UE20CS312 - Data Analytics - Worksheet 2b : Multiple Linear Regression

PES University

Reshmi Pradeep, Dept. of CSE - PES2UG20CS270

2022-09-16

```
###PROBLEM 1
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
df<-read_csv('spotify.csv')
```

```
## Rows: 195 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (13): danceability, energy, key, loudness, mode, speechiness, acousticne...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(df)
```

```
## # A tibble: 6 x 13
##   danceabil~1 energy   key loudn~2  mode speec~3 acous~4 instr~5 liven~6 valence
##         <dbl>  <dbl> <dbl>   <dbl> <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1       0.803 0.624      7   -6.76     0  0.0477  0.451  7.34e-4  0.1     0.628
## 2       0.762 0.703     10   -7.95     0  0.306   0.206  0        0.0912  0.519
## 3       0.261 0.0149     1  -27.5      1  0.0419  0.992  8.97e-1  0.102   0.0382
## 4       0.722 0.736      3   -6.99     0  0.0585  0.431  1.18e-6  0.123   0.582
## 5       0.787 0.572      1   -7.52     1  0.222   0.145  0        0.0753  0.647
## 6       0.778 0.632      8   -6.42     1  0.125   0.0404 0        0.0912  0.827
## # ... with 3 more variables: tempo <dbl>, duration_ms <dbl>,
## #   time_signature <dbl>, and abbreviated variable names 1: danceability,
## #   2: loudness, 3: speechiness, 4: acousticness, 5: instrumentalness,
## #   6: liveness
```

```
colSums(is.na(df))
```

```
##     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
```

```
df<-as.data.frame(scale(df)) #normalizing
summary(df)
```

```
##   danceability         energy              key              loudness
##  Min.   :-2.3390   Min.   :-2.44537   Min.   :-1.6097   Min.   :-5.02359
##  1st Qu.:-0.8040   1st Qu.:-0.40343   1st Qu.:-1.0241   1st Qu.:-0.07362
##  Median : 0.3155   Median : 0.07908   Median : 0.1472   Median : 0.26293
##  Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.00000
##  3rd Qu.: 0.7495   3rd Qu.: 0.76537   3rd Qu.: 0.7328   3rd Qu.: 0.55978
##  Max.   : 1.4281   Max.   : 1.37476   Max.   : 1.6112   Max.   : 1.09510
##       mode          speechiness       acousticness     instrumentalness
##  Min.   :-1.0774   Min.   :-1.0062   Min.   :-0.9947   Min.   :-0.5555
##  1st Qu.:-1.0774   1st Qu.:-0.7653   1st Qu.:-0.8632   1st Qu.:-0.5555
##  Median : 0.9234   Median :-0.4381   Median :-0.3307   Median :-0.5555
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.9234   3rd Qu.: 0.6772   3rd Qu.: 0.5764   3rd Qu.:-0.2739
##  Max.   : 0.9234   Max.   : 3.2475   Max.   : 2.1071   Max.   : 2.2432
##     liveness          valence            tempo          duration_ms
##  Min.   :-1.0885   Min.   :-1.7121   Min.   :-2.1690   Min.   :-1.8878
##  1st Qu.:-0.6082   1st Qu.:-0.8391   1st Qu.:-0.7422   1st Qu.:-0.4866
##  Median :-0.4101   Median : 0.1172   Median : 0.1357   Median :-0.1304
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.2694   3rd Qu.: 0.8363   3rd Qu.: 0.7611   3rd Qu.: 0.4014
##  Max.   : 4.5723   Max.   : 1.8169   Max.   : 2.0990   Max.   : 6.1232
##  time_signature
##  Min.   :-6.4538
##  1st Qu.: 0.1932
##  Median : 0.1932
##  Mean   : 0.0000
##  3rd Qu.: 0.1932
##  Max.   : 2.4088
```

```
#AIC is decreasing with each attribute.Even with far fewer variables, the R2 has decreased by an insign


###PROBLEM 2
model<-lm(energy~., data=df)
summary(model)
```

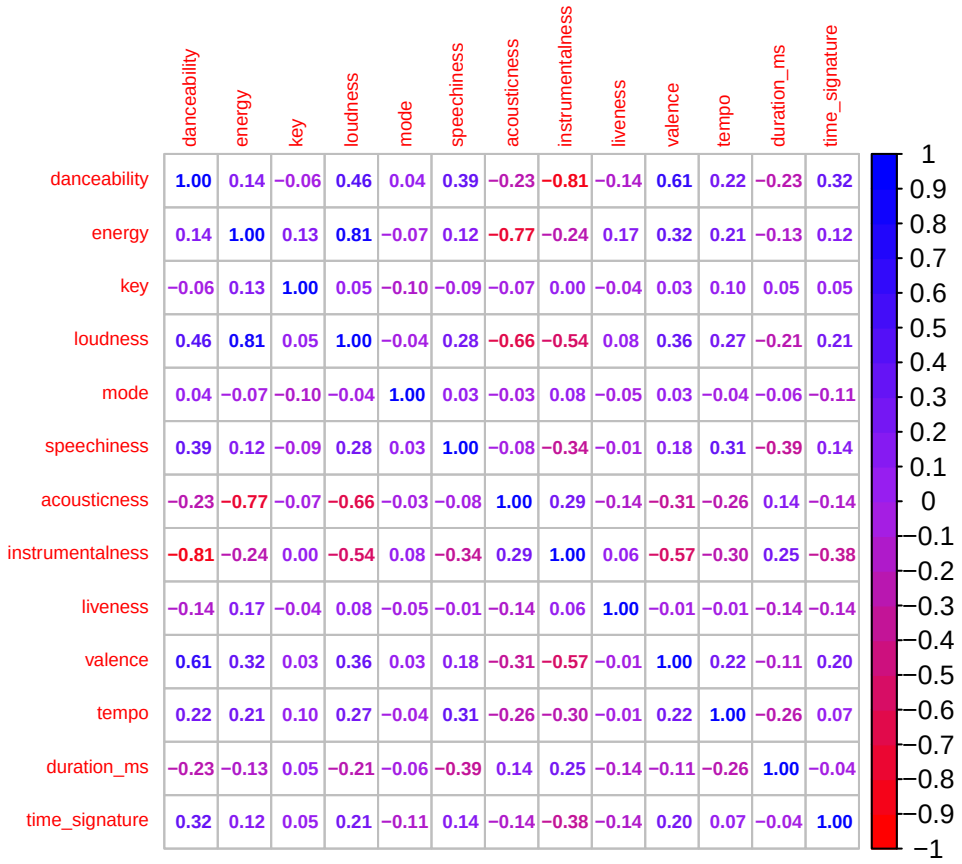
```
##
## Call:
```

```
## lm(formula = energy ~ ., data = 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
```

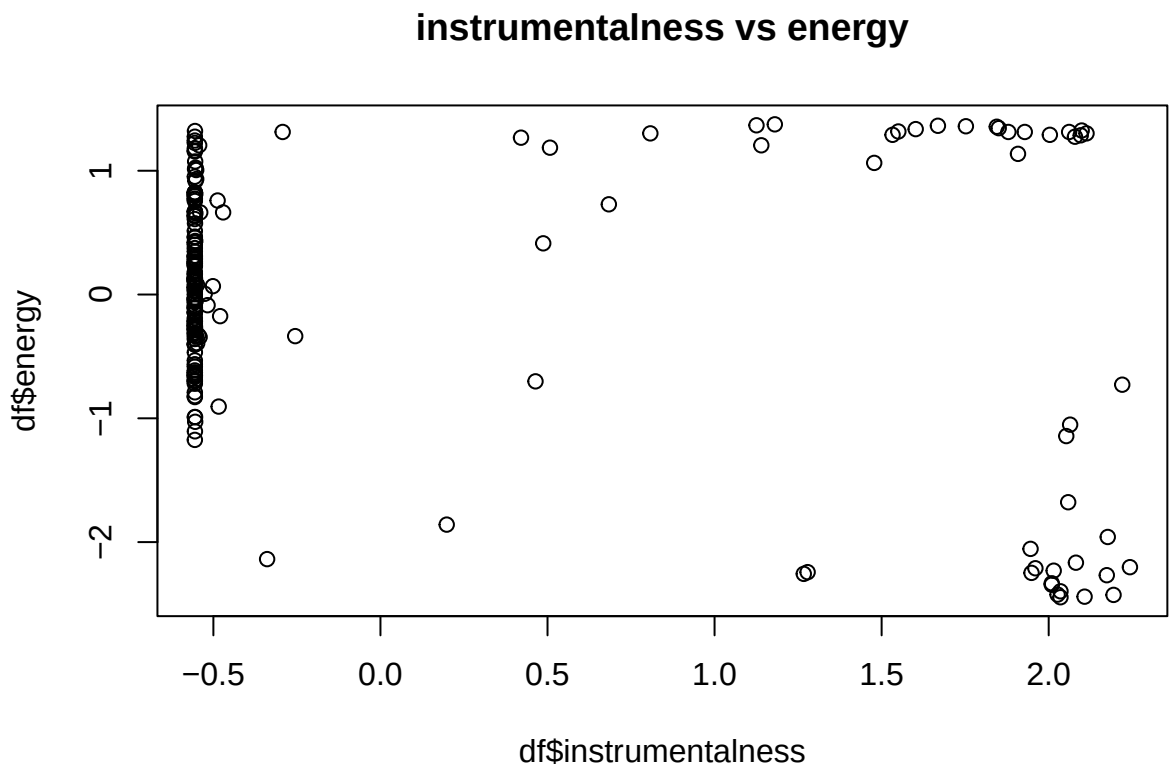
```
#The adjustment in the "Adjusted R Square" value in the summary output is a correction for the number o

###PROBLEM 3
library(corrplot)
```

```
## corrplot 0.92 loaded
```

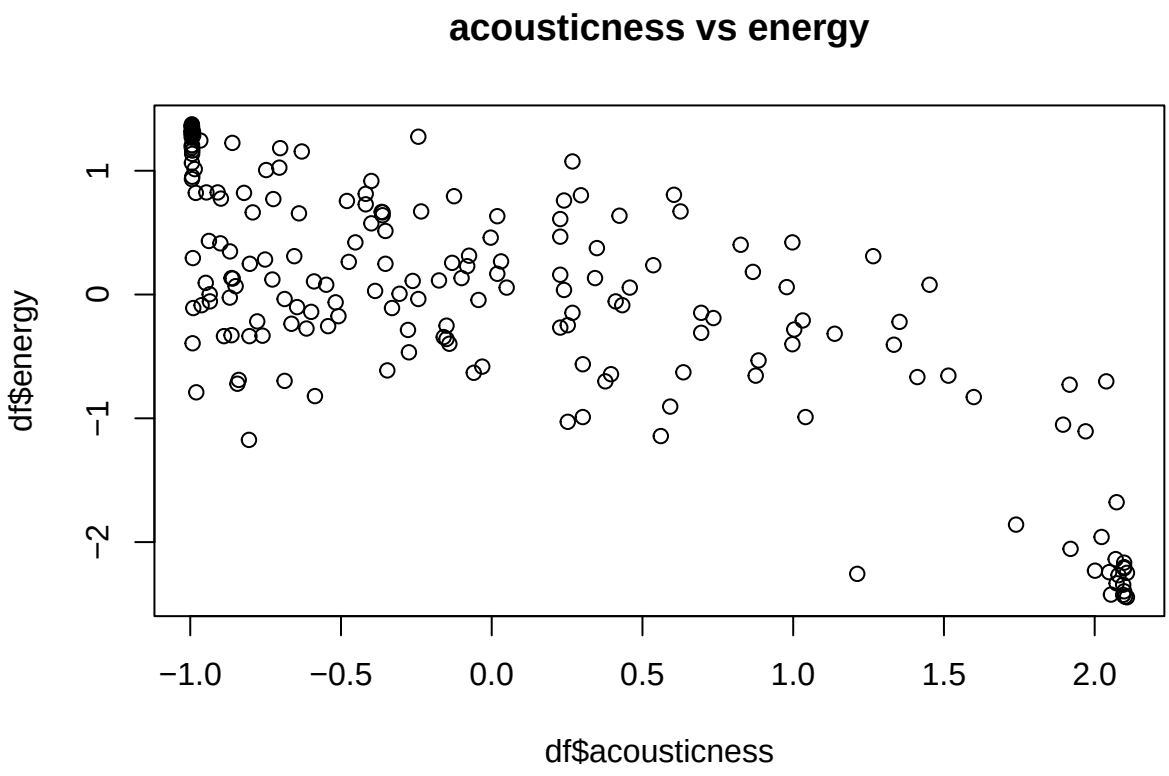
```
correl<-cor(df)
corrplot(correl,method='number',addCoef.col = 1, number.cex=0.6, tl.cex=0.6,col=colorRampPalette(c("red
```




```
#scatter plots
plot(x=df$instrumentalness,y=df$energy, main="instrumentalness vs energy")
```




```
plot(x=df$acousticness,y=df$energy, main="acousticness vs energy")
```




```
reduced<-lm(energy~loudness+acousticness, data=df) #reducing
summary(reduced)
```

```
##
## Call:
## lm(formula = energy ~ loudness + acousticness, data = 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
```

```
# p-value of the F-statistic is < 2.2e-16, which is highly significant, heaviest the predictor variable
```

```
###PROBLEM 4
anova(reduced,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
```

```
#H0: All coefficients removed from the full model are zero.
#A: At least one of the coefficients removed from the full model is non-zero.

###PROBLEM 5
library(olsrr)
```

```
##
## Attaching package: 'olsrr'
##
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
stepwise<-lm(energy~.,data=df)
summary(stepwise)
```

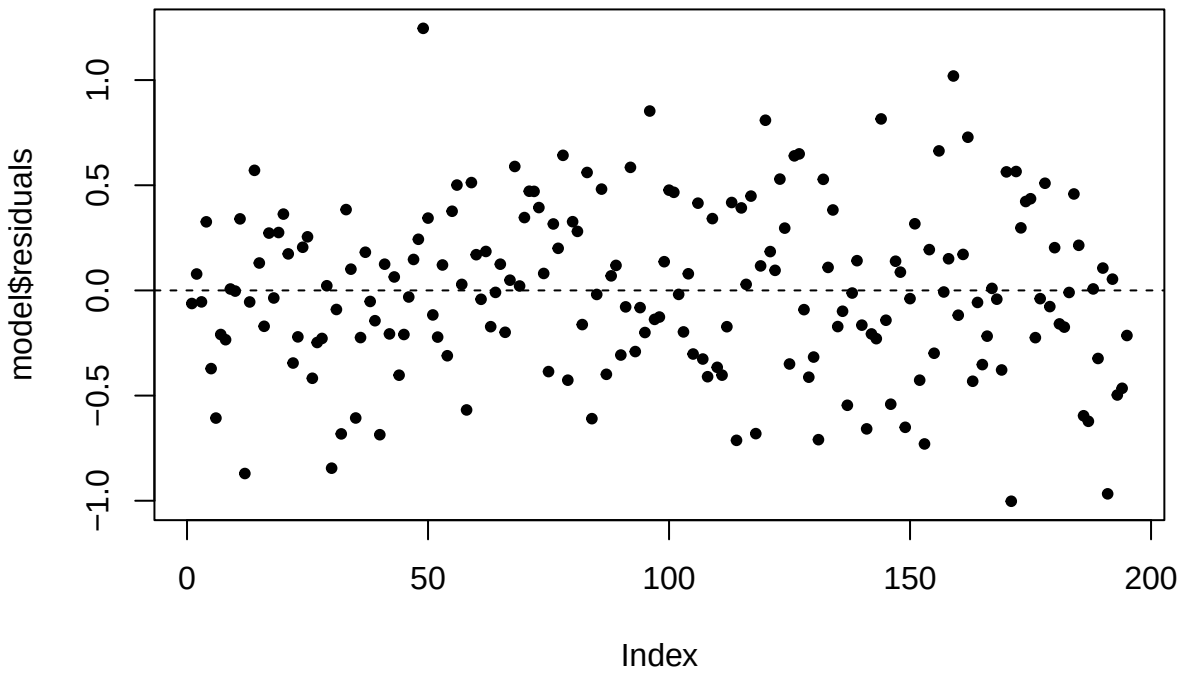
```
##
## Call:
## lm(formula = energy ~ ., data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.00232 -0.22889 -0.00973  0.27796  1.24597
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       9.156e-17  2.920e-02   0.000  1.00000
## 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
```

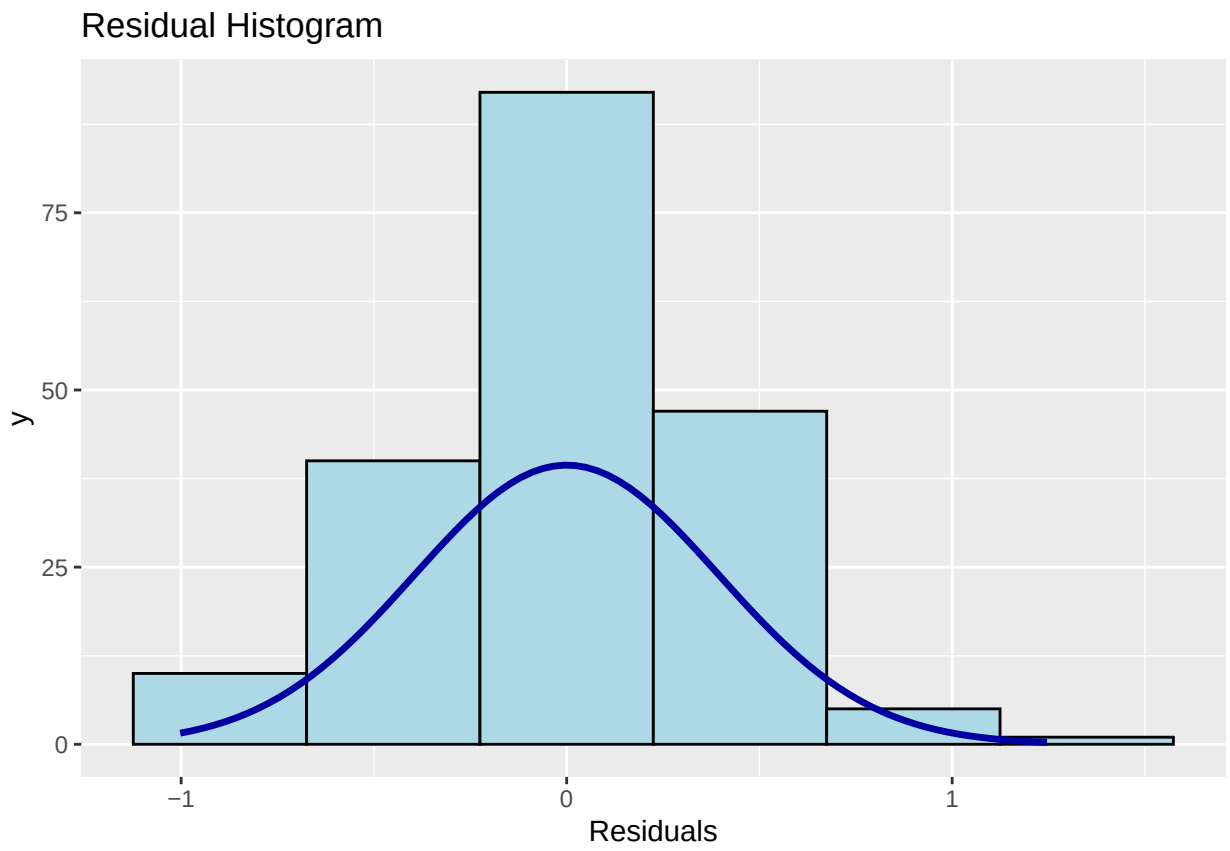
```
ols_step_both_aic(stepwise)
```

```
##
##
##                                  Stepwise Summary
## ------------------------------------------------------------------------------------
## Variable             Method       AIC       RSS      Sum Sq      R-Sq      Adj. R-Sq
## ------------------------------------------------------------------------------------
## loudness             addition    346.927    65.593    128.407    0.66189      0.66014
## acousticness         addition    283.690    46.942    147.058    0.75803      0.75551
## danceability         addition    237.092    36.587    157.413    0.81141      0.80844
## valence              addition    215.654    32.444    161.556    0.83276      0.82924
## instrumentalness     addition    212.234    31.554    162.446    0.83735      0.83305
## mode                 addition    211.005    31.036    162.964    0.84002      0.83491
## key                  addition    210.607    30.657    163.343    0.84198      0.83606
## ------------------------------------------------------------------------------------
```

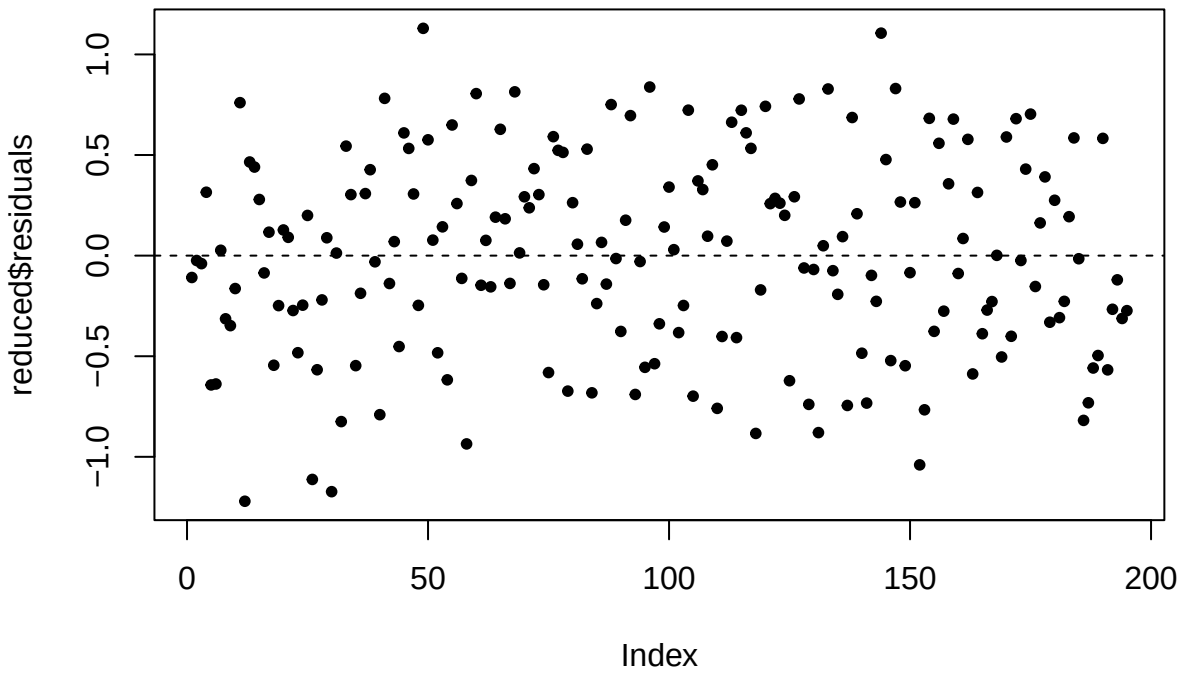
```
###PROBLEM 6
#Full model residuals
plot(model$residuals, pch=20)
abline(h=0,lty=2)
```



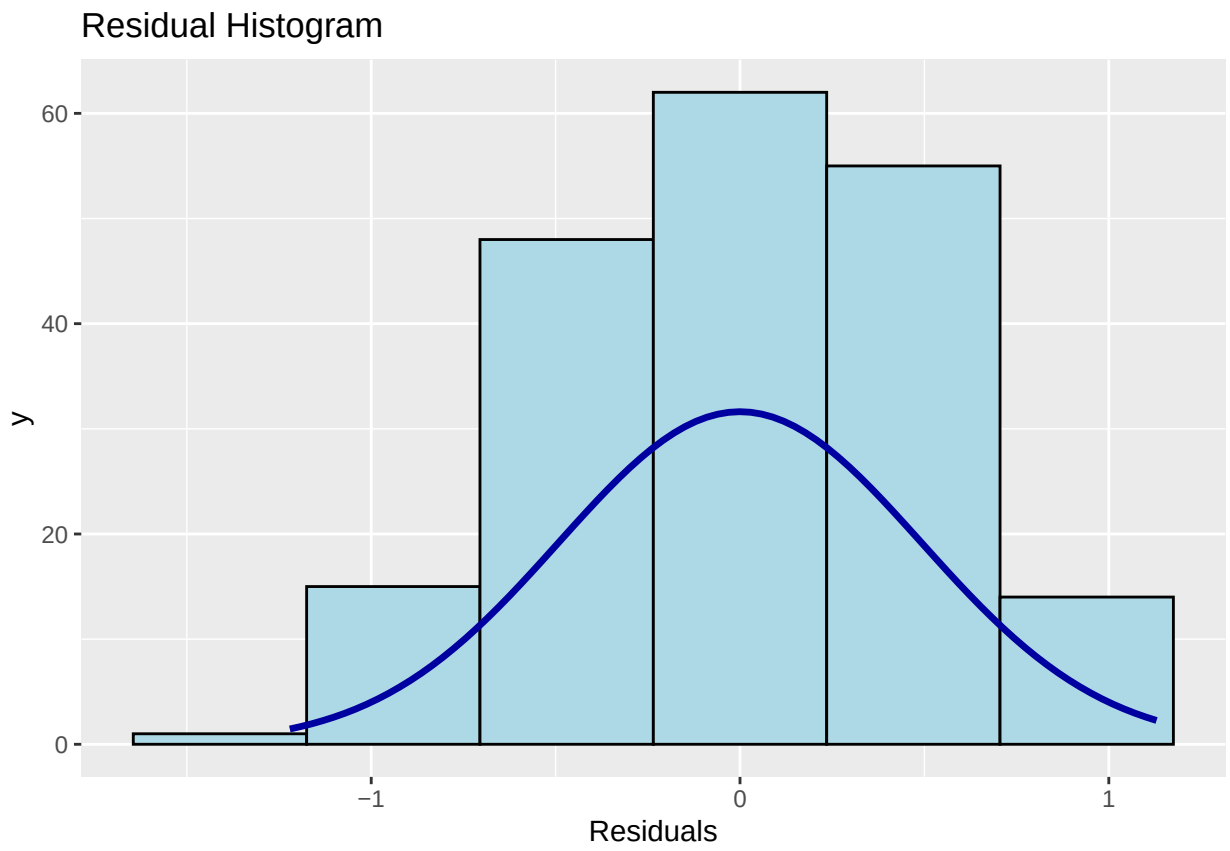

```
ols_plot_resid_hist(model)
```



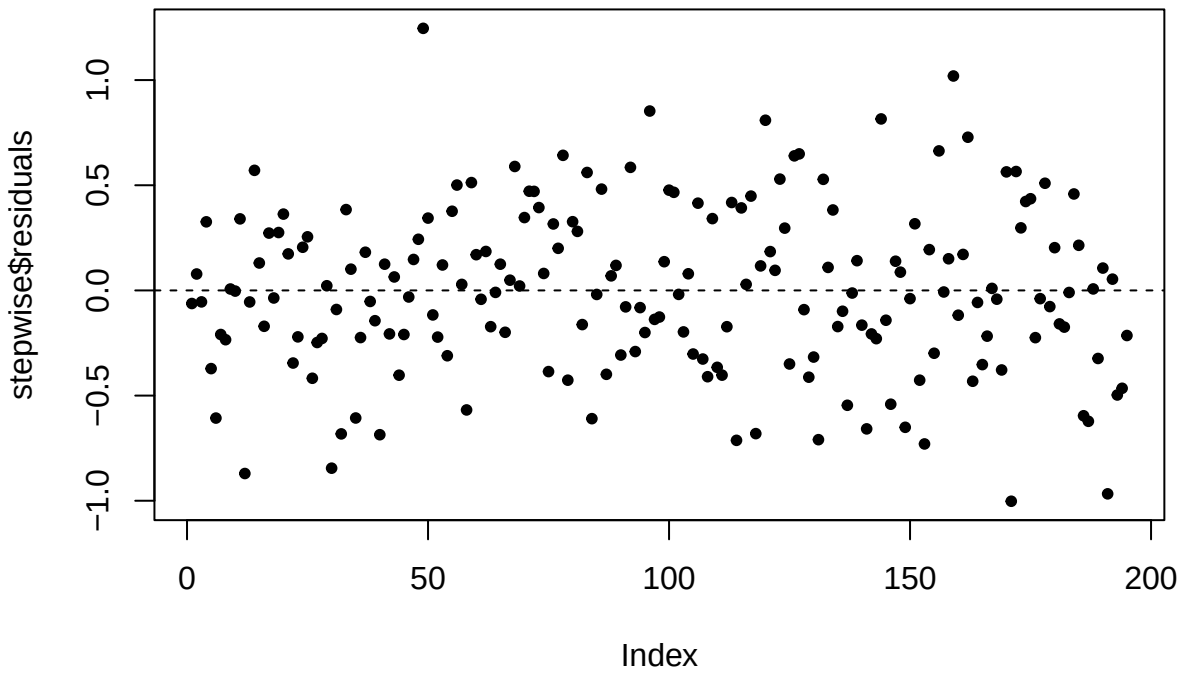

```
#Reduced mdel residuals
plot(reduced$residuals, pch=20)
abline(h=0,lty=2)
```



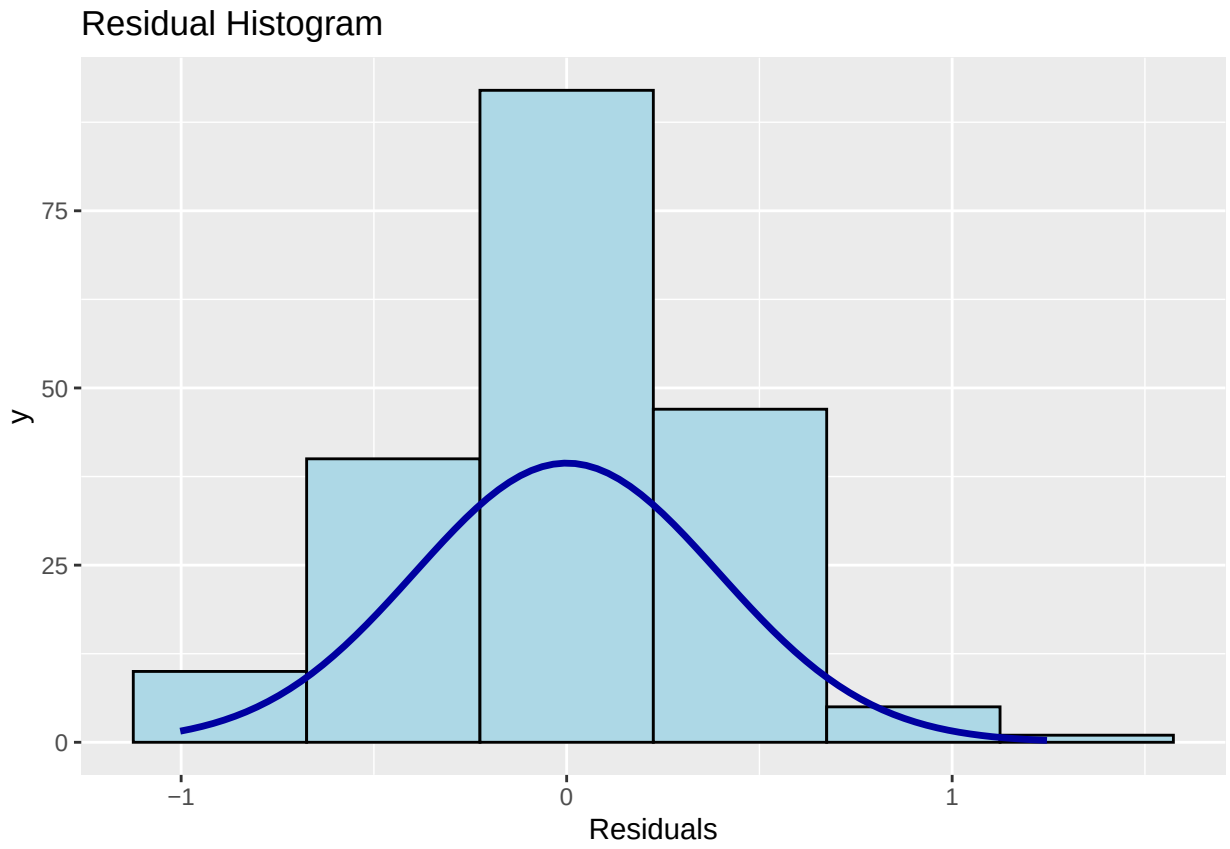

```
ols_plot_resid_hist(reduced)
```




```
#Stepwise model residuals
plot(stepwise$residuals, pch=20)
abline(h=0,lty=2)
```



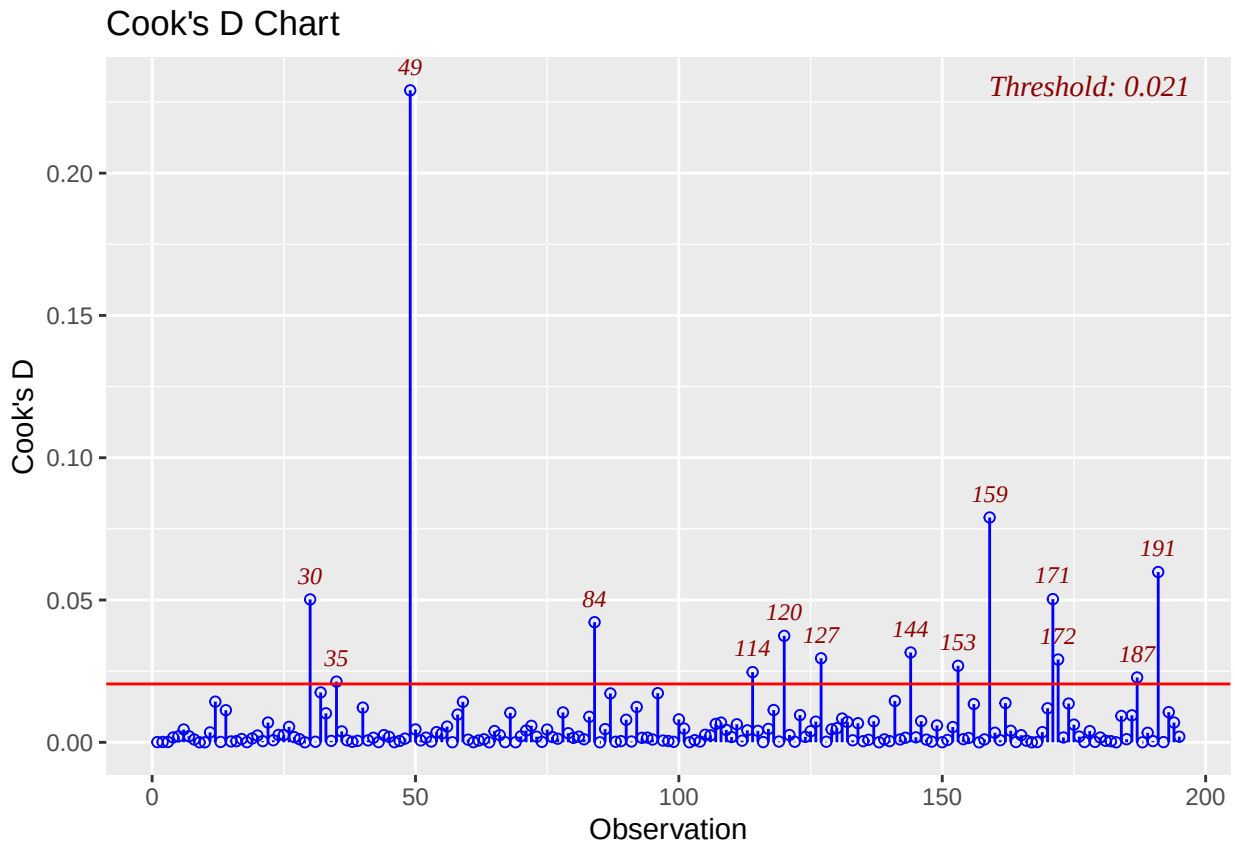

```
ols_plot_resid_hist(stepwise)
```




```
#high density of points close to the origin and a low density of points away from the origin. Hence, th

###PROBLEM 7
ols_vif_tol(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
```

```
cookdgraph<-ols_plot_cooksd_chart(model)
```




```
cooksD<-cooks.distance(model)
n<-nrow(df)
influential<-cooksD[(cooksD>4/n)]
head(influential)
```

```
##         30         35         49         84        114        120
## 0.05022185 0.02132929 0.22910565 0.04220288 0.02466182 0.03740761
```

```
names_of_influential<-names(influential)
outliers<-df[names_of_influential,]
noOutliers<-df %>% anti_join(outliers)
```

```
## Joining, by = c("danceability", "energy", "key", "loudness", "mode",
## "speechiness", "acousticness", "instrumentalness", "liveness", "valence",
## "tempo", "duration_ms", "time_signature")
```

```
newmod<-lm(energy~., data=noOutliers)
summary(newmod)
```

```
##
## Call:
## lm(formula = energy ~ ., data = noOutliers)
##
## 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
```

```
#The fit improves once the outliers are removed
```