biostats_final_combined

Miao Fu

2023-12-07

summary statistics

```
# read datafile
df = read_csv("data/Project_1_data.csv") |>
  janitor::clean_names() |>
  mutate(
    wkly_study_hours = ifelse(
      wkly_study_hours == "10-May", "5-10", wkly_study_hours)
  )|>
  na.omit()

# Transforming categorical variables to factors
df_transformed <- df |>
  mutate(
    gender = as.factor(gender),
    ethnic_group = as.factor(ethnic_group),
    parent_educ = factor(parent_educ,
                         levels= c("some high school", "high school", "associate's degree", "some colleg
    lunch_type = as.factor(lunch_type),
    test_prep = as.factor(test_prep),
    parent_marital_status = as.factor(parent_marital_status),
    practice_sport = factor(practice_sport,
                            levels = c("never", "sometimes", "regularly")),
    is_first_child = factor(is_first_child),
    transport_means = as.factor(transport_means),
    wkly_study_hours = factor(wkly_study_hours,
                              levels = c("< 5", "5-10", "> 10"))
  )

# converting categorical variable to numeric variables
df_num=df|>
  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")))
)
```

Categorical Variables

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

Numeric Variables

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

Histograms of all variables

```
png("normality_check.png", width = 1200, height = 800)
par(mfrow=c(3,5))
barplot(table(df_transformed$math_score), main='Maths Score')
barplot(table(df_transformed$writing_score), main='Writing Score')
barplot(table(df_transformed$reading_score), main='Reading Score')
barplot(table(df_transformed$gender), main='Gender')
barplot(table(df_transformed$ethnic_group), main='Ethnic Group')
barplot(table(df_transformed$lunch_type), main='Lunch Type')
barplot(table(df_transformed$test_prep), main='Test Prep')
barplot(table(df_transformed$parent_educ), main='Parent Education')
barplot(table(df_transformed$parent_marital_status), main='Parent Marital Status')
barplot(table(df_transformed$practice_sport), main='Practice Sports')
barplot(table(df_transformed$is_first_child), main='First Child')
barplot(table(df_transformed$nr_siblings), main='Siblings')
barplot(table(df_transformed$transport_means), main='Transport Means')
barplot(table(df_transformed$wkly_study_hours), main='Weekly Study Hours')
dev.off()
```

```
## pdf
##   2
```

Test the transformation for outcome variables

```
# Log, Sqrt, and Inverse transformation of outcomes
df_eda=df|>
  dplyr::select(math_score,writing_score,reading_score)|>
  mutate(
    lgMath=log(math_score),
    sqMath=sqrt(math_score),
    inMath=1/(math_score),
    lgRead=log(reading_score),
    sqRead=sqrt(reading_score),
    inRead=1/(reading_score),
    lgWrite=log(writing_score),
    sqWrite=sqrt(writing_score),
    inWrite=1/(writing_score),
  )
png("transformation_check.png", width = 1200, height = 800)
par(mfrow=c(3,3))
hist(df_eda$lgMath, main="Log(Maths Score)",xlab="Score")
hist(df_eda$sqMath, main="sq(Maths Score)",xlab="Score")
hist(df_eda$inMath, main="in(Maths Score)",xlab="Score")
hist(df_eda$lgRead, main="Log(Reading Score)",xlab="Score")
hist(df_eda$sqRead, main="sq(Reading Score)",xlab="Score")
hist(df_eda$inRead, main="in(Reading Score)",xlab="Score")
hist(df_eda$lgWrite, main="Log(Writing Score)",xlab="Score")
hist(df_eda$sqWrite, main="sq(Writing Score)",xlab="Score")
hist(df_eda$inWrite, main="in(Writing Score)",xlab="Score")
dev.off()
```

```
## pdf
##   2
```

No transformations improved the distribution. Original data were used.

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

```
## pdf
##   2
```

MLR lm()

```
# Build the MLR model for Math scores
model_math <- lm(math_score ~ gender + ethnic_group + parent_educ + lunch_type + test_prep + parent_mari
model_read <- lm(reading_score ~ gender + ethnic_group + parent_educ + lunch_type + test_prep + parent_m
model_write <- lm(writing_score ~ gender + ethnic_group + parent_educ + lunch_type + test_prep + parent_
```

MLR - Math

```
summary(model_math)
```

```
##
## Call:
## lm(formula = math_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep + parent_marital_status + practice_sport +
##     is_first_child + nr_siblings + transport_means + wkly_study_hours,
##     data = df_transformed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.916  -9.265   0.725  10.104  33.013
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      44.1006     3.6704  12.015  < 2e-16 ***
## gendermale                        5.0855     1.1386   4.467 9.61e-06 ***
## ethnic_groupgroup B              -0.1788     2.3136  -0.077  0.93841
## ethnic_groupgroup C              -0.2089     2.2149  -0.094  0.92489
## ethnic_groupgroup D               3.6247     2.2286   1.626  0.10441
## ethnic_groupgroup E              11.1752     2.4434   4.574 5.90e-06 ***
## parent_educhigh school           -0.3235     1.8015  -0.180  0.85757
## parent_educassociate's degree     4.9058     1.7728   2.767  0.00584 **
## parent_educsome college           3.1933     1.8163   1.758  0.07927 .
## parent_educbachelor's degree      6.6652     2.0763   3.210  0.00140 **
## parent_educmaster's degree        6.8096     2.5417   2.679  0.00760 **
## lunch_typestandard               12.3539     1.1771  10.495  < 2e-16 ***
## test_prepnone                    -4.7717     1.2007  -3.974 7.99e-05 ***
## parent_marital_statusmarried      5.4805     1.6170   3.389  0.00075 ***
```

```
## parent_marital_statussingle     2.1682     1.8454   1.175  0.24053
## parent_marital_statuswidowed    7.7944     3.8119   2.045  0.04134 *
## practice_sportsometimes         1.5255     1.8439   0.827  0.40838
## practice_sportregularly         1.6701     1.9046   0.877  0.38092
## is_first_childyes               1.1303     1.2125   0.932  0.35162
## nr_siblings                     0.7403     0.3844   1.926  0.05461 .
## transport_meansschool_bus      -0.4319     1.1629  -0.371  0.71050
## wkly_study_hours5-10            3.5394     1.3429   2.636  0.00863 **
## wkly_study_hours> 10            3.0384     1.7540   1.732  0.08378 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.52 on 564 degrees of freedom
## Multiple R-squared:  0.3221, Adjusted R-squared:  0.2956
## F-statistic: 12.18 on 22 and 564 DF,  p-value: < 2.2e-16
```

Coefficients and Significance Levels:

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

Residuals:

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

Model Fit:

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

R-squared:

- Multiple R-squared (0.3221): About 32.21% of the variability in math_score is explained by the model.

- Adjusted R-squared (0.2956): Adjusts the R-squared for the number of predictors, a better measure for models with multiple predictors.

Statistic & p-value

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

MLR - reading

```
summary(model_read)
```

```
##
## Call:
## lm(formula = reading_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep + parent_marital_status + practice_sport +
##     is_first_child + nr_siblings + transport_means + wkly_study_hours,
##     data = df_transformed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.754  -8.793   0.635   9.118  30.513
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     60.8028     3.5826  16.972  < 2e-16 ***
## gendermale                      -7.6725     1.1114  -6.904 1.37e-11 ***
## ethnic_groupgroup B             -1.4287     2.2582  -0.633 0.527220
## ethnic_groupgroup C             -0.8558     2.1619  -0.396 0.692355
## ethnic_groupgroup D              2.5663     2.1753   1.180 0.238600
## ethnic_groupgroup E              5.9165     2.3850   2.481 0.013402 *
## parent_educhigh school          -0.5785     1.7584  -0.329 0.742303
## parent_educassociate's degree    4.7948     1.7305   2.771 0.005776 **
## parent_educsome college          2.4082     1.7729   1.358 0.174896
## parent_educbachelor's degree     7.3496     2.0266   3.627 0.000313 ***
## parent_educmaster's degree       8.7149     2.4809   3.513 0.000479 ***
## lunch_typestandard               8.4374     1.1489   7.344 7.31e-13 ***
## test_prepnone                   -6.2822     1.1720  -5.360 1.21e-07 ***
## parent_marital_statusmarried     5.2439     1.5783   3.322 0.000950 ***
## parent_marital_statussingle      1.9235     1.8013   1.068 0.286046
## parent_marital_statuswidowed     5.5863     3.7208   1.501 0.133813
## practice_sportsometimes          0.6757     1.7998   0.375 0.707488
## practice_sportregularly         -0.6843     1.8590  -0.368 0.712923
## is_first_childyes                1.3046     1.1835   1.102 0.270780
## nr_siblings                      0.3882     0.3752   1.035 0.301309
## transport_meansschool_bus        0.2841     1.1351   0.250 0.802472
## wkly_study_hours5-10             2.6835     1.3108   2.047 0.041104 *
## wkly_study_hours> 10             1.0970     1.7121   0.641 0.521971
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 13.2 on 564 degrees of freedom
## Multiple R-squared:  0.2709, Adjusted R-squared:  0.2425
## F-statistic: 9.527 on 22 and 564 DF,  p-value: < 2.2e-16
```

Coefficients and Significance Levels:

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

Residuals:

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

Model Fit:

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

R-squared:

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

Statistic & p-value

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

MLR - writing

```
summary(model_write)
```

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

Coefficients and Significance Levels:

- **Intercept (57.808758)**: The expected value of `writing_score` when all other predictors are at their reference level or zero.

- **gendermale (-9.268845, p < 0.001)**: Being male is associated with an average decrease of 9.268845 points in `writing_score` compared to females, holding all else constant. This is statistically significant.
- **ethnic_group**: `ethnic_groupgroup D` (5.010576, p = 0.016531) and `ethnic_groupgroup E` (6.018419, p = 0.008673) are significant, suggesting students in these groups score higher in writing compared to the reference group.
- **parent_educ**: `associate's degree` (6.130783, p = 0.000239), `some college` (4.338798, p = 0.010898), `bachelor's degree` (9.217680, p = 2.62e-06), and `master's degree` (11.712279, p = 1.10e-06) are significant and associated with higher writing scores compared to the reference category.
- **lunch_typestandard (9.390698, p < 0.001)**: Students with standard lunch type score significantly higher.
- **test_prepnone (-8.754351, p < 0.001)**: Not participating in test preparation is associated with lower writing scores.
- **parent_marital_statusmarried (5.246610, p = 0.000561)**: Children of married parents score higher.
- **practice_sport**: Not significant.
- **is_first_childyes**: Not significant.
- **nr_siblings (0.546033, p = 0.129340)**: No significant association with writing scores.
- **transport_meansschool_bus**: Not significant.
- **wkly_study_hours**: Studying 5-10 hours (2.802323, p = 0.026048) shows a significant positive effect.

Residuals:

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

Model Fit:

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

R-squared:

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

Statistic & p-value

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

Cleaned datasets - updated by Nisha

```
set.seed(555)

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.integer(factor(gender)),
    ethnic_group = as.integer(factor(ethnic_group)),
    parent_educ = as.integer(factor(
      parent_educ,levels= c("some high school", "high school",
                            "associate's degree", "some college",
                            "bachelor's degree", "master's degree"))),
    lunch_type = as.integer((factor(lunch_type))),
    test_prep = as.integer((factor(test_prep))),
    parent_marital_status = as.integer((factor(parent_marital_status))),
    practice_sport = as.integer((factor(practice_sport, levels = c("never", "sometimes", "regularly")))
    is_first_child = as.integer((factor(is_first_child))),
    transport_means = as.integer((factor(transport_means))),
     wkly_study_hours = as.integer((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: Backwards Elimination

Math Score

```
mult.fit = lm(math_score ~ ., data = math_df)
summary(mult.fit)
```

```
##
## Call:
## lm(formula = math_score ~ ., data = math_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.864  -9.425   0.975  10.116  32.369
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            23.0924     5.9271   3.896 0.000109 ***
## gender                  5.3175     1.1517   4.617 4.80e-06 ***
## ethnic_group            2.7588     0.4910   5.618 3.00e-08 ***
## parent_educ             1.5290     0.3843   3.979 7.81e-05 ***
## lunch_type             12.6192     1.2000  10.516  < 2e-16 ***
## test_prep              -5.2239     1.2078  -4.325 1.80e-05 ***
## parent_marital_status   0.7239     0.8331   0.869 0.385282
## practice_sport          0.6461     0.8845   0.730 0.465418
## is_first_child          0.7792     1.2313   0.633 0.527061
```

```
## nr_siblings             0.6981     0.3884   1.797 0.072825 .
## transport_means         0.2551     1.1745   0.217 0.828116
## wkly_study_hours        2.0684     0.8749   2.364 0.018402 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.82 on 575 degrees of freedom
## Multiple R-squared:  0.2779, Adjusted R-squared:  0.2641
## F-statistic: 20.11 on 11 and 575 DF,  p-value: < 2.2e-16
```

```
# No Transport Means
step1 = update(mult.fit, . ~ . -transport_means)
summary(step1)
```

```
##
## Call:
## lm(formula = math_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep + parent_marital_status + practice_sport +
##     is_first_child + nr_siblings + wkly_study_hours, data = math_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.777  -9.369   1.069  10.160  32.206
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             23.4423     5.6992   4.113 4.47e-05 ***
## gender                   5.3217     1.1506   4.625 4.62e-06 ***
## ethnic_group             2.7555     0.4904   5.619 2.99e-08 ***
## parent_educ              1.5300     0.3839   3.985 7.61e-05 ***
## lunch_type              12.6219     1.1990  10.527  < 2e-16 ***
## test_prep               -5.2041     1.2033  -4.325 1.80e-05 ***
## parent_marital_status    0.7256     0.8324   0.872   0.3837
## practice_sport           0.6470     0.8837   0.732   0.4644
## is_first_child           0.7826     1.2301   0.636   0.5249
## nr_siblings              0.6981     0.3881   1.799   0.0726 .
## wkly_study_hours         2.0752     0.8736   2.376   0.0178 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.81 on 576 degrees of freedom
## Multiple R-squared:  0.2778, Adjusted R-squared:  0.2653
## F-statistic: 22.16 on 10 and 576 DF,  p-value: < 2.2e-16
```

```
# No Is First Child
step2 = update(step1, . ~ . -is_first_child)
summary(step2)
```

```
##
## Call:
## lm(formula = math_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep + parent_marital_status + practice_sport +
##     nr_siblings + wkly_study_hours, data = math_df)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.413  -9.258   0.787   9.904  32.464
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            25.0419     5.1120   4.899 1.25e-06 ***
## gender                  5.3263     1.1500   4.632 4.48e-06 ***
## ethnic_group            2.7538     0.4901   5.619 3.00e-08 ***
## parent_educ             1.5308     0.3837   3.989 7.48e-05 ***
## lunch_type             12.6288     1.1983  10.539  < 2e-16 ***
## test_prep              -5.2506     1.2005  -4.374 1.45e-05 ***
## parent_marital_status   0.6824     0.8292   0.823   0.4109
## practice_sport          0.6017     0.8804   0.683   0.4946
## nr_siblings             0.6776     0.3866   1.753   0.0802 .
## wkly_study_hours        2.0800     0.8731   2.382   0.0175 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.81 on 577 degrees of freedom
## Multiple R-squared:  0.2773, Adjusted R-squared:  0.266
## F-statistic:  24.6 on 9 and 577 DF,  p-value: < 2.2e-16
```

```r
# No Practice Sport
step3 = update(step2, . ~ . -practice_sport)
summary(step3)
```

```
##
## Call:
## lm(formula = math_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep + parent_marital_status + nr_siblings +
##     wkly_study_hours, data = math_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.042  -9.171   0.792  10.144  32.974
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            26.4222     4.6939   5.629 2.83e-08 ***
## gender                  5.3307     1.1494   4.638 4.36e-06 ***
## ethnic_group            2.7553     0.4899   5.624 2.90e-08 ***
## parent_educ             1.5138     0.3828   3.955 8.60e-05 ***
## lunch_type             12.6050     1.1973  10.528  < 2e-16 ***
## test_prep              -5.2588     1.1999  -4.383 1.39e-05 ***
## parent_marital_status   0.7058     0.8281   0.852   0.3944
## nr_siblings             0.6790     0.3864   1.757   0.0794 .
## wkly_study_hours        2.0891     0.8726   2.394   0.0170 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.8 on 578 degrees of freedom
## Multiple R-squared:  0.2767, Adjusted R-squared:  0.2667
```

```
## F-statistic: 27.64 on 8 and 578 DF,  p-value: < 2.2e-16
```

```
# No Parent Marital Status
step4 = update(step3, . ~ . -parent_marital_status)
summary(step4)
```

```
##
## 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
```

```
# No Number of Siblings
math_backward_manual_fit = update(step4, . ~ . -nr_siblings)
summary(math_backward_manual_fit)
```

```
##
## Call:
## lm(formula = math_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep + wkly_study_hours, data = math_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.943  -9.439   0.630  10.403  31.459
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       29.7605     4.1787   7.122 3.16e-12 ***
## gender             5.2204     1.1499   4.540 6.85e-06 ***
## ethnic_group       2.7208     0.4904   5.549 4.39e-08 ***
## parent_educ        1.5128     0.3833   3.947 8.88e-05 ***
## lunch_type        12.5868     1.1987  10.501  < 2e-16 ***
```

```
## test_prep         -5.3895     1.2000  -4.491 8.55e-06 ***
## wkly_study_hours   2.1599     0.8729   2.474   0.0136 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.82 on 580 degrees of freedom
## Multiple R-squared:  0.2718, Adjusted R-squared:  0.2642
## F-statistic: 36.08 on 6 and 580 DF,  p-value: < 2.2e-16
```

```
mean(math_backward_manual_fit$residuals^2)
```

```
## [1] 188.763
```

```
# just use one function
math_backward_func_fit = step(mult.fit, direction='backward')
```

```
## Start:  AIC=3095.24
## 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
##
##                         Df Sum of Sq    RSS    AIC
## - transport_means        1       9.0 109886 3093.3
## - is_first_child         1      76.5 109954 3093.6
## - practice_sport         1     102.0 109979 3093.8
## - parent_marital_status  1     144.3 110022 3094.0
## <none>                               109877 3095.2
## - nr_siblings            1     617.2 110495 3096.5
## - wkly_study_hours       1    1068.1 110945 3098.9
## - parent_educ            1    3025.1 112902 3109.2
## - test_prep              1    3574.7 113452 3112.0
## - gender                 1    4073.6 113951 3114.6
## - ethnic_group           1    6032.2 115910 3124.6
## - lunch_type             1   21130.5 131008 3196.5
##
## Step:  AIC=3093.29
## math_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + parent_marital_status + practice_sport + is_first_child +
##     nr_siblings + wkly_study_hours
##
##                         Df Sum of Sq    RSS    AIC
## - is_first_child         1      77.2 109964 3091.7
## - practice_sport         1     102.3 109989 3091.8
## - parent_marital_status  1     145.0 110031 3092.1
## <none>                               109886 3093.3
## - nr_siblings            1     617.2 110504 3094.6
## - wkly_study_hours       1    1076.6 110963 3097.0
## - parent_educ            1    3029.8 112916 3107.3
## - test_prep              1    3568.0 113454 3110.0
## - gender                 1    4081.2 113968 3112.7
## - ethnic_group           1    6023.6 115910 3122.6
## - lunch_type             1   21141.9 131028 3194.6
##
```

```
## Step:  AIC=3091.7
## math_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + parent_marital_status + practice_sport + nr_siblings +
##     wkly_study_hours
##
##                          Df Sum of Sq    RSS    AIC
## - practice_sport          1      89.0 110053 3090.2
## - parent_marital_status   1     129.1 110093 3090.4
## <none>                                109964 3091.7
## - nr_siblings             1     585.5 110549 3092.8
## - wkly_study_hours        1    1081.6 111045 3095.4
## - parent_educ             1    3032.9 112996 3105.7
## - test_prep               1    3645.7 113609 3108.8
## - gender                  1    4088.5 114052 3111.1
## - ethnic_group            1    6016.6 115980 3121.0
## - lunch_type              1   21166.9 131130 3193.0
##
## Step:  AIC=3090.18
## math_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + parent_marital_status + nr_siblings + wkly_study_hours
##
##                          Df Sum of Sq    RSS    AIC
## - parent_marital_status   1     138.3 110191 3088.9
## <none>                                110053 3090.2
## - nr_siblings             1     587.9 110640 3091.3
## - wkly_study_hours        1    1091.4 111144 3094.0
## - parent_educ             1    2978.5 113031 3103.9
## - test_prep               1    3657.3 113710 3107.4
## - gender                  1    4095.3 114148 3109.6
## - ethnic_group            1    6022.9 116075 3119.5
## - lunch_type              1   21105.0 131158 3191.2
##
## Step:  AIC=3088.91
## math_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + nr_siblings + wkly_study_hours
##
##                     Df Sum of Sq    RSS    AIC
## <none>                           110191 3088.9
## - nr_siblings        1     613.0 110804 3090.2
## - wkly_study_hours   1    1084.6 111275 3092.7
## - parent_educ        1    3008.2 113199 3102.7
## - test_prep          1    3708.6 113900 3106.3
## - gender             1    4054.4 114245 3108.1
## - ethnic_group       1    5977.9 116169 3117.9
## - lunch_type         1   21020.1 131211 3189.4
```

```
summary(math_backward_func_fit)
```

```
##
## 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
```

```
mean(math_backward_manual_fit$residuals^2)
```

```
## [1] 188.763
```

With manual elimination, the model we obtained was Math Score ~ Gender + Ethnic Group + Parent Education + Lunch Type + Test Prep + Weekly Study Hours.

When using the single-function method, the model obtained with the lowest AIC was Math Score ~ Gender + Ethnic Group + Parent Education + Lunch Type + Test Prep + Number of Siblings + Weekly Study Hours. Both models' MSE are equal to each other, while the manually derived model had alower adjusted R-squared value by ~ 0.3 units.

Reading Score

```
mult.fit = lm(reading_score ~ ., data = reading_df)
summary(mult.fit)
```

```
##
## Call:
## lm(formula = reading_score ~ ., data = reading_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.401  -9.051   0.404   9.807  33.637
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           61.3874     5.7518  10.673  < 2e-16 ***
## gender                -7.5190     1.1176  -6.728 4.18e-11 ***
## ethnic_group           1.8185     0.4765   3.816  0.00015 ***
## parent_educ            1.7505     0.3729   4.694 3.35e-06 ***
## lunch_type             8.6295     1.1646   7.410 4.53e-13 ***
```

```
## test_prep              -6.6530     1.1721  -5.676 2.19e-08 ***
## parent_marital_status   0.5050     0.8085   0.625  0.53247
## practice_sport         -0.6974     0.8583  -0.813  0.41684
## is_first_child          0.9079     1.1949   0.760  0.44768
## nr_siblings             0.3178     0.3770   0.843  0.39960
## transport_means         0.8813     1.1398   0.773  0.43969
## wkly_study_hours        1.0163     0.8490   1.197  0.23180
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.41 on 575 degrees of freedom
## Multiple R-squared:  0.2323, Adjusted R-squared:  0.2177
## F-statistic: 15.82 on 11 and 575 DF,  p-value: < 2.2e-16
```

```
# No Parent Marital Status
step1 = update(mult.fit, . ~ . -parent_marital_status)
summary(step1)
```

```
##
## Call:
## lm(formula = reading_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep + practice_sport + is_first_child +
##     nr_siblings + transport_means + wkly_study_hours, data = reading_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.962  -9.031   0.340   9.774  33.510
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       62.6218     5.3988  11.599  < 2e-16 ***
## gender            -7.5395     1.1166  -6.752 3.56e-11 ***
## ethnic_group       1.8103     0.4761   3.803 0.000158 ***
## parent_educ        1.7562     0.3726   4.713 3.06e-06 ***
## lunch_type         8.6086     1.1635   7.399 4.88e-13 ***
## test_prep         -6.6809     1.1706  -5.707 1.84e-08 ***
## practice_sport    -0.6789     0.8574  -0.792 0.428771
## is_first_child     0.8467     1.1902   0.711 0.477108
## nr_siblings        0.3259     0.3765   0.865 0.387178
## transport_means    0.8883     1.1391   0.780 0.435820
## wkly_study_hours   1.0114     0.8485   1.192 0.233773
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.41 on 576 degrees of freedom
## Multiple R-squared:  0.2318, Adjusted R-squared:  0.2185
## F-statistic: 17.38 on 10 and 576 DF,  p-value: < 2.2e-16
```

```
# No Is First Child
step2 = update(step1, . ~ . -is_first_child)
summary(step2)
```

```
##
```

```
## Call:
## lm(formula = reading_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep + practice_sport + nr_siblings + transport_means +
##     wkly_study_hours, data = reading_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.609  -8.970   0.378   9.579  32.976
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       64.2344     4.8978  13.115  < 2e-16 ***
## gender            -7.5327     1.1161  -6.749 3.62e-11 ***
## ethnic_group       1.8094     0.4759   3.802 0.000159 ***
## parent_educ        1.7565     0.3725   4.716 3.02e-06 ***
## lunch_type         8.6180     1.1629   7.411 4.49e-13 ***
## test_prep         -6.7298     1.1681  -5.761 1.36e-08 ***
## practice_sport    -0.7300     0.8540  -0.855 0.393012
## nr_siblings        0.3027     0.3750   0.807 0.419780
## transport_means    0.8979     1.1386   0.789 0.430634
## wkly_study_hours   1.0168     0.8481   1.199 0.231087
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.4 on 577 degrees of freedom
## Multiple R-squared:  0.2312, Adjusted R-squared:  0.2192
## F-statistic: 19.27 on 9 and 577 DF,  p-value: < 2.2e-16
```

```
# No Transport Means
step3 = update(step2, . ~ . -transport_means)
summary(step3)
```

```
##
## Call:
## lm(formula = reading_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep + practice_sport + nr_siblings + wkly_study_hours,
##     data = reading_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.293  -8.920   0.575   9.576  32.489
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       65.5028     4.6247  14.164  < 2e-16 ***
## gender            -7.5180     1.1155  -6.739 3.86e-11 ***
## ethnic_group       1.7978     0.4755   3.781 0.000172 ***
## parent_educ        1.7603     0.3723   4.728 2.85e-06 ***
## lunch_type         8.6274     1.1625   7.422 4.16e-13 ***
## test_prep         -6.6611     1.1645  -5.720 1.71e-08 ***
## practice_sport    -0.7272     0.8537  -0.852 0.394695
## nr_siblings        0.3025     0.3748   0.807 0.419988
## wkly_study_hours   1.0410     0.8473   1.229 0.219699
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.4 on 578 degrees of freedom
## Multiple R-squared:  0.2303, Adjusted R-squared:  0.2197
## F-statistic: 21.62 on 8 and 578 DF,  p-value: < 2.2e-16
```

```
# No Number of Siblings
step4 = update(step3, . ~ . -nr_siblings)
summary(step4)
```

```
##
## Call:
## lm(formula = reading_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep + practice_sport + wkly_study_hours,
##     data = reading_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.079  -8.940   0.773   9.687  32.757
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       66.2290     4.5350  14.604  < 2e-16 ***
## gender            -7.5535     1.1143  -6.779 3.00e-11 ***
## ethnic_group       1.7877     0.4752   3.762 0.000186 ***
## parent_educ        1.7569     0.3722   4.721 2.95e-06 ***
## lunch_type         8.6333     1.1621   7.429 3.94e-13 ***
## test_prep         -6.7034     1.1629  -5.764 1.33e-08 ***
## practice_sport    -0.7223     0.8534  -0.846 0.397676
## wkly_study_hours   1.0748     0.8460   1.270 0.204455
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.39 on 579 degrees of freedom
## Multiple R-squared:  0.2295, Adjusted R-squared:  0.2201
## F-statistic: 24.63 on 7 and 579 DF,  p-value: < 2.2e-16
```

```
# No Practice Sport
step5 = update(step4, . ~ . -practice_sport)
summary(step5)
```

```
##
## Call:
## lm(formula = reading_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep + wkly_study_hours, data = reading_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.551  -8.822   0.863   9.721  32.252
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       64.5007     4.0483  15.933  < 2e-16 ***
```

```
## gender            -7.5573     1.1140  -6.784 2.90e-11 ***
## ethnic_group       1.7865     0.4751   3.761 0.000187 ***
## parent_educ        1.7770     0.3713   4.786 2.16e-06 ***
## lunch_type         8.6631     1.1613   7.460 3.18e-13 ***
## test_prep         -6.6919     1.1626  -5.756 1.39e-08 ***
## wkly_study_hours   1.0638     0.8457   1.258 0.208928
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.39 on 580 degrees of freedom
## Multiple R-squared:  0.2285, Adjusted R-squared:  0.2205
## F-statistic: 28.63 on 6 and 580 DF,  p-value: < 2.2e-16
```

```
# No Weekly Study Hours
reading_backward_manual_fit = update(step5, . ~ . -wkly_study_hours)
summary(reading_backward_manual_fit)
```

```
##
## 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
```

```
mean(reading_backward_manual_fit$residuals^2)
```

```
## [1] 177.6469
```

```
# just use one function
reading_backward_func_fit = step(mult.fit, direction='backward')
```

```
## Start:  AIC=3060.01
## 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
```

```
##
##                          Df Sum of Sq    RSS    AIC
## - parent_marital_status  1      70.2 103547 3058.4
## - is_first_child         1     103.9 103581 3058.6
## - transport_means        1     107.6 103584 3058.6
## - practice_sport         1     118.8 103595 3058.7
## - nr_siblings            1     127.9 103605 3058.7
## - wkly_study_hours       1     257.8 103734 3059.5
## <none>                               103477 3060.0
## - ethnic_group           1    2621.0 106098 3072.7
## - parent_educ            1    3965.2 107442 3080.1
## - test_prep              1    5798.3 109275 3090.0
## - gender                 1    8145.0 111622 3102.5
## - lunch_type             1    9881.5 113358 3111.6
##
## Step:  AIC=3058.41
## reading_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + practice_sport + is_first_child + nr_siblings +
##     transport_means + wkly_study_hours
##
##                     Df Sum of Sq    RSS    AIC
## - is_first_child     1      91.0 103638 3056.9
## - transport_means    1     109.3 103656 3057.0
## - practice_sport     1     112.7 103660 3057.1
## - nr_siblings        1     134.6 103681 3057.2
## - wkly_study_hours   1     255.4 103802 3057.8
## <none>                           103547 3058.4
## - ethnic_group       1    2599.5 106146 3071.0
## - parent_educ        1    3993.6 107540 3078.6
## - test_prep          1    5855.5 109402 3088.7
## - gender             1    8196.5 111743 3101.1
## - lunch_type         1    9841.8 113389 3109.7
##
## Step:  AIC=3056.92
## reading_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + practice_sport + nr_siblings + transport_means +
##     wkly_study_hours
##
##                     Df Sum of Sq    RSS    AIC
## - transport_means    1     111.7 103750 3055.6
## - nr_siblings        1     117.1 103755 3055.6
## - practice_sport     1     131.2 103769 3055.7
## - wkly_study_hours   1     258.1 103896 3056.4
## <none>                           103638 3056.9
## - ethnic_group       1    2597.0 106235 3069.4
## - parent_educ        1    3994.8 107633 3077.1
## - test_prep          1    5962.1 109600 3087.8
## - gender             1    8182.2 111820 3099.5
## - lunch_type         1    9864.6 113502 3108.3
##
## Step:  AIC=3055.56
## reading_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + practice_sport + nr_siblings + wkly_study_hours
##
```

```
##                     Df Sum of Sq    RSS    AIC
## - nr_siblings        1     116.9 103866 3054.2
## - practice_sport     1     130.2 103880 3054.3
## - wkly_study_hours   1     271.0 104021 3055.1
## <none>                           103750 3055.6
## - ethnic_group       1    2566.2 106316 3067.9
## - parent_educ        1    4012.9 107762 3075.8
## - test_prep          1    5873.7 109623 3085.9
## - gender             1    8152.7 111902 3098.0
## - lunch_type         1    9887.1 113637 3107.0
##
## Step:  AIC=3054.22
## reading_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + practice_sport + wkly_study_hours
##
##                     Df Sum of Sq    RSS    AIC
## - practice_sport     1     128.5 103995 3052.9
## - wkly_study_hours   1     289.5 104156 3053.8
## <none>                           103866 3054.2
## - ethnic_group       1    2539.1 106406 3066.4
## - parent_educ        1    3997.8 107864 3074.4
## - test_prep          1    5960.5 109827 3085.0
## - gender             1    8242.7 112109 3097.0
## - lunch_type         1    9901.1 113768 3105.7
##
## Step:  AIC=3052.94
## reading_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + wkly_study_hours
##
##                     Df Sum of Sq    RSS    AIC
## - wkly_study_hours   1     283.7 104279 3052.5
## <none>                           103995 3052.9
## - ethnic_group       1    2535.8 106531 3065.1
## - parent_educ        1    4106.5 108102 3073.7
## - test_prep          1    5941.0 109936 3083.6
## - gender             1    8251.2 112246 3095.8
## - lunch_type         1    9978.6 113974 3104.7
##
## Step:  AIC=3052.54
## reading_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep
##
##                 Df Sum of Sq    RSS    AIC
## <none>                       104279 3052.5
## - ethnic_group   1    2554.7 106833 3064.8
## - parent_educ    1    4036.2 108315 3072.8
## - test_prep      1    6241.3 110520 3084.7
## - gender         1    8151.5 112430 3094.7
## - lunch_type     1    9987.0 114266 3104.2
```

```
summary(reading_backward_func_fit)
```

```
##
## 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
```

```
mean(reading_backward_func_fit$residuals^2)
```

```
## [1] 177.6469
```

With manual elimination, the model we obtained was Reading Score ~ Gender + Ethnic Group + Parent Education + Lunch Type + Test Prep.

When using the single-function method, the model obtained with the lowest AIC was Reading Score ~ Gender + Ethnic Group + Parent Education + Lunch Type + Test Prep. The one-function model had equal adjusted R-squared and MSE values.

Writing Score

```
mult.fit = lm(writing_score ~ ., data = writing_df)
summary(mult.fit)
```

```
##
## Call:
## lm(formula = writing_score ~ ., data = writing_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.711  -8.503   0.758   9.459  28.543
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           59.1030     5.5195  10.708  < 2e-16 ***
## gender                -9.1137     1.0725  -8.498  < 2e-16 ***
## ethnic_group           2.2059     0.4572   4.824 1.80e-06 ***
## parent_educ            2.3308     0.3579   6.513 1.61e-10 ***
## lunch_type             9.5265     1.1175   8.525  < 2e-16 ***
```

```
## test_prep              -8.9524     1.1247  -7.960 9.25e-15 ***
## parent_marital_status   0.7645     0.7758   0.985    0.325
## practice_sport          0.4809     0.8237   0.584    0.560
## is_first_child          0.6009     1.1466   0.524    0.600
## nr_siblings             0.4607     0.3617   1.274    0.203
## transport_means         0.7647     1.0937   0.699    0.485
## wkly_study_hours        1.1007     0.8147   1.351    0.177
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.87 on 575 degrees of freedom
## Multiple R-squared:  0.3275, Adjusted R-squared:  0.3147
## F-statistic: 25.46 on 11 and 575 DF,  p-value: < 2.2e-16
```

```
# No Is First Child
step1 = update(mult.fit, . ~ . -is_first_child)
summary(step1)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep + parent_marital_status + practice_sport +
##     nr_siblings + transport_means + wkly_study_hours, data = writing_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.434  -8.291   0.834   9.509  28.702
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             60.3211     5.0031  12.057  < 2e-16 ***
## gender                  -9.1103     1.0718  -8.500  < 2e-16 ***
## ethnic_group             2.2047     0.4570   4.825 1.80e-06 ***
## parent_educ              2.3314     0.3576   6.519 1.55e-10 ***
## lunch_type               9.5317     1.1168   8.535  < 2e-16 ***
## test_prep               -8.9887     1.1219  -8.012 6.28e-15 ***
## parent_marital_status    0.7312     0.7727   0.946    0.344
## practice_sport           0.4462     0.8205   0.544    0.587
## nr_siblings              0.4449     0.3602   1.235    0.217
## transport_means          0.7719     1.0930   0.706    0.480
## wkly_study_hours         1.1041     0.8142   1.356    0.176
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.86 on 576 degrees of freedom
## Multiple R-squared:  0.3272, Adjusted R-squared:  0.3155
## F-statistic: 28.01 on 10 and 576 DF,  p-value: < 2.2e-16
```

```
# No Practice Sport
step2 = update(step1, . ~ . -practice_sport)
summary(step2)
```

```
##
```

```
## Call:
## lm(formula = writing_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep + parent_marital_status + nr_siblings +
##     transport_means + wkly_study_hours, data = writing_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.159  -8.417   0.695   9.645  28.655
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            61.3413     4.6352  13.234  < 2e-16 ***
## gender                 -9.1071     1.0711  -8.502  < 2e-16 ***
## ethnic_group            2.2058     0.4567   4.830 1.75e-06 ***
## parent_educ             2.3188     0.3567   6.501 1.72e-10 ***
## lunch_type              9.5141     1.1156   8.528  < 2e-16 ***
## test_prep              -8.9949     1.1211  -8.023 5.79e-15 ***
## parent_marital_status   0.7486     0.7716   0.970    0.332
## nr_siblings             0.4460     0.3600   1.239    0.216
## transport_means         0.7742     1.0923   0.709    0.479
## wkly_study_hours        1.1108     0.8136   1.365    0.173
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.86 on 577 degrees of freedom
## Multiple R-squared:  0.3269, Adjusted R-squared:  0.3164
## F-statistic: 31.13 on 9 and 577 DF,  p-value: < 2.2e-16
```

```
# No Transport Means
step3 = update(step2, . ~ . -transport_means)
summary(step3)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep + parent_marital_status + nr_siblings +
##     wkly_study_hours, data = writing_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.889  -8.443   0.979   9.527  28.924
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            62.4294     4.3717  14.280  < 2e-16 ***
## gender                 -9.0942     1.0705  -8.495  < 2e-16 ***
## ethnic_group            2.1959     0.4563   4.813 1.90e-06 ***
## parent_educ             2.3220     0.3565   6.514 1.60e-10 ***
## lunch_type              9.5223     1.1151   8.540  < 2e-16 ***
## test_prep              -8.9355     1.1175  -7.996 7.04e-15 ***
## parent_marital_status   0.7535     0.7712   0.977    0.329
## nr_siblings             0.4457     0.3599   1.238    0.216
## wkly_study_hours        1.1318     0.8127   1.393    0.164
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.85 on 578 degrees of freedom
## Multiple R-squared:  0.3263, Adjusted R-squared:  0.317
## F-statistic: 34.99 on 8 and 578 DF,  p-value: < 2.2e-16
```

```
# No Parent Marital Status
step4 = update(step3, . ~ . -parent_marital_status)
summary(step4)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep + nr_siblings + wkly_study_hours,
##     data = writing_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.246  -8.263   0.690   9.167  28.855
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       64.1899     3.9829  16.116  < 2e-16 ***
## gender            -9.1252     1.0700  -8.528  < 2e-16 ***
## ethnic_group       2.1838     0.4561   4.788 2.14e-06 ***
## parent_educ        2.3296     0.3564   6.537 1.38e-10 ***
## lunch_type         9.4888     1.1145   8.514  < 2e-16 ***
## test_prep         -8.9716     1.1169  -8.033 5.35e-15 ***
## nr_siblings        0.4603     0.3595   1.280    0.201
## wkly_study_hours   1.1248     0.8126   1.384    0.167
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.85 on 579 degrees of freedom
## Multiple R-squared:  0.3252, Adjusted R-squared:  0.317
## F-statistic: 39.86 on 7 and 579 DF,  p-value: < 2.2e-16
```

```
# No Number of Siblings
step5 = update(step4, . ~ . -nr_siblings)
summary(step5)
```

```
##
## 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
```

```
# No Weekly Study Hours
writing_backward_manual_fit = update(step5, . ~ . -wkly_study_hours)
summary(writing_backward_manual_fit)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + ethnic_group + parent_educ +
##     lunch_type + test_prep, data = writing_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.594  -8.422   0.710   9.201  29.415
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   67.7575     3.5050  19.332  < 2e-16 ***
## gender        -9.1231     1.0701  -8.526  < 2e-16 ***
## ethnic_group   2.1756     0.4566   4.765 2.39e-06 ***
## parent_educ    2.3061     0.3567   6.466 2.14e-10 ***
## lunch_type     9.5016     1.1161   8.513  < 2e-16 ***
## test_prep     -9.1875     1.1125  -8.258 9.99e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.87 on 581 degrees of freedom
## Multiple R-squared:  0.3208, Adjusted R-squared:  0.315
## F-statistic: 54.89 on 5 and 581 DF,  p-value: < 2.2e-16
```

```
mean(writing_backward_manual_fit$residuals^2)
```

```
## [1] 163.9483
```

```
# just use one function
writing_backward_func_fit = step(mult.fit, direction='backward')
```

```
## Start:  AIC=3011.6
## 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
```

```
##
##                         Df Sum of Sq    RSS    AIC
## - is_first_child         1      45.5  95330 3009.9
## - practice_sport         1      56.5  95341 3009.9
## - transport_means        1      81.0  95366 3010.1
## - parent_marital_status  1     160.9  95446 3010.6
## - nr_siblings            1     268.8  95554 3011.2
## - wkly_study_hours       1     302.5  95587 3011.5
## <none>                                95285 3011.6
## - ethnic_group           1    3856.7  99142 3032.9
## - parent_educ            1    7030.2 102315 3051.4
## - test_prep              1   10498.8 105784 3070.9
## - gender                 1   11966.3 107251 3079.0
## - lunch_type             1   12042.4 107327 3079.5
##
## Step:  AIC=3009.88
## writing_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + parent_marital_status + practice_sport + nr_siblings +
##     transport_means + wkly_study_hours
##
##                         Df Sum of Sq    RSS    AIC
## - practice_sport         1      48.9  95379 3008.2
## - transport_means        1      82.6  95413 3008.4
## - parent_marital_status  1     148.2  95479 3008.8
## - nr_siblings            1     252.5  95583 3009.4
## - wkly_study_hours       1     304.4  95635 3009.8
## <none>                                95330 3009.9
## - ethnic_group           1    3852.7  99183 3031.1
## - parent_educ            1    7033.6 102364 3049.7
## - test_prep              1   10624.5 105955 3069.9
## - gender                 1   11957.7 107288 3077.2
## - lunch_type             1   12056.6 107387 3077.8
##
## Step:  AIC=3008.18
## writing_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + parent_marital_status + nr_siblings + transport_means +
##     wkly_study_hours
##
##                         Df Sum of Sq    RSS    AIC
## - transport_means        1      83.0  95462 3006.7
## - parent_marital_status  1     155.6  95535 3007.1
## - nr_siblings            1     253.7  95633 3007.7
## - wkly_study_hours       1     308.2  95688 3008.1
## <none>                                95379 3008.2
## - ethnic_group           1    3856.6  99236 3029.4
## - parent_educ            1    6987.2 102367 3047.7
## - test_prep              1   10640.3 106020 3068.3
## - gender                 1   11949.7 107329 3075.5
## - lunch_type             1   12022.2 107402 3075.9
##
## Step:  AIC=3006.69
## writing_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + parent_marital_status + nr_siblings + wkly_study_hours
##
```

```
##                         Df Sum of Sq    RSS    AIC
## - parent_marital_status  1     157.7  95620 3005.7
## - nr_siblings            1     253.3  95716 3006.2
## - wkly_study_hours       1     320.3  95783 3006.7
## <none>                                95462 3006.7
## - ethnic_group           1    3825.5  99288 3027.8
## - parent_educ            1    7007.4 102470 3046.3
## - test_prep              1   10559.2 106022 3066.3
## - gender                 1   11919.4 107382 3073.8
## - lunch_type             1   12044.3 107507 3074.4
##
## Step:  AIC=3005.66
## writing_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + nr_siblings + wkly_study_hours
##
##                     Df Sum of Sq    RSS    AIC
## - nr_siblings        1     270.7  95891 3005.3
## - wkly_study_hours   1     316.4  95936 3005.6
## <none>                           95620 3005.7
## - ethnic_group       1    3786.2  99406 3026.4
## - parent_educ        1    7057.0 102677 3045.5
## - test_prep          1   10656.5 106277 3065.7
## - lunch_type         1   11971.1 107591 3072.9
## - gender             1   12011.2 107631 3073.1
##
## Step:  AIC=3005.32
## writing_score ~ gender + ethnic_group + parent_educ + lunch_type +
##     test_prep + wkly_study_hours
##
##                     Df Sum of Sq    RSS    AIC
## <none>                           95891 3005.3
## - wkly_study_hours   1     346.8  96238 3005.4
## - ethnic_group       1    3735.6  99626 3025.8
## - parent_educ        1    7025.2 102916 3044.8
## - test_prep          1   10832.0 106723 3066.1
## - lunch_type         1   11993.5 107884 3072.5
## - gender             1   12172.7 108064 3073.5
```

```
summary(writing_backward_func_fit)
```

```
##
## 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
```

```
mean(writing_backward_func_fit$residuals^2)
```

```
## [1] 163.3575
```

With manual elimination, the model we obtained was Writing Score ~ Gender + Ethnic Group + Parent Education + Lunch Type + Test Prep.

When using the single-function method, the model obtained with the lowest AIC was Writing Score ~ Gender + Ethnic Group + Parent Education + Lunch Type + Test Prep + Weekly Study Hours. Both models had equal adjusted R-squared values and MSEs within 0.6 points of each other.

Step-wise: Forward Elimination

Math Score

```
mult.fit = lm(math_score ~ ., data = math_df)

### Step 1:  Fit simple linear regressions for all variables,look for the variable with lowest p-value
fit1 = lm(math_score ~ gender, data = step_df)
summary(fit1)
```

```
##
## Call:
## lm(formula = math_score ~ gender, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.933 -10.393   0.147  11.107  36.067
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   58.014      2.029  28.596  < 2e-16 ***
## gender         5.920      1.312   4.511  7.8e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.85 on 585 degrees of freedom
## Multiple R-squared:  0.03362,    Adjusted R-squared:  0.03196
## F-statistic: 20.35 on 1 and 585 DF,  p-value: 7.801e-06
```

```r
fit2 = lm(math_score ~ ethnic_group, data = step_df)
summary(fit2)
```

```
##
## Call:
## lm(formula = math_score ~ ethnic_group, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.134  -9.256   0.744  10.805  39.112
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   56.7656     1.8778  30.230  < 2e-16 ***
## ethnic_group   3.1227     0.5553   5.624  2.9e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.71 on 585 degrees of freedom
## Multiple R-squared:  0.05129,    Adjusted R-squared:  0.04967
## F-statistic: 31.62 on 1 and 585 DF,  p-value: 2.9e-08
```

```r
fit3 = lm(math_score ~ parent_educ, data = step_df)
summary(fit3)
```

```
##
## Call:
## lm(formula = math_score ~ parent_educ, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.753  -9.889   0.823  11.399  35.247
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  62.3292     1.4986  41.591   <2e-16 ***
## parent_educ   1.4240     0.4408   3.231   0.0013 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.99 on 585 degrees of freedom
## Multiple R-squared:  0.01753,    Adjusted R-squared:  0.01585
## F-statistic: 10.44 on 1 and 585 DF,  p-value: 0.001304
```

```r
fit4 = lm(math_score ~ lunch_type, data = step_df)
summary(fit4)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type, data = step_df)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -58.286 -10.286   0.787  10.787  41.714
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   45.360      2.212   20.51   <2e-16 ***
## lunch_type    12.926      1.288   10.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.9 on 585 degrees of freedom
## Multiple R-squared:  0.1468, Adjusted R-squared:  0.1454
## F-statistic: 100.7 on 1 and 585 DF,  p-value: < 2.2e-16
```

```
fit5 = lm(math_score ~ parent_marital_status, data = step_df)
summary(fit5)
```

```
##
## Call:
## lm(formula = math_score ~ parent_marital_status, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.017 -10.235  -0.017  11.374  33.765
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            65.8442     2.1578  30.514   <2e-16 ***
## parent_marital_status   0.3911     0.9647   0.405    0.685
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.13 on 585 degrees of freedom
## Multiple R-squared:  0.0002809,  Adjusted R-squared:  -0.001428
## F-statistic: 0.1644 on 1 and 585 DF,  p-value: 0.6853
```

```
fit6 = lm(math_score ~ practice_sport, data = step_df)
summary(fit6)
```

```
##
## Call:
## lm(formula = math_score ~ practice_sport, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.873 -10.344   0.127  11.391  33.391
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     66.0793     2.4052  27.473   <2e-16 ***
## practice_sport   0.2647     1.0247   0.258    0.796
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 16.13 on 585 degrees of freedom
## Multiple R-squared:  0.000114,   Adjusted R-squared:  -0.001595
## F-statistic: 0.06671 on 1 and 585 DF,  p-value: 0.7963
```

```
fit7 = lm(math_score ~ is_first_child, data = step_df)
summary(fit7)
```

```
##
## Call:
## lm(formula = math_score ~ is_first_child, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.008 -10.008  -0.008  11.005  34.005
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      64.982      2.464  26.371   <2e-16 ***
## is_first_child    1.013      1.418   0.714    0.475
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.12 on 585 degrees of freedom
## Multiple R-squared:  0.000871,   Adjusted R-squared:  -0.0008369
## F-statistic:  0.51 on 1 and 585 DF,  p-value: 0.4754
```

```
fit8 = lm(math_score ~ nr_siblings, data = step_df)
summary(fit8)
```

```
##
## Call:
## lm(formula = math_score ~ nr_siblings, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.271 -10.234   0.112  11.037  33.803
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  65.1969     1.1675  55.841   <2e-16 ***
## nr_siblings   0.6914     0.4487   1.541    0.124
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.09 on 585 degrees of freedom
## Multiple R-squared:  0.004042,   Adjusted R-squared:  0.00234
## F-statistic: 2.374 on 1 and 585 DF,  p-value: 0.1239
```

```
fit9 = lm(math_score ~ transport_means, data = step_df)
summary(fit9)
```

```
##
## Call:
## lm(formula = math_score ~ transport_means, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.696 -10.646   0.304  11.304  33.354
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     66.59705    2.29558  29.011   <2e-16 ***
## transport_means  0.04924    1.36467   0.036    0.971
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.13 on 585 degrees of freedom
## Multiple R-squared:  2.226e-06,  Adjusted R-squared:  -0.001707
## F-statistic: 0.001302 on 1 and 585 DF,  p-value: 0.9712
```

```
fit10 = lm(math_score ~ wkly_study_hours, data = step_df)
summary(fit10)
```

```
##
## Call:
## lm(formula = math_score ~ wkly_study_hours, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -69.545  -9.902   0.098  11.598  33.742
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         61.614      2.038  30.236  < 2e-16 ***
## wkly_study_hours     2.644      1.007   2.627  0.00885 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.03 on 585 degrees of freedom
## Multiple R-squared:  0.01166,    Adjusted R-squared:  0.009966
## F-statistic: 6.899 on 1 and 585 DF,  p-value: 0.008851
```

```
fit11 = lm(math_score ~ test_prep, data = step_df)
summary(fit11)
```

```
##
## Call:
## lm(formula = math_score ~ test_prep, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.715  -9.715  -0.250  11.285  35.285
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   75.785      2.353  32.214  < 2e-16 ***
## test_prep     -5.535      1.373  -4.032 6.26e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.91 on 585 degrees of freedom
## Multiple R-squared:  0.02704,    Adjusted R-squared:  0.02538
## F-statistic: 16.26 on 1 and 585 DF,  p-value: 6.26e-05
```

```
# Enter first the one with the lowest p-value: Lunch Type
forward1 = lm(math_score ~ lunch_type, data = step_df)
first = summary(forward1)|> broom::tidy()


### Step 2: Enter the one with the lowest p-value in the rest
fit1 = update(forward1, . ~ . +gender)
summary(fit1)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + gender, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.033 -10.487   0.692  10.335  38.692
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   38.127      2.744  13.894  < 2e-16 ***
## lunch_type    12.632      1.271   9.939  < 2e-16 ***
## gender         5.274      1.216   4.336 1.71e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.67 on 584 degrees of freedom
## Multiple R-squared:  0.1734, Adjusted R-squared:  0.1706
## F-statistic: 61.27 on 2 and 584 DF,  p-value: < 2.2e-16
```

```
fit2 = update(forward1, . ~ . +ethnic_group)
summary(fit2)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.960  -9.607   0.552  10.353  36.199
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   36.5522     2.6533  13.776  < 2e-16 ***
```

```
## lunch_type    12.6467     1.2560  10.069  < 2e-16 ***
## ethnic_group   2.9205     0.5134   5.688 2.03e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.51 on 584 degrees of freedom
## Multiple R-squared:  0.1916, Adjusted R-squared:  0.1888
## F-statistic: 69.22 on 2 and 584 DF,  p-value: < 2.2e-16
```

```
fit3 = update(forward1, . ~ . +parent_educ)
summary(fit3)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + parent_educ, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.223 -10.409   0.879  10.845  41.828
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  40.7657     2.5292  16.118  < 2e-16 ***
## lunch_type   12.9827     1.2752  10.181  < 2e-16 ***
## parent_educ   1.4745     0.4066   3.627 0.000312 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.74 on 584 degrees of freedom
## Multiple R-squared:  0.1656, Adjusted R-squared:  0.1628
## F-statistic: 57.96 on 2 and 584 DF,  p-value: < 2.2e-16
```

```
fit4 = update(forward1, . ~ . +parent_marital_status)
summary(fit4)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + parent_marital_status,
##     data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.875 -10.434   0.865  10.845  41.826
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            43.8110     2.9633  14.785   <2e-16 ***
## lunch_type             12.9612     1.2895  10.052   <2e-16 ***
## parent_marital_status   0.7009     0.8920   0.786    0.432
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.9 on 584 degrees of freedom
```

```
## Multiple R-squared:  0.1477, Adjusted R-squared:  0.1448
## F-statistic: 50.61 on 2 and 584 DF,  p-value: < 2.2e-16
```

```
fit5 = update(forward1, . ~ . +practice_sport)
summary(fit5)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + practice_sport, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.678 -10.536   0.919  10.893  41.322
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     44.0952     3.1203  14.132   <2e-16 ***
## lunch_type      12.9480     1.2896  10.041   <2e-16 ***
## practice_sport   0.5449     0.9475   0.575    0.565
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.91 on 584 degrees of freedom
## Multiple R-squared:  0.1473, Adjusted R-squared:  0.1444
## F-statistic: 50.45 on 2 and 584 DF,  p-value: < 2.2e-16
```

```
fit6 = update(forward1, . ~ . +is_first_child)
summary(fit6)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + is_first_child, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.571 -10.529   0.513  10.513  42.279
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     43.9538     3.0975  14.190   <2e-16 ***
## lunch_type      12.9159     1.2890  10.020   <2e-16 ***
## is_first_child   0.8508     1.3113   0.649    0.517
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.9 on 584 degrees of freedom
## Multiple R-squared:  0.1474, Adjusted R-squared:  0.1445
## F-statistic:  50.5 on 2 and 584 DF,  p-value: < 2.2e-16
```

```
fit7 = update(forward1, . ~ . +nr_siblings)
summary(fit7)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + nr_siblings, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.877 -10.432   0.756  10.529  41.123
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  43.9040     2.3785  18.459   <2e-16 ***
## lunch_type   12.9222     1.2864  10.045   <2e-16 ***
## nr_siblings   0.6836     0.4147   1.648   0.0998 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.87 on 584 degrees of freedom
## Multiple R-squared:  0.1508, Adjusted R-squared:  0.1479
## F-statistic: 51.84 on 2 and 584 DF,  p-value: < 2.2e-16
```

```
fit8 = update(forward1, . ~ . +transport_means)
summary(fit8)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + transport_means, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.246 -10.246   0.754  10.827  41.652
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      45.5226     2.9870  15.240   <2e-16 ***
## lunch_type       12.9274     1.2895  10.025   <2e-16 ***
## transport_means  -0.1021     1.2617  -0.081    0.936
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.91 on 584 degrees of freedom
## Multiple R-squared:  0.1468, Adjusted R-squared:  0.1439
## F-statistic: 50.26 on 2 and 584 DF,  p-value: < 2.2e-16
```

```
fit9 = update(forward1, . ~ . +wkly_study_hours)
summary(fit9)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + wkly_study_hours, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.136 -10.433   0.947  10.485  41.485
```

```
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       40.3568     2.8257  14.282  < 2e-16 ***
## lunch_type        12.9173     1.2807  10.086  < 2e-16 ***
## wkly_study_hours   2.6206     0.9297   2.819  0.00498 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.81 on 584 degrees of freedom
## Multiple R-squared:  0.1583, Adjusted R-squared:  0.1554
## F-statistic: 54.91 on 2 and 584 DF,  p-value: < 2.2e-16
```

```
fit10 = update(forward1, . ~ . +test_prep)
summary(fit10)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + test_prep, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.071 -10.175   0.899  10.377  38.003
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   54.819      2.965   18.49  < 2e-16 ***
## lunch_type    13.104      1.266   10.35  < 2e-16 ***
## test_prep     -5.926      1.264   -4.69 3.41e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.64 on 584 degrees of freedom
## Multiple R-squared:  0.1778, Adjusted R-squared:  0.175
## F-statistic: 63.14 on 2 and 584 DF,  p-value: < 2.2e-16
```

```
# Enter the one with the lowest p-value: Ethnic Group
forward2 = update(forward1, . ~ . +ethnic_group)
summary(fit2)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.960  -9.607   0.552  10.353  36.199
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   36.5522     2.6533  13.776  < 2e-16 ***
## lunch_type    12.6467     1.2560  10.069  < 2e-16 ***
## ethnic_group   2.9205     0.5134   5.688 2.03e-08 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.51 on 584 degrees of freedom
## Multiple R-squared:  0.1916, Adjusted R-squared:  0.1888
## F-statistic: 69.22 on 2 and 584 DF,  p-value: < 2.2e-16
```

```
### Step 3: Enter the one with the lowest p-value in the rest
fit1 = update(forward2, . ~ . +gender)
summary(fit1)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + gender,
##     data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.717  -9.148   0.015   9.845  34.370
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   29.3670     3.0731   9.556  < 2e-16 ***
## lunch_type    12.3544     1.2382   9.978  < 2e-16 ***
## ethnic_group   2.9138     0.5054   5.765 1.32e-08 ***
## gender         5.2540     1.1842   4.437 1.09e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.29 on 583 degrees of freedom
## Multiple R-squared:  0.218,  Adjusted R-squared:  0.214
## F-statistic: 54.18 on 3 and 583 DF,  p-value: < 2.2e-16
```

```
fit2 = update(forward2, . ~ . +parent_educ)
summary(fit2)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + parent_educ,
##     data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.234  -9.790   0.258  10.355  36.529
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   32.8065     2.8632  11.458  < 2e-16 ***
## lunch_type    12.7088     1.2456  10.203  < 2e-16 ***
## ethnic_group   2.7998     0.5104   5.486 6.14e-08 ***
## parent_educ    1.3189     0.3978   3.315 0.000972 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 14.39 on 583 degrees of freedom
## Multiple R-squared:  0.2066, Adjusted R-squared:  0.2025
## F-statistic:  50.6 on 3 and 583 DF,  p-value: < 2.2e-16
```

```
fit3 = update(forward2, . ~ . +parent_marital_status)
summary(fit3)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + parent_marital_status,
##     data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.657  -9.579   0.554  10.482  36.308
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            34.6762     3.2998  10.509  < 2e-16 ***
## lunch_type             12.6870     1.2568  10.094  < 2e-16 ***
## ethnic_group            2.9334     0.5136   5.711 1.79e-08 ***
## parent_marital_status   0.8312     0.8691   0.956    0.339
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.51 on 583 degrees of freedom
## Multiple R-squared:  0.1929, Adjusted R-squared:  0.1887
## F-statistic: 46.44 on 3 and 583 DF,  p-value: < 2.2e-16
```

```
fit4 = update(forward2, . ~ . +practice_sport)
summary(fit4)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + practice_sport,
##     data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.356  -9.474   0.605  10.400  35.802
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     35.2709     3.4130  10.334  < 2e-16 ***
## lunch_type      12.6688     1.2573  10.076  < 2e-16 ***
## ethnic_group     2.9209     0.5137   5.686 2.06e-08 ***
## practice_sport   0.5514     0.9231   0.597    0.551
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.52 on 583 degrees of freedom
## Multiple R-squared:  0.1921, Adjusted R-squared:  0.188
## F-statistic: 46.21 on 3 and 583 DF,  p-value: < 2.2e-16
```

```r
fit5 = update(forward2, . ~ . +is_first_child)
summary(fit5)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + is_first_child,
##     data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.244  -9.840   0.756  10.162  36.762
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     35.1542     3.3914  10.366  < 2e-16 ***
## lunch_type      12.6365     1.2567  10.055  < 2e-16 ***
## ethnic_group     2.9203     0.5137   5.685 2.07e-08 ***
## is_first_child   0.8461     1.2775   0.662    0.508
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.52 on 583 degrees of freedom
## Multiple R-squared:  0.1922, Adjusted R-squared:  0.1881
## F-statistic: 46.25 on 3 and 583 DF,  p-value: < 2.2e-16
```

```r
fit6 = update(forward2, . ~ . +nr_siblings)
summary(fit6)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + nr_siblings,
##     data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.600  -9.548   0.345  10.095  35.510
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   34.8937     2.7968  12.476  < 2e-16 ***
## lunch_type    12.6400     1.2535  10.084  < 2e-16 ***
## ethnic_group   2.9450     0.5125   5.746 1.47e-08 ***
## nr_siblings    0.7439     0.4039   1.842    0.066 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.48 on 583 degrees of freedom
## Multiple R-squared:  0.1963, Adjusted R-squared:  0.1922
## F-statistic: 47.46 on 3 and 583 DF,  p-value: < 2.2e-16
```

```r
fit7 = update(forward2, . ~ . +transport_means)
summary(fit7)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + transport_means,
##     data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.006  -9.633   0.620  10.310  36.265
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      36.3653     3.3263  10.933  < 2e-16 ***
## lunch_type       12.6452     1.2572  10.058  < 2e-16 ***
## ethnic_group      2.9220     0.5141   5.684 2.08e-08 ***
## transport_means   0.1147     1.2297   0.093    0.926
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.52 on 583 degrees of freedom
## Multiple R-squared:  0.1916, Adjusted R-squared:  0.1875
## F-statistic: 46.07 on 3 and 583 DF,  p-value: < 2.2e-16
```

```
fit8 = update(forward2, . ~ . +wkly_study_hours)
summary(fit8)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + wkly_study_hours,
##     data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.754  -9.175   0.814   9.769  36.003
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       31.6953     3.1451  10.078  < 2e-16 ***
## lunch_type        12.6394     1.2485  10.123  < 2e-16 ***
## ethnic_group       2.9056     0.5104   5.693 1.98e-08 ***
## wkly_study_hours   2.5674     0.9057   2.835  0.00474 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.43 on 583 degrees of freedom
## Multiple R-squared:  0.2026, Adjusted R-squared:  0.1985
## F-statistic: 49.38 on 3 and 583 DF,  p-value: < 2.2e-16
```

```
fit9 = update(forward2, . ~ . +test_prep)
summary(fit9)
```

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

```
##     data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.800  -9.572   1.075  10.329  32.657
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   45.9342     3.2818  13.997  < 2e-16 ***
## lunch_type    12.8245     1.2345  10.389  < 2e-16 ***
## ethnic_group   2.8753     0.5044   5.700 1.91e-08 ***
## test_prep     -5.7920     1.2310  -4.705 3.17e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.26 on 583 degrees of freedom
## Multiple R-squared:  0.2212, Adjusted R-squared:  0.2172
## F-statistic: 55.19 on 3 and 583 DF,  p-value: < 2.2e-16
```

```
# Enter the one with the lowest p-value: Test Prep
forward3 = update(forward2, . ~ . + test_prep)
summary(forward3)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep,
##     data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.800  -9.572   1.075  10.329  32.657
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   45.9342     3.2818  13.997  < 2e-16 ***
## lunch_type    12.8245     1.2345  10.389  < 2e-16 ***
## ethnic_group   2.8753     0.5044   5.700 1.91e-08 ***
## test_prep     -5.7920     1.2310  -4.705 3.17e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.26 on 583 degrees of freedom
## Multiple R-squared:  0.2212, Adjusted R-squared:  0.2172
## F-statistic: 55.19 on 3 and 583 DF,  p-value: < 2.2e-16
```

```
### Step 4: Enter the one with the lowest p-value in the rest
fit1 = update(forward3, . ~ . +gender)
summary(fit1)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     gender, data = step_df)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.734 -10.013   0.383   9.970  32.599
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    38.700      3.642  10.627  < 2e-16 ***
## lunch_type     12.538      1.218  10.294  < 2e-16 ***
## ethnic_group    2.870      0.497   5.776 1.25e-08 ***
## test_prep      -5.573      1.214  -4.591 5.40e-06 ***
## gender          5.031      1.165   4.317 1.86e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.05 on 582 degrees of freedom
## Multiple R-squared:  0.2454, Adjusted R-squared:  0.2402
## F-statistic: 47.31 on 4 and 582 DF,  p-value: < 2.2e-16
```

```
fit2 = update(forward3, . ~ . +parent_educ)
summary(fit2)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     parent_educ, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.947  -9.121   1.049  10.283  32.938
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   42.2241     3.4203  12.345  < 2e-16 ***
## lunch_type    12.8917     1.2230  10.541  < 2e-16 ***
## ethnic_group   2.7498     0.5010   5.489 6.04e-08 ***
## test_prep     -5.8907     1.2197  -4.830 1.75e-06 ***
## parent_educ    1.3626     0.3905   3.489 0.000521 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.12 on 582 degrees of freedom
## Multiple R-squared:  0.2371, Adjusted R-squared:  0.2319
## F-statistic: 45.23 on 4 and 582 DF,  p-value: < 2.2e-16
```

```
fit3 = update(forward3, . ~ . +parent_marital_status)
summary(fit3)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     parent_marital_status, data = step_df)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.400  -9.670   1.073  10.241  32.769
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            44.2993     3.8406  11.534  < 2e-16 ***
## lunch_type             12.8575     1.2355  10.407  < 2e-16 ***
## ethnic_group            2.8864     0.5048   5.718 1.72e-08 ***
## test_prep              -5.7590     1.2320  -4.674 3.67e-06 ***
## parent_marital_status   0.7007     0.8544   0.820    0.413
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.26 on 582 degrees of freedom
## Multiple R-squared:  0.2221, Adjusted R-squared:  0.2167
## F-statistic: 41.54 on 4 and 582 DF,  p-value: < 2.2e-16
```

```
fit4 = update(forward3, . ~ . +practice_sport)
summary(fit4)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     practice_sport, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.155  -9.277   1.001  10.183  32.311
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     44.7832     3.9179  11.430  < 2e-16 ***
## lunch_type      12.8438     1.2358  10.393  < 2e-16 ***
## ethnic_group     2.8757     0.5048   5.697 1.94e-08 ***
## test_prep       -5.7822     1.2319  -4.694 3.35e-06 ***
## practice_sport   0.4885     0.9070   0.539     0.59
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.27 on 582 degrees of freedom
## Multiple R-squared:  0.2216, Adjusted R-squared:  0.2162
## F-statistic: 41.42 on 4 and 582 DF,  p-value: < 2.2e-16
```

```
fit5 = update(forward3, . ~ . +is_first_child)
summary(fit5)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     is_first_child, data = step_df)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -55.99  -9.71   1.07  10.31  33.03
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     45.0129     3.9442  11.412  < 2e-16 ***
## lunch_type      12.8173     1.2355  10.374  < 2e-16 ***
## ethnic_group     2.8753     0.5048   5.696 1.95e-08 ***
## test_prep       -5.7640     1.2337  -4.672 3.70e-06 ***
## is_first_child   0.5302     1.2571   0.422    0.673
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.27 on 582 degrees of freedom
## Multiple R-squared:  0.2214, Adjusted R-squared:  0.2161
## F-statistic: 41.38 on 4 and 582 DF,  p-value: < 2.2e-16
```

```
fit6 = update(forward3, . ~ . +nr_siblings)
summary(fit6)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     nr_siblings, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.400  -9.375   1.049  10.075  32.111
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   44.3155     3.4213  12.953  < 2e-16 ***
## lunch_type    12.8156     1.2327  10.396  < 2e-16 ***
## ethnic_group   2.8976     0.5039   5.750 1.44e-08 ***
## test_prep     -5.6936     1.2307  -4.626 4.59e-06 ***
## nr_siblings    0.6545     0.3975   1.647      0.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.24 on 582 degrees of freedom
## Multiple R-squared:  0.2248, Adjusted R-squared:  0.2195
## F-statistic: 42.19 on 4 and 582 DF,  p-value: < 2.2e-16
```

```
fit7 = update(forward3, . ~ . +transport_means)
summary(fit7)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     transport_means, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -55.993  -9.355   0.889  10.216  32.935
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      45.1465     3.7597  12.008  < 2e-16 ***
## lunch_type       12.8187     1.2354  10.376  < 2e-16 ***
## ethnic_group      2.8817     0.5050   5.706 1.84e-08 ***
## test_prep        -5.8298     1.2350  -4.720 2.95e-06 ***
## transport_means   0.5211     1.2110   0.430    0.667
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.27 on 582 degrees of freedom
## Multiple R-squared:  0.2214, Adjusted R-squared:  0.2161
## F-statistic: 41.38 on 4 and 582 DF,  p-value: < 2.2e-16
```

```
fit8 = update(forward3, . ~ . +wkly_study_hours)
summary(fit8)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     wkly_study_hours, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.284  -9.309   0.857   9.743  32.666
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       41.3369     3.7728  10.957  < 2e-16 ***
## lunch_type        12.8095     1.2293  10.420  < 2e-16 ***
## ethnic_group       2.8648     0.5023   5.703 1.88e-08 ***
## test_prep         -5.5040     1.2315  -4.469 9.44e-06 ***
## wkly_study_hours   2.1836     0.8954   2.439    0.015 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.2 on 582 degrees of freedom
## Multiple R-squared:  0.2291, Adjusted R-squared:  0.2238
## F-statistic: 43.23 on 4 and 582 DF,  p-value: < 2.2e-16
```

```
# Enter the one with the lowest p-value: Gender
forward4 = update(forward3, . ~ . + gender)
summary(forward4)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     gender, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -53.734 -10.013   0.383   9.970  32.599
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    38.700      3.642  10.627  < 2e-16 ***
## lunch_type     12.538      1.218  10.294  < 2e-16 ***
## ethnic_group    2.870      0.497   5.776 1.25e-08 ***
## test_prep      -5.573      1.214  -4.591 5.40e-06 ***
## gender          5.031      1.165   4.317 1.86e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.05 on 582 degrees of freedom
## Multiple R-squared:  0.2454, Adjusted R-squared:  0.2402
## F-statistic: 47.31 on 4 and 582 DF,  p-value: < 2.2e-16
```

```
### Step 5: Enter the one with the lowest p-value in the rest
fit1 = update(forward4, . ~ . +parent_educ)
summary(fit1)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     gender + parent_educ, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.514  -9.535   0.837  10.037  30.067
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   34.2503     3.7807   9.059  < 2e-16 ***
## lunch_type    12.5942     1.2039  10.461  < 2e-16 ***
## ethnic_group   2.7340     0.4925   5.552 4.31e-08 ***
## test_prep     -5.6676     1.2000  -4.723 2.92e-06 ***
## gender         5.3234     1.1542   4.612 4.91e-06 ***
## parent_educ    1.4796     0.3847   3.846 0.000133 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.88 on 581 degrees of freedom
## Multiple R-squared:  0.2641, Adjusted R-squared:  0.2578
## F-statistic:  41.7 on 5 and 581 DF,  p-value: < 2.2e-16
```

```
fit2 = update(forward4, . ~ . +parent_marital_status)
summary(fit2)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     gender + parent_marital_status, data = step_df)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -54.421  -9.827   0.520  10.165  32.709
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            36.7294     4.1646   8.820  < 2e-16 ***
## lunch_type             12.5744     1.2187  10.318  < 2e-16 ***
## ethnic_group            2.8835     0.4972   5.800 1.09e-08 ***
## test_prep              -5.5329     1.2146  -4.555 6.38e-06 ***
## gender                  5.0681     1.1660   4.347 1.63e-05 ***
## parent_marital_status   0.8215     0.8420   0.976     0.33
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.05 on 581 degrees of freedom
## Multiple R-squared:  0.2466, Adjusted R-squared:  0.2401
## F-statistic: 38.03 on 5 and 581 DF,  p-value: < 2.2e-16
```

```
fit3 = update(forward4, . ~ . +practice_sport)
summary(fit3)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     gender + practice_sport, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.066  -9.907   0.260  10.048  32.249
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     37.6369     4.2012   8.959  < 2e-16 ***
## lunch_type      12.5562     1.2193  10.298  < 2e-16 ***
## ethnic_group     2.8709     0.4973   5.773 1.27e-08 ***
## test_prep       -5.5643     1.2148  -4.581 5.68e-06 ***
## gender           5.0253     1.1661   4.310 1.92e-05 ***
## practice_sport   0.4544     0.8936   0.509    0.611
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.06 on 581 degrees of freedom
## Multiple R-squared:  0.2457, Adjusted R-squared:  0.2392
## F-statistic: 37.85 on 5 and 581 DF,  p-value: < 2.2e-16
```

```
fit4 = update(forward4, . ~ . +is_first_child)
summary(fit4)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     gender + is_first_child, data = step_df)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.902  -9.853   0.541  10.069  32.437
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     37.8925     4.2227   8.973  < 2e-16 ***
## lunch_type      12.5318     1.2190  10.280  < 2e-16 ***
## ethnic_group     2.8706     0.4974   5.772 1.28e-08 ***
## test_prep       -5.5487     1.2165  -4.561 6.21e-06 ***
## gender           5.0255     1.1662   4.309 1.92e-05 ***
## is_first_child   0.4690     1.2386   0.379    0.705
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.06 on 581 degrees of freedom
## Multiple R-squared:  0.2455, Adjusted R-squared:  0.239
## F-statistic: 37.82 on 5 and 581 DF,  p-value: < 2.2e-16
```

```
fit5 = update(forward4, . ~ . +nr_siblings)
summary(fit5)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     gender + nr_siblings, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.359  -9.453   0.056  10.049  34.165
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   36.8121     3.7775   9.745  < 2e-16 ***
## lunch_type    12.5237     1.2156  10.303  < 2e-16 ***
## ethnic_group   2.8949     0.4962   5.835 8.96e-09 ***
## test_prep     -5.4619     1.2129  -4.503 8.10e-06 ***
## gender         5.1091     1.1637   4.390 1.35e-05 ***
## nr_siblings    0.7178     0.3917   1.833   0.0674 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.02 on 581 degrees of freedom
## Multiple R-squared:  0.2497, Adjusted R-squared:  0.2432
## F-statistic: 38.67 on 5 and 581 DF,  p-value: < 2.2e-16
```

```
fit6 = update(forward4, . ~ . +transport_means)
summary(fit6)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     gender + transport_means, data = step_df)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.896 -10.039   0.299   9.852  32.878
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      38.0574     4.0535   9.389  < 2e-16 ***
## lunch_type       12.5335     1.2190  10.282  < 2e-16 ***
## ethnic_group      2.8759     0.4976   5.780 1.22e-08 ***
## test_prep        -5.6049     1.2179  -4.602 5.14e-06 ***
## gender            5.0233     1.1663   4.307 1.94e-05 ***
## transport_means   0.4322     1.1933   0.362    0.717
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.06 on 581 degrees of freedom
## Multiple R-squared:  0.2455, Adjusted R-squared:  0.239
## F-statistic: 37.81 on 5 and 581 DF,  p-value: < 2.2e-16
```

```
fit7 = update(forward4, . ~ . +wkly_study_hours)
summary(fit7)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     gender + wkly_study_hours, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.095  -9.508   0.175  10.116  34.314
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       34.5555     4.0481   8.536  < 2e-16 ***
## lunch_type        12.5298     1.2135  10.325  < 2e-16 ***
## ethnic_group       2.8609     0.4952   5.778 1.24e-08 ***
## test_prep         -5.3087     1.2148  -4.370 1.47e-05 ***
## gender             4.9272     1.1618   4.241 2.59e-05 ***
## wkly_study_hours   2.0392     0.8833   2.309   0.0213 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.99 on 581 degrees of freedom
## Multiple R-squared:  0.2522, Adjusted R-squared:  0.2458
## F-statistic: 39.19 on 5 and 581 DF,  p-value: < 2.2e-16
```

```
# Enter the one with the lowest p-value: Parent Education
forward5 = update(forward4, . ~ . + parent_educ)
summary(forward5)
```

```
##
## Call:
```

```
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     gender + parent_educ, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.514  -9.535   0.837  10.037  30.067
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   34.2503     3.7807   9.059  < 2e-16 ***
## lunch_type    12.5942     1.2039  10.461  < 2e-16 ***
## ethnic_group   2.7340     0.4925   5.552 4.31e-08 ***
## test_prep     -5.6676     1.2000  -4.723 2.92e-06 ***
## gender         5.3234     1.1542   4.612 4.91e-06 ***
## parent_educ    1.4796     0.3847   3.846 0.000133 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.88 on 581 degrees of freedom
## Multiple R-squared:  0.2641, Adjusted R-squared:  0.2578
## F-statistic:  41.7 on 5 and 581 DF,  p-value: < 2.2e-16
```

```
### Step 6: Enter the one with the lowest p-value in the rest
fit1 = update(forward5, . ~ . +parent_marital_status)
summary(fit1)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     gender + parent_educ + parent_marital_status, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.161  -9.524   0.562  10.079  30.168
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            32.4681     4.2647   7.613 1.09e-13 ***
## lunch_type             12.6273     1.2047  10.482  < 2e-16 ***
## ethnic_group            2.7467     0.4928   5.574 3.82e-08 ***
## test_prep              -5.6302     1.2009  -4.688 3.44e-06 ***
## gender                  5.3563     1.1550   4.637 4.36e-06 ***
## parent_educ             1.4721     0.3849   3.825 0.000145 ***
## parent_marital_status   0.7524     0.8325   0.904 0.366512
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.88 on 580 degrees of freedom
## Multiple R-squared:  0.2651, Adjusted R-squared:  0.2575
## F-statistic: 34.88 on 6 and 580 DF,  p-value: < 2.2e-16
```

```
fit2 = update(forward5, . ~ . +practice_sport)
summary(fit2)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     gender + parent_educ + practice_sport, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.967  -9.382   0.825   9.975  29.595
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     32.6121     4.3479   7.501 2.40e-13 ***
## lunch_type      12.6220     1.2049  10.475  < 2e-16 ***
## ethnic_group     2.7329     0.4927   5.547 4.42e-08 ***
## test_prep       -5.6556     1.2005  -4.711 3.09e-06 ***
## gender           5.3193     1.1547   4.607 5.03e-06 ***
## parent_educ      1.4986     0.3857   3.886 0.000114 ***
## practice_sport   0.6758     0.8848   0.764 0.445283
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.89 on 580 degrees of freedom
## Multiple R-squared:  0.2648, Adjusted R-squared:  0.2572
## F-statistic: 34.82 on 6 and 580 DF,  p-value: < 2.2e-16
```

```
fit3 = update(forward5, . ~ . +is_first_child)
summary(fit3)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     gender + parent_educ + is_first_child, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.673  -9.645   0.774   9.999  30.373
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     33.5052     4.3272   7.743 4.35e-14 ***
## lunch_type      12.5885     1.2050  10.447  < 2e-16 ***
## ethnic_group     2.7342     0.4929   5.548 4.41e-08 ***
## test_prep       -5.6448     1.2026  -4.694 3.35e-06 ***
## gender           5.3185     1.1552   4.604 5.10e-06 ***
## parent_educ      1.4786     0.3850   3.840 0.000136 ***
## is_first_child   0.4345     1.2242   0.355 0.722808
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.89 on 580 degrees of freedom
## Multiple R-squared:  0.2643, Adjusted R-squared:  0.2566
## F-statistic: 34.72 on 6 and 580 DF,  p-value: < 2.2e-16
```

```r
fit4 = update(forward5, . ~ . +nr_siblings)
summary(fit4)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     gender + parent_educ + nr_siblings, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.136  -9.280   0.816  10.107  31.298
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   32.2786     3.9112   8.253 1.05e-15 ***
## lunch_type    12.5799     1.2012  10.472  < 2e-16 ***
## ethnic_group   2.7582     0.4915   5.611 3.11e-08 ***
## test_prep     -5.5537     1.1988  -4.633 4.46e-06 ***
## gender         5.4060     1.1524   4.691 3.39e-06 ***
## parent_educ    1.4896     0.3839   3.880 0.000116 ***
## nr_siblings    0.7382     0.3871   1.907 0.056968 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.85 on 580 degrees of freedom
## Multiple R-squared:  0.2687, Adjusted R-squared:  0.2611
## F-statistic: 35.51 on 6 and 580 DF,  p-value: < 2.2e-16
```

```r
fit5 = update(forward5, . ~ . +transport_means)
summary(fit5)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     gender + parent_educ + transport_means, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.661  -9.669   0.663   9.877  30.273
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     33.6893     4.1648   8.089 3.53e-15 ***
## lunch_type      12.5902     1.2049  10.449  < 2e-16 ***
## ethnic_group     2.7389     0.4931   5.555 4.25e-08 ***
## test_prep       -5.6954     1.2040  -4.730 2.82e-06 ***
## gender           5.3167     1.1553   4.602 5.15e-06 ***
## parent_educ      1.4782     0.3850   3.839 0.000137 ***
## transport_means  0.3801     1.1795   0.322 0.747350
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 13.89 on 580 degrees of freedom
## Multiple R-squared:  0.2642, Adjusted R-squared:  0.2566
## F-statistic: 34.71 on 6 and 580 DF,  p-value: < 2.2e-16
```

```
fit6 = update(forward5, . ~ . +wkly_study_hours)
summary(fit6)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     gender + parent_educ + wkly_study_hours, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.943  -9.439   0.630  10.403  31.459
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       29.7605     4.1787   7.122 3.16e-12 ***
## lunch_type        12.5868     1.1987  10.501  < 2e-16 ***
## ethnic_group       2.7208     0.4904   5.549 4.39e-08 ***
## test_prep         -5.3895     1.2000  -4.491 8.55e-06 ***
## gender             5.2204     1.1499   4.540 6.85e-06 ***
## parent_educ        1.5128     0.3833   3.947 8.88e-05 ***
## wkly_study_hours   2.1599     0.8729   2.474   0.0136 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.82 on 580 degrees of freedom
## Multiple R-squared:  0.2718, Adjusted R-squared:  0.2642
## F-statistic: 36.08 on 6 and 580 DF,  p-value: < 2.2e-16
```

```
# Enter the one with the lowest p-value: Weekly Study Hours
forward6 = update(forward5, . ~ . + wkly_study_hours)
summary(forward6)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     gender + parent_educ + wkly_study_hours, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.943  -9.439   0.630  10.403  31.459
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       29.7605     4.1787   7.122 3.16e-12 ***
## lunch_type        12.5868     1.1987  10.501  < 2e-16 ***
## ethnic_group       2.7208     0.4904   5.549 4.39e-08 ***
## test_prep         -5.3895     1.2000  -4.491 8.55e-06 ***
## gender             5.2204     1.1499   4.540 6.85e-06 ***
## parent_educ        1.5128     0.3833   3.947 8.88e-05 ***
```

```
## wkly_study_hours   2.1599     0.8729   2.474   0.0136 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.82 on 580 degrees of freedom
## Multiple R-squared:  0.2718, Adjusted R-squared:  0.2642
## F-statistic: 36.08 on 6 and 580 DF,  p-value: < 2.2e-16
```

```
### Step 7: Enter the one with the lowest p-value in the rest
fit1 = update(forward6, . ~ . +parent_marital_status)
summary(fit1)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     gender + parent_educ + wkly_study_hours + parent_marital_status,
##     data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.608  -9.405   0.583  10.152  31.582
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            27.9334     4.6228   6.043 2.72e-09 ***
## lunch_type             12.6206     1.1994  10.523  < 2e-16 ***
## ethnic_group            2.7336     0.4906   5.572 3.87e-08 ***
## test_prep              -5.3507     1.2009  -4.456 1.00e-05 ***
## gender                  5.2537     1.1506   4.566 6.08e-06 ***
## parent_educ             1.5052     0.3834   3.926 9.68e-05 ***
## wkly_study_hours        2.1654     0.8731   2.480   0.0134 *
## parent_marital_status   0.7665     0.8289   0.925   0.3555
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.82 on 579 degrees of freedom
## Multiple R-squared:  0.2729, Adjusted R-squared:  0.2641
## F-statistic: 31.04 on 7 and 579 DF,  p-value: < 2.2e-16
```

```
fit2 = update(forward6, . ~ . +practice_sport)
summary(fit2)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     gender + parent_educ + wkly_study_hours + practice_sport,
##     data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.363  -9.358   0.643  10.342  30.921
##
## Coefficients:
```

```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       28.2228     4.6818   6.028 2.95e-09 ***
## lunch_type        12.6133     1.1997  10.514  < 2e-16 ***
## ethnic_group       2.7197     0.4906   5.544 4.49e-08 ***
## test_prep         -5.3794     1.2006  -4.481 8.97e-06 ***
## gender             5.2170     1.1504   4.535 7.01e-06 ***
## parent_educ        1.5307     0.3842   3.984 7.64e-05 ***
## wkly_study_hours   2.1502     0.8734   2.462   0.0141 *
## practice_sport     0.6427     0.8811   0.729   0.4660
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.83 on 579 degrees of freedom
## Multiple R-squared:  0.2724, Adjusted R-squared:  0.2636
## F-statistic: 30.97 on 7 and 579 DF,  p-value: < 2.2e-16
```

```
fit3 = update(forward6, . ~ . +is_first_child)
summary(fit3)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     gender + parent_educ + wkly_study_hours + is_first_child,
##     data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.096  -9.567   0.708  10.528  31.314
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       29.0370     4.6723   6.215 9.84e-10 ***
## lunch_type        12.5812     1.1997  10.487  < 2e-16 ***
## ethnic_group       2.7209     0.4907   5.545 4.48e-08 ***
## test_prep         -5.3675     1.2026  -4.463 9.71e-06 ***
## gender             5.2157     1.1509   4.532 7.10e-06 ***
## parent_educ        1.5119     0.3836   3.941 9.09e-05 ***
## wkly_study_hours   2.1588     0.8736   2.471   0.0138 *
## is_first_child     0.4232     1.2189   0.347   0.7286
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.83 on 579 degrees of freedom
## Multiple R-squared:  0.2719, Adjusted R-squared:  0.2631
## F-statistic: 30.89 on 7 and 579 DF,  p-value: < 2.2e-16
```

```
fit4 = update(forward6, . ~ . +nr_siblings)
summary(fit4)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     gender + parent_educ + wkly_study_hours + nr_siblings, data = step_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 ***
## lunch_type        12.5737     1.1964  10.510  < 2e-16 ***
## ethnic_group       2.7439     0.4896   5.605 3.23e-08 ***
## test_prep         -5.2926     1.1989  -4.414 1.21e-05 ***
## gender             5.3017     1.1486   4.616 4.83e-06 ***
## parent_educ        1.5210     0.3826   3.976 7.90e-05 ***
## wkly_study_hours   2.0825     0.8723   2.387   0.0173 *
## nr_siblings        0.6927     0.3860   1.795   0.0732 .
## ---
## 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
```

```
fit5 = update(forward6, . ~ . +transport_means)
summary(fit5)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     gender + parent_educ + wkly_study_hours + transport_means,
##     data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.041  -9.471   0.579  10.519  31.633
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       29.3703     4.5026   6.523 1.50e-10 ***
## lunch_type        12.5840     1.1997  10.489  < 2e-16 ***
## ethnic_group       2.7243     0.4910   5.549 4.39e-08 ***
## test_prep         -5.4106     1.2044  -4.493 8.50e-06 ***
## gender             5.2159     1.1510   4.531 7.12e-06 ***
## parent_educ        1.5117     0.3836   3.941 9.12e-05 ***
## wkly_study_hours   2.1524     0.8742   2.462   0.0141 *
## transport_means    0.2749     1.1752   0.234   0.8151
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.83 on 579 degrees of freedom
## Multiple R-squared:  0.2718, Adjusted R-squared:  0.263
## F-statistic: 30.88 on 7 and 579 DF,  p-value: < 2.2e-16
```

```r
# P-value of all new added variables are larger than 0.05, which means that they
# are not significant predictor, and we stop here.

math_forward_manual_fit = lm(math_score ~ lunch_type + ethnic_group + test_prep +
    gender + parent_educ + wkly_study_hours, data = step_df)
summary(math_forward_manual_fit)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     gender + parent_educ + wkly_study_hours, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.943  -9.439   0.630  10.403  31.459
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       29.7605     4.1787   7.122 3.16e-12 ***
## lunch_type        12.5868     1.1987  10.501  < 2e-16 ***
## ethnic_group       2.7208     0.4904   5.549 4.39e-08 ***
## test_prep         -5.3895     1.2000  -4.491 8.55e-06 ***
## gender             5.2204     1.1499   4.540 6.85e-06 ***
## parent_educ        1.5128     0.3833   3.947 8.88e-05 ***
## wkly_study_hours   2.1599     0.8729   2.474   0.0136 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.82 on 580 degrees of freedom
## Multiple R-squared:  0.2718, Adjusted R-squared:  0.2642
## F-statistic: 36.08 on 6 and 580 DF,  p-value: < 2.2e-16
```

```r
mean(math_forward_manual_fit$residuals^2)
```

```
## [1] 188.763
```

```r
# fit using one function
intercept_only <- lm (math_score ~ 1, data = math_df)
math_forward_func_fit = step(intercept_only, direction = "forward", scope = formula(mult.fit))
```

```
## Start:  AIC=3264.33
## math_score ~ 1
##
##                        Df Sum of Sq    RSS    AIC
## + lunch_type            1   22340.6 129816 3173.1
## + ethnic_group          1    7803.7 144353 3235.4
## + gender                1    5114.8 147042 3246.3
## + test_prep             1    4114.3 148042 3250.2
## + parent_educ           1    2667.1 149489 3256.0
## + wkly_study_hours      1    1773.4 150383 3259.5
## + nr_siblings           1     615.0 151541 3264.0
## <none>                              152157 3264.3
```

```
## + is_first_child         1     132.5 152024 3265.8
## + parent_marital_status  1      42.7 152114 3266.2
## + practice_sport         1      17.3 152139 3266.3
## + transport_means        1       0.3 152156 3266.3
##
## Step:  AIC=3173.12
## math_score ~ lunch_type
##
##                          Df Sum of Sq    RSS    AIC
## + ethnic_group            1    6815.3 123001 3143.5
## + test_prep               1    4711.5 125104 3153.4
## + gender                  1    4049.1 125767 3156.5
## + parent_educ             1    2859.4 126956 3162.0
## + wkly_study_hours        1    1742.6 128073 3167.2
## + nr_siblings             1     601.2 129215 3172.4
## <none>                                 129816 3173.1
## + parent_marital_status   1     137.1 129679 3174.5
## + is_first_child          1      93.5 129722 3174.7
## + practice_sport          1      73.5 129742 3174.8
## + transport_means         1       1.5 129814 3175.1
##
## Step:  AIC=3143.47
## math_score ~ lunch_type + ethnic_group
##
##                          Df Sum of Sq    RSS    AIC
## + test_prep               1    4499.7 118501 3123.6
## + gender                  1    4017.7 118983 3126.0
## + parent_educ             1    2276.0 120725 3134.5
## + wkly_study_hours        1    1672.4 121328 3137.4
## + nr_siblings             1     711.4 122289 3142.1
## <none>                                 123001 3143.5
## + parent_marital_status   1     192.7 122808 3144.6
## + is_first_child          1      92.5 122908 3145.0
## + practice_sport          1      75.2 122925 3145.1
## + transport_means         1       1.8 122999 3145.5
##
## Step:  AIC=3123.59
## math_score ~ lunch_type + ethnic_group + test_prep
##
##                          Df Sum of Sq    RSS    AIC
## + gender                  1    3676.9 114824 3107.1
## + parent_educ             1    2428.1 116073 3113.4
## + wkly_study_hours        1    1198.6 117302 3119.6
## + nr_siblings             1     549.5 117951 3122.9
## <none>                                 118501 3123.6
## + parent_marital_status   1     136.8 118364 3124.9
## + practice_sport          1      59.0 118442 3125.3
## + transport_means         1      37.7 118463 3125.4
## + is_first_child          1      36.2 118465 3125.4
##
## Step:  AIC=3107.09
## math_score ~ lunch_type + ethnic_group + test_prep + gender
##
##                          Df Sum of Sq    RSS    AIC
```

```
## + parent_educ            1   2850.64 111973 3094.3
## + wkly_study_hours       1   1043.73 113780 3103.7
## + nr_siblings            1    659.98 114164 3105.7
## <none>                                114824 3107.1
## + parent_marital_status  1    187.79 114636 3108.1
## + practice_sport         1     51.08 114773 3108.8
## + is_first_child         1     28.33 114796 3108.9
## + transport_means        1     25.92 114798 3109.0
##
## Step:  AIC=3094.33
## math_score ~ lunch_type + ethnic_group + test_prep + gender +
##     parent_educ
##
##                          Df Sum of Sq    RSS    AIC
## + wkly_study_hours       1   1169.53 110804 3090.2
## + nr_siblings            1    697.97 111275 3092.7
## <none>                                111973 3094.3
## + parent_marital_status  1    157.45 111816 3095.5
## + practice_sport         1    112.52 111861 3095.7
## + is_first_child         1     24.31 111949 3096.2
## + transport_means        1     20.05 111953 3096.2
##
## Step:  AIC=3090.17
## math_score ~ lunch_type + ethnic_group + test_prep + gender +
##     parent_educ + wkly_study_hours
##
##                          Df Sum of Sq    RSS    AIC
## + nr_siblings            1    613.00 110191 3088.9
## <none>                                110804 3090.2
## + parent_marital_status  1    163.41 110640 3091.3
## + practice_sport         1    101.74 110702 3091.6
## + is_first_child         1     23.06 110781 3092.1
## + transport_means        1     10.47 110793 3092.1
##
## Step:  AIC=3088.91
## math_score ~ lunch_type + ethnic_group + test_prep + gender +
##     parent_educ + wkly_study_hours + nr_siblings
##
##                          Df Sum of Sq    RSS    AIC
## <none>                                110191 3088.9
## + parent_marital_status  1   138.323 110053 3090.2
## + practice_sport         1    98.281 110093 3090.4
## + is_first_child         1    48.607 110142 3090.7
## + transport_means        1    10.595 110180 3090.9
```

```
summary(math_forward_func_fit)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     gender + parent_educ + wkly_study_hours + nr_siblings, 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 ***
## lunch_type        12.5737     1.1964  10.510  < 2e-16 ***
## ethnic_group       2.7439     0.4896   5.605 3.23e-08 ***
## test_prep         -5.2926     1.1989  -4.414 1.21e-05 ***
## gender             5.3017     1.1486   4.616 4.83e-06 ***
## parent_educ        1.5210     0.3826   3.976 7.90e-05 ***
## wkly_study_hours   2.0825     0.8723   2.387   0.0173 *
## nr_siblings        0.6927     0.3860   1.795   0.0732 .
## ---
## 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
```

```
mean(math_forward_func_fit$residuals^2)
```

```
## [1] 187.7187
```

The model we obtained is Math Score ~ Lunch Type + Ethnic Group + Test Prep + Gender + Parent Education + Weekly Study Hours.

When using the single-function method, the model obtained with the lowest AIC was Math Score ~ Lunch Type + Ethnic Group + Test Prep + Gender + Parent Education + Weekly Study Hours + Number of Siblings. This method resulted in a model that had a slightly lower MSE by a difference of about 1 point and approximately the same adjusted R-squared values (difference of < 0.3 units).

Reading Score

```
mult.fit = lm(reading_score ~ ., data = reading_df)

### Step 1:  Fit simple linear regressions for all variables,look for the variable with lowest p-value
fit1 = lm(reading_score ~ gender, data = step_df)
summary(fit1)
```

```
##
## Call:
## lm(formula = reading_score ~ gender, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.152 -10.018  -0.152  10.915  33.982
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   80.286      1.888  42.521  < 2e-16 ***
## gender        -7.134      1.221  -5.841  8.6e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 14.76 on 585 degrees of freedom
## Multiple R-squared:  0.05511,    Adjusted R-squared:  0.05349
## F-statistic: 34.12 on 1 and 585 DF,  p-value: 8.604e-09
```

```
fit2 = lm(reading_score ~ ethnic_group, data = step_df)
summary(fit2)
```

```
##
## Call:
## lm(formula = reading_score ~ ethnic_group, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.477 -10.415   0.398  10.523  34.773
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   63.1022     1.7901  35.251  < 2e-16 ***
## ethnic_group   2.1251     0.5293   4.015 6.73e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.97 on 585 degrees of freedom
## Multiple R-squared:  0.02681,    Adjusted R-squared:  0.02515
## F-statistic: 16.12 on 1 and 585 DF,  p-value: 6.732e-05
```

```
fit3 = lm(reading_score ~ parent_educ, data = step_df)
summary(fit3)
```

```
##
## Call:
## lm(formula = reading_score ~ parent_educ, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.850 -10.691   0.363  10.756  34.150
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  63.9035     1.3965   45.76  < 2e-16 ***
## parent_educ   1.9468     0.4107    4.74 2.69e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.9 on 585 degrees of freedom
## Multiple R-squared:  0.03698,    Adjusted R-squared:  0.03534
## F-statistic: 22.47 on 1 and 585 DF,  p-value: 2.688e-06
```

```
fit4 = lm(reading_score ~ lunch_type, data = step_df)
summary(fit4)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.558  -9.558   0.294  10.442  35.442
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   56.410      2.178  25.898  < 2e-16 ***
## lunch_type     8.148      1.269   6.422 2.79e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.67 on 585 degrees of freedom
## Multiple R-squared:  0.06585,    Adjusted R-squared:  0.06425
## F-statistic: 41.24 on 1 and 585 DF,  p-value: 2.791e-10
```

```
fit5 = lm(reading_score ~ parent_marital_status, data = step_df)
summary(fit5)
```

```
##
## Call:
## lm(formula = reading_score ~ parent_marital_status, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.351 -10.195   0.649  11.227  30.805
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            68.6169     2.0306  33.792   <2e-16 ***
## parent_marital_status   0.5780     0.9078   0.637    0.525
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.17 on 585 degrees of freedom
## Multiple R-squared:  0.0006925,  Adjusted R-squared:  -0.001016
## F-statistic: 0.4054 on 1 and 585 DF,  p-value: 0.5246
```

```
fit6 = lm(reading_score ~ practice_sport, data = step_df)
summary(fit6)
```

```
##
## Call:
## lm(formula = reading_score ~ practice_sport, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.008 -10.135   0.739  10.992  30.992
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     72.3876     2.2614  32.011   <2e-16 ***
## practice_sport  -1.1265     0.9634  -1.169    0.243
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.16 on 585 degrees of freedom
## Multiple R-squared:  0.002332,   Adjusted R-squared:  0.0006262
## F-statistic: 1.367 on 1 and 585 DF,  p-value: 0.2428
```

```r
fit7 = lm(reading_score ~ is_first_child, data = step_df)
summary(fit7)
```

```
##
## Call:
## lm(formula = reading_score ~ is_first_child, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.256 -10.256   0.744  11.744  30.995
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      67.755      2.319  29.223   <2e-16 ***
## is_first_child    1.250      1.334   0.937    0.349
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.17 on 585 degrees of freedom
## Multiple R-squared:  0.001499,   Adjusted R-squared:  -0.000208
## F-statistic: 0.8781 on 1 and 585 DF,  p-value: 0.3491
```

```r
fit8 = lm(reading_score ~ nr_siblings, data = step_df)
summary(fit8)
```

```
##
## Call:
## lm(formula = reading_score ~ nr_siblings, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.241 -10.199   0.676  11.217  31.134
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  68.8661     1.1000  62.603   <2e-16 ***
## nr_siblings   0.4583     0.4228   1.084    0.279
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.16 on 585 degrees of freedom
## Multiple R-squared:  0.002004,   Adjusted R-squared:  0.0002984
## F-statistic: 1.175 on 1 and 585 DF,  p-value: 0.2788
```

```r
fit9 = lm(reading_score ~ transport_means, data = step_df)
summary(fit9)
```

```
##
## Call:
## lm(formula = reading_score ~ transport_means, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.989  -9.989   0.376  11.193  30.376
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      69.2601     2.1605  32.057   <2e-16 ***
## transport_means   0.3644     1.2844   0.284    0.777
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.18 on 585 degrees of freedom
## Multiple R-squared:  0.0001376,  Adjusted R-squared:  -0.001572
## F-statistic: 0.08048 on 1 and 585 DF,  p-value: 0.7767
```

```r
fit10 = lm(reading_score ~ wkly_study_hours, data = step_df)
summary(fit10)
```

```
##
## Call:
## lm(formula = reading_score ~ wkly_study_hours, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.127 -10.127   1.053  11.553  30.233
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       67.5878     1.9267   35.08   <2e-16 ***
## wkly_study_hours   1.1797     0.9517    1.24    0.216
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.16 on 585 degrees of freedom
## Multiple R-squared:  0.00262,    Adjusted R-squared:  0.0009147
## F-statistic: 1.537 on 1 and 585 DF,  p-value: 0.2156
```

```r
fit11 = lm(reading_score ~ test_prep, data = step_df)
summary(fit11)
```

```
##
## Call:
## lm(formula = reading_score ~ test_prep, data = step_df)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -50.644  -9.861   1.139  10.748  32.356
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   80.077      2.201   36.38  < 2e-16 ***
## test_prep     -6.217      1.284   -4.84 1.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.88 on 585 degrees of freedom
## Multiple R-squared:  0.03851,    Adjusted R-squared:  0.03686
## F-statistic: 23.43 on 1 and 585 DF,  p-value: 1.662e-06
```

```
# Enter first the one with the lowest p-value: Lunch Type
forward1 = lm(reading_score ~ lunch_type, data = step_df)
summary(forward1)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.558  -9.558   0.294  10.442  35.442
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   56.410      2.178  25.898  < 2e-16 ***
## lunch_type     8.148      1.269   6.422 2.79e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.67 on 585 degrees of freedom
## Multiple R-squared:  0.06585,    Adjusted R-squared:  0.06425
## F-statistic: 41.24 on 1 and 585 DF,  p-value: 2.791e-10
```

```
### Step 2: Enter the one with the lowest p-value in the rest
fit1 = update(forward1, . ~ . +gender)
summary(fit1)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.793  -9.292   0.779  10.207  39.779
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   66.795      2.653  25.175  < 2e-16 ***
## lunch_type     8.570      1.229   6.974 8.37e-12 ***
```

```
## gender        -7.572      1.176  -6.438 2.53e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.19 on 584 degrees of freedom
## Multiple R-squared:  0.1278, Adjusted R-squared:  0.1248
## F-statistic: 42.77 on 2 and 584 DF,  p-value: < 2.2e-16
```

```
fit2 = update(forward1, . ~ . +ethnic_group)
summary(fit2)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + ethnic_group, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.335  -9.333  -0.287  10.664  34.663
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   50.3851     2.6504  19.010  < 2e-16 ***
## lunch_type     7.9566     1.2547   6.342 4.56e-10 ***
## ethnic_group   1.9978     0.5128   3.896 0.000109 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.5 on 584 degrees of freedom
## Multiple R-squared:  0.08951,    Adjusted R-squared:  0.08639
## F-statistic: 28.71 on 2 and 584 DF,  p-value: 1.283e-12
```

```
fit3 = update(forward1, . ~ . +parent_educ)
summary(fit3)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + parent_educ, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.447  -9.436   0.372  10.505  35.595
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  50.2445     2.4668  20.369  < 2e-16 ***
## lunch_type    8.2236     1.2437   6.612 8.57e-11 ***
## parent_educ   1.9788     0.3965   4.990 7.97e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.38 on 584 degrees of freedom
## Multiple R-squared:  0.1041, Adjusted R-squared:  0.101
## F-statistic: 33.91 on 2 and 584 DF,  p-value: 1.165e-14
```

```r
fit4 = update(forward1, . ~ . +parent_marital_status)
summary(fit4)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + parent_marital_status,
##     data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.208  -9.528   0.379  10.606  36.339
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             54.7005     2.9181  18.745  < 2e-16 ***
## lunch_type               8.1864     1.2698   6.447 2.39e-10 ***
## parent_marital_status    0.7737     0.8784   0.881    0.379
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.67 on 584 degrees of freedom
## Multiple R-squared:  0.06709,    Adjusted R-squared:  0.06389
## F-statistic:    21 on 2 and 584 DF,  p-value: 1.561e-09
```

```r
fit5 = update(forward1, . ~ . +practice_sport)
summary(fit5)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + practice_sport, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.875  -9.777   0.064  11.015  36.125
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     58.6184     3.0712  19.086  < 2e-16 ***
## lunch_type       8.1097     1.2693   6.389 3.41e-10 ***
## practice_sport  -0.9510     0.9326  -1.020    0.308
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.67 on 584 degrees of freedom
## Multiple R-squared:  0.06751,    Adjusted R-squared:  0.06432
## F-statistic: 21.14 on 2 and 584 DF,  p-value: 1.368e-09
```

```r
fit6 = update(forward1, . ~ . +is_first_child)
summary(fit6)
```

```
##
## Call:
```

```
## lm(formula = reading_score ~ lunch_type + is_first_child, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.943  -9.943   0.057  10.923  36.206
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      54.512      3.050  17.875  < 2e-16 ***
## lunch_type        8.134      1.269   6.409 3.02e-10 ***
## is_first_child    1.148      1.291   0.890    0.374
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.67 on 584 degrees of freedom
## Multiple R-squared:  0.06711,    Adjusted R-squared:  0.06392
## F-statistic: 21.01 on 2 and 584 DF,  p-value: 1.549e-09
```

```
fit7 = update(forward1, . ~ . +nr_siblings)
summary(fit7)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + nr_siblings, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.950  -9.595  -0.043  10.657  36.410
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  55.4448     2.3455  23.639  < 2e-16 ***
## lunch_type    8.1452     1.2686   6.421 2.81e-10 ***
## nr_siblings   0.4533     0.4090   1.109    0.268
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.67 on 584 degrees of freedom
## Multiple R-squared:  0.06781,    Adjusted R-squared:  0.06462
## F-statistic: 21.24 on 2 and 584 DF,  p-value: 1.245e-09
```

```
fit8 = update(forward1, . ~ . +transport_means)
summary(fit8)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + transport_means, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.665  -9.541   0.190  10.604  35.604
##
## Coefficients:
```

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       55.983      2.942  19.030  < 2e-16 ***
## lunch_type         8.145      1.270   6.413 2.94e-10 ***
## transport_means    0.269      1.243   0.216    0.829
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.68 on 584 degrees of freedom
## Multiple R-squared:  0.06592,    Adjusted R-squared:  0.06273
## F-statistic: 20.61 on 2 and 584 DF,  p-value: 2.246e-09
```

```
fit9 = update(forward1, . ~ . +wkly_study_hours)
summary(fit9)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + wkly_study_hours, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.825  -9.232   0.196  11.031  35.340
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       54.1859     2.7980  19.366  < 2e-16 ***
## lunch_type         8.1438     1.2681   6.422 2.79e-10 ***
## wkly_study_hours   1.1651     0.9206   1.266    0.206
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.66 on 584 degrees of freedom
## Multiple R-squared:  0.06841,    Adjusted R-squared:  0.06522
## F-statistic: 21.44 on 2 and 584 DF,  p-value: 1.033e-09
```

```
fit10 = update(forward1, . ~ . +test_prep)
summary(fit10)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + test_prep, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.142  -9.312   0.517  10.222  31.393
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   66.731      2.907  22.952  < 2e-16 ***
## lunch_type     8.341      1.242   6.717 4.41e-11 ***
## test_prep     -6.466      1.239  -5.218 2.52e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 14.35 on 584 degrees of freedom
## Multiple R-squared:  0.1075, Adjusted R-squared:  0.1044
## F-statistic: 35.16 on 2 and 584 DF,  p-value: 3.829e-15
```

```
# Enter the one with the lowest p-value: Gender
forward2 = update(forward1, . ~ . +gender)
summary(fit2)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + ethnic_group, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.335  -9.333  -0.287  10.664  34.663
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   50.3851     2.6504  19.010  < 2e-16 ***
## lunch_type     7.9566     1.2547   6.342 4.56e-10 ***
## ethnic_group   1.9978     0.5128   3.896 0.000109 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.5 on 584 degrees of freedom
## Multiple R-squared:  0.08951,    Adjusted R-squared:  0.08639
## F-statistic: 28.71 on 2 and 584 DF,  p-value: 1.283e-12
```

```
### Step 3: Enter the one with the lowest p-value in the rest
fit1 = update(forward2, . ~ . +ethnic_group)
summary(fit1)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + ethnic_group,
##     data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.575  -8.989   0.425  10.621  35.996
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   60.7592     3.0128  20.167  < 2e-16 ***
## lunch_type     8.3786     1.2138   6.903 1.34e-11 ***
## gender        -7.5858     1.1609  -6.534 1.39e-10 ***
## ethnic_group   2.0076     0.4955   4.052 5.77e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.01 on 583 degrees of freedom
## Multiple R-squared:  0.1516, Adjusted R-squared:  0.1473
## F-statistic: 34.74 on 3 and 583 DF,  p-value: < 2.2e-16
```

```r
fit2 = update(forward2, . ~ . +parent_educ)
summary(fit2)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + parent_educ,
##     data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.857  -9.511   0.684   9.884  39.712
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  60.6267     2.9148  20.799  < 2e-16 ***
## lunch_type    8.6194     1.2071   7.141 2.77e-12 ***
## gender       -7.2082     1.1578  -6.226 9.16e-10 ***
## parent_educ   1.8194     0.3852   4.724 2.90e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.94 on 583 degrees of freedom
## Multiple R-squared:  0.1599, Adjusted R-squared:  0.1556
## F-statistic: 36.99 on 3 and 583 DF,  p-value: < 2.2e-16
```

```r
fit3 = update(forward2, . ~ . +parent_marital_status)
summary(fit3)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + parent_marital_status,
##     data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.286  -9.439   1.115  10.061  39.860
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            65.4297     3.2818  19.937  < 2e-16 ***
## lunch_type              8.5985     1.2300   6.991 7.52e-12 ***
## gender                 -7.5453     1.1772  -6.410 3.01e-10 ***
## parent_marital_status   0.6012     0.8501   0.707     0.48
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.2 on 583 degrees of freedom
## Multiple R-squared:  0.1285, Adjusted R-squared:  0.124
## F-statistic: 28.65 on 3 and 583 DF,  p-value: < 2.2e-16
```

```r
fit4 = update(forward2, . ~ . +practice_sport)
summary(fit4)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + practice_sport,
##     data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.144  -9.527   0.987   9.973  40.417
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     68.8601     3.3707  20.429  < 2e-16 ***
## lunch_type       8.5334     1.2294   6.941 1.04e-11 ***
## gender          -7.5607     1.1762  -6.428 2.69e-10 ***
## practice_sport  -0.8962     0.9020  -0.994    0.321
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.19 on 583 degrees of freedom
## Multiple R-squared:  0.1292, Adjusted R-squared:  0.1247
## F-statistic: 28.84 on 3 and 583 DF,  p-value: < 2.2e-16
```

```
fit5 = update(forward2, . ~ . +is_first_child)
summary(fit5)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + is_first_child,
##     data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.221  -9.298   0.811  10.295  40.625
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      64.736      3.348  19.337  < 2e-16 ***
## lunch_type        8.556      1.229   6.962 9.07e-12 ***
## gender           -7.588      1.176  -6.451 2.33e-10 ***
## is_first_child    1.259      1.249   1.008    0.314
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.19 on 583 degrees of freedom
## Multiple R-squared:  0.1293, Adjusted R-squared:  0.1248
## F-statistic: 28.85 on 3 and 583 DF,  p-value: < 2.2e-16
```

```
fit6 = update(forward2, . ~ . +nr_siblings)
summary(fit6)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + nr_siblings,
```

```
##     data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.092  -9.393   0.876  10.073  39.442
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  65.9647     2.8018  23.544  < 2e-16 ***
## lunch_type    8.5657     1.2290   6.970 8.62e-12 ***
## gender       -7.5340     1.1770  -6.401 3.17e-10 ***
## nr_siblings   0.3653     0.3959   0.923    0.357
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.19 on 583 degrees of freedom
## Multiple R-squared:  0.129,  Adjusted R-squared:  0.1245
## F-statistic: 28.79 on 3 and 583 DF,  p-value: < 2.2e-16
```

```
fit7 = update(forward2, . ~ . +transport_means)
summary(fit7)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + transport_means,
##     data = step_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -50.95  -9.44   1.01  10.06  40.01
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      66.1997     3.2578  20.320  < 2e-16 ***
## lunch_type        8.5656     1.2299   6.965 8.91e-12 ***
## gender           -7.5770     1.1771  -6.437 2.55e-10 ***
## transport_means   0.3787     1.2018   0.315    0.753
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.2 on 583 degrees of freedom
## Multiple R-squared:  0.1279, Adjusted R-squared:  0.1234
## F-statistic:  28.5 on 3 and 583 DF,  p-value: < 2.2e-16
```

```
fit8 = update(forward2, . ~ . +wkly_study_hours)
summary(fit8)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + wkly_study_hours,
##     data = step_df)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -52.360  -9.194   0.891  10.101  39.701
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       64.2095     3.1118  20.634  < 2e-16 ***
## lunch_type         8.5696     1.2273   6.983 7.91e-12 ***
## gender            -7.6506     1.1757  -6.508 1.65e-10 ***
## wkly_study_hours   1.4107     0.8904   1.584    0.114
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.17 on 583 degrees of freedom
## Multiple R-squared:  0.1315, Adjusted R-squared:  0.127
## F-statistic: 29.42 on 3 and 583 DF,  p-value: < 2.2e-16
```

```
fit9 = update(forward2, . ~ . +test_prep)
summary(fit9)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + test_prep,
##     data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.365  -9.154  -0.154  10.259  35.673
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   78.035      3.251  24.004  < 2e-16 ***
## lunch_type     8.789      1.198   7.339 7.27e-13 ***
## gender        -7.845      1.147  -6.842 1.98e-11 ***
## test_prep     -6.807      1.194  -5.700 1.90e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.82 on 583 degrees of freedom
## Multiple R-squared:  0.1738, Adjusted R-squared:  0.1695
## F-statistic: 40.88 on 3 and 583 DF,  p-value: < 2.2e-16
```

```
# Enter the one with the lowest p-value: Test Prep
forward3 = update(forward2, . ~ . + test_prep)
summary(forward3)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + test_prep,
##     data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.365  -9.154  -0.154  10.259  35.673
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   78.035      3.251  24.004  < 2e-16 ***
## lunch_type     8.789      1.198   7.339 7.27e-13 ***
## gender        -7.845      1.147  -6.842 1.98e-11 ***
## test_prep     -6.807      1.194  -5.700 1.90e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.82 on 583 degrees of freedom
## Multiple R-squared:  0.1738, Adjusted R-squared:  0.1695
## F-statistic: 40.88 on 3 and 583 DF,  p-value: < 2.2e-16
```

```
### Step 4: Enter the one with the lowest p-value in the rest
fit1 = update(forward3, . ~ . +ethnic_group)
summary(fit1)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + test_prep +
##     ethnic_group, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.185  -9.696   0.309  10.143  32.043
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   72.0069     3.5370  20.358  < 2e-16 ***
## lunch_type     8.5998     1.1830   7.269 1.17e-12 ***
## gender        -7.8550     1.1318  -6.940 1.05e-11 ***
## test_prep     -6.7166     1.1790  -5.697 1.94e-08 ***
## ethnic_group   1.9554     0.4827   4.051 5.79e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.64 on 582 degrees of freedom
## Multiple R-squared:  0.1965, Adjusted R-squared:  0.1909
## F-statistic: 35.57 on 4 and 582 DF,  p-value: < 2.2e-16
```

```
fit2 = update(forward3, . ~ . +parent_educ)
summary(fit2)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + test_prep +
##     parent_educ, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.299  -9.218   0.751   9.968  35.537
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  71.9074     3.4168  21.045  < 2e-16 ***
## lunch_type    8.8432     1.1740   7.532 1.91e-13 ***
## gender       -7.4774     1.1264  -6.638 7.29e-11 ***
## test_prep    -6.9179     1.1709  -5.908 5.89e-09 ***
## parent_educ   1.8616     0.3745   4.971 8.78e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.55 on 582 degrees of freedom
## Multiple R-squared:  0.2074, Adjusted R-squared:  0.202
## F-statistic: 38.08 on 4 and 582 DF,  p-value: < 2.2e-16
```

```
fit3 = update(forward3, . ~ . +parent_marital_status)
summary(fit3)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + test_prep +
##     parent_marital_status, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.737  -9.103  -0.278  10.348  35.746
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            76.9929     3.7914  20.307  < 2e-16 ***
## lunch_type              8.8095     1.1989   7.348 6.86e-13 ***
## gender                 -7.8246     1.1480  -6.816 2.34e-11 ***
## test_prep              -6.7856     1.1956  -5.675 2.19e-08 ***
## parent_marital_status   0.4435     0.8287   0.535    0.593
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.83 on 582 degrees of freedom
## Multiple R-squared:  0.1742, Adjusted R-squared:  0.1685
## F-statistic: 30.69 on 4 and 582 DF,  p-value: < 2.2e-16
```

```
fit4 = update(forward3, . ~ . +practice_sport)
summary(fit4)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + test_prep +
##     practice_sport, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.657  -9.459   0.208  10.426  36.350
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     80.2982     3.8445  20.887  < 2e-16 ***
## lunch_type       8.7502     1.1979   7.305 9.20e-13 ***
## gender          -7.8337     1.1465  -6.833 2.10e-11 ***
## test_prep       -6.8260     1.1941  -5.716 1.74e-08 ***
## practice_sport  -0.9684     0.8786  -1.102    0.271
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.82 on 582 degrees of freedom
## Multiple R-squared:  0.1755, Adjusted R-squared:  0.1699
## F-statistic: 30.97 on 4 and 582 DF,  p-value: < 2.2e-16
```

```
fit5 = update(forward3, . ~ . +is_first_child)
summary(fit5)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + test_prep +
##     is_first_child, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.685  -9.416   0.207  10.392  36.301
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      76.499      3.870  19.768  < 2e-16 ***
## lunch_type        8.777      1.198   7.326 7.97e-13 ***
## gender           -7.854      1.147  -6.847 1.92e-11 ***
## test_prep        -6.760      1.196  -5.651 2.50e-08 ***
## is_first_child    0.892      1.218   0.732    0.464
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.83 on 582 degrees of freedom
## Multiple R-squared:  0.1746, Adjusted R-squared:  0.1689
## F-statistic: 30.77 on 4 and 582 DF,  p-value: < 2.2e-16
```

```
fit6 = update(forward3, . ~ . +nr_siblings)
summary(fit6)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + test_prep +
##     nr_siblings, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.590  -9.317  -0.259  10.162  35.459
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  77.3862     3.3955  22.791  < 2e-16 ***
## lunch_type    8.7848     1.1982   7.332 7.65e-13 ***
## gender       -7.8168     1.1480  -6.809 2.45e-11 ***
## test_prep    -6.7676     1.1963  -5.657 2.41e-08 ***
## nr_siblings   0.2569     0.3862   0.665    0.506
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.83 on 582 degrees of freedom
## Multiple R-squared:  0.1744, Adjusted R-squared:  0.1687
## F-statistic: 30.74 on 4 and 582 DF,  p-value: < 2.2e-16
```

```
fit7 = update(forward3, . ~ . +transport_means)
summary(fit7)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + test_prep +
##     transport_means, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.692  -9.409   0.168  10.344  36.162
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      76.7815     3.6696  20.924  < 2e-16 ***
## lunch_type        8.7812     1.1981   7.329 7.78e-13 ***
## gender           -7.8595     1.1473  -6.851 1.88e-11 ***
## test_prep        -6.8710     1.1978  -5.736 1.56e-08 ***
## transport_means   0.8653     1.1733   0.737    0.461
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.83 on 582 degrees of freedom
## Multiple R-squared:  0.1746, Adjusted R-squared:  0.1689
## F-statistic: 30.77 on 4 and 582 DF,  p-value: < 2.2e-16
```

```
fit8 = update(forward3, . ~ . +wkly_study_hours)
summary(fit8)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + test_prep +
##     wkly_study_hours, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.466  -9.300  -0.300   9.722  35.695
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      76.0874     3.7085  20.517  < 2e-16 ***
```

```
## lunch_type         8.7848     1.1974   7.336 7.40e-13 ***
## gender            -7.8932     1.1473  -6.880 1.55e-11 ***
## test_prep         -6.6835     1.1994  -5.572 3.85e-08 ***
## wkly_study_hours   0.9514     0.8722   1.091    0.276
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.82 on 582 degrees of freedom
## Multiple R-squared:  0.1755, Adjusted R-squared:  0.1698
## F-statistic: 30.97 on 4 and 582 DF,  p-value: < 2.2e-16
```

```
# Enter the one with the lowest p-value: Parent Education
forward4 = update(forward3, . ~ . + parent_educ)
summary(forward4)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + test_prep +
##     parent_educ, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.299  -9.218   0.751   9.968  35.537
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  71.9074     3.4168  21.045  < 2e-16 ***
## lunch_type    8.8432     1.1740   7.532 1.91e-13 ***
## gender       -7.4774     1.1264  -6.638 7.29e-11 ***
## test_prep    -6.9179     1.1709  -5.908 5.89e-09 ***
## parent_educ   1.8616     0.3745   4.971 8.78e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.55 on 582 degrees of freedom
## Multiple R-squared:  0.2074, Adjusted R-squared:  0.202
## F-statistic: 38.08 on 4 and 582 DF,  p-value: < 2.2e-16
```

```
### Step 5: Enter the one with the lowest p-value in the rest
fit1 = update(forward4, . ~ . +ethnic_group)
summary(fit1)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + test_prep +
##     parent_educ + ethnic_group, data = step_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 ***
## lunch_type     8.6667     1.1618   7.459 3.18e-13 ***
## gender        -7.5066     1.1139  -6.739 3.84e-11 ***
## test_prep     -6.8289     1.1580  -5.897 6.28e-09 ***
## parent_educ    1.7606     0.3713   4.742 2.66e-06 ***
## ethnic_group   1.7930     0.4753   3.773 0.000178 ***
## ---
## 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
```

```
fit2 = update(forward4, . ~ . +parent_marital_status)
summary(fit2)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + test_prep +
##     parent_educ + parent_marital_status, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.611  -9.180   0.629   9.997  35.598
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            71.0626     3.9046  18.200  < 2e-16 ***
## lunch_type              8.8598     1.1754   7.538 1.85e-13 ***
## gender                 -7.4614     1.1278  -6.616 8.39e-11 ***
## test_prep              -6.9001     1.1724  -5.886 6.71e-09 ***
## parent_educ             1.8582     0.3748   4.958 9.38e-07 ***
## parent_marital_status   0.3641     0.8126   0.448    0.654
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.56 on 581 degrees of freedom
## Multiple R-squared:  0.2077, Adjusted R-squared:  0.2009
## F-statistic: 30.47 on 5 and 581 DF,  p-value: < 2.2e-16
```

```
fit3 = update(forward4, . ~ . +practice_sport)
summary(fit3)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + test_prep +
##     parent_educ + practice_sport, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.833  -9.145   0.826  10.176  36.026
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     73.5975     4.0101  18.353  < 2e-16 ***
## lunch_type       8.8147     1.1749   7.502 2.36e-13 ***
## gender          -7.4732     1.1268  -6.632 7.57e-11 ***
## test_prep       -6.9303     1.1713  -5.917 5.62e-09 ***
## parent_educ      1.8421     0.3754   4.907 1.20e-06 ***
## practice_sport  -0.6958     0.8635  -0.806    0.421
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.55 on 581 degrees of freedom
## Multiple R-squared:  0.2083, Adjusted R-squared:  0.2015
## F-statistic: 30.58 on 5 and 581 DF,  p-value: < 2.2e-16
```

```
fit4 = update(forward4, . ~ . +is_first_child)
summary(fit4)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + test_prep +
##     parent_educ + is_first_child, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.608  -9.313   0.700  10.098  36.135
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     70.4526     3.9845  17.682  < 2e-16 ***
## lunch_type       8.8320     1.1746   7.519 2.10e-13 ***
## gender          -7.4870     1.1270  -6.643 7.06e-11 ***
## test_prep       -6.8734     1.1731  -5.859 7.79e-09 ***
## parent_educ      1.8596     0.3747   4.963 9.11e-07 ***
## is_first_child   0.8488     1.1944   0.711    0.478
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.55 on 581 degrees of freedom
## Multiple R-squared:  0.2081, Adjusted R-squared:  0.2013
## F-statistic: 30.54 on 5 and 581 DF,  p-value: < 2.2e-16
```

```
fit5 = update(forward4, . ~ . +nr_siblings)
summary(fit5)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + test_prep +
##     parent_educ + nr_siblings, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.540  -9.254   0.737  10.246  35.299
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  71.1703     3.5545  20.023  < 2e-16 ***
## lunch_type    8.8386     1.1745   7.526 2.01e-13 ***
## gender       -7.4452     1.1277  -6.602 9.14e-11 ***
## test_prep    -6.8742     1.1727  -5.862 7.69e-09 ***
## parent_educ   1.8659     0.3747   4.980 8.39e-07 ***
## nr_siblings   0.2862     0.3786   0.756     0.45
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.55 on 581 degrees of freedom
## Multiple R-squared:  0.2082, Adjusted R-squared:  0.2014
## F-statistic: 30.56 on 5 and 581 DF,  p-value: < 2.2e-16
```

```
fit6 = update(forward4, . ~ . +transport_means)
summary(fit6)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + test_prep +
##     parent_educ + transport_means, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.611  -9.130   0.694  10.057  35.997
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      70.7390     3.7978  18.626  < 2e-16 ***
## lunch_type        8.8358     1.1746   7.523 2.05e-13 ***
## gender           -7.4916     1.1271  -6.647 6.91e-11 ***
## test_prep        -6.9777     1.1744  -5.941 4.88e-09 ***
## parent_educ       1.8591     0.3747   4.962 9.18e-07 ***
## transport_means   0.8121     1.1502   0.706     0.48
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.55 on 581 degrees of freedom
## Multiple R-squared:  0.2081, Adjusted R-squared:  0.2013
## F-statistic: 30.54 on 5 and 581 DF,  p-value: < 2.2e-16
```

```
fit7 = update(forward4, . ~ . +wkly_study_hours)
summary(fit7)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + test_prep +
##     parent_educ + wkly_study_hours, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.535  -9.074   0.847   9.928  35.562
```

```
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       69.6040     3.8570  18.046  < 2e-16 ***
## lunch_type         8.8388     1.1734   7.533 1.91e-13 ***
## gender            -7.5299     1.1266  -6.684 5.46e-11 ***
## test_prep         -6.7761     1.1754  -5.765 1.33e-08 ***
## parent_educ        1.8781     0.3745   5.015 7.06e-07 ***
## wkly_study_hours   1.0987     0.8552   1.285    0.199
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.54 on 581 degrees of freedom
## Multiple R-squared:  0.2097, Adjusted R-squared:  0.2029
## F-statistic: 30.83 on 5 and 581 DF,  p-value: < 2.2e-16
```

```
# Enter the one with the lowest p-value: Ethnic Group
forward5 = update(forward4, . ~ . + ethnic_group)
summary(forward5)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + test_prep +
##     parent_educ + ethnic_group, data = step_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 ***
## lunch_type     8.6667     1.1618   7.459 3.18e-13 ***
## gender        -7.5066     1.1139  -6.739 3.84e-11 ***
## test_prep     -6.8289     1.1580  -5.897 6.28e-09 ***
## parent_educ    1.7606     0.3713   4.742 2.66e-06 ***
## ethnic_group   1.7930     0.4753   3.773 0.000178 ***
## ---
## 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
```

```
### Step 6: Enter the one with the lowest p-value in the rest
fit1 = update(forward5, . ~ . +parent_marital_status)
summary(fit1)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + test_prep +
##     parent_educ + ethnic_group + parent_marital_status, data = step_df)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.741  -8.668   0.860  10.107  32.276
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            65.6456     4.1173  15.944  < 2e-16 ***
## lunch_type              8.6865     1.1631   7.469 2.99e-13 ***
## gender                 -7.4869     1.1151  -6.714 4.51e-11 ***
## test_prep              -6.8065     1.1594  -5.871 7.31e-09 ***
## parent_educ             1.7561     0.3716   4.726 2.88e-06 ***
## ethnic_group            1.8006     0.4757   3.785  0.00017 ***
## parent_marital_status   0.4502     0.8037   0.560  0.57557
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.4 on 580 degrees of freedom
## Multiple R-squared:  0.2268, Adjusted R-squared:  0.2188
## F-statistic: 28.36 on 6 and 580 DF,  p-value: < 2.2e-16
```

```
fit2 = update(forward5, . ~ . +practice_sport)
summary(fit2)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + test_prep +
##     parent_educ + ethnic_group + practice_sport, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.882  -9.319   0.845  10.001  32.709
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     68.4230     4.1955  16.309  < 2e-16 ***
## lunch_type       8.6377     1.1627   7.429 3.93e-13 ***
## gender          -7.5024     1.1142  -6.734 3.99e-11 ***
## test_prep       -6.8414     1.1584  -5.906 5.98e-09 ***
## parent_educ      1.7408     0.3721   4.678 3.61e-06 ***
## ethnic_group     1.7943     0.4754   3.774 0.000177 ***
## practice_sport  -0.7058     0.8538  -0.827 0.408769
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.4 on 580 degrees of freedom
## Multiple R-squared:  0.2273, Adjusted R-squared:  0.2193
## F-statistic: 28.44 on 6 and 580 DF,  p-value: < 2.2e-16
```

```
fit3 = update(forward5, . ~ . +is_first_child)
summary(fit3)
```

```
##
## Call:
```

```
## lm(formula = reading_score ~ lunch_type + gender + test_prep +
##     parent_educ + ethnic_group + is_first_child, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.665  -9.187   0.886  10.213  32.816
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     65.2492     4.1744  15.631  < 2e-16 ***
## lunch_type       8.6555     1.1624   7.446 3.50e-13 ***
## gender          -7.5162     1.1144  -6.745 3.72e-11 ***
## test_prep       -6.7841     1.1602  -5.848 8.33e-09 ***
## parent_educ      1.7587     0.3714   4.735 2.76e-06 ***
## ethnic_group     1.7934     0.4755   3.772 0.000179 ***
## is_first_child   0.8529     1.1810   0.722 0.470481
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.4 on 580 degrees of freedom
## Multiple R-squared:  0.2271, Adjusted R-squared:  0.2191
## F-statistic:  28.4 on 6 and 580 DF,  p-value: < 2.2e-16
```

```
fit4 = update(forward5, . ~ . +nr_siblings)
summary(fit4)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + test_prep +
##     parent_educ + ethnic_group + nr_siblings, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.626  -8.902   0.850   9.677  32.420
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   65.8499     3.7839  17.403  < 2e-16 ***
## lunch_type     8.6605     1.1621   7.452 3.35e-13 ***
## gender        -7.4705     1.1149  -6.701 4.92e-11 ***
## test_prep     -6.7791     1.1597  -5.845 8.44e-09 ***
## parent_educ    1.7650     0.3714   4.752 2.54e-06 ***
## ethnic_group   1.8036     0.4755   3.793 0.000165 ***
## nr_siblings    0.3228     0.3744   0.862 0.388988
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.4 on 580 degrees of freedom
## Multiple R-squared:  0.2274, Adjusted R-squared:  0.2194
## F-statistic: 28.45 on 6 and 580 DF,  p-value: < 2.2e-16
```

```
fit5 = update(forward5, . ~ . +transport_means)
summary(fit5)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + test_prep +
##     parent_educ + ethnic_group + transport_means, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.717  -8.855   0.790   9.849  32.728
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      65.3191     4.0171  16.260  < 2e-16 ***
## lunch_type        8.6569     1.1622   7.449 3.44e-13 ***
## gender           -7.5233     1.1144  -6.751 3.56e-11 ***
## test_prep        -6.8978     1.1613  -5.940 4.93e-09 ***
## parent_educ       1.7571     0.3714   4.731 2.81e-06 ***
## ethnic_group      1.8051     0.4756   3.795 0.000163 ***
## transport_means   0.9440     1.1377   0.830 0.407036
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.4 on 580 degrees of freedom
## Multiple R-squared:  0.2273, Adjusted R-squared:  0.2193
## F-statistic: 28.44 on 6 and 580 DF,  p-value: < 2.2e-16
```

```
fit6 = update(forward5, . ~ . +wkly_study_hours)
summary(fit6)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + test_prep +
##     parent_educ + ethnic_group + wkly_study_hours, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.551  -8.822   0.863   9.721  32.252
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       64.5007     4.0483  15.933  < 2e-16 ***
## lunch_type         8.6631     1.1613   7.460 3.18e-13 ***
## gender            -7.5573     1.1140  -6.784 2.90e-11 ***
## test_prep         -6.6919     1.1626  -5.756 1.39e-08 ***
## parent_educ        1.7770     0.3713   4.786 2.16e-06 ***
## ethnic_group       1.7865     0.4751   3.761 0.000187 ***
## wkly_study_hours   1.0638     0.8457   1.258 0.208928
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.39 on 580 degrees of freedom
## Multiple R-squared:  0.2285, Adjusted R-squared:  0.2205
## F-statistic: 28.63 on 6 and 580 DF,  p-value: < 2.2e-16
```

```r
# P-value of all new added variables are larger than 0.05, which means that they
# are not significant predictor, and we stop here.

# The model we obtained is Reading Score ~ Lunch Type + Gender + Test Prep +
# Parent Education + Ethnic Group

reading_forward_manual_fit = lm(reading_score ~ lunch_type + gender + test_prep +
                      parent_educ + ethnic_group, data = step_df)
summary(reading_forward_manual_fit)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + test_prep +
##     parent_educ + ethnic_group, data = step_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 ***
## lunch_type     8.6667     1.1618   7.459 3.18e-13 ***
## gender        -7.5066     1.1139  -6.739 3.84e-11 ***
## test_prep     -6.8289     1.1580  -5.897 6.28e-09 ***
## parent_educ    1.7606     0.3713   4.742 2.66e-06 ***
## ethnic_group   1.7930     0.4753   3.773 0.000178 ***
## ---
## 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
```

```r
mean(reading_forward_manual_fit$residuals^2)
```

```
## [1] 177.6469
```

```r
# fit using one function
intercept_only <- lm (reading_score ~ 1, data = reading_df)
reading_forward_func_fit = step(intercept_only, direction = "forward", scope = formula(mult.fit))
```

```
## Start:  AIC=3193.22
## reading_score ~ 1
##
##                        Df Sum of Sq    RSS    AIC
## + lunch_type            1    8876.3 125920 3155.2
## + gender                1    7428.6 127368 3161.9
## + test_prep             1    5190.3 129606 3172.2
## + parent_educ           1    4985.3 129811 3173.1
## + ethnic_group          1    3614.0 131182 3179.3
## <none>                              134796 3193.2
```

```
## + wkly_study_hours       1     353.1 134443 3193.7
## + practice_sport         1     314.3 134482 3193.9
## + nr_siblings            1     270.2 134526 3194.0
## + is_first_child         1     202.0 134594 3194.3
## + parent_marital_status  1      93.3 134703 3194.8
## + transport_means        1      18.5 134778 3195.1
##
## Step:  AIC=3155.24
## reading_score ~ lunch_type
##
##                          Df Sum of Sq    RSS    AIC
## + gender                  1    8344.3 117576 3117.0
## + test_prep               1    5609.0 120311 3130.5
## + parent_educ             1    5149.8 120770 3132.7
## + ethnic_group            1    3189.3 122731 3142.2
## <none>                                 125920 3155.2
## + wkly_study_hours        1     344.5 125575 3155.6
## + nr_siblings             1     264.4 125655 3156.0
## + practice_sport          1     223.8 125696 3156.2
## + is_first_child          1     170.4 125749 3156.4
## + parent_marital_status   1     167.1 125753 3156.5
## + transport_means         1      10.1 125910 3157.2
##
## Step:  AIC=3116.99
## reading_score ~ lunch_type + gender
##
##                          Df Sum of Sq    RSS    AIC
## + test_prep               1    6206.4 111369 3087.2
## + parent_educ             1    4334.5 113241 3096.9
## + ethnic_group            1    3220.4 114355 3102.7
## + wkly_study_hours        1     504.0 117072 3116.5
## <none>                                 117576 3117.0
## + is_first_child          1     204.6 117371 3118.0
## + practice_sport          1     198.7 117377 3118.0
## + nr_siblings             1     171.4 117404 3118.1
## + parent_marital_status   1     100.8 117475 3118.5
## + transport_means         1      20.0 117556 3118.9
##
## Step:  AIC=3087.16
## reading_score ~ lunch_type + gender + test_prep
##
##                          Df Sum of Sq    RSS    AIC
## + parent_educ             1    4535.8 106833 3064.8
## + ethnic_group            1    3054.3 108315 3072.8
## <none>                                 111369 3087.2
## + practice_sport          1     232.0 111137 3087.9
## + wkly_study_hours        1     227.2 111142 3088.0
## + transport_means         1     104.0 111265 3088.6
## + is_first_child          1     102.5 111267 3088.6
## + nr_siblings             1      84.6 111285 3088.7
## + parent_marital_status   1      54.8 111314 3088.9
##
## Step:  AIC=3064.75
## reading_score ~ lunch_type + gender + test_prep + parent_educ
```

```
##
##                         Df Sum of Sq    RSS    AIC
## + ethnic_group           1   2554.69 104279 3052.5
## <none>                               106833 3064.8
## + wkly_study_hours       1    302.64 106531 3065.1
## + practice_sport         1    119.26 106714 3066.1
## + nr_siblings            1    104.94 106728 3066.2
## + is_first_child         1     92.77 106741 3066.2
## + transport_means        1     91.59 106742 3066.2
## + parent_marital_status  1     36.91 106796 3066.6
##
## Step:  AIC=3052.54
## reading_score ~ lunch_type + gender + test_prep + parent_educ +
##     ethnic_group
##
##                         Df Sum of Sq    RSS    AIC
## <none>                               104279 3052.5
## + wkly_study_hours       1   283.719 103995 3052.9
## + nr_siblings            1   133.453 104145 3053.8
## + transport_means        1   123.629 104155 3053.8
## + practice_sport         1   122.719 104156 3053.8
## + is_first_child         1    93.682 104185 3054.0
## + parent_marital_status  1    56.388 104222 3054.2
```

```
summary(reading_forward_func_fit)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + test_prep +
##     parent_educ + ethnic_group, 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 ***
## lunch_type     8.6667     1.1618   7.459 3.18e-13 ***
## gender        -7.5066     1.1139  -6.739 3.84e-11 ***
## test_prep     -6.8289     1.1580  -5.897 6.28e-09 ***
## parent_educ    1.7606     0.3713   4.742 2.66e-06 ***
## ethnic_group   1.7930     0.4753   3.773 0.000178 ***
## ---
## 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
```

```
mean(reading_forward_func_fit$residuals^2)
```

```
## [1] 177.6469
```

The model we obtained is Reading Score ~ Lunch Type + Gender + Test Prep + Parent Education + Ethnic Group.

When using the single-function method, the model obtained with the lowest AIC was Reading Score ~ Lunch Type + Gender + Test Prep + Parent Education + Ethnic Group. Both models have equal MSE and adjusted R-squared values.

Writing Score

```
mult.fit = lm(writing_score ~ ., data = writing_df)

### Step 1:  Fit simple linear regressions for all variables,look for the variable with lowest p-value
fit1 = lm(writing_score ~ gender, data = step_df)
summary(fit1)
```

```
##
## Call:
## lm(formula = writing_score ~ gender, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -62.943 -10.221   1.057  11.057  35.779
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   81.665      1.912  42.714  < 2e-16 ***
## gender        -8.722      1.237  -7.053 4.96e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.94 on 585 degrees of freedom
## Multiple R-squared:  0.07837,    Adjusted R-squared:  0.07679
## F-statistic: 49.74 on 1 and 585 DF,  p-value: 4.964e-12
```

```
fit2 = lm(writing_score ~ ethnic_group, data = step_df)
summary(fit2)
```

```
##
## Call:
## lm(formula = writing_score ~ ethnic_group, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.452 -10.868   0.548  11.548  33.716
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   60.6995     1.8250  33.259  < 2e-16 ***
## ethnic_group   2.5842     0.5397   4.788 2.13e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.27 on 585 degrees of freedom
## Multiple R-squared:  0.03772,    Adjusted R-squared:  0.03607
## F-statistic: 22.93 on 1 and 585 DF,  p-value: 2.132e-06
```

```r
fit3 = lm(writing_score ~ parent_educ, data = step_df)
summary(fit3)
```

```
##
## Call:
## lm(formula = writing_score ~ parent_educ, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.707 -10.533   0.172  11.232  36.293
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  61.1771     1.4150   43.23  < 2e-16 ***
## parent_educ   2.5301     0.4162    6.08 2.18e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.09 on 585 degrees of freedom
## Multiple R-squared:  0.05943,    Adjusted R-squared:  0.05782
## F-statistic: 36.96 on 1 and 585 DF,  p-value: 2.176e-09
```

```r
fit4 = lm(writing_score ~ lunch_type, data = step_df)
summary(fit4)
```

```
##
## Call:
## lm(formula = writing_score ~ lunch_type, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.165  -9.584   0.835  10.997  36.835
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   54.327      2.224  24.430  < 2e-16 ***
## lunch_type     8.838      1.295   6.822 2.24e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.98 on 585 degrees of freedom
## Multiple R-squared:  0.0737, Adjusted R-squared:  0.07212
## F-statistic: 46.54 on 1 and 585 DF,  p-value: 2.242e-11
```

```r
fit5 = lm(writing_score ~ parent_marital_status, data = step_df)
summary(fit5)
```

```
##
## Call:
## lm(formula = writing_score ~ parent_marital_status, data = step_df)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -59.767 -10.774   0.226  10.726  32.218
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            66.7887     2.0806  32.101   <2e-16 ***
## parent_marital_status   0.9928     0.9301   1.067    0.286
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.55 on 585 degrees of freedom
## Multiple R-squared:  0.001944,   Adjusted R-squared:  0.0002376
## F-statistic: 1.139 on 1 and 585 DF,  p-value: 0.2862
```

```
fit6 = lm(writing_score ~ practice_sport, data = step_df)
summary(fit6)
```

```
##
## Call:
## lm(formula = writing_score ~ practice_sport, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.898 -10.902   0.102  10.102  31.102
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    68.91193    2.32122  29.688   <2e-16 ***
## practice_sport -0.00476    0.98893  -0.005    0.996
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.56 on 585 degrees of freedom
## Multiple R-squared:  3.96e-08,   Adjusted R-squared:  -0.001709
## F-statistic: 2.317e-05 on 1 and 585 DF,  p-value: 0.9962
```

```
fit7 = lm(writing_score ~ is_first_child, data = step_df)
summary(fit7)
```

```
##
## Call:
## lm(formula = writing_score ~ is_first_child, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.182 -11.182  -0.182  10.677  31.677
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     67.4636     2.3781  28.368   <2e-16 ***
## is_first_child   0.8594     1.3688   0.628     0.53
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 15.56 on 585 degrees of freedom
## Multiple R-squared:  0.0006734,  Adjusted R-squared:  -0.001035
## F-statistic: 0.3942 on 1 and 585 DF,  p-value: 0.5304
```

```
fit8 = lm(writing_score ~ nr_siblings, data = step_df)
summary(fit8)
```

```
##
## Call:
## lm(formula = writing_score ~ nr_siblings, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.469 -10.639   0.531  10.851  32.511
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   67.489      1.127  59.898   <2e-16 ***
## nr_siblings    0.660      0.433   1.524    0.128
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.53 on 585 degrees of freedom
## Multiple R-squared:  0.003955,   Adjusted R-squared:  0.002252
## F-statistic: 2.323 on 1 and 585 DF,  p-value: 0.128
```

```
fit9 = lm(writing_score ~ transport_means, data = step_df)
summary(fit9)
```

```
##
## Call:
## lm(formula = writing_score ~ transport_means, data = step_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -58.93 -10.86   0.14  10.14  31.14
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     68.79315    2.21527  31.054   <2e-16 ***
## transport_means  0.06711    1.31692   0.051    0.959
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.56 on 585 degrees of freedom
## Multiple R-squared:  4.439e-06,  Adjusted R-squared:  -0.001705
## F-statistic: 0.002597 on 1 and 585 DF,  p-value: 0.9594
```

```
fit10 = lm(writing_score ~ wkly_study_hours, data = step_df)
summary(fit10)
```

```
##
## Call:
## lm(formula = writing_score ~ wkly_study_hours, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.381 -10.654   0.346  10.346  32.346
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       66.2897     1.9747  33.569   <2e-16 ***
## wkly_study_hours   1.3638     0.9754   1.398    0.163
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.54 on 585 degrees of freedom
## Multiple R-squared:  0.003331,   Adjusted R-squared:  0.001627
## F-statistic: 1.955 on 1 and 585 DF,  p-value: 0.1626
```

```
fit11 = lm(writing_score ~ test_prep, data = step_df)
summary(fit11)
```

```
##
## Call:
## lm(formula = writing_score ~ test_prep, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.902 -10.365   1.098  10.635  34.098
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   82.828      2.222  37.276  < 2e-16 ***
## test_prep     -8.463      1.297  -6.527 1.46e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.03 on 585 degrees of freedom
## Multiple R-squared:  0.06788,    Adjusted R-squared:  0.06629
## F-statistic:  42.6 on 1 and 585 DF,  p-value: 1.455e-10
```

```
# Enter first the one with the lowest p-value: Gender
forward1 = lm(writing_score ~ gender, data = step_df)
summary(forward1)
```

```
##
## Call:
## lm(formula = writing_score ~ gender, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -62.943 -10.221   1.057  11.057  35.779
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   81.665      1.912  42.714  < 2e-16 ***
## gender        -8.722      1.237  -7.053 4.96e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.94 on 585 degrees of freedom
## Multiple R-squared:  0.07837,    Adjusted R-squared:  0.07679
## F-statistic: 49.74 on 1 and 585 DF,  p-value: 4.964e-12
```

```
### Step 2: Enter the one with the lowest p-value in the rest
fit1 = update(forward1, . ~ . +ethnic_group)
summary(fit1)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + ethnic_group, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -62.505  -9.545   0.098  10.042  31.042
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   73.4485     2.4855  29.550  < 2e-16 ***
## gender        -8.7532     1.2118  -7.223 1.59e-12 ***
## ethnic_group   2.6032     0.5175   5.030 6.53e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.64 on 584 degrees of freedom
## Multiple R-squared:  0.1166, Adjusted R-squared:  0.1136
## F-statistic: 38.56 on 2 and 584 DF,  p-value: < 2.2e-16
```

```
fit2 = update(forward1, . ~ . +parent_educ)
summary(fit2)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + parent_educ, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.908  -9.667   1.092  10.506  32.092
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  73.8127     2.2951  32.162  < 2e-16 ***
## gender       -8.2503     1.2058  -6.842 1.98e-11 ***
## parent_educ   2.3460     0.4017   5.840 8.65e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 14.54 on 584 degrees of freedom
## Multiple R-squared:  0.1292, Adjusted R-squared:  0.1262
## F-statistic: 43.33 on 2 and 584 DF,  p-value: < 2.2e-16
```

```
fit3 = update(forward1, . ~ . +parent_marital_status)
summary(fit3)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + parent_marital_status,
##     data = step_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -63.61 -10.14   0.86  11.17  35.08
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            79.9463     2.7411  29.166  < 2e-16 ***
## gender                 -8.6860     1.2376  -7.018 6.25e-12 ***
## parent_marital_status   0.7829     0.8945   0.875    0.382
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.94 on 584 degrees of freedom
## Multiple R-squared:  0.07958,    Adjusted R-squared:  0.07642
## F-statistic: 25.25 on 2 and 584 DF,  p-value: 3.051e-11
```

```
fit4 = update(forward1, . ~ . +practice_sport)
summary(fit4)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + practice_sport, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -62.979 -10.208   1.021  11.069  35.744
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    81.55789    2.86260  28.491  < 2e-16 ***
## gender         -8.72276    1.23778  -7.047 5.17e-12 ***
## practice_sport  0.04786    0.95023   0.050     0.96
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.95 on 584 degrees of freedom
## Multiple R-squared:  0.07837,    Adjusted R-squared:  0.07522
## F-statistic: 24.83 on 2 and 584 DF,  p-value: 4.469e-11
```

```r
fit5 = update(forward1, . ~ . +is_first_child)
summary(fit5)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + is_first_child, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.274 -10.538   0.726  10.726  35.462
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     80.0245     2.8956  27.636  < 2e-16 ***
## gender          -8.7356     1.2373  -7.060 4.73e-12 ***
## is_first_child   0.9924     1.3151   0.755    0.451
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.95 on 584 degrees of freedom
## Multiple R-squared:  0.07927,    Adjusted R-squared:  0.07611
## F-statistic: 25.14 on 2 and 584 DF,  p-value: 3.366e-11
```

```r
fit6 = update(forward1, . ~ . +nr_siblings)
summary(fit6)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + nr_siblings, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.397 -10.338   0.827  11.044  33.032
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  80.3850     2.1355  37.643  < 2e-16 ***
## gender       -8.6648     1.2366  -7.007 6.73e-12 ***
## nr_siblings   0.5590     0.4165   1.342     0.18
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.93 on 584 degrees of freedom
## Multiple R-squared:  0.0812, Adjusted R-squared:  0.07806
## F-statistic: 25.81 on 2 and 584 DF,  p-value: 1.821e-11
```

```r
fit7 = update(forward1, . ~ . +transport_means)
summary(fit7)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + transport_means, data = step_df)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.022 -10.197   0.978  10.978  35.902
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      81.3488     2.7755  29.310  < 2e-16 ***
## gender           -8.7252     1.2379  -7.049 5.12e-12 ***
## transport_means   0.1991     1.2655   0.157    0.875
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.95 on 584 degrees of freedom
## Multiple R-squared:  0.07841,    Adjusted R-squared:  0.07525
## F-statistic: 24.84 on 2 and 584 DF,  p-value: 4.419e-11
```

```
fit8 = update(forward1, . ~ . +wkly_study_hours)
summary(fit8)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + wkly_study_hours, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.774  -9.664   0.874  10.874  35.689
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       78.6452     2.5677  30.629  < 2e-16 ***
## gender            -8.8145     1.2356  -7.134  2.9e-12 ***
## wkly_study_hours   1.6476     0.9371   1.758   0.0793 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.91 on 584 degrees of freedom
## Multiple R-squared:  0.08322,    Adjusted R-squared:  0.08008
## F-statistic: 26.51 on 2 and 584 DF,  p-value: 9.578e-12
```

```
fit9 = update(forward1, . ~ . +test_prep)
summary(fit9)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + test_prep, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.967  -9.747   1.033   9.143  32.095
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   96.714      2.795  34.605  < 2e-16 ***
## gender        -9.063      1.188  -7.629  9.7e-14 ***
## test_prep     -8.842      1.239  -7.139  2.8e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.34 on 584 degrees of freedom
## Multiple R-squared:  0.1523, Adjusted R-squared:  0.1494
## F-statistic: 52.48 on 2 and 584 DF,  p-value: < 2.2e-16
```

```
fit10 = update(forward1, . ~ . +lunch_type)
summary(fit10)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + lunch_type, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.095  -9.170   0.754   9.754  37.105
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   66.944      2.668  25.087  < 2e-16 ***
## gender        -9.200      1.183  -7.777 3.37e-14 ***
## lunch_type     9.351      1.236   7.566 1.51e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.27 on 584 degrees of freedom
## Multiple R-squared:  0.1606, Adjusted R-squared:  0.1578
## F-statistic: 55.88 on 2 and 584 DF,  p-value: < 2.2e-16
```

```
# Enter the one with the lowest p-value: Lunch Type
forward2 = update(forward1, . ~ . +lunch_type)
summary(fit2)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + parent_educ, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.908  -9.667   1.092  10.506  32.092
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  73.8127     2.2951  32.162  < 2e-16 ***
## gender       -8.2503     1.2058  -6.842 1.98e-11 ***
## parent_educ   2.3460     0.4017   5.840 8.65e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 14.54 on 584 degrees of freedom
## Multiple R-squared:  0.1292, Adjusted R-squared:  0.1262
## F-statistic: 43.33 on 2 and 584 DF,  p-value: < 2.2e-16
```

```
### Step 3: Enter the one with the lowest p-value in the rest
fit1 = update(forward1, . ~ . +ethnic_group)
summary(fit1)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + ethnic_group, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -62.505  -9.545   0.098  10.042  31.042
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   73.4485     2.4855  29.550  < 2e-16 ***
## gender        -8.7532     1.2118  -7.223 1.59e-12 ***
## ethnic_group   2.6032     0.5175   5.030 6.53e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.64 on 584 degrees of freedom
## Multiple R-squared:  0.1166, Adjusted R-squared:  0.1136
## F-statistic: 38.56 on 2 and 584 DF,  p-value: < 2.2e-16
```

```
fit2 = update(forward1, . ~ . +parent_educ)
summary(fit2)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + parent_educ, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.908  -9.667   1.092  10.506  32.092
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  73.8127     2.2951  32.162  < 2e-16 ***
## gender       -8.2503     1.2058  -6.842 1.98e-11 ***
## parent_educ   2.3460     0.4017   5.840 8.65e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.54 on 584 degrees of freedom
## Multiple R-squared:  0.1292, Adjusted R-squared:  0.1262
## F-statistic: 43.33 on 2 and 584 DF,  p-value: < 2.2e-16
```

```r
fit3 = update(forward1, . ~ . +parent_marital_status)
summary(fit3)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + parent_marital_status,
##     data = step_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -63.61 -10.14   0.86  11.17  35.08
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            79.9463     2.7411  29.166  < 2e-16 ***
## gender                 -8.6860     1.2376  -7.018 6.25e-12 ***
## parent_marital_status   0.7829     0.8945   0.875    0.382
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.94 on 584 degrees of freedom
## Multiple R-squared:  0.07958,    Adjusted R-squared:  0.07642
## F-statistic: 25.25 on 2 and 584 DF,  p-value: 3.051e-11
```

```r
fit4 = update(forward1, . ~ . +practice_sport)
summary(fit4)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + practice_sport, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -62.979 -10.208   1.021  11.069  35.744
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    81.55789    2.86260  28.491  < 2e-16 ***
## gender         -8.72276    1.23778  -7.047 5.17e-12 ***
## practice_sport  0.04786    0.95023   0.050     0.96
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.95 on 584 degrees of freedom
## Multiple R-squared:  0.07837,    Adjusted R-squared:  0.07522
## F-statistic: 24.83 on 2 and 584 DF,  p-value: 4.469e-11
```

```r
fit5 = update(forward1, . ~ . +is_first_child)
summary(fit5)
```

```
##
## Call:
```

```
## lm(formula = writing_score ~ gender + is_first_child, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.274 -10.538   0.726  10.726  35.462
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     80.0245     2.8956  27.636  < 2e-16 ***
## gender          -8.7356     1.2373  -7.060 4.73e-12 ***
## is_first_child   0.9924     1.3151   0.755    0.451
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.95 on 584 degrees of freedom
## Multiple R-squared:  0.07927,    Adjusted R-squared:  0.07611
## F-statistic: 25.14 on 2 and 584 DF,  p-value: 3.366e-11
```

```
fit6 = update(forward1, . ~ . +nr_siblings)
summary(fit6)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + nr_siblings, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.397 -10.338   0.827  11.044  33.032
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  80.3850     2.1355  37.643  < 2e-16 ***
## gender       -8.6648     1.2366  -7.007 6.73e-12 ***
## nr_siblings   0.5590     0.4165   1.342     0.18
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.93 on 584 degrees of freedom
## Multiple R-squared:  0.0812, Adjusted R-squared:  0.07806
## F-statistic: 25.81 on 2 and 584 DF,  p-value: 1.821e-11
```

```
fit7 = update(forward1, . ~ . +transport_means)
summary(fit7)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + transport_means, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.022 -10.197   0.978  10.978  35.902
##
## Coefficients:
```

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      81.3488     2.7755  29.310  < 2e-16 ***
## gender           -8.7252     1.2379  -7.049 5.12e-12 ***
## transport_means   0.1991     1.2655   0.157    0.875
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.95 on 584 degrees of freedom
## Multiple R-squared:  0.07841,    Adjusted R-squared:  0.07525
## F-statistic: 24.84 on 2 and 584 DF,  p-value: 4.419e-11
```

```
fit8 = update(forward1, . ~ . +wkly_study_hours)
summary(fit8)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + wkly_study_hours, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.774  -9.664   0.874  10.874  35.689
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       78.6452     2.5677  30.629  < 2e-16 ***
## gender            -8.8145     1.2356  -7.134  2.9e-12 ***
## wkly_study_hours   1.6476     0.9371   1.758   0.0793 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.91 on 584 degrees of freedom
## Multiple R-squared:  0.08322,    Adjusted R-squared:  0.08008
## F-statistic: 26.51 on 2 and 584 DF,  p-value: 9.578e-12
```

```
fit9 = update(forward1, . ~ . +test_prep)
summary(fit9)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + test_prep, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.967  -9.747   1.033   9.143  32.095
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   96.714      2.795  34.605  < 2e-16 ***
## gender        -9.063      1.188  -7.629  9.7e-14 ***
## test_prep     -8.842      1.239  -7.139  2.8e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 14.34 on 584 degrees of freedom
## Multiple R-squared:  0.1523, Adjusted R-squared:  0.1494
## F-statistic: 52.48 on 2 and 584 DF,  p-value: < 2.2e-16
```

```
# Enter the one with the lowest p-value: Test Prep
forward3 = update(forward2, . ~ . + test_prep)
summary(forward3)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + lunch_type + test_prep,
##     data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.831  -8.831   0.523  10.523  31.585
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   82.055      3.197  25.665  < 2e-16 ***
## gender        -9.567      1.128  -8.484  < 2e-16 ***
## lunch_type     9.645      1.178   8.189 1.67e-15 ***
## test_prep     -9.151      1.174  -7.791 3.05e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.59 on 583 degrees of freedom
## Multiple R-squared:  0.2398, Adjusted R-squared:  0.2359
## F-statistic:  61.3 on 3 and 583 DF,  p-value: < 2.2e-16
```

```
### Step 4: Enter the one with the lowest p-value in the rest
fit1 = update(forward3, . ~ . +ethnic_group)
summary(fit1)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + lunch_type + test_prep +
##     ethnic_group, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.611  -8.743   0.158   9.306  30.331
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   74.6928     3.4518  21.639  < 2e-16 ***
## gender        -9.5795     1.1045  -8.673  < 2e-16 ***
## lunch_type     9.4139     1.1545   8.154 2.17e-15 ***
## test_prep     -9.0404     1.1506  -7.857 1.91e-14 ***
## ethnic_group   2.3883     0.4711   5.070 5.36e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 13.31 on 582 degrees of freedom
## Multiple R-squared:  0.2719, Adjusted R-squared:  0.2669
## F-statistic: 54.35 on 4 and 582 DF,  p-value: < 2.2e-16
```

```
fit2 = update(forward3, . ~ . +parent_educ)
summary(fit2)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + lunch_type + test_prep +
##     parent_educ, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.527  -8.462   1.122   9.869  31.408
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  74.0611     3.3057  22.404  < 2e-16 ***
## gender       -9.0877     1.0898  -8.339 5.44e-16 ***
## lunch_type    9.7157     1.1358   8.554  < 2e-16 ***
## test_prep    -9.2954     1.1328  -8.206 1.48e-15 ***
## parent_educ   2.4286     0.3623   6.703 4.83e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.11 on 582 degrees of freedom
## Multiple R-squared:  0.2943, Adjusted R-squared:  0.2894
## F-statistic: 60.67 on 4 and 582 DF,  p-value: < 2.2e-16
```

```
fit3 = update(forward3, . ~ . +parent_marital_status)
summary(fit3)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + lunch_type + test_prep +
##     parent_marital_status, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.489  -8.733   0.295  10.431  31.714
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            80.2119     3.7267  21.524  < 2e-16 ***
## gender                 -9.5312     1.1284  -8.447 2.40e-16 ***
## lunch_type              9.6811     1.1785   8.215 1.38e-15 ***
## test_prep              -9.1130     1.1752  -7.754 4.00e-14 ***
## parent_marital_status   0.7844     0.8146   0.963    0.336
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.59 on 582 degrees of freedom
```

```
## Multiple R-squared:  0.241,  Adjusted R-squared:  0.2358
## F-statistic:  46.2 on 4 and 582 DF,  p-value: < 2.2e-16
```

```
fit4 = update(forward3, . ~ . +practice_sport)
summary(fit4)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + lunch_type + test_prep +
##     practice_sport, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.946  -8.831   0.403  10.517  31.475
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     81.6893     3.7848  21.584  < 2e-16 ***
## gender          -9.5689     1.1287  -8.478  < 2e-16 ***
## lunch_type       9.6514     1.1793   8.184 1.74e-15 ***
## test_prep       -9.1478     1.1756  -7.781 3.29e-14 ***
## practice_sport   0.1566     0.8649   0.181    0.856
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.6 on 582 degrees of freedom
## Multiple R-squared:  0.2398, Adjusted R-squared:  0.2346
## F-statistic: 45.91 on 4 and 582 DF,  p-value: < 2.2e-16
```

```
fit5 = update(forward3, . ~ . +is_first_child)
summary(fit5)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + lunch_type + test_prep +
##     is_first_child, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.971  -8.971   0.389  10.403  31.857
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     81.3888     3.8074  21.376  < 2e-16 ***
## gender          -9.5713     1.1286  -8.480  < 2e-16 ***
## lunch_type       9.6400     1.1788   8.178 1.82e-15 ***
## test_prep       -9.1306     1.1771  -7.757 3.91e-14 ***
## is_first_child   0.3871     1.1987   0.323    0.747
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.6 on 582 degrees of freedom
## Multiple R-squared:  0.2399, Adjusted R-squared:  0.2347
## F-statistic: 45.93 on 4 and 582 DF,  p-value: < 2.2e-16
```

```r
fit6 = update(forward3, . ~ . +nr_siblings)
summary(fit6)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + lunch_type + test_prep +
##     nr_siblings, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.183  -8.822   0.580  10.432  31.251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  81.0394     3.3375  24.282  < 2e-16 ***
## gender       -9.5231     1.1283  -8.440 2.53e-16 ***
## lunch_type    9.6385     1.1777   8.184 1.74e-15 ***
## test_prep    -9.0891     1.1758  -7.730 4.75e-14 ***
## nr_siblings   0.4023     0.3796   1.060     0.29
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.59 on 582 degrees of freedom
## Multiple R-squared:  0.2413, Adjusted R-squared:  0.236
## F-statistic: 46.26 on 4 and 582 DF,  p-value: < 2.2e-16
```

```r
fit7 = update(forward3, . ~ . +transport_means)
summary(fit7)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + lunch_type + test_prep +
##     transport_means, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.114  -8.753   0.370  10.598  32.008
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      80.9705     3.6093  22.434  < 2e-16 ***
## gender           -9.5797     1.1284  -8.490  < 2e-16 ***
## lunch_type        9.6383     1.1784   8.179 1.81e-15 ***
## test_prep        -9.2062     1.1782  -7.814 2.60e-14 ***
## transport_means   0.7488     1.1540   0.649    0.517
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.6 on 582 degrees of freedom
## Multiple R-squared:  0.2403, Adjusted R-squared:  0.2351
## F-statistic: 46.03 on 4 and 582 DF,  p-value: < 2.2e-16
```

```r
fit8 = update(forward3, . ~ . +wkly_study_hours)
summary(fit8)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + lunch_type + test_prep +
##     wkly_study_hours, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.020  -8.634   0.366  10.344  31.608
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       79.9537     3.6464  21.927  < 2e-16 ***
## gender            -9.6192     1.1281  -8.527  < 2e-16 ***
## lunch_type         9.6405     1.1774   8.188 1.69e-15 ***
## test_prep         -9.0175     1.1793  -7.646 8.60e-14 ***
## wkly_study_hours   1.0266     0.8576   1.197    0.232
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.59 on 582 degrees of freedom
## Multiple R-squared:  0.2417, Adjusted R-squared:  0.2364
## F-statistic: 46.37 on 4 and 582 DF,  p-value: < 2.2e-16
```

```r
# Enter the one with the lowest p-value: Parent Education
forward4 = update(forward3, . ~ . + parent_educ)
summary(forward4)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + lunch_type + test_prep +
##     parent_educ, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.527  -8.462   1.122   9.869  31.408
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  74.0611     3.3057  22.404  < 2e-16 ***
## gender       -9.0877     1.0898  -8.339 5.44e-16 ***
## lunch_type    9.7157     1.1358   8.554  < 2e-16 ***
## test_prep    -9.2954     1.1328  -8.206 1.48e-15 ***
## parent_educ   2.4286     0.3623   6.703 4.83e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.11 on 582 degrees of freedom
## Multiple R-squared:  0.2943, Adjusted R-squared:  0.2894
## F-statistic: 60.67 on 4 and 582 DF,  p-value: < 2.2e-16
```

```r
### Step 5: Enter the one with the lowest p-value in the rest
fit1 = update(forward4, . ~ . +ethnic_group)
summary(fit1)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + lunch_type + test_prep +
##     parent_educ + ethnic_group, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.594  -8.422   0.710   9.201  29.415
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   67.7575     3.5050  19.332  < 2e-16 ***
## gender        -9.1231     1.0701  -8.526  < 2e-16 ***
## lunch_type     9.5016     1.1161   8.513  < 2e-16 ***
## test_prep     -9.1875     1.1125  -8.258 9.99e-16 ***
## parent_educ    2.3061     0.3567   6.466 2.14e-10 ***
## ethnic_group   2.1756     0.4566   4.765 2.39e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.87 on 581 degrees of freedom
## Multiple R-squared:  0.3208, Adjusted R-squared:  0.315
## F-statistic: 54.89 on 5 and 581 DF,  p-value: < 2.2e-16
```

```r
fit2 = update(forward4, . ~ . +parent_marital_status)
summary(fit2)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + lunch_type + test_prep +
##     parent_educ + parent_marital_status, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.111  -8.640   1.036   9.795  31.521
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            72.4812     3.7758  19.196  < 2e-16 ***
## gender                 -9.0578     1.0906  -8.305 7.02e-16 ***
## lunch_type              9.7468     1.1367   8.575  < 2e-16 ***
## test_prep              -9.2621     1.1337  -8.170 1.94e-15 ***
## parent_educ             2.4224     0.3625   6.683 5.49e-11 ***
## parent_marital_status   0.6809     0.7858   0.867    0.387
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.11 on 581 degrees of freedom
## Multiple R-squared:  0.2952, Adjusted R-squared:  0.2891
```

```
## F-statistic: 48.67 on 5 and 581 DF,  p-value: < 2.2e-16
```

```
fit3 = update(forward4, . ~ . +practice_sport)
summary(fit3)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + lunch_type + test_prep +
##     parent_educ + practice_sport, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.874  -8.565   1.021   9.907  31.045
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     72.8024     3.8806  18.761  < 2e-16 ***
## gender          -9.0908     1.0904  -8.337 5.53e-16 ***
## lunch_type       9.7369     1.1370   8.564  < 2e-16 ***
## test_prep       -9.2862     1.1335  -8.192 1.64e-15 ***
## parent_educ      2.4431     0.3633   6.725 4.20e-11 ***
## practice_sport   0.5182     0.8356   0.620    0.535
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.11 on 581 degrees of freedom
## Multiple R-squared:  0.2947, Adjusted R-squared:  0.2887
## F-statistic: 48.56 on 5 and 581 DF,  p-value: < 2.2e-16
```

```
fit4 = update(forward4, . ~ . +is_first_child)
summary(fit4)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + lunch_type + test_prep +
##     parent_educ + is_first_child, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.647  -8.375   1.027   9.788  31.641
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     73.4943     3.8563  19.058  < 2e-16 ***
## gender          -9.0914     1.0907  -8.335 5.62e-16 ***
## lunch_type       9.7114     1.1368   8.542  < 2e-16 ***
## test_prep       -9.2781     1.1353  -8.172 1.90e-15 ***
## parent_educ      2.4278     0.3626   6.695 5.08e-11 ***
## is_first_child   0.3307     1.1560   0.286    0.775
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.12 on 581 degrees of freedom
```

```
## Multiple R-squared:  0.2944, Adjusted R-squared:  0.2883
## F-statistic: 48.48 on 5 and 581 DF,  p-value: < 2.2e-16
```

```
fit5 = update(forward4, . ~ . +nr_siblings)
summary(fit5)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + lunch_type + test_prep +
##     parent_educ + nr_siblings, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.897  -8.679   0.964   9.568  31.042
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  72.9266     3.4363  21.222  < 2e-16 ***
## gender       -9.0381     1.0902  -8.291 7.85e-16 ***
## lunch_type    9.7086     1.1354   8.551  < 2e-16 ***
## test_prep    -9.2282     1.1337  -8.140 2.42e-15 ***
## parent_educ   2.4353     0.3622   6.723 4.25e-11 ***
## nr_siblings   0.4404     0.3660   1.203    0.229
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.1 on 581 degrees of freedom
## Multiple R-squared:  0.296,  Adjusted R-squared:   0.29
## F-statistic: 48.86 on 5 and 581 DF,  p-value: < 2.2e-16
```

```
fit6 = update(forward4, . ~ . +transport_means)
summary(fit6)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + lunch_type + test_prep +
##     parent_educ + transport_means, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.788  -8.350   1.076   9.626  31.792
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      73.0836     3.6747  19.888  < 2e-16 ***
## gender           -9.0996     1.0906  -8.344 5.25e-16 ***
## lunch_type        9.7095     1.1365   8.543  < 2e-16 ***
## test_prep        -9.3455     1.1364  -8.224 1.29e-15 ***
## parent_educ       2.4265     0.3625   6.693 5.14e-11 ***
## transport_means   0.6794     1.1129   0.610    0.542
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 13.12 on 581 degrees of freedom
## Multiple R-squared:  0.2947, Adjusted R-squared:  0.2887
## F-statistic: 48.56 on 5 and 581 DF,  p-value: < 2.2e-16
```

```
fit7 = update(forward4, . ~ . +wkly_study_hours)
summary(fit7)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + lunch_type + test_prep +
##     parent_educ + wkly_study_hours, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.897  -8.394   0.862   9.572  31.435
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       71.5065     3.7299  19.171  < 2e-16 ***
## gender            -9.1459     1.0894  -8.395 3.56e-16 ***
## lunch_type         9.7108     1.1347   8.558  < 2e-16 ***
## test_prep         -9.1382     1.1367  -8.039 5.08e-15 ***
## parent_educ        2.4469     0.3622   6.756 3.44e-11 ***
## wkly_study_hours   1.2185     0.8270   1.473    0.141
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.09 on 581 degrees of freedom
## Multiple R-squared:  0.2969, Adjusted R-squared:  0.2909
## F-statistic: 49.07 on 5 and 581 DF,  p-value: < 2.2e-16
```

```
# Enter the one with the lowest p-value: Ethnic Group
forward5 = update(forward4, . ~ . + ethnic_group)
summary(forward5)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + lunch_type + test_prep +
##     parent_educ + ethnic_group, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.594  -8.422   0.710   9.201  29.415
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   67.7575     3.5050  19.332  < 2e-16 ***
## gender        -9.1231     1.0701  -8.526  < 2e-16 ***
## lunch_type     9.5016     1.1161   8.513  < 2e-16 ***
## test_prep     -9.1875     1.1125  -8.258 9.99e-16 ***
## parent_educ    2.3061     0.3567   6.466 2.14e-10 ***
## ethnic_group   2.1756     0.4566   4.765 2.39e-06 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.87 on 581 degrees of freedom
## Multiple R-squared:  0.3208, Adjusted R-squared:  0.315
## F-statistic: 54.89 on 5 and 581 DF,  p-value: < 2.2e-16
```

```
### Step 6: Enter the one with the lowest p-value in the rest
fit1 = update(forward5, . ~ . +parent_marital_status)
summary(fit1)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + lunch_type + test_prep +
##     parent_educ + ethnic_group + parent_marital_status, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.269  -8.482   0.835   9.426  29.473
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            65.8964     3.9529  16.670  < 2e-16 ***
## gender                 -9.0888     1.0706  -8.490  < 2e-16 ***
## lunch_type              9.5362     1.1166   8.540  < 2e-16 ***
## test_prep              -9.1484     1.1131  -8.219 1.35e-15 ***
## parent_educ             2.2982     0.3567   6.442 2.47e-10 ***
## ethnic_group            2.1888     0.4567   4.792 2.10e-06 ***
## parent_marital_status   0.7856     0.7717   1.018    0.309
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.87 on 580 degrees of freedom
## Multiple R-squared:  0.322,  Adjusted R-squared:  0.315
## F-statistic: 45.92 on 6 and 580 DF,  p-value: < 2.2e-16
```

```
fit2 = update(forward5, . ~ . +practice_sport)
summary(fit2)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + lunch_type + test_prep +
##     parent_educ + ethnic_group + practice_sport, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.933  -8.411   0.853   9.389  29.030
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     66.5308     4.0315  16.503  < 2e-16 ***
## gender          -9.1262     1.0706  -8.524  < 2e-16 ***
## lunch_type       9.5224     1.1173   8.523  < 2e-16 ***
## test_prep       -9.1785     1.1132  -8.245 1.11e-15 ***
```

```
## parent_educ      2.3203     0.3576   6.489 1.86e-10 ***
## ethnic_group     2.1747     0.4568   4.761 2.44e-06 ***
## practice_sport   0.5060     0.8204   0.617    0.538
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.88 on 580 degrees of freedom
## Multiple R-squared:  0.3213, Adjusted R-squared:  0.3142
## F-statistic: 45.75 on 6 and 580 DF,  p-value: < 2.2e-16
```

```
fit3 = update(forward5, . ~ . +is_first_child)
summary(fit3)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + lunch_type + test_prep +
##     parent_educ + ethnic_group + is_first_child, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.716  -8.536   0.728   9.239  29.303
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     67.1816     4.0118  16.746  < 2e-16 ***
## gender          -9.1269     1.0710  -8.522  < 2e-16 ***
## lunch_type       9.4971     1.1171   8.501  < 2e-16 ***
## test_prep       -9.1698     1.1150  -8.224 1.29e-15 ***
## parent_educ      2.3054     0.3570   6.458 2.24e-10 ***
## ethnic_group     2.1757     0.4569   4.762 2.43e-06 ***
## is_first_child   0.3357     1.1350   0.296    0.767
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.88 on 580 degrees of freedom
## Multiple R-squared:  0.3209, Adjusted R-squared:  0.3139
## F-statistic: 45.68 on 6 and 580 DF,  p-value: < 2.2e-16
```

```
fit4 = update(forward5, . ~ . +nr_siblings)
summary(fit4)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + lunch_type + test_prep +
##     parent_educ + ethnic_group + nr_siblings, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.002  -8.116   0.877   9.056  28.949
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   66.4622     3.6317  18.301  < 2e-16 ***
```

```
## gender        -9.0688     1.0701  -8.475  < 2e-16 ***
## lunch_type     9.4922     1.1154   8.510  < 2e-16 ***
## test_prep     -9.1126     1.1131  -8.187 1.71e-15 ***
## parent_educ    2.3127     0.3564   6.488 1.87e-10 ***
## ethnic_group   2.1915     0.4564   4.802 2.01e-06 ***
## nr_siblings    0.4849     0.3594   1.349    0.178
## ---
## 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:  0.3229, Adjusted R-squared:  0.3159
## F-statistic: 46.11 on 6 and 580 DF,  p-value: < 2.2e-16
```

```
fit5 = update(forward5, . ~ . +transport_means)
summary(fit5)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + lunch_type + test_prep +
##     parent_educ + ethnic_group + transport_means, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.917  -8.476   0.575   9.219  29.122
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      66.5192     3.8595  17.235  < 2e-16 ***
## gender           -9.1379     1.0706  -8.535  < 2e-16 ***
## lunch_type        9.4928     1.1166   8.502  < 2e-16 ***
## test_prep        -9.2487     1.1157  -8.289 7.95e-16 ***
## parent_educ       2.3030     0.3568   6.454 2.30e-10 ***
## ethnic_group      2.1863     0.4569   4.785 2.18e-06 ***
## transport_means   0.8391     1.0930   0.768    0.443
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.87 on 580 degrees of freedom
## Multiple R-squared:  0.3215, Adjusted R-squared:  0.3145
## F-statistic: 45.81 on 6 and 580 DF,  p-value: < 2.2e-16
```

```
fit6 = update(forward5, . ~ . +wkly_study_hours)
summary(fit6)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + lunch_type + test_prep +
##     parent_educ + ethnic_group + wkly_study_hours, data = step_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 ***
## lunch_type         9.4976     1.1151   8.517  < 2e-16 ***
## test_prep         -9.0360     1.1163  -8.094 3.40e-15 ***
## parent_educ        2.3242     0.3566   6.519 1.54e-10 ***
## ethnic_group       2.1684     0.4562   4.753 2.53e-06 ***
## 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
```

```r
# P-value of all new added variables are larger than 0.05, which means that they
# are not significant predictor, and we stop here.

# The model we obtained is Writing Score ~ Gender + Lunch Type + Test Prep +
# Parent Education + Ethnic Group

writing_forward_manual_fit = lm(writing_score ~ gender + lunch_type + test_prep +
                     parent_educ + ethnic_group, data = step_df)
summary(writing_forward_manual_fit)
```

```
##
## Call:
## lm(formula = writing_score ~ gender + lunch_type + test_prep +
##     parent_educ + ethnic_group, data = step_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.594  -8.422   0.710   9.201  29.415
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   67.7575     3.5050  19.332  < 2e-16 ***
## gender        -9.1231     1.0701  -8.526  < 2e-16 ***
## lunch_type     9.5016     1.1161   8.513  < 2e-16 ***
## test_prep     -9.1875     1.1125  -8.258 9.99e-16 ***
## parent_educ    2.3061     0.3567   6.466 2.14e-10 ***
## ethnic_group   2.1756     0.4566   4.765 2.39e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.87 on 581 degrees of freedom
## Multiple R-squared:  0.3208, Adjusted R-squared:  0.315
## F-statistic: 54.89 on 5 and 581 DF,  p-value: < 2.2e-16
```

```r
mean(writing_forward_manual_fit$residuals^2)
```

```
## [1] 163.9483
```

```r
# fit using one function
intercept_only <- lm (writing_score ~ 1, data = writing_df)
writing_forward_func_fit = step(intercept_only, direction = "forward", scope = formula(mult.fit))
```

```
## Start:  AIC=3222.53
## writing_score ~ 1
##
##                          Df Sum of Sq    RSS    AIC
## + gender                  1   11104.5 130592 3176.6
## + lunch_type              1   10442.9 131253 3179.6
## + test_prep               1    9618.7 132078 3183.3
## + parent_educ             1    8420.5 133276 3188.6
## + ethnic_group            1    5344.3 136352 3202.0
## + nr_siblings             1     560.4 141136 3222.2
## <none>                                 141696 3222.5
## + wkly_study_hours        1     472.0 141224 3222.6
## + parent_marital_status   1     275.4 141421 3223.4
## + is_first_child          1      95.4 141601 3224.1
## + transport_means         1       0.6 141696 3224.5
## + practice_sport          1       0.0 141696 3224.5
##
## Step:  AIC=3176.62
## writing_score ~ gender
##
##                          Df Sum of Sq    RSS    AIC
## + lunch_type              1   11657.0 118935 3123.7
## + test_prep               1   10482.9 120109 3129.5
## + parent_educ             1    7206.7 123385 3145.3
## + ethnic_group            1    5423.0 125169 3153.7
## + wkly_study_hours        1     687.6 129904 3175.5
## <none>                                 130592 3176.6
## + nr_siblings             1     401.5 130190 3176.8
## + parent_marital_status   1     171.1 130421 3177.9
## + is_first_child          1     127.2 130465 3178.0
## + transport_means         1       5.5 130586 3178.6
## + practice_sport          1       0.6 130591 3178.6
##
## Step:  AIC=3123.74
## writing_score ~ gender + lunch_type
##
##                          Df Sum of Sq    RSS    AIC
## + test_prep               1   11216.2 107719 3067.6
## + parent_educ             1    7367.0 111568 3088.2
## + ethnic_group            1    4829.4 114105 3101.4
## + wkly_study_hours        1     686.4 118248 3122.3
## <none>                                 118935 3123.7
## + nr_siblings             1     385.6 118549 3123.8
## + parent_marital_status   1     276.7 118658 3124.4
## + is_first_child          1     100.6 118834 3125.2
## + practice_sport          1      15.9 118919 3125.7
## + transport_means         1       1.3 118933 3125.7
##
## Step:  AIC=3067.59
```

```
## writing_score ~ gender + lunch_type + test_prep
##
##                          Df Sum of Sq    RSS    AIC
## + parent_educ             1    7719.9  99999 3025.9
## + ethnic_group            1    4556.1 103162 3044.2
## <none>                                 107719 3067.6
## + wkly_study_hours        1     264.6 107454 3068.2
## + nr_siblings             1     207.4 107511 3068.5
## + parent_marital_status   1     171.3 107547 3068.7
## + transport_means         1      77.9 107641 3069.2
## + is_first_child          1      19.3 107699 3069.5
## + practice_sport          1       6.1 107713 3069.6
##
## Step:  AIC=3025.94
## writing_score ~ gender + lunch_type + test_prep + parent_educ
##
##                          Df Sum of Sq   RSS    AIC
## + ethnic_group            1    3761.0 96238 3005.4
## + wkly_study_hours        1     372.2 99626 3025.8
## <none>                                 99999 3025.9
## + nr_siblings             1     248.6 99750 3026.5
## + parent_marital_status   1     129.1 99870 3027.2
## + practice_sport          1      66.2 99933 3027.6
## + transport_means         1      64.1 99935 3027.6
## + is_first_child          1      14.1 99985 3027.9
##
## Step:  AIC=3005.44
## writing_score ~ gender + lunch_type + test_prep + parent_educ +
##     ethnic_group
##
##                          Df Sum of Sq   RSS    AIC
## + wkly_study_hours        1    346.82 95891 3005.3
## <none>                                 96238 3005.4
## + nr_siblings             1    301.18 95936 3005.6
## + parent_marital_status   1    171.69 96066 3006.4
## + transport_means         1     97.68 96140 3006.8
## + practice_sport          1     63.09 96175 3007.1
## + is_first_child          1     14.52 96223 3007.3
##
## Step:  AIC=3005.32
## writing_score ~ gender + lunch_type + test_prep + parent_educ +
##     ethnic_group + wkly_study_hours
##
##                          Df Sum of Sq   RSS    AIC
## <none>                                 95891 3005.3
## + nr_siblings             1   270.733 95620 3005.7
## + parent_marital_status   1   175.061 95716 3006.2
## + transport_means         1    84.868 95806 3006.8
## + practice_sport          1    58.664 95832 3007.0
## + is_first_child          1    13.991 95877 3007.2
```

```
summary(writing_forward_func_fit)
```

```
##
```

```
## Call:
## lm(formula = writing_score ~ gender + lunch_type + test_prep +
##     parent_educ + ethnic_group + 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 ***
## lunch_type         9.4976     1.1151   8.517  < 2e-16 ***
## test_prep         -9.0360     1.1163  -8.094 3.40e-15 ***
## parent_educ        2.3242     0.3566   6.519 1.54e-10 ***
## ethnic_group       2.1684     0.4562   4.753 2.53e-06 ***
## 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
```

```
mean(writing_forward_func_fit$residuals^2)
```

```
## [1] 163.3575
```

The model we obtained is Writing Score ~ Lunch Type + Ethnic Group + Test Prep + Gender + Parent Education + Weekly Study Hours.

When using the single-function method, the model obtained with the lowest AIC was Writing Score ~ Gender + Lunch Type + Test Prep + Parent Education + Ethnic Group + Weekly Study Hours. Both models had approximately equal adjusted R-squared values, while the MSE of the one-line model was about 3.5 units lower.

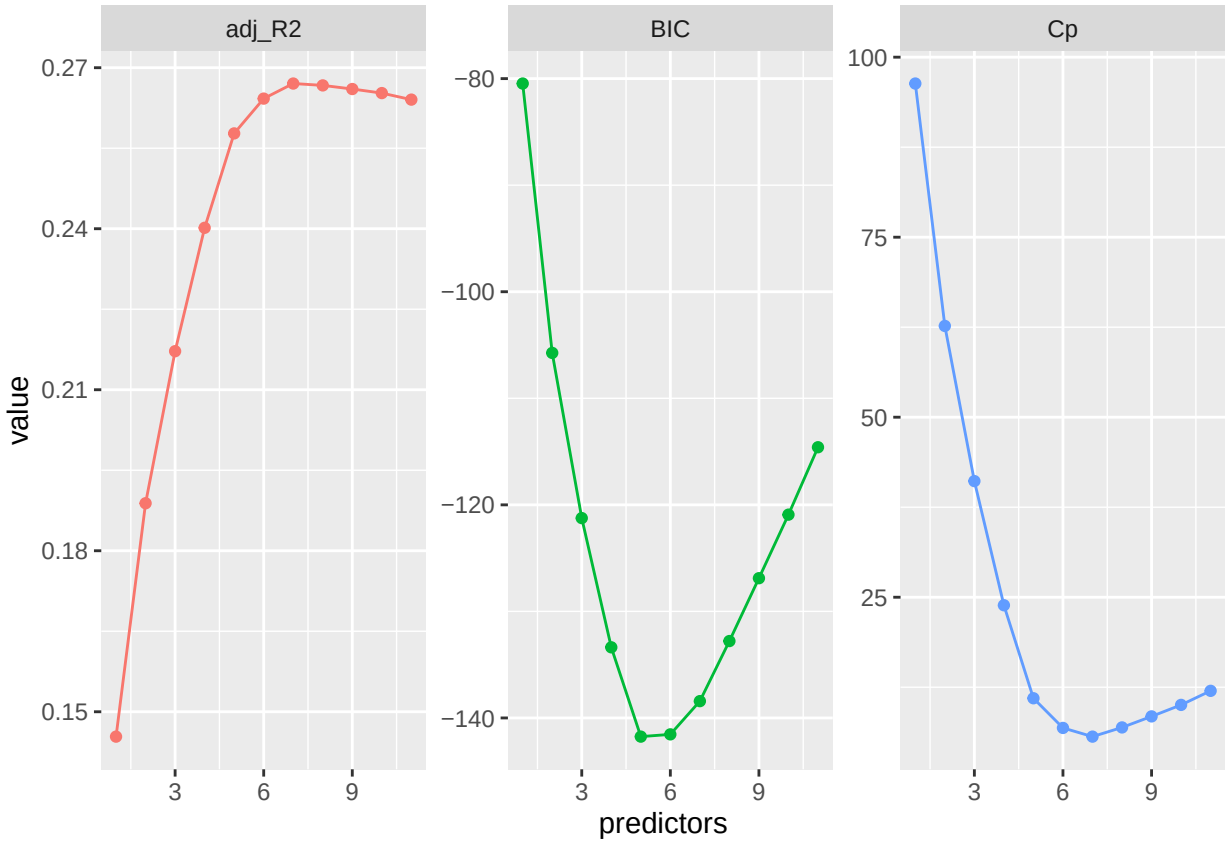
Criteria-based approach - Adjusted R^2, Cp, BIC

(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")
```




```
results$which[7,]|>print()
```

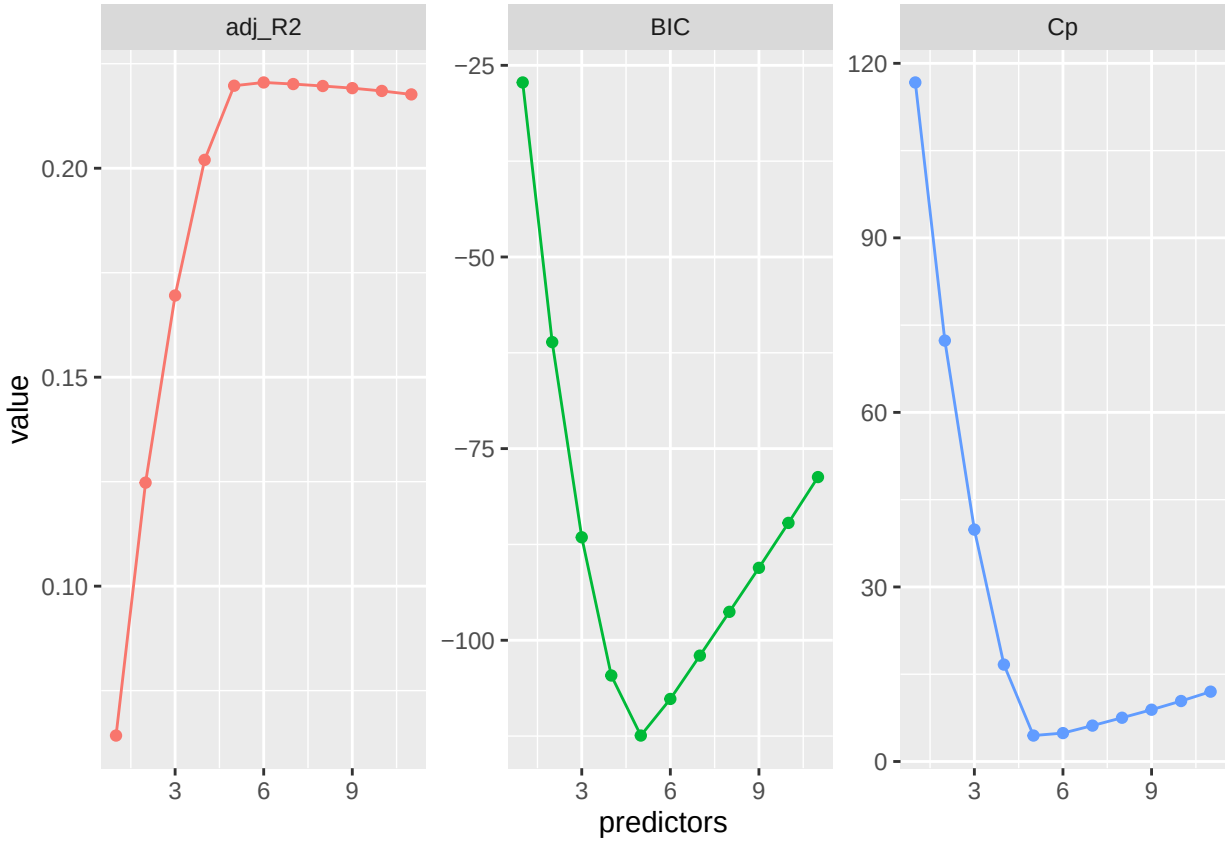
```
##           (Intercept)                gender          ethnic_group
##                  TRUE                  TRUE                  TRUE
##           parent_educ           lunch_type             test_prep
##                  TRUE                  TRUE                  TRUE
## parent_marital_status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

```
math_criteria_fit = lm(math_score ~ gender + ethnic_group + parent_educ + lunch_type+ test_prep + nr_sib

ggsave("math_criteria_plots.png")
```

To predict math score, the adjusted R^2 statistic, Cp, and BIC plots in combination show that a 7-variable model is optimal. The predictors selected are: gender, ethnic_group, parent_educ, lunch_type, test_prep, nr_siblings, and wkly study_hours.

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")
```




```
results$which[5,]|>print()
```

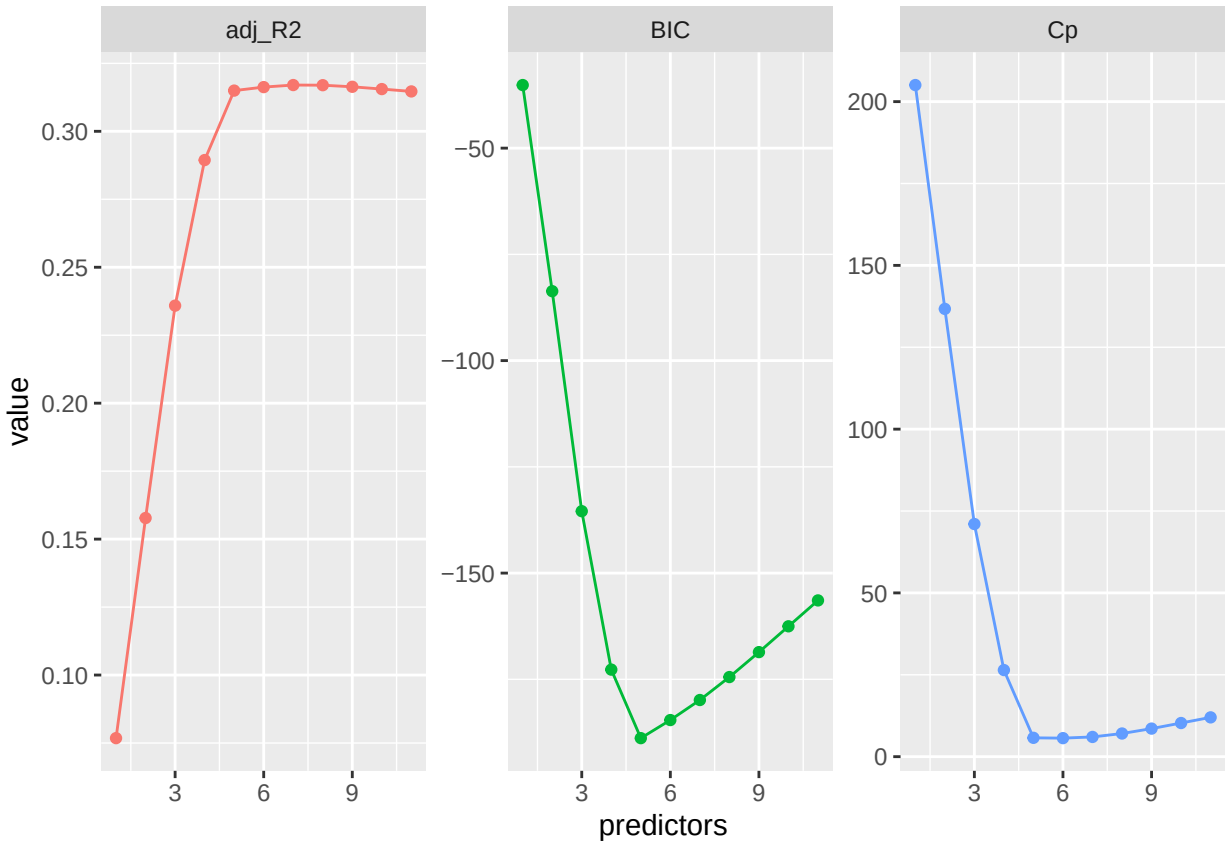
```
##           (Intercept)                gender          ethnic_group
##                  TRUE                  TRUE                  TRUE
##           parent_educ           lunch_type             test_prep
##                  TRUE                  TRUE                  TRUE
## parent_marital_status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

```
reading_criteria_fit = lm(reading_score ~ gender + ethnic_group + parent_educ + lunch_type+ test_prep, d

ggsave("reading_criteria_plots.png")
```

To predict reading score, the adjusted R^2 statistic and Cp and BIC plots shows that a 5-variable model is optimal. The predictors selected are: gender, ethnic_group, parent_educ, lunch_type, test_prep.

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) %>%
  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")
```




```
results$which[5,] |>print()
```

```
##           (Intercept)                gender          ethnic_group
```

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

```
writing_criteria_fit = lm(writing_score ~ gender + ethnic_group + parent_educ + lunch_type + test_prep,

ggsave("writing_criteria_plots.png")
```

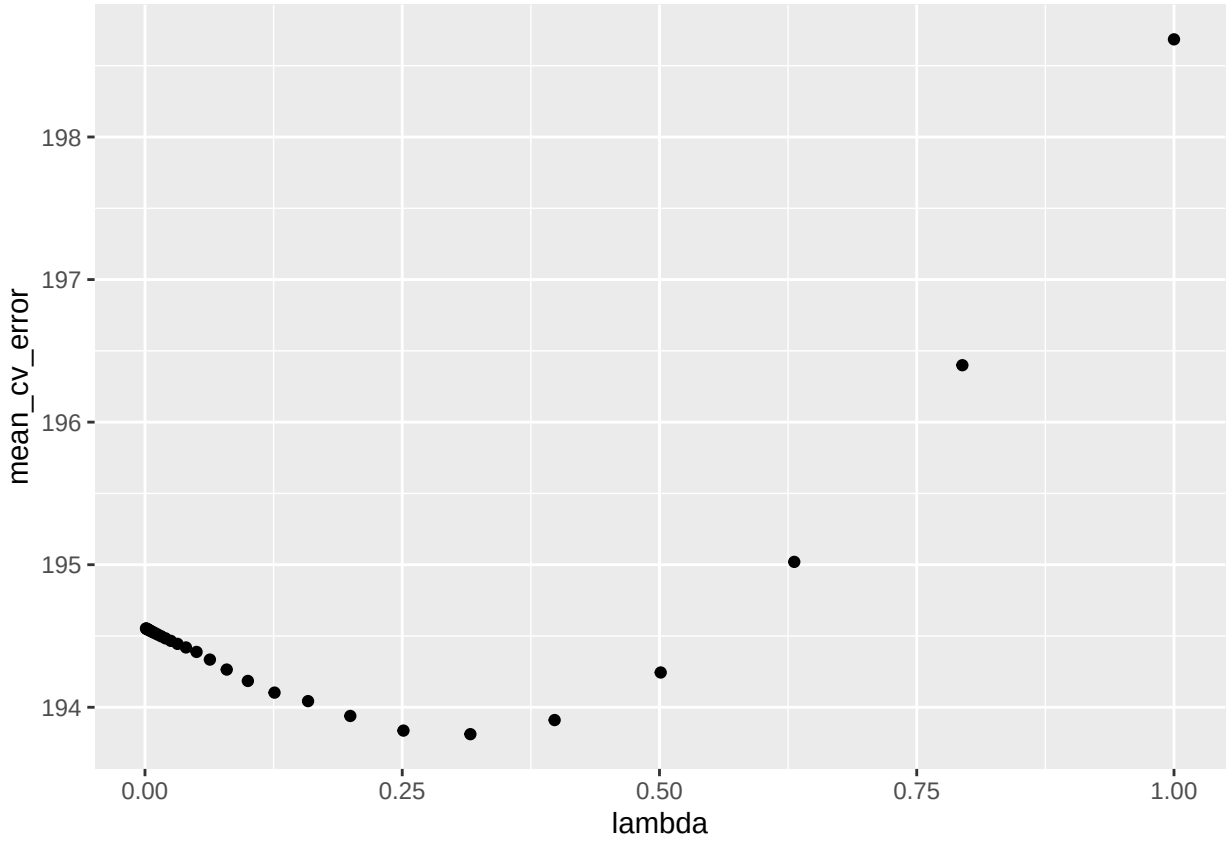
To predict writing score, the adjusted R^2, Cp, and BIC statistics show that a 5-variable model is optimal. The predictors selected are: gender, ethnic_group, parent_educ, lunch_type, test_prep

Limitation: noting that the plots maximum and minimum are not that obvious.

LASSO approach -

Maths score:

```
# Find the best lambda
math_lasso=step_df|>
  dplyr::select(-reading_score,-writing_score)|>
  dplyr::select(math_score,everything())

lambda_seq = 10^seq(-3, 0, by = .1)
set.seed(1)
cv_object = cv.glmnet(as.matrix(math_lasso[2:12]),math_lasso$math_score, lambda = lambda_seq, nfolds = 5

tibble(lambda = cv_object$lambda,
       mean_cv_error = cv_object$cvm) |>
  ggplot(aes(x = lambda, y = mean_cv_error)) +
  geom_point()
```



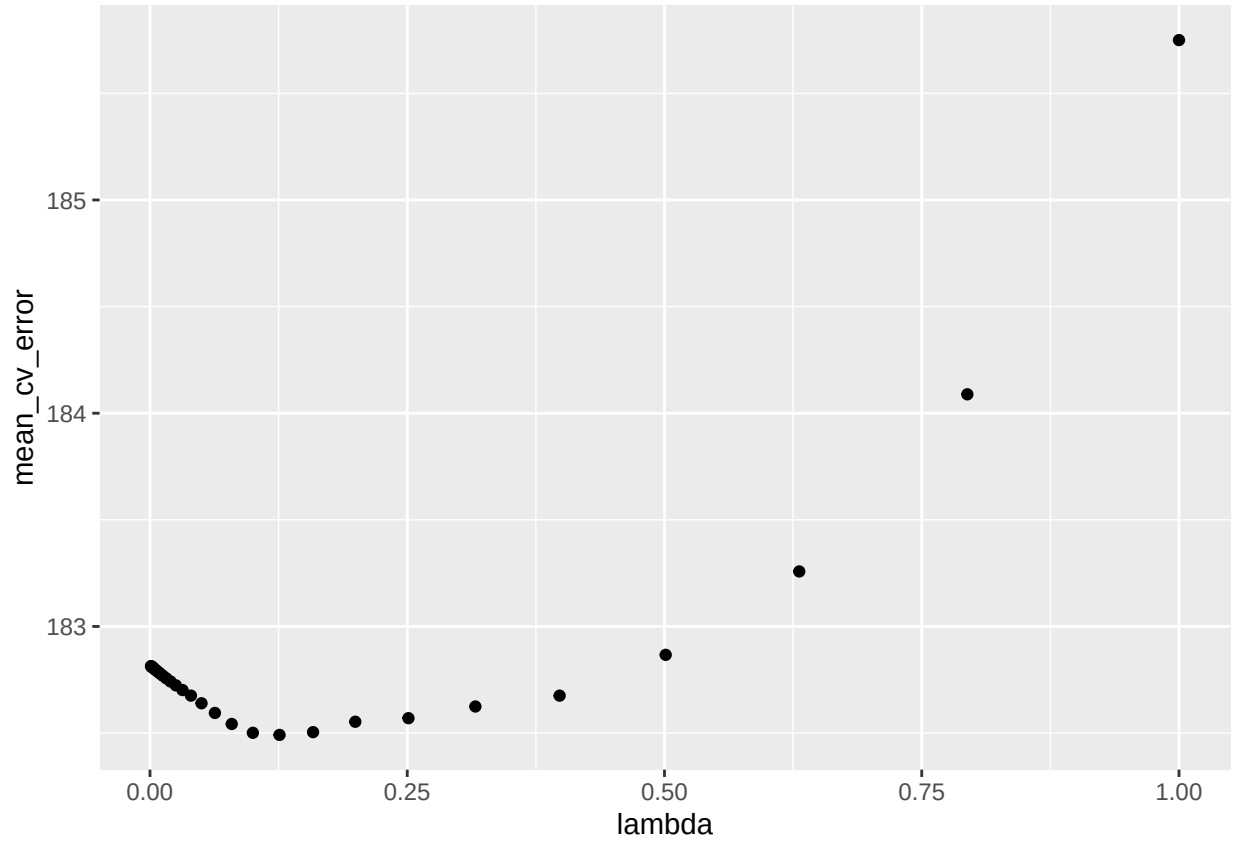

```
# Use the best lambda to model
math_model_lasso=glmnet(as.matrix(math_lasso[2:12]),math_lasso$math_score,lambda=cv_object$lambda.min)
coef(math_model_lasso)
```

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

Reading score:

```
read_lasso=step_df|>
  dplyr::select(-math_score,-writing_score)|>
  dplyr::select(reading_score,everything())
```

```
lambda_seq = 10^seq(-3, 0, by = .1)
set.seed(2)
cv_object = cv.glmnet(as.matrix(read_lasso[2:12]),read_lasso$reading_score, lambda = lambda_seq, nfolds

tibble(lambda = cv_object$lambda,
       mean_cv_error = cv_object$cvm) |>
  ggplot(aes(x = lambda, y = mean_cv_error)) +
  geom_point()
```



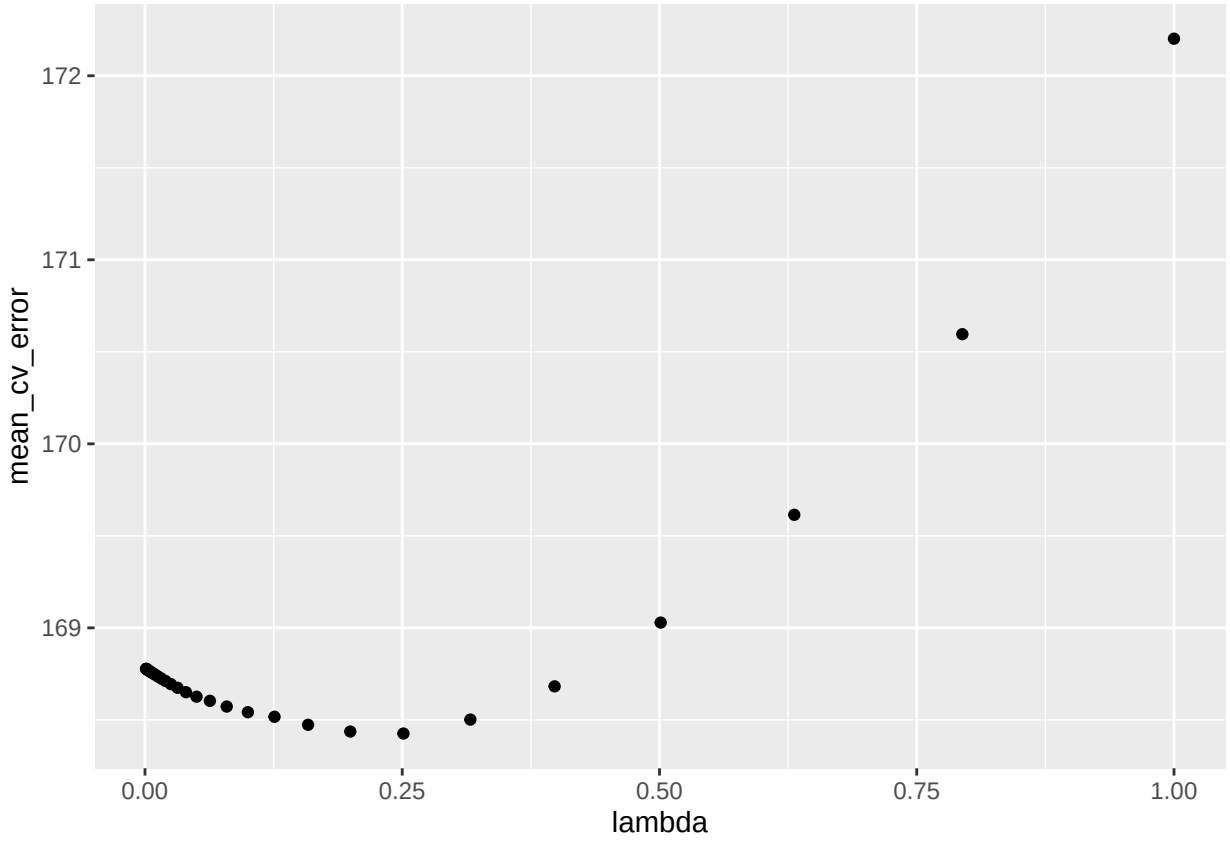

```
# Use the best lambda to model
read_model_lasso=glmnet(as.matrix(read_lasso[2:12]),read_lasso$reading_score,lambda=cv_object$lambda.mi
coef(read_model_lasso)
```

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

```
## transport_means         0.6090418
## wkly_study_hours        0.8425385
```

Writing score:

```
write_lasso=df_num|>
  dplyr::select(-reading_score,-math_score)|>
  dplyr::select(writing_score,everything())

lambda_seq = 10^seq(-3, 0, by = .1)
set.seed(2)
cv_object = cv.glmnet(as.matrix(write_lasso[2:12]),write_lasso$writing_score, lambda = lambda_seq, nfol

tibble(lambda = cv_object$lambda,
       mean_cv_error = cv_object$cvm) |>
  ggplot(aes(x = lambda, y = mean_cv_error)) +
  geom_point()
```




```
# Use the best lambda to model
write_model_lasso=glmnet(as.matrix(write_lasso[2:12]),write_lasso$writing_score,lambda=cv_object$lambda
coef(write_model_lasso)
```

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                                s0
## (Intercept)          64.24045111
```

```
## gender                -8.58684295
## ethnic_group           2.00086017
## parent_educ            2.16675381
## lunch_type             8.95045903
## test_prep             -8.47239858
## parent_marital_status  0.40572033
## practice_sport         0.04910284
## is_first_child         .
## nr_siblings            0.29894240
## transport_means        0.22598323
## wkly_study_hours       0.76604805
```

Cross Validation

Here are the summary of all the models that have been created in this project.

```
# Clean out the variables used in stepwise analysis
var_names = step_df |> dplyr::select(!ends_with("score"))  |> colnames()

math_theoretical_fit = lm(math_score ~ gender + ethnic_group + parent_educ + lunch_type + test_prep + pa

reading_theoretical_fit = lm(reading_score ~ gender + ethnic_group + parent_educ + lunch_type + test_pre

writing_theoretical_fit = lm(writing_score ~ gender + ethnic_group + parent_educ + lunch_type + test_pre

models_report_df = rbind(
  math_theoretical_fit,
  math_backward_manual_fit,
  math_backward_func_fit,
  math_forward_manual_fit,
  math_forward_func_fit,
  math_criteria_fit,

  reading_theoretical_fit,
  reading_backward_manual_fit,
  reading_backward_func_fit,
  reading_forward_manual_fit,
  reading_forward_func_fit,
  reading_criteria_fit,

  writing_theoretical_fit,
  writing_backward_manual_fit,
  writing_backward_func_fit,
  writing_forward_manual_fit,
  writing_forward_func_fit,
  writing_criteria_fit)

models_report_df_rownames = models_report_df |> row.names()
models_report_df_colnames = models_report_df |> colnames()

models_report_df = models_report_df |>
  as.data.frame() |>
```

subject	method	(Intercept)	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	theoretical	X	X	X	X	X	X	X	X	X	X	X	X
math	backward_manual	X	X	X	X	X	X	NA	NA	NA	NA	NA	X
math	backward_func	X	X	X	X	X	X	NA	NA	NA	X	NA	X
math	forward_manual	X	X	X	X	X	X	NA	NA	NA	NA	NA	X
math	forward_func	X	X	X	X	X	X	NA	NA	NA	X	NA	X
math	criteria	X	X	X	X	X	X	NA	NA	NA	X	NA	NA
reading	theoretical	X	X	X	X	X	X	X	X	X	X	X	X
reading	backward_manual	X	X	X	X	X	X	NA	NA	NA	NA	NA	NA
reading	backward_func	X	X	X	X	X	X	NA	NA	NA	NA	NA	NA
reading	forward_manual	X	X	X	X	X	X	NA	NA	NA	NA	NA	NA
reading	forward_func	X	X	X	X	X	X	NA	NA	NA	NA	NA	NA
reading	criteria	X	X	X	X	X	X	NA	NA	NA	NA	NA	NA
writing	theoretical	X	X	X	X	X	X	X	X	X	X	X	X
writing	backward_manual	X	X	X	X	X	X	NA	NA	NA	NA	NA	NA
writing	backward_func	X	X	X	X	X	X	NA	NA	NA	NA	NA	X
writing	forward_manual	X	X	X	X	X	X	NA	NA	NA	NA	NA	NA
writing	forward_func	X	X	X	X	X	X	NA	NA	NA	NA	NA	X
writing	criteria	X	X	X	X	X	X	NA	NA	NA	NA	NA	NA

```
  cbind(models_report_df_rownames) |>
  rename(model_name = models_report_df_rownames) |>
  dplyr::select( model_name, coefficients, terms) |>
  mutate(coefficients = map(coefficients, \(coef) map(coef, ~ "X"))) |>
  unnest_wider(coefficients) |>
  mutate(
    subject = map(model_name, \(i) str_extract(i, "^[:alpha:]+")),
    method = map(model_name, \(i) str_extract(i, "(?<=[:alpha:]_).+(?=_fit)"))
  ) |>
  dplyr::select(-model_name) |>
  relocate(subject,method, terms)

models_report_df |>
  dplyr::select(-terms) |>
  knitr::kable()%>%
kable_styling("striped", full_width = F) %>%
  row_spec(0, angle = -90)
```

We will be performing cross validation to select the best model resulted from the models above.

Method from the lecture code

```
set.seed(1)
# Use 5-fold validation and create the training sets
train = trainControl(method = "cv", number = 5)

cv_lecture_df = models_report_df |>
  dplyr::select(subject, method, terms) |>
  mutate(model = map(terms, \(formula) train(as.formula(formula),
                    data = step_df,
                    trControl = train,
                    method = 'lm',
                    na.action = na.pass)),
         RMSE = map(model, \(i) i$results$RMSE),
         model = map(model, \(i) i$finalModel)
         )
```

Method using crossv_mc

```
set.seed(1)
cv_ds_df =
  modelr::crossv_mc(step_df, 100) |>
  mutate(
    train = map(train, as_tibble),
    test = map(test, as_tibble)) |>
  mutate(
    fits = map(train, \(i) cv_lecture_df |> transpose() |> as.list())) |>
  unnest(fits) |>
  unnest_wider(fits, strict = TRUE, names_repair = "minimal") |>
  mutate(
    cv_model = map2(train, terms, \(df, i) lm(as.formula(i), data = df)),
    cv_rmse = map2(cv_model, test, \(mod, df) rmse(mod,df)),
    cv_rmse = as.numeric(cv_rmse),
    method = as.character(method)
  )
```

Notice how `practice_sport` and `transport_means` are not selected in any of the model selections methods. This will be reported at the effect modifier section.

Cross Validation - Math

Method from the lecture codes

```
math_caret_df = cv_lecture_df |>
  filter(subject == "math")

math_caret_df |>
  dplyr::select(method, RMSE) |>
  knitr::kable()
```

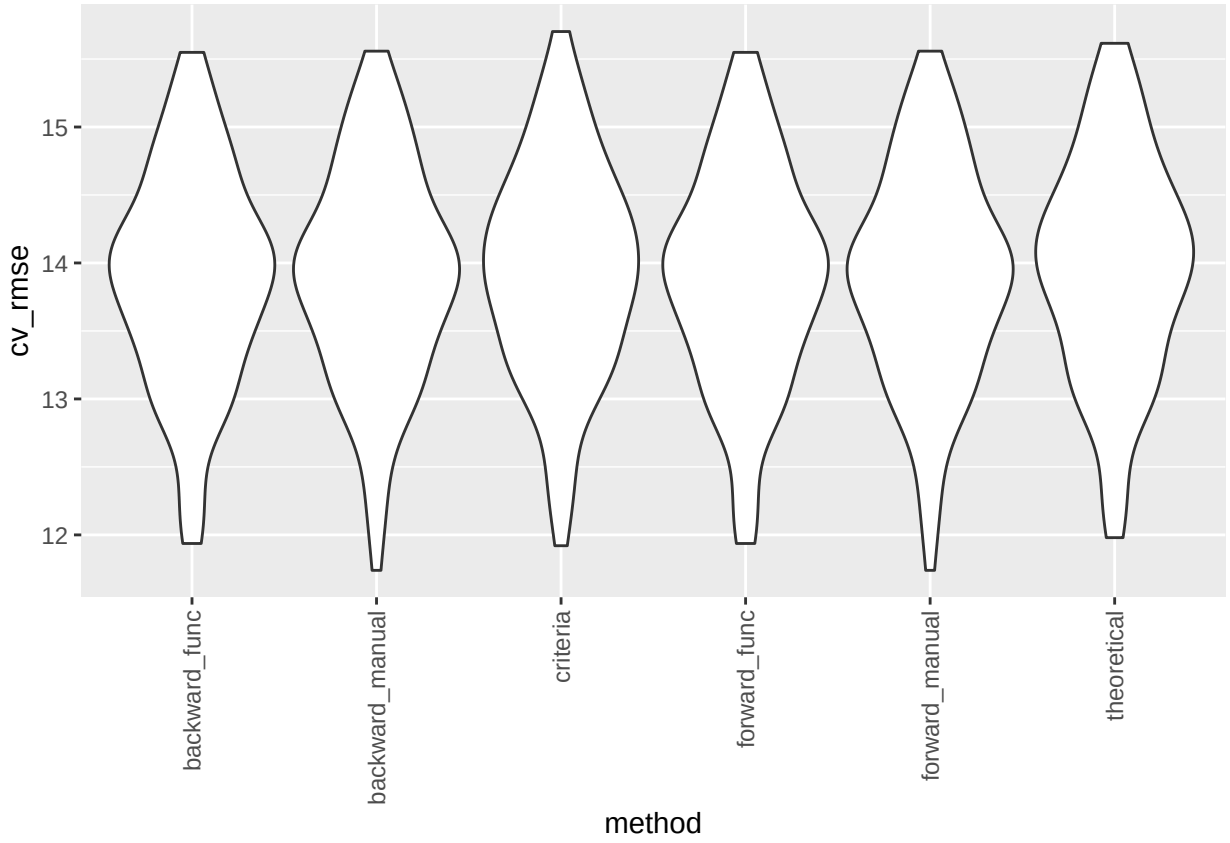
method	RMSE
theoretical	13.96975
backward_manual	13.82973
backward_func	13.84664
forward_manual	13.88238
forward_func	13.89506
criteria	13.9035

```
math_caret_df = math_caret_df |>
  filter(RMSE == min(math_caret_df$RMSE |> unlist()))

# Only one value
math_best_fit = math_caret_df$model[[1]]
```

The model with the best RMSE for Math is math_score ~ gender + ethnic_group + parent_educ + lunch_type + test_prep + wkly_study_hours, which uses backward_manual as a method of approach.

Method using crossv_mc

```
cv_ds_df |>
  filter(subject == "math") |>
  group_by(method) |>
  ggplot(aes(x = method, y = cv_rmse)) +
  geom_violin()+ theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



```r
cv_ds_df |>
  filter(subject == "math") |>
  group_by(method) |>
  summarize(average_rmse = mean(cv_rmse)) |>
  knitr::kable()
```

method	average_rmse
backward_func	13.88519
backward_manual	13.89653
criteria	13.94264
forward_func	13.88519
forward_manual	13.89653
theoretical	13.96494

We noticed that the the best model is the one that uses forward elimination with one line code and backward elimination with one line code. The model is math_score ~ lunch_type + ethnic_group + test_prep + gender + parent_educ + wkly_study_hours + nr_siblings

Cross Validation - Reading

Method from the lecture codes

```r
reading_caret_df = cv_lecture_df |>
  filter(subject == "reading")

reading_caret_df |>
  dplyr::select(method, RMSE) |>
  knitr::kable()
```

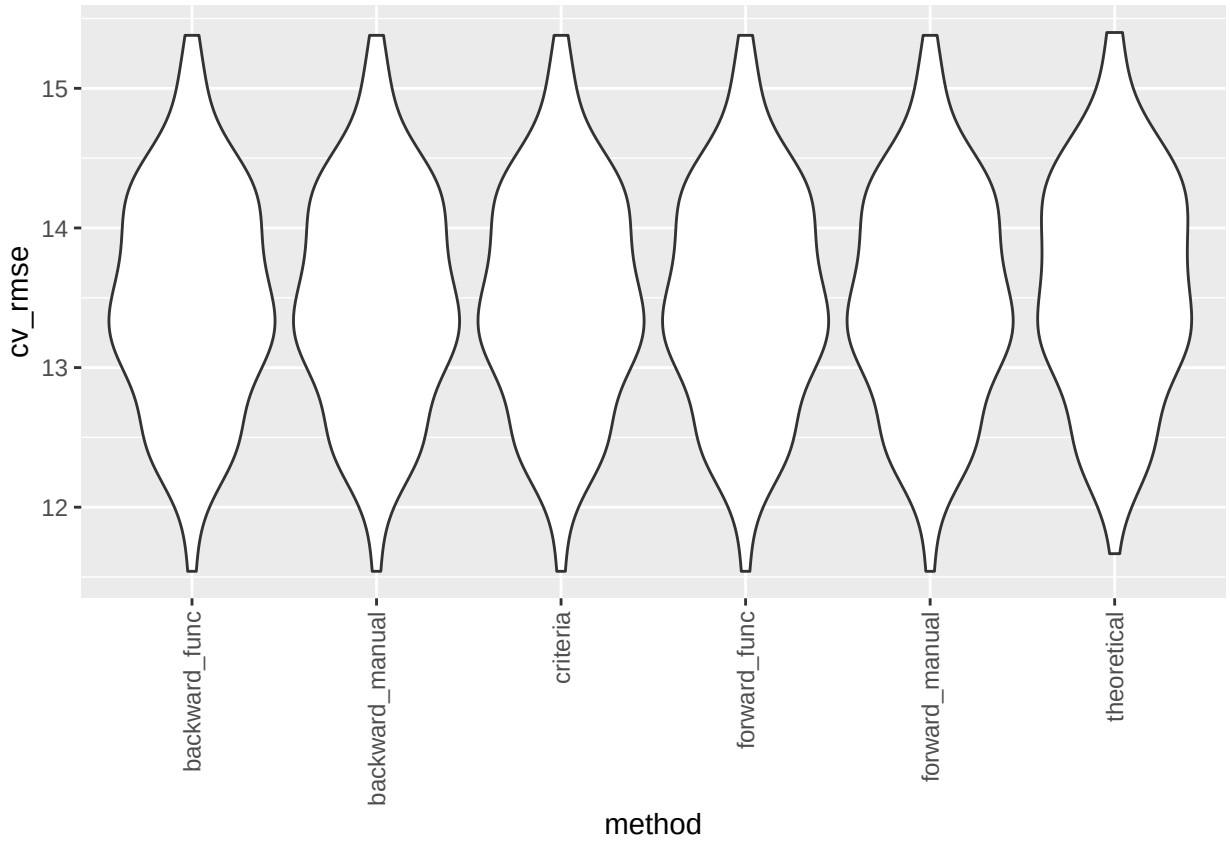
method	RMSE
theoretical	13.53786
backward_manual	13.42876
backward_func	13.53256
forward_manual	13.42892
forward_func	13.40176
criteria	13.53827

```r
reading_caret_df = reading_caret_df |>
  filter(RMSE == min(reading_caret_df$RMSE |> unlist()))

# Only one value
reading_best_fit = reading_caret_df$model[[1]]
```

The model with the best MSE for Reading is reading_score ~ lunch_type + gender + test_prep + parent_educ + ethnic_group, which uses forward_func as a method of approach.

Method using crossv_mc

```
cv_ds_df |>
  filter(subject == "reading") |>
  group_by(method) |>
  ggplot(aes(x = method, y = cv_rmse)) +
  geom_violin()+ theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```




```
cv_ds_df |>
  filter(subject == "reading") |>
  group_by(method) |>
  summarize(average_rmse = mean(cv_rmse)) |>
  knitr::kable()
```

method	average_rmse
backward_func	13.49136
backward_manual	13.49136
criteria	13.49136
forward_func	13.49136
forward_manual	13.49136
theoretical	13.57974

We noticed that the the best model is the model that is picked by forward and backward elimination method and criterion based approach. The model is reading_score ~ lunch_type + gender + test_prep + parent_educ + ethnic_group

Cross Validation - Writing

Method from the lecture codes

```
writing_caret_df = cv_lecture_df |>
  filter(subject == "writing")

writing_caret_df |>
  dplyr::select(method, RMSE) |>
  knitr::kable()
```

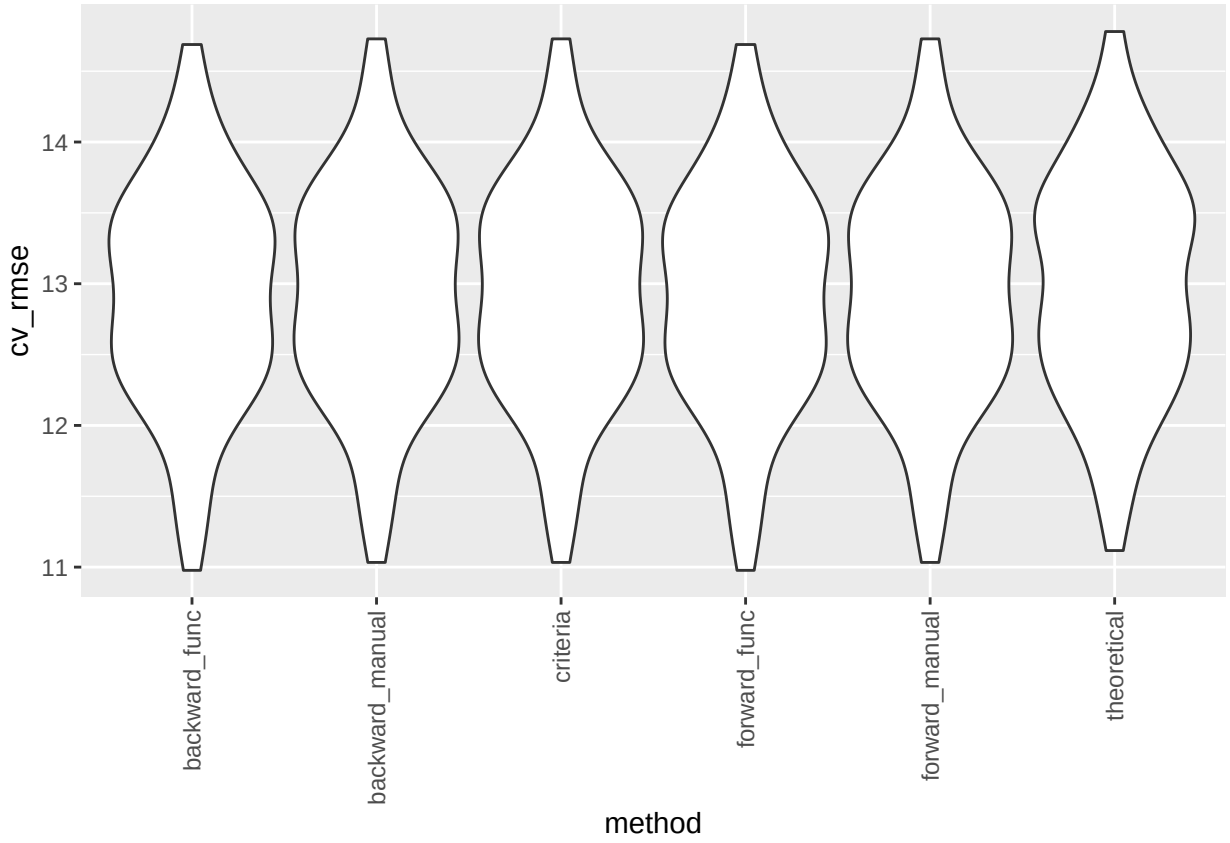
method	RMSE
theoretical	12.95331
backward_manual	12.93239
backward_func	12.90413
forward_manual	12.88047
forward_func	12.90165
criteria	12.9846

```
writing_caret_df = writing_caret_df |>
  filter(RMSE == min(writing_caret_df$RMSE |> unlist()))

# Only one value
writing_best_fit = writing_caret_df$model[[1]]
```

The model with the best RMSE for Writing is writing_score ~ gender + lunch_type + test_prep + parent_educ + ethnic_group, which uses forward_manual as a method of approach.

Method using crossv_mc

```
cv_ds_df |>
  filter(subject == "writing") |>
  group_by(method) |>
  ggplot(aes(x = method, y = cv_rmse)) +
  geom_violin()+ theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```




```
cv_ds_df |>
  filter(subject == "writing") |>
  group_by(method) |>
  summarize(average_rmse = mean(cv_rmse)) |>
  knitr::kable()
```

method	average_rmse
backward_func	12.91828
backward_manual	12.92630
criteria	12.92630
forward_func	12.91828
forward_manual	12.92630
theoretical	12.99816

We noticed that the the best model is the one that uses forward elimination with one line code and backward elimination with one line code. The model is writing_score ~ gender + lunch_type + test_prep + parent_educ + ethnic_group + wkly_study_hours

Effect Modifier

```
math_best_cv_terms = models_report_df |> filter(subject == "math", method == "forward_func") |> pull(te
reading_best_cv_terms = models_report_df |> filter(subject == "reading", method == "forward_func") |> pu
writing_best_cv_terms = models_report_df |> filter(subject == "writing", method == "forward_func") |> pu
```

Table 1: Math: Effect Modifiers

term	df	sumsq	meansq	statistic	p.value
gender:parent_educ:wkly_study_hours	1	1013.717	1013.717	6.018977	0.0144781

Table 2: Math: Effect Modifiers

term	df	sumsq	meansq	statistic	p.value
gender:parent_educ:wkly_study_hours	1	1013.717	1013.717	6.018977	0.0144781

```r
lm(as.formula(gsub("\\+", "*", writing_best_cv_terms)), data = step_df) |>
  anova() |>
  broom::tidy() |>
  filter(str_detect(term, ":")) |>
  filter(p.value < 0.05) |>
  knitr::kable(caption = "Math: Effect Modifiers")
```

```r
lm(as.formula(gsub("\\+", "*", writing_best_cv_terms)), data = step_df) |>
  anova() |>
  broom::tidy() |>
  filter(str_detect(term, ":")) |>
  filter(p.value < 0.05) |>
  knitr::kable(caption = "Math: Effect Modifiers")
```

```r
lm(as.formula(gsub("\\+", "*", writing_best_cv_terms)), data = step_df) |>
  anova() |>
  broom::tidy() |>
  filter(str_detect(term, ":")) |>
  filter(p.value < 0.05) |>
  knitr::kable(caption = "Math: Effect Modifiers")
```

```r
lm(math_score ~ parent_educ * nr_siblings, data = df_transformed) |>
  broom::tidy() |>
  knitr::kable(caption = "Math: Effect Modifiers")
```

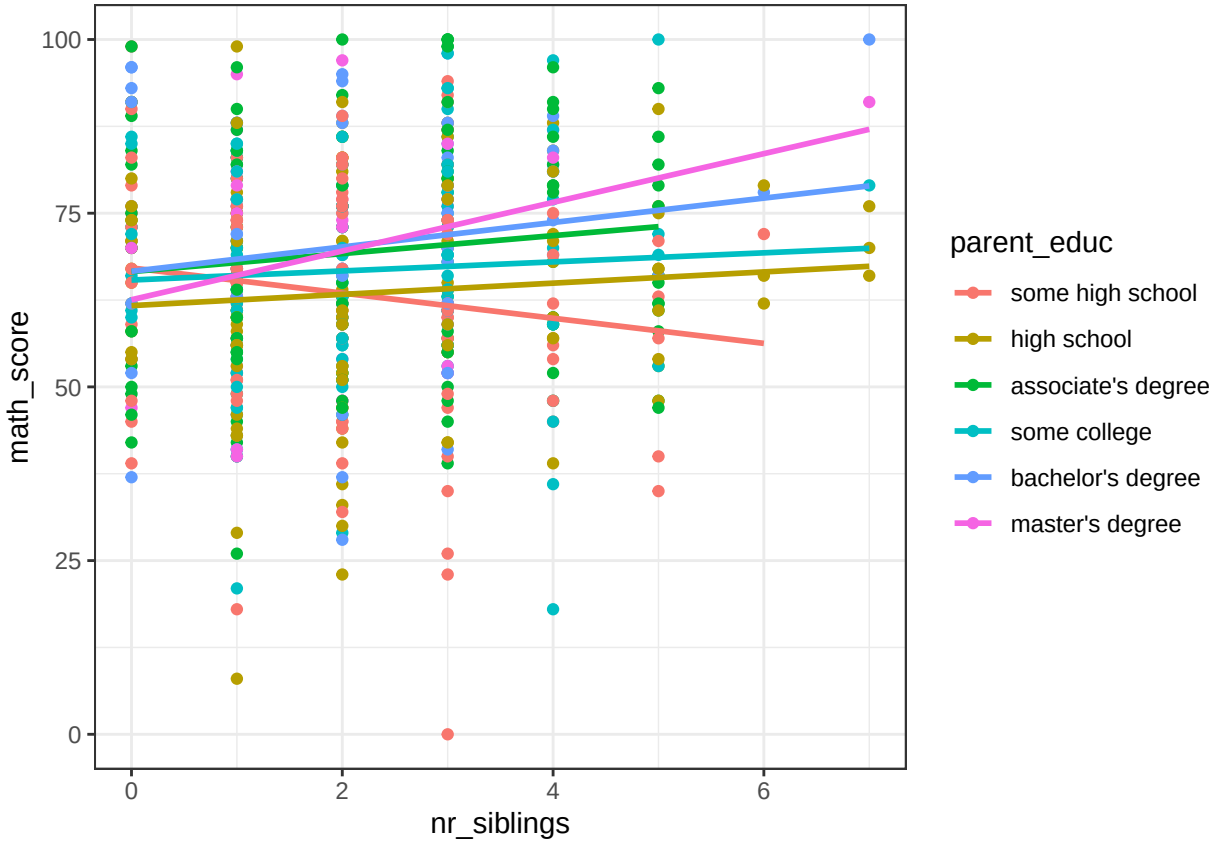
```r
df_transformed |>
  ggplot(aes(x = nr_siblings, y = math_score, color = parent_educ)) +
  geom_point() +
  geom_smooth(method="lm", se=F, aes(group = parent_educ, color = parent_educ)) +
  theme_bw()
```

Table 3: Math: Effect Modifiers

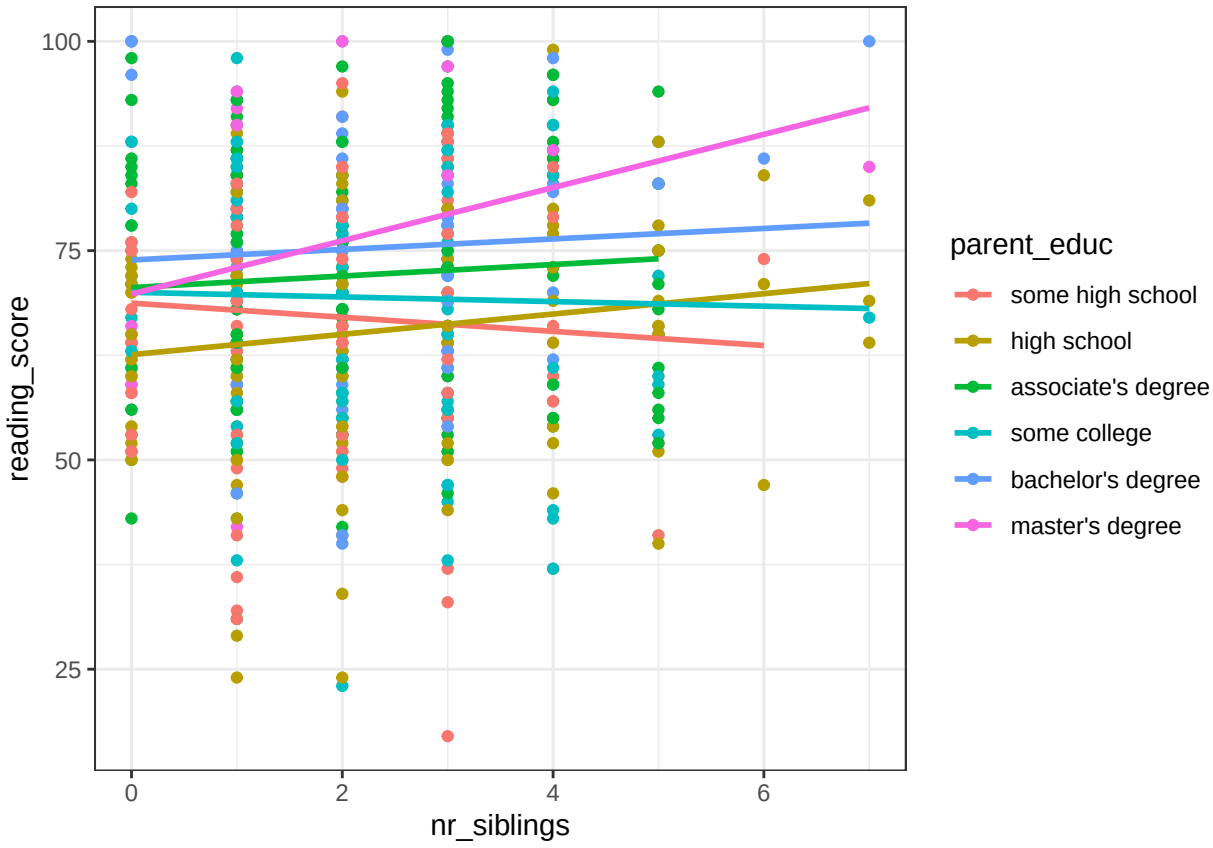
term	df	sumsq	meansq	statistic	p.value
gender:parent_educ:wkly_study_hours	1	1013.717	1013.717	6.018977	0.0144781

Table 4: Math: Effect Modifiers

term	estimate	std.error	statistic	p.value
(Intercept)	67.1235922	2.607864	25.7389193	0.0000000
parent_educhigh school	-5.4269635	3.568956	-1.5206027	0.1289093
parent_educassociate's degree	-0.5301543	3.619042	-0.1464902	0.8835857
parent_educsome college	-1.7421489	3.820951	-0.4559464	0.6486007
parent_educbachelor's degree	-0.4937420	4.209905	-0.1172810	0.9066783
parent_educmaster's degree	-4.5795427	5.155579	-0.8882694	0.3747673
nr_siblings	-1.8097637	1.088621	-1.6624371	0.0969701
parent_educhigh school:nr_siblings	2.6183145	1.381791	1.8948696	0.0586111
parent_educassociate's degree:nr_siblings	3.1036097	1.432380	2.1667504	0.0306639
parent_educsome college:nr_siblings	2.4621644	1.544827	1.5938123	0.1115276
parent_educbachelor's degree:nr_siblings	3.5667749	1.704282	2.0928313	0.0368019
parent_educmaster's degree:nr_siblings	5.3136277	2.232714	2.3798968	0.0176425



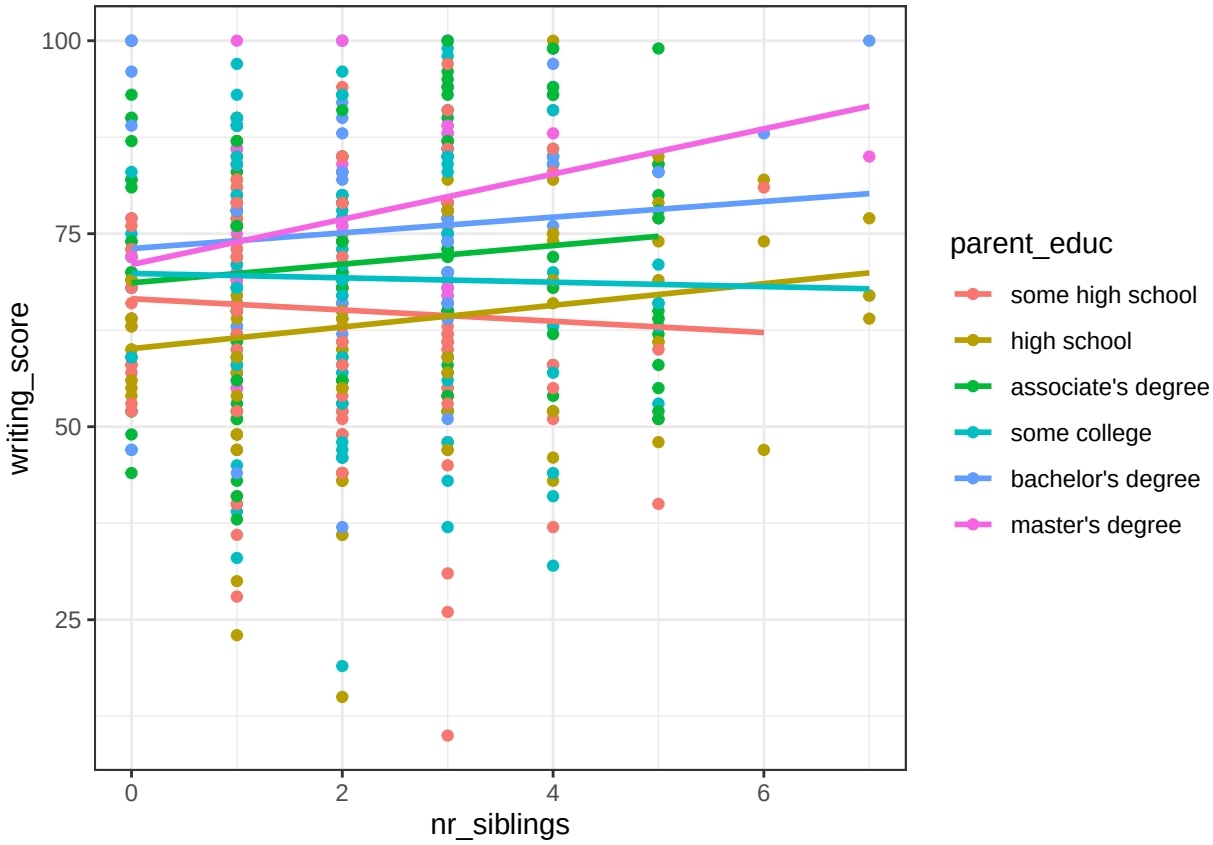

```
df_transformed |>
  ggplot(aes(x = nr_siblings, y = reading_score, group = parent_educ, color = parent_educ)) +
  geom_point() +
  geom_smooth(method="lm", se=F) +
  theme_bw()
```




```
df_transformed |>
  ggplot(aes(x = nr_siblings, y = writing_score, color = parent_educ)) +
  geom_point() +
  geom_smooth(method="lm", se=F, aes(group = parent_educ, color = parent_educ)) +
  theme_bw()
```

Table 5: math: full under CV

term	estimate	std.error	statistic	p.value
(Intercept)	28.0712943	4.2756226	6.565429	0.0000000
lunch_type	12.5737137	1.1964102	10.509534	0.0000000
ethnic_group	2.7439281	0.4895913	5.604528	0.0000000
test_prep	-5.2926304	1.1989408	-4.414422	0.0000121
gender	5.3016818	1.1486346	4.615638	0.0000048
parent_educ	1.5210046	0.3825700	3.975754	0.0000790
wkly_study_hours	2.0824941	0.8723498	2.387224	0.0172960
nr_siblings	0.6926991	0.3859641	1.794724	0.0732191



Confounder

Confounding - Math

```
# math
lm(as.formula(math_best_cv_terms), data = step_df) |>
  broom::tidy() |>
  knitr::kable(caption = "math: full under CV")
```

Table 6: math: without Gender

term	estimate	std.error	statistic	p.value
(Intercept)	35.8588529	3.9968205	8.971845	0.0000000
lunch_type	12.8699814	1.2154184	10.588931	0.0000000
ethnic_group	2.7563514	0.4980796	5.533957	0.0000000
test_prep	-5.5052544	1.2188454	-4.516778	0.0000076
parent_educ	1.4061946	0.3883852	3.620618	0.0003196
wkly_study_hours	2.2357475	0.8868446	2.521014	0.0119687
nr_siblings	0.6224711	0.3923565	1.586494	0.1131724

Table 7: math: without Lunch Type

term	estimate	std.error	statistic	p.value
(Intercept)	46.5624715	4.2486221	10.959429	0.0000000
ethnic_group	2.9517064	0.5333555	5.534219	0.0000000
test_prep	-4.8719574	1.3064508	-3.729155	0.0002110
gender	5.9493280	1.2505281	4.757453	0.0000025
parent_educ	1.4728518	0.4170780	3.531358	0.0004463
wkly_study_hours	2.1024658	0.9511021	2.210558	0.0274560
nr_siblings	0.7174494	0.4208005	1.704963	0.0887368

```
lm(as.formula(gsub("gender", "", math_best_cv_terms)), data = step_df) |> #
  broom::tidy() |>
  knitr::kable(caption = "math: without Gender")
```

```
lm(as.formula(gsub("lunch_type", "", math_best_cv_terms)), data = step_df) |>
  broom::tidy() |>
  knitr::kable(caption = "math: without Lunch Type")
```

```
lm(as.formula(gsub("test_prep", "", math_best_cv_terms)), data = step_df) |> #
  broom::tidy() |>
  knitr::kable(caption = "math: without Test Prep")
```

```
lm(as.formula(gsub("\\+ parent_educ", "", math_best_cv_terms)), data = step_df) |> #
  broom::tidy() |>
  knitr::kable(caption = "math: without Parent Education")
```

Table 8: math: without Test Prep

term	estimate	std.error	statistic	p.value
(Intercept)	18.4754298	3.7400301	4.939915	0.0000010
lunch_type	12.3973873	1.2146505	10.206547	0.0000000
ethnic_group	2.7880487	0.4972292	5.607171	0.0000000
gender	5.4965040	1.1659354	4.714244	0.0000030
parent_educ	1.4931204	0.3885663	3.842640	0.0001352
wkly_study_hours	2.4337489	0.8824495	2.757947	0.0059999
nr_siblings	0.7694383	0.3916691	1.964511	0.0499479

Table 9: math: without Parent Education

term	estimate	std.error	statistic	p.value
(Intercept)	32.9359482	4.1487671	7.938732	0.0000000
lunch_type	12.5167463	1.2114984	10.331624	0.0000000
ethnic_group	2.8841686	0.4945127	5.832345	0.0000000
test_prep	-5.2139275	1.2139826	-4.294895	0.0000205
gender	5.0047625	1.1607425	4.311691	0.0000190
wkly_study_hours	1.9631840	0.8828917	2.223584	0.0265609
nr_siblings	0.6744634	0.3908320	1.725712	0.0849319

Table 10: math: without Ethnic Group

term	estimate	std.error	statistic	p.value
(Intercept)	36.0431862	4.1364418	8.713573	0.0000000
lunch_type	12.8444847	1.2263736	10.473550	0.0000000
test_prep	-5.4298042	1.2297146	-4.415500	0.0000120
gender	5.3370727	1.1783449	4.529296	0.0000072
parent_educ	1.6754839	0.3914515	4.280183	0.0000218
wkly_study_hours	2.1423461	0.8948602	2.394057	0.0169798
nr_siblings	0.6356448	0.3958156	1.605912	0.1088376

```
lm(as.formula(gsub("ethnic_group", "", math_best_cv_terms)), data = step_df) |>
  broom::tidy() |>
  knitr::kable(caption = "math: without Ethnic Group")
```

```
lm(as.formula(gsub("\\+ wkly_study_hours", "", math_best_cv_terms)), data = step_df) |>
  broom::tidy() |>
  knitr::kable(caption = "math: without Weekly Study Hours")
```

```
lm(as.formula(gsub("\\+ nr_siblings", "", math_best_cv_terms)), data = step_df) |>
  broom::tidy() |>
  knitr::kable(caption = "math: without Number of Siblings")
```

Removing `test_prep` will increase `wkly_study_hours` by 0.0288184. Therefore test prep could be a confounder for weekly study hours in hours

Table 11: math: without Weekly Study Hours

term	estimate	std.error	statistic	p.value
(Intercept)	32.2785820	3.9112360	8.252783	0.0000000
lunch_type	12.5799355	1.2012438	10.472425	0.0000000
ethnic_group	2.7582360	0.4915337	5.611489	0.0000000
test_prep	-5.5536954	1.1987695	-4.632830	0.0000045
gender	5.4060487	1.1524423	4.690949	0.0000034
parent_educ	1.4895871	0.3838892	3.880252	0.0001163
nr_siblings	0.7382445	0.3870506	1.907359	0.0569678

Table 12: math: without Number of Siblings

term	estimate	std.error	statistic	p.value
(Intercept)	29.760538	4.1787141	7.121937	0.0000000
lunch_type	12.586815	1.1986764	10.500595	0.0000000
ethnic_group	2.720752	0.4903572	5.548511	0.0000000
test_prep	-5.389546	1.2000152	-4.491231	0.0000085
gender	5.220415	1.1499372	4.539739	0.0000069
parent_educ	1.512845	0.3832748	3.947155	0.0000888
wkly_study_hours	2.159886	0.8729500	2.474238	0.0136371

Table 13: reading: full under CV

term	estimate	std.error	statistic	p.value
(Intercept)	66.712128	3.6484858	18.284881	0.00e+00
lunch_type	8.666712	1.1618414	7.459462	0.00e+00
gender	-7.506622	1.1138734	-6.739206	0.00e+00
test_prep	-6.828881	1.1580310	-5.896976	0.00e+00
parent_educ	1.760626	0.3712699	4.742173	2.70e-06
ethnic_group	1.793048	0.4752616	3.772761	1.78e-04

Confounding - Reading

```
lm(as.formula(reading_best_cv_terms), data = step_df) |>
  broom::tidy() |>
  knitr::kable(caption = "reading: full under CV")
```

```
lm(as.formula(gsub("gender", "", reading_best_cv_terms)), data = step_df) |> #
  broom::tidy() |>
  knitr::kable(caption = "reading: without Gender")
```

```
lm(as.formula(gsub("lunch_type", "", reading_best_cv_terms)), data = step_df) |>
  broom::tidy() |>
  knitr::kable(caption = "reading: without Lunch Type")
```

```
lm(as.formula(gsub("test_prep", "", reading_best_cv_terms)), data = step_df) |> #
  broom::tidy() |>
  knitr::kable(caption = "reading: without Test Prep")
```

Table 14: reading: without Gender

term	estimate	std.error	statistic	p.value
(Intercept)	55.468110	3.3661817	16.478050	0.0000000
lunch_type	8.247206	1.2036296	6.851947	0.0000000
test_prep	-6.514335	1.2004315	-5.426661	0.0000001
parent_educ	1.925653	0.3843380	5.010310	0.0000007
ethnic_group	1.770801	0.4930517	3.591512	0.0003566

Table 15: reading: without Lunch Type

term	estimate	std.error	statistic	p.value
(Intercept)	79.532319	3.3660323	23.627913	0.0000000
gender	-7.061453	1.1633176	-6.070099	0.0000000
test_prep	-6.543314	1.2105129	-5.405407	0.0000001
parent_educ	1.726992	0.3882795	4.447807	0.0000104
ethnic_group	1.935796	0.4966690	3.897558	0.0001085

Table 16: reading: without Test Prep

term	estimate	std.error	statistic	p.value
(Intercept)	55.415717	3.1939548	17.350188	0.0000000
lunch_type	8.440218	1.1944245	7.066347	0.0000000
gender	-7.241882	1.1448065	-6.325857	0.0000000
parent_educ	1.715858	0.3818108	4.494000	0.0000084
ethnic_group	1.850135	0.4887557	3.785399	0.0001694

```
lm(as.formula(gsub("\\+ parent_educ", "", reading_best_cv_terms)), data = step_df) |> #
  broom::tidy() |>
  knitr::kable(caption = "reading: without Parent Education")
```

```
lm(as.formula(gsub("\\+ ethnic_group", "", reading_best_cv_terms)), data = step_df) |>
  broom::tidy() |>
  knitr::kable(caption = "reading: without Ethnic Group")
```

We don't have any confounders for reading score

Confounding - Writing

```
# Writing
lm(as.formula(writing_best_cv_terms), data = step_df) |>
  broom::tidy() |>
  knitr::kable(caption = "Writing: full under CV")
```

```
lm(as.formula(gsub("gender", "", writing_best_cv_terms)), data = step_df) |> #
  broom::tidy() |>
  knitr::kable(caption = "Writing: without Gender")
```

Table 17: reading: without Parent Education

term	estimate	std.error	statistic	p.value
(Intercept)	72.006949	3.5369813	20.358306	0.00e+00
lunch_type	8.599800	1.1830081	7.269434	0.00e+00
gender	-7.855014	1.1317801	-6.940407	0.00e+00
test_prep	-6.716589	1.1789687	-5.697004	0.00e+00
ethnic_group	1.955447	0.4826977	4.051080	5.79e-05

Table 18: reading: without Ethnic Group

term	estimate	std.error	statistic	p.value
(Intercept)	71.907404	3.4168330	21.045045	0e+00
lunch_type	8.843208	1.1740235	7.532394	0e+00
gender	-7.477432	1.1264389	-6.638116	0e+00
test_prep	-6.917873	1.1708798	-5.908269	0e+00
parent_educ	1.861556	0.3744912	4.970894	9e-07

Table 19: Writing: full under CV

term	estimate	std.error	statistic	p.value
(Intercept)	65.312482	3.8873510	16.801283	0.0000000
gender	-9.179189	1.0697572	-8.580629	0.0000000
lunch_type	9.497555	1.1150980	8.517238	0.0000000
test_prep	-9.036032	1.1163435	-8.094312	0.0000000
parent_educ	2.324218	0.3565507	6.518618	0.0000000
ethnic_group	2.168352	0.4561667	4.753419	0.0000025
wkly_study_hours	1.176186	0.8120831	1.448357	0.1480577

Table 20: Writing: without Gender

term	estimate	std.error	statistic	p.value
(Intercept)	52.1053044	3.7861927	13.761926	0.0000000
lunch_type	8.9861099	1.1810512	7.608569	0.0000000
test_prep	-8.6843663	1.1832649	-7.339326	0.0000000
parent_educ	2.5218719	0.3773897	6.682407	0.0000000
ethnic_group	2.1427345	0.4838283	4.428708	0.0000113
wkly_study_hours	0.9240399	0.8607815	1.073489	0.2834972

Table 21: Writing: without Lunch Type

term	estimate	std.error	statistic	p.value
(Intercept)	79.325906	3.7325524	21.252456	0.0000000
gender	-8.692165	1.1320889	-7.677988	0.0000000
test_prep	-8.720881	1.1824309	-7.375383	0.0000000
parent_educ	2.287624	0.3778387	6.054500	0.0000000
ethnic_group	2.324677	0.4830459	4.812539	0.0000019
wkly_study_hours	1.193361	0.8606285	1.386616	0.1660908

Table 22: Writing: without Test Prep

term	estimate	std.error	statistic	p.value
(Intercept)	49.216447	3.5208066	13.978742	0.0000000
gender	-8.861298	1.1268300	-7.863917	0.0000000
lunch_type	9.198386	1.1747364	7.830171	0.0000000
parent_educ	2.274969	0.3757718	6.054123	0.0000000
ethnic_group	2.239439	0.4807388	4.658328	0.0000040
wkly_study_hours	1.791765	0.8522239	2.102458	0.0359429

```
lm(as.formula(gsub("lunch_type", "", writing_best_cv_terms)), data = step_df) |>
  broom::tidy() |>
  knitr::kable(caption = "Writing: without Lunch Type")
```

```
lm(as.formula(gsub("test_prep", "", writing_best_cv_terms)), data = step_df) |> #
  broom::tidy() |>
  knitr::kable(caption = "Writing: without Test Prep")
```

```
lm(as.formula(gsub("\\+ parent_educ", "", writing_best_cv_terms)), data = step_df) |> #
  broom::tidy() |>
  knitr::kable(caption = "Writing: without Parent Education")
```

```
lm(as.formula(gsub("ethnic_group", "", writing_best_cv_terms)), data = step_df) |>
  broom::tidy() |>
  knitr::kable(caption = "Writing: without Ethnic Group")
```

Table 23: Writing: without Parent Education

term	estimate	std.error	statistic	p.value
(Intercept)	72.6791532	3.8499827	18.877787	0.0000000
gender	-9.6297006	1.1049839	-8.714788	0.0000000
lunch_type	9.4099646	1.1541452	8.153189	0.0000000
test_prep	-8.9118520	1.1553500	-7.713552	0.0000000
ethnic_group	2.3836134	0.4709357	5.061441	0.0000006
wkly_study_hours	0.9907309	0.8400647	1.179351	0.2387413

Table 24: Writing: without Ethnic Group

term	estimate	std.error	statistic	p.value
(Intercept)	71.506520	3.7298885	19.171222	0.0000000
gender	-9.145910	1.0894332	-8.395108	0.0000000
lunch_type	9.710824	1.1347127	8.557958	0.0000000
test_prep	-9.138195	1.1366900	-8.039303	0.0000000
parent_educ	2.446910	0.3621638	6.756362	0.0000000
wkly_study_hours	1.218475	0.8269878	1.473390	0.1411875

Table 25: Writing: without Weekly Study Hours

term	estimate	std.error	statistic	p.value
(Intercept)	67.757450	3.5049950	19.331682	0.0e+00
gender	-9.123124	1.0700661	-8.525758	0.0e+00
lunch_type	9.501565	1.1161475	8.512822	0.0e+00
test_prep	-9.187450	1.1124869	-8.258479	0.0e+00
parent_educ	2.306126	0.3566683	6.465745	0.0e+00
ethnic_group	2.175590	0.4565701	4.765073	2.4e-06

```
lm(as.formula(gsub("\\+ wkly_study_hours", "", writing_best_cv_terms)), data = step_df) |>
  broom::tidy() |>
  knitr::kable(caption = "Writing: without Weekly Study Hours")
```

As we can see, removing `gender` will lower `wkly_study_hours` by 0.2142857, removing `test_prep` will increase `wkly_study_hours` by 0.5238095, and removing `parent_educ` will lower `wkly_study_hours` by 0.1573129.

Hence, `gender`, `test_prep`, `parent_educ` could be potential confounder for `wkly_study_hours`

Final model

Math

```
math_final = lm(math_score ~ lunch_type + ethnic_group + test_prep + gender + parent_educ + wkly_study_
reading_final = lm(reading_score ~ lunch_type + gender + test_prep + parent_educ + ethnic_group, data =
writing_final = lm(writing_score ~ gender + lunch_type + test_prep + parent_educ + ethnic_group + wkly_
```

```
summary(math_final)
```

```
##
## Call:
## lm(formula = math_score ~ lunch_type + ethnic_group + test_prep +
##     gender + parent_educ + wkly_study_hours + nr_siblings, data = df_num)
##
## 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 ***
## lunch_type        12.5737     1.1964  10.510  < 2e-16 ***
## ethnic_group       2.7439     0.4896   5.605 3.23e-08 ***
## test_prep         -5.2926     1.1989  -4.414 1.21e-05 ***
## gender             5.3017     1.1486   4.616 4.83e-06 ***
## parent_educ        1.5210     0.3826   3.976 7.90e-05 ***
## wkly_study_hours   2.0825     0.8723   2.387   0.0173 *
## nr_siblings        0.6927     0.3860   1.795   0.0732 .
## ---
## 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
```

$\hat{MathScore} = 28.0713 + 12.5737 * \text{Lunch Type} + 2.7439 * \text{Ethnic Group} - 5.2926 * \text{Test Prep} + 5.3017 * \text{Gender} + 1.5210 * \text{Parent Education} + 2.0825 * \text{Weekly Study Hours} + 1.5210 * \text{Number of Sibilings}$

Reading

```
summary(reading_final)
```

```
##
## Call:
## lm(formula = reading_score ~ lunch_type + gender + test_prep +
##     parent_educ + ethnic_group, data = df_num)
##
## 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 ***
## lunch_type     8.6667     1.1618   7.459 3.18e-13 ***
## gender        -7.5066     1.1139  -6.739 3.84e-11 ***
## test_prep     -6.8289     1.1580  -5.897 6.28e-09 ***
## parent_educ    1.7606     0.3713   4.742 2.66e-06 ***
## ethnic_group   1.7930     0.4753   3.773 0.000178 ***
## ---
## 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
```

$\hat{ReadingScore} = 66.7121 + 8.6667 * \text{Lunch Type} - 7.5066 * \text{Gender} - 6.8289 * \text{Test Prep} + 1.7606 * \text{Parent Education} + 1.7930 * \text{Ethnic Group}$

Writing

```
summary(writing_final)
```

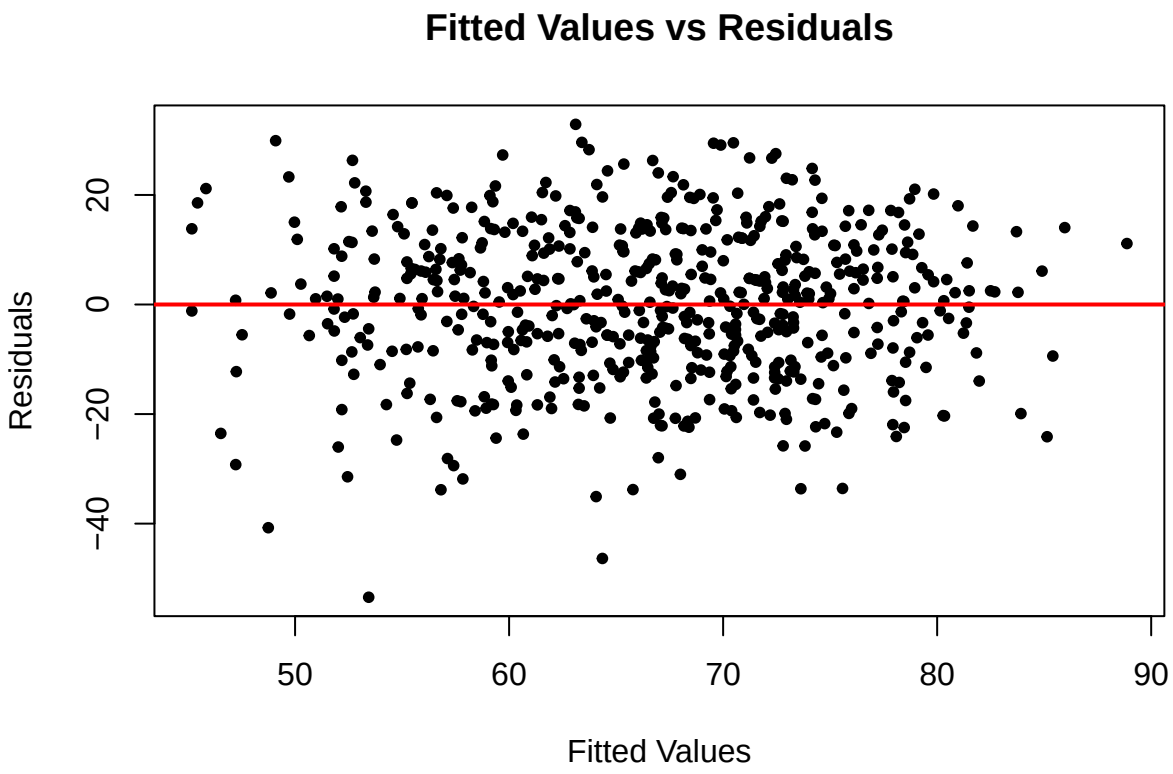
```
##
## Call:
## lm(formula = writing_score ~ gender + lunch_type + test_prep +
##     parent_educ + ethnic_group + wkly_study_hours, data = df_num)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.917  -8.391   0.613   9.143  29.293
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       65.3125     3.8874  16.801  < 2e-16 ***
## gender            -9.1792     1.0698  -8.581  < 2e-16 ***
## lunch_type         9.4976     1.1151   8.517  < 2e-16 ***
## test_prep         -9.0360     1.1163  -8.094 3.40e-15 ***
## parent_educ        2.3242     0.3566   6.519 1.54e-10 ***
## ethnic_group       2.1684     0.4562   4.753 2.53e-06 ***
## 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
```

$\hat{Writing Score} = 65.3125 + 9.4976 * \text{Lunch Type} - 9.1792 * \text{Gender} - 9.0360 * \text{Test Prep} + 2.3242 * \text{Parent Education} + 2.1684 * \text{Ethnic Group} + 1.1762 * \text{Weekly Study Hours}$

Check Assumptions

Math

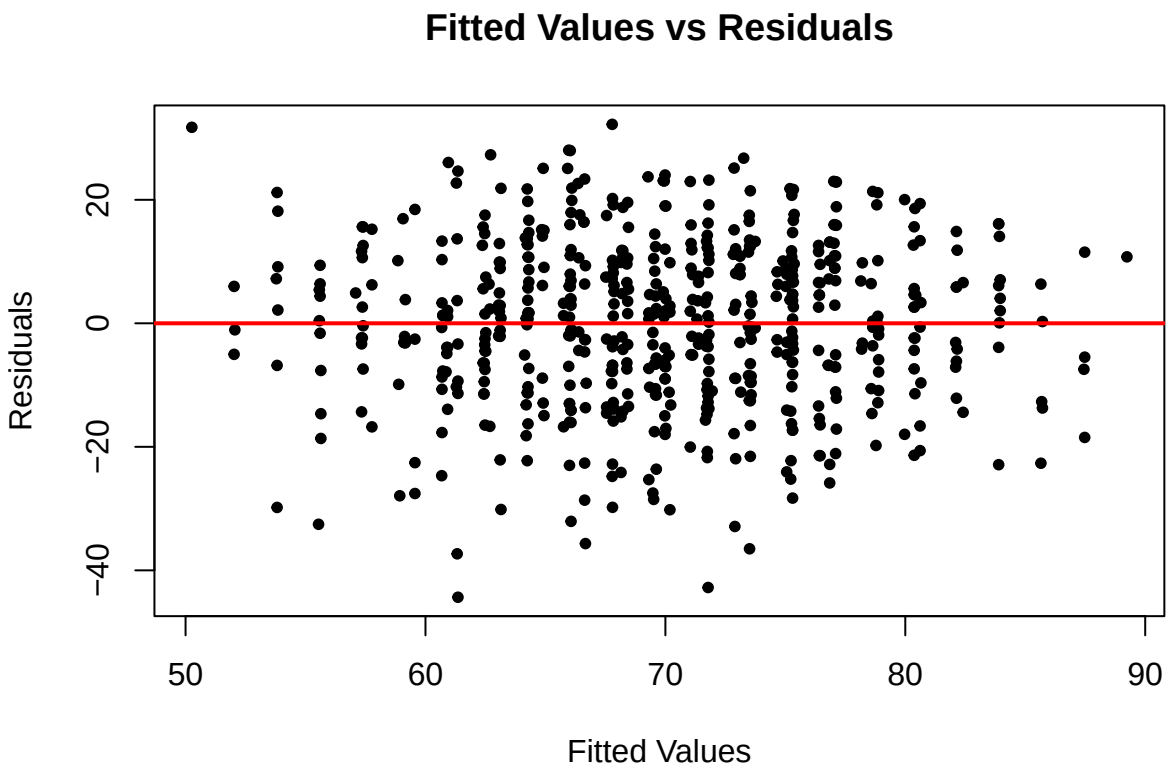
```
plot(math_final$fitted.values, math_final$residuals,
     xlab = "Fitted Values",
     ylab = "Residuals",
     main = "Fitted Values vs Residuals",
     pch = 20)

# Adding a horizontal line at 0
abline(h = 0, col = "red", lwd = 2)
```



The residuals show no pattern when plotted against independent variables indicates the relationship between the independent variables and the dependent variable is linear. Also, the residuals are independent of each other. The variance of errors is also constant. These attributes ensure the reliability and validity of the predictive model.

Reading

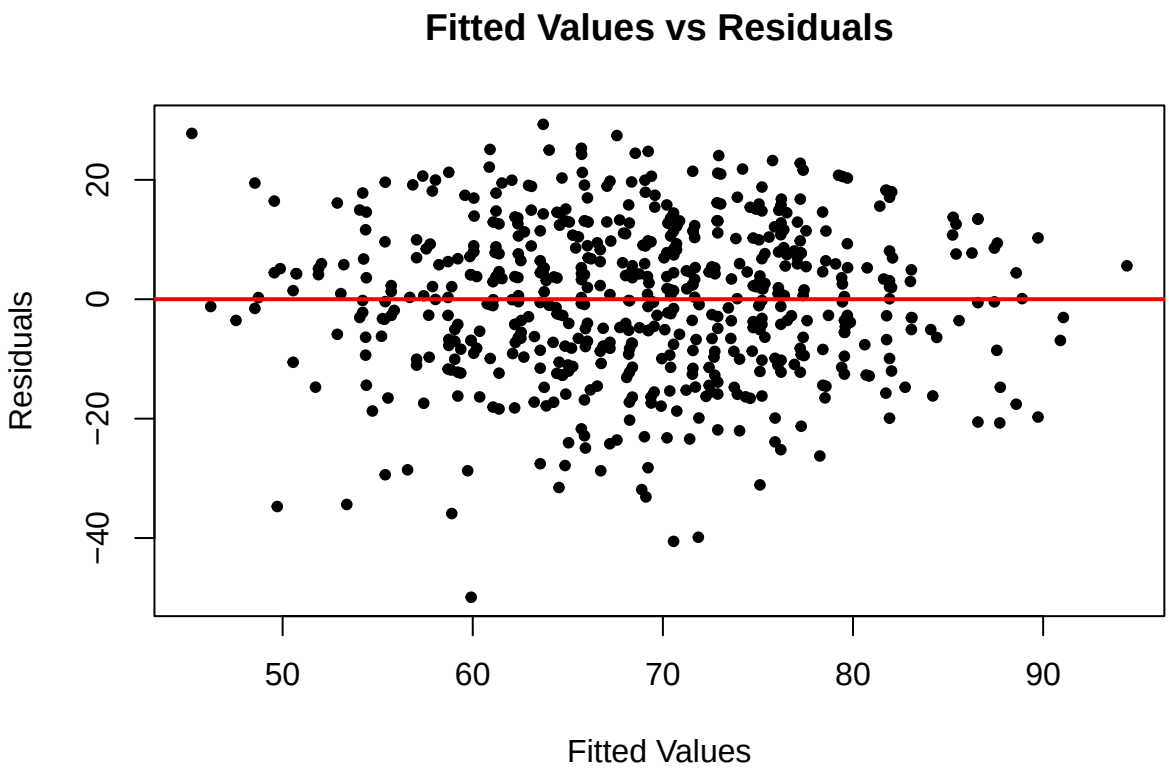
```
plot(reading_final$fitted.values, reading_final$residuals,
     xlab = "Fitted Values",
     ylab = "Residuals",
     main = "Fitted Values vs Residuals",
     pch = 20)

# Adding a horizontal line at 0
abline(h = 0, col = "red", lwd = 2)
```



The residuals show no pattern when plotted against independent variables indicates the relationship between the independent variables and the dependent variable is linear. Also, the residuals are independent of each other. The variance of errors is also constant. These attributes ensure the reliability and validity of the predictive model.

Writing

```
plot(writing_final$fitted.values, writing_final$residuals,
     xlab = "Fitted Values",
     ylab = "Residuals",
     main = "Fitted Values vs Residuals",
     pch = 20)

# Adding a horizontal line at 0
abline(h = 0, col = "red", lwd = 2)
```



The residuals show no pattern when plotted against independent variables indicates the relationship between the independent variables and the dependent variable is linear. Also, the residuals are independent of each other. The variance of errors is also constant. These attributes ensure the reliability and validity of the predictive model.

Performance

```
set.seed(123)

# Create trainControl object for 5-fold cross-validation
control <- trainControl(method = "cv", number = 5)

math_model = train(math_score ~ lunch_type + ethnic_group + test_prep + gender + parent_educ + wkly_stu
                   data = df_num,
                   method = "lm",
                   trControl = control)

reading_model = train(reading_score ~ lunch_type + gender + test_prep + parent_educ + ethnic_group,
                      data = df_num,
                      method = "lm",
                      trControl = control)
writing_model = train(writing_score ~ gender + lunch_type + test_prep + parent_educ + ethnic_group + wk
                      data = df_num,
                      method = "lm",
```

```
                      trControl = control)
math_model$results |>
  knitr::kable()
```

intercept	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
TRUE	13.79213	0.2682465	11.14582	0.99906	0.0203526	0.6808633

```
reading_model$results |>
  knitr::kable()
```

intercept	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
TRUE	13.42626	0.21953	10.90863	0.3631765	0.0430995	0.4314327

```
writing_model$results |>
  knitr::kable()
```

intercept	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
TRUE	12.89674	0.3118989	10.37449	0.8072032	0.0628649	0.7391721

The math, reading, writing model explain 26.8%, 22.0%, 31.2% of the score's variance respectively, with a RMSE of 13.7, 13.4, 12.9 and a MAE of 11.1, 10.9, 10.4. These models have an average absolute difference 11.2, 10.9, 10.4 between true score and predicted score. Overall, these three multiple linear regression model indicates a reasonably good fit and predictive accuracy.

Leverage one score

Effect of adding Writing, Reading scores on Maths model

```
maths_enhance=lm(math_score~reading_score+writing_score+lunch_type+ethnic_group+test_prep+gender+parent
mse_maths_enhance=mean((df_num$math_score-predict(maths_enhance,newdata=df_num))^2)
mse_maths=mean((df_num$math_score-predict(math_forward_func_fit,newdata=df_num))^2)
```

Effect of adding Maths, Writing scores on Reading model

```
reading_enhance=lm(reading_score~math_score+writing_score+lunch_type+gender+test_prep+parent_educ+ethni
mse_reading_enhance=mean((df_num$reading_score-predict(reading_enhance,newdata=df_num))^2)
mse_reading=mean((df_num$reading_score-predict(reading_criteria_fit,newdata=df_num))^2)
```

Effect of adding Maths, Reading scores on Writing model

```
writing_enhance=lm(writing_score~reading_score+math_score+lunch_type+gender+test_prep+parent_educ+ethni
mse_writing_enhance=mean((df_num$writing_score-predict(writing_enhance,newdata=df_num))^2)
mse_writing=mean((df_num$writing_score-predict(writing_forward_func_fit,newdata=df_num))^2)
```

Combined table

```r
tibble(
  model_name=c("maths_reading+writing","maths_original","reading_maths+writing","reading_original","wri
  MSE=c(mse_maths_enhance,mse_maths,mse_reading_enhance,mse_reading,mse_writing_enhance,mse_writing)
)|>knitr::kable()
```

model_name	MSE
maths_reading+writing	31.67574
maths_original	187.71873
reading_maths+writing	16.36549
reading_original	177.64685
writing_maths+reading	12.79564
writing_original	163.35751

We can see that the MSE all significantly decreased after adding other scores to fit one score's model, indicating that leveraging other scores to enhance one score's model is possible and successful.

Test potential overfitting issue - maths example

```r
# Split data into training and test sets
set.seed(123)
train_index = createDataPartition(df_num$math_score, p = 0.7, list = FALSE)
train_data = df_num[train_index, ]
test_data = df_num[-train_index, ]

# Train the model on the training set
maths_enhance = lm(math_score ~ reading_score + writing_score + lunch_type + ethnic_group + test_prep +

# Make predictions on the test set
predictions = predict(maths_enhance, newdata = test_data)

# Evaluate model performance on the test set (MSPE)
mse_test = mean((test_data$math_score - predictions)^2)
print(paste("MSE on Test Set:", mse_test))
```

```
## [1] "MSE on Test Set: 31.2708965516677"
```

```r
# Perform k-fold cross-validation (e.g., 5-fold)
set.seed(123)
folds = createFolds(df_num$math_score, k = 5, list = TRUE)

mse_cv = numeric(length(folds))

for (i in seq_along(folds)) {
  train_indices = unlist(folds[-i])
  test_indices = folds[[i]]

  train_data_cv = df_num[train_indices, ]
  test_data_cv = df_num[test_indices, ]

  model_cv = lm(math_score ~ reading_score + writing_score + lunch_type + ethnic_group + test_prep + ge
```

```
  predictions_cv = predict(model_cv, newdata = test_data_cv)
  mse_cv[i] = mean((test_data_cv$math_score - predictions_cv)^2)
}

mean_mse_cv = mean(mse_cv)
print(paste("Mean MSE across Folds:", mean_mse_cv))
```

```
## [1] "Mean MSE across Folds: 32.5279607340191"
```

The model performs as well on test set and also across folds. Adding other scores to one score's best fit model did enhance the model.