ncl - variable selection

2023-12-04

Cleaned datasets

```
step_df = read_csv("data/Project_1_data.csv") |>
  drop_na() |> janitor::clean_names() |>
  mutate(
    wkly_study_hours = ifelse(
      wkly_study_hours == "10-May", "5-10", wkly_study_hours)
  )|>
  mutate(
    gender = as.numeric(factor(gender)),
    ethnic_group = as.numeric(factor(ethnic_group)),
    parent_educ = as.numeric(factor(
      parent_educ,levels= c("some high school", "high school",
                            "associate's degree", "some college",
                            "bachelor's degree", "master's degree"))),
    lunch_type = as.numeric(factor(lunch_type)),
    test_prep = as.numeric(factor(test_prep)),
    parent_marital_status = as.numeric(factor(parent_marital_status)),
    practice_sport = as.numeric(
      factor(practice_sport, levels = c("never", "sometimes", "regularly"))),
    is_first_child = as.numeric(factor(is_first_child)),
    transport_means = as.numeric(as.factor(transport_means)),
     wkly_study_hours = as.numeric(factor(wkly_study_hours,
                              levels = c("< 5", "5-10", "> 10")))
  )

math_df = dplyr::select(step_df, -c(reading_score, writing_score))

reading_df = dplyr::select(step_df, -c(math_score, writing_score))

writing_df = dplyr::select(step_df, -c(reading_score, math_score))
```

Step-wise + criteria-based: stepAIC()

Math Score

```
math_mlr <- lm(math_score ~., data = math_df)

mathstep.model <- stepAIC(math_mlr, direction = "both",
                      trace = FALSE)
summary(mathstep.model)
```

```
##
## Call:
## lm(formula = math_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep + nr_siblings + wkly_study_hours,
##     data = math_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.440  -8.894   0.776  10.134  32.889
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       28.0713     4.2756   6.565 1.15e-10 ***
## gender             5.3017     1.1486   4.616 4.83e-06 ***
## ethnic_group       2.7439     0.4896   5.605 3.23e-08 ***
## parent_educ        1.5210     0.3826   3.976 7.90e-05 ***
## lunch_type        12.5737     1.1964  10.510  < 2e-16 ***
## test_prep         -5.2926     1.1989  -4.414 1.21e-05 ***
## nr_siblings        0.6927     0.3860   1.795   0.0732 .
## wkly_study_hours   2.0825     0.8723   2.387   0.0173 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.8 on 579 degrees of freedom
## Multiple R-squared:  0.2758, Adjusted R-squared:  0.2671
## F-statistic:  31.5 on 7 and 579 DF,  p-value: < 2.2e-16
```

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

Reading Score

```r
reading_mlr <- lm(reading_score ~., data = reading_df)

readstep.model <- stepAIC(reading_mlr, direction = "both",
                      trace = FALSE)
summary(readstep.model)
```

```
##
## Call:
## lm(formula = reading_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep, data = reading_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.354  -8.959   0.802   9.901  32.216
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   66.7121     3.6485  18.285  < 2e-16 ***
## gender        -7.5066     1.1139  -6.739 3.84e-11 ***
```

```
## ethnic_group   1.7930     0.4753   3.773 0.000178 ***
## parent_educ    1.7606     0.3713   4.742 2.66e-06 ***
## lunch_type     8.6667     1.1618   7.459 3.18e-13 ***
## test_prep     -6.8289     1.1580  -5.897 6.28e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.4 on 581 degrees of freedom
## Multiple R-squared:  0.2264, Adjusted R-squared:  0.2197
## F-statistic: 34.01 on 5 and 581 DF,  p-value: < 2.2e-16
```

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

Writing Score

```r
writing_mlr <- lm(writing_score ~., data = writing_df)

writestep.model <- stepAIC(writing_mlr, direction = "both",
                      trace = FALSE)
summary(writestep.model)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep + wkly_study_hours, data = writing_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.917  -8.391   0.613   9.143  29.293
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       65.3125     3.8874  16.801  < 2e-16 ***
## gender            -9.1792     1.0698  -8.581  < 2e-16 ***
## ethnic_group       2.1684     0.4562   4.753 2.53e-06 ***
## parent_educ        2.3242     0.3566   6.519 1.54e-10 ***
## lunch_type         9.4976     1.1151   8.517  < 2e-16 ***
## test_prep         -9.0360     1.1163  -8.094 3.40e-15 ***
## wkly_study_hours   1.1762     0.8121   1.448    0.148
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.86 on 580 degrees of freedom
## Multiple R-squared:  0.3233, Adjusted R-squared:  0.3163
## F-statistic: 46.18 on 6 and 580 DF,  p-value: < 2.2e-16
```

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

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

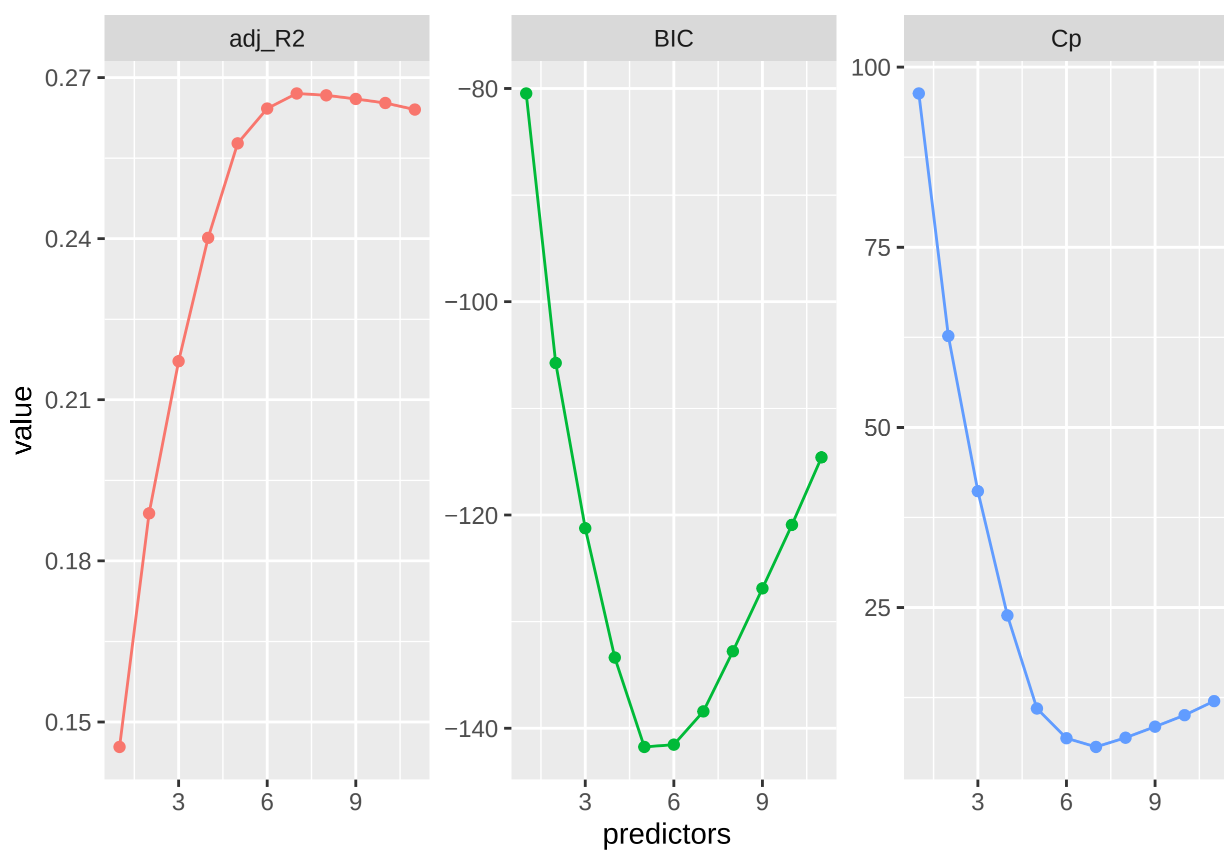
Criteria-based approach - Adjusted R^2, Cp, and BIC

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

Math Score

```
# perform best subset selection
best_subset <- regsubsets(math_score ~ ., math_df, nvmax = 11)
results <- summary(best_subset)

# extract and plot results
tibble(predictors = 1:11,
       adj_R2 = results$adjr2,
       Cp = results$cp,
       BIC = results$bic) |>
  gather(statistic, value, -predictors) |>
  ggplot(aes(predictors, value, color = statistic)) +
  geom_line(show.legend = F) +
  geom_point(show.legend = F) +
  facet_wrap(~ statistic, scales = "free")
```

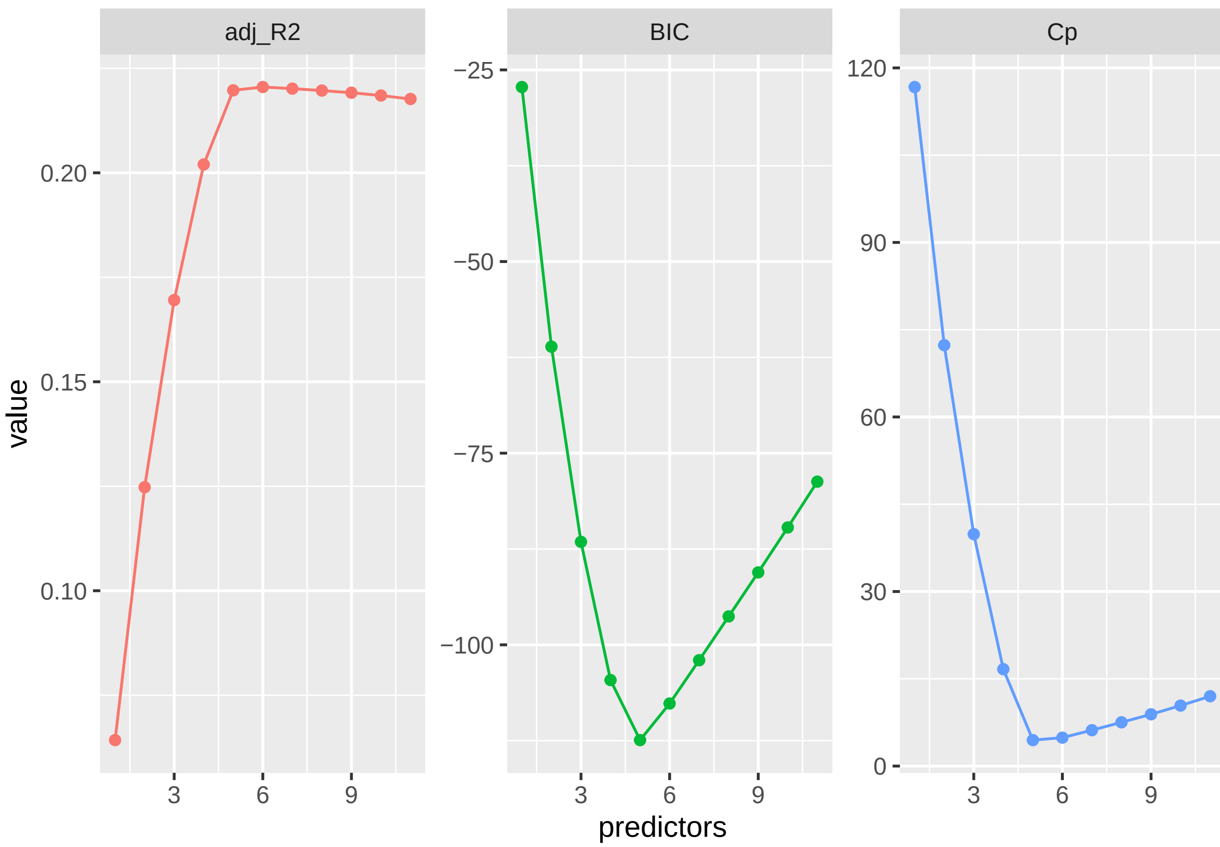


To predict math score, the adjusted R^2 statistic shows that a 7-variable is model is optimal, while the BIC statistic points to a 5-variable model. The C_p suggests a 7-variable model as well.

Reading Score

```
best_subset <- regsubsets(reading_score ~ ., reading_df, nvmax = 11)
results <- summary(best_subset)

tibble(predictors = 1:11,
       adj_R2 = results$adjr2,
       Cp = results$cp,
       BIC = results$bic) %>%
  gather(statistic, value, -predictors) %>%
  ggplot(aes(predictors, value, color = statistic)) +
  geom_line(show.legend = F) +
  geom_point(show.legend = F) +
  facet_wrap(~ statistic, scales = "free")
```



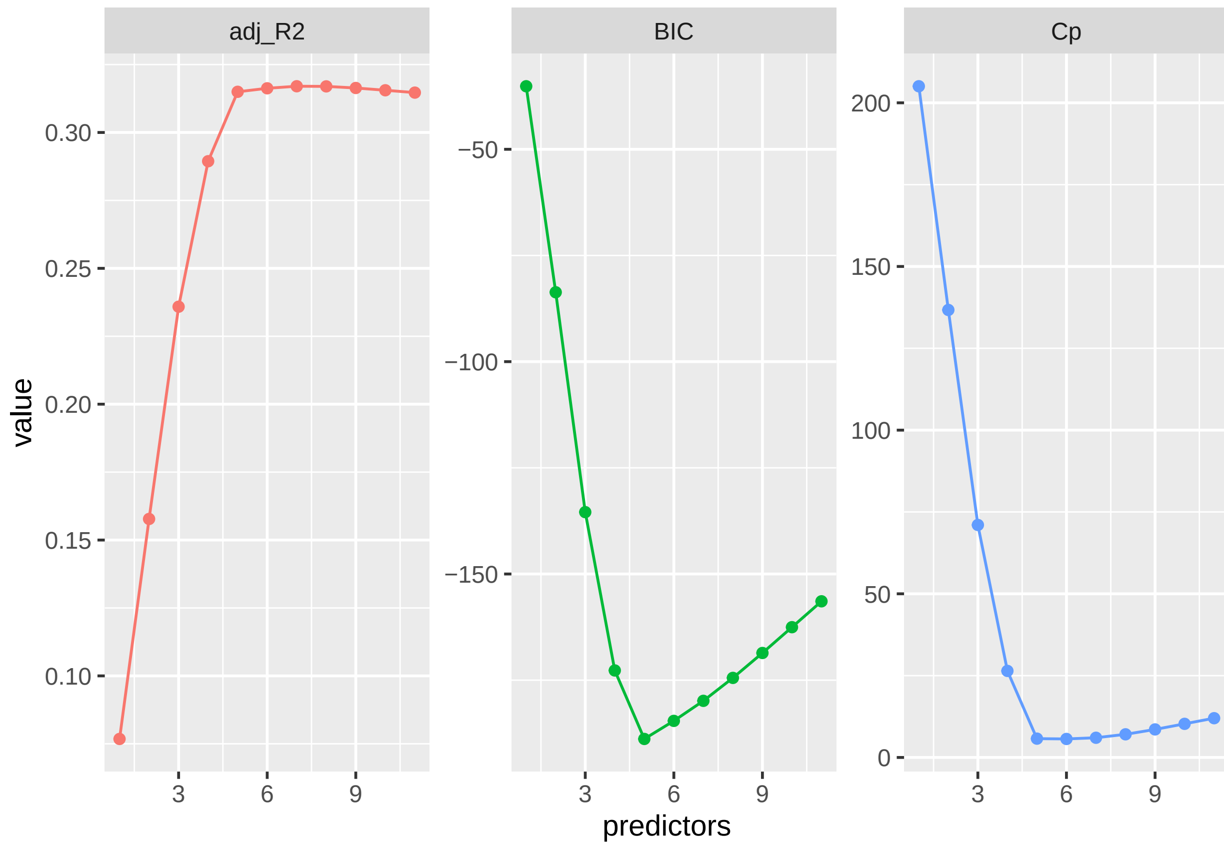
To predict reading score, the adjusted R^2 statistic shows that 6 or 7-variable is model is optimal, while the BIC statistic points to a 5-variable model. The C_p seems to suggest a 6 or 7-variable model as well.

Writing Score

```
best_subset <- regsubsets(writing_score ~ ., writing_df, nvmax = 11)
results <- summary(best_subset)

tibble(predictors = 1:11,
       adj_R2 = results$adjr2,
       Cp = results$cp,
       BIC = results$bic) %>%
```

```r
  gather(statistic, value, -predictors) %>%
  ggplot(aes(predictors, value, color = statistic)) +
  geom_line(show.legend = F) +
  geom_point(show.legend = F) +
  facet_wrap(~ statistic, scales = "free")
```



To predict writing score, the adjusted R^2 statistic shows that a 7 or 8-variable is model is optimal, while the BIC statistic points to a 5-variable model. The C_p suggests a 7-variable model as well.

LASSO approach -

When lambda = 5, the model will tend to have fewer predictors due to the larger penalty. The number of predictors present in the model will increase as lambda decreases; lambda = 1 tends to have about half of the total predictors (~ 6-7) and lambda = 0.1 typically contains all of the available predictors.

Math score (3):

```r
# fit a LASSO with lambda = 5
fit_5 <- glmnet(as.matrix(dplyr::select(math_df, 1:11)), math_df$math_score,
                lambda = 5)
coef(fit_5)
```

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                              s0
## (Intercept)           62.636459
```

```
## gender                 .
## ethnic_group           .
## parent_educ            .
## lunch_type             2.449792
## test_prep              .
## parent_marital_status  .
## practice_sport         .
## is_first_child         .
## nr_siblings            .
## transport_means        .
## wkly_study_hours       .
```

```r
# fit a LASSO with lambda = 1
fit_1 <- glmnet(as.matrix(dplyr::select(math_df, 1:11)), math_df$math_score,
                lambda = 1)
coef(fit_1)
```

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                                s0
## (Intercept)           39.75179929
## gender                 3.33577736
## ethnic_group           1.98807803
## parent_educ            0.79935701
## lunch_type            10.58669309
## test_prep             -3.47963577
## parent_marital_status  .
## practice_sport         .
## is_first_child         .
## nr_siblings            0.02818102
## transport_means        .
## wkly_study_hours       0.78384148
```

```r
# fit a LASSO with lambda = 0.1
fit_0.1 <- glmnet(as.matrix(dplyr::select(math_df, 1:11)), math_df$math_score,
                  lambda = 0.1)
coef(fit_0.1)
```

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                                s0
## (Intercept)           25.76418765
## gender                 5.12095024
## ethnic_group           2.67800923
## parent_educ            1.45498040
## lunch_type            12.41224863
## test_prep             -5.05068605
## parent_marital_status  0.57160682
## practice_sport         0.47516351
## is_first_child         0.52820267
## nr_siblings            0.62819618
## transport_means        0.04781629
## wkly_study_hours       1.94664266
```

The LASSO model fitted with $\lambda = 5$ reduced all of the predictors' coefficients to zero, except for lunch type which had a coefficient of 2.45. The model fitted with $\lambda = 1$ selected for gender, ethnic group, parent education level, lunch type, test prep, number of siblings, and weekly study hours. The $\lambda = 0.1$ model maintains coefficient values similar in range to those of $\lambda = 1$ model and the corresponding step-wise-AIC model above.

Reading score (3):

```
# fit a LASSO with lambda = 5
fit_5 <- glmnet(as.matrix(dplyr::select(reading_df, 1:11)),
                reading_df$reading_score, lambda = 5)
coef(fit_5)
```

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                              s0
## (Intercept)           69.84668
## gender                 0.00000
## ethnic_group           .
## parent_educ            .
## lunch_type             .
## test_prep              .
## parent_marital_status  .
## practice_sport         .
## is_first_child         .
## nr_siblings            .
## transport_means        .
## wkly_study_hours       .
```

```
# fit a LASSO with lambda = 1
fit_1 <- glmnet(as.matrix(dplyr::select(reading_df, 1:11)),
                reading_df$reading_score, lambda = 1)
coef(fit_1)
```

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                               s0
## (Intercept)           67.856478
## gender                -5.413114
## ethnic_group           1.038989
## parent_educ            1.156891
## lunch_type             6.436282
## test_prep             -4.576598
## parent_marital_status  .
## practice_sport         .
## is_first_child         .
## nr_siblings            .
## transport_means        .
## wkly_study_hours       .
```

```
# fit a LASSO with lambda = 0.1
fit_0.1 <- glmnet(as.matrix(dplyr::select(reading_df, 1:11)),
                  reading_df$reading_score, lambda = 0.1)
coef(fit_0.1)
```

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                                 s0
```

```
## (Intercept)          62.6817128
## gender               -7.3115878
## ethnic_group          1.7361352
## parent_educ           1.6939077
## lunch_type            8.4122905
## test_prep            -6.4563567
## parent_marital_status  0.3512184
## practice_sport       -0.5579691
## is_first_child        0.6891704
## nr_siblings           0.2531866
## transport_means       0.6650453
## wkly_study_hours      0.8782682
```

The LASSO model fitted with $\lambda = 5$ reduced all of the predictors' coefficients to zero. The model fitted with $\lambda = 1$ selected for gender, ethnic group, parent education level, lunch type, and test prep. The $\lambda = 0.1$ model maintains coefficient values similar in range to those of $\lambda = 1$ model and the corresponding step-wise-AIC model above.

Writing score (3):

```
# fit a LASSO with lambda = 5
fit_5 <- glmnet(as.matrix(writing_df[1:11]),
                writing_df$writing_score, lambda = 5)
coef(fit_5)
```

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                             s0
## (Intercept)          68.90119
## gender                0.00000
## ethnic_group          .
## parent_educ           .
## lunch_type            .
## test_prep             .
## parent_marital_status .
## practice_sport        .
## is_first_child        .
## nr_siblings           .
## transport_means       .
## wkly_study_hours      .
```

```
# fit a LASSO with lambda = 1
fit_1 <- glmnet(as.matrix(dplyr::select(writing_df, 1:11)),
                writing_df$writing_score, lambda = 1)
coef(fit_1)
```

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                               s0
## (Intercept)          68.901800
## gender               -7.029616
## ethnic_group          1.421530
## parent_educ           1.702391
## lunch_type            7.271135
## test_prep            -6.935167
```

```
## parent_marital_status  .
## practice_sport         .
## is_first_child         .
## nr_siblings            .
## transport_means        .
## wkly_study_hours       .
```

```r
# fit a LASSO with lambda = 0.1
fit_0.1 <- glmnet(as.matrix(dplyr::select(writing_df, 1:11)),
                  writing_df$writing_score, lambda = 0.1)
coef(fit_0.1)
```

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                               s0
## (Intercept)           61.1775776
## gender                -8.9038811
## ethnic_group           2.1242331
## parent_educ            2.2655357
## lunch_type             9.2972877
## test_prep             -8.7621667
## parent_marital_status  0.6208785
## practice_sport         0.3081889
## is_first_child         0.3472107
## nr_siblings            0.3959246
## transport_means        0.5503885
## wkly_study_hours       0.9675351
```

The LASSO model fitted with $\lambda = 5$ reduced all of the predictors' coefficients to zero. The model fitted with $\lambda = 1$ selected for gender, ethnic group, parent education level, lunch type, and test prep. The $\lambda = 0.1$ model maintains coefficient values similar in range to those of $\lambda = 1$ model and the corresponding step-wise-AIC model above.