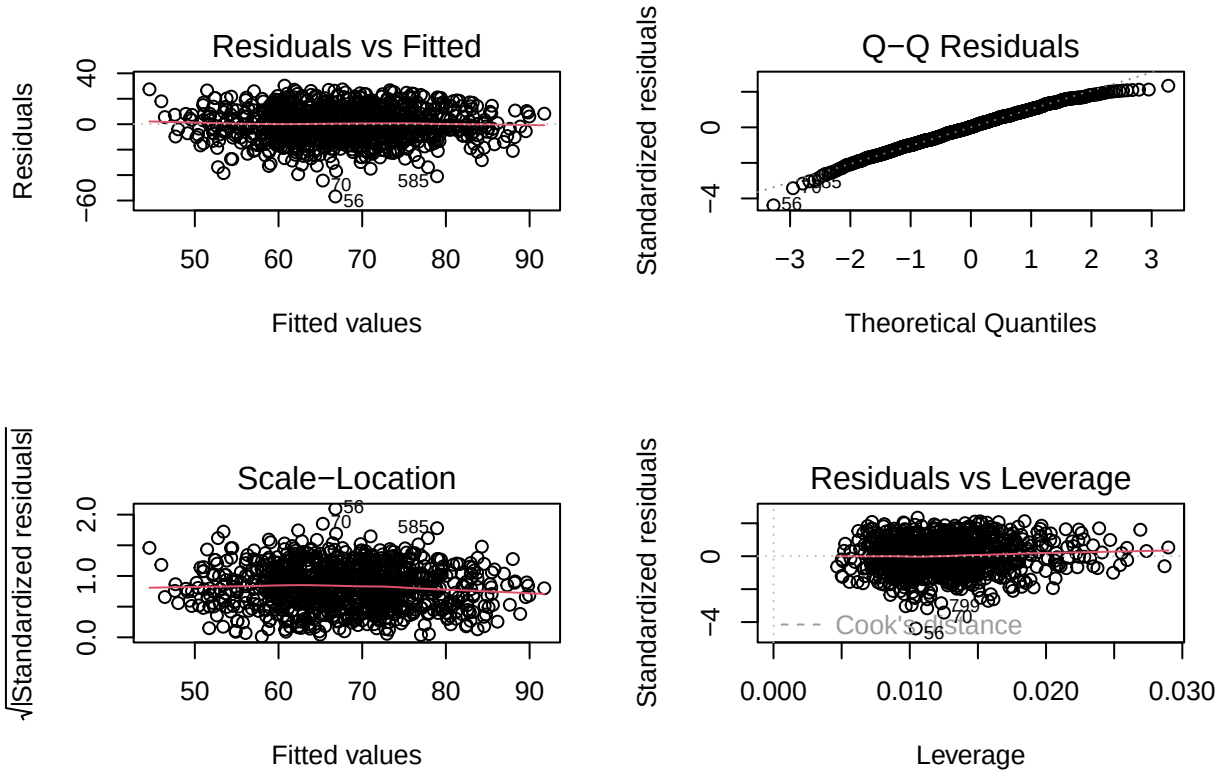



```r
# Writing
model_writing_full = lm(writing_score ~ .-reading_score -math_score, data = data)
```

```r
summary(model_writing_full)
```

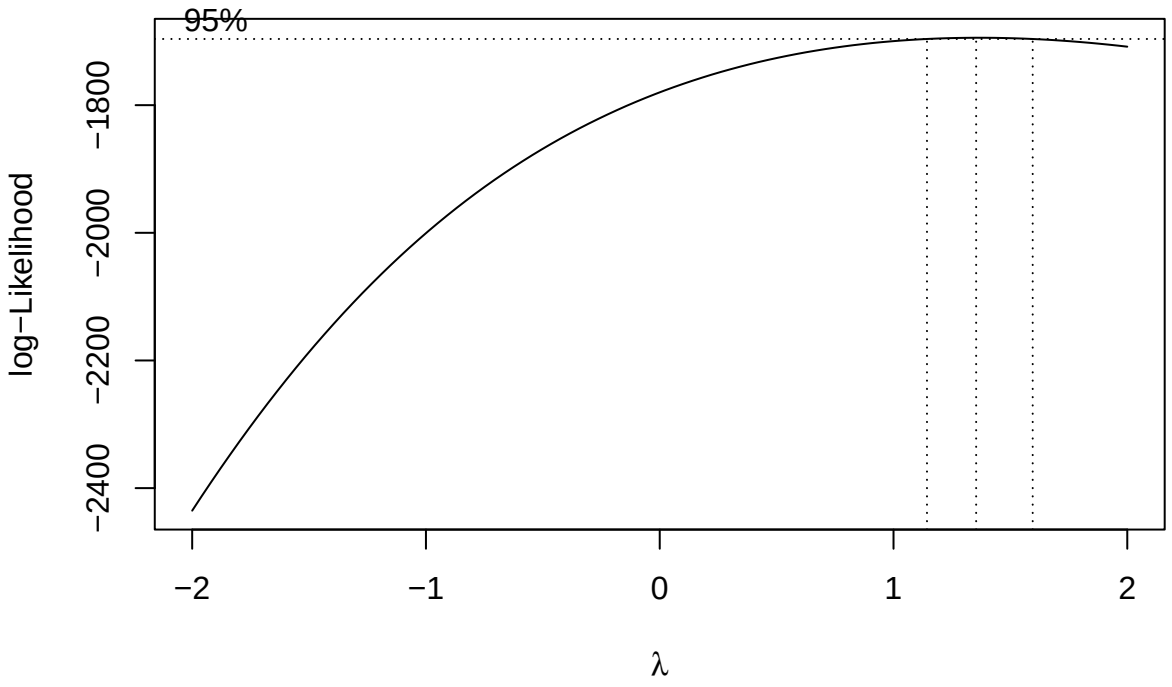
```
##
## Call:
## lm(formula = writing_score ~ . - reading_score - math_score,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.797  -8.970   0.215   9.004  30.278
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            51.2858     2.1468  23.890  < 2e-16 ***
## gender                  9.2050     0.8479  10.856  < 2e-16 ***
## ethnic_group            1.9744     0.3773   5.233 2.06e-07 ***
## parent_educ             2.2854     0.5738   3.983 7.33e-05 ***
## lunch_type             -8.1526     0.8910  -9.150  < 2e-16 ***
## test_prep               9.5146     0.8973  10.603  < 2e-16 ***
## parent_marital_status  -1.5514     0.3976  -3.902 0.000102 ***
## practice_sport          1.0852     0.6548   1.657 0.097798 .
## is_first_child          2.3026     0.9106   2.529 0.011615 *
## nr_siblings             0.1425     0.2963   0.481 0.630829
## transport_means        -1.3569     0.8838  -1.535 0.125045
## wkly_study_hours        1.2950     0.6641   1.950 0.051454 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.01 on 936 degrees of freedom
## Multiple R-squared:  0.2953, Adjusted R-squared:  0.287
## F-statistic: 35.66 on 11 and 936 DF,  p-value: < 2.2e-16
```

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



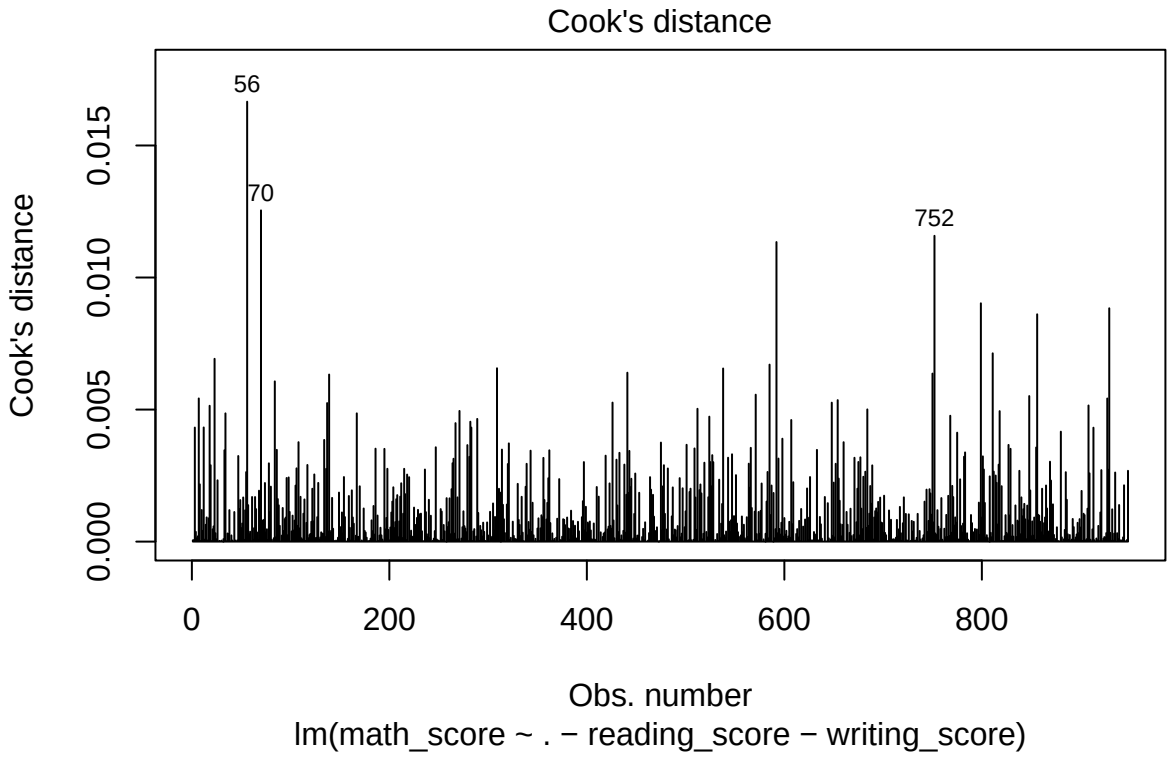
Transformation

```
boxcox(model_reading_full)
```

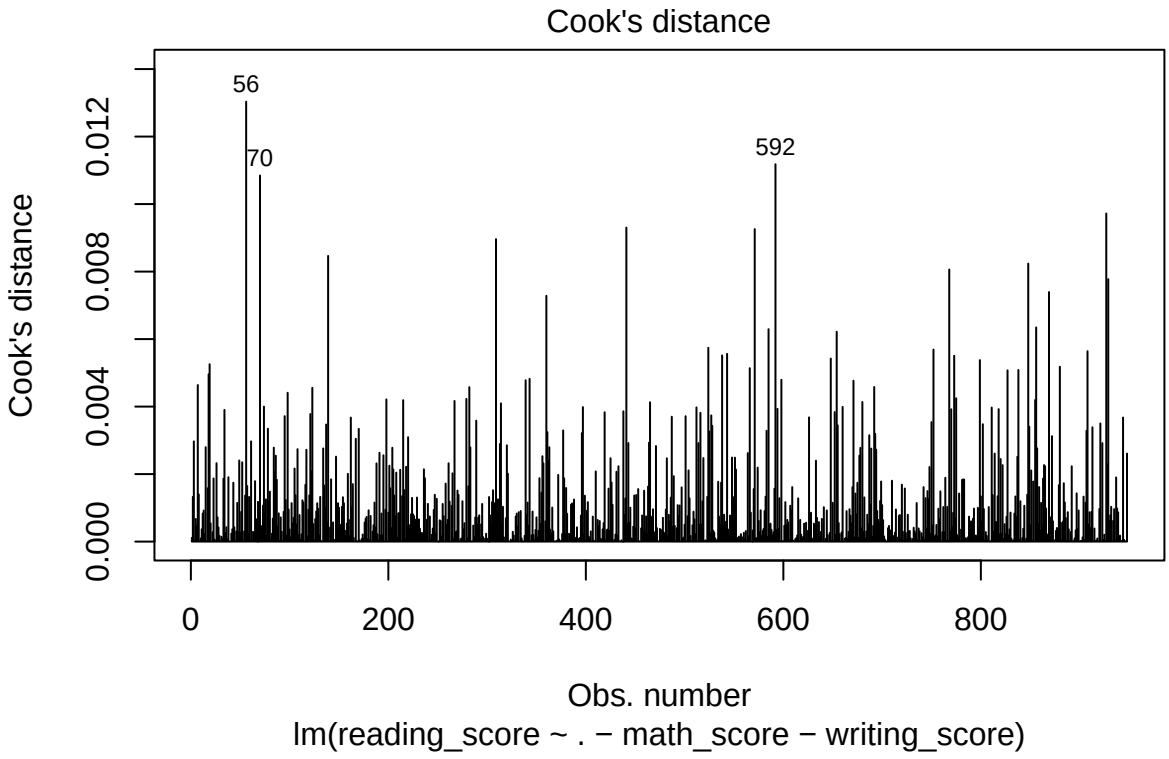


Outlier and influence points

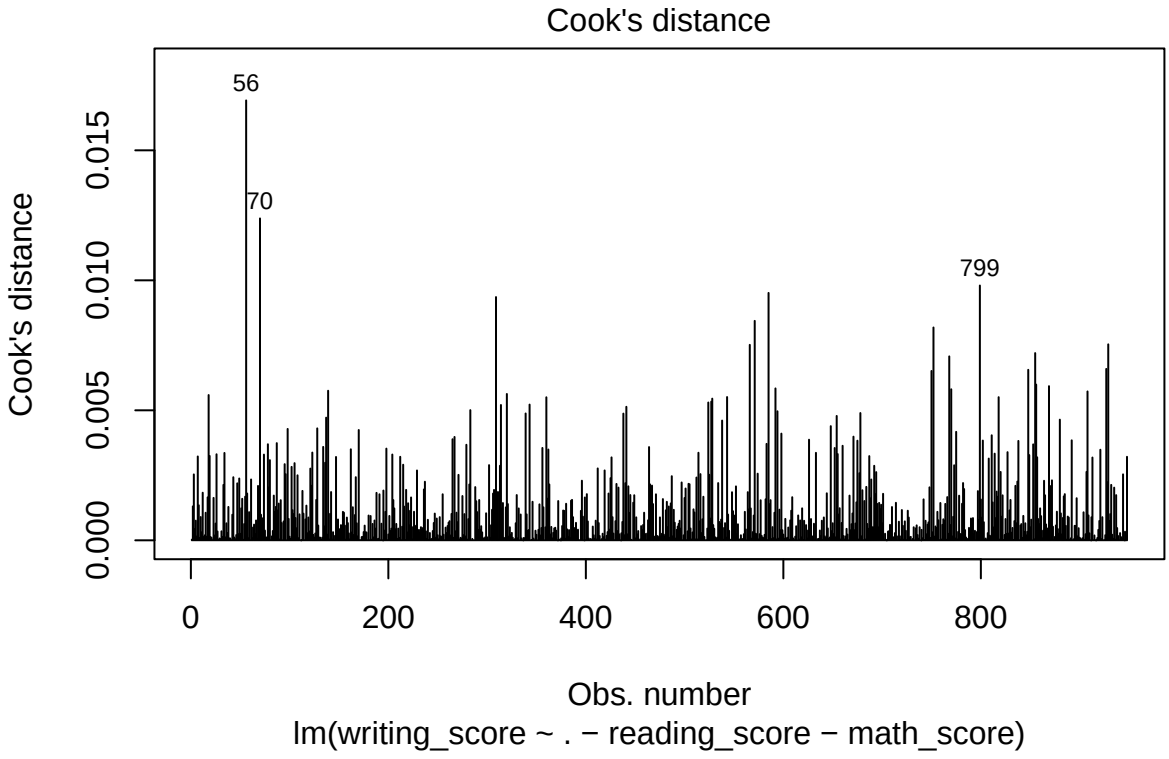
```
plot(model_math_full, which = 4)
```




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




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



Multicollinearity

```
# check VIF
performance::check_collinearity(model_math_full)
```

```
## # Check for Multicollinearity
##
## Low Correlation
##
##                   Term  VIF       VIF 95% CI Increased SE Tolerance
##                 gender 1.01 [1.00,  1066.42]         1.00      0.99
##           ethnic_group 1.01 [1.00,    25.44]         1.00      0.99
##            parent_educ 1.02 [1.00,     1.89]         1.01      0.98
##             lunch_type 1.01 [1.00,     7.40]         1.00      0.99
##              test_prep 1.01 [1.00,     4.97]         1.01      0.99
##  parent_marital_status 1.01 [1.00,     8.59]         1.00      0.99
##         practice_sport 1.01 [1.00,     9.58]         1.00      0.99
##         is_first_child 1.03 [1.00,     1.30]         1.01      0.97
##            nr_siblings 1.03 [1.00,     1.29]         1.01      0.97
##        transport_means 1.00 [1.00, 6.55e+12]         1.00      1.00
##       wkly_study_hours 1.02 [1.00,     1.51]         1.01      0.98
##  Tolerance 95% CI
##      [0.00, 1.00]
##      [0.04, 1.00]
##      [0.53, 1.00]
##      [0.14, 1.00]
##      [0.20, 1.00]
##      [0.12, 1.00]
##      [0.10, 1.00]
##      [0.77, 1.00]
##      [0.77, 1.00]
##      [0.00, 1.00]
##      [0.66, 1.00]
```

```
performance::check_collinearity(model_reading_full)
```

```
## # Check for Multicollinearity
##
## Low Correlation
##
##                   Term  VIF       VIF 95% CI Increased SE Tolerance
##                 gender 1.01 [1.00,  1066.42]         1.00      0.99
##           ethnic_group 1.01 [1.00,    25.44]         1.00      0.99
##            parent_educ 1.02 [1.00,     1.89]         1.01      0.98
##             lunch_type 1.01 [1.00,     7.40]         1.00      0.99
##              test_prep 1.01 [1.00,     4.97]         1.01      0.99
##  parent_marital_status 1.01 [1.00,     8.59]         1.00      0.99
##         practice_sport 1.01 [1.00,     9.58]         1.00      0.99
##         is_first_child 1.03 [1.00,     1.30]         1.01      0.97
##            nr_siblings 1.03 [1.00,     1.29]         1.01      0.97
##        transport_means 1.00 [1.00, 6.55e+12]         1.00      1.00
##       wkly_study_hours 1.02 [1.00,     1.51]         1.01      0.98
```

```
##  Tolerance 95% CI
##      [0.00, 1.00]
##      [0.04, 1.00]
##      [0.53, 1.00]
##      [0.14, 1.00]
##      [0.20, 1.00]
##      [0.12, 1.00]
##      [0.10, 1.00]
##      [0.77, 1.00]
##      [0.77, 1.00]
##      [0.00, 1.00]
##      [0.66, 1.00]
```

```
performance::check_collinearity(model_writing_full)
```

```
## # Check for Multicollinearity
##
## Low Correlation
##
##                   Term  VIF       VIF 95% CI Increased SE Tolerance
##                 gender 1.01 [1.00,  1066.42]         1.00      0.99
##           ethnic_group 1.01 [1.00,    25.44]         1.00      0.99
##            parent_educ 1.02 [1.00,     1.89]         1.01      0.98
##             lunch_type 1.01 [1.00,     7.40]         1.00      0.99
##              test_prep 1.01 [1.00,     4.97]         1.01      0.99
##  parent_marital_status 1.01 [1.00,     8.59]         1.00      0.99
##         practice_sport 1.01 [1.00,     9.58]         1.00      0.99
##         is_first_child 1.03 [1.00,     1.30]         1.01      0.97
##            nr_siblings 1.03 [1.00,     1.29]         1.01      0.97
##        transport_means 1.00 [1.00, 6.55e+12]         1.00      1.00
##       wkly_study_hours 1.02 [1.00,     1.51]         1.01      0.98
##  Tolerance 95% CI
##      [0.00, 1.00]
##      [0.04, 1.00]
##      [0.53, 1.00]
##      [0.14, 1.00]
##      [0.20, 1.00]
##      [0.12, 1.00]
##      [0.10, 1.00]
##      [0.77, 1.00]
##      [0.77, 1.00]
##      [0.00, 1.00]
##      [0.66, 1.00]
```

Model building for math

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

```
## Start:  AIC=4950.67
```

```
## 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 Sum of Sq    RSS    AIC
## - nr_siblings            1     160.3 171476 4949.6
## - transport_means        1     198.3 171514 4949.8
## <none>                               171316 4950.7
## - parent_educ            1     733.9 172050 4952.7
## - practice_sport         1     764.3 172080 4952.9
## - is_first_child         1    1292.1 172608 4955.8
## - wkly_study_hours       1    1878.3 173194 4959.0
## - parent_marital_status  1    2175.3 173491 4960.6
## - gender                 1    5998.3 177314 4981.3
## - test_prep              1    6695.8 178011 4985.0
## - ethnic_group           1    7949.0 179265 4991.7
## - lunch_type             1   26218.1 197534 5083.7
##
## Step:  AIC=4949.56
## 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 Sum of Sq    RSS    AIC
## - transport_means        1     195.5 171671 4948.6
## <none>                               171476 4949.6
## - parent_educ            1     707.0 172183 4951.5
## - practice_sport         1     756.7 172233 4951.7
## - is_first_child         1    1195.3 172671 4954.1
## - wkly_study_hours       1    1976.6 173453 4958.4
## - parent_marital_status  1    2193.5 173669 4959.6
## - gender                 1    5942.3 177418 4979.9
## - test_prep              1    6712.9 178189 4984.0
## - ethnic_group           1    7932.5 179408 4990.4
## - lunch_type             1   26102.5 197578 5081.9
##
## Step:  AIC=4948.64
## math_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + parent_marital_status + practice_sport + is_first_child +
##     wkly_study_hours
##
##                         Df Sum of Sq    RSS    AIC
## <none>                               171671 4948.6
## - parent_educ            1     699.1 172371 4950.5
## - practice_sport         1     748.4 172420 4950.8
## - is_first_child         1    1197.7 172869 4953.2
## - wkly_study_hours       1    1972.6 173644 4957.5
## - parent_marital_status  1    2197.9 173869 4958.7
## - gender                 1    5946.1 177618 4978.9
## - test_prep              1    6670.0 178341 4982.8
## - ethnic_group           1    7862.7 179534 4989.1
## - lunch_type             1   26112.8 197784 5080.9
```

```
##
## 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)
##
## Coefficients:
##           (Intercept)                 gender           ethnic_group
##                58.410                 -5.022                  2.570
##           parent_educ             lunch_type              test_prep
##                 1.165                -11.056                  5.630
## parent_marital_status         practice_sport         is_first_child
##                -1.432                  1.376                  2.400
##      wkly_study_hours
##                 2.259
```

```
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
## -58.040  -8.690  -0.140   9.655  32.095
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            58.4098     2.0894  27.955  < 2e-16 ***
## gender                 -5.0215     0.8810  -5.700 1.60e-08 ***
## ethnic_group            2.5696     0.3920   6.554 9.20e-11 ***
## parent_educ             1.1646     0.5959   1.954 0.050951 .
## lunch_type            -11.0563     0.9256 -11.945  < 2e-16 ***
## test_prep               5.6298     0.9326   6.037 2.26e-09 ***
## parent_marital_status  -1.4320     0.4132  -3.465 0.000553 ***
## practice_sport          1.3763     0.6806   2.022 0.043441 *
## is_first_child          2.4003     0.9383   2.558 0.010678 *
## wkly_study_hours        2.2594     0.6882   3.283 0.001065 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.53 on 938 degrees of freedom
## Multiple R-squared:  0.2484, Adjusted R-squared:  0.2412
## F-statistic: 34.45 on 9 and 938 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, al
coef(model_math_lasso)
```

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                                s0
## (Intercept)            58.1473174
## gender                 -5.0129779
## ethnic_group            2.5710502
## parent_educ             1.1687362
## lunch_type            -11.0531032
## test_prep               5.6137057
## parent_marital_status  -1.4097074
## practice_sport          1.3623978
## is_first_child          2.4732206
## nr_siblings             0.2751105
## transport_means        -0.9207641
## wkly_study_hours        2.1903871
```

```
model_math_lasso$dev.ratio
```

```
## [1] 0.2499478
```

Model building for reading

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

```
## Start:  AIC=4899.49
## 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            1       0.2 162312 4897.5
## - practice_sport         1       0.8 162313 4897.5
## <none>                               162312 4899.5
## - transport_means        1     531.8 162844 4900.6
## - wkly_study_hours       1     690.8 163003 4901.5
## - is_first_child         1    1286.9 163599 4905.0
## - parent_educ            1    1913.5 164225 4908.6
## - parent_marital_status  1    2316.8 164629 4910.9
## - ethnic_group           1    3155.1 165467 4915.7
## - test_prep              1   10233.2 172545 4955.4
## - lunch_type             1   11598.3 173910 4962.9
```

```
## - gender                 1   12529.0 174841 4968.0
##
## Step:  AIC=4897.49
## 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         1       0.8 162313 4895.5
## <none>                               162312 4897.5
## - transport_means        1     532.0 162844 4898.6
## - wkly_study_hours       1     693.3 163005 4899.5
## - is_first_child         1    1313.9 163626 4903.1
## - parent_educ            1    1918.2 164230 4906.6
## - parent_marital_status  1    2316.7 164629 4908.9
## - ethnic_group           1    3155.6 165468 4913.7
## - test_prep              1   10233.2 172545 4953.4
## - lunch_type             1   11617.5 173930 4961.0
## - gender                 1   12538.7 174851 4966.0
##
## Step:  AIC=4895.49
## 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
## <none>                               162313 4895.5
## - transport_means        1     532.5 162845 4896.6
## - wkly_study_hours       1     693.1 163006 4897.5
## - is_first_child         1    1315.6 163629 4901.1
## - parent_educ            1    1937.1 164250 4904.7
## - parent_marital_status  1    2316.0 164629 4906.9
## - ethnic_group           1    3157.1 165470 4911.8
## - test_prep              1   10232.6 172545 4951.5
## - lunch_type             1   11647.6 173960 4959.2
## - gender                 1   12538.0 174851 4964.0

##
## 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)
##
## Coefficients:
##           (Intercept)                 gender           ethnic_group
##                56.718                  7.292                 1.629
##           parent_educ             lunch_type              test_prep
##                 1.932                 -7.377                 6.974
## parent_marital_status         is_first_child        transport_means
##                -1.470                  2.515                -1.567
##      wkly_study_hours
##                 1.339
```

```r
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
## -50.479  -9.377   0.140   9.409  33.532
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              56.7180     1.8310  30.976  < 2e-16 ***
## gender                    7.2917     0.8566   8.512  < 2e-16 ***
## ethnic_group              1.6290     0.3814   4.271 2.14e-05 ***
## parent_educ               1.9324     0.5776   3.346 0.000853 ***
## lunch_type               -7.3773     0.8992  -8.204 7.60e-16 ***
## test_prep                 6.9744     0.9070   7.690 3.72e-14 ***
## parent_marital_status    -1.4699     0.4018  -3.658 0.000268 ***
## is_first_child            2.5152     0.9122   2.757 0.005940 **
## transport_means          -1.5669     0.8933  -1.754 0.079729 .
## wkly_study_hours          1.3393     0.6692   2.001 0.045647 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.15 on 938 degrees of freedom
## Multiple R-squared:  0.2175, Adjusted R-squared:   0.21
## F-statistic: 28.96 on 9 and 938 DF,  p-value: < 2.2e-16
```

```r
# 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] 0.1258925
```

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

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                               s0
## (Intercept)           57.558779
## gender                 7.037495
## ethnic_group           1.524557
```

```
## parent_educ            1.761090
## lunch_type            -7.119427
## test_prep              6.735154
## parent_marital_status -1.342417
## practice_sport         .
## is_first_child         2.248298
## nr_siblings            .
## transport_means       -1.288309
## wkly_study_hours       1.150146
```

```
model_reading_lasso$dev.ratio
```

```
## [1] 0.2168212
```

Model building for writing

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

```
## Start:  AIC=4877.12
## 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            1      39.1 158565 4875.3
## <none>                                158526 4877.1
## - transport_means        1     399.2 158925 4877.5
## - practice_sport         1     465.2 158991 4877.9
## - wkly_study_hours       1     644.1 159170 4879.0
## - is_first_child         1    1082.9 159609 4881.6
## - parent_marital_status  1    2579.0 161105 4890.4
## - parent_educ            1    2687.0 161213 4891.0
## - ethnic_group           1    4637.2 163163 4902.4
## - lunch_type             1   14178.3 172705 4956.3
## - test_prep              1   19042.2 177568 4982.7
## - gender                 1   19960.5 178487 4987.5
##
## Step:  AIC=4875.35
## 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
## <none>                                158565 4875.3
## - transport_means        1     397.3 158963 4875.7
## - practice_sport         1     462.2 159028 4876.1
## - wkly_study_hours       1     673.1 159238 4877.4
## - is_first_child         1    1047.4 159613 4879.6
```

```
## - parent_marital_status  1    2589.1 161154 4888.7
## - parent_educ            1    2664.8 161230 4889.1
## - ethnic_group           1    4631.0 163196 4900.6
## - lunch_type             1   14142.7 172708 4954.3
## - test_prep              1   19057.3 177623 4980.9
## - gender                 1   20036.7 178602 4986.2

##
## 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)
##
## Coefficients:
##           (Intercept)                 gender           ethnic_group
##                51.626                  9.218                 1.973
##           parent_educ             lunch_type              test_prep
##                 2.274                 -8.137                 9.518
## parent_marital_status         practice_sport         is_first_child
##                -1.554                  1.082                 2.245
##       transport_means       wkly_study_hours
##                -1.354                  1.320
```

```
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
## -56.695  -8.841   0.236   9.143  30.535
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            51.6264     2.0257  25.486  < 2e-16 ***
## gender                  9.2179     0.8471  10.881  < 2e-16 ***
## ethnic_group            1.9730     0.3772   5.231 2.08e-07 ***
## parent_educ             2.2740     0.5730   3.968 7.79e-05 ***
## lunch_type             -8.1367     0.8901  -9.142  < 2e-16 ***
## test_prep               9.5181     0.8969  10.612  < 2e-16 ***
## parent_marital_status  -1.5542     0.3974  -3.911 9.84e-05 ***
## practice_sport          1.0817     0.6545   1.653   0.0987 .
## is_first_child          2.2446     0.9022   2.488   0.0130 *
## transport_means        -1.3535     0.8834  -1.532   0.1258
## wkly_study_hours        1.3199     0.6618   1.994   0.0464 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 13.01 on 937 degrees of freedom
## Multiple R-squared:  0.2951, Adjusted R-squared:  0.2876
## F-statistic: 39.23 on 10 and 937 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] 0.03162278
```

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

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                                s0
## (Intercept)           51.6340508
## gender                 9.1437299
## ethnic_group           1.9475822
## parent_educ            2.2363843
## lunch_type            -8.0817362
## test_prep              9.4555121
## parent_marital_status -1.5203082
## practice_sport         1.0286104
## is_first_child         2.2240245
## nr_siblings            0.1176168
## transport_means       -1.2855322
## wkly_study_hours       1.2521530
```

```
model_writing_lasso$dev.ratio
```

```
## [1] 0.29526
```