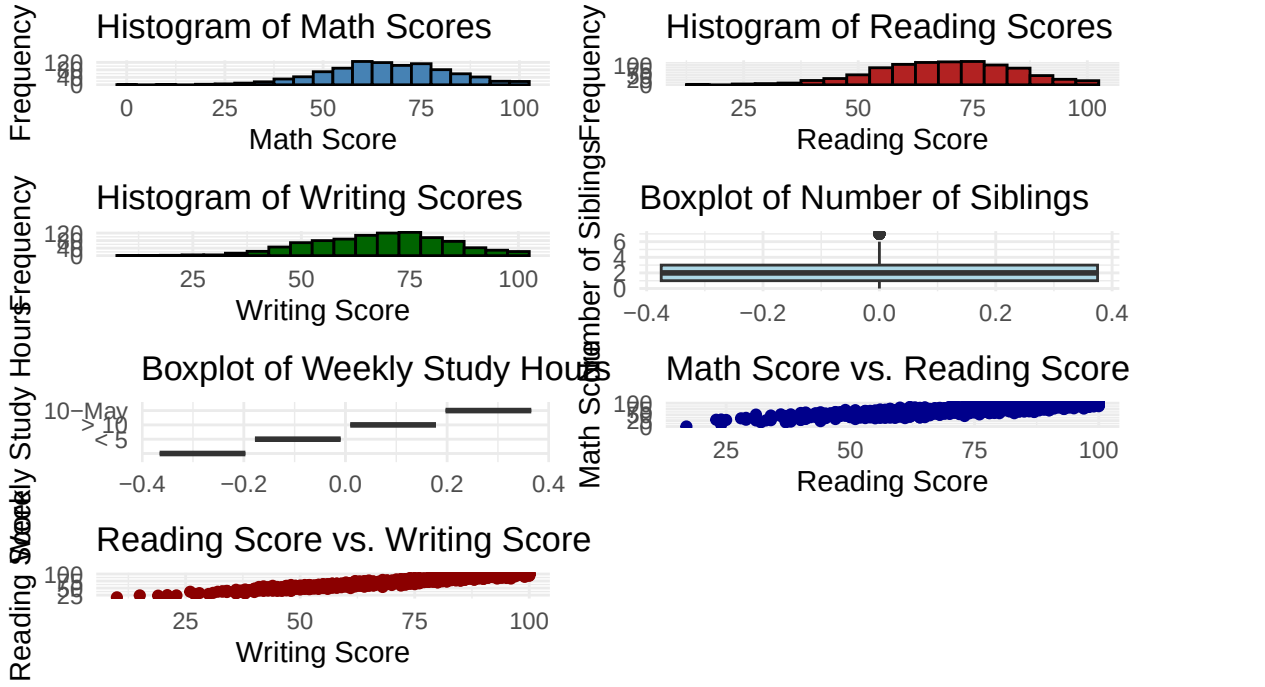
bm-project

group-27

2024-12-19

```
data <- read.csv("Project_1_data.csv")

hist_math <- ggplot(data, aes(x = MathScore)) +
  geom_histogram(binwidth = 5, fill = "steelblue", color = "black") +
  labs(title = "Histogram of Math Scores", x = "Math Score", y = "Frequency")

hist_reading <- ggplot(data, aes(x = ReadingScore)) +
  geom_histogram(binwidth = 5, fill = "firebrick", color = "black") +
  labs(title = "Histogram of Reading Scores", x = "Reading Score", y = "Frequency")

hist_writing <- ggplot(data, aes(x = WritingScore)) +
  geom_histogram(binwidth = 5, fill = "darkgreen", color = "black") +
  labs(title = "Histogram of Writing Scores", x = "Writing Score", y = "Frequency")
box_nr_siblings <- ggplot(data, aes(y = NrSiblings)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Boxplot of Number of Siblings", y = "Number of Siblings")

box_wkly_study_hours <- ggplot(data, aes(y = WklyStudyHours)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Boxplot of Weekly Study Hours", y = "Weekly Study Hours")
scatter_math_reading <- ggplot(data, aes(x = ReadingScore, y = MathScore)) +
  geom_point(color = "darkblue") +
  labs(title = "Math Score vs. Reading Score", x = "Reading Score", y = "Math Score")

scatter_reading_writing <- ggplot(data, aes(x = WritingScore, y = ReadingScore)) +
  geom_point(color = "darkred") +
  labs(title = "Reading Score vs. Writing Score", x = "Writing Score", y = "Reading Score")
# Convert categorical variables to factors if needed
data <- data %>%
  mutate(across(c(Gender, EthnicGroup, ParentEduc, LunchType, TestPrep,
                  ParentMaritalStatus, PracticeSport, IsFirstChild,
                  TransportMeans, WklyStudyHours), as.factor))

pairwise_plots <- ggpairs(data,
  columns = c("MathScore", "ReadingScore", "WritingScore", "NrSiblings", "WklyStudyHours"),
  aes(color = Gender),
  lower = list(continuous = "smooth"),
  upper = list(continuous = "cor"),
  diag = list(continuous = "densityDiag"))
grid.arrange(
  hist_math, hist_reading, hist_writing,
  box_nr_siblings, box_wkly_study_hours,
```

```
  scatter_math_reading, scatter_reading_writing,
  ncol = 2
)
```




```
data$Gender <- as.factor(data$Gender)
data$EthnicGroup <- as.factor(data$EthnicGroup)
data$ParentEduc <- as.factor(data$ParentEduc)
data$LunchType <- as.factor(data$LunchType)
data$TestPrep <- as.factor(data$TestPrep)
data$ParentMaritalStatus <- as.factor(data$ParentMaritalStatus)
data$PracticeSport <- as.factor(data$PracticeSport)
data$IsFirstChild <- as.factor(data$IsFirstChild)
data$TransportMeans <- as.factor(data$TransportMeans)
data$WklyStudyHours <- as.factor(data$WklyStudyHours)

data <- na.omit(data)
```

```
data[data == "" | data == " "] <- NA
data <- na.omit(data)
```

```
data_dict <- tibble(
  Variable = c(
    "Gender", "EthnicGroup", "ParentEduc", "LunchType", "TestPrep",
    "ParentMaritalStatus", "PracticeSport", "IsFirstChild", "NrSiblings",
    "TransportMeans", "WklyStudyHours", "MathScore", "ReadingScore", "WritingScore"
  ),
  Description = c(
    "Gender of the student (male/female)",
    "Ethnic group of the student (group A to E)",
    "Parent(s) education background (from some_highschool to master's degree)",
    "School lunch type (standard or free/reduced)",
    "Test preparation course followed (completed or none)",
    "Parent(s) marital status (married/single/widowed/divorced)",
```

```
    "How often the student practices sport (never/sometimes/regularly)",
    "If the child is the first child in the family (yes/no)",
    "Number of siblings the student has (0 to 7)",
    "Means of transport to school (schoolbus/private)",
    "Weekly self-study hours (less that 5 hours; between 5 and 10 hours; more than
10 hours)",
    "Math test score (0-100)",
    "Reading test score (0-100)",
    "Writing test score (0-100)"
  )
)

data_dict %>%
  knitr::kable(caption = "Data Dictionary") %>%
  kable_styling() %>%
  column_spec(1, width = "12em") %>%
  column_spec(2, width = "32em")
```

Table 1: Data Dictionary

Variable	Description
Gender	Gender of the student (male/female)
EthnicGroup	Ethnic group of the student (group A to E)
ParentEduc	Parent(s) education background (from some_highschool to master's degree)
LunchType	School lunch type (standard or free/reduced)
TestPrep	Test preparation course followed (completed or none)
ParentMaritalStatus	Parent(s) marital status (married/single/widowed/divorced)
PracticeSport	How often the student practices sport (never/sometimes/regularly)
IsFirstChild	If the child is the first child in the family (yes/no)
NrSiblings	Number of siblings the student has (0 to 7)
TransportMeans	Means of transport to school (schoolbus/private)
WklyStudyHours	Weekly self-study hours (less that 5 hours; between 5 and 10 hours; more than 10 hours)
MathScore	Math test score (0-100)
ReadingScore	Reading test score (0-100)
WritingScore	Writing test score (0-100)

Table 2: Summary Statistics for all Numeric Variables

Variable Name	Mean	SD	Median	IQR	Max	Min
NrSiblings	2.139693	1.481712	2	2	7	0
MathScore	66.676320	16.113744	67	22	100	0
ReadingScore	69.846678	15.166662	70	21	100	17
WritingScore	68.901192	15.550000	69	21	100	10

```
categorical_table <- data |>
  summarize(
    gender_Male = sum(Gender == "male", na.rm = TRUE),
    gender_Female = sum(Gender == "female", na.rm = TRUE),
```

```
    ethnicgroup_A = sum(EthnicGroup == "group A", na.rm = TRUE),
    ethnicgroup_B = sum(EthnicGroup == "group B", na.rm = TRUE),
    ethnicgroup_C = sum(EthnicGroup == "group C", na.rm = TRUE),
    ethnicgroup_D = sum(EthnicGroup == "group D", na.rm = TRUE),
    ethnicgroup_E = sum(EthnicGroup == "group E", na.rm = TRUE),

    parenteduc_SomeHighSchool = sum(ParentEduc == "some college", na.rm = TRUE),
    parenteduc_HighSchool = sum(ParentEduc == "some high School", na.rm = TRUE),
    parenteduc_Associates = sum(ParentEduc == "associate's degree high school", na.rm = TRUE),
    parenteduc_Bachelors = sum(ParentEduc == "bachelor's degree", na.rm = TRUE),
    parenteduc_Masters = sum(ParentEduc == "master's degree", na.rm = TRUE),

    lunchtime_Standard = sum(LunchType == "standard", na.rm = TRUE),
    lunchtime_FreeReduced = sum(LunchType == "free/reduced", na.rm = TRUE),

    testprep_Completed = sum(TestPrep == "completed", na.rm = TRUE),
    testprep_None = sum(TestPrep == "none", na.rm = TRUE),

    parentmaritalstatus_Married = sum(ParentMaritalStatus == "married", na.rm = TRUE),
    parentmaritalstatus_Single = sum(ParentMaritalStatus == "single", na.rm = TRUE),
    parentmaritalstatus_Widowed = sum(ParentMaritalStatus == "widowed", na.rm = TRUE),
    parentmaritalstatus_Divorced = sum(ParentMaritalStatus == "divorced", na.rm = TRUE),

    practicesport_Never = sum(PracticeSport == "never", na.rm = TRUE),
    practicesport_Sometimes = sum(PracticeSport == "sometimes", na.rm = TRUE),
    practicesport_Regularly = sum(PracticeSport == "regularly", na.rm = TRUE),

    isfirstchild_Yes = sum(IsFirstChild == "yes", na.rm = TRUE),
    isfirstchild_No = sum(IsFirstChild == "no", na.rm = TRUE),

    transportmeans_SchoolBus = sum(TransportMeans == "school_bus", na.rm = TRUE),
    transportmeans_Private = sum(TransportMeans == "private", na.rm = TRUE),

    wklystudyhours_LessThan5 = sum(WklyStudyHours == "< 5", na.rm = TRUE),
    wklystudyhours_5to10 = sum(WklyStudyHours == "10-May", na.rm = TRUE),
    wklystudyhours_MoreThan10 = sum(WklyStudyHours == "> 10", na.rm = TRUE)
  )


categorical_final <- data.frame(
  Variable = c("Gender Male", "Gender Female",
               "EthnicGroup A", "EthnicGroup B", "EthnicGroup C", "EthnicGroup D", "EthnicGroup E",
               "ParentEduc Some High School", "ParentEduc High School", "ParentEduc Associates",
               "ParentEduc Bachelors", "ParentEduc Masters",
               "LunchType Standard", "LunchType Free/Reduced",
               "TestPrep Completed", "TestPrep None",
               "ParentMaritalStatus Married", "ParentMaritalStatus Single", "ParentMaritalStatus Widowe
               "PracticeSport Never", "PracticeSport Sometimes", "PracticeSport Regularly",
               "IsFirstChild Yes", "IsFirstChild No",
               "TransportMeans SchoolBus", "TransportMeans Private",
               "WklyStudyHours Less than 5", "WklyStudyHours 5-10", "WklyStudyHours More than 10"),
  Count = c(
    categorical_table$gender_Male, categorical_table$gender_Female,
```

```
    categorical_table$ethnicgroup_A, categorical_table$ethnicgroup_B, categorical_table$ethnicgroup_C, c
    categorical_table$parenteduc_SomeHighSchool, categorical_table$parenteduc_HighSchool, categorical_ta
    categorical_table$parenteduc_Bachelors, categorical_table$parenteduc_Masters,
    categorical_table$lunchtime_Standard, categorical_table$lunchtime_FreeReduced,
    categorical_table$testprep_Completed, categorical_table$testprep_None,
    categorical_table$parentmaritalstatus_Married, categorical_table$parentmaritalstatus_Single, categor
    categorical_table$practicesport_Never, categorical_table$practicesport_Sometimes, categorical_table$
    categorical_table$isfirstchild_Yes, categorical_table$isfirstchild_No,
    categorical_table$transportmeans_SchoolBus, categorical_table$transportmeans_Private,
    categorical_table$wklystudyhours_LessThan5, categorical_table$wklystudyhours_5to10, categorical_tabl
  ),
  Proportion = round(c(
    categorical_table$gender_Male / nrow(data), categorical_table$gender_Female / nrow(data),
    categorical_table$ethnicgroup_A / nrow(data), categorical_table$ethnicgroup_B / nrow(data), categori
    categorical_table$ethnicgroup_D / nrow(data), categorical_table$ethnicgroup_E / nrow(data),
    categorical_table$parenteduc_SomeHighSchool / nrow(data), categorical_table$parenteduc_HighSchool /
    categorical_table$parenteduc_Associates / nrow(data), categorical_table$parenteduc_Bachelors / nrow
    categorical_table$parenteduc_Masters / nrow(data),
    categorical_table$lunchtime_Standard / nrow(data), categorical_table$lunchtime_FreeReduced / nrow(da
    categorical_table$testprep_Completed / nrow(data), categorical_table$testprep_None / nrow(data),
    categorical_table$parentmaritalstatus_Married / nrow(data), categorical_table$parentmaritalstatus_S
    categorical_table$parentmaritalstatus_Widowed / nrow(data), categorical_table$parentmaritalstatus_D
    categorical_table$practicesport_Never / nrow(data), categorical_table$practicesport_Sometimes / nrow
    categorical_table$practicesport_Regularly / nrow(data),
    categorical_table$isfirstchild_Yes / nrow(data), categorical_table$isfirstchild_No / nrow(data),
    categorical_table$transportmeans_SchoolBus / nrow(data), categorical_table$transportmeans_Private /
    categorical_table$wklystudyhours_LessThan5 / nrow(data), categorical_table$wklystudyhours_5to10 / nr
    categorical_table$wklystudyhours_MoreThan10 / nrow(data)
  ), 4)
)

knitr::kable(categorical_final, col.names = c("Variable Name and Levels", "Count", "Proportion"),
             caption = "Summary Statistics for all Categorical Variables", format = "pipe")
```

Table 3: Summary Statistics for all Categorical Variables

Variable Name and Levels	Count	Proportion
Gender Male	272	0.4634
Gender Female	315	0.5366
EthnicGroup A	50	0.0852
EthnicGroup B	123	0.2095
EthnicGroup C	174	0.2964
EthnicGroup D	155	0.2641
EthnicGroup E	85	0.1448
ParentEduc Some High School	116	0.1976
ParentEduc High School	0	0.0000
ParentEduc Associates	0	0.0000
ParentEduc Bachelors	71	0.1210
ParentEduc Masters	39	0.0664
LunchType Standard	381	0.6491
LunchType Free/Reduced	206	0.3509
TestPrep Completed	208	0.3543
TestPrep None	379	0.6457

Variable Name and Levels	Count	Proportion
ParentMaritalStatus Married	343	0.5843
ParentMaritalStatus Single	137	0.2334
ParentMaritalStatus Widowed	15	0.0256
ParentMaritalStatus Divorced	92	0.1567
PracticeSport Never	68	0.1158
PracticeSport Sometimes	301	0.5128
PracticeSport Regularly	218	0.3714
IsFirstChild Yes	395	0.6729
IsFirstChild No	192	0.3271
TransportMeans SchoolBus	358	0.6099
TransportMeans Private	229	0.3901
WklyStudyHours Less than 5	154	0.2624
WklyStudyHours 5-10	329	0.5605
WklyStudyHours More than 10	104	0.1772

```
math_full_model <- lm(MathScore ~ Gender + EthnicGroup + ParentEduc + LunchType + TestPrep +
                        ParentMaritalStatus + PracticeSport + IsFirstChild + NrSiblings +
                        TransportMeans + WklyStudyHours, data = data)
math_null_model = lm(MathScore ~ 1, data = data)
math_selected_model = step(math_null_model,
scope = list(lower = formula(math_null_model),
upper = formula(math_full_model)))
```

```
## Start:  AIC=3264.33
## MathScore ~ 1
##
##                       Df Sum of Sq    RSS    AIC
## + LunchType            1   22340.6 129816 3173.1
## + EthnicGroup          4   11630.1 140526 3225.7
## + Gender               1    5114.8 147042 3246.3
## + TestPrep             1    4114.3 148042 3250.2
## + ParentEduc           5    4397.1 147759 3257.1
## + WklyStudyHours       2    2365.3 149791 3259.1
## + ParentMaritalStatus  3    2625.8 149531 3260.1
## + NrSiblings           1     615.0 151541 3264.0
## <none>                             152157 3264.3
## + IsFirstChild         1     132.5 152024 3265.8
## + TransportMeans       1       0.3 152156 3266.3
## + PracticeSport        2      17.8 152139 3268.3
##
## Step:  AIC=3173.12
## MathScore ~ LunchType
##
##                       Df Sum of Sq    RSS    AIC
## + EthnicGroup          4   10097.8 119718 3133.6
## + TestPrep             1    4711.5 125104 3153.4
## + Gender               1    4049.1 125767 3156.5
## + ParentEduc           5    4657.6 125158 3161.7
## + ParentMaritalStatus  3    2481.0 127335 3167.8
## + WklyStudyHours       2    2008.6 127807 3168.0
## + NrSiblings           1     601.2 129215 3172.4
## <none>                             129816 3173.1
```

```
## + IsFirstChild         1      93.5 129722 3174.7
## + TransportMeans       1       1.5 129814 3175.1
## + PracticeSport        2      76.4 129739 3176.8
## - LunchType            1   22340.6 152157 3264.3
##
## Step:  AIC=3133.59
## MathScore ~ LunchType + EthnicGroup
##
##                       Df Sum of Sq    RSS    AIC
## + TestPrep             1    4077.4 115641 3115.2
## + Gender               1    3574.9 116143 3117.8
## + ParentMaritalStatus  3    3208.1 116510 3123.6
## + ParentEduc           5    3901.2 115817 3124.1
## + WklyStudyHours       2    1623.3 118095 3129.6
## + NrSiblings           1     669.1 119049 3132.3
## <none>                             119718 3133.6
## + IsFirstChild         1      82.1 119636 3135.2
## + TransportMeans       1       1.2 119717 3135.6
## + PracticeSport        2     178.0 119540 3136.7
## - EthnicGroup          4   10097.8 129816 3173.1
## - LunchType            1   20808.3 140526 3225.7
##
## Step:  AIC=3115.25
## MathScore ~ LunchType + EthnicGroup + TestPrep
##
##                       Df Sum of Sq    RSS    AIC
## + Gender               1    3258.7 112382 3100.5
## + ParentMaritalStatus  3    3343.5 112297 3104.0
## + ParentEduc           5    3694.7 111946 3106.2
## + WklyStudyHours       2    1226.5 114414 3113.0
## + NrSiblings           1     527.9 115113 3114.6
## <none>                             115641 3115.2
## + IsFirstChild         1      34.0 115607 3117.1
## + TransportMeans       1      12.8 115628 3117.2
## + PracticeSport        2     113.8 115527 3118.7
## - TestPrep             1    4077.4 119718 3133.6
## - EthnicGroup          4    9463.6 125104 3153.4
## - LunchType            1   21399.9 137041 3212.9
##
## Step:  AIC=3100.47
## MathScore ~ LunchType + EthnicGroup + TestPrep + Gender
##
##                       Df Sum of Sq    RSS    AIC
## + ParentEduc           5    4081.3 108301 3088.8
## + ParentMaritalStatus  3    3157.1 109225 3089.7
## + WklyStudyHours       2    1243.9 111138 3097.9
## + NrSiblings           1     631.9 111750 3099.2
## <none>                             112382 3100.5
## + IsFirstChild         1      24.9 112357 3102.3
## + TransportMeans       1       7.4 112375 3102.4
## + PracticeSport        2     118.4 112264 3103.9
## - Gender               1    3258.7 115641 3115.2
## - TestPrep             1    3761.1 116143 3117.8
## - EthnicGroup          4    9023.7 121406 3137.8
```

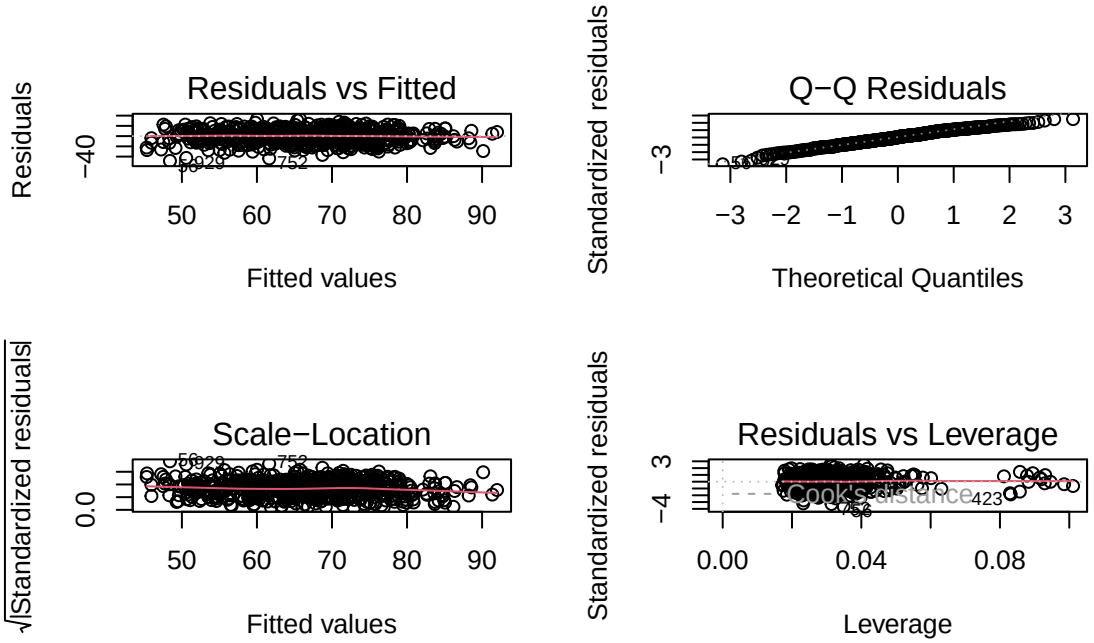
```
## - LunchType            1   20463.2 132845 3196.7
##
## Step:  AIC=3088.76
## MathScore ~ LunchType + EthnicGroup + TestPrep + Gender + ParentEduc
##
##                        Df Sum of Sq    RSS    AIC
## + ParentMaritalStatus  3    2912.6 105388 3078.8
## + WklyStudyHours       2    1385.3 106915 3085.2
## + NrSiblings           1     681.7 107619 3087.0
## <none>                             108301 3088.8
## + IsFirstChild         1      47.0 108254 3090.5
## + TransportMeans       1       2.2 108298 3090.7
## + PracticeSport        2     172.1 108129 3091.8
## - ParentEduc           5    4081.3 112382 3100.5
## - TestPrep             1    3515.1 111816 3105.5
## - Gender               1    3645.3 111946 3106.2
## - EthnicGroup          4    8255.3 116556 3123.9
## - LunchType            1   20674.9 128976 3189.3
##
## Step:  AIC=3078.75
## MathScore ~ LunchType + EthnicGroup + TestPrep + Gender + ParentEduc +
##     ParentMaritalStatus
##
##                        Df Sum of Sq    RSS    AIC
## + WklyStudyHours       2    1297.1 104091 3075.5
## + NrSiblings           1     582.0 104806 3077.5
## <none>                             105388 3078.8
## + IsFirstChild         1     118.2 105270 3080.1
## + TransportMeans       1      11.1 105377 3080.7
## + PracticeSport        2     153.5 105235 3081.9
## - ParentMaritalStatus  3    2912.6 108301 3088.8
## - ParentEduc           5    3836.8 109225 3089.7
## - Gender               1    3444.7 108833 3095.6
## - TestPrep             1    3637.0 109025 3096.7
## - EthnicGroup          4    8892.7 114281 3118.3
## - LunchType            1   20565.4 125953 3181.4
##
## Step:  AIC=3075.48
## MathScore ~ LunchType + EthnicGroup + TestPrep + Gender + ParentEduc +
##     ParentMaritalStatus + WklyStudyHours
##
##                        Df Sum of Sq    RSS    AIC
## + NrSiblings           1     629.8 103461 3073.9
## <none>                             104091 3075.5
## + IsFirstChild         1      96.4 103995 3076.9
## + TransportMeans       1      17.3 104074 3077.4
## + PracticeSport        2     131.1 103960 3078.7
## - WklyStudyHours       2    1297.1 105388 3078.8
## - ParentMaritalStatus  3    2824.4 106915 3085.2
## - ParentEduc           5    3972.8 108064 3087.5
## - TestPrep             1    3265.4 107356 3091.6
## - Gender               1    3507.7 107599 3092.9
## - EthnicGroup          4    8640.2 112731 3114.3
## - LunchType            1   20147.8 124239 3177.3
```

```
##
## Step:  AIC=3073.92
## MathScore ~ LunchType + EthnicGroup + TestPrep + Gender + ParentEduc +
##     ParentMaritalStatus + WklyStudyHours + NrSiblings
##
##                       Df Sum of Sq    RSS    AIC
## <none>                             103461 3073.9
## + IsFirstChild         1     142.1 103319 3075.1
## - NrSiblings           1     629.8 104091 3075.5
## + TransportMeans       1      18.0 103443 3075.8
## + PracticeSport        2     130.1 103331 3077.2
## - WklyStudyHours       2    1344.8 104806 3077.5
## - ParentMaritalStatus  3    2726.0 106187 3083.2
## - ParentEduc           5    4053.0 107514 3086.5
## - TestPrep             1    3115.5 106577 3089.3
## - Gender               1    3650.4 107112 3092.3
## - EthnicGroup          4    8726.4 112188 3113.5
## - LunchType            1   20082.2 123543 3176.1
```

```
math_selected_model
```

```
##
## Call:
## lm(formula = MathScore ~ LunchType + EthnicGroup + TestPrep +
##     Gender + ParentEduc + ParentMaritalStatus + WklyStudyHours +
##     NrSiblings, data = data)
##
## Coefficients:
##                  (Intercept)            LunchTypestandard
##                     50.96821                     12.32626
##          EthnicGroupgroup B           EthnicGroupgroup C
##                     -0.12625                     -0.06815
##          EthnicGroupgroup D           EthnicGroupgroup E
##                      3.71549                     11.18161
##                TestPrepnone                   Gendermale
##                     -4.92090                      5.08556
## ParentEducbachelor's degree        ParentEduchigh school
##                      1.71868                     -5.06167
##   ParentEducmaster's degree       ParentEducsome college
##                      1.87883                     -1.59208
##  ParentEducsome high school   ParentMaritalStatusmarried
##                     -4.87887                      5.41133
##   ParentMaritalStatussingle   ParentMaritalStatuswidowed
##                      2.13481                      7.48771
##          WklyStudyHours> 10         WklyStudyHours10-May
##                      3.04378                      3.60274
##                  NrSiblings
##                      0.71090
```

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




```
par(mfrow = c(1, 1))
```

```
reading_full_model <- lm(ReadingScore ~ Gender + EthnicGroup + ParentEduc + LunchType + TestPrep +
                         ParentMaritalStatus + PracticeSport + IsFirstChild + NrSiblings +
                         TransportMeans + WklyStudyHours, data = data)
reading_null_model = lm(ReadingScore ~ 1, data = data)
reading_selected_model = step(reading_null_model,
scope = list(lower = formula(reading_null_model),
upper = formula(reading_full_model)))
```

```
## Start:  AIC=3193.22
## ReadingScore ~ 1
##
##                        Df Sum of Sq    RSS    AIC
## + LunchType             1    8876.3 125920 3155.2
## + Gender                1    7428.6 127368 3161.9
## + ParentEduc            5    7361.4 127435 3170.3
## + TestPrep              1    5190.3 129606 3172.2
## + EthnicGroup           4    4266.3 130530 3182.3
## + ParentMaritalStatus   3    1950.9 132845 3190.7
## + WklyStudyHours        2    1301.7 133494 3191.5
## <none>                               134796 3193.2
## + NrSiblings            1     270.2 134526 3194.0
## + IsFirstChild          1     202.0 134594 3194.3
## + TransportMeans        1      18.5 134778 3195.1
## + PracticeSport         2     442.6 134354 3195.3
##
## Step:  AIC=3155.24
## ReadingScore ~ LunchType
##
##                        Df Sum of Sq    RSS    AIC
## + Gender                1    8344.3 117576 3117.0
## + ParentEduc            5    7664.9 118255 3128.4
## + TestPrep              1    5609.0 120311 3130.5
```

```
## + EthnicGroup          4    3696.5 122223 3145.7
## + ParentMaritalStatus  3    1889.7 124030 3152.4
## + WklyStudyHours       2    1008.3 124912 3154.5
## <none>                              125920 3155.2
## + NrSiblings           1     264.4 125655 3156.0
## + IsFirstChild         1     170.4 125749 3156.4
## + TransportMeans       1      10.1 125910 3157.2
## + PracticeSport        2     388.1 125532 3157.4
## - LunchType            1    8876.3 134796 3193.2
##
## Step:  AIC=3116.99
## ReadingScore ~ LunchType + Gender
##
##                        Df Sum of Sq    RSS    AIC
## + TestPrep              1    6206.4 111369 3087.2
## + ParentEduc            5    6612.7 110963 3093.0
## + EthnicGroup           4    4042.9 113533 3104.4
## + ParentMaritalStatus   3    2015.2 115560 3112.8
## + WklyStudyHours        2     963.4 116612 3116.2
## <none>                              117576 3117.0
## + IsFirstChild          1     204.6 117371 3118.0
## + NrSiblings            1     171.4 117404 3118.1
## + TransportMeans        1      20.0 117556 3118.9
## + PracticeSport         2     331.5 117244 3119.3
## - Gender                1    8344.3 125920 3155.2
## - LunchType             1    9792.0 127368 3161.9
##
## Step:  AIC=3087.16
## ReadingScore ~ LunchType + Gender + TestPrep
##
##                        Df Sum of Sq    RSS    AIC
## + ParentEduc            5    6036.7 105333 3064.4
## + EthnicGroup           4    3694.0 107675 3075.4
## + ParentMaritalStatus   3    2156.8 109212 3081.7
## <none>                              111369 3087.2
## + WklyStudyHours        2     688.2 110681 3087.5
## + TransportMeans        1     104.0 111265 3088.6
## + IsFirstChild          1     102.5 111267 3088.6
## + NrSiblings            1      84.6 111285 3088.7
## + PracticeSport         2     311.4 111058 3089.5
## - TestPrep              1    6206.4 117576 3117.0
## - Gender                1    8941.7 120311 3130.5
## - LunchType             1   10288.6 121658 3137.0
##
## Step:  AIC=3064.44
## ReadingScore ~ LunchType + Gender + TestPrep + ParentEduc
##
##                        Df Sum of Sq    RSS    AIC
## + EthnicGroup           4    3179.3 102153 3054.4
## + ParentMaritalStatus   3    2007.8 103325 3059.2
## + WklyStudyHours        2     883.2 104449 3063.5
## <none>                              105333 3064.4
## + IsFirstChild          1     138.3 105194 3065.7
## + NrSiblings            1     133.5 105199 3065.7
```

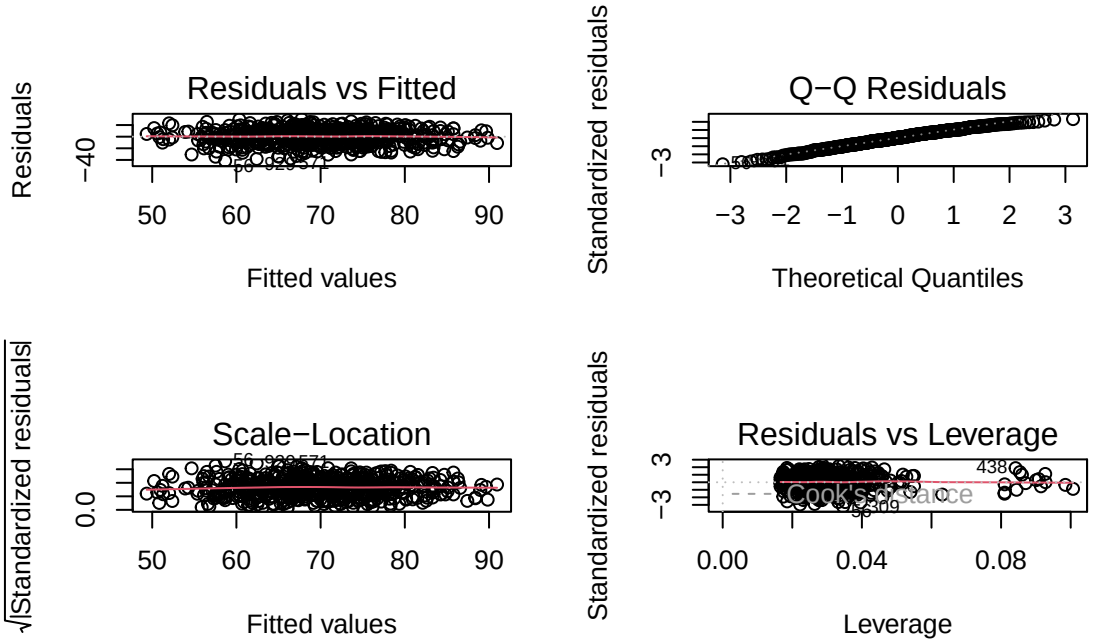
```
## + TransportMeans       1      31.8 105301 3066.3
## + PracticeSport        2     170.2 105162 3067.5
## - ParentEduc           5    6036.7 111369 3087.2
## - TestPrep             1    5630.4 110963 3093.0
## - Gender               1    7881.2 113214 3104.8
## - LunchType            1   10471.1 115804 3118.1
##
## Step:  AIC=3054.45
## ReadingScore ~ LunchType + Gender + TestPrep + ParentEduc + EthnicGroup
##
##                        Df Sum of Sq    RSS    AIC
## + ParentMaritalStatus  3    2469.0  99684 3046.1
## + WklyStudyHours       2     785.9 101367 3053.9
## <none>                             102153 3054.4
## + NrSiblings           1     152.7 102000 3055.6
## + IsFirstChild         1     146.8 102006 3055.6
## + TransportMeans       1      38.2 102115 3056.2
## + PracticeSport        2     263.7 101889 3056.9
## - EthnicGroup          4    3179.3 105333 3064.4
## - ParentEduc           5    5522.0 107675 3075.4
## - TestPrep             1    5395.7 107549 3082.7
## - Gender               1    8375.0 110528 3098.7
## - LunchType            1    9908.7 112062 3106.8
##
## Step:  AIC=3046.09
## ReadingScore ~ LunchType + Gender + TestPrep + ParentEduc + EthnicGroup +
##     ParentMaritalStatus
##
##                        Df Sum of Sq    RSS    AIC
## + WklyStudyHours       2     761.9  98922 3045.6
## <none>                             99684 3046.1
## + IsFirstChild         1     242.8  99441 3046.7
## + NrSiblings           1     117.2  99567 3047.4
## + TransportMeans       1      20.7  99663 3048.0
## + PracticeSport        2     261.2  99423 3048.6
## - ParentMaritalStatus  3    2469.0 102153 3054.4
## - EthnicGroup          4    3640.5 103325 3059.2
## - ParentEduc           5    5358.7 105043 3066.8
## - TestPrep             1    5573.9 105258 3076.0
## - Gender               1    8652.5 108337 3092.9
## - LunchType            1    9804.0 109488 3099.2
##
## Step:  AIC=3045.59
## ReadingScore ~ LunchType + Gender + TestPrep + ParentEduc + EthnicGroup +
##     ParentMaritalStatus + WklyStudyHours
##
##                        Df Sum of Sq    RSS    AIC
## <none>                             98922 3045.6
## - WklyStudyHours       2     761.9  99684 3046.1
## + IsFirstChild         1     213.5  98709 3046.3
## + NrSiblings           1     156.3  98766 3046.7
## + TransportMeans       1      18.3  98904 3047.5
## + PracticeSport        2     265.8  98656 3048.0
## - ParentMaritalStatus  3    2445.0 101367 3053.9
```

```
## - EthnicGroup          4    3541.6 102464 3058.2
## - ParentEduc           5    5525.2 104448 3067.5
## - TestPrep             1    5317.8 104240 3074.3
## - Gender               1    8438.8 107361 3091.6
## - LunchType            1    9508.6 108431 3097.5
```

```
reading_selected_model
```

```
##
## Call:
## lm(formula = ReadingScore ~ LunchType + Gender + TestPrep + ParentEduc +
##     EthnicGroup + ParentMaritalStatus + WklyStudyHours, data = data)
##
## Coefficients:
##                  (Intercept)            LunchTypestandard
##                      67.7283                       8.4814
##                   Gendermale                TestPrepnone
##                      -7.7224                      -6.4210
## ParentEducbachelor's degree        ParentEduchigh school
##                       2.4853                      -5.2903
##   ParentEducmaster's degree       ParentEducsome college
##                       4.0736                      -2.4173
##  ParentEducsome high school           EthnicGroupgroup B
##                      -4.9095                      -1.4318
##          EthnicGroupgroup C           EthnicGroupgroup D
##                      -0.7388                       2.5473
##          EthnicGroupgroup E   ParentMaritalStatusmarried
##                       5.8112                       5.1453
##   ParentMaritalStatussingle   ParentMaritalStatuswidowed
##                       1.7892                       5.4490
##          WklyStudyHours> 10         WklyStudyHours10-May
##                       1.2128                       2.6738
```

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




```
par(mfrow = c(1, 1))
```

```
writing_full_model <- lm(WritingScore ~ Gender + EthnicGroup + ParentEduc + LunchType + TestPrep +
                         ParentMaritalStatus + PracticeSport + IsFirstChild + NrSiblings +
                         TransportMeans + WklyStudyHours, data = data)
writing_null_model = lm(WritingScore ~ 1, data = data)
writing_selected_model = step(writing_null_model,
scope = list(lower = formula(writing_null_model),
upper = formula(writing_full_model)))
```

```
## Start:  AIC=3222.53
## WritingScore ~ 1
##
##                       Df Sum of Sq    RSS    AIC
## + Gender               1   11104.5 130592 3176.6
## + LunchType            1   10442.9 131253 3179.6
## + TestPrep             1    9618.7 132078 3183.3
## + ParentEduc           5   11133.9 130562 3184.5
## + EthnicGroup          4    5484.2 136212 3207.4
## + WklyStudyHours       2    1531.9 140164 3220.1
## + ParentMaritalStatus  3    1929.1 139767 3220.5
## + NrSiblings           1     560.4 141136 3222.2
## <none>                             141696 3222.5
## + IsFirstChild         1      95.4 141601 3224.1
## + TransportMeans       1       0.6 141696 3224.5
## + PracticeSport        2     162.8 141533 3225.9
##
## Step:  AIC=3176.62
## WritingScore ~ Gender
##
##                       Df Sum of Sq    RSS    AIC
## + LunchType            1   11657.0 118935 3123.7
## + TestPrep             1   10482.9 120109 3129.5
## + ParentEduc           5    9612.1 120980 3141.7
```

```
## + EthnicGroup          4    5921.4 124670 3157.4
## + ParentMaritalStatus  3    2060.8 128531 3173.3
## + WklyStudyHours       2    1464.7 129127 3174.0
## <none>                             130592 3176.6
## + NrSiblings           1     401.5 130190 3176.8
## + IsFirstChild         1     127.2 130465 3178.0
## + TransportMeans       1       5.5 130586 3178.6
## + PracticeSport        2     125.5 130466 3180.1
## - Gender               1   11104.5 141696 3222.5
##
## Step:  AIC=3123.74
## WritingScore ~ Gender + LunchType
##
##                       Df Sum of Sq    RSS    AIC
## + TestPrep             1   11216.2 107719 3067.6
## + ParentEduc           5    9909.5 109025 3082.7
## + EthnicGroup          4    5256.1 113679 3105.2
## + ParentMaritalStatus  3    2028.6 116906 3119.6
## + WklyStudyHours       2    1162.9 117772 3122.0
## <none>                             118935 3123.7
## + NrSiblings           1     385.6 118549 3123.8
## + IsFirstChild         1     100.6 118834 3125.2
## + TransportMeans       1       1.3 118933 3125.7
## + PracticeSport        2     179.9 118755 3126.8
## - LunchType            1   11657.0 130592 3176.6
## - Gender               1   12318.6 131253 3179.6
##
## Step:  AIC=3067.59
## WritingScore ~ Gender + LunchType + TestPrep
##
##                       Df Sum of Sq    RSS    AIC
## + ParentEduc           5    9190.5  98528 3025.2
## + EthnicGroup          4    5091.3 102627 3047.2
## + ParentMaritalStatus  3    2161.1 105557 3061.7
## + WklyStudyHours       2     743.3 106975 3067.5
## <none>                             107719 3067.6
## + NrSiblings           1     207.4 107511 3068.5
## + TransportMeans       1      77.9 107641 3069.2
## + IsFirstChild         1      19.3 107699 3069.5
## + PracticeSport        2      92.7 107626 3071.1
## - TestPrep             1   11216.2 118935 3123.7
## - LunchType            1   12390.3 120109 3129.5
## - Gender               1   13298.6 121017 3133.9
##
## Step:  AIC=3025.24
## WritingScore ~ Gender + LunchType + TestPrep + ParentEduc
##
##                       Df Sum of Sq    RSS    AIC
## + EthnicGroup          4    4313.2  94215 3007.0
## + ParentMaritalStatus  3    1944.7  96583 3019.5
## + WklyStudyHours       2     974.0  97554 3023.4
## <none>                             98528 3025.2
## + NrSiblings           1     292.8  98235 3025.5
## + IsFirstChild         1      34.2  98494 3027.0
```

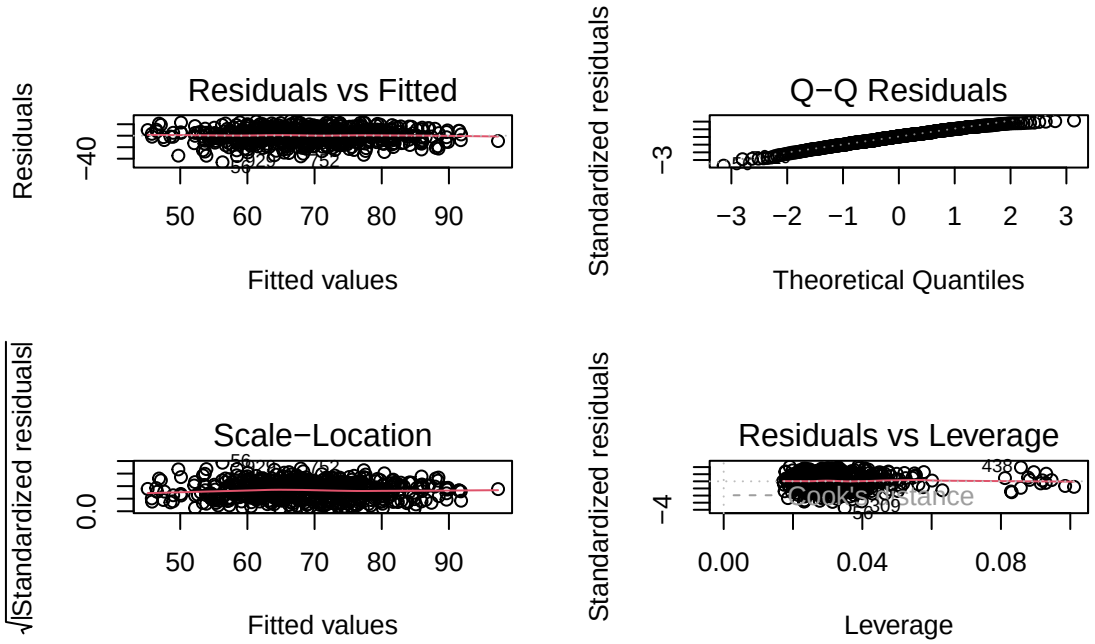
```
## + TransportMeans       1      18.0  98510 3027.1
## + PracticeSport        2     177.4  98351 3028.2
## - ParentEduc           5    9190.5 107719 3067.6
## - TestPrep             1   10497.2 109025 3082.7
## - Gender               1   11662.6 110191 3088.9
## - LunchType            1   12645.1 111173 3094.1
##
## Step:  AIC=3006.97
## WritingScore ~ Gender + LunchType + TestPrep + ParentEduc + EthnicGroup
##
##                        Df Sum of Sq    RSS    AIC
## + ParentMaritalStatus  3    2545.7  91669 2996.9
## + WklyStudyHours       2     864.8  93350 3005.6
## + NrSiblings           1     339.6  93875 3006.8
## <none>                              94215 3007.0
## + IsFirstChild         1      48.3  94167 3008.7
## + TransportMeans       1      32.9  94182 3008.8
## + PracticeSport        2     195.3  94020 3009.8
## - EthnicGroup          4    4313.2  98528 3025.2
## - ParentEduc           5    8412.4 102627 3047.2
## - TestPrep             1   10486.4 104701 3066.9
## - LunchType            1   12063.8 106279 3075.7
## - Gender               1   12303.4 106518 3077.0
##
## Step:  AIC=2996.89
## WritingScore ~ Gender + LunchType + TestPrep + ParentEduc + EthnicGroup +
##     ParentMaritalStatus
##
##                        Df Sum of Sq    RSS    AIC
## + WklyStudyHours       2     827.3  90842 2995.6
## <none>                              91669 2996.9
## + NrSiblings           1     276.3  91393 2997.1
## + IsFirstChild         1     115.3  91554 2998.2
## + TransportMeans       1      16.4  91653 2998.8
## + PracticeSport        2     173.6  91496 2999.8
## - ParentMaritalStatus  3    2545.7  94215 3007.0
## - EthnicGroup          4    4914.3  96583 3019.5
## - ParentEduc           5    8166.8  99836 3037.0
## - TestPrep             1   10682.5 102352 3059.6
## - LunchType            1   11982.5 103652 3067.0
## - Gender               1   12612.0 104281 3070.6
##
## Step:  AIC=2995.57
## WritingScore ~ Gender + LunchType + TestPrep + ParentEduc + EthnicGroup +
##     ParentMaritalStatus + WklyStudyHours
##
##                        Df Sum of Sq    RSS    AIC
## + NrSiblings           1     336.4  90505 2995.4
## <none>                              90842 2995.6
## - WklyStudyHours       2     827.3  91669 2996.9
## + IsFirstChild         1      94.7  90747 2996.9
## + TransportMeans       1      14.0  90828 2997.5
## + PracticeSport        2     153.3  90689 2998.6
## - ParentMaritalStatus  3    2508.2  93350 3005.6
```

```
## - EthnicGroup          4    4803.2  95645 3017.8
## - ParentEduc           5    8364.6  99206 3037.3
## - TestPrep             1   10274.5 101116 3056.5
## - LunchType            1   11644.0 102486 3064.4
## - Gender               1   12344.4 103186 3068.4
##
## Step:  AIC=2995.39
## WritingScore ~ Gender + LunchType + TestPrep + ParentEduc + EthnicGroup +
##     ParentMaritalStatus + WklyStudyHours + NrSiblings
##
##                       Df Sum of Sq    RSS    AIC
## <none>                               90505 2995.4
## - NrSiblings           1     336.4  90842 2995.6
## + IsFirstChild         1     127.2  90378 2996.6
## - WklyStudyHours       2     887.4  91393 2997.1
## + TransportMeans       1      13.5  90492 2997.3
## + PracticeSport        2     152.5  90353 2998.4
## - ParentMaritalStatus  3    2442.3  92948 3005.0
## - EthnicGroup          4    4860.1  95366 3018.1
## - ParentEduc           5    8467.3  98973 3037.9
## - TestPrep             1   10063.8 100569 3055.3
## - LunchType            1   11607.5 102113 3064.2
## - Gender               1   12107.9 102613 3067.1
```

```
writing_selected_model
```

```
##
## Call:
## lm(formula = WritingScore ~ Gender + LunchType + TestPrep + ParentEduc +
##     EthnicGroup + ParentMaritalStatus + WklyStudyHours + NrSiblings,
##     data = data)
##
## Coefficients:
##                  (Intercept)                   Gendermale
##                     66.29079                     -9.26189
##           LunchTypestandard                 TestPrepnone
##                      9.37121                     -8.84425
## ParentEducbachelor's degree        ParentEduchigh school
##                      3.04329                     -6.28698
##   ParentEducmaster's degree       ParentEducsome college
##                      5.51361                     -1.76781
##  ParentEducsome high school           EthnicGroupgroup B
##                     -6.15730                     -1.30296
##          EthnicGroupgroup C           EthnicGroupgroup D
##                      0.09471                      5.08525
##          EthnicGroupgroup E   ParentMaritalStatusmarried
##                      5.98545                      5.21090
##   ParentMaritalStatussingle   ParentMaritalStatuswidowed
##                      2.12419                      6.61763
##          WklyStudyHours> 10         WklyStudyHours10-May
##                      1.22990                      2.87616
##                  NrSiblings
##                      0.51954
```

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




```
par(mfrow = c(1, 1))
```

Summary of the the predit model

```
model1 <- lm(formula = ReadingScore ~ Gender + EthnicGroup + ParentEduc +
               LunchType + TestPrep + ParentMaritalStatus + IsFirstChild, data = data)

model2 <- lm(formula = MathScore ~ Gender + EthnicGroup + ParentEduc +
               LunchType + TestPrep + ParentMaritalStatus + IsFirstChild + WklyStudyHours,
             data = data)

model3 <- lm(formula = WritingScore ~ Gender + EthnicGroup + ParentEduc +
               LunchType + TestPrep + ParentMaritalStatus + WklyStudyHours, data = data)
```

```
summary(data$MathScore)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   56.00   67.00   66.68   78.00  100.00
```

```
data$MathScore_shifted <- data$MathScore + 1
model2_shifted <- lm(MathScore_shifted ~ Gender + EthnicGroup + ParentEduc +
                       LunchType + TestPrep + ParentMaritalStatus + IsFirstChild +
                       WklyStudyHours, data = data)
```

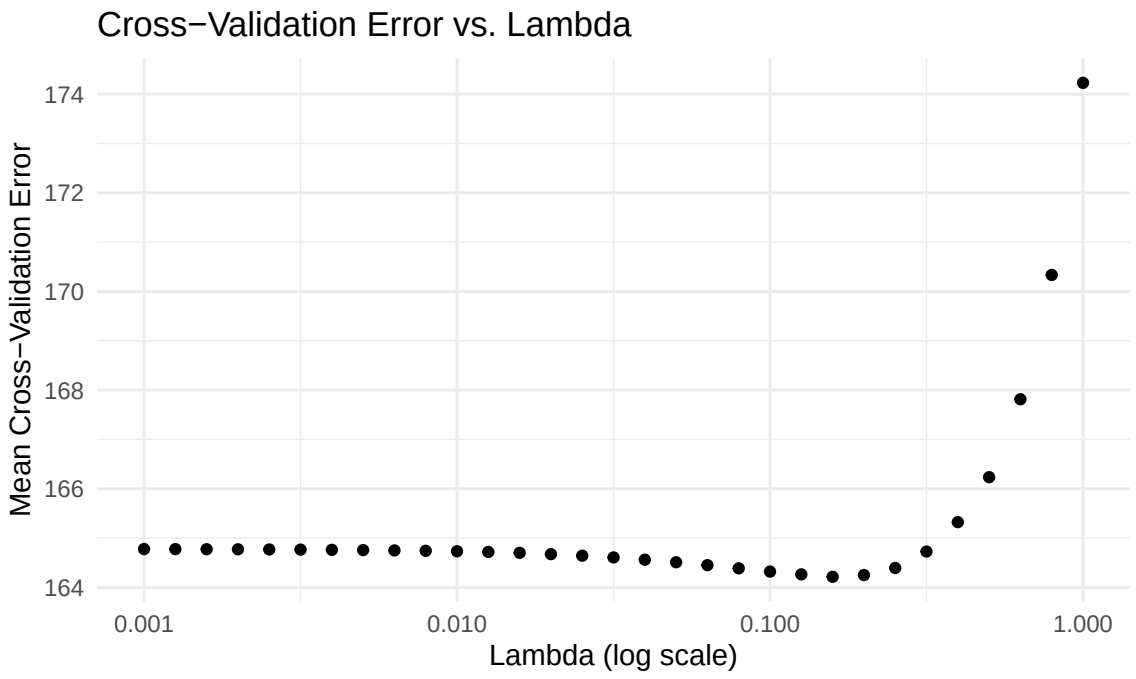
Lasso for Writing

```
library(glmnet)
library(ggplot2)
```

```r
library(tibble)

set.seed(2024)

lambda_seq <- 10^seq(-3, 0, by = 0.1)

x <- model.matrix(WritingScore ~ Gender + EthnicGroup + ParentEduc + LunchType + TestPrep +
                  ParentMaritalStatus + PracticeSport + IsFirstChild + NrSiblings +
                  TransportMeans + WklyStudyHours, data = data)[, -1]

cv_object <- cv.glmnet(x, data$WritingScore, lambda = lambda_seq, nfolds = 5)

cv_object
```

```
##
## Call:  cv.glmnet(x = x, y = data$WritingScore, lambda = lambda_seq,      nfolds = 5)
##
## Measure: Mean-Squared Error
##
##     Lambda Index Measure     SE Nonzero
## min 0.1585     9   164.2  9.783      20
## 1se 0.7943     2   170.3 11.968      14
```

```r
tibble(lambda = cv_object$lambda, mean_cv_error = cv_object$cvm) %>%
  ggplot(aes(x = lambda, y = mean_cv_error)) +
  geom_point() +
  scale_x_log10() +
  labs(title = "Cross-Validation Error vs. Lambda",
       x = "Lambda (log scale)", y = "Mean Cross-Validation Error")
```




```r
min_lambda <- cv_object$lambda.min
min_lambda
```

```
## [1] 0.1584893
```

```
fit_bestcv <- glmnet(x, data$WritingScore, lambda = min_lambda)

coef(fit_bestcv)
```

```
## 31 x 1 sparse Matrix of class "dgCMatrix"
##                                         s0
## (Intercept)                   6.049302e+01
## Gendermale                   -8.931856e+00
## EthnicGroupgroup A            .
## EthnicGroupgroup B           -1.281703e+00
## EthnicGroupgroup C            .
## EthnicGroupgroup D            4.455087e+00
## EthnicGroupgroup E            5.442714e+00
## ParentEducassociate's degree  1.340075e+00
## ParentEducbachelor's degree   4.325147e+00
## ParentEduchigh school        -4.449226e+00
## ParentEducmaster's degree     6.573872e+00
## ParentEducsome college        .
## ParentEducsome high school   -4.211452e+00
## LunchTypestandard             9.056628e+00
## TestPrepcompleted             8.459357e+00
## TestPrepnone                  .
## ParentMaritalStatusdivorced  -1.952763e+00
## ParentMaritalStatusmarried    2.717609e+00
## ParentMaritalStatussingle     .
## ParentMaritalStatuswidowed    3.577020e+00
## PracticeSportnever           -1.093892e+00
## PracticeSportregularly        .
## PracticeSportsometimes        .
## IsFirstChildno               -6.304157e-01
## IsFirstChildyes               1.607411e-13
## NrSiblings                    4.286121e-01
## TransportMeansprivate         .
## TransportMeansschool_bus      .
## WklyStudyHours< 5            -1.081949e+00
## WklyStudyHours> 10            .
## WklyStudyHours10-May          1.374683e+00
```

```
writing_lasso_formula <- WritingScore ~ Gender + EthnicGroup + ParentEduc + LunchType
writing_selected_model <- lm(
  formula = WritingScore ~ Gender + EthnicGroup + ParentEduc + LunchType,
  data = data
)

summary(writing_selected_model)
```

```
##
## Call:
## lm(formula = WritingScore ~ Gender + EthnicGroup + ParentEduc +
##     LunchType, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

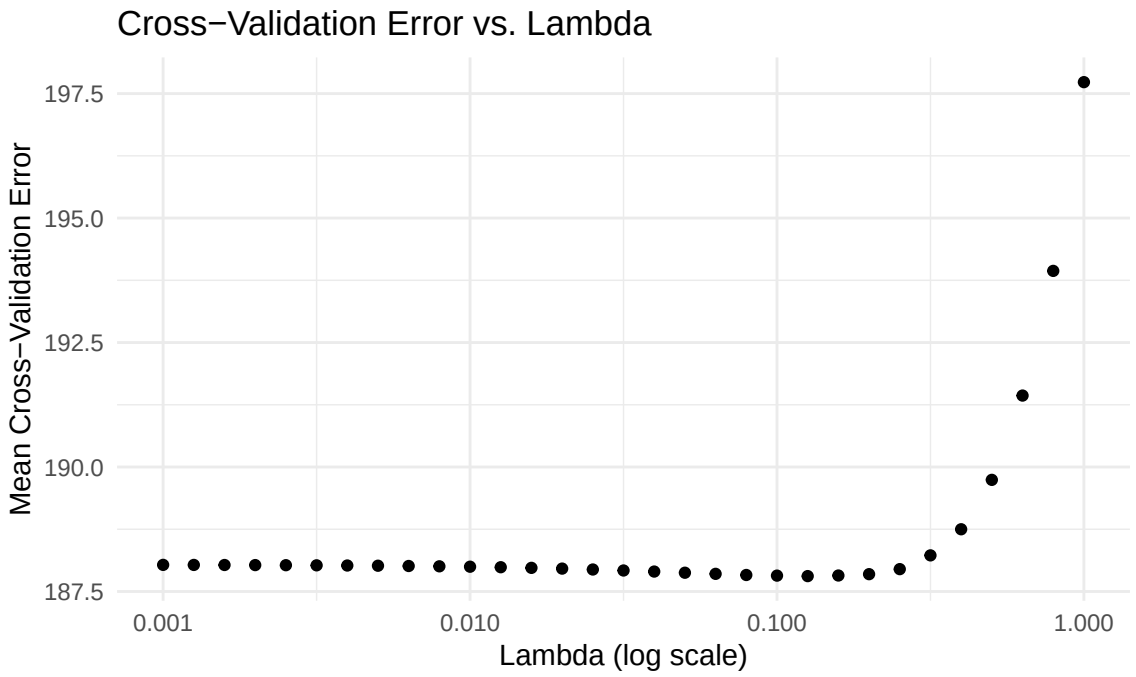
```
## -51.975  -8.968   1.041   9.708  29.880
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    67.454      2.435  27.701  < 2e-16 ***
## Gendermale                     -8.894      1.130  -7.870 1.78e-14 ***
## EthnicGroupgroup B             -1.292      2.276  -0.568 0.570392
## EthnicGroupgroup C              0.675      2.186   0.309 0.757653
## EthnicGroupgroup D              4.239      2.205   1.923 0.055000 .
## EthnicGroupgroup E              6.561      2.425   2.706 0.007014 **
## ParentEducbachelor's degree     3.505      2.005   1.748 0.080953 .
## ParentEduchigh school          -7.072      1.721  -4.109 4.56e-05 ***
## ParentEducmaster's degree       4.250      2.479   1.715 0.086925 .
## ParentEducsome college         -2.505      1.737  -1.442 0.149919
## ParentEducsome high school     -6.154      1.759  -3.498 0.000506 ***
## LunchTypestandard               9.232      1.171   7.887 1.57e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.49 on 575 degrees of freedom
## Multiple R-squared:  0.2611, Adjusted R-squared:  0.247
## F-statistic: 18.47 on 11 and 575 DF,  p-value: < 2.2e-16
```

Lasso for Math

```
set.seed(2024)

lambda_seq <- 10^seq(-3, 0, by = 0.1)

x <- model.matrix(MathScore ~ Gender + EthnicGroup + ParentEduc + LunchType + TestPrep +
                  ParentMaritalStatus + PracticeSport + IsFirstChild + NrSiblings +
                  TransportMeans + WklyStudyHours, data = data)[, -1]

cv_object <- cv.glmnet(x, data$MathScore, lambda = lambda_seq, nfolds = 5)

cv_object
```

```
##
## Call:  cv.glmnet(x = x, y = data$MathScore, lambda = lambda_seq, nfolds = 5)
##
## Measure: Mean-Squared Error
##
##     Lambda Index Measure    SE Nonzero
## min 0.1259    10   187.8 13.07      22
## 1se 1.0000     1   197.7 15.78      11
```

```
tibble(lambda = cv_object$lambda, mean_cv_error = cv_object$cvm) %>%
  ggplot(aes(x = lambda, y = mean_cv_error)) +
  geom_point() +
  scale_x_log10() +
  labs(title = "Cross-Validation Error vs. Lambda",
       x = "Lambda (log scale)", y = "Mean Cross-Validation Error")
```




```
min_lambda <- cv_object$lambda.min
min_lambda
```

```
## [1] 0.1258925
```

```
fit_bestcv <- glmnet(x, data$MathScore, lambda = min_lambda)

coef(fit_bestcv)
```

```
## 31 x 1 sparse Matrix of class "dgCMatrix"
##                                       s0
## (Intercept)                 5.096655e+01
## Gendermale                  4.837613e+00
## EthnicGroupgroup A          .
## EthnicGroupgroup B         -1.029591e-02
## EthnicGroupgroup C          .
## EthnicGroupgroup D          3.347301e+00
## EthnicGroupgroup E          1.090092e+01
## ParentEducassociate's degree  1.299984e+00
## ParentEducbachelor's degree   2.940548e+00
## ParentEduchigh school        -3.354740e+00
## ParentEducmaster's degree     2.898704e+00
## ParentEducsome college        .
## ParentEducsome high school   -3.106260e+00
## LunchTypestandard             1.210664e+01
## TestPrepcompleted             4.605550e+00
## TestPrepnone                  .
## ParentMaritalStatusdivorced  -1.990602e+00
## ParentMaritalStatusmarried    3.011776e+00
## ParentMaritalStatussingle     .
## ParentMaritalStatuswidowed    4.696536e+00
## PracticeSportnever           -1.100817e+00
## PracticeSportregularly        .
```

```
## PracticeSportsometimes        .
## IsFirstChildno               -7.748471e-01
## IsFirstChildyes               9.411899e-14
## NrSiblings                    6.374608e-01
## TransportMeansprivate         1.092178e-01
## TransportMeansschool_bus     -3.349586e-14
## WklyStudyHours< 5            -2.970340e+00
## WklyStudyHours> 10            .
## WklyStudyHours10-May          2.795436e-01
```

```
math_lasso_formula <- MathScore ~ Gender + EthnicGroup + ParentEduc + LunchType + TestPrep +
                      ParentMaritalStatus + PracticeSport + IsFirstChild + NrSiblings +
                      TransportMeans + WklyStudyHours

math_selected_model <- lm(
  formula = math_lasso_formula,
  data = data
)

summary(math_selected_model)
```

```
##
## Call:
## lm(formula = math_lasso_formula, data = data)
##
## 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)                   49.0064     3.7750  12.982  < 2e-16 ***
## Gendermale                     5.0855     1.1386   4.467 9.61e-06 ***
## EthnicGroupgroup B            -0.1788     2.3136  -0.077  0.93841
## EthnicGroupgroup C            -0.2089     2.2149  -0.094  0.92489
## EthnicGroupgroup D             3.6247     2.2286   1.626  0.10441
## EthnicGroupgroup E            11.1752     2.4434   4.574 5.90e-06 ***
## ParentEducbachelor's degree    1.7594     2.0219   0.870  0.38458
## ParentEduchigh school         -5.2293     1.7463  -2.994  0.00287 **
## ParentEducmaster's degree      1.9038     2.5136   0.757  0.44912
## ParentEducsome college        -1.7126     1.7556  -0.976  0.32973
## ParentEducsome high school    -4.9058     1.7728  -2.767  0.00584 **
## LunchTypestandard             12.3539     1.1771  10.495  < 2e-16 ***
## TestPrepnone                  -4.7717     1.2007  -3.974 7.99e-05 ***
## ParentMaritalStatusmarried     5.4805     1.6170   3.389  0.00075 ***
## ParentMaritalStatussingle      2.1682     1.8454   1.175  0.24053
## ParentMaritalStatuswidowed     7.7944     3.8119   2.045  0.04134 *
## PracticeSportregularly         1.6701     1.9046   0.877  0.38092
## PracticeSportsometimes         1.5255     1.8439   0.827  0.40838
## IsFirstChildyes                1.1303     1.2125   0.932  0.35162
## NrSiblings                     0.7403     0.3844   1.926  0.05461 .
## TransportMeansschool_bus      -0.4319     1.1629  -0.371  0.71050
## WklyStudyHours> 10             3.0384     1.7540   1.732  0.08378 .
## WklyStudyHours10-May           3.5394     1.3429   2.636  0.00863 **
## ---
```

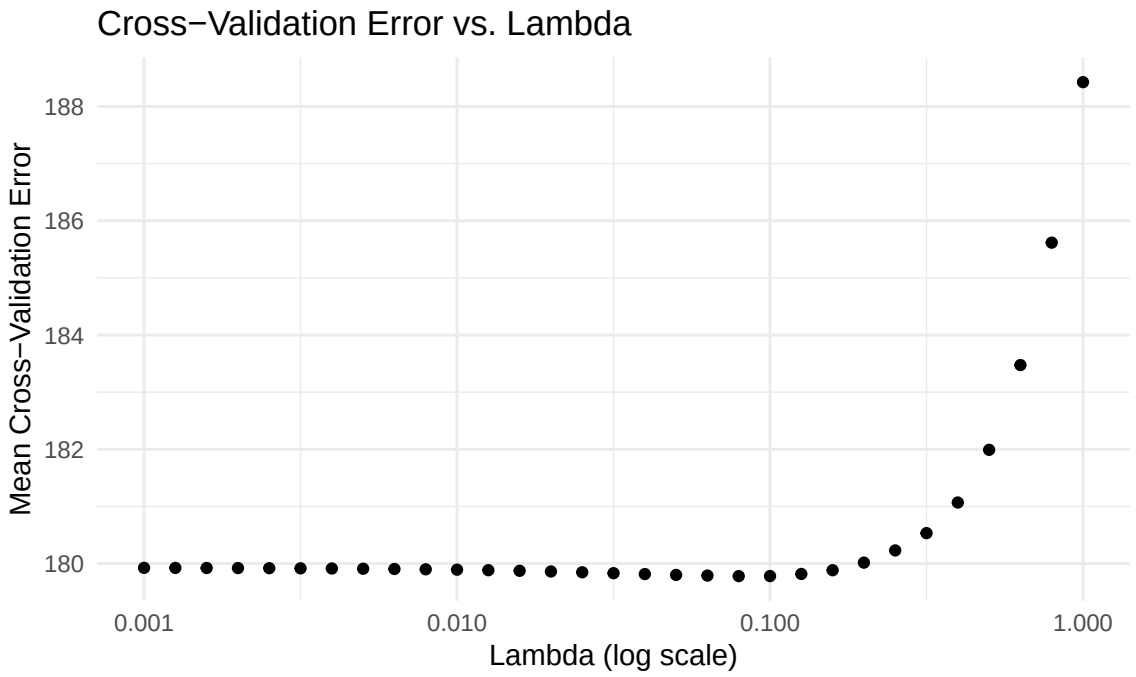
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.52 on 564 degrees of freedom
## Multiple R-squared:  0.3221, Adjusted R-squared:  0.2956
## F-statistic: 12.18 on 22 and 564 DF,  p-value: < 2.2e-16
```

Lasso for Reading

```
set.seed(2024)

lambda_seq <- 10^seq(-3, 0, by = 0.1)

x <- model.matrix(ReadingScore ~ Gender + EthnicGroup + ParentEduc + LunchType + TestPrep +
                  ParentMaritalStatus + PracticeSport + IsFirstChild + NrSiblings +
                  TransportMeans + WklyStudyHours, data = data)[, -1]

cv_object <- cv.glmnet(x, data$ReadingScore, lambda = lambda_seq, nfolds = 5)

cv_object
```

```
##
## Call:  cv.glmnet(x = x, y = data$ReadingScore, lambda = lambda_seq,      nfolds = 5)
##
## Measure: Mean-Squared Error
##
##     Lambda Index Measure    SE Nonzero
## min 0.0794    12   179.8 10.25      23
## 1se 1.0000     1   188.4 11.87      13
```

```
tibble(lambda = cv_object$lambda, mean_cv_error = cv_object$cvm) %>%
  ggplot(aes(x = lambda, y = mean_cv_error)) +
  geom_point() +
  scale_x_log10() +
  labs(title = "Cross-Validation Error vs. Lambda",
       x = "Lambda (log scale)", y = "Mean Cross-Validation Error")
```




```
min_lambda <- cv_object$lambda.min
min_lambda
```

```
## [1] 0.07943282
```

```
fit_bestcv <- glmnet(x, data$ReadingScore, lambda = min_lambda)

coef(fit_bestcv)
```

```
## 31 x 1 sparse Matrix of class "dgCMatrix"
##                                        s0
## (Intercept)                  6.178868e+01
## Gendermale                  -7.490495e+00
## EthnicGroupgroup A           .
## EthnicGroupgroup B          -1.155661e+00
## EthnicGroupgroup C          -5.599657e-01
## EthnicGroupgroup D           2.509873e+00
## EthnicGroupgroup E           5.858709e+00
## ParentEducassociate's degree 2.161557e+00
## ParentEducbachelor's degree  4.667263e+00
## ParentEduchigh school       -2.917463e+00
## ParentEducmaster's degree    5.944661e+00
## ParentEducsome college       .
## ParentEducsome high school  -2.327561e+00
## LunchTypestandard            8.275192e+00
## TestPrepcompleted            6.124144e+00
## TestPrepnone                 .
## ParentMaritalStatusdivorced -1.802243e+00
## ParentMaritalStatusmarried   3.127710e+00
## ParentMaritalStatussingle    .
## ParentMaritalStatuswidowed   3.090673e+00
## PracticeSportnever           .
## PracticeSportregularly      -6.347620e-01
```

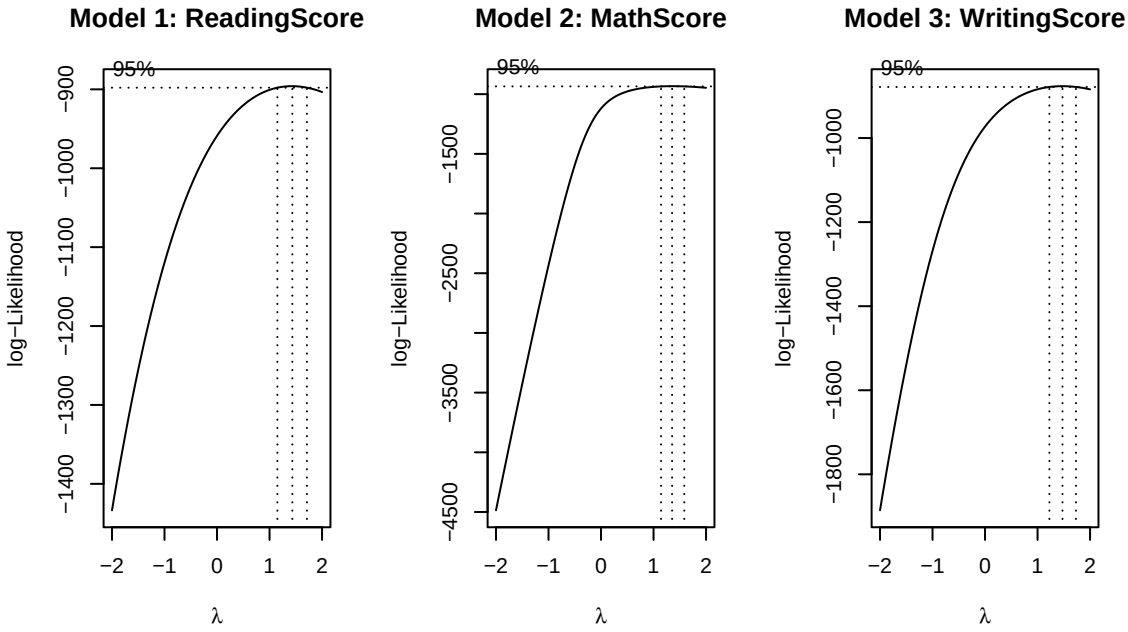
```
## PracticeSportsometimes        5.726530e-01
## IsFirstChildno               -1.106325e+00
## IsFirstChildyes               3.793391e-14
## NrSiblings                    3.298797e-01
## TransportMeansprivate        -1.513493e-01
## TransportMeansschool_bus      .
## WklyStudyHours< 5            -1.064847e+00
## WklyStudyHours> 10            .
## WklyStudyHours10-May          1.450228e+00
```

```r
reading_lasso_formula <- ReadingScore ~ Gender + EthnicGroup + ParentEduc + LunchType + TestPrep +
                         ParentMaritalStatus + PracticeSport + IsFirstChild + NrSiblings +
                         TransportMeans + WklyStudyHours

reading_selected_model <- lm(
  formula = reading_lasso_formula,
  data = data
)

summary(reading_selected_model)
```

```
##
## Call:
## lm(formula = reading_lasso_formula, data = data)
##
## 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)                   65.5976     3.6847  17.803  < 2e-16 ***
## Gendermale                    -7.6725     1.1114  -6.904 1.37e-11 ***
## EthnicGroupgroup B            -1.4287     2.2582  -0.633  0.52722
## EthnicGroupgroup C            -0.8558     2.1619  -0.396  0.69236
## EthnicGroupgroup D             2.5663     2.1753   1.180  0.23860
## EthnicGroupgroup E             5.9165     2.3850   2.481  0.01340 *
## ParentEducbachelor's degree    2.5549     1.9735   1.295  0.19600
## ParentEduchigh school         -5.3732     1.7046  -3.152  0.00171 **
## ParentEducmaster's degree      3.9202     2.4535   1.598  0.11065
## ParentEducsome college        -2.3866     1.7136  -1.393  0.16424
## ParentEducsome high school    -4.7948     1.7305  -2.771  0.00578 **
## LunchTypestandard              8.4374     1.1489   7.344 7.31e-13 ***
## TestPrepnone                  -6.2822     1.1720  -5.360 1.21e-07 ***
## ParentMaritalStatusmarried     5.2439     1.5783   3.322  0.00095 ***
## ParentMaritalStatussingle      1.9235     1.8013   1.068  0.28605
## ParentMaritalStatuswidowed     5.5863     3.7208   1.501  0.13381
## PracticeSportregularly        -0.6843     1.8590  -0.368  0.71292
## PracticeSportsometimes         0.6757     1.7998   0.375  0.70749
## IsFirstChildyes                1.3046     1.1835   1.102  0.27078
## NrSiblings                     0.3882     0.3752   1.035  0.30131
## TransportMeansschool_bus       0.2841     1.1351   0.250  0.80247
## WklyStudyHours> 10             1.0970     1.7121   0.641  0.52197
## WklyStudyHours10-May           2.6835     1.3108   2.047  0.04110 *
## ---
```

```
## 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
```

```r
# Set up a 1-row, 3-column layout
par(mfrow = c(1, 3))

# Box-Cox analysis for model1
boxcox(model1, lambda = seq(-2, 2, by = 0.1))  # Range of lambda values
title("Model 1: ReadingScore")

# Box-Cox analysis for model2
boxcox(model2_shifted, lambda = seq(-2, 2, by = 0.1))  # Range of lambda values
title("Model 2: MathScore")

# Box-Cox analysis for model3
boxcox(model3, lambda = seq(-2, 2, by = 0.1))  # Range of lambda values
title("Model 3: WritingScore")
```

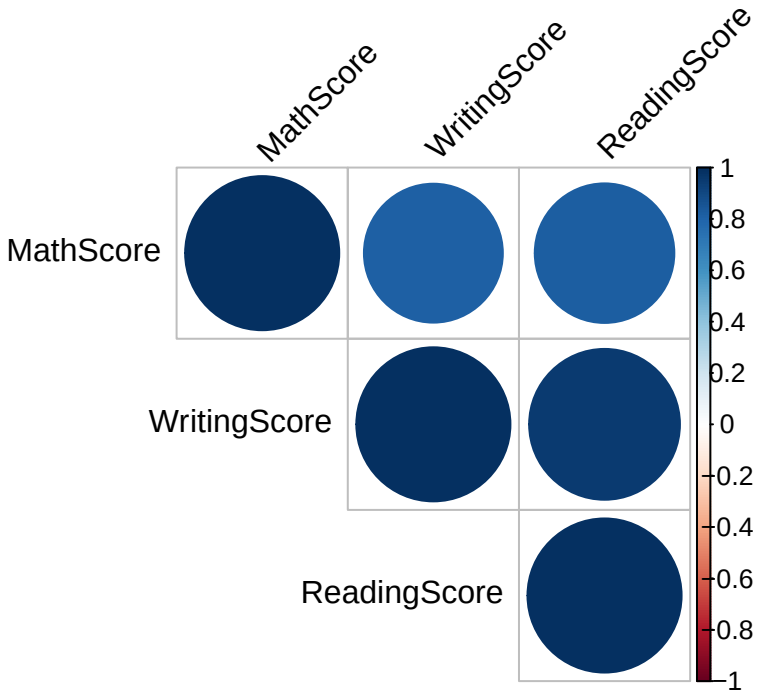



```r
# Reset graphical parameters (optional)
par(mfrow = c(1, 1))
```

```r
correlation_matrix <- cor(data[, c("MathScore", "WritingScore", "ReadingScore")], use = "complete.obs")

print(correlation_matrix)
```

```
##              MathScore WritingScore ReadingScore
## MathScore    1.0000000    0.8124605    0.8201729
## WritingScore 0.8124605    1.0000000    0.9577284
## ReadingScore 0.8201729    0.9577284    1.0000000
```

```r
library(corrplot)
corrplot(correlation_matrix, method = "circle", type = "upper", tl.col = "black", tl.srt = 45)
```



Cross-Validation

```
# CV for Math Score
set.seed(2024)
train_control = trainControl(method = "cv", number = 10)
cv_model = train(MathScore ~ Gender + EthnicGroup + ParentEduc +
                LunchType + TestPrep + ParentMaritalStatus + IsFirstChild + WklyStudyHours,
              data = data,
method = "lm",
trControl = train_control)
print(cv_model)
```

```
## Linear Regression
##
## 587 samples
##   8 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 528, 528, 529, 528, 528, 529, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   13.77219  0.2749112  11.19126
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
# CV for Reading Score
set.seed(2024)
train_control = trainControl(method = "cv", number = 10)
cv_model = train(ReadingScore ~ Gender + EthnicGroup + ParentEduc +
                LunchType + TestPrep + ParentMaritalStatus + IsFirstChild,
                data = data,
```

```
method = "lm",
trControl = train_control)
print(cv_model)
```

```
## Linear Regression
##
## 587 samples
##   7 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 529, 529, 528, 528, 528, 529, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   13.30402  0.2405314  10.79295
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
# CV for Reading Score
set.seed(2024)
train_control = trainControl(method = "cv", number = 10)
cv_model = train(WritingScore ~ Gender + EthnicGroup + ParentEduc +
                 LunchType + TestPrep + ParentMaritalStatus + WklyStudyHours,
                 data = data,
method = "lm",
trControl = train_control)
print(cv_model)
```

```
## Linear Regression
##
## 587 samples
##   7 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 529, 529, 527, 529, 527, 528, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   12.73832  0.3339129  10.26367
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

‘‘‘