PES University, Bangalore

UE20CS312 - Data Analytics

Worksheet 2b : Multiple Linear Regression

Solution Set

Course Anchor : Dr. Gowri Srinivasa

Prepared by : Nishanth M S - nishanthmsathish.23@gmail.com

Multiple Linear Regression

Multiple Linear Regression (MLR) is a statistical technique that uses several explanatory variables to predict the outcome of response variable.The goal of MLR is to model **a linear relationship** between explanatory(independent) variables and response(dependent) variables.

Data Dictionary

The data required for this worksheet can be downloaded from this GitHub Link. The data was obtained from this dataset from Kaggle. The dataset contains features of songs on Spotify collected using Spotify API.The features are as follows :

-acousticness : A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.

-danceability : Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.

-duration_ms : The duration of track in milliseconds.

-energy : Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity.Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.

-instrumentalness : Predicts whether a track contains no vocals.The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.

-key : The key the track is in. Integers map to pitches using standard Pitch Class notation

-liveness : Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.

-loudness : The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db.

-mode : Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.

-speechiness : Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.

-tempo : The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.

-time_signature : An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure).

-valence : A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

Throughout the course of this worksheet , our response variable is energy. We shall try and apply the concepts learnt in class to predict the energy of a song using the other features of a song.

Libraries used

-tidyverse

-corrplot

-olsrr : documentation

Points

The problems for this worksheet is for a total of 10 points and the weightage is not uniformly distributed.

- *Problem 1* : 0.5 points
- *Problem 2* : 2 points
- *Problem 3* : 2 points
- *Problem 4* : 1 point
- *Problem 5* : 1.5 points
- *Problem 6* : 1 point
- *Problem 7* : 2 points

Loading the Dataset

After downloading the dataset and ensuring the working directory is right , we read the csv into the dataframe.

```
library(tidyverse)
spotify_df <- read_csv('spotify.csv')
```

Problem-1 (0.5 Points)

Check for missing values in the dataset and normalize the dataset.

```
colSums(is.na(spotify_df)) #There are no missing values in the dataset
```

```
##     danceability           energy              key         loudness
##                0                0                0                0
##             mode      speechiness     acousticness instrumentalness
##                0                0                0                0
##         liveness          valence            tempo      duration_ms
##                0                0                0                0
##   time_signature
##                0
```

```
#Normalizing the dataset
spotify_df <- as.data.frame(scale(spotify_df))
```

Problem-2 (2 Points)

Fit a linear model to predict the *energy* rating using *all* other attributes.Get the summary of the model and explain the results in detail.[*Hint* : Use the lm() function. Click here To get the documentation of the same.]

```
full_model <- lm(energy ~ . , data=spotify_df)
summary(full_model)
```

```
##
## Call:
## lm(formula = energy ~ ., data = spotify_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.00232 -0.22889 -0.00973  0.27796  1.24597
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       9.156e-17  2.920e-02   0.000  1.00000
## danceability     -2.751e-01  5.555e-02  -4.952 1.67e-06 ***
## key               4.970e-02  3.009e-02   1.652  0.10030
## loudness          7.015e-01  4.561e-02  15.381  < 2e-16 ***
## mode             -4.794e-02  3.034e-02  -1.580  0.11582
## speechiness       2.359e-02  3.519e-02   0.670  0.50343
## acousticness     -3.435e-01  4.136e-02  -8.306 2.21e-14 ***
## instrumentalness  1.493e-01  5.577e-02   2.677  0.00811 **
## liveness          2.004e-02  3.100e-02   0.646  0.51880
## valence           2.046e-01  3.884e-02   5.269 3.85e-07 ***
## tempo            -2.395e-02  3.295e-02  -0.727  0.46817
## duration_ms      -1.865e-02  3.303e-02  -0.565  0.57298
## time_signature    2.409e-02  3.220e-02   0.748  0.45535
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

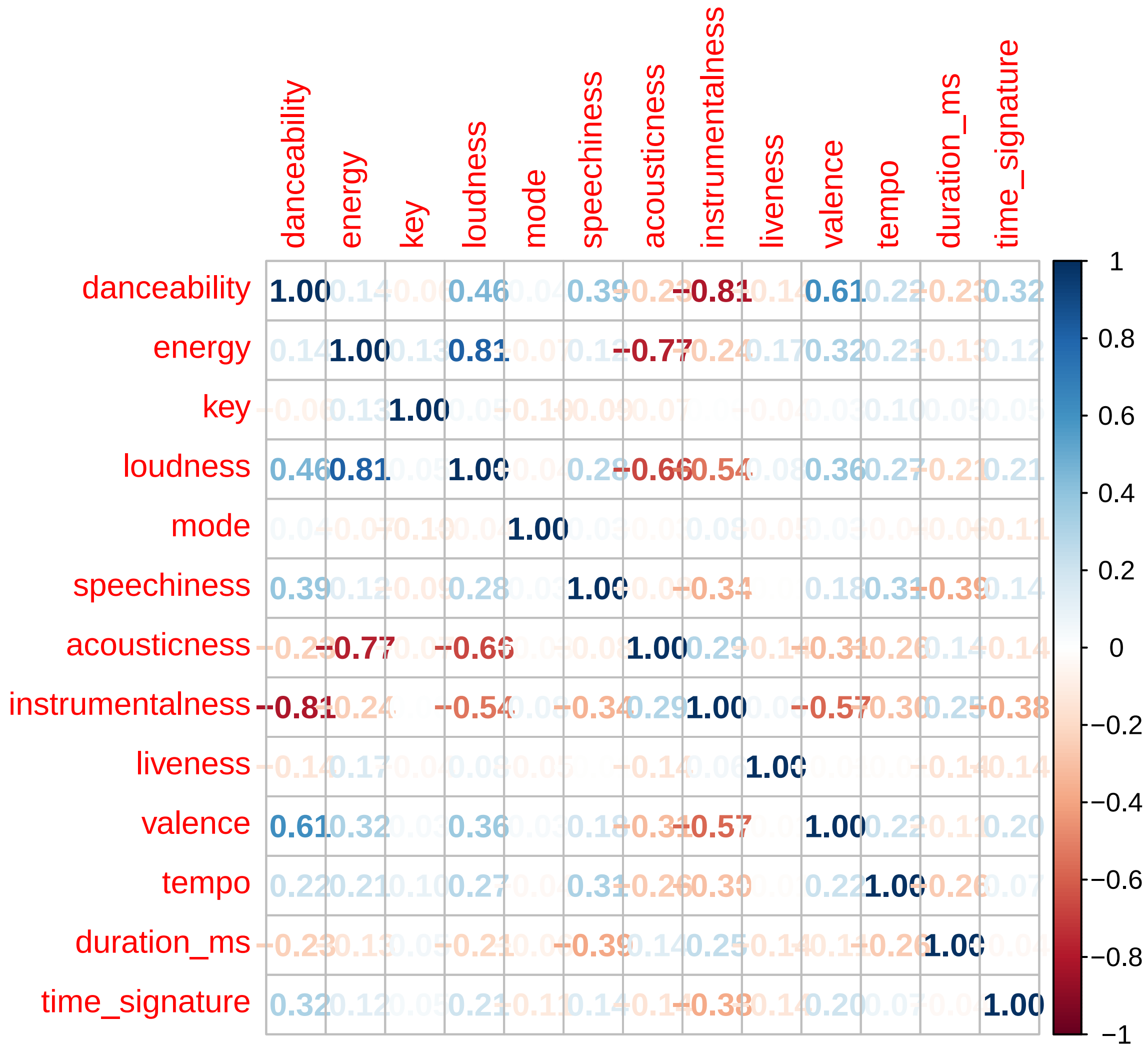
```
##
## Residual standard error: 0.4077 on 182 degrees of freedom
## Multiple R-squared:  0.844,  Adjusted R-squared:  0.8338
## F-statistic: 82.08 on 12 and 182 DF,  p-value: < 2.2e-16
```

- The F-statistic is 82.08 whose p-value is less than 2.2e-16. This indicates that the regression model as a whole is highly statistically significant.
- The *Estimate* column denotes the regression coefficients or the *Beta(s)*. Negatively correlated attributes have negative *Beta(s)* and positively correlated attributes have positive *Beta(s)*.
- For a given predictor , the t-statistic evaluates whether or not there is significant association between the predictor and the outcome variable, that is whether the beta coefficient of the predictor is significantly different from zero. The corresponding p-values are given.
- Thus the predictors *danceability , loudness , acousticness and valence* are statistically significant at a significance level of 0.001 and *instrumentalness* is statistically significant at 0.01.
- The **goodness of fit** of the model , R-squared is 0.844. It measures the total variablity explained by the model , ie , 84.4% of variability in energy is explained by the model.
- The adjusted R-squared , more valuable metric in multi linear regression which accounts for the number of independent predictors in the model is 0.8338.

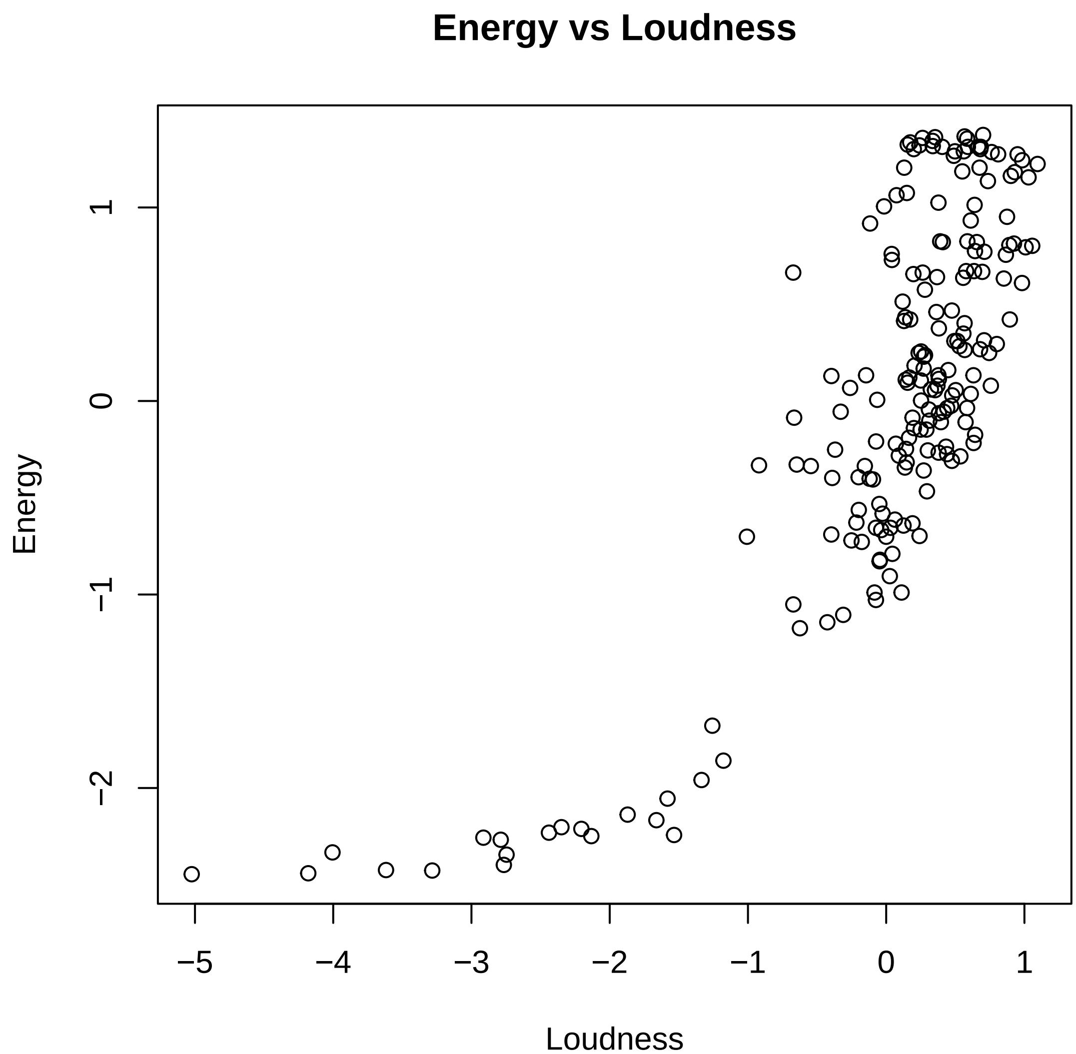
Problem-3 (2 points)

With the help of a correlogram and scatter plots, choose the features you think are important and model an MLR. Justify your choice and explain the new findings.

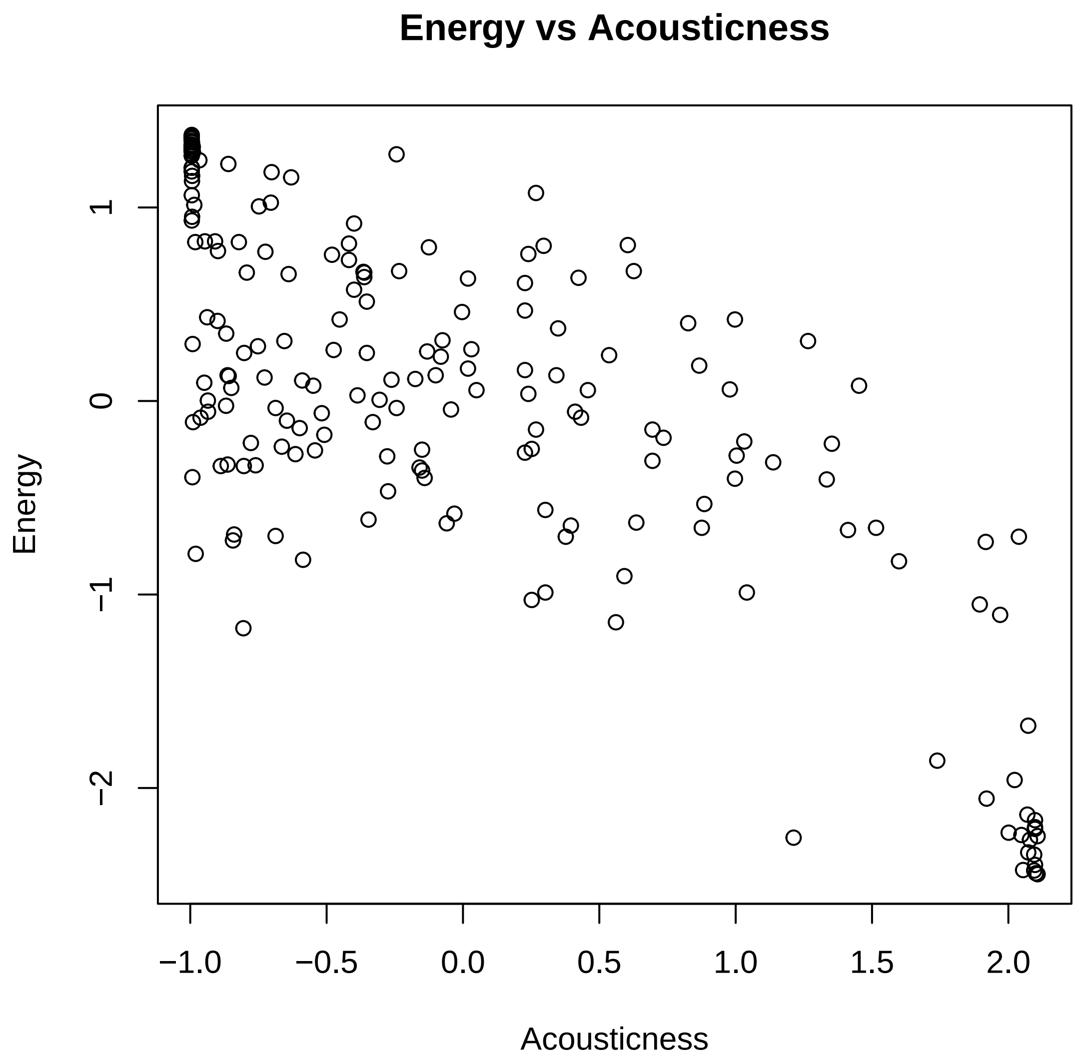
```r
library(corrplot)
correlation <- cor(spotify_df)
corrplot(correlation, method = 'number')
```




```r
plot(x = spotify_df$loudness,y = spotify_df$energy,
   xlab = "Loudness",
   ylab = "Energy",
   main = "Energy vs Loudness"
)
```




```
plot(x = spotify_df$acousticness,y = spotify_df$energy,
   xlab = "Acousticness",
   ylab = "Energy",
   main = "Energy vs Acousticness"
   )
```



NOTE : This question is subjective. I have made a choice of choosing only 2 predictors which has a strong correlation. Both the predictors have an absolute value of correlation higher than 0.7 . The scatter plots indicate a definitive trend. Moreover , from *Question-2* , i observed that these two predictors had a high statistical relationship with the dependent variable. Thus I have chosen these predictors.

Time to see how good my choices are!!

```
reduced_model <- lm(energy ~ loudness + acousticness , data = spotify_df)
summary(reduced_model)
```

```
##
## Call:
## lm(formula = energy ~ loudness + acousticness, data = spotify_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.22073 -0.34349  0.00132  0.34870  1.12953
```

```
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.001e-16  3.541e-02   0.000        1
## loudness      5.375e-01  4.753e-02  11.308  < 2e-16 ***
## acousticness -4.152e-01  4.753e-02  -8.734  1.2e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4945 on 192 degrees of freedom
## Multiple R-squared:  0.758,  Adjusted R-squared:  0.7555
## F-statistic: 300.7 on 2 and 192 DF,  p-value: < 2.2e-16
```

As one can observe , merely choosing the most correlated predictors doesn't help as the goodness of fit is 0.758 . This means I must have included more predictors. But wait , is there a statistical way to confirm this? Oh well , onto the next question for that!

Problem-4 (1 Point)

Conduct a partial F-test to determine if the attributes not chosen by you in *Problem-3* are significant to predict the energy.What are the null and alternate hypotheses? [*Hint* : Use the anova function between models created in *Problem-2* and *Problem-3*]

```
anova(reduced_model,full_model)
```

```
## Analysis of Variance Table
##
## Model 1: energy ~ loudness + acousticness
## Model 2: energy ~ danceability + key + loudness + mode + speechiness +
##     acousticness + instrumentalness + liveness + valence + tempo +
##     duration_ms + time_signature
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    192 46.942
## 2    182 30.257 10    16.686 10.037 2.416e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null hypothesis : The coefficients of regression for the variables not chosen are 0 ie the variables not chosen are not significant.

Alternate hypothesis : The coefficients of regression for the variables not chosen are not 0 ie the variables not chosen are statistically significant to determine energy.

The output shows the results of the partial F-test. For F = 10.037 , the corresponding p value is 2.416e-13 , thus, at 1% (and 5%) level of significance , the null hypothesis can be rejected. Thus in this scenario , the attributes not chosen turns out to be significant.

Problem-5 (1.5 Points)

AIC - Akaike Information Criterion is used to compare different models and determine the best fit for the data. The best-fit model according to AIC is the one that explains greatest amount of variation using the fewest number of attributes. Check this resource to learn more about AIC.

Build a model based on AIC using Stepwise AIC regression. Elucidate your observations from the new model. (*Hint* : Use an appropriate function in olsrr package.)

```
library(olsrr)
stepwise <- ols_step_both_aic(full_model, progress = TRUE , details = TRUE)
```

```
## Stepwise Selection Method
## -------------------------
##
## Candidate Terms:
##
## 1 . danceability
## 2 . key
## 3 . loudness
## 4 . mode
## 5 . speechiness
## 6 . acousticness
## 7 . instrumentalness
## 8 . liveness
## 9 . valence
## 10 . tempo
## 11 . duration_ms
## 12 . time_signature
##
##  Step 0: AIC = 556.3835
##  energy ~ 1
##
##
## Variables Entered/Removed:
##
##                               Enter New Variables
## ------------------------------------------------------------------------------
## Variable            DF      AIC      Sum Sq       RSS      R-Sq     Adj. R-Sq
## ------------------------------------------------------------------------------
## loudness             1    346.927    128.407     65.593    0.662        0.660
## acousticness         1    381.220    115.796     78.204    0.597        0.595
## valence              1    537.400     19.792    174.208    0.102        0.097
## instrumentalness     1    546.671     11.309    182.691    0.058        0.053
## tempo                1    549.163      8.960    185.040    0.046        0.041
## liveness             1    552.901      5.379    188.621    0.028        0.023
## danceability         1    554.678      3.651    190.349    0.019        0.014
## duration_ms          1    554.822      3.511    190.489    0.018        0.013
## key                  1    555.047      3.291    190.709    0.017        0.012
## time_signature       1    555.365      2.980    191.020    0.015        0.010
## speechiness          1    555.419      2.927    191.073    0.015        0.010
## mode                 1    557.471      0.905    193.095    0.005        0.000
## ------------------------------------------------------------------------------
##
## - loudness added
##
##
##  Step 1 : AIC = 346.9275
##  energy ~ loudness
##
```

```
##                              Enter New Variables
## -------------------------------------------------------------------------------
## Variable            DF      AIC      Sum Sq      RSS      R-Sq     Adj. R-Sq
## -------------------------------------------------------------------------------
## acousticness         1    283.690    147.058    46.942    0.758        0.756
## danceability         1    304.782    141.695    52.305    0.730        0.728
## instrumentalness     1    314.757    138.950    55.050    0.716        0.713
## speechiness          1    341.940    130.716    63.284    0.674        0.670
## liveness             1    342.675    130.477    63.523    0.673        0.669
## key                  1    343.959    130.057    63.943    0.670        0.667
## time_signature       1    347.696    128.820    65.180    0.664        0.661
## mode                 1    348.243    128.637    65.363    0.663        0.660
## duration_ms          1    348.257    128.632    65.368    0.663        0.660
## valence              1    348.555    128.532    65.468    0.663        0.659
## tempo                1    348.884    128.422    65.578    0.662        0.658
## -------------------------------------------------------------------------------
##
## - acousticness added
##
##
##  Step 2 : AIC = 283.6903
##  energy ~ loudness + acousticness
##
##                        Remove Existing Variables
## ---------------------------------------------------------------------------
## Variable        DF      AIC      Sum Sq      RSS      R-Sq     Adj. R-Sq
## ---------------------------------------------------------------------------
## acousticness     1    346.927    128.407    65.593    0.662        0.660
## loudness         1    381.220    115.796    78.204    0.597        0.595
## ---------------------------------------------------------------------------
##
##                              Enter New Variables
## -------------------------------------------------------------------------------
## Variable            DF      AIC      Sum Sq      RSS      R-Sq     Adj. R-Sq
## -------------------------------------------------------------------------------
## danceability         1    237.092    157.413    36.587    0.811        0.808
## instrumentalness     1    249.354    155.038    38.962    0.799        0.796
## key                  1    280.791    148.222    45.778    0.764        0.760
## liveness             1    282.074    147.920    46.080    0.762        0.759
## speechiness          1    282.379    147.848    46.152    0.762        0.758
## mode                 1    283.075    147.683    46.317    0.761        0.758
## time_signature       1    283.839    147.501    46.499    0.760        0.757
## tempo                1    284.379    147.372    46.628    0.760        0.756
## duration_ms          1    284.715    147.292    46.708    0.759        0.755
## valence              1    285.653    147.067    46.933    0.758        0.754
## -------------------------------------------------------------------------------
##
## - danceability added
##
##
##  Step 3 : AIC = 237.0918
##  energy ~ loudness + acousticness + danceability
##
##                        Remove Existing Variables
```

```
## -----------------------------------------------------------------------
## Variable        DF      AIC      Sum Sq      RSS      R-Sq     Adj. R-Sq
## -----------------------------------------------------------------------
## danceability     1    283.690    147.058    46.942    0.758        0.756
## acousticness     1    304.782    141.695    52.305    0.730        0.728
## loudness         1    382.238    116.188    77.812    0.599        0.595
## -----------------------------------------------------------------------
##
##                              Enter New Variables
## ---------------------------------------------------------------------------
## Variable            DF      AIC      Sum Sq      RSS      R-Sq     Adj. R-Sq
## ---------------------------------------------------------------------------
## valence              1    215.654    161.556    32.444    0.833        0.829
## key                  1    235.753    158.034    35.966    0.815        0.811
## instrumentalness     1    236.925    157.817    36.183    0.813        0.810
## mode                 1    237.532    157.704    36.296    0.813        0.809
## liveness             1    238.470    157.529    36.471    0.812        0.808
## time_signature       1    238.941    157.441    36.559    0.812        0.808
## tempo                1    238.966    157.437    36.563    0.812        0.808
## speechiness          1    239.070    157.417    36.583    0.811        0.807
## duration_ms          1    239.086    157.414    36.586    0.811        0.807
## ---------------------------------------------------------------------------
##
## - valence added
##
##
##  Step 4 : AIC = 215.6543
##  energy ~ loudness + acousticness + danceability + valence
##
##                        Remove Existing Variables
## -----------------------------------------------------------------------
## Variable        DF      AIC      Sum Sq      RSS      R-Sq     Adj. R-Sq
## -----------------------------------------------------------------------
## valence          1    237.092    157.413    36.587    0.811        0.808
## acousticness     1    276.308    149.263    44.737    0.769        0.766
## danceability     1    285.653    147.067    46.933    0.758        0.754
## loudness         1    375.553    119.578    74.422    0.616        0.610
## -----------------------------------------------------------------------
##
##                              Enter New Variables
## ---------------------------------------------------------------------------
## Variable            DF      AIC      Sum Sq      RSS      R-Sq     Adj. R-Sq
## ---------------------------------------------------------------------------
## instrumentalness     1    212.234    162.446    31.554    0.837        0.833
## key                  1    215.223    161.958    32.042    0.835        0.830
## mode                 1    215.870    161.852    32.148    0.834        0.830
## tempo                1    217.171    161.637    32.363    0.833        0.829
## liveness             1    217.336    161.609    32.391    0.833        0.829
## speechiness          1    217.378    161.602    32.398    0.833        0.829
## time_signature       1    217.458    161.589    32.411    0.833        0.829
## duration_ms          1    217.528    161.577    32.423    0.833        0.828
## ---------------------------------------------------------------------------
##
## - instrumentalness added
```

```
##
##
##  Step 5 : AIC = 212.2341
##  energy ~ loudness + acousticness + danceability + valence + instrumentalness
##
##                         Remove Existing Variables
## ---------------------------------------------------------------------------
## Variable            DF      AIC      Sum Sq      RSS      R-Sq     Adj. R-Sq
## ---------------------------------------------------------------------------
## instrumentalness     1    215.654    161.556    32.444    0.833        0.829
## valence              1    236.925    157.817    36.183    0.813        0.810
## danceability         1    239.721    157.295    36.705    0.811        0.807
## acousticness         1    272.294    150.622    43.378    0.776        0.772
## loudness             1    375.313    120.428    73.572    0.621        0.613
## ---------------------------------------------------------------------------
##
##                            Enter New Variables
## -------------------------------------------------------------------------
## Variable          DF      AIC      Sum Sq      RSS      R-Sq     Adj. R-Sq
## -------------------------------------------------------------------------
## mode               1    211.005    162.964    31.036    0.840        0.835
## key                1    211.356    162.908    31.092    0.840        0.835
## time_signature     1    213.333    162.591    31.409    0.838        0.833
## liveness           1    213.785    162.518    31.482    0.838        0.833
## speechiness        1    213.895    162.501    31.499    0.838        0.832
## duration_ms        1    213.920    162.497    31.503    0.838        0.832
## tempo              1    214.109    162.466    31.534    0.837        0.832
## -------------------------------------------------------------------------
##
## - mode added
##
##
##  Step 6 : AIC = 211.0053
##  energy ~ loudness + acousticness + danceability + valence + instrumentalness + mode
##
##                         Remove Existing Variables
## ---------------------------------------------------------------------------
## Variable            DF      AIC      Sum Sq      RSS      R-Sq     Adj. R-Sq
## ---------------------------------------------------------------------------
## mode                 1    212.234    162.446    31.554    0.837        0.833
## instrumentalness     1    215.870    161.852    32.148    0.834        0.830
## danceability         1    235.412    158.463    35.537    0.817        0.812
## valence              1    236.564    158.253    35.747    0.816        0.811
## acousticness         1    273.228    150.858    43.142    0.778        0.772
## loudness             1    374.348    121.538    72.462    0.626        0.617
## ---------------------------------------------------------------------------
##
##                            Enter New Variables
## -------------------------------------------------------------------------
## Variable          DF      AIC      Sum Sq      RSS      R-Sq     Adj. R-Sq
## -------------------------------------------------------------------------
## key                1    210.607    163.343    30.657    0.842        0.836
## time_signature     1    212.391    163.062    30.938    0.841        0.835
## duration_ms        1    212.509    163.043    30.957    0.840        0.834
```

```
## speechiness        1    212.570    163.033    30.967    0.840        0.834
## liveness           1    212.628    163.024    30.976    0.840        0.834
## tempo              1    212.854    162.988    31.012    0.840        0.834
## ------------------------------------------------------------------------
##
## - key added
##
##
##  Step 7 : AIC = 210.6068
##  energy ~ loudness + acousticness + danceability + valence + instrumentalness + mode + key
##
##                          Remove Existing Variables
## --------------------------------------------------------------------------
## Variable             DF      AIC      Sum Sq      RSS      R-Sq     Adj. R-Sq
## --------------------------------------------------------------------------
## key                   1    211.005    162.964    31.036    0.840        0.835
## mode                  1    211.356    162.908    31.092    0.840        0.835
## instrumentalness      1    215.798    162.192    31.808    0.836        0.831
## danceability          1    233.547    159.161    34.839    0.820        0.815
## valence               1    235.302    158.846    35.154    0.819        0.813
## acousticness          1    272.812    151.389    42.611    0.780        0.773
## loudness              1    374.763    122.125    71.875    0.630        0.618
## --------------------------------------------------------------------------
##
##                             Enter New Variables
## ------------------------------------------------------------------------
## Variable           DF      AIC      Sum Sq      RSS      R-Sq     Adj. R-Sq
## ------------------------------------------------------------------------
## speechiness        1    211.994    163.440    30.560    0.842        0.836
## duration_ms        1    211.997    163.439    30.561    0.842        0.836
## liveness           1    212.057    163.430    30.570    0.842        0.836
## time_signature     1    212.098    163.423    30.577    0.842        0.836
## tempo              1    212.331    163.387    30.613    0.842        0.835
## ------------------------------------------------------------------------
##
##
## No more variables to be added or removed.
##
## Final Model Output
## ------------------
##
##                            Model Summary
## ------------------------------------------------------------------
## R                       0.918       RMSE                    0.405
## R-Squared               0.842       Coef. Var        9.628172e+17
## Adj. R-Squared          0.836       MSE                     0.164
## Pred R-Squared          0.824       MAE                     0.315
## ------------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                 ANOVA
## -------------------------------------------------------------------
```

```
##                Sum of
##               Squares         DF    Mean Square       F         Sig.
## --------------------------------------------------------------------
## Regression    163.343          7         23.335    142.338    0.0000
## Residual       30.657        187          0.164
## Total         194.000        194
## --------------------------------------------------------------------
##
##                                      Parameter Estimates
## ----------------------------------------------------------------------------------------------
##            model      Beta    Std. Error    Std. Beta      t        Sig      lower     upper
## ----------------------------------------------------------------------------------------------
##      (Intercept)     0.000         0.029                  0.000    1.000    -0.057     0.057
##         loudness     0.708         0.045        0.708    15.856    0.000     0.620     0.796
##     acousticness    -0.342         0.040       -0.342    -8.539    0.000    -0.421    -0.263
##     danceability    -0.268         0.053       -0.268    -5.051    0.000    -0.373    -0.163
##          valence     0.200         0.038        0.200     5.238    0.000     0.125     0.276
## instrumentalness     0.142         0.054        0.142     2.650    0.009     0.036     0.247
##             mode    -0.049         0.030       -0.049    -1.629    0.105    -0.108     0.010
##              key     0.045         0.030        0.045     1.521    0.130    -0.013     0.103
## ----------------------------------------------------------------------------------------------
```

The step-wise selection method showed how the attribute selection happened at every step. Whichever feature resulted in a lower AIC , that was added to the list.

```
stepwise_model <- lm(energy ~ loudness + acousticness + danceability + valence + instrumentalness + mode
summary(stepwise_model)
```

```
##
## Call:
## lm(formula = energy ~ loudness + acousticness + danceability +
##     valence + instrumentalness + mode + key, data = spotify_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.05662 -0.24874 -0.01126  0.27930  1.25974
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.999e-17  2.900e-02   0.000  1.00000
## loudness         7.075e-01  4.462e-02  15.856  < 2e-16 ***
## acousticness    -3.420e-01  4.005e-02  -8.539 4.63e-15 ***
## danceability    -2.681e-01  5.308e-02  -5.051 1.04e-06 ***
## valence          2.003e-01  3.825e-02   5.238 4.35e-07 ***
## instrumentalness  1.418e-01  5.351e-02   2.650  0.00873 **
## mode            -4.863e-02  2.985e-02  -1.629  0.10491
## key              4.488e-02  2.950e-02   1.521  0.12988
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4049 on 187 degrees of freedom
## Multiple R-squared:  0.842,  Adjusted R-squared:  0.8361
## F-statistic: 142.3 on 7 and 187 DF,  p-value: < 2.2e-16
```

Although the "full_model" has a marginally higher R-Squared , AIC favours the most simple model which does a good job. Without adding redundant attributes , the resulting model is much simpler and has (almost) a similar performance!

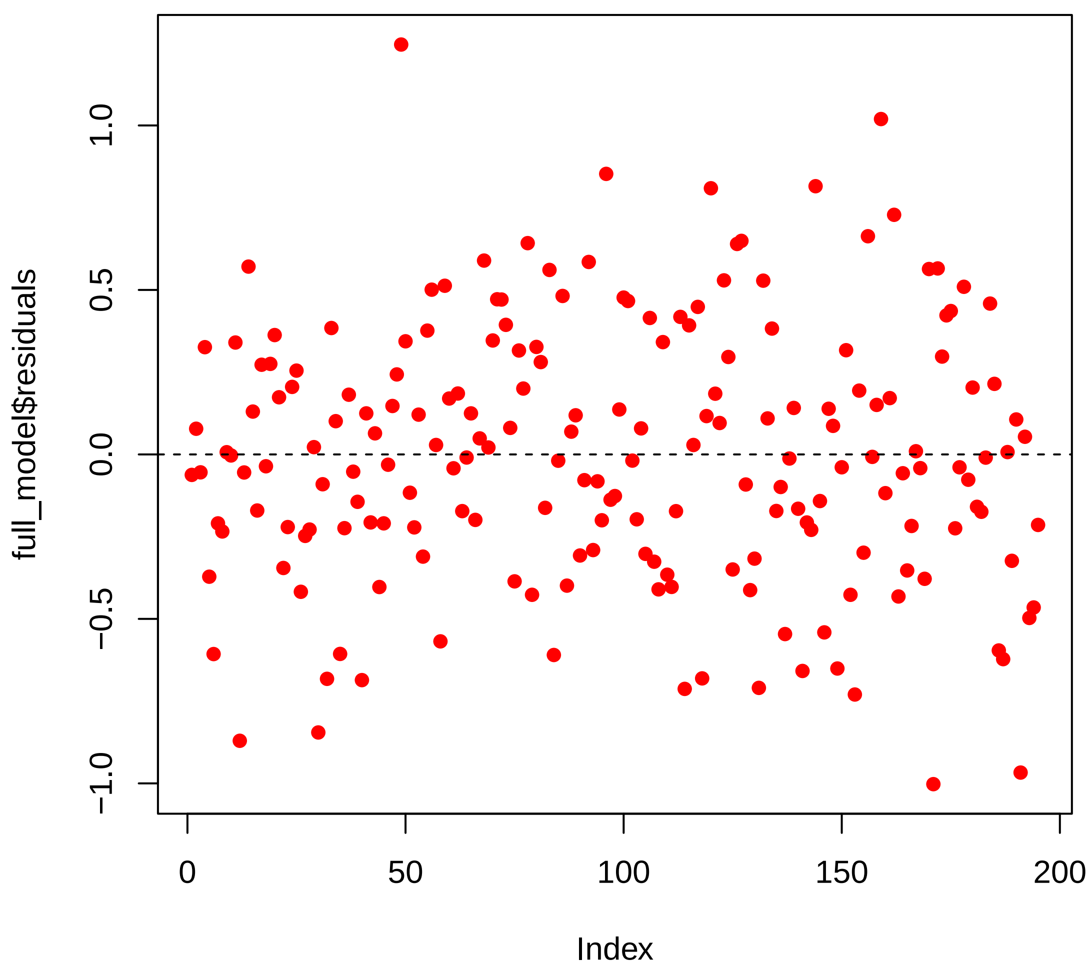
Problem-6 (1 Point)

Plot the residuals of the models built till now and comment on it satisfying the assumptions of MLR.

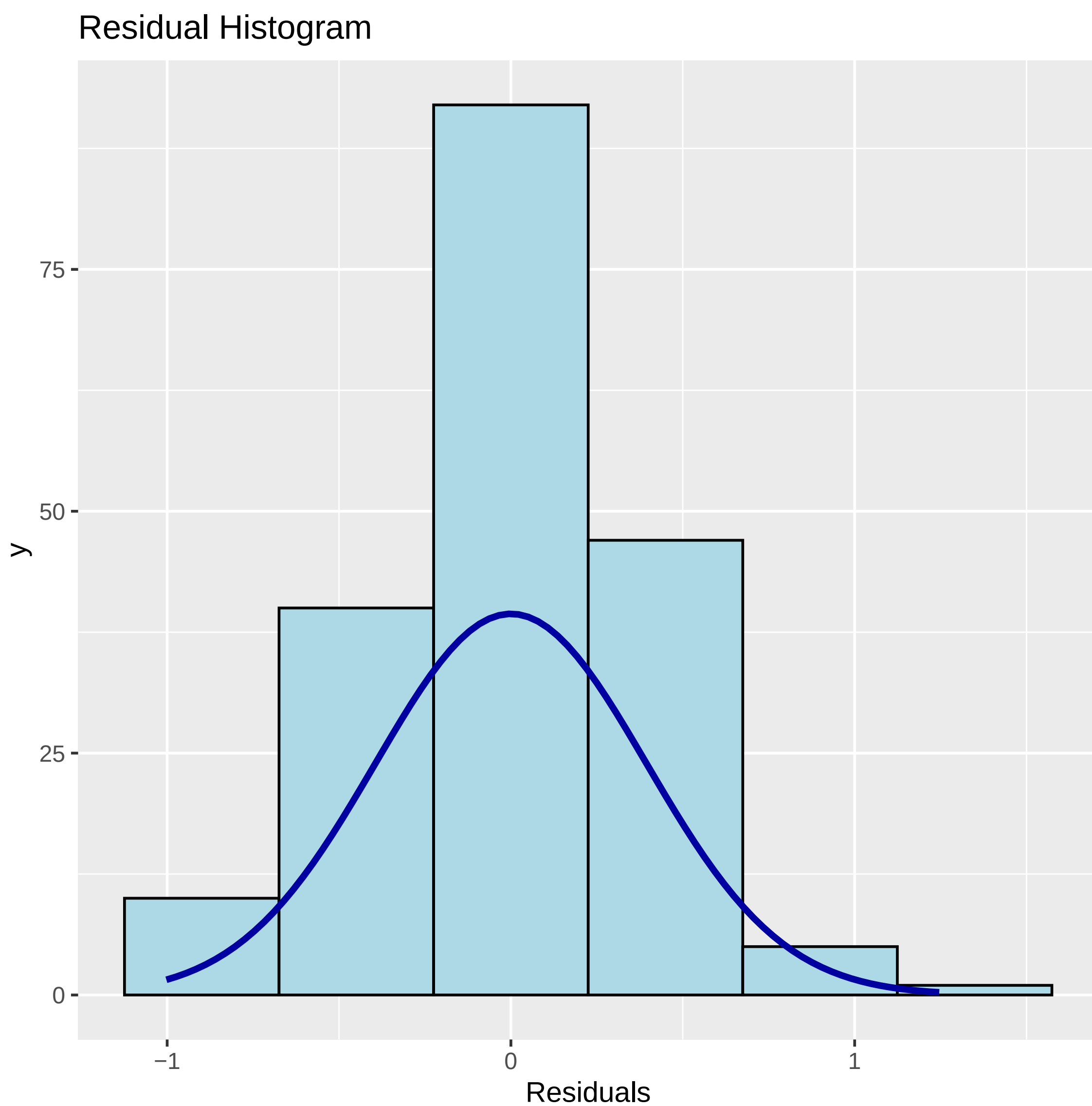
```
print("full_model residual plots")
```

```
## [1] "full_model residual plots"
```

```
plot(full_model$residuals, pch = 16, col = "red")
abline(h = 0, lty = 2)
```



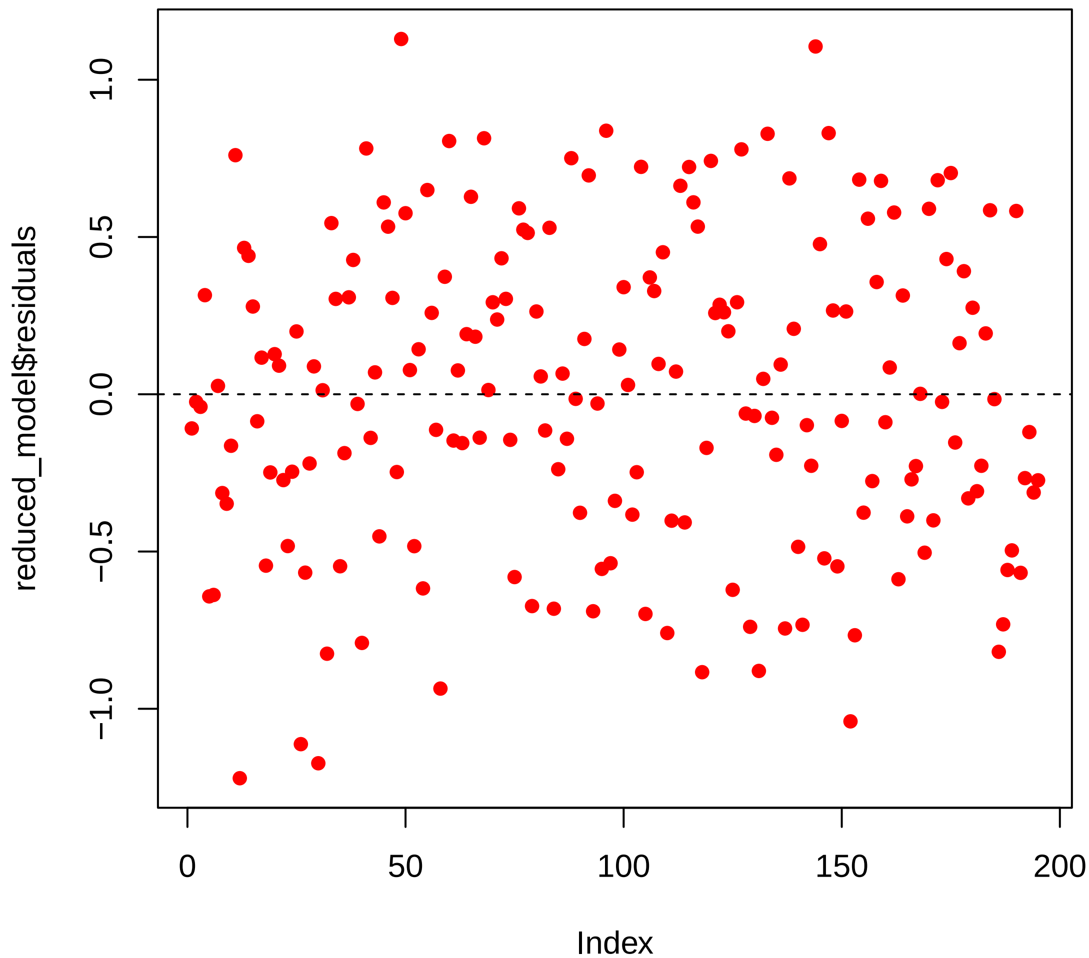
```
ols_plot_resid_hist(full_model)
```



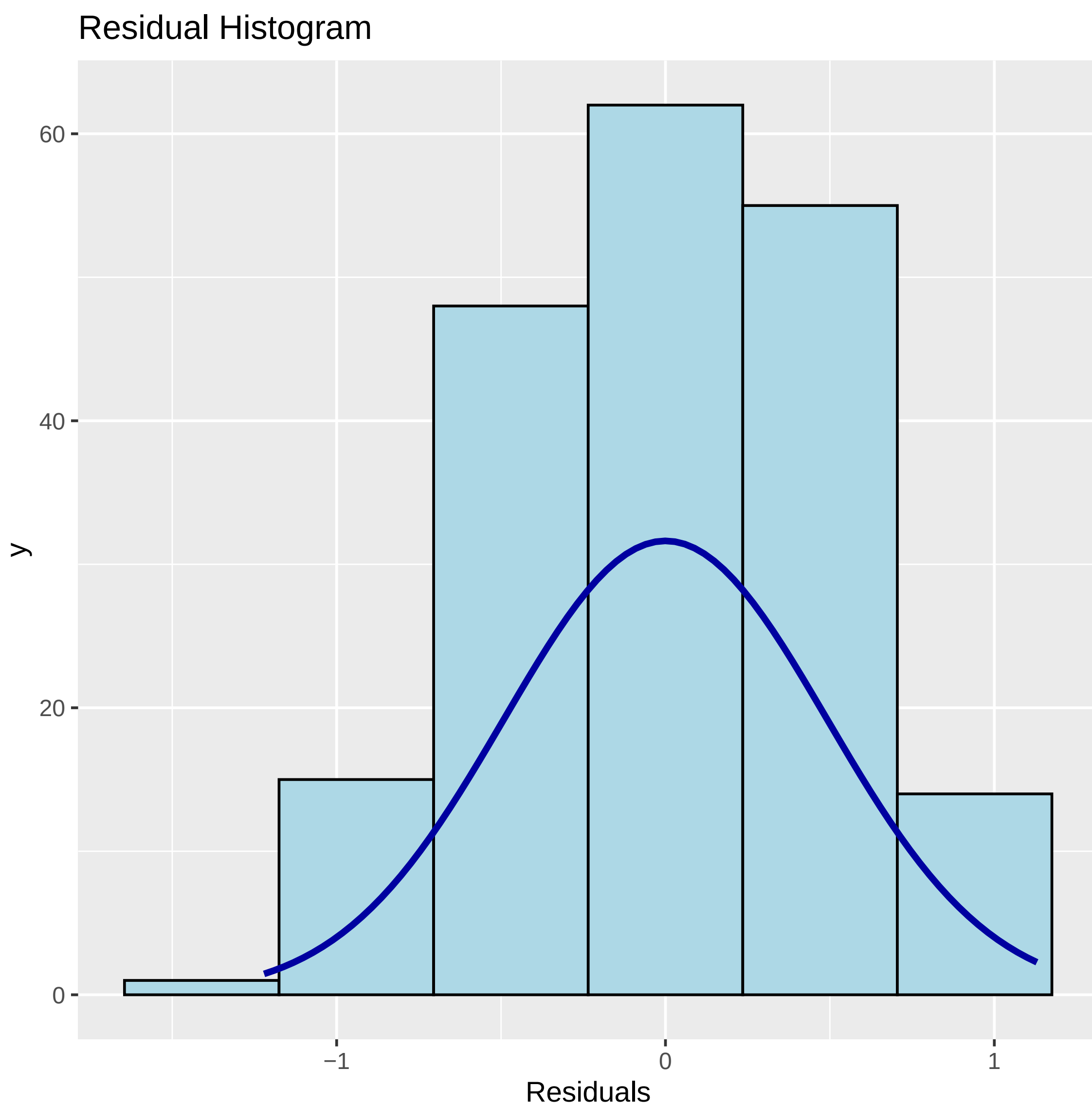

```
print("reduced_model residual plots")
```

```
## [1] "reduced_model residual plots"
```

```
plot(reduced_model$residuals, pch = 16, col = "red")
abline(h = 0, lty = 2)
```



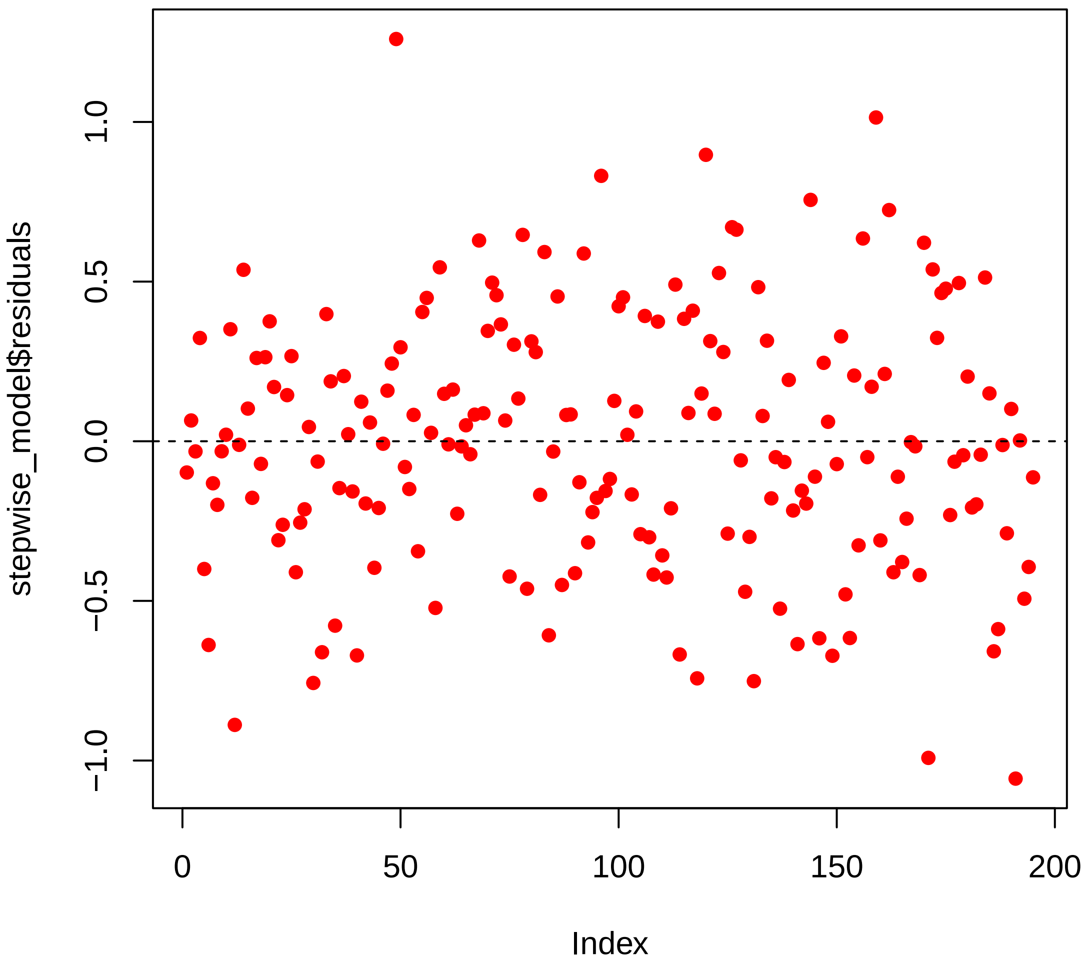

```
ols_plot_resid_hist(reduced_model)
```



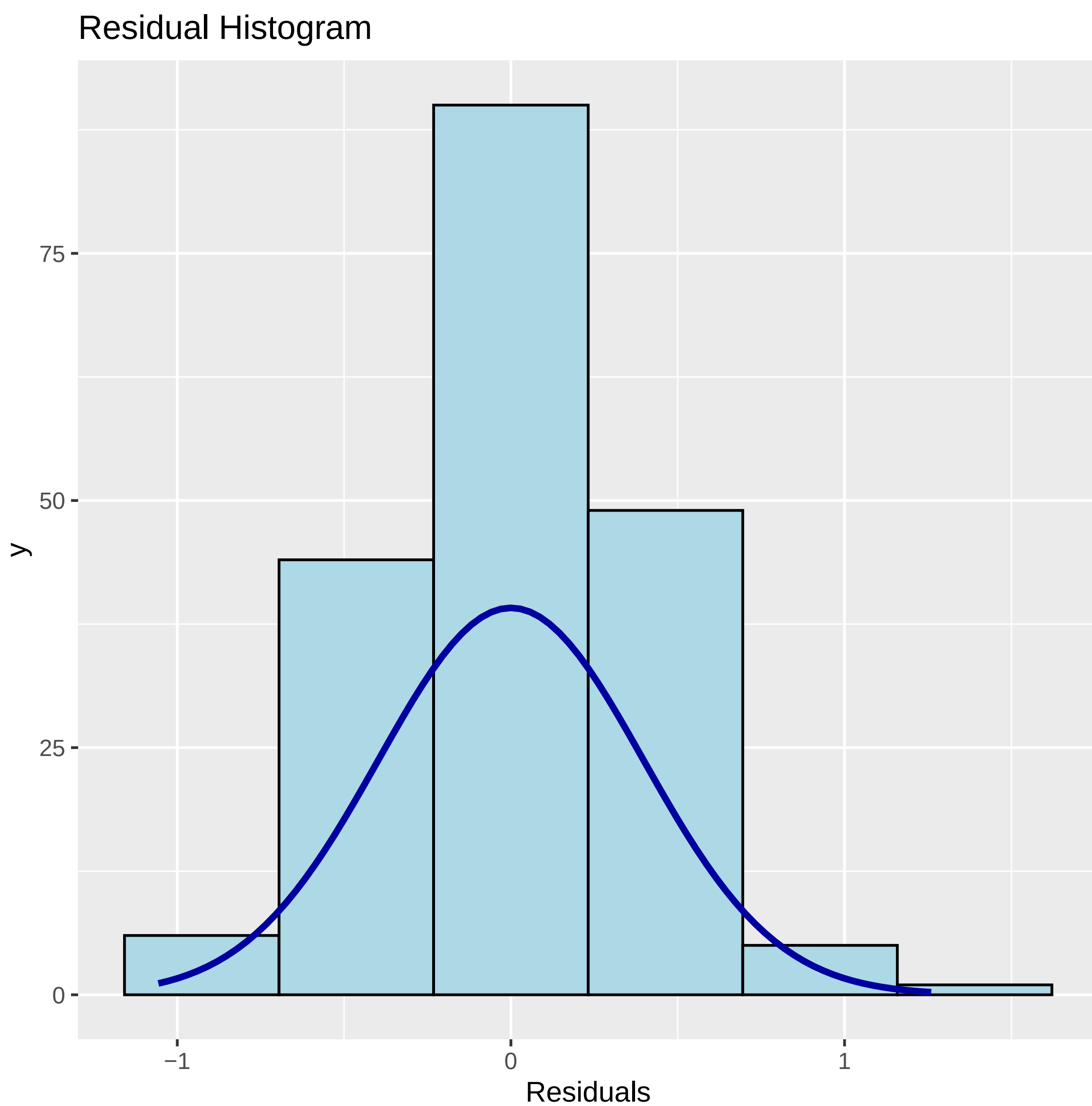

```r
print("stepwise_model residual plots")
```

```
## [1] "stepwise_model residual plots"
```

```r
plot(stepwise_model$residuals, pch = 16, col = "red")
abline(h = 0, lty = 2)
```




```
ols_plot_resid_hist(stepwise_model)
```



From the plots we can observe that the residuals of all the models are homoscedastic and approximately follow normal distribution.

Problem-7 (2 Points)

For the model built in ***Problem-2*** , determine the presence of multicollinearity using VIF. Determine if there are outliers in the data using Cook's Distance. If you find any , remove the outliers and fit the model for *Problem-2* and see if the fit improves. [*Hint* : All the relevant functions can be found in *olsrr* package. An observation can be termed as an outlier if it has a Cook's distance of more than 4/n where n is the number of records.]

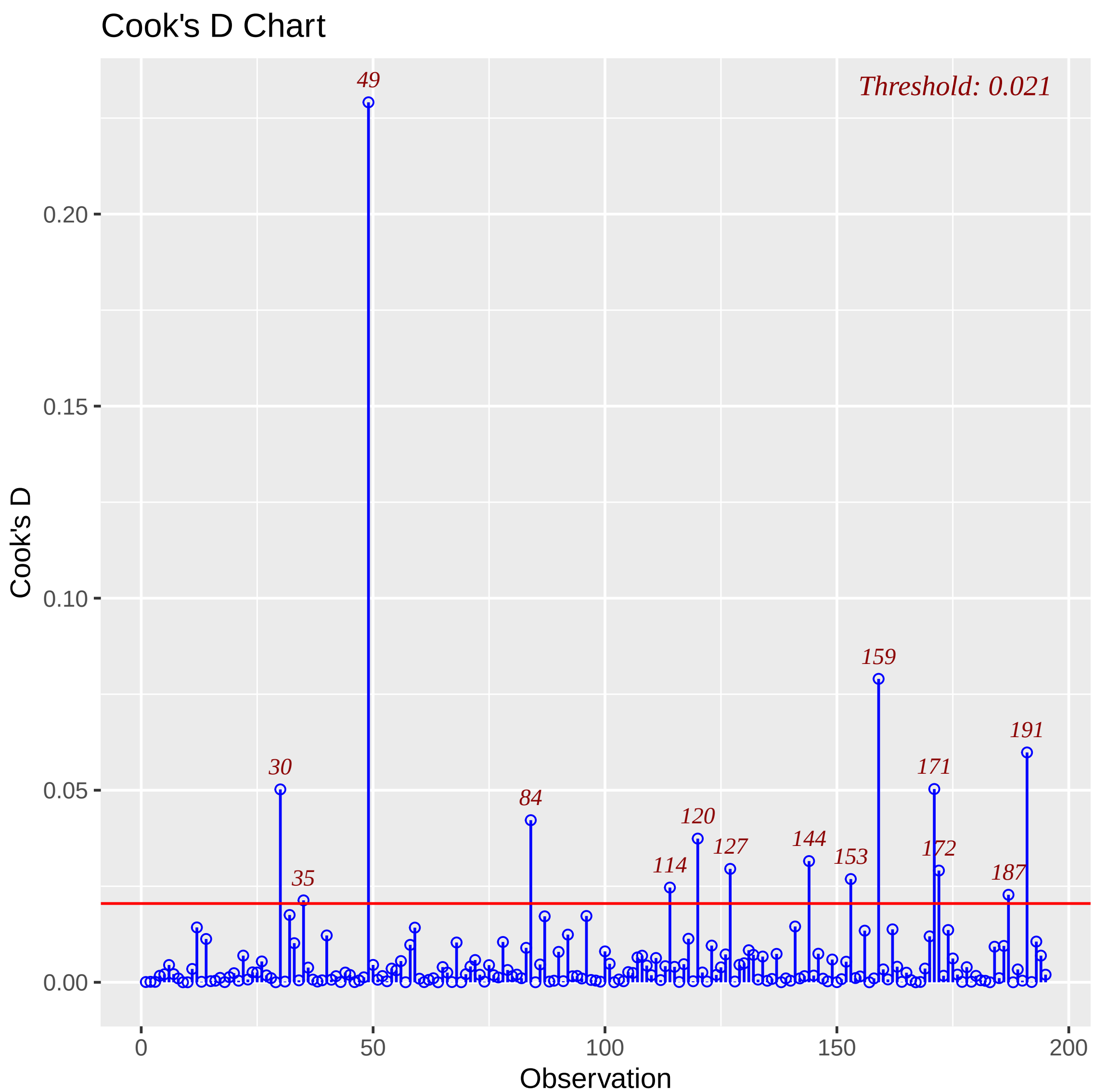
```
ols_vif_tol(full_model)
```

```
##          Variables Tolerance      VIF
## 1     danceability 0.2776703 3.601393
## 2              key 0.9467671 1.056226
```

```
## 3           loudness 0.4119898 2.427245
## 4               mode 0.9308390 1.074300
## 5        speechiness 0.6921660 1.444740
## 6       acousticness 0.5009458 1.996224
## 7   instrumentalness 0.2755568 3.629016
## 8           liveness 0.8914397 1.121781
## 9            valence 0.5680642 1.760364
## 10             tempo 0.7892957 1.266952
## 11       duration_ms 0.7855373 1.273014
## 12    time_signature 0.8262918 1.210226
```

We can conclude the absence of multicolinearity as the VIF is less than 5 for all attributes.

```r
cookd <- ols_plot_cooksd_chart(full_model)
```



The threshold is calculated by the formula 4/n which is 4/195 which is rounded to 0.021 . Let's remove the 14 outliers and see if we can achieve better fit.

```r
new_df <- spotify_df[-c(30,35,49,84,114,120,127,144,153,159,171,172,187,191),] #removing outliers
new_full_model <- lm(energy ~ . , data = new_df)
summary(new_full_model)
```

```
##
## Call:
## lm(formula = energy ~ ., data = new_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.76364 -0.20836  0.01581  0.23506  0.95145
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.001128   0.025283  -0.045 0.964458
## danceability     -0.258483   0.052291  -4.943 1.85e-06 ***
## key               0.088181   0.026094   3.379 0.000903 ***
## loudness          0.838411   0.045399  18.468  < 2e-16 ***
## mode             -0.012666   0.026559  -0.477 0.634036
## speechiness      -0.004528   0.032087  -0.141 0.887947
## acousticness     -0.280188   0.037293  -7.513 3.26e-12 ***
## instrumentalness  0.199483   0.051442   3.878 0.000151 ***
## liveness          0.028416   0.027232   1.043 0.298230
## valence           0.187216   0.033329   5.617 7.90e-08 ***
## tempo            -0.018193   0.029627  -0.614 0.540008
## duration_ms      -0.059788   0.028685  -2.084 0.038647 *
## time_signature    0.036680   0.028430   1.290 0.198761
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.337 on 168 degrees of freedom
## Multiple R-squared:  0.8778, Adjusted R-squared:  0.8691
## F-statistic: 100.6 on 12 and 168 DF,  p-value: < 2.2e-16
```

As you can observe , after removing the outliers , the goodness of fit has improved!