# Understanding Pop Music in 2018

*Eva Ma*

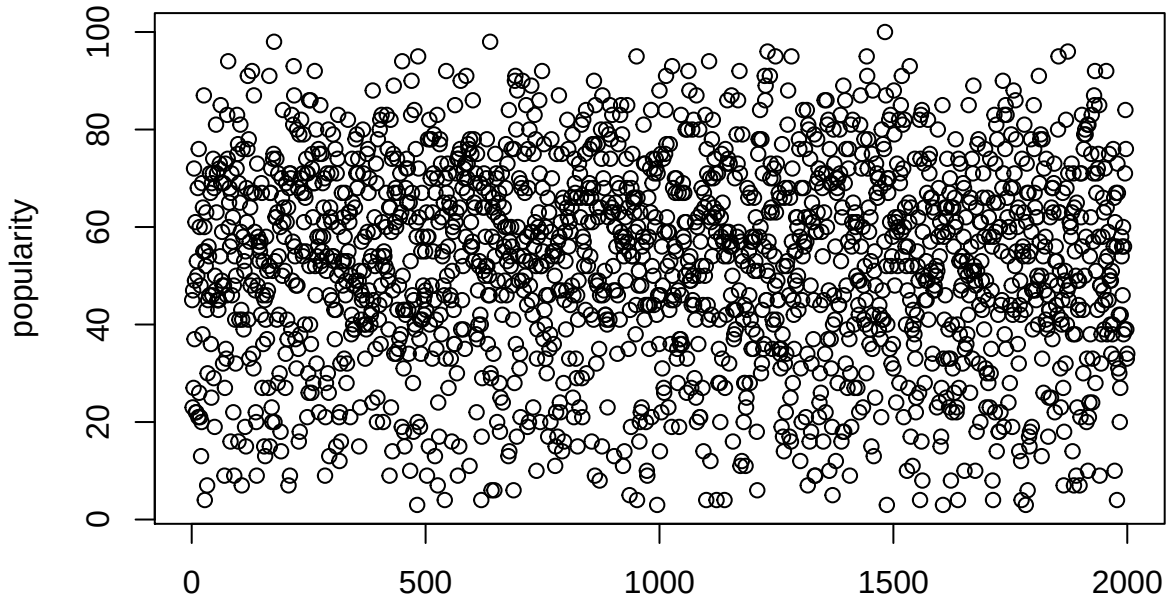
*11/8/2018*

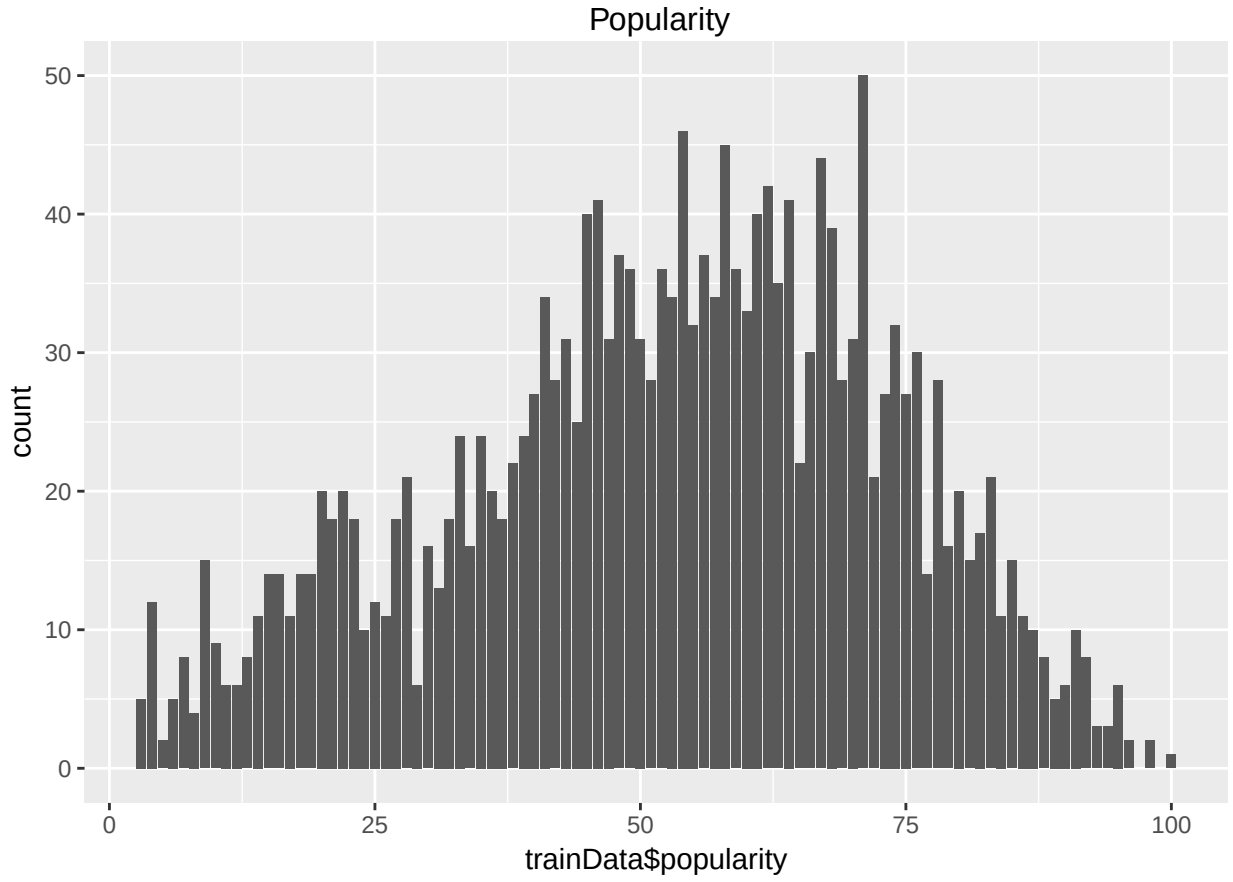
## Data Collection

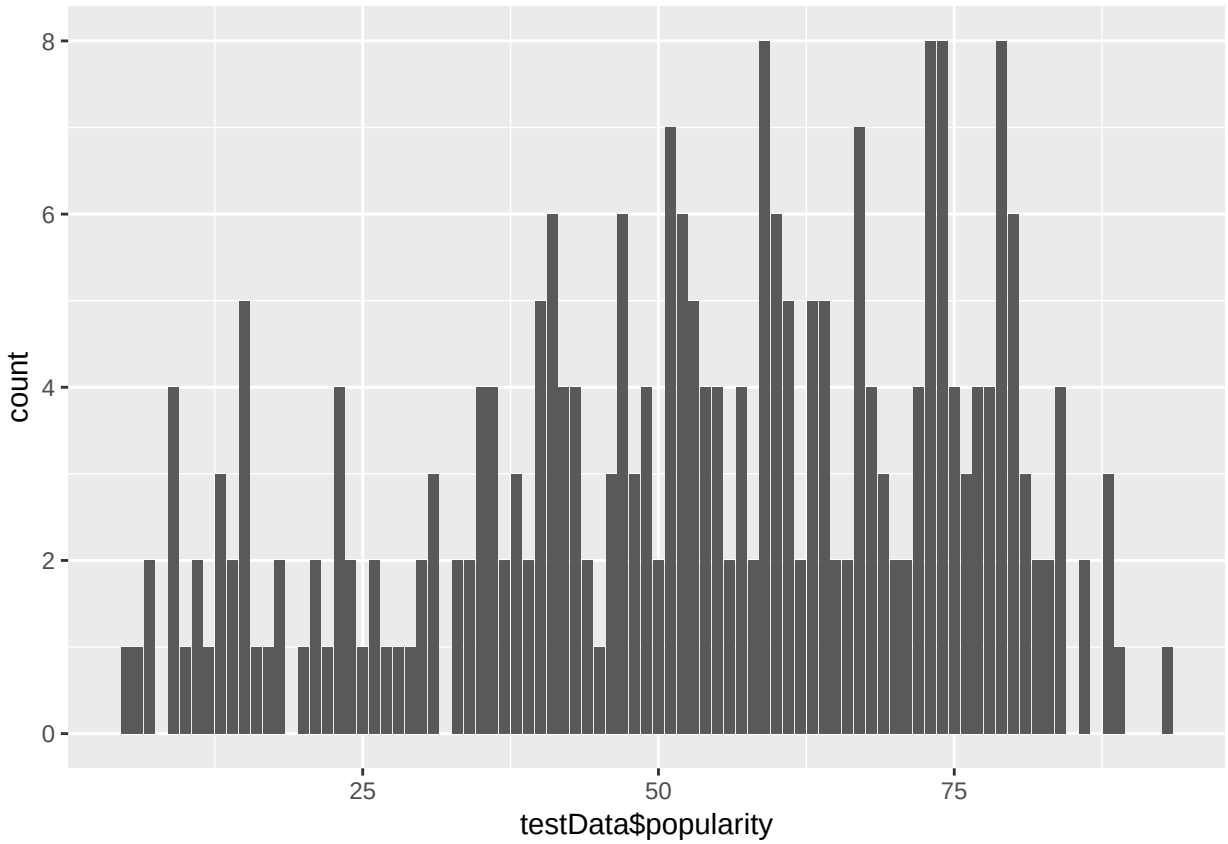
2260 unique songs was obtained from the playlists "Today's Top Hits", 'Pop Rising', 'Hot Rhythmic', 'Mega Hit Mix', 'New Music Friday', 'Hit Rewind', 'Teen Party', 'Guilty Pleasures', 'Women of Pop', 'Soft Pop Hits', 'African Heat', 'Acoustic Hits', 'Fresh & Chill', 'Bedroom Pop', 'Everyday Favorites', 'Global X', 'Contemporary Blend', 'Fangirls Run the World', 'Singled Out', 'Left of Center', 'Afropop', 'Pop Sauce', 'Mellow Pop', 'Wa-oh-wa-oh!', 'Out Now', 'Pop Royalty', 'Workday: Pop', 'Now Hear This', 'Certified Gold', 'Crowd Pleasers', 'LA Pops', 'LADY GAGA / JOANNE', 'Pop Matters', 'Retro Pop', "Tomorrow's Hits", 'All A Cappella', 'Yalla Araby', 'Persian Essentials', 'Radio 1 Playlist (BBC)', 'Wild Cards: Winter Mix', 'Arab X', 'Fresh Finds: Poptronix', 'The GRAMMYs Official Playlist', 'Pop Chile', "Today's Top Egyptian Hits", "Today's Top Maghreb Hits".

```
##   liveness   tempo energy speechiness mode instrumentalness
## 1   0.2760 120.966  0.768      0.0360    1         4.49e-05
## 2   0.1150 103.968  0.767      0.1860    0         0.00e+00
## 3   0.1370 113.981  0.780      0.0623    0         7.07e-06
## 4   0.0398  99.974  0.678      0.0514    0         2.12e-05
## 5   0.0505 139.943  0.545      0.0625    1         2.89e-04
## 6   0.1570 119.953  0.890      0.0405    0         0.00e+00
##                                     name popularity acousticness loudness
## 1                       Dance In The Dark         45     2.99e-05   -6.211
## 2                             Yamen Yasar         23     4.30e-01   -2.073
## 3                               Creatures         47     2.17e-01   -4.313
## 4                           Mafeesh Menha         27     2.50e-01   -7.162
## 5                                   Boys         72     6.46e-02   -5.192
## 6 Why Are We So Broken (feat. blink-182)         37     2.63e-03   -5.016
##   valence danceability
## 1  0.0986        0.645
## 2  0.8900        0.784
## 3  0.5500        0.745
## 4  0.8410        0.820
## 5  0.5250        0.867
## 6  0.3820        0.587
```

```
## [1] 20.64241
```


popularity
0
20
40
60
80
100
0
500
1000
1500
2000


Popularity
count
0
10
20
30
40
50
0
25
50
75
100
trainData$popularity




The data set covers a range of popularities with a standard deviation of 20.64241. The data set is slightly more centered towards the right tail.

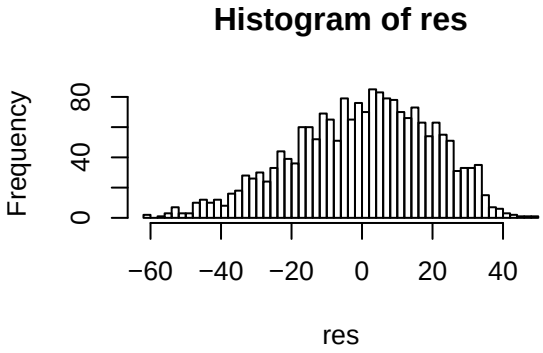
The data set is split into a training set and a testings set.

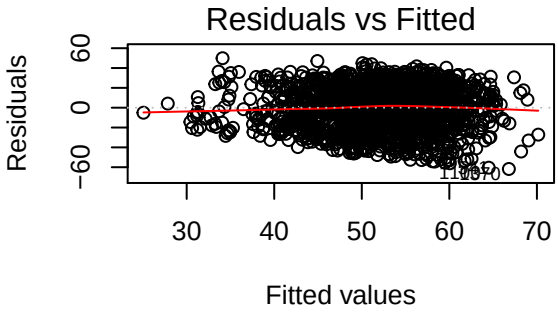
## Naive MLR Model

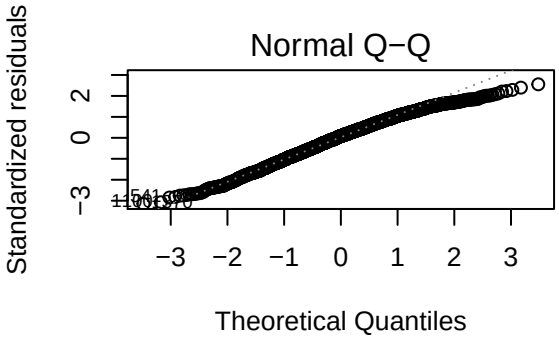
To investigate what factors influence a song's popularity, we first run a naive MLR model, where the regressors are energy, valence, liveness, tempo, speechiness, instrumentalness, acousticness, loudness, and danceability.

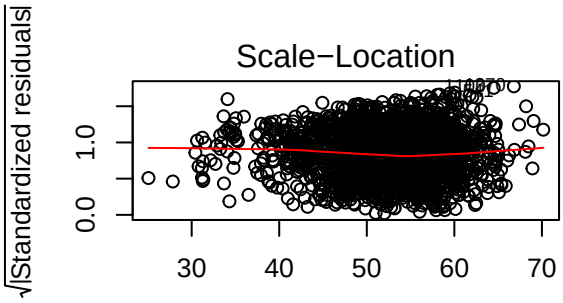
```
##
## Call:
## lm(formula = (popularity) ~ energy + valence + liveness + tempo +
##     speechiness + instrumentalness + acousticness + loudness +
##     danceability)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.795 -13.490   1.701  14.663  49.899
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       7.496e+01  5.121e+00  14.637  < 2e-16 ***
## energy           -1.790e+01  4.207e+00  -4.254 2.19e-05 ***
## valence          -1.646e+01  2.374e+00  -6.933 5.53e-12 ***
## liveness         -8.647e+00  3.312e+00  -2.611  0.00911 **
## tempo            -2.375e-04  1.629e-02  -0.015  0.98837
```

```
## speechiness       2.112e+00  5.426e+00   0.389  0.69718
## instrumentalness -1.631e+00  3.987e+00  -0.409  0.68257
## acousticness     -8.506e+00  2.164e+00  -3.931 8.76e-05 ***
## loudness          1.878e+00  2.516e-01   7.463 1.26e-13 ***
## danceability      2.138e+01  3.539e+00   6.042 1.82e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.71 on 1990 degrees of freedom
## Multiple R-squared:  0.09236,    Adjusted R-squared:  0.08825
## F-statistic:  22.5 on 9 and 1990 DF,  p-value: < 2.2e-16
```

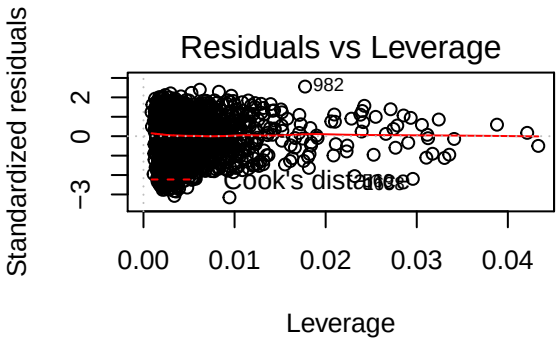


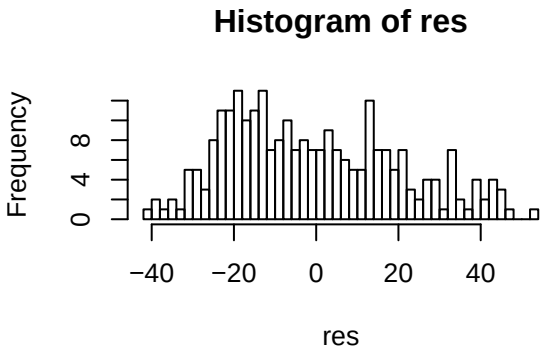




```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  resid(naive_mlr)/sigma(naive_mlr)
## D = 0.03722, p-value = 0.007843
## alternative hypothesis: two-sided
```



The estimates indicate that there is a negative relationship between popularity and energy, valence, liveness, tempo, instrumentalness, acousticness. Most of the results make sense. For example, generally, people prefer to play music recorded in studios, so songs with low liveness are more popular. In 2018, popular music is increasingly electronic, so it makes sense that acousticness have a negative coefficient. Also, catchy music often have lyrics, which results in a low instrumentalness.

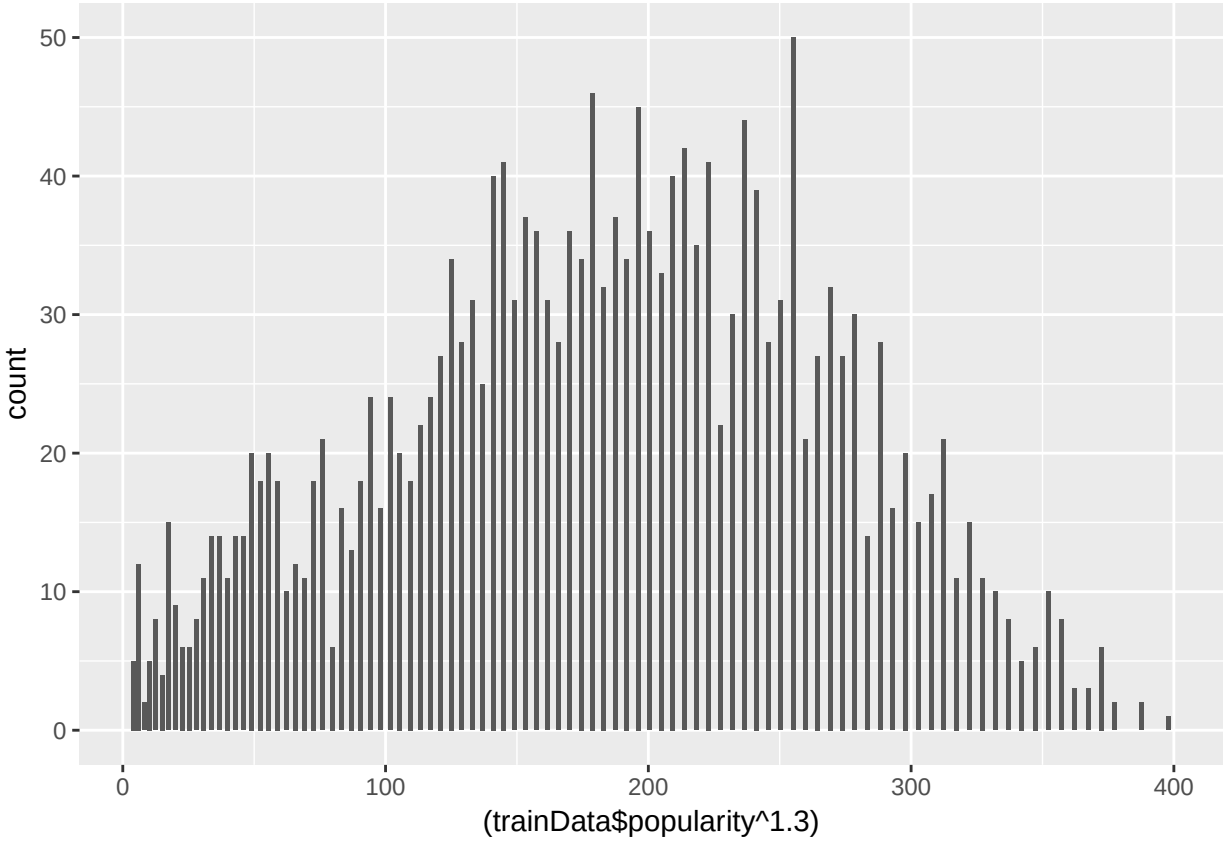
However, it is unclear why energy, valence, and tempo also have negative coefficients. It is assumed that generally pop music is energetic and happy.

In terms of standard error of each estimates, it is very small and therefore desired. It follows that most estimates are highly significant, except for tempo, speechiness, and instrumentalness.

The $R^2$ value is relatively small, which means that only a small portion of the variability in observations is explained by this model.

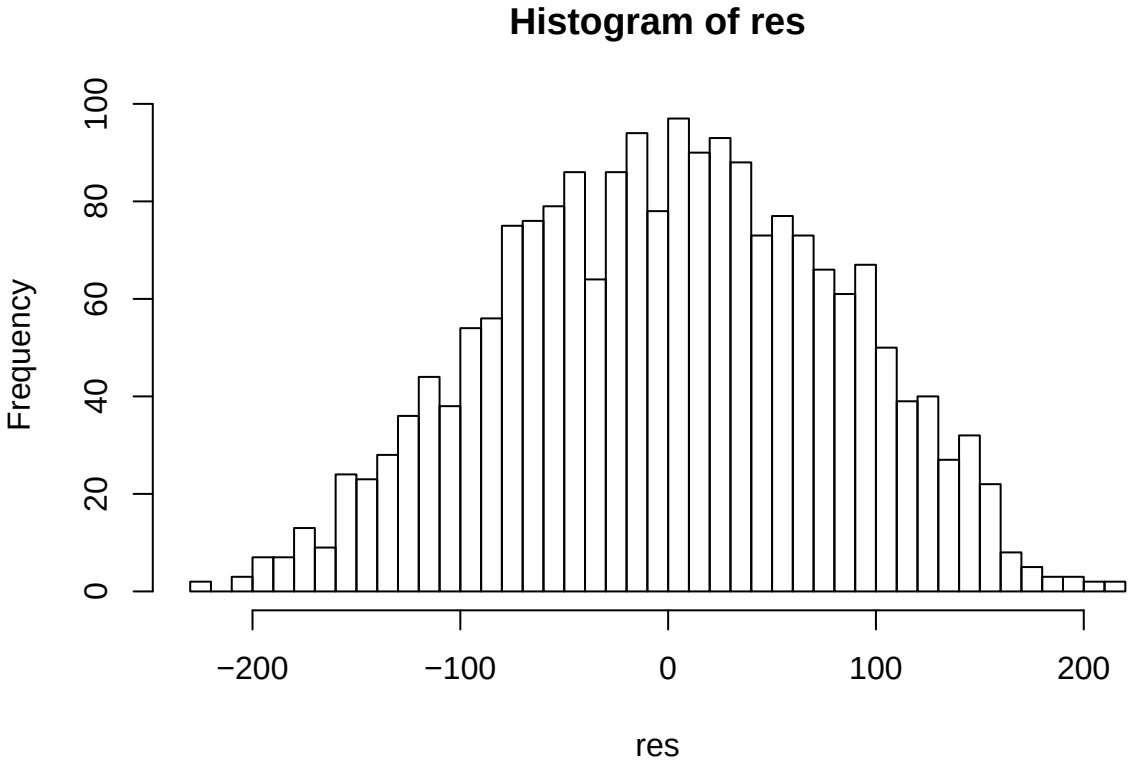
Our F statistics, however, is high significant. It shows that our model is highly significant.

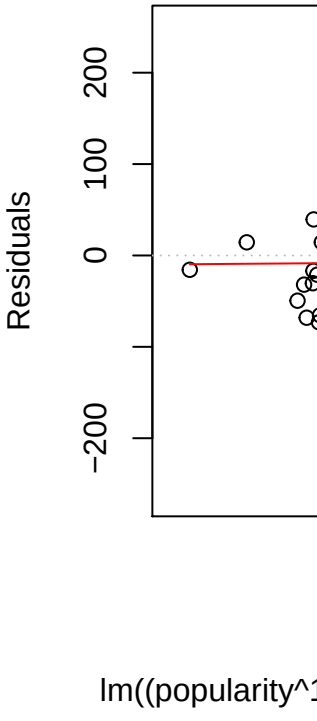
Next, we will focus on improving the $R^2$ value.

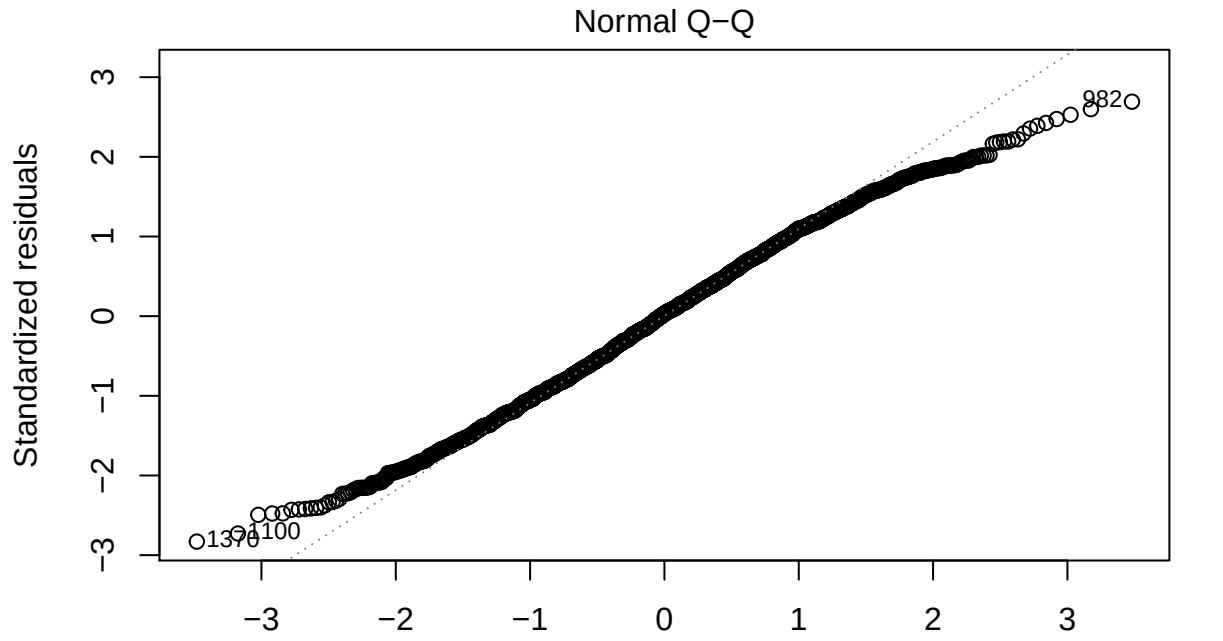


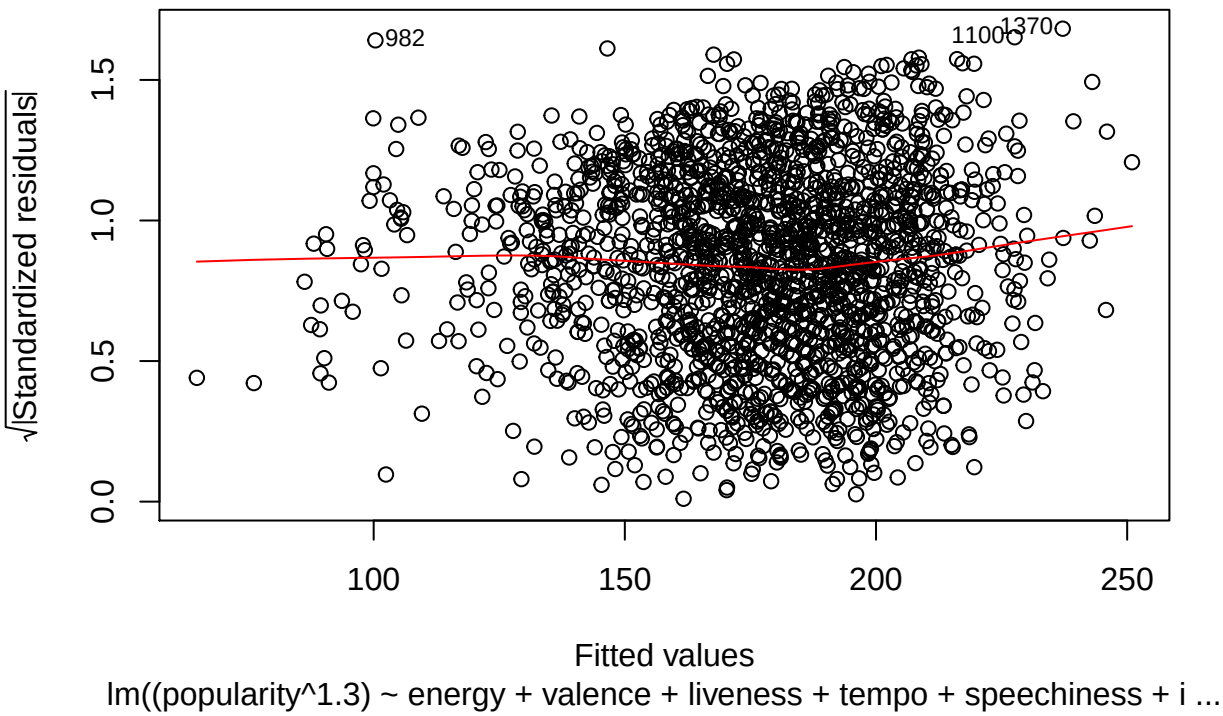

```
##
## Call:
## lm(formula = (popularity^1.3) ~ energy + valence + liveness +
##     tempo + speechiness + instrumentalness + acousticness + loudness +
##     danceability)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -229.070  -59.354    2.293   60.016  217.020
##
## Coefficients:
```

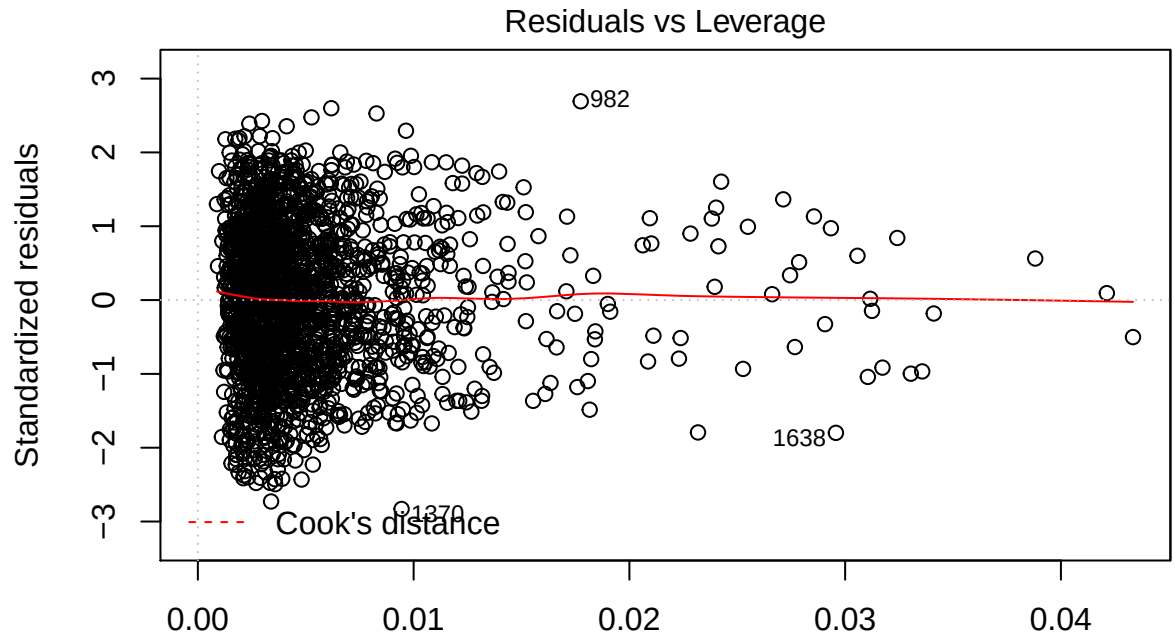
```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      269.798236  21.131204  12.768  < 2e-16 ***
## energy           -72.792865  17.360787  -4.193 2.87e-05 ***
## valence          -69.089088   9.797541  -7.052 2.43e-12 ***
## liveness         -33.663446  13.667177  -2.463 0.013859 *
## tempo              0.006854   0.067225   0.102 0.918797
## speechiness        6.786891  22.389726   0.303 0.761826
## instrumentalness -10.368091  16.449525  -0.630 0.528572
## acousticness     -34.281698   8.928764  -3.839 0.000127 ***
## loudness           7.823709   1.038207   7.536 7.33e-14 ***
## danceability      88.193384  14.603822   6.039 1.84e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 81.33 on 1990 degrees of freedom
## Multiple R-squared:  0.093,  Adjusted R-squared:  0.0889
## F-statistic: 22.67 on 9 and 1990 DF,  p-value: < 2.2e-16
```

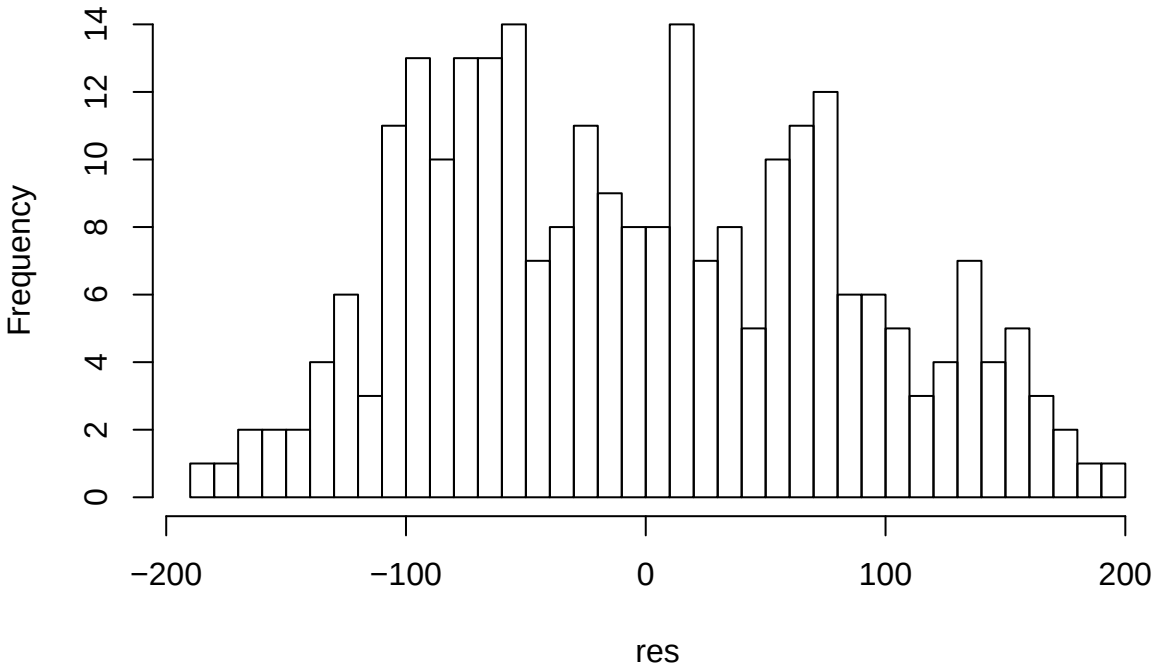



Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm((popularity^1.3) ~ energy + valence + liveness + tempo + speechiness + i ...
982
1100
1370

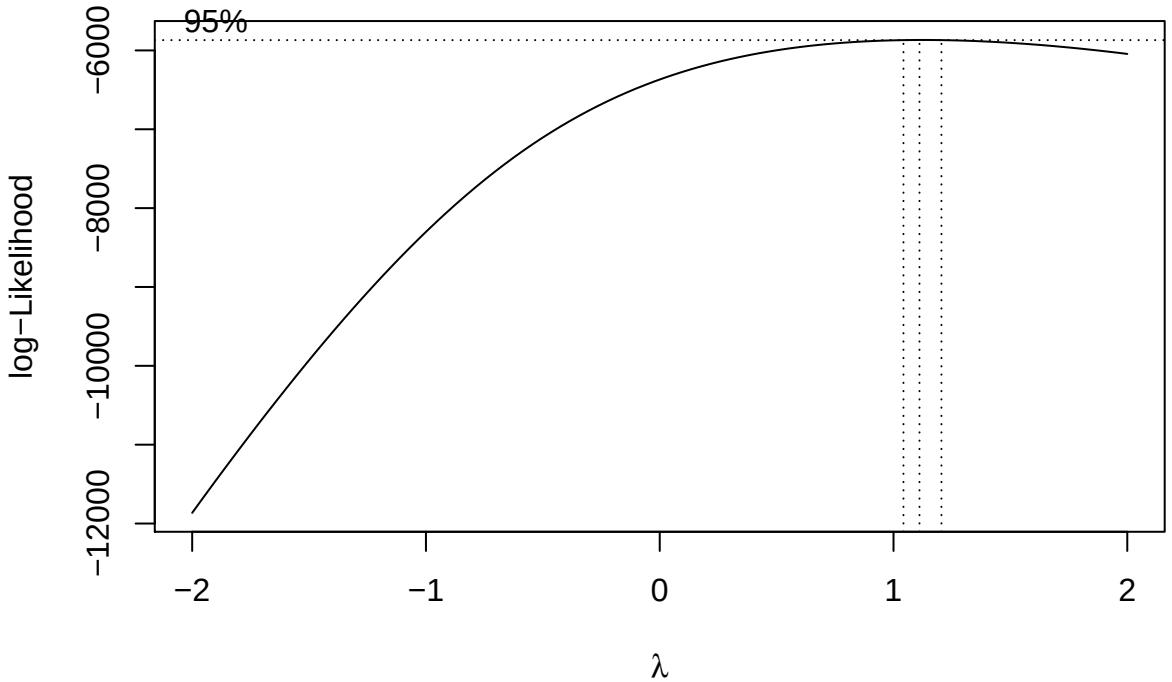

Scale−Location
√|Standardized residuals|
Fitted values
lm((popularity^1.3) ~ energy + valence + liveness + tempo + speechiness + i ...
982
1100
1370

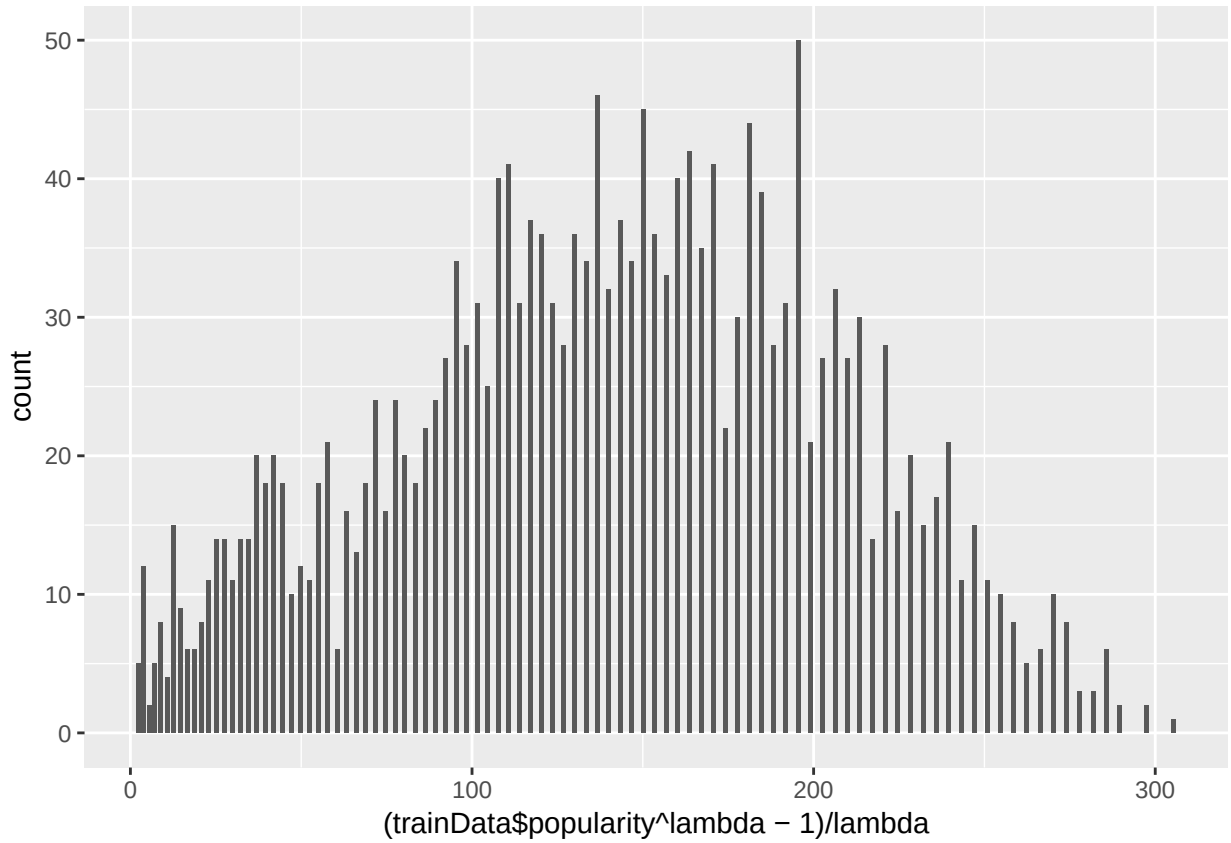



```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  resid(trans_mlr)/sigma(trans_mlr)
## D = 0.024286, p-value = 0.1888
## alternative hypothesis: two-sided
```

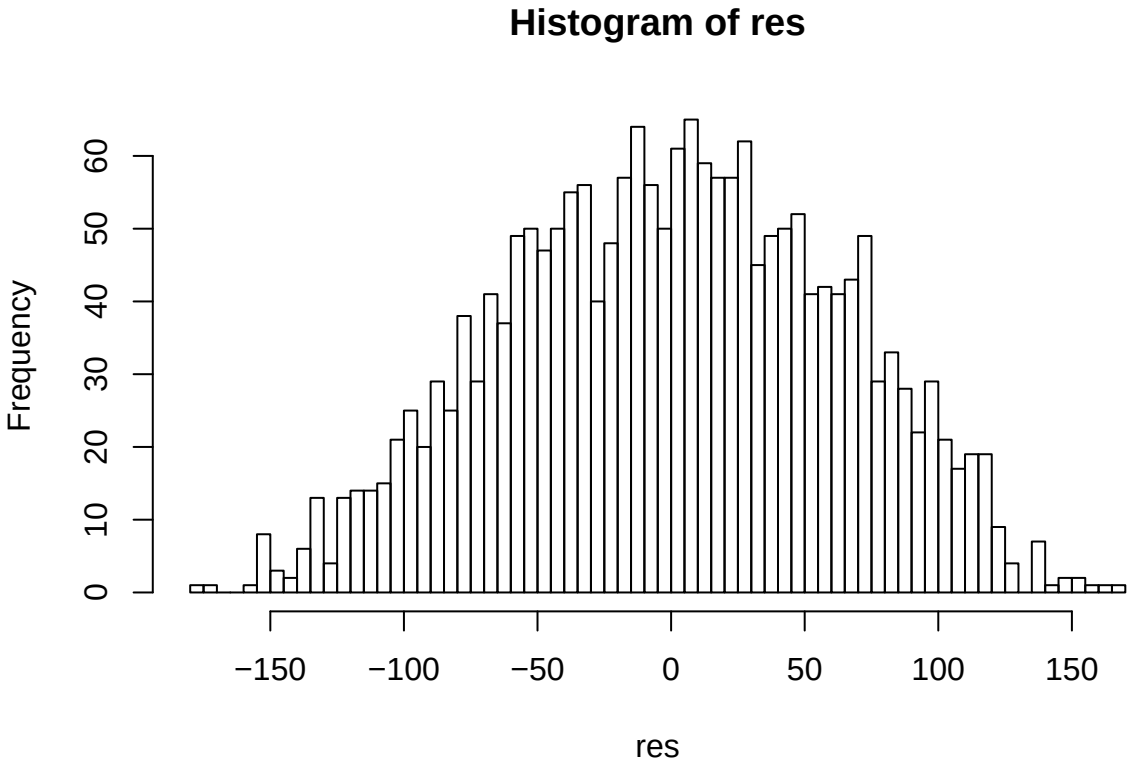
```
## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :
##  extra argument 'optimize' will be disregarded
```

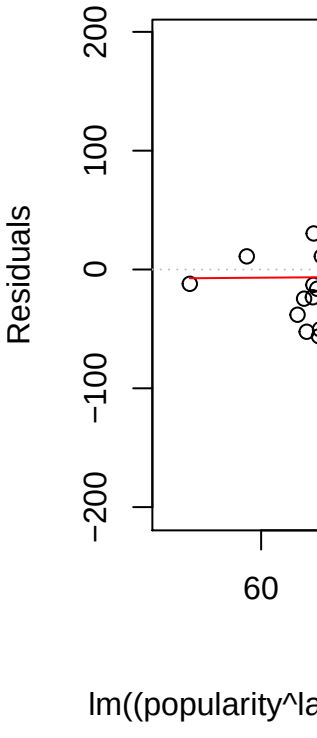


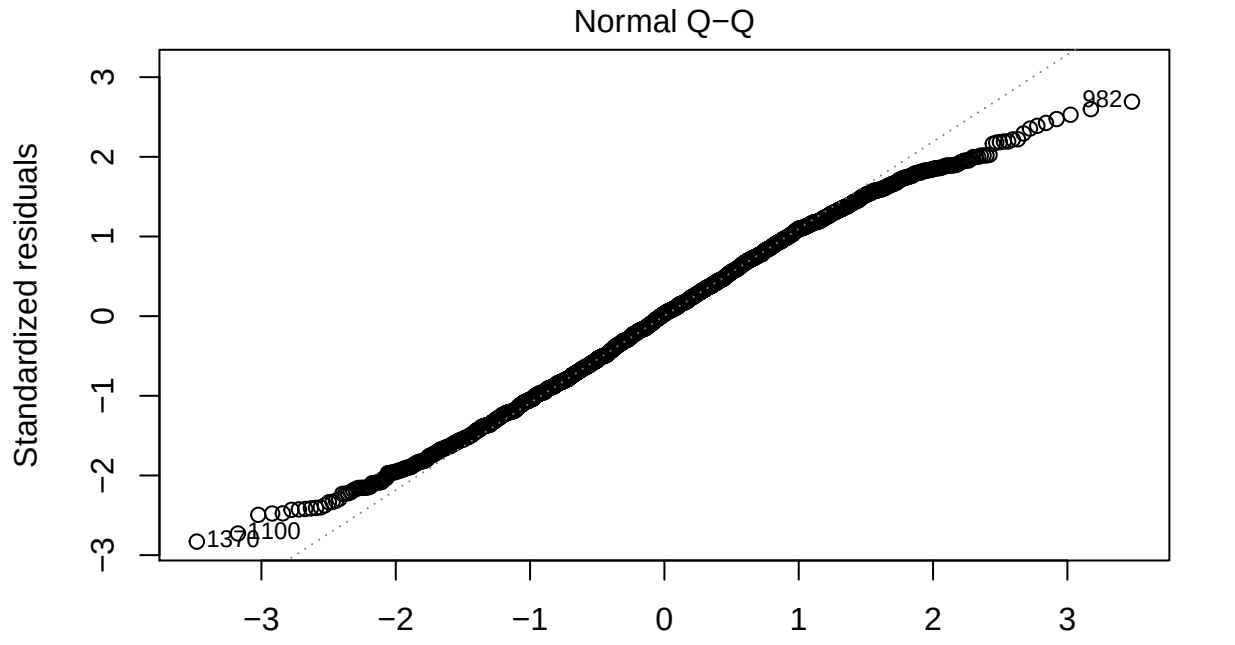



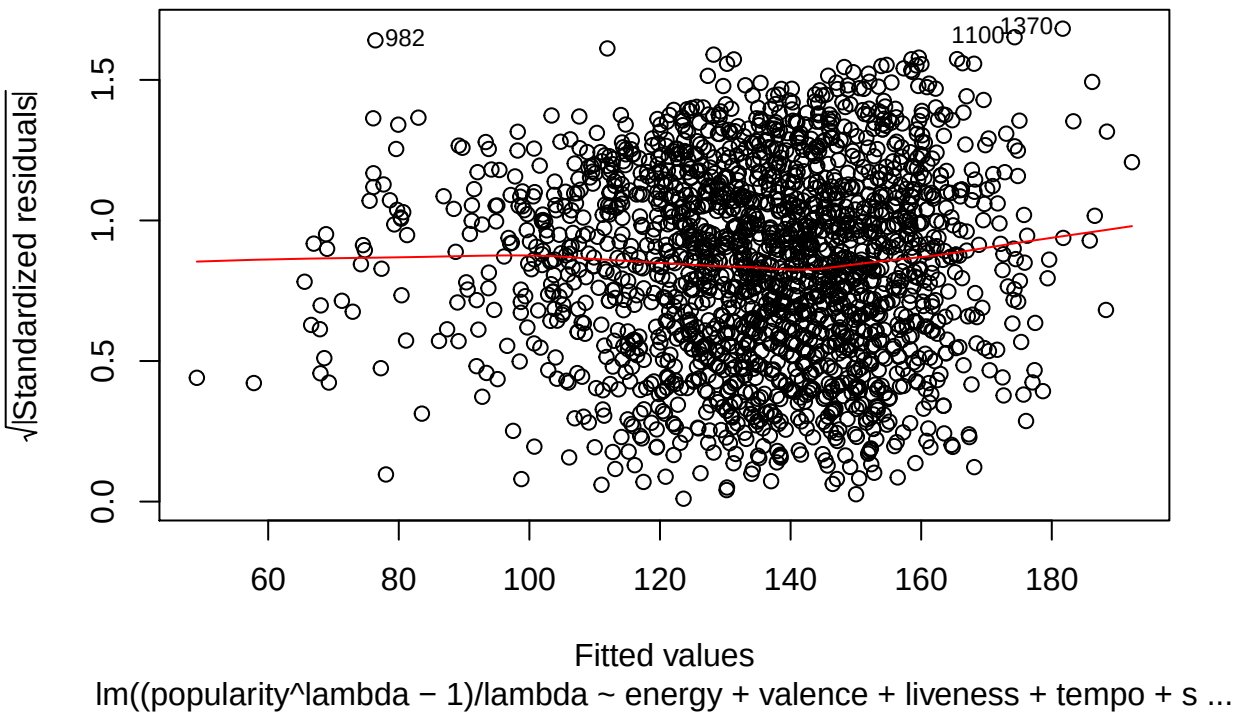

```
##
## Call:
## lm(formula = (popularity^lambda - 1)/lambda ~ energy + valence +
```

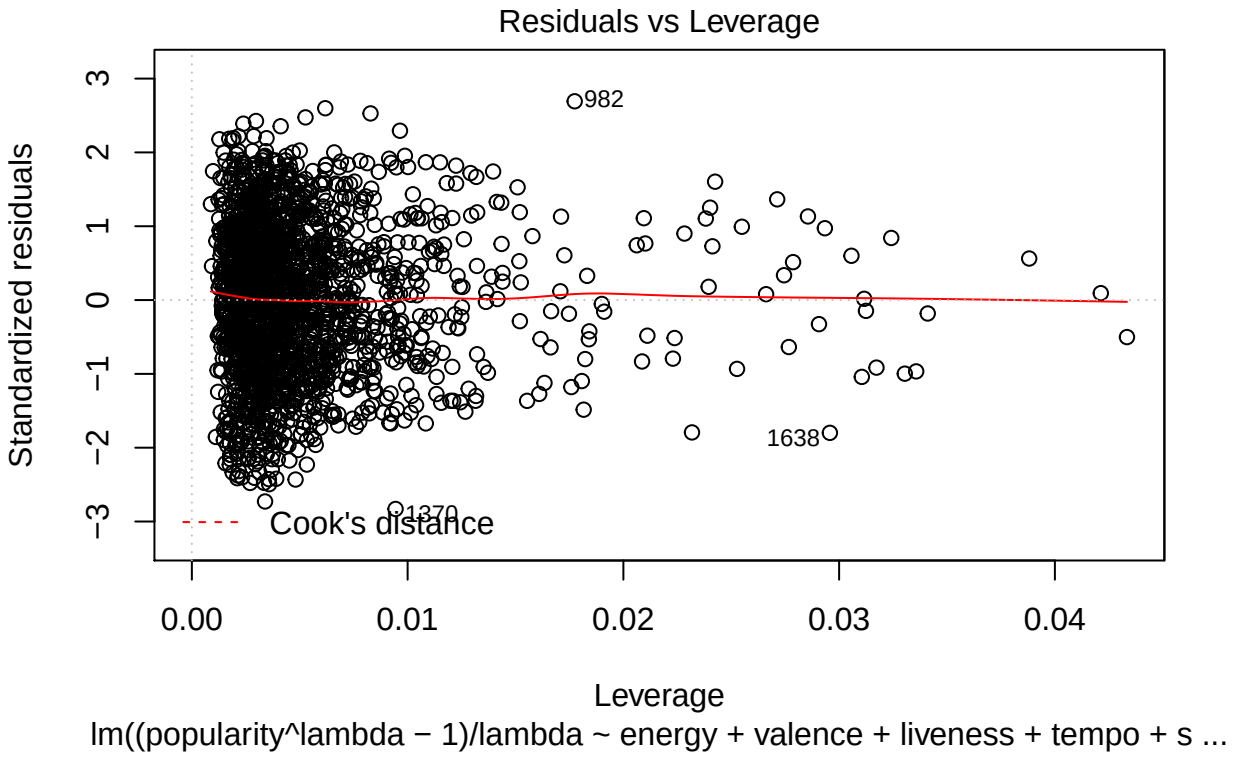
```
##     liveness + tempo + speechiness + instrumentalness + acousticness +
##     loudness + danceability)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -176.208  -45.657    1.764   46.166  166.939
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      206.767874  16.254772  12.720  < 2e-16 ***
## energy           -55.994512  13.354452  -4.193 2.87e-05 ***
## valence          -53.145452   7.536570  -7.052 2.43e-12 ***
## liveness         -25.894959  10.513213  -2.463 0.013859 *
## tempo              0.005273   0.051711   0.102 0.918797
## speechiness        5.220685  17.222866   0.303 0.761826
## instrumentalness  -7.975455  12.653481  -0.630 0.528572
## acousticness     -26.370537   6.868280  -3.839 0.000127 ***
## loudness           6.018238   0.798621   7.536 7.33e-14 ***
## danceability      67.841064  11.233709   6.039 1.84e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62.56 on 1990 degrees of freedom
## Multiple R-squared:  0.093,  Adjusted R-squared:  0.0889
## F-statistic: 22.67 on 9 and 1990 DF,  p-value: < 2.2e-16
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm((popularity^lambda − 1)/lambda ~ energy + valence + liveness + tempo + s ...
982
1100
1370


Scale−Location
√|Standardized residuals|
Fitted values
lm((popularity^lambda − 1)/lambda ~ energy + valence + liveness + tempo + s ...
982
1100
1370




```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  resid(bc_trans_mlr)/sigma(bc_trans_mlr)
## D = 0.024286, p-value = 0.1888
## alternative hypothesis: two-sided
```

## A Closer Look at the Dataset - Visualization

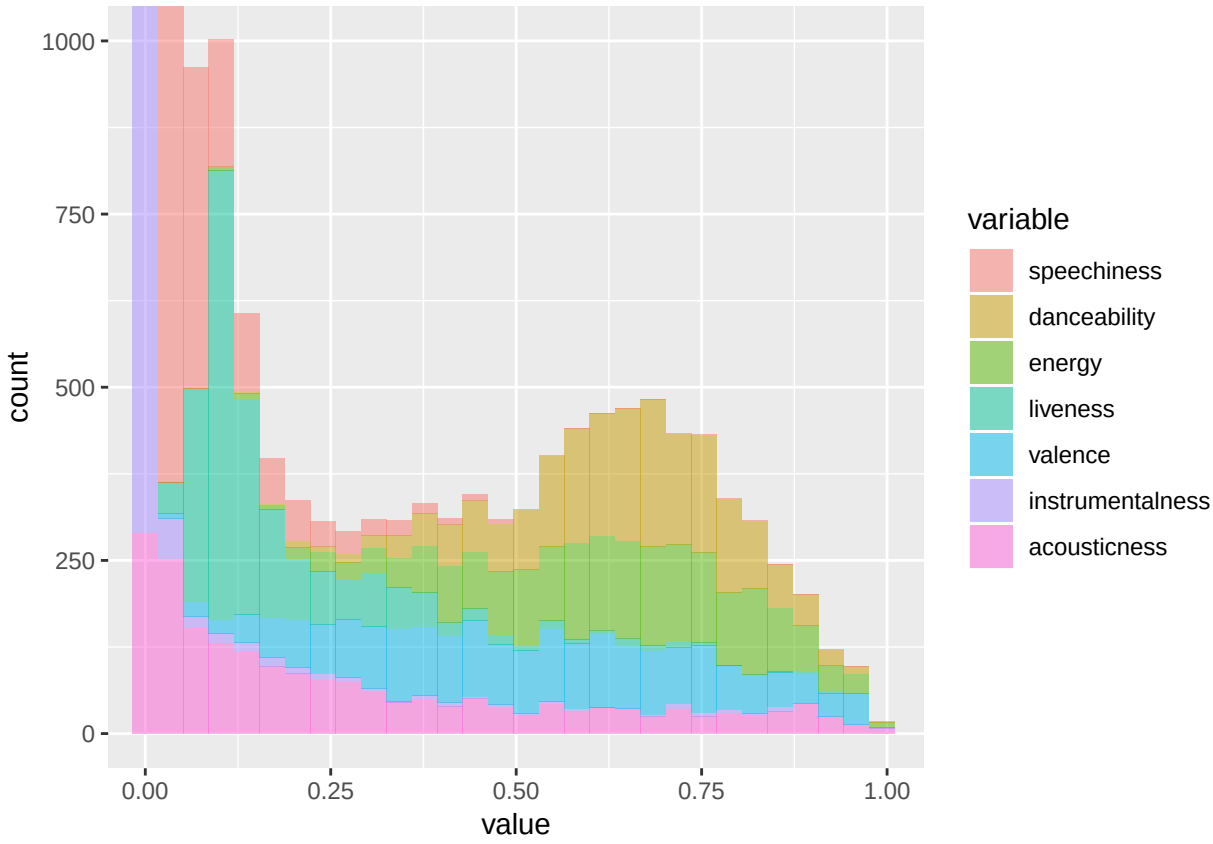
Let's first visualization the distribution of the regressor speechiness, tempo and instrumentalness.

Below is a histogram for the distribution of speechiness, danceability, energy, liveness, valence, and instrumentalness.

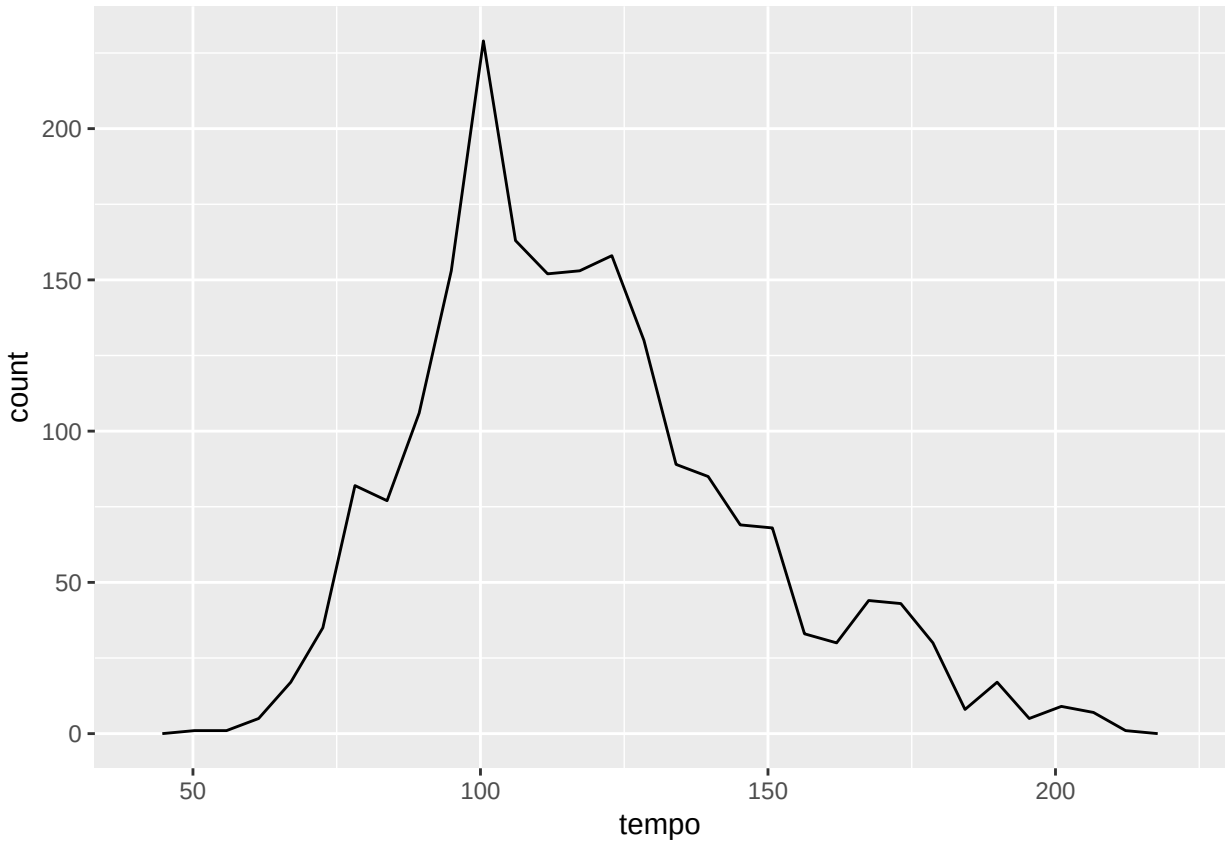
We can see that speechiness, instrumentalness, acousticness, and liveness are heavily centered around 0.

```
## No id variables; using all as measure variables
```

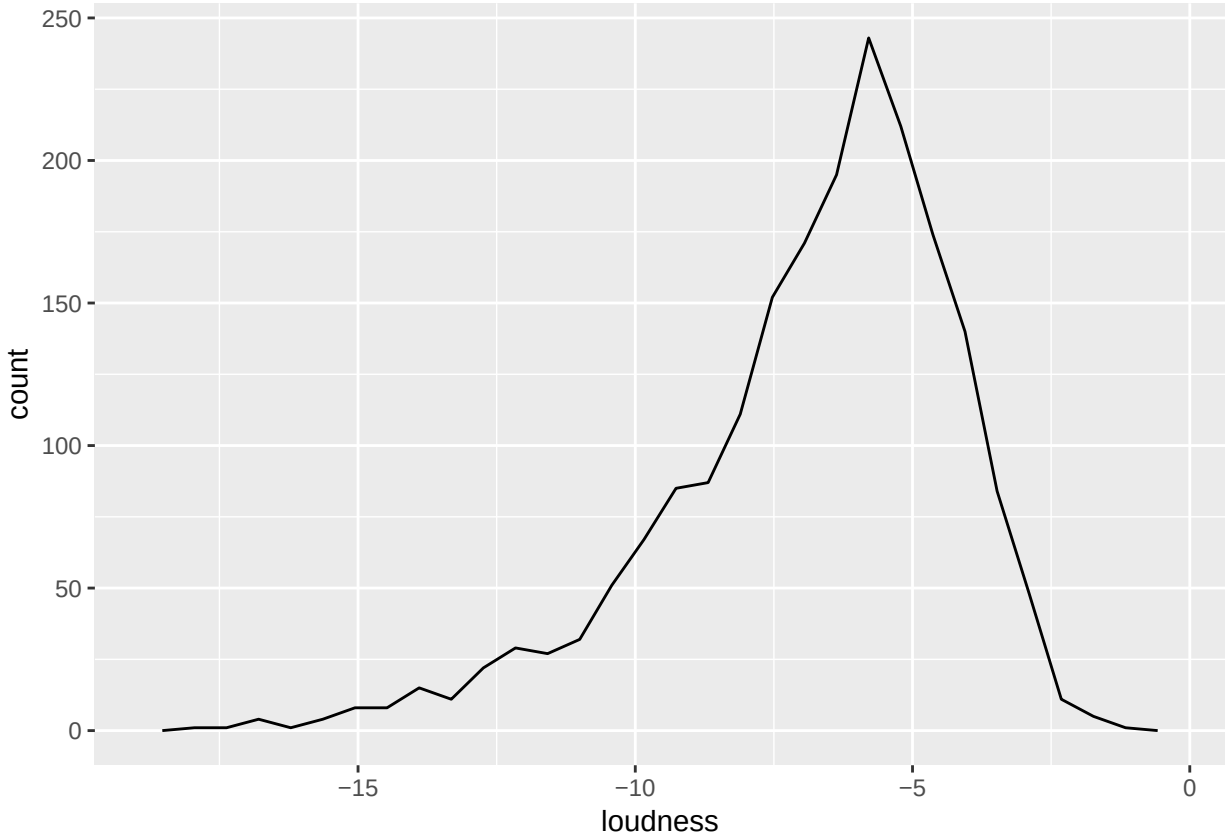
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

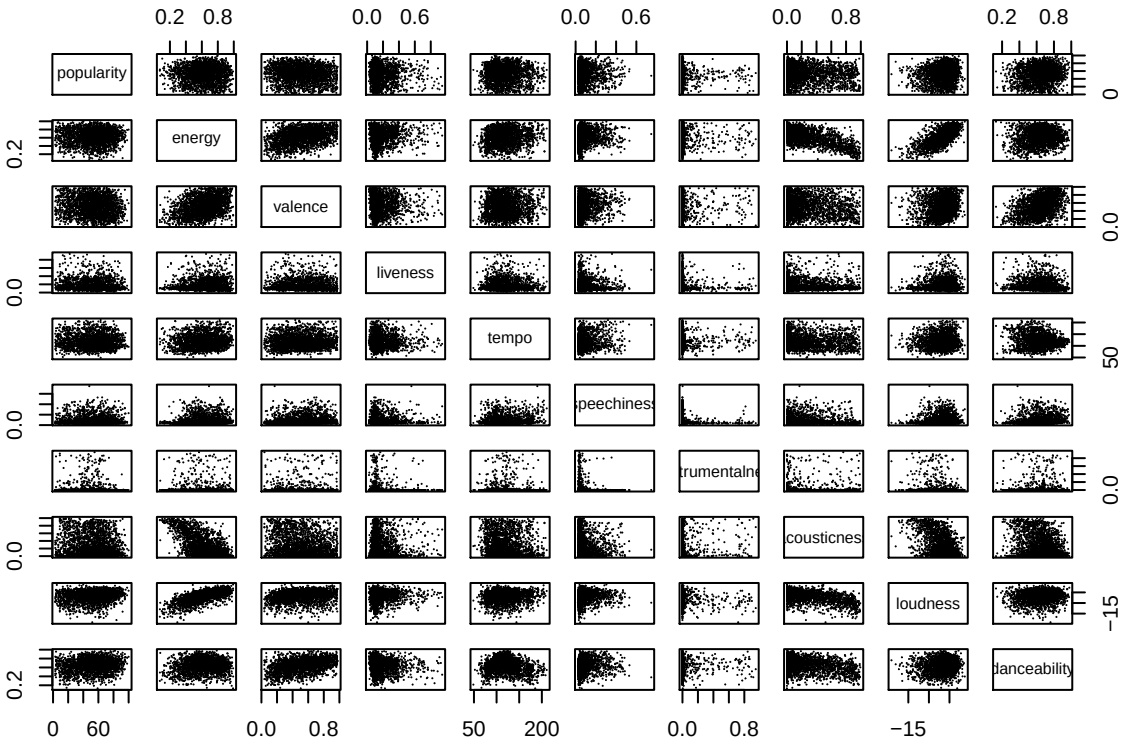



```
## Saving 6.5 x 4.5 in image
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




The pair plot comfirmed our observation that some regressor does not follow a normal distribution. Also, there little correlation between the regressors, except between energy and loudness.

## Are the variables correlated?

The correlation matrix and the VIF are attached below.

```
##                       energy    valence    liveness       tempo
## energy            1.00000000  0.3727164  0.11254158  0.11004218
## valence           0.37271641  1.0000000  0.02685500  0.01175350
## liveness          0.11254158  0.0268550  1.00000000  0.01089756
## tempo             0.11004218  0.0117535  0.01089756  1.00000000
## speechiness       0.11249669  0.1295321  0.02714273  0.11784701
## instrumentalness -0.04463278 -0.0103007 -0.00895431  0.02546522
## acousticness     -0.63203800 -0.1625865 -0.03123947 -0.07336780
## loudness          0.70101084  0.2111887  0.04728143  0.05572854
## danceability      0.11929062  0.4049074 -0.08808117 -0.16440472
##                  speechiness instrumentalness acousticness    loudness
## energy            0.11249669      -0.04463278  -0.63203800  0.70101084
## valence           0.12953212      -0.01030070  -0.16258652  0.21118869
## liveness          0.02714273      -0.00895431  -0.03123947  0.04728143
## tempo             0.11784701       0.02546522  -0.07336780  0.05572854
## speechiness       1.00000000      -0.06465612  -0.07627825  0.04038528
## instrumentalness -0.06465612       1.00000000   0.03509993 -0.19519831
## acousticness     -0.07627825       0.03509993   1.00000000 -0.50335926
## loudness          0.04038528      -0.19519831  -0.50335926  1.00000000
## danceability      0.10034899       0.00391788  -0.22994586  0.13646074
##                  danceability
## energy             0.11929062
## valence            0.40490736
## liveness          -0.08808117
## tempo             -0.16440472
## speechiness        0.10034899
## instrumentalness   0.00391788
## acousticness      -0.22994586
## loudness           0.13646074
## danceability       1.00000000
```

```
##           energy          valence         liveness            tempo
##         2.995352         1.450687         1.027293         1.069369
##      speechiness instrumentalness     acousticness         loudness
##         1.051210         1.067450         1.823955         2.143040
##     danceability
##         1.367619
```

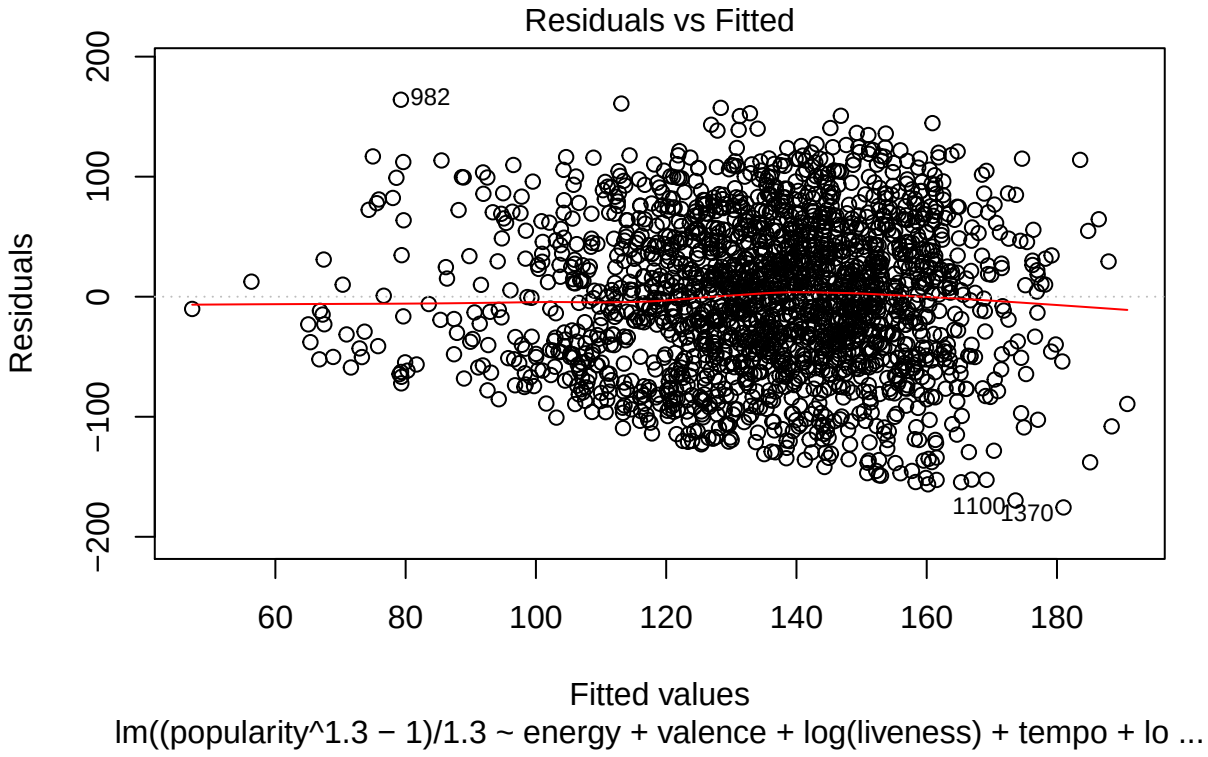
The variance inflation factors are small - most of them below 2 except for energy ( 2.995352). It suggests that the model does not suffer from multicollinearity.

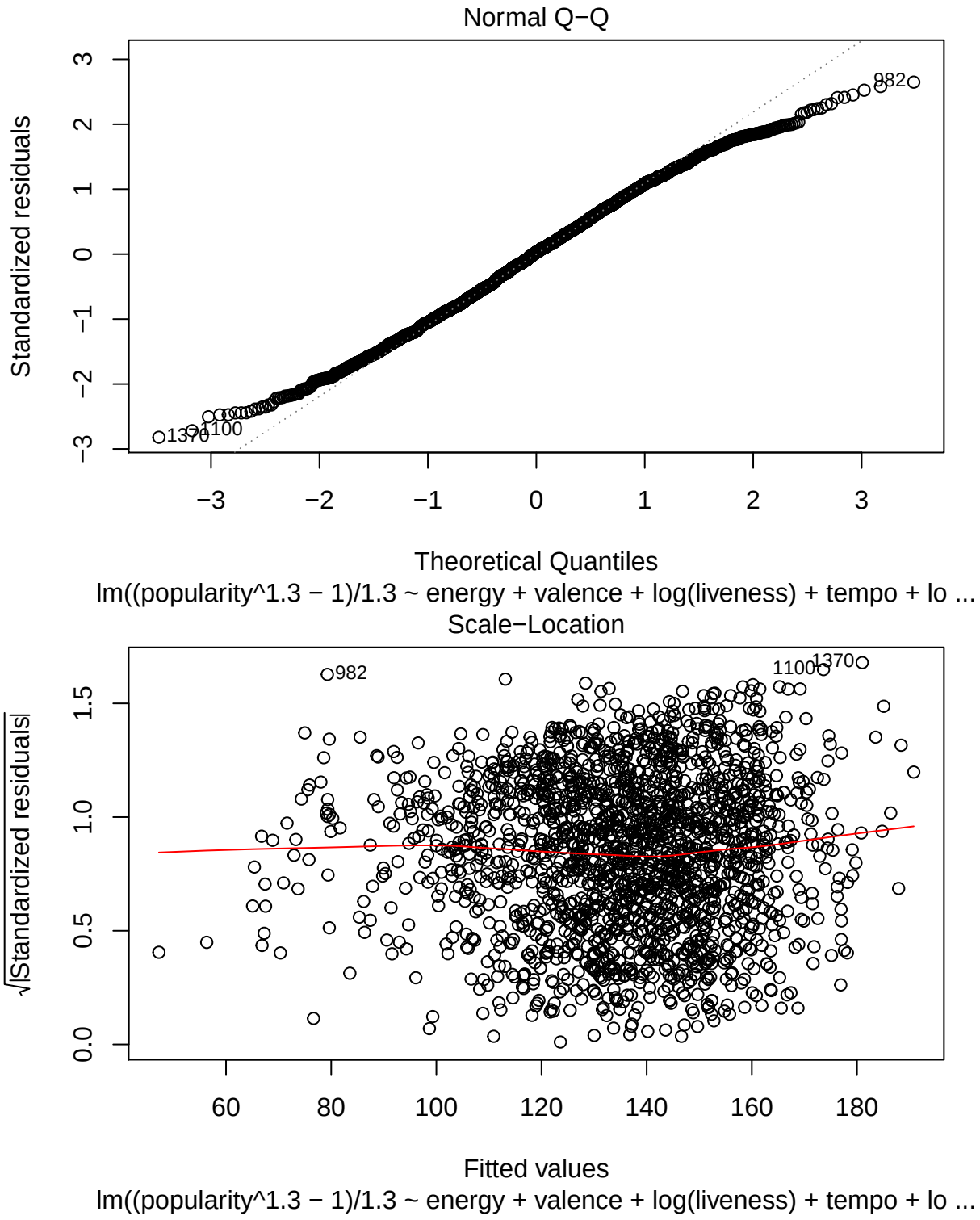
## Log Transformation

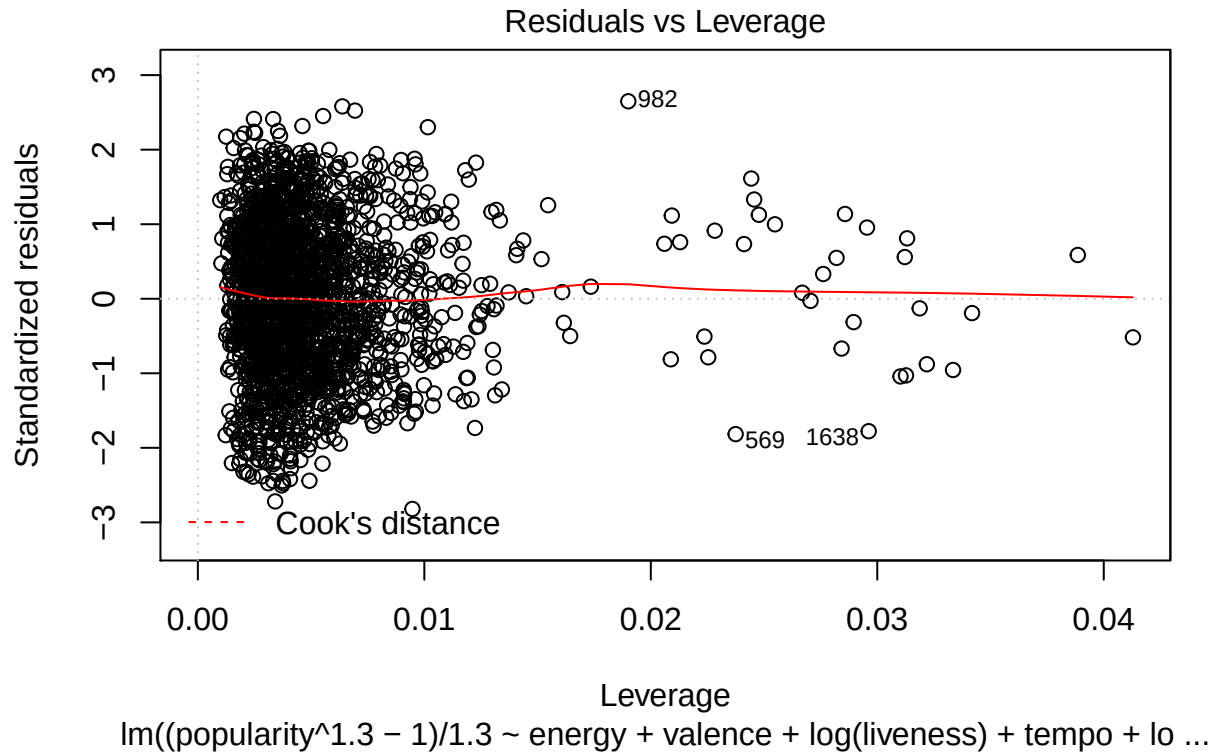
From the plots above, we can see a few regressors are not normally distributed like liveness and speechiness. A log transformation is applied to both variables.

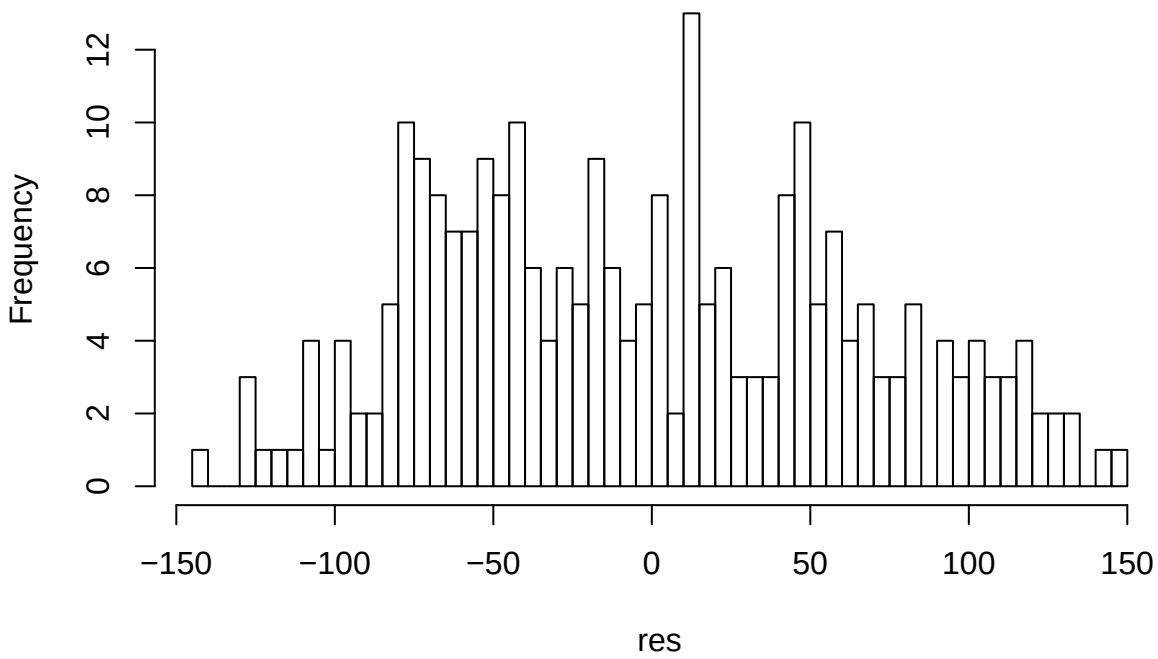
```
##
## Call:
## lm(formula = (popularity^1.3 - 1)/1.3 ~ energy + valence + log(liveness) +
##     tempo + log(speechiness) + instrumentalness + acousticness +
##     loudness + danceability)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -175.54  -46.08    1.89   45.94  164.09
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      194.136877  19.038247  10.197  < 2e-16 ***
## energy           -56.221713  13.390758  -4.199 2.80e-05 ***
## valence          -53.481201   7.530899  -7.102 1.71e-12 ***
## log(liveness)     -6.251298   2.318216  -2.697 0.007064 **
## tempo              0.003513   0.051729   0.068 0.945864
## log(speechiness)   0.892653   2.098486   0.425 0.670606
## instrumentalness  -8.286410  12.664369  -0.654 0.512988
## acousticness     -26.011240   6.869857  -3.786 0.000157 ***
## loudness           6.040478   0.798016   7.569 5.70e-14 ***
## danceability      66.800868  11.327930   5.897 4.34e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62.54 on 1990 degrees of freedom
## Multiple R-squared:  0.09358,    Adjusted R-squared:  0.08948
## F-statistic: 22.83 on 9 and 1990 DF,  p-value: < 2.2e-16
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm((popularity^1.3 − 1)/1.3 ~ energy + valence + log(liveness) + tempo + lo ...
Scale−Location
√|Standardized residuals|
982
1100
1370
Fitted values
lm((popularity^1.3 − 1)/1.3 ~ energy + valence + log(liveness) + tempo + lo ...



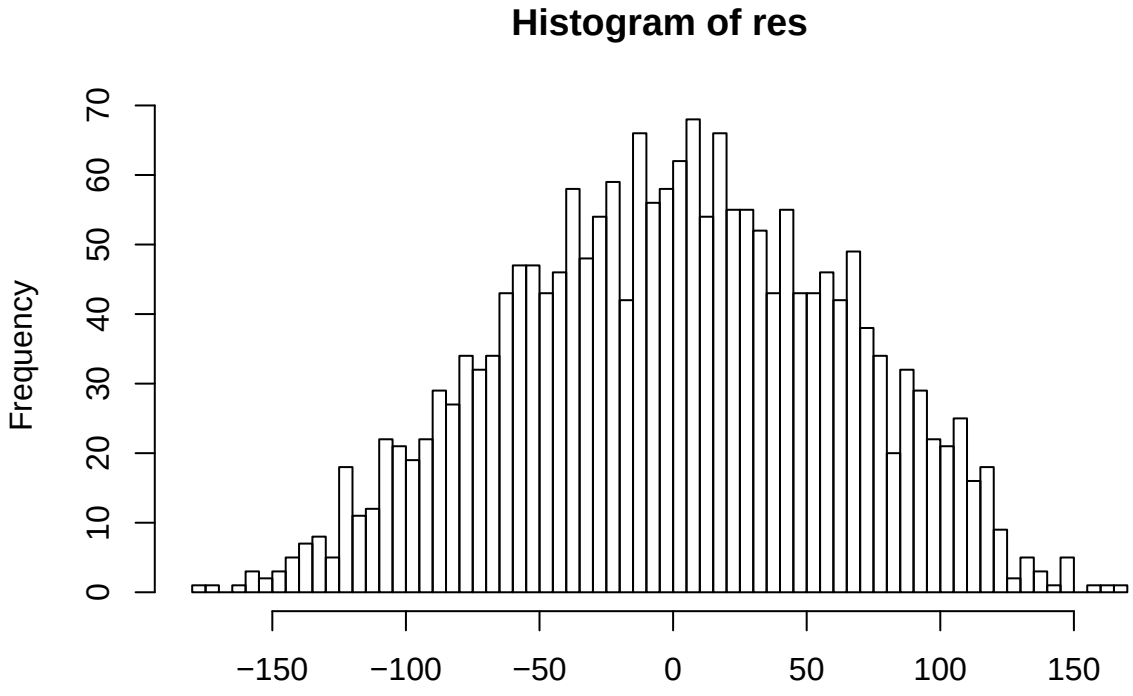

As we can see, $R^2$ improved while maintaining a significant model.

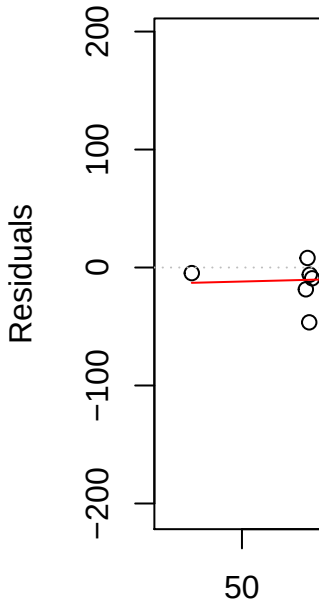
## Categorical Analysis

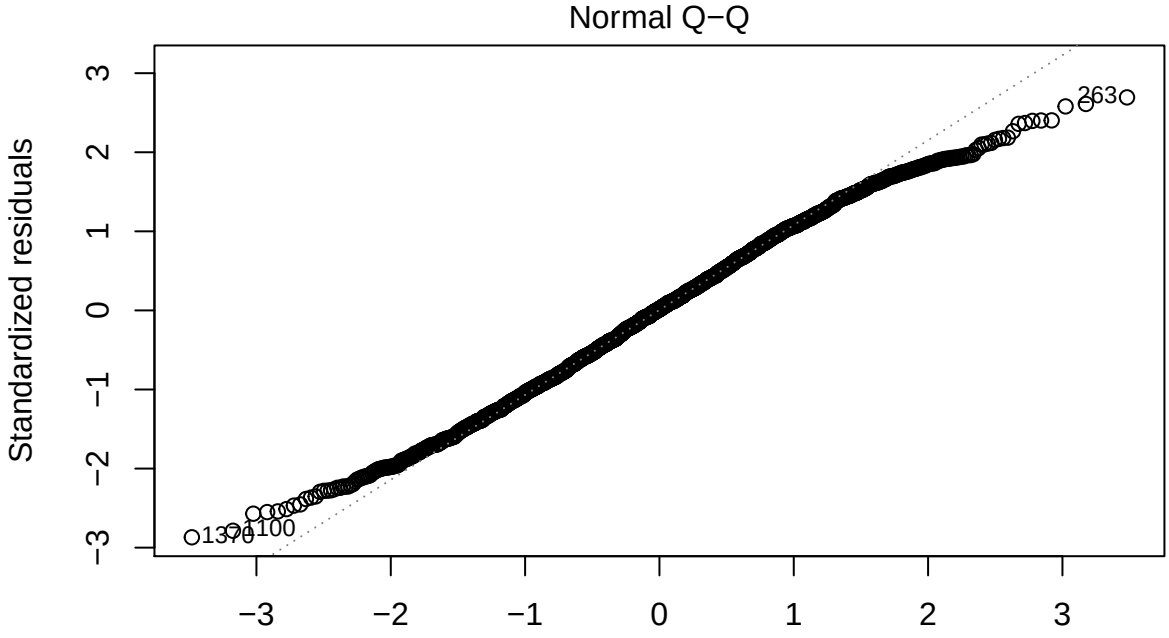
Now we try to add a categorical variable, mode to see if $R^2$ can be further improved. Mode encodes major scale as 1 and minor as 0. As shown below, $R^2$ is increased 7.68% from the naive MLR.

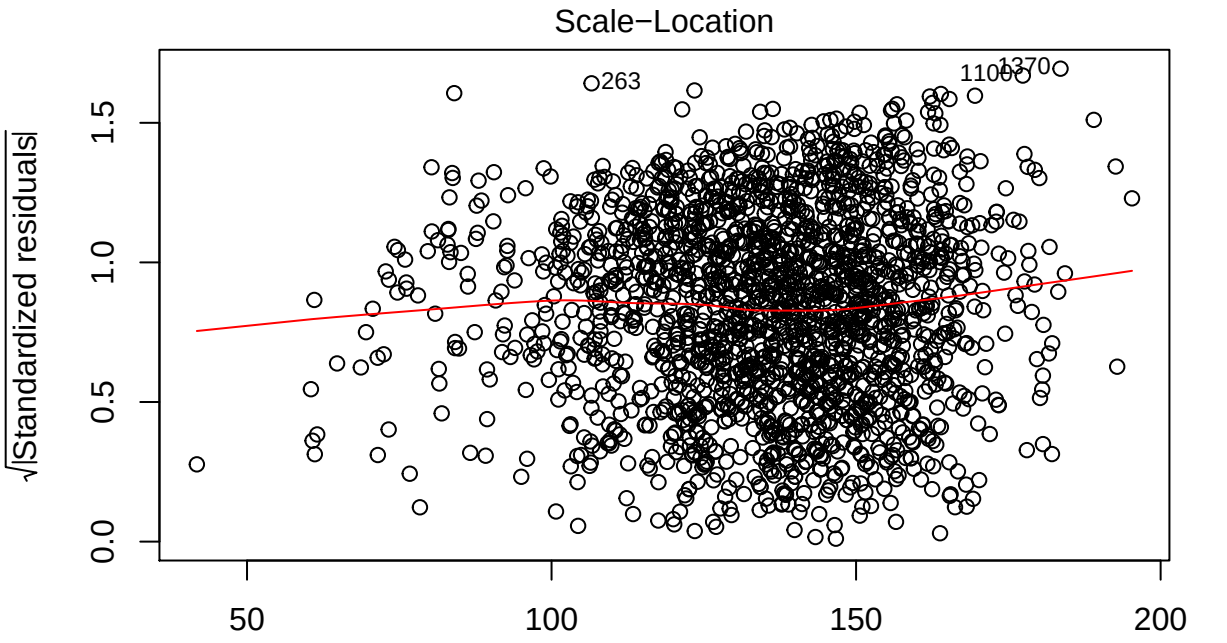
```
## F
##    0    1
##  818 1182
```

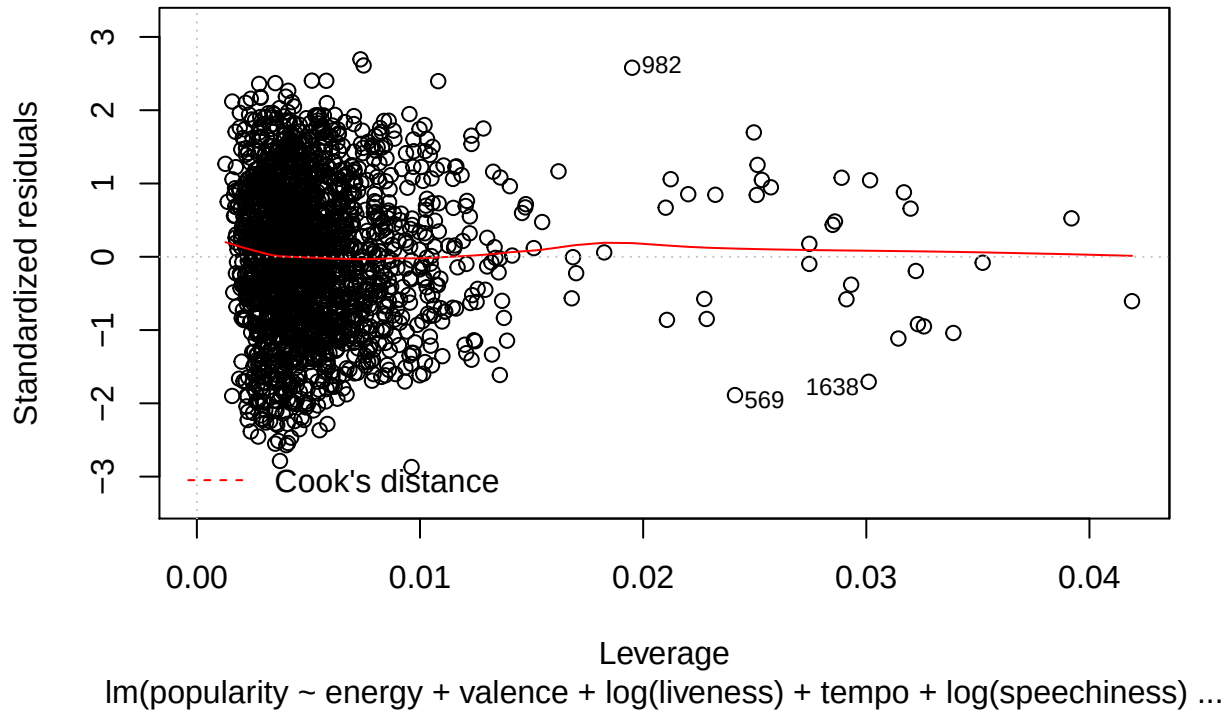
```
##
## Call:
## lm(formula = popularity ~ energy + valence + log(liveness) +
##     tempo + log(speechiness) + instrumentalness + acousticness +
##     loudness + danceability + mode)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -178.099  -44.685    1.022   45.429  167.439
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      187.2958    19.0936   9.809  < 2e-16 ***
## energy           -54.0742    13.3701  -4.044 5.45e-05 ***
## valence          -53.0072     7.5122  -7.056 2.35e-12 ***
## log(liveness)     -5.8902     2.3145  -2.545  0.01101 *
## tempo              0.0013     0.0516   0.025  0.97990
## log(speechiness)   1.4812     2.1000   0.705  0.48069
## instrumentalness  -8.3387    12.6308  -0.660  0.50921
## acousticness     -26.4130     6.8526  -3.854  0.00012 ***
## loudness           5.8750     0.7974   7.368 2.53e-13 ***
## danceability      68.4358    11.3080   6.052 1.71e-09 ***
## mode               9.7783     2.8701   3.407  0.00067 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62.38 on 1989 degrees of freedom
## Multiple R-squared:  0.09884,    Adjusted R-squared:  0.0943
## F-statistic: 21.81 on 10 and 1989 DF,  p-value: < 2.2e-16
```

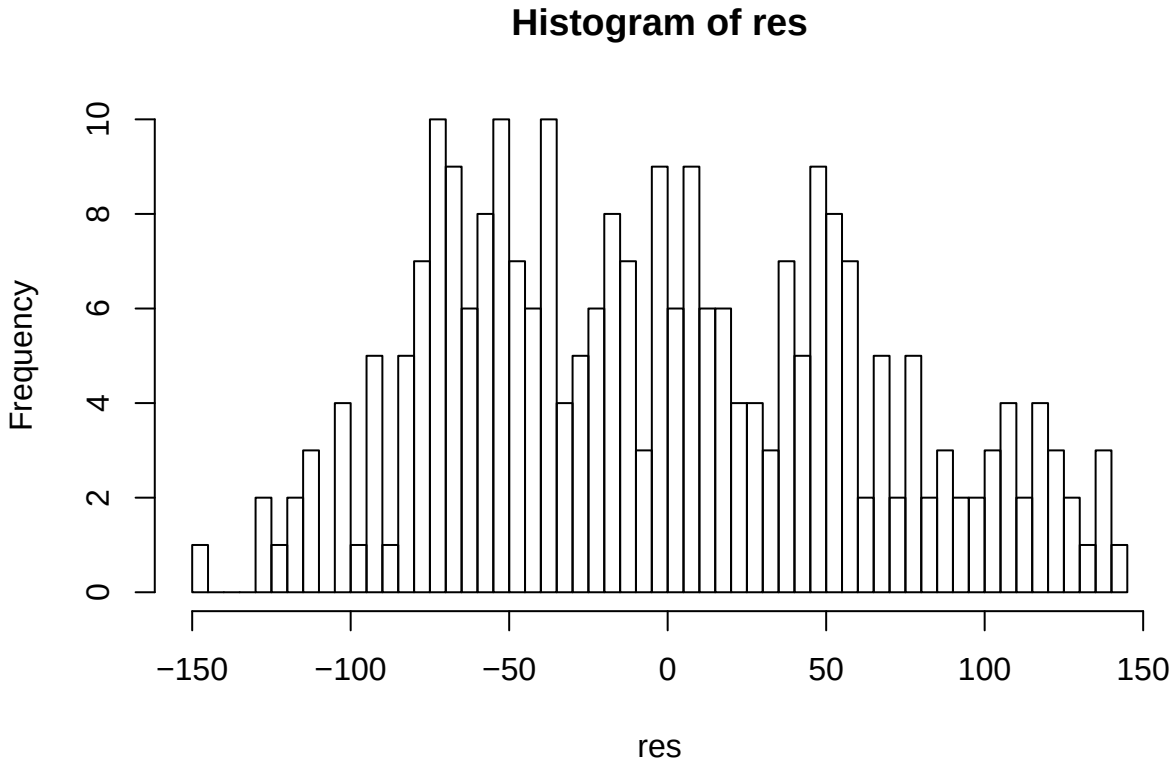

Histogram of res
Frequency
0
10
20
30
40
50
60
70
−150
−100
−50
0
50
100
150
res


Residuals
200
100
0
−100
−200
50
lm(popularity ~


Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
263
1370
100
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(popularity ~ energy + valence + log(liveness) + tempo + log(speechiness) ...


Scale−Location
√|Standardized residuals|
263
1370
0.0
0.5
1.0
1.5
50
100
150
200
Fitted values
lm(popularity ~ energy + valence + log(liveness) + tempo + log(speechiness) ...


Residuals vs Leverage
Standardized residuals
982
569
1638
Cook's distance
−3
−2
−1
0
1
2
3
0.00
0.01
0.02
0.03
0.04
Leverage
lm(popularity ~ energy + valence + log(liveness) + tempo + log(speechiness) ...




The difference between the popularity of a song resulting from changing from a minor scale to a major scale is 3.15.

### Removing insignificant variables

```
## Start:  AIC=16543.67
## popularity ~ energy + valence + log(liveness) + tempo + log(speechiness) +
##     instrumentalness + acousticness + loudness + danceability +
##     mode
##
##                    Df Sum of Sq     RSS   AIC
## - tempo             1         2 7738639 16542
## - instrumentalness  1      1696 7740333 16542
## - log(speechiness)  1      1936 7740572 16542
## <none>                          7738637 16544
## - log(liveness)     1     25199 7763836 16548
## - mode              1     45161 7783798 16553
## - acousticness      1     57803 7796440 16557
## - energy            1     63642 7802279 16558
## - danceability      1    142502 7881139 16578
## - valence           1    193716 7932353 16591
## - loudness          1    211207 7949844 16596
##
## Step:  AIC=16541.67
## popularity ~ energy + valence + log(liveness) + log(speechiness) +
##     instrumentalness + acousticness + loudness + danceability +
##     mode
##
##                    Df Sum of Sq     RSS   AIC
```

```
## - instrumentalness  1      1693 7740333 16540
## - log(speechiness)  1      1987 7740627 16540
## <none>                          7738639 16542
## - log(liveness)     1     25231 7763870 16546
## - mode              1     45177 7783816 16551
## - acousticness      1     57994 7796634 16555
## - energy            1     63690 7802330 16556
## - danceability      1    148454 7887094 16578
## - valence           1    194046 7932685 16589
## - loudness          1    211206 7949845 16594
##
## Step:  AIC=16540.11
## popularity ~ energy + valence + log(liveness) + log(speechiness) +
##     acousticness + loudness + danceability + mode
##
##                     Df Sum of Sq     RSS   AIC
## - log(speechiness)  1      2334 7742667 16539
## <none>                          7740333 16540
## - log(liveness)     1     24943 7765276 16544
## - mode              1     45146 7785479 16550
## - acousticness      1     57856 7798188 16553
## - energy            1     67265 7807597 16555
## - danceability      1    147437 7887770 16576
## - valence           1    193478 7933811 16588
## - loudness          1    233382 7973715 16598
##
## Step:  AIC=16538.71
## popularity ~ energy + valence + log(liveness) + acousticness +
##     loudness + danceability + mode
##
##                 Df Sum of Sq     RSS   AIC
## <none>                       7742667 16539
## - log(liveness)  1     24669 7767336 16543
## - mode           1     43776 7786443 16548
## - acousticness   1     58064 7800731 16552
## - energy         1     65381 7808048 16554
## - danceability   1    153980 7896647 16576
## - valence        1    191786 7934452 16586
## - loudness       1    231784 7974451 16596

## Start:  AIC=16605.28
## popularity ~ energy + valence + log(liveness) + tempo + log(speechiness) +
##     instrumentalness + acousticness + loudness + danceability +
##     mode
##
##                     Df Sum of Sq     RSS   AIC
## - tempo              1         2 7738639 16598
## - instrumentalness  1      1696 7740333 16598
## - log(speechiness)  1      1936 7740572 16598
## - log(liveness)     1     25199 7763836 16604
## <none>                          7738637 16605
## - mode              1     45161 7783798 16609
## - acousticness      1     57803 7796440 16613
## - energy            1     63642 7802279 16614
```

```
## - danceability      1    142502 7881139 16634
## - valence           1    193716 7932353 16647
## - loudness          1    211207 7949844 16652
##
## Step:  AIC=16597.68
## popularity ~ energy + valence + log(liveness) + log(speechiness) +
##     instrumentalness + acousticness + loudness + danceability +
##     mode
##
##                     Df Sum of Sq     RSS   AIC
## - instrumentalness  1      1693 7740333 16590
## - log(speechiness)  1      1987 7740627 16591
## - log(liveness)     1     25231 7763870 16597
## <none>                          7738639 16598
## - mode              1     45177 7783816 16602
## - acousticness      1     57994 7796634 16605
## - energy            1     63690 7802330 16606
## - danceability      1    148454 7887094 16628
## - valence           1    194046 7932685 16640
## - loudness          1    211206 7949845 16644
##
## Step:  AIC=16590.51
## popularity ~ energy + valence + log(liveness) + log(speechiness) +
##     acousticness + loudness + danceability + mode
##
##                     Df Sum of Sq     RSS   AIC
## - log(speechiness)  1      2334 7742667 16584
## - log(liveness)     1     24943 7765276 16589
## <none>                          7740333 16590
## - mode              1     45146 7785479 16594
## - acousticness      1     57856 7798188 16598
## - energy            1     67265 7807597 16600
## - danceability      1    147437 7887770 16621
## - valence           1    193478 7933811 16632
## - loudness          1    233382 7973715 16642
##
## Step:  AIC=16583.52
## popularity ~ energy + valence + log(liveness) + acousticness +
##     loudness + danceability + mode
##
##                 Df Sum of Sq     RSS   AIC
## - log(liveness)  1     24669 7767336 16582
## <none>                       7742667 16584
## - mode           1     43776 7786443 16587
## - acousticness   1     58064 7800731 16591
## - energy         1     65381 7808048 16593
## - danceability   1    153980 7896647 16615
## - valence        1    191786 7934452 16625
## - loudness       1    231784 7974451 16635
##
## Step:  AIC=16582.28
## popularity ~ energy + valence + acousticness + loudness + danceability +
##     mode
##
```

```
##                Df Sum of Sq     RSS   AIC
## <none>                      7767336 16582
## - mode          1     47088 7814424 16587
## - acousticness  1     61459 7828795 16590
## - energy        1     73438 7840774 16594
## - danceability  1    169156 7936491 16618
## - valence       1    191673 7959009 16623
## - loudness      1    234141 8001477 16634
```

```
## Start:  AIC=16731.8
## popularity ~ 1
##
##                    Df Sum of Sq     RSS   AIC
## + loudness          1    235221 8352149 16678
## + acousticness      1    158871 8428500 16696
## + danceability      1    126569 8460801 16704
## + valence           1    105201 8482170 16709
## + mode              1     54394 8532977 16721
## + log(liveness)     1     52608 8534762 16722
## + instrumentalness  1     25303 8562068 16728
## <none>                          8587371 16732
## + energy            1      5878 8581493 16732
## + tempo             1      3652 8583719 16733
## + log(speechiness)  1      1561 8585810 16733
##
## Step:  AIC=16678.25
## popularity ~ loudness
##
##                    Df Sum of Sq     RSS   AIC
## + valence           1    190638 8161512 16634
## + energy            1    136334 8215815 16647
## + danceability      1     85450 8266700 16660
## + log(liveness)     1     62141 8290008 16665
## + mode              1     54696 8297453 16667
## + acousticness      1     31954 8320196 16673
## <none>                          8352149 16678
## + tempo             1      7673 8344476 16678
## + instrumentalness  1      4311 8347838 16679
## + log(speechiness)  1        76 8352074 16680
##
## Step:  AIC=16634.07
## popularity ~ loudness + valence
##
##                    Df Sum of Sq     RSS   AIC
## + danceability      1    251256 7910256 16574
## + log(liveness)     1     64697 8096815 16620
## + energy            1     58269 8103243 16622
## + acousticness      1     43395 8118117 16625
## + mode              1     40971 8120541 16626
## <none>                          8161512 16634
## + tempo             1      7674 8153837 16634
## + log(speechiness)  1      3671 8157841 16635
## + instrumentalness  1      2663 8158849 16635
##
```

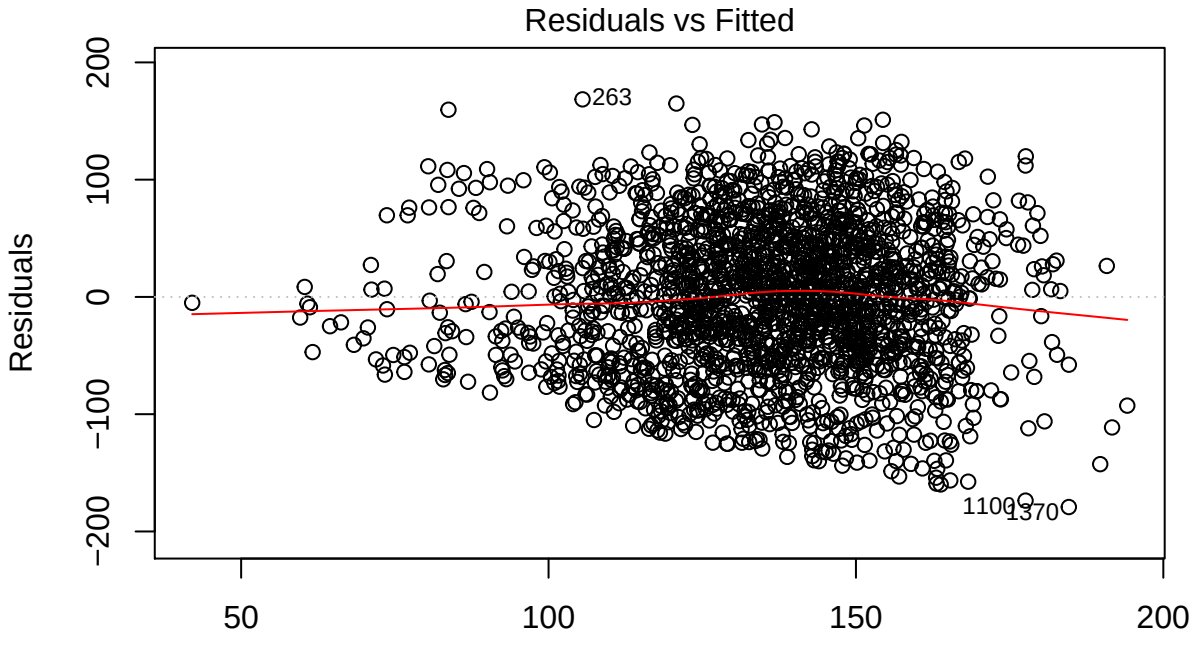
```
## Step:  AIC=16573.54
## popularity ~ loudness + valence + danceability
##
##                     Df Sum of Sq     RSS   AIC
## + mode               1     51678 7858578 16562
## + log(liveness)      1     37101 7873155 16566
## + energy             1     35996 7874260 16566
## + acousticness       1     14796 7895460 16572
## <none>                           7910256 16574
## + instrumentalness   1      3833 7906423 16575
## + tempo              1        51 7910205 16576
## + log(speechiness)   1        17 7910239 16576
##
## Step:  AIC=16562.43
## popularity ~ loudness + valence + danceability + mode
##
##                     Df Sum of Sq     RSS   AIC
## + log(liveness)      1     32667 7825912 16556
## + energy             1     29783 7828795 16557
## + acousticness       1     17805 7840774 16560
## <none>                           7858578 16562
## + instrumentalness   1      3821 7854757 16564
## + log(speechiness)   1       634 7857944 16564
## + tempo              1        65 7858513 16564
##
## Step:  AIC=16556.1
## popularity ~ loudness + valence + danceability + mode + log(liveness)
##
##                     Df Sum of Sq     RSS   AIC
## + energy             1   25180.8 7800731 16552
## + acousticness       1   17864.3 7808048 16554
## <none>                           7825912 16556
## + instrumentalness   1    4145.1 7821767 16557
## + log(speechiness)   1     915.9 7824996 16558
## + tempo              1      18.9 7825893 16558
##
## Step:  AIC=16551.65
## popularity ~ loudness + valence + danceability + mode + log(liveness) +
##     energy
##
##                     Df Sum of Sq     RSS   AIC
## + acousticness       1     58064 7742667 16539
## <none>                           7800731 16552
## + log(speechiness)   1      2543 7798188 16553
## + instrumentalness   1      1904 7798827 16553
## + tempo              1       346 7800385 16554
##
## Step:  AIC=16538.71
## popularity ~ loudness + valence + danceability + mode + log(liveness) +
##     energy + acousticness
##
##                     Df Sum of Sq     RSS   AIC
## <none>                           7742667 16539
## + log(speechiness)   1   2334.22 7740333 16540
```

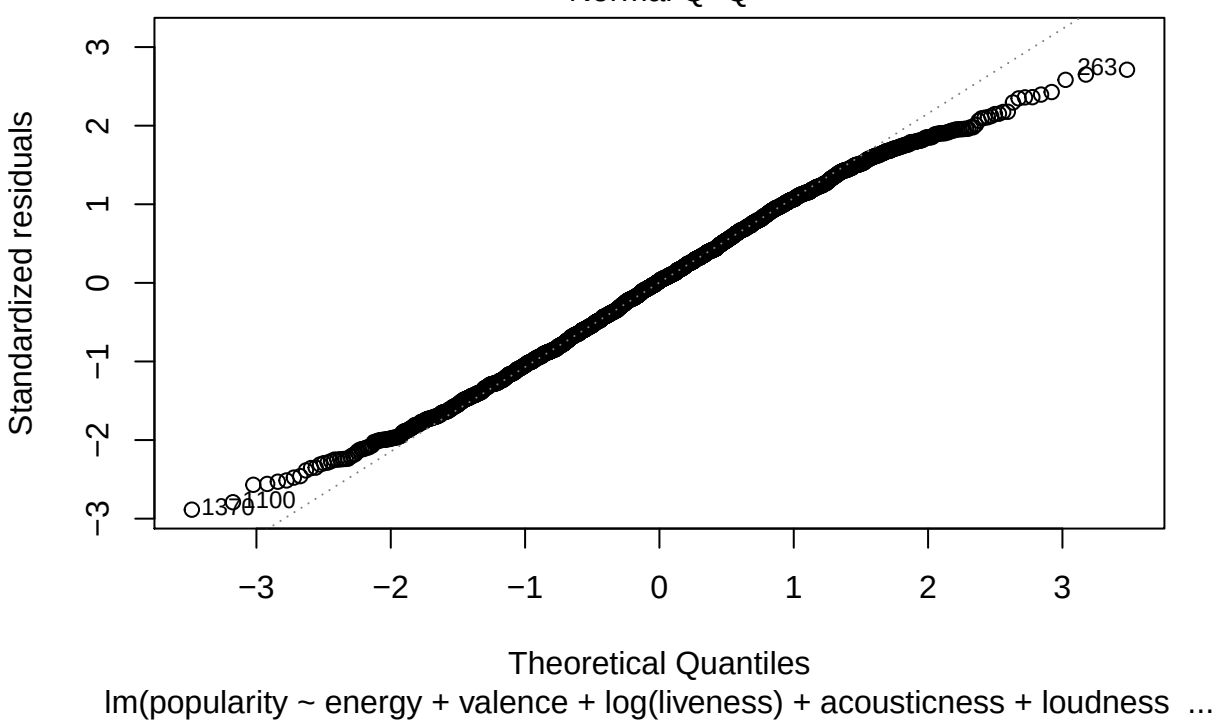
```
## + instrumentalness  1   2040.26 7740627 16540
## + tempo             1     36.22 7742631 16541
```

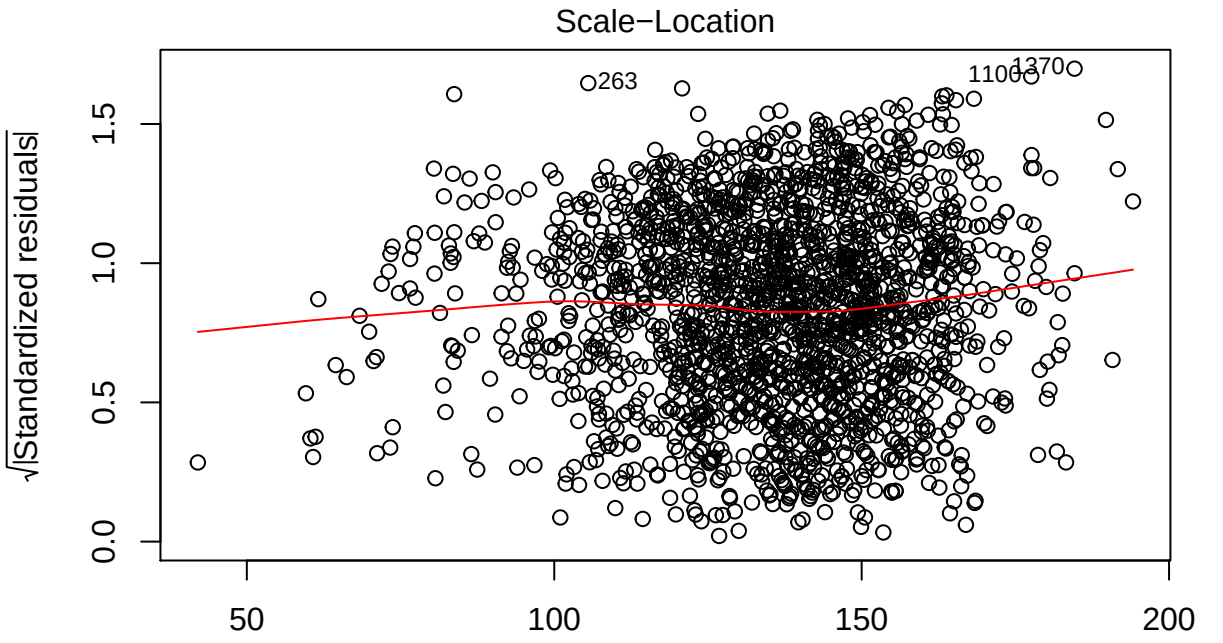
```
## Start:  AIC=16737.4
## popularity ~ 1
##
##                     Df Sum of Sq     RSS   AIC
## + loudness           1    235221 8352149 16690
## + acousticness       1    158871 8428500 16708
## + danceability       1    126569 8460801 16715
## + valence            1    105201 8482170 16720
## + mode               1     54394 8532977 16732
## + log(liveness)      1     52608 8534762 16733
## <none>                           8587371 16737
## + instrumentalness   1     25303 8562068 16739
## + energy             1      5878 8581493 16744
## + tempo              1      3652 8583719 16744
## + log(speechiness)   1      1561 8585810 16745
##
## Step:  AIC=16689.46
## popularity ~ loudness
##
##                     Df Sum of Sq     RSS   AIC
## + valence            1    190638 8161512 16651
## + energy             1    136334 8215815 16664
## + danceability       1     85450 8266700 16676
## + log(liveness)      1     62141 8290008 16682
## + mode               1     54696 8297453 16684
## + acousticness       1     31954 8320196 16689
## <none>                           8352149 16690
## + tempo              1      7673 8344476 16695
## + instrumentalness   1      4311 8347838 16696
## + log(speechiness)   1        76 8352074 16697
##
## Step:  AIC=16650.88
## popularity ~ loudness + valence
##
##                     Df Sum of Sq     RSS   AIC
## + danceability       1    251256 7910256 16596
## + log(liveness)      1     64697 8096815 16643
## + energy             1     58269 8103243 16644
## + acousticness       1     43395 8118117 16648
## + mode               1     40971 8120541 16648
## <none>                           8161512 16651
## + tempo              1      7674 8153837 16657
## + log(speechiness)   1      3671 8157841 16658
## + instrumentalness   1      2663 8158849 16658
##
## Step:  AIC=16595.94
## popularity ~ loudness + valence + danceability
##
##                     Df Sum of Sq     RSS   AIC
## + mode               1     51678 7858578 16590
## + log(liveness)      1     37101 7873155 16594
```

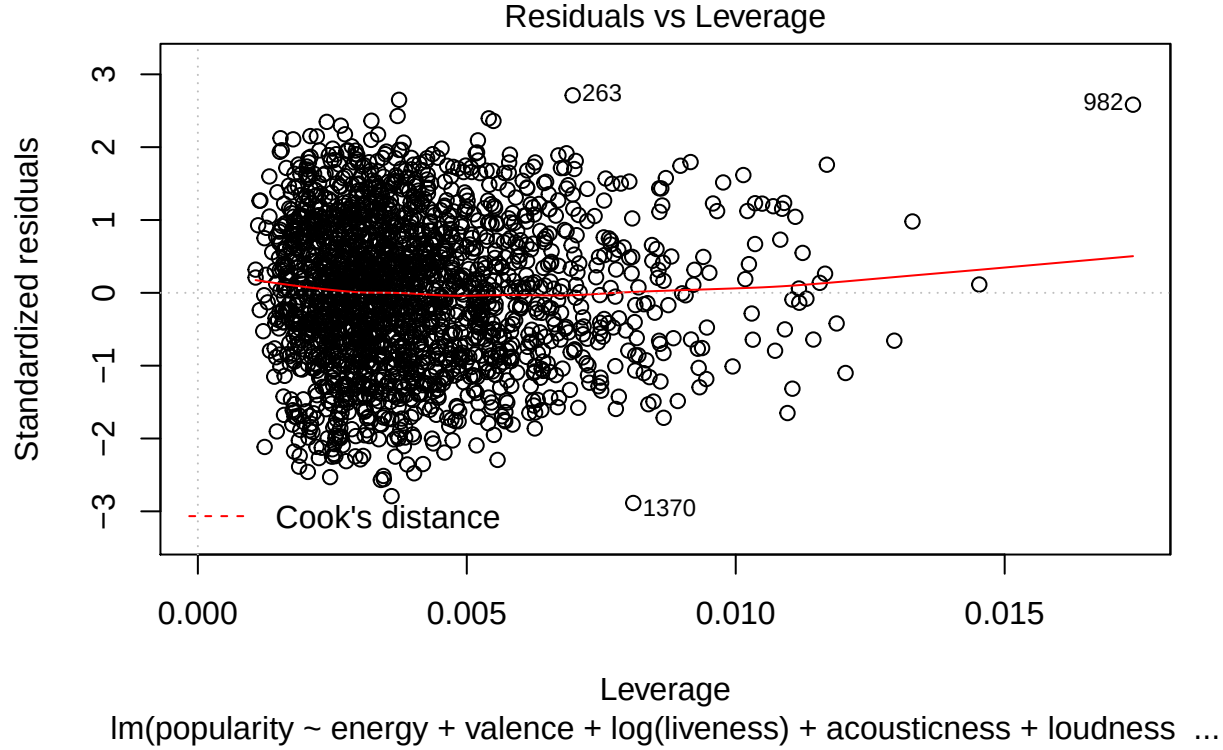
```
## + energy            1     35996 7874260 16594
## <none>                          7910256 16596
## + acousticness      1     14796 7895460 16600
## + instrumentalness  1      3833 7906423 16603
## + tempo             1        51 7910205 16604
## + log(speechiness)  1        17 7910239 16604
##
## Step:  AIC=16590.43
## popularity ~ loudness + valence + danceability + mode
##
##                     Df Sum of Sq     RSS   AIC
## + log(liveness)      1     32667 7825912 16590
## <none>                           7858578 16590
## + energy             1     29783 7828795 16590
## + acousticness       1     17805 7840774 16594
## + instrumentalness   1      3821 7854757 16597
## + log(speechiness)   1       634 7857944 16598
## + tempo              1        65 7858513 16598
##
## Step:  AIC=16589.7
## popularity ~ loudness + valence + danceability + mode + log(liveness)
##
##                     Df Sum of Sq     RSS   AIC
## <none>                           7825912 16590
## + energy             1   25180.8 7800731 16591
## + acousticness       1   17864.3 7808048 16593
## + instrumentalness   1    4145.1 7821767 16596
## + log(speechiness)   1     915.9 7824996 16597
## + tempo              1      18.9 7825893 16597
##
## Call:
## lm(formula = popularity ~ energy + valence + log(liveness) +
##     acousticness + loudness + danceability + mode)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -179.165  -44.846    1.626   45.435  168.475
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   183.4591    15.4576  11.869  < 2e-16 ***
## energy        -54.0635    13.1820  -4.101 4.27e-05 ***
## valence       -52.6091     7.4895  -7.024 2.94e-12 ***
## log(liveness)  -5.8232     2.3115  -2.519 0.011837 *
## acousticness  -26.4360     6.8398  -3.865 0.000115 ***
## loudness        5.9733     0.7735   7.722 1.80e-14 ***
## danceability   69.0998    10.9786   6.294 3.79e-10 ***
## mode            9.5942     2.8589   3.356 0.000806 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62.34 on 1992 degrees of freedom
## Multiple R-squared:  0.09837,    Adjusted R-squared:  0.0952
```

```
## F-statistic: 31.05 on 7 and 1992 DF,  p-value: < 2.2e-16
```





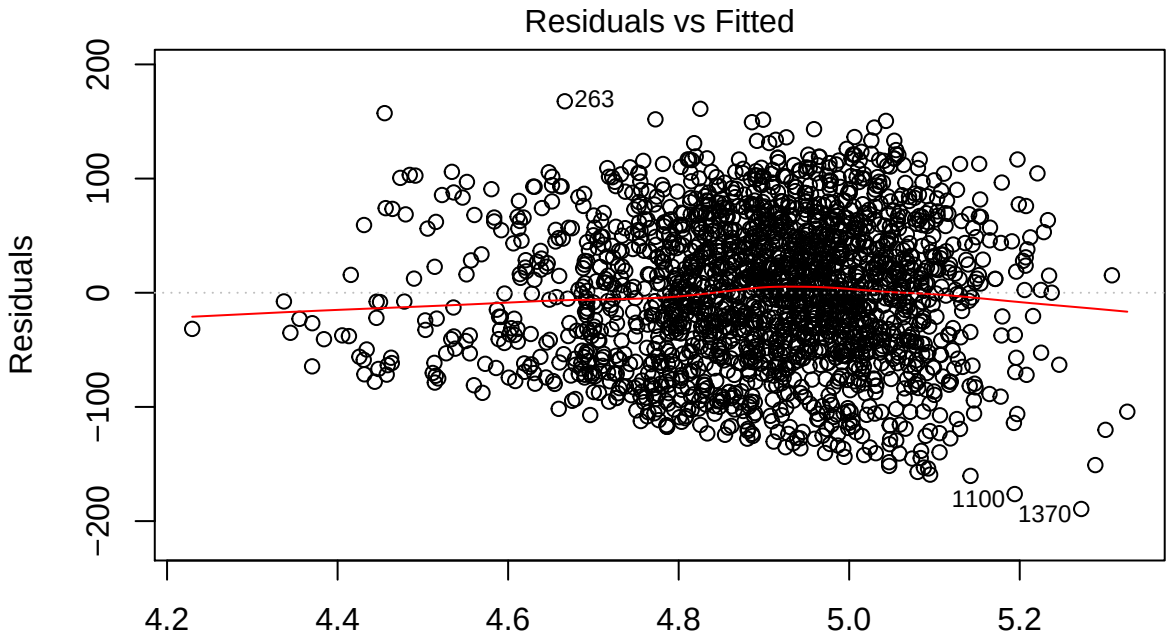

Backward AIC and BIC result: popularity ~ energy + valence + log(liveness) + acousticness + loudness + danceability + mode

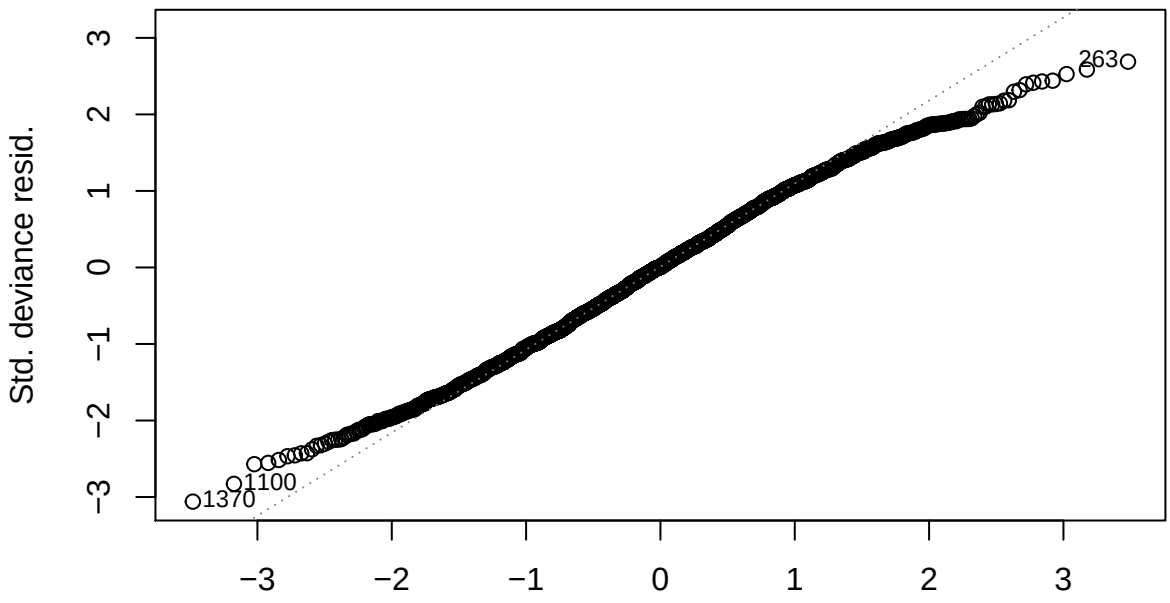
Forward AIC: popularity ~ energy + valence + log(liveness) + acousticness + loudness + danceability + mode

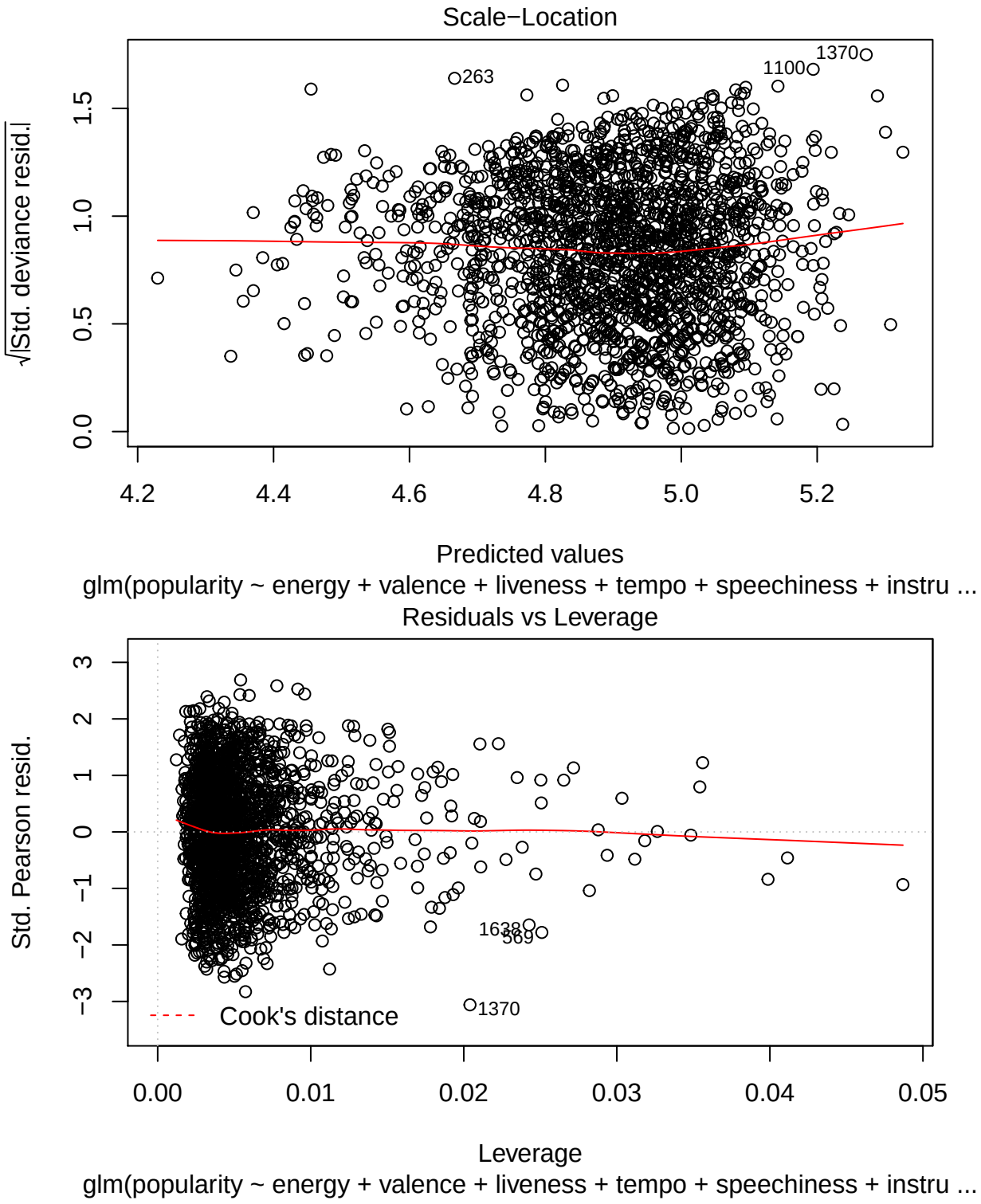
Forward BIC : popularity ~ loudness + valence + danceability + mode + log(liveness) + energy + acousticness

## GLM

```
##
## Call:
## glm(formula = popularity ~ energy + valence + liveness + tempo +
##     speechiness + instrumentalness + acousticness + loudness +
##     danceability + mode, family = gaussian(link = "log"))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -189.373   -44.938     0.473    46.311   167.690
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.400e+00  1.248e-01  43.255  < 2e-16 ***
## energy           -4.008e-01  9.879e-02  -4.057 5.16e-05 ***
## valence          -3.704e-01  5.534e-02  -6.693 2.84e-11 ***
## liveness         -1.617e-01  8.199e-02  -1.973 0.048661 *
## tempo            -5.595e-06  3.868e-04  -0.014 0.988462
## speechiness       1.063e-01  1.275e-01   0.834 0.404302
## instrumentalness -1.056e-01  1.086e-01  -0.972 0.331007
## acousticness     -1.843e-01  5.204e-02  -3.542 0.000407 ***
## loudness          4.687e-02  6.353e-03   7.378 2.34e-13 ***
## danceability      4.675e-01  8.294e-02   5.636 1.99e-08 ***
## mode              6.084e-02  2.121e-02   2.868 0.004170 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 3910.523)
##
##     Null deviance: 8587371  on 1999  degrees of freedom
## Residual deviance: 7778048  on 1989  degrees of freedom
## AIC: 22232
##
## Number of Fisher Scoring iterations: 6
```

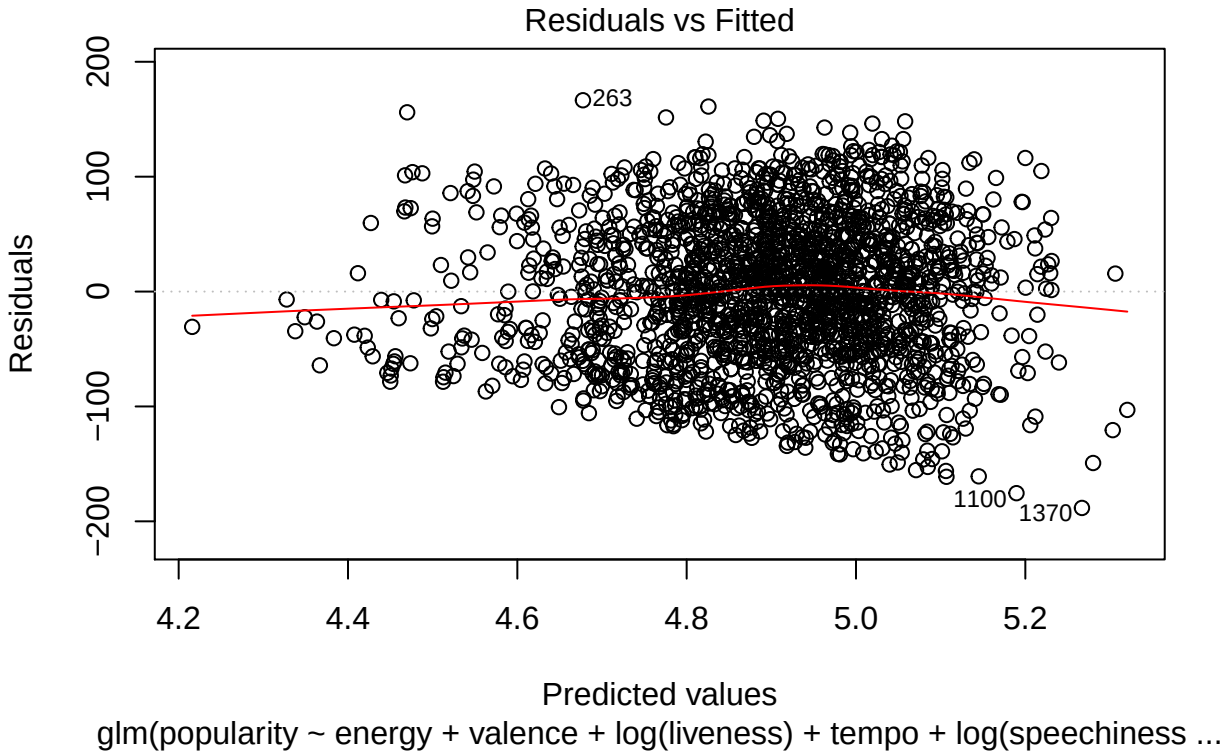

Residuals vs Fitted
Residuals
200
100
0
−100
−200
263
1100
1370
4.2
4.4
4.6
4.8
5.0
5.2
Predicted values
glm(popularity ~ energy + valence + liveness + tempo + speechiness + instru ...

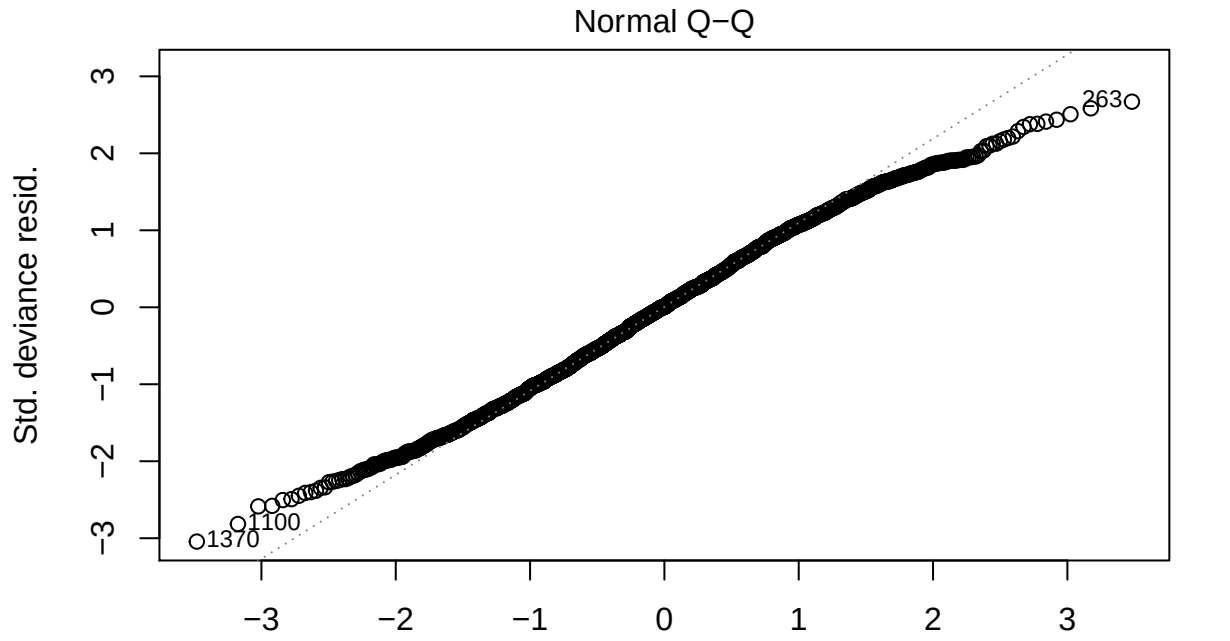

Normal Q−Q
Std. deviance resid.
3
2
1
0
−1
−2
−3
263
1100
1370
−3
−2
−1
0
1
2
3
Theoretical Quantiles
glm(popularity ~ energy + valence + liveness + tempo + speechiness + instru ...

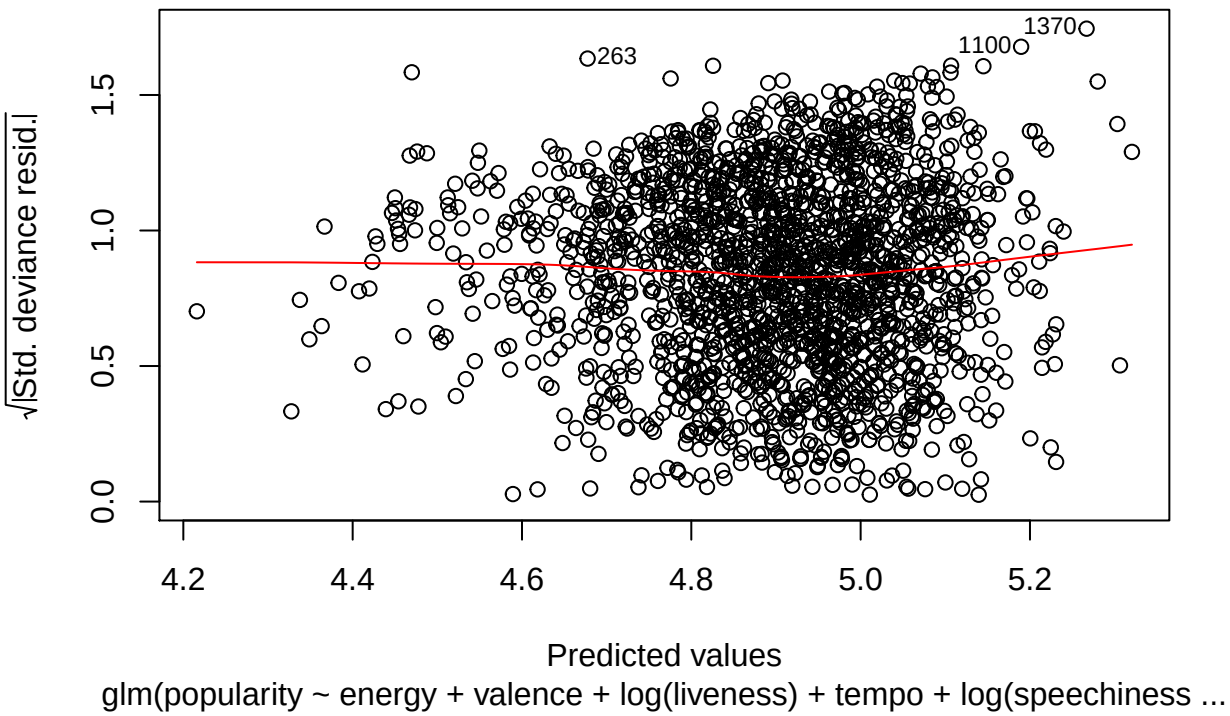


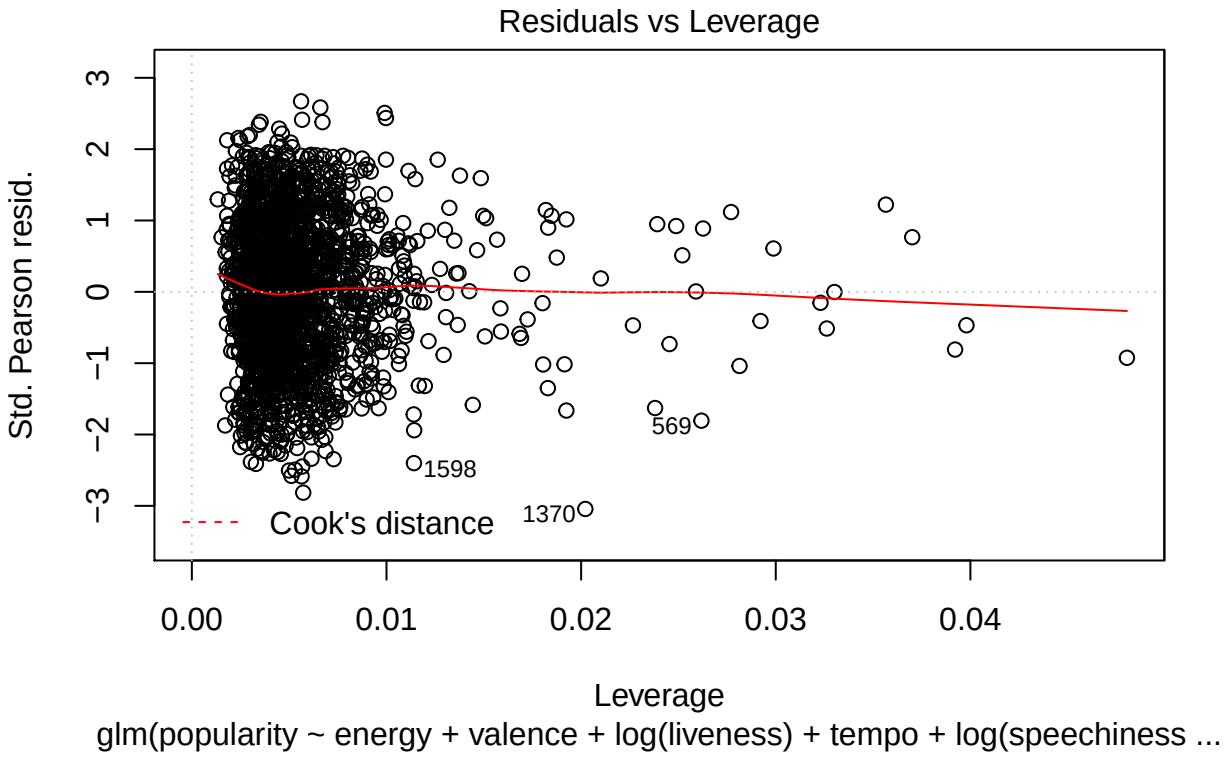

```
##
## Call:
## glm(formula = popularity ~ energy + valence + log(liveness) +
##     tempo + log(speechiness) + instrumentalness + acousticness +
##     loudness + danceability + mode, family = gaussian(link = "log"))
##
```

```
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -188.343   -45.431     0.473    46.323   166.508
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.361e+00  1.452e-01  36.913  < 2e-16 ***
## energy           -4.041e-01  9.899e-02  -4.082 4.64e-05 ***
## valence          -3.743e-01  5.532e-02  -6.767 1.72e-11 ***
## log(liveness)    -3.847e-02  1.732e-02  -2.221 0.026462 *
## tempo            -2.322e-05  3.867e-04  -0.060 0.952130
## log(speechiness)  1.545e-02  1.547e-02   0.999 0.317948
## instrumentalness -1.055e-01  1.086e-01  -0.972 0.331133
## acousticness     -1.813e-01  5.203e-02  -3.485 0.000503 ***
## loudness          4.698e-02  6.342e-03   7.408 1.89e-13 ***
## danceability      4.569e-01  8.366e-02   5.462 5.31e-08 ***
## mode              6.104e-02  2.123e-02   2.875 0.004079 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 3908.22)
##
##     Null deviance: 8587371  on 1999  degrees of freedom
## Residual deviance: 7773467  on 1989  degrees of freedom
## AIC: 22230
##
## Number of Fisher Scoring iterations: 6
```



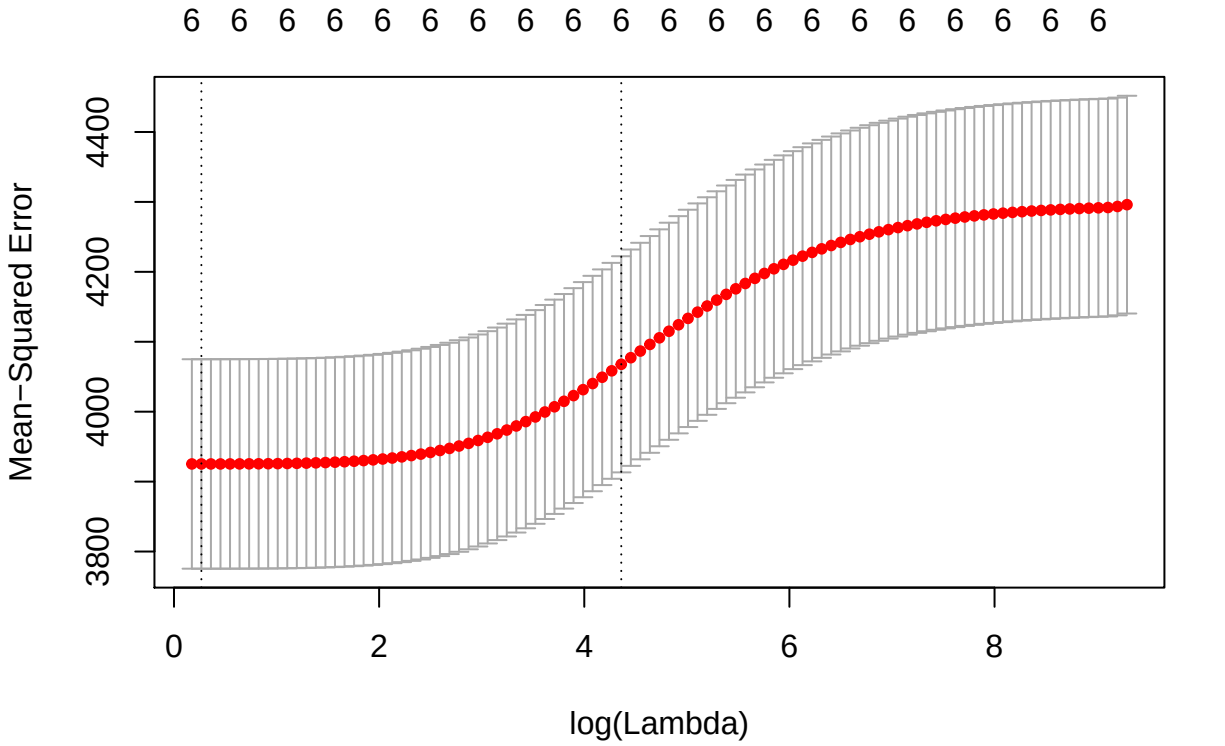

Normal Q−Q
Std. deviance resid.
Theoretical Quantiles
glm(popularity ~ energy + valence + log(liveness) + tempo + log(speechiness ...
263
1100
1370


Scale−Location
√|Std. deviance resid.|
Predicted values
glm(popularity ~ energy + valence + log(liveness) + tempo + log(speechiness ...
263
1100
1370



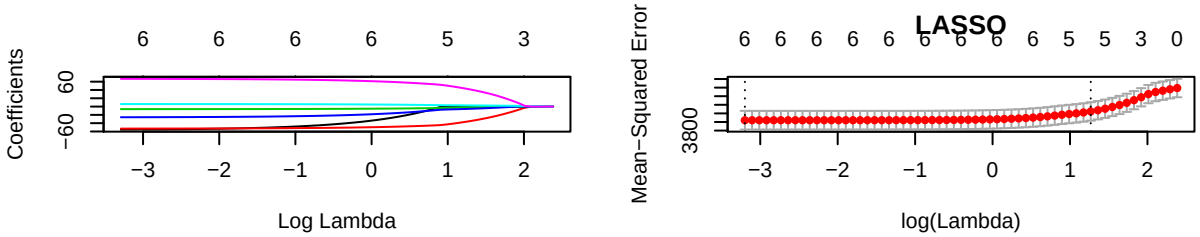

### Ridge, LASSO, and Elastic Net Models

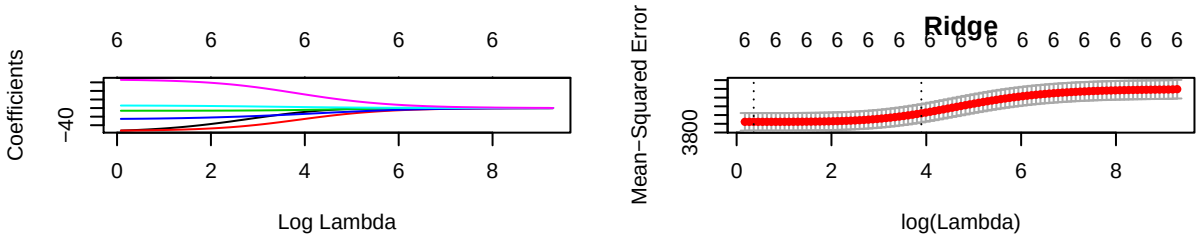
```
## [1] 1.306265
```

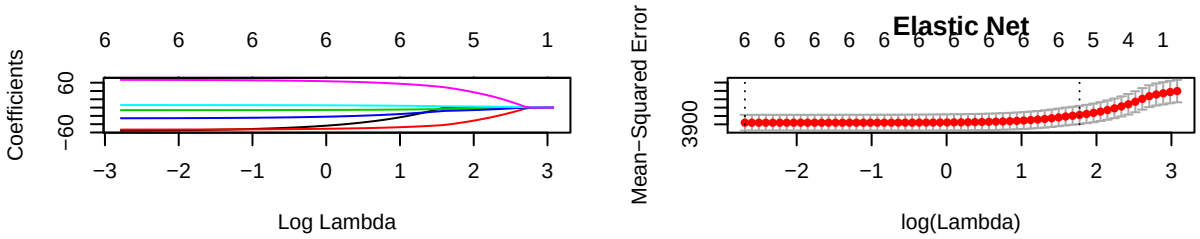



```
## 7 x 1 sparse Matrix of class "dgCMatrix"
```

```
##                              1
## (Intercept)      142.193738
## energy            -5.576769
## valence          -20.246872
##                   -3.593482
## acousticness     -11.638673
## loudness           1.860652
## danceability      25.872250
```




```
## 8 x 1 sparse Matrix of class "dgCMatrix"
##                              1
## (Intercept)      194.744549
## liveness         -24.273827
## energy           -50.286777
## mode               9.548689
## acousticness     -25.950814
## loudness           5.707692
## valence          -51.541311
## danceability      68.391426
```

```
##       RMSE    Rsquare
## 1 65.37451 0.07692129
```

```
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info =
## trainInfo, : There were missing values in resampled performance measures.
```

```
## 8 x 1 sparse Matrix of class "dgCMatrix"
##                              1
## (Intercept)      197.573271
## liveness         -24.036395
```

```
## energy       -53.082645
## mode           9.582212
## acousticness -26.461968
## loudness       5.902543
## valence      -52.237991
## danceability  69.453568
```

```
##       RMSE    Rsquare
## 1 65.37211 0.07732247
```

```
## 8 x 1 sparse Matrix of class "dgCMatrix"
##                       1
## (Intercept)  196.598935
## liveness     -24.078865
## energy       -52.106389
## mode           9.563593
## acousticness -26.259438
## loudness       5.839207
## valence      -52.040707
## danceability  69.134107
```

```
##       RMSE    Rsquare
## 1 65.37383 0.07716293
```

```
##
## Call:
## summary.resamples(object = ., metric = "RMSE")
##
## Models: ridge, lasso, elastic
## Number of resamples: 10
##
## RMSE
##             Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
## ridge   58.47129 61.36450 61.81626 62.39844 63.15454 67.01711    0
## lasso   59.27477 61.10752 62.92250 62.51121 63.55660 65.70513    0
## elastic 59.86233 60.64360 61.66329 62.44687 64.11779 66.14420    0
```

The ridge regression yields the most optimum RMSE with the coefficients:

(Intercept) 70.587562856 \ liveness -8.166892612 \ tempo -0.001324982 \ energy -15.921887134 \ speechiness 3.058979282 \ mode 3.114763830 \ instrumentalness -1.940615809 \ acousticness -8.374153049 \ loudness 1.745995798 \ valence -16.069317150 \ danceability 21.587672148 \