Build and Evaluate Multiple Linear Regression

Dr. Nishat Mohammad

03/02/2024

Load data:

```
# load student.mat file
path= "./student_performance/student/student-mat.csv"
student_mat <- read.table(path, sep=";",header=TRUE)

any(is.na(student_mat))
```

```
## [1] FALSE
```

```
str(student_mat)
```

```
## 'data.frame':    395 obs. of  33 variables:
##  $ school    : chr  "GP" "GP" "GP" "GP" ...
##  $ sex       : chr  "F" "F" "F" "F" ...
##  $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
##  $ address   : chr  "U" "U" "U" "U" ...
##  $ famsize   : chr  "GT3" "GT3" "LE3" "GT3" ...
##  $ Pstatus   : chr  "A" "T" "T" "T" ...
##  $ Medu      : int  4 1 1 4 3 4 2 4 3 3 ...
##  $ Fedu      : int  4 1 1 2 3 3 2 4 2 4 ...
##  $ Mjob      : chr  "at_home" "at_home" "at_home" "health" ...
##  $ Fjob      : chr  "teacher" "other" "other" "services" ...
##  $ reason    : chr  "course" "course" "other" "home" ...
##  $ guardian  : chr  "mother" "father" "mother" "mother" ...
##  $ traveltime: int  2 1 1 1 1 1 1 2 1 1 ...
##  $ studytime : int  2 2 2 3 2 2 2 2 2 2 ...
##  $ failures  : int  0 0 3 0 0 0 0 0 0 0 ...
##  $ schoolsup : chr  "yes" "no" "yes" "no" ...
##  $ famsup    : chr  "no" "yes" "no" "yes" ...
##  $ paid      : chr  "no" "no" "yes" "yes" ...
##  $ activities: chr  "no" "no" "no" "yes" ...
##  $ nursery   : chr  "yes" "no" "yes" "yes" ...
##  $ higher    : chr  "yes" "yes" "yes" "yes" ...
##  $ internet  : chr  "no" "yes" "yes" "yes" ...
##  $ romantic  : chr  "no" "no" "no" "yes" ...
##  $ famrel    : int  4 5 4 3 4 5 4 4 4 5 ...
##  $ freetime  : int  3 3 3 2 3 4 4 1 2 5 ...
##  $ goout     : int  4 3 2 2 2 2 4 4 2 1 ...
```

```
##  $ Dalc       : int  1 1 2 1 1 1 1 1 1 1 ...
##  $ Walc       : int  1 1 3 1 2 2 1 1 1 1 ...
##  $ health     : int  3 3 3 5 5 5 3 1 1 5 ...
##  $ absences   : int  6 4 10 2 4 10 0 6 0 0 ...
##  $ G1         : int  5 5 7 15 6 15 12 6 16 14 ...
##  $ G2         : int  6 5 8 14 10 15 12 5 18 15 ...
##  $ G3         : int  6 6 10 15 10 15 11 6 19 15 ...
```

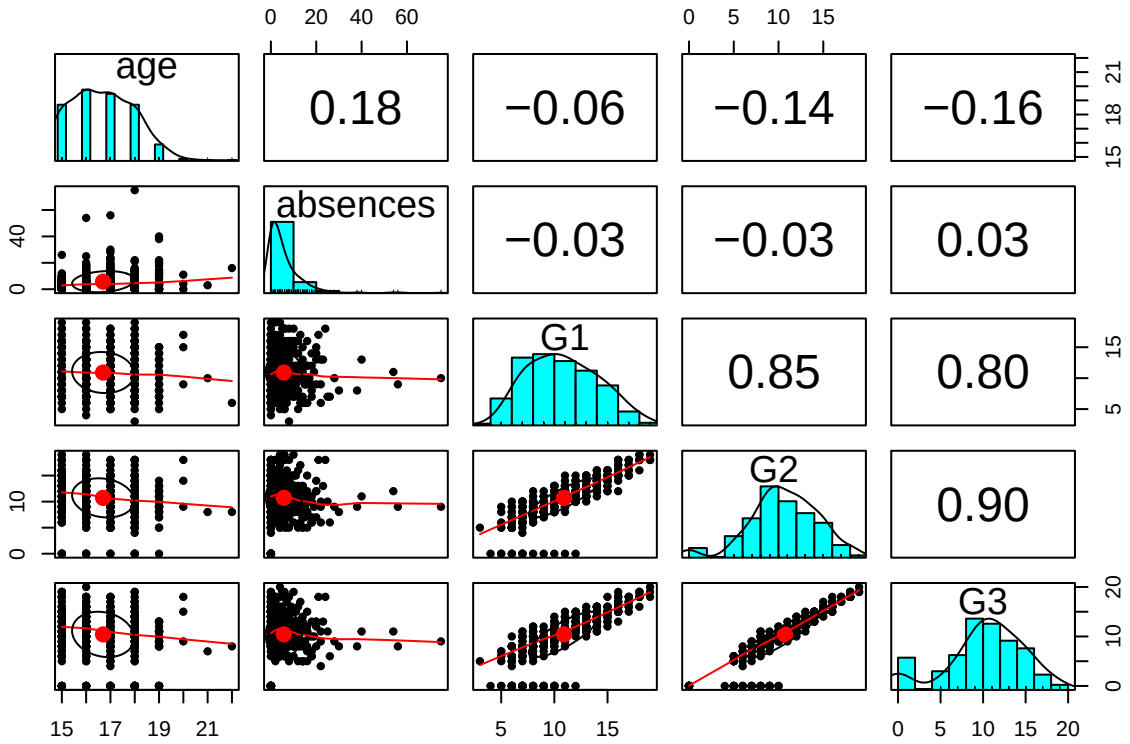
This is the student math data, with 395 rows of observations on 33 columns of variables all of type either integer or character.

1. Visualize Correlation Between Age, Absences and Grades:

Create scatter plots and pairwise correlations between age, absences, G1, and G2 and final grade (G3) using the pairs.panels() function in R.

```
library(psych)

pairs.panels(student_mat[c("age", "absences", "G1", "G2", "G3")])
```



2. Mutiple Regression Model:

Build a multiple regression model predicting final math grade (G3) using as many features as you like but you must use at least four. Include at least one categorical variables and be sure to encode it properly using

a method of your choice. Select the features that you believe are useful – you do not have to include all features.

```
# Use the absences, activities and the first two grades as features to predict final grade
summary(student_mat[c("absences", "activities", "G1", "G2")])
```

```
##     absences       activities             G1              G2
##  Min.   : 0.000   Length:395         Min.   : 3.00   Min.   : 0.00
##  1st Qu.: 0.000   Class :character   1st Qu.: 8.00   1st Qu.: 9.00
##  Median : 4.000   Mode  :character   Median :11.00   Median :11.00
##  Mean   : 5.709                      Mean   :10.91   Mean   :10.71
##  3rd Qu.: 8.000                      3rd Qu.:13.00   3rd Qu.:13.00
##  Max.   :75.000                      Max.   :19.00   Max.   :19.00
```

```
# Factor encode the activities
student_mat$activities <- ifelse(student_mat$activities=="yes", 1, 0)

# Create the model
final_grade_model <- lm(G3 ~ activities+absences+G1+G2, data=student_mat)

# Look at the model
summary(final_grade_model)
```

```
##
## Call:
## lm(formula = G3 ~ activities + absences + G1 + G2, data = student_mat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.2233 -0.3684  0.2795  0.9771  3.7877
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.95013    0.35018  -5.569 4.78e-08 ***
## activities  -0.27957    0.19303  -1.448  0.14834
## absences     0.03615    0.01206   2.997  0.00290 **
## G1           0.15666    0.05555   2.820  0.00504 **
## G2           0.98864    0.04900  20.176  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.915 on 390 degrees of freedom
## Multiple R-squared:  0.8271, Adjusted R-squared:  0.8253
## F-statistic: 466.5 on 4 and 390 DF,  p-value: < 2.2e-16
```

The summary of the model shows: 1. The Residual differences between the actual values and predicted one based on the some descriptive statistics measures like min max, median and 1st and 3rd quantiles.

2. The Coefficients for each feature have been displayed with the corresponding standard error and t-values with the p-values for the t tests. according to this model the G2 is very significant in contributing to the outcome of the final grade while G1 and absences are significant but to a lesser extent as depicted by the number of * assigned after the p-values which code for the significance level. Very obvious is that "activities" does not have significance according to our model.

3. The Multiple R-squared and its adjusted counterpart are both very close, implying that the features I selected can account for only around 82% of the proportion of variance in the final grades.

4. The F-statisic has a very low p-value of 2.2e-16, indicating that this model is significant.

3. Stepwise Backward Elimintaion:

Using the model from (2), use stepwise backward elimination to remove all non-significant variables and then state the final model as an equation. State the backward elimination measure you applied (p-value, AIC, Adjusted R2). This tutorial shows how to use various feature elimination techniques.

```
# Create a new model without activities
final_grade_model2 <- lm(G3 ~ absences + G1 + G2, data=student_mat)
summary(final_grade_model2)
```

```
##
## Call:
## lm(formula = G3 ~ absences + G1 + G2, data = student_mat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3616 -0.3559  0.3163  0.9642  3.9242
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.06747    0.34116  -6.060  3.2e-09 ***
## absences     0.03635    0.01208   3.010  0.00278 **
## G1           0.15452    0.05561   2.779  0.00572 **
## G2           0.98838    0.04907  20.142  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.917 on 391 degrees of freedom
## Multiple R-squared:  0.8262, Adjusted R-squared:  0.8249
## F-statistic: 619.5 on 3 and 391 DF,  p-value: < 2.2e-16
```

Taking off the " "activities based on the fact that the p-value of the test showed that is is not significant. the resulting model has changed slightly only on the Residuals. R- squared values remain similar to the previous model as well. the overall pvalue of the F-statistics remains the same and thus this model is significant.

Let us look out for interaction between extracurricular activites and absences with anothe model below:

```
final_grade_model3 <- lm(G3 ~ G1 + G2+ absences*activities, data=student_mat)
summary(final_grade_model3)
```

```
##
## Call:
## lm(formula = G3 ~ G1 + G2 + absences * activities, data = student_mat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1035 -0.3615  0.3132  0.9733  3.7014
##
```

```
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1.88129    0.35255  -5.336 1.62e-07 ***
## G1                   0.16089    0.05553   2.897  0.00397 **
## G2                   0.98659    0.04894  20.159  < 2e-16 ***
## absences             0.02023    0.01599   1.266  0.20641
## activities          -0.49004    0.23768  -2.062  0.03989 *
## absences:activities  0.03681    0.02433   1.513  0.13110
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.911 on 389 degrees of freedom
## Multiple R-squared:  0.8281, Adjusted R-squared:  0.8259
## F-statistic: 374.9 on 5 and 389 DF,  p-value: < 2.2e-16
```

this 3rd model shows that absences and absences:activities do not have significance, while activities now seem to have sligth significance. All other. parameters remain the same.

Let us take of all features except the G1 and G2 with the code chunk below:

```
final_grade_model4 <- lm(G3 ~ G1 + G2, data=student_mat)
summary(final_grade_model4)
```

```
##
## Call:
## lm(formula = G3 ~ G1 + G2, data = student_mat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5713 -0.3888  0.2885  0.9725  3.7089
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.83001    0.33531  -5.458 8.57e-08 ***
## G1           0.15327    0.05618   2.728  0.00665 **
## G2           0.98687    0.04957  19.909  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.937 on 392 degrees of freedom
## Multiple R-squared:  0.8222, Adjusted R-squared:  0.8213
## F-statistic: 906.1 on 2 and 392 DF,  p-value: < 2.2e-16
```

The fourth model also shows similar stats, similar R-squared values and pvalues. It would be best to compare the models with ANOVA and look for the model with lowest p value on ANOVA test.

```
anov_2_1 <- anova(final_grade_model2, final_grade_model)

anov_3_1 <- anova(final_grade_model3, final_grade_model)
anov_3_1
```

```
## Analysis of Variance Table
##
```

```
## Model 1: G3 ~ G1 + G2 + absences * activities
## Model 2: G3 ~ activities + absences + G1 + G2
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    389 1421.3
## 2    390 1429.7 -1   -8.3638 2.2891 0.1311
```

```
anov_4_1 <- anova(final_grade_model4, final_grade_model)
anov_4_1
```

```
## Analysis of Variance Table
##
## Model 1: G3 ~ G1 + G2
## Model 2: G3 ~ activities + absences + G1 + G2
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    392 1470.7
## 2    390 1429.7  2    41.005 5.5928 0.00403 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The anova results foe comparing all the models against the first have been doen and the conclusion is that the first model is the best model. based on the final anova test between the 4th and 1st model with p value of 0.00403 having ** meaning this is significant anova comparison.

3. Get the 95% CI:

Calculate the 95% confidence interval for a prediction – you may choose any data you wish for some new student.

```
absences<-c(2)
activities <- c(0)
G1 <- c(10)
G2 <- c(15)
nw_candid <- data.frame(absences, activities, G1, G2)

final_grade_nw_candidate <- predict(final_grade_model, nw_candid)
final_grade_nw_candidate
```

```
##        1
## 14.51838
```

```
stats4model<-summary(final_grade_model)

# Get the lower limit of confidence interval
lower_CI <- final_grade_nw_candidate - 1.96 * stats4model$sigma

# Get the upper limit of Confidence interval
upper_CI <- final_grade_nw_candidate + 1.96 * stats4model$sigma
```

The final grade for th demo new candidate is 14.5183842. The 95% confidence interval for this is 10.7656782, 18.2710903.

4. RMSE for the Model:

What is the RMSE for this model – use the entire data set for both training and validation. You may find the residuals() function useful. Alternatively, you can inspect the model object, e.g., if your model is in the variable m, then the residuals (errors) are in m$residuals and your predicted values (fitted values) are in m$fitted.values.

```
# Predict using all the data
final_grade_prediction <- predict(final_grade_model, student_mat[c("absences", "activities", "G1", "G2"
)

# Get RMSE
RMSE_stud_mat <- sqrt(mean((final_grade_prediction - student_mat$G3)^2))
RMSE_stud_mat
```

```
## [1] 1.902489
```

```
# Another way to get the RMSE
Same_RMSE <- sqrt(mean((stats4model$residuals)^2))
```

The RMSE is for this model is 1.9024893. I calculated this form scratch and also using the residuals form the model summary.

5. Reconsider Missing Values, Outliers and Normality of the Data:

We did not consider outliers, manage missing values, nor check for normality of the included features. This is important, so return to the data set and check for missing values and use an appropriate strategy to deal with them, check that all features are reasonably normally distributed – and, if not, apply a transform (e.g., log-transform), and, finally, consider outliers as statistical learning algorithms are sensitive to outliers. Next, rebuild your regression model using appropriate features.

```
# Look at the data again
summary(student_mat)
```

```
##     school              sex                 age         address
##  Length:395         Length:395         Min.   :15.0   Length:395
##  Class :character   Class :character   1st Qu.:16.0   Class :character
##  Mode  :character   Mode  :character   Median :17.0   Mode  :character
##                                        Mean   :16.7
##                                        3rd Qu.:18.0
##                                        Max.   :22.0
##    famsize            Pstatus              Medu            Fedu
##  Length:395         Length:395         Min.   :0.000   Min.   :0.000
##  Class :character   Class :character   1st Qu.:2.000   1st Qu.:2.000
##  Mode  :character   Mode  :character   Median :3.000   Median :2.000
##                                        Mean   :2.749   Mean   :2.522
##                                        3rd Qu.:4.000   3rd Qu.:3.000
##                                        Max.   :4.000   Max.   :4.000
##      Mjob               Fjob              reason            guardian
##  Length:395         Length:395         Length:395         Length:395
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
```

```
##
##
##    traveltime      studytime        failures       schoolsup
##  Min.   :1.000   Min.   :1.000   Min.   :0.0000   Length:395
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   Class :character
##  Median :1.000   Median :2.000   Median :0.0000   Mode  :character
##  Mean   :1.448   Mean   :2.035   Mean   :0.3342
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000
##  Max.   :4.000   Max.   :4.000   Max.   :3.0000
##     famsup              paid             activities       nursery
##  Length:395         Length:395         Min.   :0.0000   Length:395
##  Class :character   Class :character   1st Qu.:0.0000   Class :character
##  Mode  :character   Mode  :character   Median :1.0000   Mode  :character
##                                        Mean   :0.5089
##                                        3rd Qu.:1.0000
##                                        Max.   :1.0000
##     higher            internet           romantic            famrel
##  Length:395         Length:395         Length:395         Min.   :1.000
##  Class :character   Class :character   Class :character   1st Qu.:4.000
##  Mode  :character   Mode  :character   Mode  :character   Median :4.000
##                                                           Mean   :3.944
##                                                           3rd Qu.:5.000
##                                                           Max.   :5.000
##     freetime         goout            Dalc            Walc
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000
##  Median :3.000   Median :3.000   Median :1.000   Median :2.000
##  Mean   :3.235   Mean   :3.109   Mean   :1.481   Mean   :2.291
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
##      health         absences            G1              G2
##  Min.   :1.000   Min.   : 0.000   Min.   : 3.00   Min.   : 0.00
##  1st Qu.:3.000   1st Qu.: 0.000   1st Qu.: 8.00   1st Qu.: 9.00
##  Median :4.000   Median : 4.000   Median :11.00   Median :11.00
##  Mean   :3.554   Mean   : 5.709   Mean   :10.91   Mean   :10.71
##  3rd Qu.:5.000   3rd Qu.: 8.000   3rd Qu.:13.00   3rd Qu.:13.00
##  Max.   :5.000   Max.   :75.000   Max.   :19.00   Max.   :19.00
##        G3
##  Min.   : 0.00
##  1st Qu.: 8.00
##  Median :11.00
##  Mean   :10.42
##  3rd Qu.:14.00
##  Max.   :20.00
```

```
str(student_mat)
```

```
## 'data.frame':    395 obs. of  33 variables:
##  $ school    : chr  "GP" "GP" "GP" "GP" ...
##  $ sex       : chr  "F" "F" "F" "F" ...
##  $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
##  $ address   : chr  "U" "U" "U" "U" ...
##  $ famsize   : chr  "GT3" "GT3" "LE3" "GT3" ...
##  $ Pstatus   : chr  "A" "T" "T" "T" ...
```

```
##  $ Medu      : int  4 1 1 4 3 4 2 4 3 3 ...
##  $ Fedu      : int  4 1 1 2 3 3 2 4 2 4 ...
##  $ Mjob      : chr  "at_home" "at_home" "at_home" "health" ...
##  $ Fjob      : chr  "teacher" "other" "other" "services" ...
##  $ reason    : chr  "course" "course" "other" "home" ...
##  $ guardian  : chr  "mother" "father" "mother" "mother" ...
##  $ traveltime: int  2 1 1 1 1 1 1 2 1 1 ...
##  $ studytime : int  2 2 2 3 2 2 2 2 2 2 ...
##  $ failures  : int  0 0 3 0 0 0 0 0 0 0 ...
##  $ schoolsup : chr  "yes" "no" "yes" "no" ...
##  $ famsup    : chr  "no" "yes" "no" "yes" ...
##  $ paid      : chr  "no" "no" "yes" "yes" ...
##  $ activities: num  0 0 0 1 0 1 0 0 0 1 ...
##  $ nursery   : chr  "yes" "no" "yes" "yes" ...
##  $ higher    : chr  "yes" "yes" "yes" "yes" ...
##  $ internet  : chr  "no" "yes" "yes" "yes" ...
##  $ romantic  : chr  "no" "no" "no" "yes" ...
##  $ famrel    : int  4 5 4 3 4 5 4 4 4 5 ...
##  $ freetime  : int  3 3 3 2 3 4 4 1 2 5 ...
##  $ goout     : int  4 3 2 2 2 2 4 4 2 1 ...
##  $ Dalc      : int  1 1 2 1 1 1 1 1 1 1 ...
##  $ Walc      : int  1 1 3 1 2 2 1 1 1 1 ...
##  $ health    : int  3 3 3 5 5 5 3 1 1 5 ...
##  $ absences  : int  6 4 10 2 4 10 0 6 0 0 ...
##  $ G1        : int  5 5 7 15 6 15 12 6 16 14 ...
##  $ G2        : int  6 5 8 14 10 15 12 5 18 15 ...
##  $ G3        : int  6 6 10 15 10 15 11 6 19 15 ...
```

```
# Show the absence of missing values again
any(is.na(student_mat))
```

```
## [1] FALSE
```

```
# Check normality for integer type columns
int_columns <- student_mat[sapply(student_mat, is.integer)]
head(int_columns)
```

```
##   age Medu Fedu traveltime studytime failures famrel freetime goout Dalc Walc
## 1  18    4    4          2         2        0      4        3     4    1    1
## 2  17    1    1          1         2        0      5        3     3    1    1
## 3  15    1    1          1         2        3      4        3     2    2    3
## 4  15    4    2          1         3        0      3        2     2    1    1
## 5  16    3    3          1         2        0      4        3     2    1    2
## 6  16    4    3          1         2        0      5        4     2    1    2
##   health absences G1 G2 G3
## 1      3        6  5  6  6
## 2      3        4  5  5  6
## 3      3       10  7  8 10
## 4      5        2 15 14 15
## 5      5        4  6 10 10
## 6      5       10 15 15 15
```

First look at the data again, there are no missing values, also there are several variables that are ordinal categorical. let us choose only the integer data that is continuous to check for normality.

5.1. Shapiro-Wilk Test, Visualization, Skewness:

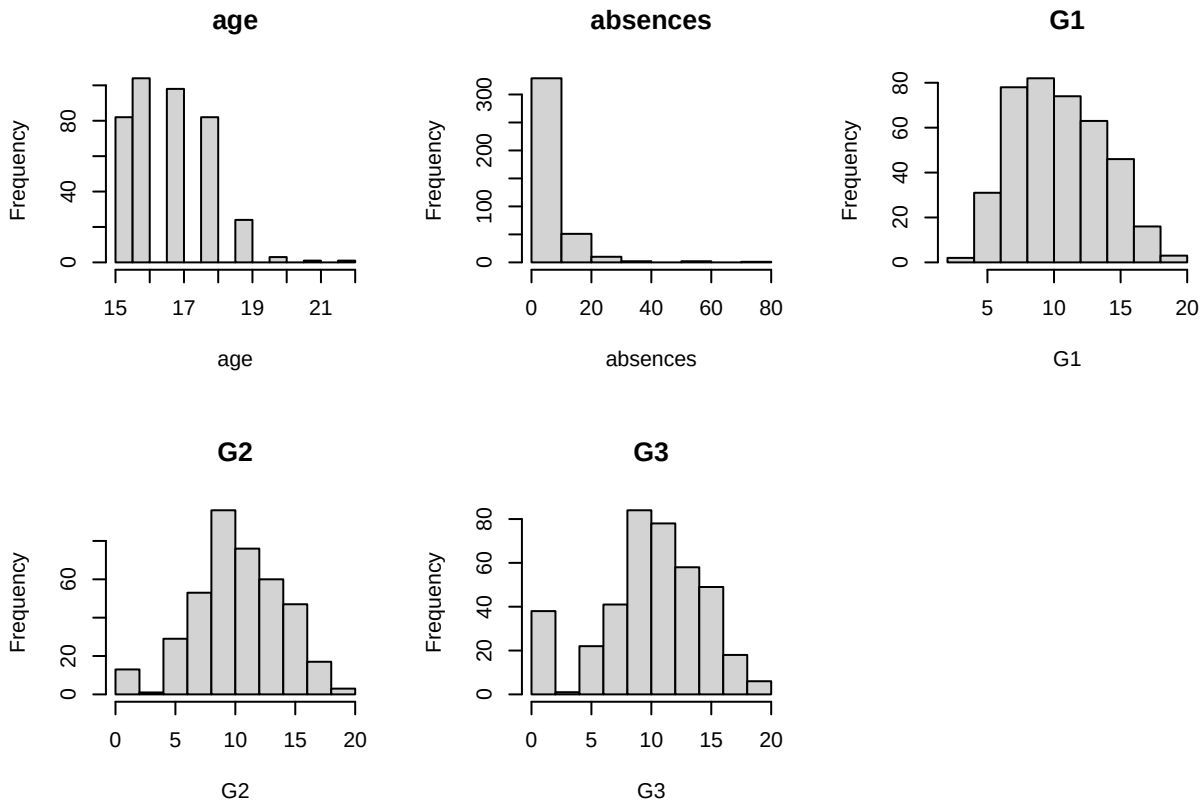
```
cont_int_columns <- student_mat[, c("age", "absences", "G1", "G2", "G3")]

# Shapiro test
shapiro_tests <- sapply(cont_int_columns, shapiro.test)
shapiro_tests
```

```
##           age                           absences
## statistic 0.9105932                     0.6668336
## p.value   1.588761e-14                  4.140363e-27
## method    "Shapiro-Wilk normality test" "Shapiro-Wilk normality test"
## data.name "X[[i]]"                      "X[[i]]"
##           G1                            G2
## statistic 0.9749134                     0.9691415
## p.value   2.454158e-06                  2.08396e-07
## method    "Shapiro-Wilk normality test" "Shapiro-Wilk normality test"
## data.name "X[[i]]"                      "X[[i]]"
##           G3
## statistic 0.9287298
## p.value   8.835916e-13
## method    "Shapiro-Wilk normality test"
## data.name "X[[i]]"
```

```
# Visualize with histograms
par(mfrow=c(2, 3))
for (col in names(cont_int_columns)) {
  hist(cont_int_columns[[col]], main = col, xlab = col)
}
# Check for skewness
library(e1071)
skewnez <- sapply(cont_int_columns, skewness)
skewnez
```

```
##        age   absences         G1         G2         G3
##  0.4627348  3.6437406  0.2387889 -0.4283726 -0.7271171
```

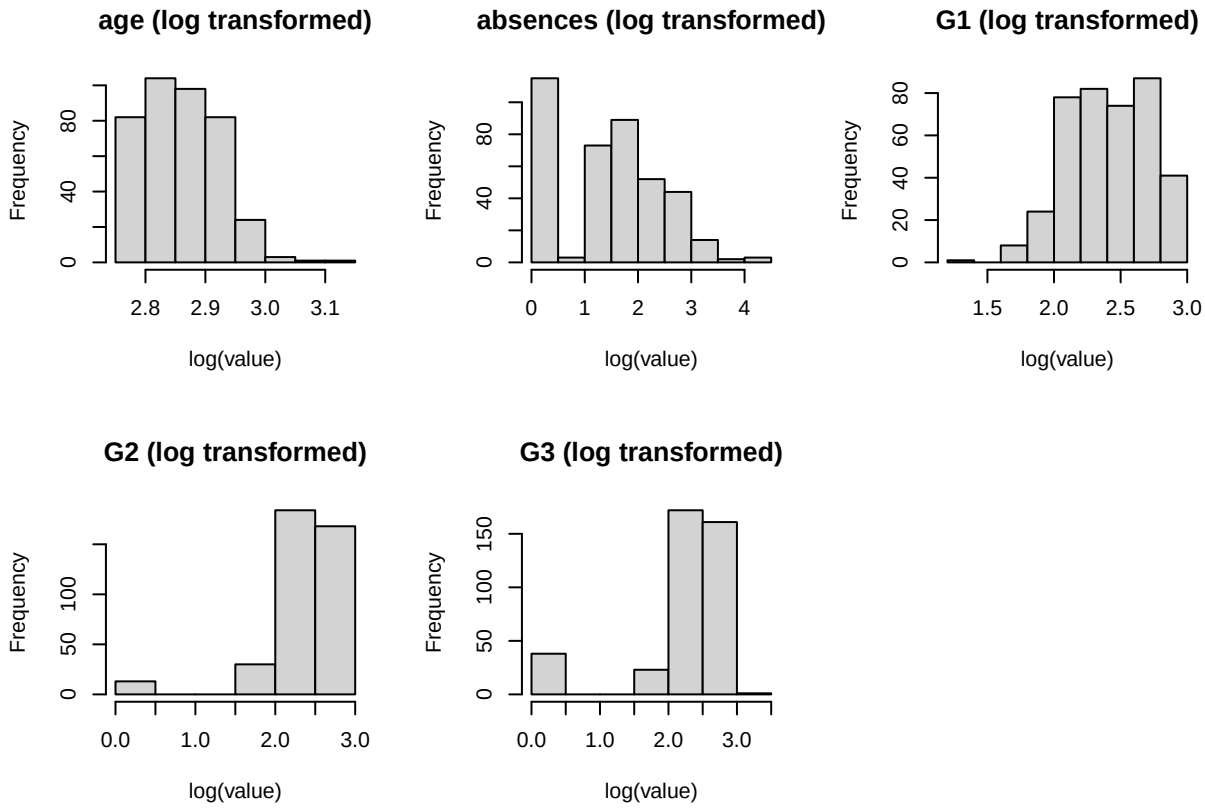


The Histograms show that none of the variables are perfectly normally distributed, this can also be seen with the results of Shapiro Tests. I checked for the skewness as well which shped the G2 having a slight left skew and G3 moderately skewed to the left, G1 and age are slightly right skewed, while absences are heavily right skewed (shows that majority of students were punctual).

5.2. Log Transformation:

```
log_transform_columns <- lapply(cont_int_columns, function(x) log(x + 1))

# Plot histograms for each log-transformed variable
par(mfrow=c(2, 3))  # Set up the layout for multiple plots
for (i in 1:length(log_transform_columns)) {
  hist(log_transform_columns[[i]],
       main = paste(names(log_transform_columns)[i],
                    "(log transformed)"),
       xlab = "log(value)")
}
```



The varibales have now been log transformed and visualized again. we can go ahead to filter out outliers.

5.3. Handle Outliers:

```
# Quantile method in a function
str(log_transform_columns)
```

```
## List of 5
##  $ age     : num [1:395] 2.94 2.89 2.77 2.77 2.83 ...
##  $ absences: num [1:395] 1.95 1.61 2.4 1.1 1.61 ...
##  $ G1      : num [1:395] 1.79 1.79 2.08 2.77 1.95 ...
##  $ G2      : num [1:395] 1.95 1.79 2.2 2.71 2.4 ...
##  $ G3      : num [1:395] 1.95 1.95 2.4 2.77 2.4 ...
```

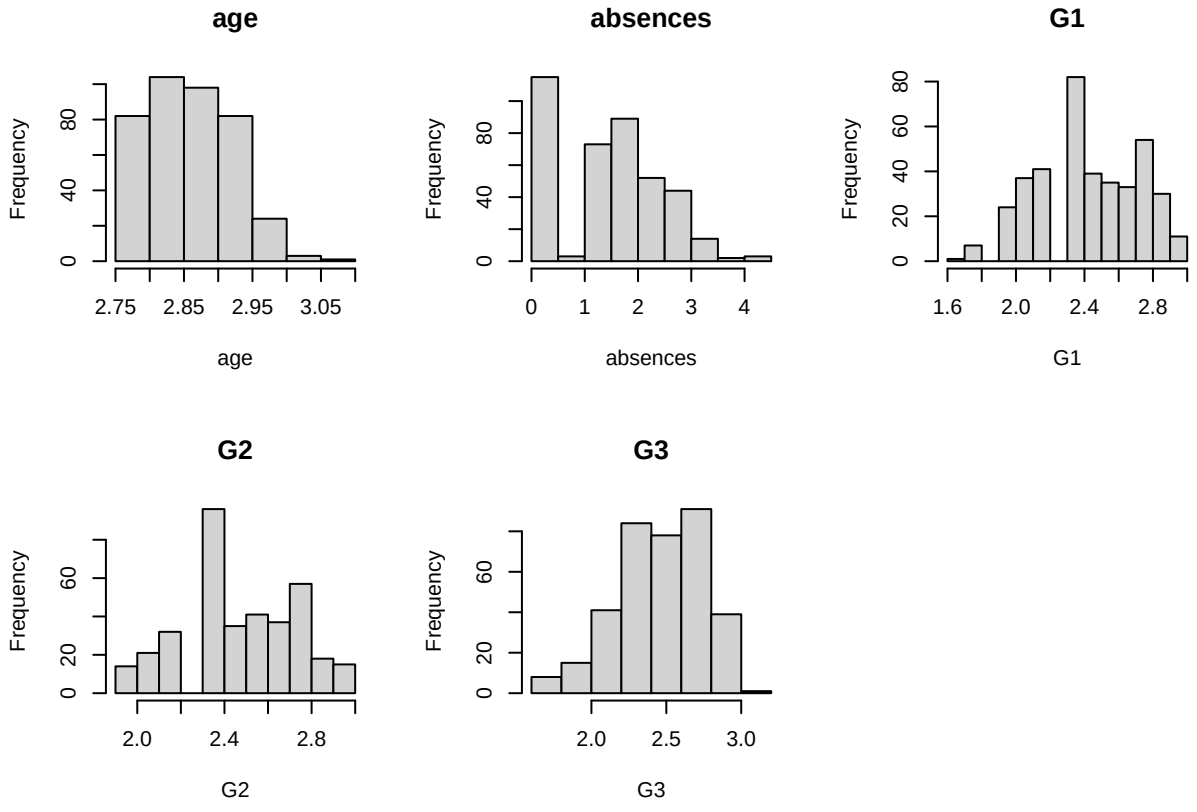
```
handle_with_quantile <- function(x, threshold = 1.5) {
  # Convert x to atomic vector
  x <- unlist(x)
  # Calculate quartiles
  q <- quantile(x, probs = c(0.25, 0.75), na.rm = TRUE)
  # Calculate interquartile range
  iqr <- diff(q)
  # Calculate lower and upper bounds
  lower_bound <- q[1] - threshold * iqr
  upper_bound <- q[2] + threshold * iqr
```

```r
  # Identify outliers
  outliers <- which(x < lower_bound | x > upper_bound)
  # Replace outliers with NA
  x[outliers] <- NA

  return(x)
}
clean_data <- lapply(log_transform_columns, handle_with_quantile)
str(clean_data)
```

```
## List of 5
##  $ age     : num [1:395] 2.94 2.89 2.77 2.77 2.83 ...
##  $ absences: num [1:395] 1.95 1.61 2.4 1.1 1.61 ...
##  $ G1      : num [1:395] 1.79 1.79 2.08 2.77 1.95 ...
##  $ G2      : num [1:395] 1.95 NA 2.2 2.71 2.4 ...
##  $ G3      : num [1:395] 1.95 1.95 2.4 2.77 2.4 ...
```

```r
par(mfrow=c(2, 3))
for (i in 1:length(clean_data)) {
  hist(clean_data[[i]], main = names(log_transform_columns)[i], xlab = names(log_transform_columns)[i])
}
```



5.4. Remodel with Linear Regression:

```
clean_data<- na.omit(clean_data)
clean_data1 <- as.data.frame(clean_data)

latest_model <- lm(G3 ~ ., data = clean_data1)
summary(latest_model)
```

```
##
## Call:
## lm(formula = G3 ~ ., data = clean_data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34875 -0.02703 -0.00739  0.05114  0.26685
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.038169   0.189142  -0.202 0.840193
## age          0.031991   0.063365   0.505 0.613980
## absences    -0.005012   0.004386  -1.143 0.254024
## G1           0.120130   0.034753   3.457 0.000616 ***
## G2           0.866353   0.037576  23.056  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07824 on 342 degrees of freedom
##   (48 observations deleted due to missingness)
## Multiple R-squared:  0.9057, Adjusted R-squared:  0.9046
## F-statistic:   821 on 4 and 342 DF,  p-value: < 2.2e-16
```

Here we checked the model using the cleaned data, the new model uses the age, absences, G1 and G2 features to predict the G3. It shows that both G1 and G2 are strongly significant for the prediction of the final grades. The R squared and its adjusted counterpart are similar in values but have increased from the previous model to 90%. the model p-value is 2.2e-16 implying this model is significantly strong.