bm-project

YISU

2024-12-19

```
data = read.csv("Project_1_data.csv")
```

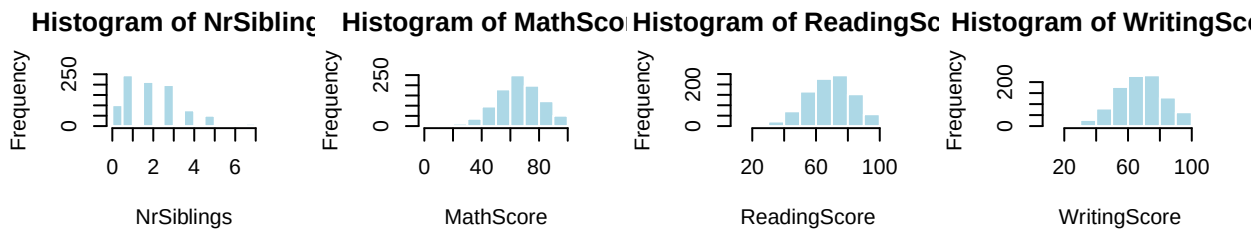
```
# Load necessary library
library(ggplot2)

# Load your dataset
data <- read.csv("Project_1_data.csv")

# Select only numeric variables
numeric_data <- data[sapply(data, is.numeric)]

# Set up the plotting area to accommodate all 11 histograms
par(mfrow = c(3, 4)) # 3 rows and 4 columns (will leave one empty space for 11 variables)

# Create histograms for each numeric variable
for (variable in names(numeric_data)) {
  hist(numeric_data[[variable]],
       main = paste("Histogram of", variable),
       col = "lightblue",
       xlab = variable,
       border = "white")
}
par(mfrow = c(1, 1))
```




```
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)
```

```r
# Handle missing values by removing rows with NA
data <- na.omit(data)

# Define the full model for MathScore
math_full_model <- lm(MathScore ~ Gender + EthnicGroup + ParentEduc + LunchType + TestPrep +
                        ParentMaritalStatus + PracticeSport + IsFirstChild + NrSiblings +
                        TransportMeans + WklyStudyHours, data = data)

# Perform stepwise model selection
math_selected_model <- stepAIC(math_full_model, direction = "both", trace = FALSE)
math_selected_model
```

```
##
## Call:
## lm(formula = MathScore ~ Gender + EthnicGroup + ParentEduc +
##     LunchType + TestPrep + ParentMaritalStatus + IsFirstChild +
##     WklyStudyHours, data = data)
##
## Coefficients:
##                   (Intercept)                     Gendermale
##                       51.2910                         5.0885
##           EthnicGroupgroup A            EthnicGroupgroup B
##                       -1.2758                         0.1118
##           EthnicGroupgroup C            EthnicGroupgroup D
##                       -0.2774                         3.5351
##           EthnicGroupgroup E  ParentEducassociate's degree
##                        8.5754                         4.7280
##  ParentEducbachelor's degree         ParentEduchigh school
##                        6.0476                        -0.8361
##    ParentEducmaster's degree        ParentEducsome college
##                        6.4044                         3.7222
##   ParentEducsome high school             LunchTypestandard
##                       -0.5013                        11.0841
##            TestPrepcompleted                    TestPrepnone
##                        4.4017                        -0.9571
##  ParentMaritalStatusdivorced    ParentMaritalStatusmarried
##                       -0.5039                         3.2310
##    ParentMaritalStatussingle    ParentMaritalStatuswidowed
##                        0.1923                         4.2876
##               IsFirstChildno               IsFirstChildyes
##                       -1.1316                         0.8604
##            WklyStudyHours< 5            WklyStudyHours> 10
##                       -4.7980                        -1.0807
##         WklyStudyHours10-May
##                       -1.4518
```

```r
# Display the summary of the selected model
cat("\nSelected Model for MathScore:\n")
```

```
##
## Selected Model for MathScore:
```

```r
summary(math_selected_model)
```

```
##
```

```
## Call:
## lm(formula = MathScore ~ Gender + EthnicGroup + ParentEduc +
##     LunchType + TestPrep + ParentMaritalStatus + IsFirstChild +
##     WklyStudyHours, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.527  -8.997   0.498   9.626  30.285
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    51.2910     5.1078  10.042  < 2e-16 ***
## Gendermale                      5.0885     0.8992   5.659 2.06e-08 ***
## EthnicGroupgroup A             -1.2758     2.4139  -0.529 0.597264
## EthnicGroupgroup B              0.1118     2.0927   0.053 0.957422
## EthnicGroupgroup C             -0.2774     1.9986  -0.139 0.889628
## EthnicGroupgroup D              3.5351     2.0268   1.744 0.081489 .
## EthnicGroupgroup E              8.5754     2.2003   3.897 0.000105 ***
## ParentEducassociate's degree    4.7280     2.1442   2.205 0.027708 *
## ParentEducbachelor's degree     6.0476     2.3515   2.572 0.010280 *
## ParentEduchigh school          -0.8361     2.1666  -0.386 0.699661
## ParentEducmaster's degree       6.4044     2.6695   2.399 0.016643 *
## ParentEducsome college          3.7222     2.1417   1.738 0.082561 .
## ParentEducsome high school     -0.5013     2.1831  -0.230 0.818444
## LunchTypestandard              11.0841     0.9381  11.815  < 2e-16 ***
## TestPrepcompleted               4.4017     2.0171   2.182 0.029357 *
## TestPrepnone                   -0.9571     1.9433  -0.493 0.622484
## ParentMaritalStatusdivorced    -0.5039     2.3146  -0.218 0.827721
## ParentMaritalStatusmarried      3.2310     2.0944   1.543 0.123277
## ParentMaritalStatussingle       0.1923     2.2226   0.087 0.931062
## ParentMaritalStatuswidowed      4.2876     3.4046   1.259 0.208240
## IsFirstChildno                 -1.1316     2.6597  -0.425 0.670602
## IsFirstChildyes                 0.8604     2.6059   0.330 0.741341
## WklyStudyHours< 5              -4.7980     2.3751  -2.020 0.043676 *
## WklyStudyHours> 10             -1.0807     2.4903  -0.434 0.664422
## WklyStudyHours10-May           -1.4518     2.3020  -0.631 0.528423
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.36 on 877 degrees of freedom
## Multiple R-squared:  0.2819, Adjusted R-squared:  0.2623
## F-statistic: 14.35 on 24 and 877 DF,  p-value: < 2.2e-16
```

```r
# Make predictions using the selected model
math_predictions <- predict(math_selected_model, newdata = data)

# Combine actual and predicted values for MathScore
math_results <- data.frame(
  Actual_MathScore = data$MathScore,
  Predicted_MathScore = math_predictions
)

# Save results to a CSV file
write.csv(math_results, "math_score_predictions.csv", row.names = FALSE)
```

```r
reading_full_model <- lm(ReadingScore ~ Gender + EthnicGroup + ParentEduc + LunchType + TestPrep +
                         ParentMaritalStatus + PracticeSport + IsFirstChild + NrSiblings +
                         TransportMeans + WklyStudyHours, data = data)

# Perform stepwise model selection
reading_selected_model <- stepAIC(reading_full_model, direction = "both", trace = FALSE)
reading_selected_model
```

```
##
## Call:
## lm(formula = ReadingScore ~ Gender + EthnicGroup + ParentEduc +
##     LunchType + TestPrep + ParentMaritalStatus + IsFirstChild,
##     data = data)
##
## Coefficients:
##                   (Intercept)                     Gendermale
##                      58.63929                       -7.15237
##           EthnicGroupgroup A            EthnicGroupgroup B
##                       2.11437                        1.61464
##           EthnicGroupgroup C            EthnicGroupgroup D
##                       2.10744                        4.82071
##           EthnicGroupgroup E  ParentEducassociate's degree
##                       6.46862                        5.14291
##  ParentEducbachelor's degree         ParentEduchigh school
##                       6.13542                       -0.87311
##    ParentEducmaster's degree        ParentEducsome college
##                       9.02888                        3.27435
##   ParentEducsome high school             LunchTypestandard
##                       0.05839                        7.56883
##            TestPrepcompleted                   TestPrepnone
##                       6.15804                       -0.75034
##  ParentMaritalStatusdivorced    ParentMaritalStatusmarried
##                      -1.08867                        2.67064
##    ParentMaritalStatussingle    ParentMaritalStatuswidowed
##                      -0.53098                        3.47406
##               IsFirstChildno               IsFirstChildyes
##                      -1.44383                        0.71327
```

```r
# Display the summary of the selected model
cat("\nSelected Model for Reading:\n")
```

```
##
## Selected Model for Reading:
```

```r
summary(reading_selected_model)
```

```
##
## Call:
## lm(formula = ReadingScore ~ Gender + EthnicGroup + ParentEduc +
##     LunchType + TestPrep + ParentMaritalStatus + IsFirstChild,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.237  -8.650   1.161   9.385  29.517
```

```
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  58.63929    4.50285  13.023  < 2e-16 ***
## Gendermale                   -7.15237    0.88045  -8.124 1.53e-15 ***
## EthnicGroupgroup A            2.11437    2.35196   0.899 0.368908
## EthnicGroupgroup B            1.61464    2.04598   0.789 0.430219
## EthnicGroupgroup C            2.10744    1.94808   1.082 0.279636
## EthnicGroupgroup D            4.82071    1.97981   2.435 0.015092 *
## EthnicGroupgroup E            6.46862    2.15021   3.008 0.002701 **
## ParentEducassociate's degree  5.14291    2.09638   2.453 0.014351 *
## ParentEducbachelor's degree   6.13542    2.29642   2.672 0.007685 **
## ParentEduchigh school        -0.87311    2.11963  -0.412 0.680499
## ParentEducmaster's degree     9.02888    2.60907   3.461 0.000565 ***
## ParentEducsome college        3.27435    2.09170   1.565 0.117848
## ParentEducsome high school    0.05839    2.13514   0.027 0.978191
## LunchTypestandard             7.56883    0.91763   8.248 5.83e-16 ***
## TestPrepcompleted             6.15804    1.96245   3.138 0.001758 **
## TestPrepnone                 -0.75034    1.89705  -0.396 0.692548
## ParentMaritalStatusdivorced  -1.08867    2.26625  -0.480 0.631072
## ParentMaritalStatusmarried    2.67064    2.05082   1.302 0.193177
## ParentMaritalStatussingle    -0.53098    2.17544  -0.244 0.807227
## ParentMaritalStatuswidowed    3.47406    3.33026   1.043 0.297151
## IsFirstChildno               -1.44383    2.60496  -0.554 0.579541
## IsFirstChildyes               0.71327    2.55239   0.279 0.779964
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.09 on 880 degrees of freedom
## Multiple R-squared:  0.2389, Adjusted R-squared:  0.2207
## F-statistic: 13.15 on 21 and 880 DF,  p-value: < 2.2e-16
```

```r
# Make predictions using the selected model
reading_predictions <- predict(reading_selected_model, newdata = data)

# Combine actual and predicted values for MathScore
reading_results <- data.frame(
  Actual_ReadingScore = data$ReadingScore,
  Predicted_ReadingScore = reading_predictions
)

# Save results to a CSV file
write.csv(reading_results, "reading_score_predictions.csv", row.names = FALSE)
```

```r
writing_full_model <- lm(WritingScore ~ Gender + EthnicGroup + ParentEduc + LunchType + TestPrep +
                         ParentMaritalStatus + PracticeSport + IsFirstChild + NrSiblings +
                         TransportMeans + WklyStudyHours, data = data)

# Perform stepwise model selection
writing_selected_model <- stepAIC(writing_full_model, direction = "both", trace = FALSE)
writing_selected_model
```

```
##
## Call:
## lm(formula = WritingScore ~ Gender + EthnicGroup + ParentEduc +
```

```
##     LunchType + TestPrep + ParentMaritalStatus + WklyStudyHours,
##     data = data)
##
## Coefficients:
##                   (Intercept)                     Gendermale
##                      59.36546                       -9.23816
##           EthnicGroupgroup A             EthnicGroupgroup B
##                       0.63448                        0.66800
##           EthnicGroupgroup C             EthnicGroupgroup D
##                       1.27259                        5.76199
##           EthnicGroupgroup E  ParentEducassociate's degree
##                       5.03801                        4.48934
##  ParentEducbachelor's degree         ParentEduchigh school
##                       6.69286                       -2.30978
##    ParentEducmaster's degree        ParentEducsome college
##                       9.76964                        3.19346
##   ParentEducsome high school             LunchTypestandard
##                      -1.80483                        8.31180
##            TestPrepcompleted                   TestPrepnone
##                       7.72567                       -1.82437
##  ParentMaritalStatusdivorced    ParentMaritalStatusmarried
##                      -0.34162                        3.20838
##    ParentMaritalStatussingle    ParentMaritalStatuswidowed
##                      -0.05569                        3.15168
##            WklyStudyHours< 5            WklyStudyHours> 10
##                      -2.17429                       -0.16869
##         WklyStudyHours10-May
##                       0.34306
```

```
# Display the summary of the selected model
cat("\nSelected Model for WritingScore:\n")
```

```
##
## Selected Model for WritingScore:
```

```
summary(writing_selected_model)
```

```
##
## Call:
## lm(formula = WritingScore ~ Gender + EthnicGroup + ParentEduc +
##     LunchType + TestPrep + ParentMaritalStatus + WklyStudyHours,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.784  -7.859   0.772   8.712  32.001
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 59.36546    4.20505  14.118  < 2e-16 ***
## Gendermale                  -9.23816    0.85274 -10.834  < 2e-16 ***
## EthnicGroupgroup A           0.63448    2.28724   0.277 0.781539
## EthnicGroupgroup B           0.66800    1.98226   0.337 0.736205
## EthnicGroupgroup C           1.27259    1.89387   0.672 0.501790
## EthnicGroupgroup D           5.76199    1.92143   2.999 0.002787 **
```

```
## EthnicGroupgroup E            5.03801    2.08690   2.414 0.015977 *
## ParentEducassociate's degree  4.48934    2.03180   2.210 0.027394 *
## ParentEducbachelor's degree   6.69286    2.22839   3.003 0.002745 **
## ParentEduchigh school        -2.30978    2.05029  -1.127 0.260235
## ParentEducmaster's degree     9.76964    2.52630   3.867 0.000118 ***
## ParentEducsome college        3.19346    2.02703   1.575 0.115516
## ParentEducsome high school   -1.80483    2.06990  -0.872 0.383478
## LunchTypestandard             8.31180    0.88936   9.346  < 2e-16 ***
## TestPrepcompleted             7.72567    1.91429   4.036 5.92e-05 ***
## TestPrepnone                 -1.82437    1.84365  -0.990 0.322669
## ParentMaritalStatusdivorced  -0.34162    2.19391  -0.156 0.876295
## ParentMaritalStatusmarried    3.20838    1.98472   1.617 0.106336
## ParentMaritalStatussingle    -0.05569    2.10631  -0.026 0.978913
## ParentMaritalStatuswidowed    3.15168    3.23112   0.975 0.329623
## WklyStudyHours< 5            -2.17429    2.25368  -0.965 0.334924
## WklyStudyHours> 10           -0.16869    2.36248  -0.071 0.943092
## WklyStudyHours10-May          0.34306    2.18355   0.157 0.875192
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.68 on 879 degrees of freedom
## Multiple R-squared:  0.3347, Adjusted R-squared:  0.3181
## F-statistic:  20.1 on 22 and 879 DF,  p-value: < 2.2e-16
```

```r
# Make predictions using the selected model
writing_predictions <- predict(writing_selected_model, newdata = data)

# Combine actual and predicted values for MathScore
writing_results <- data.frame(
  Actual_WritingScore = data$WritingScore,
  Predicted_WritingScore = writing_predictions
)

# Save results to a CSV file
write.csv(writing_results, "writing_score_predictions.csv", row.names = FALSE)
```