Predicting Song Popularity - Model Testing

STAT 420, Summer 2023, UIUC - Final Data Project

Soumya Nanda, Jonas Jansen, Noam Isachar

Extracted this file for faster computations on model finding. The final models should be included into the report file.

```
library(knitr)
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(MASS)
library(gridExtra)
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(car)
```

```
## Loading required package: carData
```

Data preparation

```
data <- read.csv("../data/spotify_data.csv")
#str(data)
```

```
data$mode <- ifelse(data$mode == 1, "major", "minor")
data$duration_m <- data$duration_ms / 1000 / 60

data$year <- factor(data$year)
data$genre <- factor(data$genre)
data$key <- factor(data$key)
data$mode <- factor(data$mode)
data$time_signature <- factor(data$time_signature)

data <- subset(data, select = -c(X, track_id, duration_ms))
```

```
#str(data)
```

```
data <- data[data$duration_m <= 40, ]
#summary(data$duration_m)
```

Model finding

Prepare test and Train data.

For an easier model fitting, let's create a data set without `artist_name` and `track_name`.

```
fit_data = subset(data, select = -c(artist_name, track_name))
```

Split data into test and train. Using 70% train data and 30% test data.

```
set.seed(42)

# TODO Increase the train data size to 0.7
ratio = 0.1
sample_size = floor(ratio * nrow(fit_data))

data_idx = sample(nrow(fit_data), sample_size)
data_trn = fit_data[data_idx, ]
data_tst = fit_data[-data_idx, ]
```

Model Test Stats

Create a function, that is performing several test on the model and return the results.

```
check_model = function(model, test_data){
  sample_idx = sample(length(resid(model)), 5000)
  residuals_sample = resid(model)[sample_idx]

  # shapiro
  shapiro =  shapiro.test(residuals_sample)$p.value

  # bp test
  bptest = bptest(model)$p.value

  # leverage
  high_leverage_count = sum(hatvalues(model) > 2 * mean(hatvalues(model)))

  # outliers
  outliers_count = length(rstandard(model)[abs(rstandard(model)) > 2])

  # influence
  influence_count = sum(cooks.distance(model) > 4 / length(cooks.distance(model)))

  # vif
  vif_values <- car::vif(model)
  predictor_names = names(model$coefficients)[-1]
  high_vif_predictors = predictor_names[vif_values > 5]
  high_vif_count = sum(vif_values > 5)

  # loocv_rmse
```

```
  loocv_rmse = sqrt(mean((resid(model) / (1 - hatvalues(model))) ^ 2))

  # adjusted r squared
  adjusted_r_squared = summary(model)$adj.r.squared

  # TODO: Some test to test against the test_data

  # Creating a data frame to store the results
  results = data.frame(
    shapiro = shapiro,
    bptest = bptest,
    high_leverage_count = high_leverage_count,
    outliers_count = outliers_count,
    influence_count = influence_count,
    high_vif_count = high_vif_count,
    # TODO: Proper format list of high vif predictors.
    #high_vif_predictors = high_vif_predictors,
    loocv_rmse = loocv_rmse,
    adjusted_r_squared = adjusted_r_squared
  )

  results

}
```

First model. Full additive.

First try, fit an full additive model with popularity as response.

```
model_add_full = lm(popularity ~ ., data = data_trn)
```

Let's do some first tests.

```
summary(model_add_full)
```

```
##
## Call:
## lm(formula = popularity ~ ., data = data_trn)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -59.04  -7.14  -1.58   5.45  68.81
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             18.13769    1.04619   17.34  < 2e-16 ***
## year2001                 0.47625    0.22464    2.12  0.03400 *
## year2002                 0.49182    0.22133    2.22  0.02628 *
## year2003                 1.49085    0.22323    6.68  2.4e-11 ***
## year2004                 1.19187    0.22273    5.35  8.8e-08 ***
## year2005                 2.69317    0.22146   12.16  < 2e-16 ***
## year2006                 2.49645    0.21897   11.40  < 2e-16 ***
## year2007                 3.24087    0.21963   14.76  < 2e-16 ***
## year2008                 3.48570    0.21764   16.02  < 2e-16 ***
## year2009                 4.73724    0.21871   21.66  < 2e-16 ***
```

```
## year2010                 4.80476    0.21743   22.10  < 2e-16 ***
## year2011                 6.15165    0.21869   28.13  < 2e-16 ***
## year2012                 6.65118    0.21017   31.65  < 2e-16 ***
## year2013                 7.71820    0.21232   36.35  < 2e-16 ***
## year2014                 8.93639    0.21269   42.02  < 2e-16 ***
## year2015                10.32488    0.21436   48.17  < 2e-16 ***
## year2016                11.20768    0.22776   49.21  < 2e-16 ***
## year2017                12.96487    0.20991   61.76  < 2e-16 ***
## year2018                13.86487    0.20876   66.41  < 2e-16 ***
## year2019                15.02936    0.20963   71.70  < 2e-16 ***
## year2020                16.28389    0.21126   77.08  < 2e-16 ***
## year2021                17.24776    0.21313   80.93  < 2e-16 ***
## year2022                20.67819    0.21272   97.21  < 2e-16 ***
## year2023                 8.97961    0.22951   39.12  < 2e-16 ***
## genreafrobeat          -10.78597    0.38148  -28.27  < 2e-16 ***
## genrealt-rock           20.49670    0.33042   62.03  < 2e-16 ***
## genreambient             5.76998    0.33835   17.05  < 2e-16 ***
## genreblack-metal        -5.00178    0.34100  -14.67  < 2e-16 ***
## genreblues               4.07967    0.33173   12.30  < 2e-16 ***
## genrebreakbeat         -14.49233    0.38539  -37.60  < 2e-16 ***
## genrecantopop           -6.11917    0.35251  -17.36  < 2e-16 ***
## genrechicago-house     -15.93515    0.53009  -30.06  < 2e-16 ***
## genrechill               6.94967    0.33393   20.81  < 2e-16 ***
## genreclassical          11.27956    0.34579   32.62  < 2e-16 ***
## genreclub               -8.36557    0.35876  -23.32  < 2e-16 ***
## genrecomedy             -7.42162    0.40499  -18.33  < 2e-16 ***
## genrecountry            14.80063    0.33910   43.65  < 2e-16 ***
## genredance              23.80275    0.34379   69.24  < 2e-16 ***
## genredancehall          -3.76501    0.34003  -11.07  < 2e-16 ***
## genredeath-metal         2.28882    0.35154    6.51  7.5e-11 ***
## genredeep-house         -0.39727    0.35301   -1.13  0.26042
## genredetroit-techno    -16.08585    0.57494  -27.98  < 2e-16 ***
## genredisco              -0.04258    0.35124   -0.12  0.90352
## genredrum-and-bass      -7.81871    0.37282  -20.97  < 2e-16 ***
## genredub                -1.80985    0.34028   -5.32  1.0e-07 ***
## genredubstep           -12.58236    0.53260  -23.62  < 2e-16 ***
## genreedm                 6.64925    0.39893   16.67  < 2e-16 ***
## genreelectro            15.03636    0.38898   38.66  < 2e-16 ***
## genreelectronic          5.79201    0.41849   13.84  < 2e-16 ***
## genreemo                 5.30725    0.32889   16.14  < 2e-16 ***
## genrefolk               15.27735    0.34334   44.50  < 2e-16 ***
## genreforro              -7.51087    0.33984  -22.10  < 2e-16 ***
## genrefrench              8.77723    0.33902   25.89  < 2e-16 ***
## genrefunk                7.84987    0.36020   21.79  < 2e-16 ***
## genregarage             -0.83018    0.34330   -2.42  0.01560 *
## genregerman              7.36794    0.34459   21.38  < 2e-16 ***
## genregospel              1.03582    0.32405    3.20  0.00139 **
## genregoth               -6.06938    0.34474  -17.61  < 2e-16 ***
## genregrindcore         -15.40708    0.38244  -40.29  < 2e-16 ***
## genregroove             -4.87543    0.36777  -13.26  < 2e-16 ***
## genreguitar             -5.17765    0.34688  -14.93  < 2e-16 ***
## genrehard-rock           2.42226    0.37180    6.51  7.3e-11 ***
## genrehardcore            7.88110    0.36052   21.86  < 2e-16 ***
## genrehardstyle          -6.13129    0.38456  -15.94  < 2e-16 ***
```

```
## genreheavy-metal       -14.85071    0.36685  -40.48  < 2e-16 ***
## genrehip-hop            26.41858    0.35981   73.42  < 2e-16 ***
## genrehouse               7.71991    0.52047   14.83  < 2e-16 ***
## genreindian             -7.37688    0.32954  -22.39  < 2e-16 ***
## genreindie-pop          16.82212    0.39722   42.35  < 2e-16 ***
## genreindustrial         -6.66544    0.36196  -18.41  < 2e-16 ***
## genrejazz               13.03350    0.35395   36.82  < 2e-16 ***
## genrek-pop               9.55826    0.32809   29.13  < 2e-16 ***
## genremetal              20.22007    0.46920   43.09  < 2e-16 ***
## genremetalcore          -1.62887    0.48518   -3.36  0.00079 ***
## genreminimal-techno     -8.81593    0.37427  -23.56  < 2e-16 ***
## genrenew-age            -3.27787    0.33604   -9.75  < 2e-16 ***
## genreopera              -9.54134    0.35379  -26.97  < 2e-16 ***
## genreparty             -10.57270    0.40702  -25.98  < 2e-16 ***
## genrepiano               3.00405    0.36366    8.26  < 2e-16 ***
## genrepop                37.46479    0.48091   77.90  < 2e-16 ***
## genrepop-film           -0.92984    0.34822   -2.67  0.00758 **
## genrepower-pop         -12.64263    0.34492  -36.65  < 2e-16 ***
## genreprogressive-house  -1.51572    0.41050   -3.69  0.00022 ***
## genrepsych-rock         -1.30002    0.36880   -3.52  0.00042 ***
## genrepunk               12.85365    0.49174   26.14  < 2e-16 ***
## genrepunk-rock           1.71508    0.44792    3.83  0.00013 ***
## genrerock               27.97208    0.64125   43.62  < 2e-16 ***
## genrerock-n-roll        -9.48483    0.35287  -26.88  < 2e-16 ***
## genreromance           -17.30067    0.47555  -36.38  < 2e-16 ***
## genresad                11.72513    0.49151   23.86  < 2e-16 ***
## genresalsa              -6.15707    0.34759  -17.71  < 2e-16 ***
## genresamba              -7.92695    0.33861  -23.41  < 2e-16 ***
## genresertanejo           1.80137    0.34232    5.26  1.4e-07 ***
## genreshow-tunes         -8.74079    0.39217  -22.29  < 2e-16 ***
## genresinger-songwriter   5.04359    0.35682   14.13  < 2e-16 ***
## genreska                -3.38968    0.36907   -9.18  < 2e-16 ***
## genresleep               1.08512    0.36698    2.96  0.00311 **
## genresongwriter          1.55863    1.43300    1.09  0.27674
## genresoul               11.94908    0.42284   28.26  < 2e-16 ***
## genrespanish             5.35406    0.33296   16.08  < 2e-16 ***
## genreswedish             2.05635    0.38609    5.33  1.0e-07 ***
## genretango             -14.90005    0.35395  -42.10  < 2e-16 ***
## genretechno             -2.93577    0.45679   -6.43  1.3e-10 ***
## genretrance             -1.53020    0.41398   -3.70  0.00022 ***
## genretrip-hop          -11.51838    0.38605  -29.84  < 2e-16 ***
## danceability             4.13108    0.25343   16.30  < 2e-16 ***
## energy                  -1.24329    0.25760   -4.83  1.4e-06 ***
## key1                     0.09459    0.13608    0.70  0.48697
## key2                     0.14908    0.13147    1.13  0.25683
## key3                     0.46318    0.19612    2.36  0.01819 *
## key4                     0.11913    0.14462    0.82  0.41010
## key5                     0.08508    0.14318    0.59  0.55236
## key6                     0.38212    0.15275    2.50  0.01237 *
## key7                    -0.12890    0.12738   -1.01  0.31158
## key8                     0.60441    0.15487    3.90  9.5e-05 ***
## key9                    -0.08379    0.13324   -0.63  0.52941
## key10                    0.14433    0.15298    0.94  0.34545
## key11                    0.15481    0.14625    1.06  0.28980
```

```
## loudness                 0.21211    0.01109   19.12  < 2e-16 ***
## modeminor                0.49478    0.06793    7.28  3.3e-13 ***
## speechiness             -1.84928    0.35623   -5.19  2.1e-07 ***
## acousticness             0.65422    0.15762    4.15  3.3e-05 ***
## instrumentalness        -2.38223    0.11618  -20.51  < 2e-16 ***
## liveness                -0.12283    0.17024   -0.72  0.47060
## valence                 -2.89828    0.16143  -17.95  < 2e-16 ***
## tempo                    0.00138    0.00111    1.23  0.21690
## time_signature1         -5.71186    1.03013   -5.54  2.9e-08 ***
## time_signature3         -5.45888    0.99631   -5.48  4.3e-08 ***
## time_signature4         -4.91070    0.99377   -4.94  7.8e-07 ***
## time_signature5         -5.50952    1.01391   -5.43  5.5e-08 ***
## duration_m              -0.37886    0.01674  -22.63  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.4 on 115774 degrees of freedom
## Multiple R-squared:  0.57,   Adjusted R-squared:  0.57
## F-statistic: 1.18e+03 on 130 and 115774 DF,  p-value: <2e-16
```

What we can see is: - p-value is very low - most variables are siginficant. Except some categorical keys. - RSS? - Adjusted R => Okay, but could be better.

Model seems decent, but we can do better.

Trying to find a smaller good model. Using both AIC or BIC running backward did not lead to smaller models.

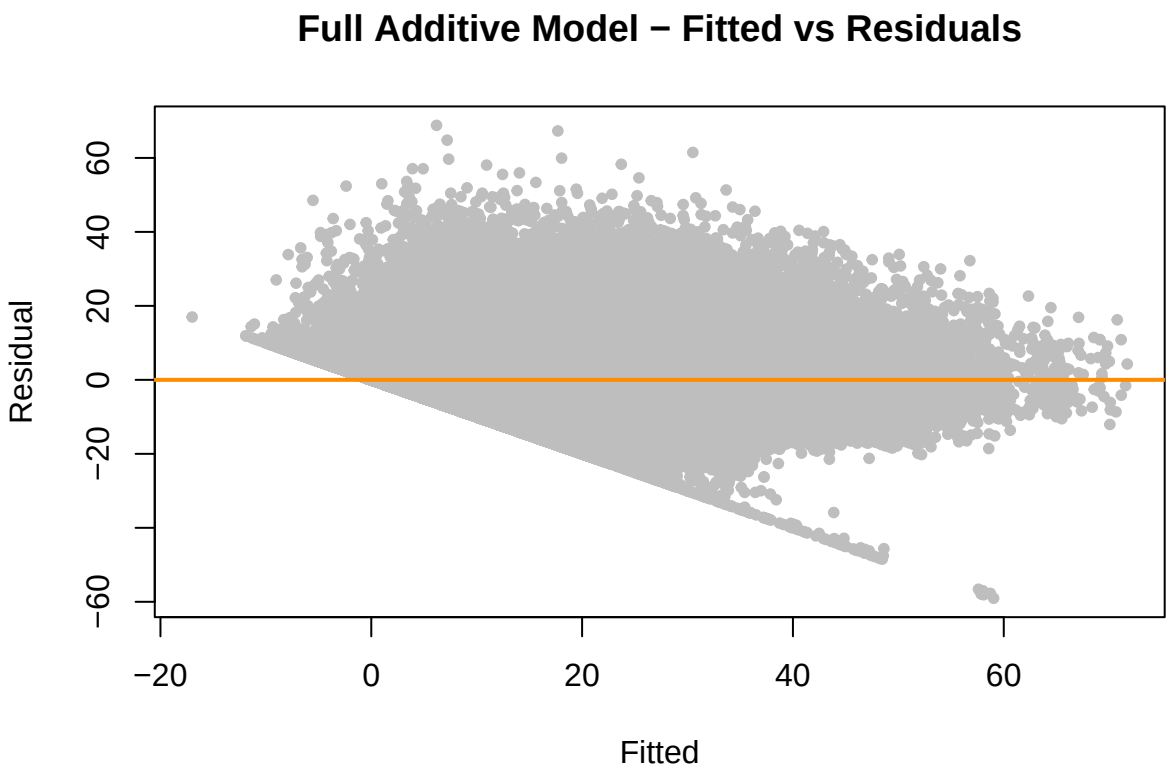
Let's have a look into the test results:

```
results = check_model(model_add_full, data_tst)
results
```

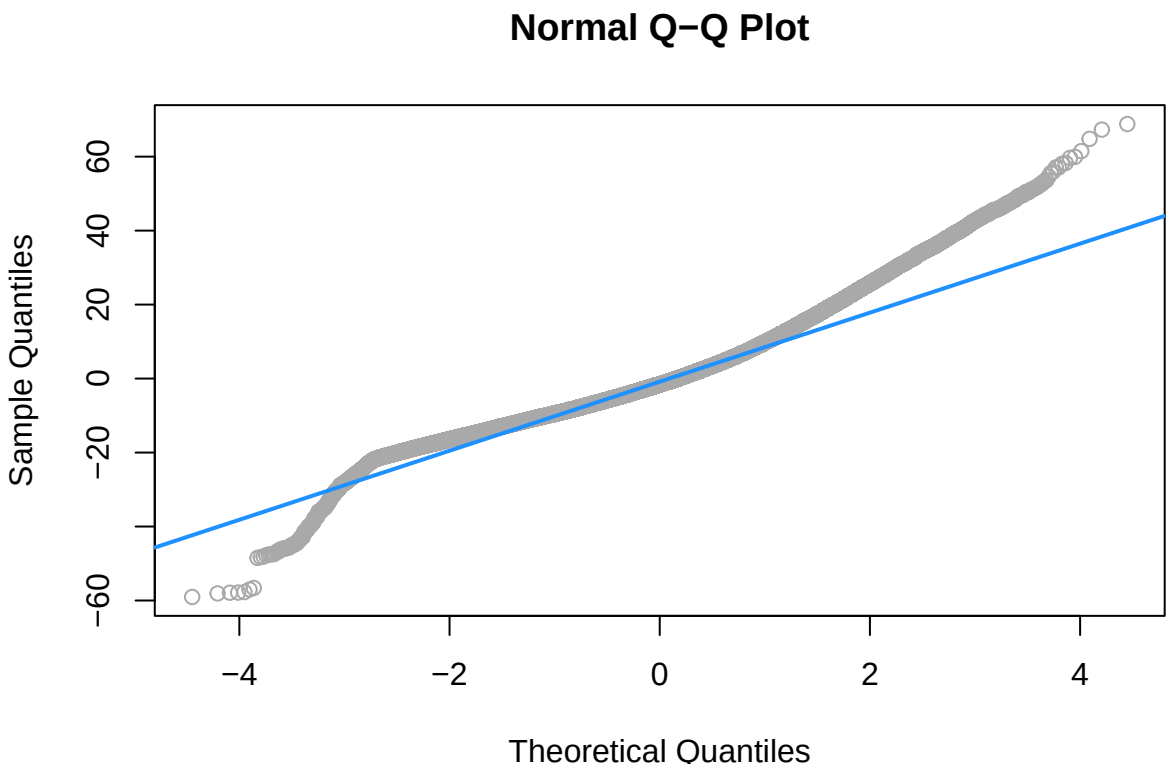
```
##      shapiro bptest high_leverage_count outliers_count influence_count
## BP 3.217e-37      0                2606           5575            5496
##    high_vif_count loocv_rmse adjusted_r_squared
## BP              5      10.43             0.5698
```

The test results make show that the normality and constant variance and normal assumption. The Plots Fitted vs Residuals underline that.

```
plot(fitted(model_add_full), resid(model_add_full), col = "grey", pch = 20,
     xlab = "Fitted", ylab = "Residual",
     main = "Full Additive Model - Fitted vs Residuals")
abline(h = 0, col = "darkorange", lwd = 2)
```




```
qqnorm(resid(model_add_full), col = "darkgrey")
qqline(resid(model_add_full), col = "dodgerblue", lwd = 2)
```



The model needs adjustments. Transformations could be helpful to fullfill the assumptions.

2. Smaller addtive models, but more explainable

In terms of finding a better explainable model, we leave out the year and the gerne.

```
model_add_noyear = lm(popularity ~ . -year, data = data_trn)
model_add_noyear_nogenre = lm(popularity ~ . -genre -year, data = data_trn)
model_add_nogenre = lm(popularity ~ . -genre, data = data_trn)
```

Let's plot the model's summary

```
summary(model_add_noyear)
```

```
##
## Call:
## lm(formula = popularity ~ . - year, data = data_trn)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -59.48  -8.21  -1.82   6.81  68.04
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              28.33941    1.17957   24.03  < 2e-16 ***
## genreafrobeat            -7.47974    0.43496  -17.20  < 2e-16 ***
## genrealt-rock            20.55754    0.37722   54.50  < 2e-16 ***
## genreambient              6.16510    0.38618   15.96  < 2e-16 ***
## genreblack-metal         -5.07420    0.38928  -13.03  < 2e-16 ***
## genreblues                4.98158    0.37868   13.16  < 2e-16 ***
## genrebreakbeat          -13.18303    0.43989  -29.97  < 2e-16 ***
## genrecantopop            -5.96595    0.40243  -14.82  < 2e-16 ***
## genrechicago-house      -12.62818    0.60472  -20.88  < 2e-16 ***
## genrechill                7.41760    0.38118   19.46  < 2e-16 ***
## genreclassical           11.58007    0.39475   29.33  < 2e-16 ***
## genreclub                -6.80286    0.40942  -16.62  < 2e-16 ***
## genrecomedy              -7.31787    0.46222  -15.83  < 2e-16 ***
## genrecountry             15.19842    0.38702   39.27  < 2e-16 ***
## genredance               23.46692    0.39234   59.81  < 2e-16 ***
## genredancehall           -4.10094    0.38812  -10.57  < 2e-16 ***
## genredeath-metal          1.67582    0.40122    4.18  3.0e-05 ***
## genredeep-house           0.30612    0.40290    0.76  0.44738
## genredetroit-techno     -13.17945    0.65593  -20.09  < 2e-16 ***
## genredisco                0.04114    0.40101    0.10  0.91828
## genredrum-and-bass       -7.20492    0.42558  -16.93  < 2e-16 ***
## genredub                 -2.33999    0.38848   -6.02  1.7e-09 ***
## genredubstep            -14.56842    0.60704  -24.00  < 2e-16 ***
## genreedm                  6.45635    0.45536   14.18  < 2e-16 ***
## genreelectro             13.09064    0.44384   29.49  < 2e-16 ***
## genreelectronic           6.29104    0.47777   13.17  < 2e-16 ***
## genreemo                  4.85381    0.37544   12.93  < 2e-16 ***
## genrefolk                15.97470    0.39196   40.76  < 2e-16 ***
## genreforro               -6.67356    0.38789  -17.20  < 2e-16 ***
## genrefrench               7.96734    0.38701   20.59  < 2e-16 ***
## genrefunk                 8.34252    0.41123   20.29  < 2e-16 ***
## genregarage              -0.66498    0.39191   -1.70  0.08974 .
## genregerman               7.63520    0.39339   19.41  < 2e-16 ***
## genregospel               1.66369    0.36993    4.50  6.9e-06 ***
```

```
## genregoth               -5.65866    0.39350  -14.38  < 2e-16 ***
## genregrindcore         -15.76204    0.43651  -36.11  < 2e-16 ***
## genregroove             -5.26606    0.41985  -12.54  < 2e-16 ***
## genreguitar             -6.01189    0.39597  -15.18  < 2e-16 ***
## genrehard-rock           3.11525    0.42443    7.34  2.2e-13 ***
## genrehardcore            7.29706    0.41155   17.73  < 2e-16 ***
## genrehardstyle          -4.22108    0.43883   -9.62  < 2e-16 ***
## genreheavy-metal       -13.60278    0.41872  -32.49  < 2e-16 ***
## genrehip-hop            26.54443    0.41081   64.62  < 2e-16 ***
## genrehouse               5.04464    0.59385    8.49  < 2e-16 ***
## genreindian             -6.33330    0.37617  -16.84  < 2e-16 ***
## genreindie-pop          16.81476    0.45341   37.09  < 2e-16 ***
## genreindustrial         -5.83133    0.41323  -14.11  < 2e-16 ***
## genrejazz               13.31734    0.40410   32.96  < 2e-16 ***
## genrek-pop               9.49605    0.37458   25.35  < 2e-16 ***
## genremetal              21.71747    0.53559   40.55  < 2e-16 ***
## genremetalcore          -2.07516    0.55381   -3.75  0.00018 ***
## genreminimal-techno     -8.04177    0.42723  -18.82  < 2e-16 ***
## genrenew-age            -3.14804    0.38362   -8.21  2.3e-16 ***
## genreopera              -8.91310    0.40380  -22.07  < 2e-16 ***
## genreparty              -9.72652    0.46461  -20.93  < 2e-16 ***
## genrepiano               4.34333    0.41510   10.46  < 2e-16 ***
## genrepop                37.13582    0.54897   67.65  < 2e-16 ***
## genrepop-film           -0.05991    0.39740   -0.15  0.88017
## genrepower-pop         -12.37122    0.39368  -31.42  < 2e-16 ***
## genreprogressive-house  -1.05304    0.46843   -2.25  0.02458 *
## genrepsych-rock         -0.69887    0.42094   -1.66  0.09686 .
## genrepunk               13.51078    0.56117   24.08  < 2e-16 ***
## genrepunk-rock           2.30075    0.51118    4.50  6.8e-06 ***
## genrerock               28.58645    0.73199   39.05  < 2e-16 ***
## genrerock-n-roll        -8.84461    0.40273  -21.96  < 2e-16 ***
## genreromance           -16.97063    0.54246  -31.28  < 2e-16 ***
## genresad                17.51150    0.55971   31.29  < 2e-16 ***
## genresalsa              -4.74502    0.39662  -11.96  < 2e-16 ***
## genresamba              -7.10591    0.38645  -18.39  < 2e-16 ***
## genresertanejo           1.98266    0.39069    5.07  3.9e-07 ***
## genreshow-tunes         -7.55477    0.44744  -16.88  < 2e-16 ***
## genresinger-songwriter   5.70952    0.40719   14.02  < 2e-16 ***
## genreska                -2.50017    0.42121   -5.94  2.9e-09 ***
## genresleep               2.51420    0.41869    6.00  1.9e-09 ***
## genresongwriter         -4.88451    1.62673   -3.00  0.00268 **
## genresoul               12.82605    0.48259   26.58  < 2e-16 ***
## genrespanish             5.16168    0.38000   13.58  < 2e-16 ***
## genreswedish             1.84675    0.44068    4.19  2.8e-05 ***
## genretango             -14.68761    0.40392  -36.36  < 2e-16 ***
## genretechno             -2.92184    0.52137   -5.60  2.1e-08 ***
## genretrance             -2.07728    0.47235   -4.40  1.1e-05 ***
## genretrip-hop          -10.30490    0.44047  -23.40  < 2e-16 ***
## danceability             8.07750    0.28821   28.03  < 2e-16 ***
## energy                  -1.13499    0.29406   -3.86  0.00011 ***
## key1                     0.38549    0.15534    2.48  0.01308 *
## key2                    -0.06889    0.15009   -0.46  0.64623
## key3                     0.60787    0.22391    2.71  0.00663 **
## key4                    -0.08407    0.16511   -0.51  0.61060
```

```
## key5                    0.20879    0.16345    1.28  0.20147
## key6                    0.58671    0.17439    3.36  0.00077 ***
## key7                   -0.38010    0.14543   -2.61  0.00896 **
## key8                    0.99577    0.17679    5.63  1.8e-08 ***
## key9                   -0.41479    0.15210   -2.73  0.00639 **
## key10                   0.09451    0.17465    0.54  0.58843
## key11                   0.23741    0.16696    1.42  0.15504
## loudness                0.29242    0.01262   23.17  < 2e-16 ***
## modeminor               0.82611    0.07753   10.66  < 2e-16 ***
## speechiness            -0.54851    0.40659   -1.35  0.17733
## acousticness            0.77666    0.17994    4.32  1.6e-05 ***
## instrumentalness       -2.10215    0.13258  -15.86  < 2e-16 ***
## liveness               -0.45047    0.19432   -2.32  0.02044 *
## valence                -7.67127    0.18181  -42.19  < 2e-16 ***
## tempo                   0.00573    0.00127    4.51  6.6e-06 ***
## time_signature1        -6.63081    1.17589   -5.64  1.7e-08 ***
## time_signature3        -6.55362    1.13724   -5.76  8.3e-09 ***
## time_signature4        -5.80790    1.13434   -5.12  3.1e-07 ***
## time_signature5        -6.15193    1.15730   -5.32  1.1e-07 ***
## duration_m             -0.80754    0.01892  -42.69  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.9 on 115797 degrees of freedom
## Multiple R-squared:  0.44,   Adjusted R-squared:  0.439
## F-statistic:  849 on 107 and 115797 DF,  p-value: <2e-16
```

```
summary(model_add_noyear_nogenre)
```

```
##
## Call:
## lm(formula = popularity ~ . - genre - year, data = data_trn)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -32.73 -12.35  -2.61  10.10  70.59
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      33.96902    1.46628   23.17  < 2e-16 ***
## danceability     10.69407    0.31159   34.32  < 2e-16 ***
## energy           -7.83096    0.35307  -22.18  < 2e-16 ***
## key1              0.63738    0.19873    3.21   0.0013 **
## key2             -0.15999    0.19231   -0.83   0.4054
## key3              1.24677    0.28687    4.35  1.4e-05 ***
## key4             -0.01144    0.21144   -0.05   0.9568
## key5              0.42237    0.20956    2.02   0.0439 *
## key6              1.07901    0.22338    4.83  1.4e-06 ***
## key7             -0.75025    0.18645   -4.02  5.7e-05 ***
## key8              1.57541    0.22646    6.96  3.5e-12 ***
## key9             -0.63039    0.19482   -3.24   0.0012 **
## key10             0.30888    0.22376    1.38   0.1675
## key11             0.49057    0.21384    2.29   0.0218 *
## loudness          0.30838    0.01457   21.16  < 2e-16 ***
## modeminor         0.69466    0.09781    7.10  1.2e-12 ***
```

```
## speechiness      -4.64145    0.39197  -11.84  < 2e-16 ***
## acousticness     -3.21679    0.20564  -15.64  < 2e-16 ***
## instrumentalness -6.21284    0.14796  -41.99  < 2e-16 ***
## liveness         -2.77485    0.24363  -11.39  < 2e-16 ***
## valence          -9.33052    0.21482  -43.43  < 2e-16 ***
## tempo             0.00215    0.00159    1.35   0.1773
## time_signature1  -3.44926    1.49780   -2.30   0.0213 *
## time_signature3  -3.12957    1.44655   -2.16   0.0305 *
## time_signature4  -1.96210    1.44216   -1.36   0.1737
## time_signature5  -3.12074    1.47367   -2.12   0.0342 *
## duration_m       -1.04424    0.02272  -45.95  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.3 on 115878 degrees of freedom
## Multiple R-squared:  0.0767, Adjusted R-squared:  0.0765
## F-statistic:  370 on 26 and 115878 DF,  p-value: <2e-16
```

```
summary(model_add_nogenre)
```

```
##
## Call:
## lm(formula = popularity ~ . - genre, data = data_trn)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -36.27 -10.57  -2.15   8.94  69.56
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      24.88363    1.38812   17.93  < 2e-16 ***
## year2001          0.22206    0.30622    0.73   0.4683
## year2002          0.70210    0.30268    2.32   0.0204 *
## year2003          1.28161    0.30526    4.20  2.7e-05 ***
## year2004          1.28178    0.30460    4.21  2.6e-05 ***
## year2005          2.24709    0.30274    7.42  1.2e-13 ***
## year2006          1.93748    0.29917    6.48  9.4e-11 ***
## year2007          2.69106    0.30014    8.97  < 2e-16 ***
## year2008          2.54978    0.29730    8.58  < 2e-16 ***
## year2009          3.35161    0.29853   11.23  < 2e-16 ***
## year2010          3.86176    0.29697   13.00  < 2e-16 ***
## year2011          4.90177    0.29853   16.42  < 2e-16 ***
## year2012          5.19501    0.28689   18.11  < 2e-16 ***
## year2013          6.33876    0.28986   21.87  < 2e-16 ***
## year2014          7.26229    0.29031   25.02  < 2e-16 ***
## year2015          8.46380    0.29232   28.95  < 2e-16 ***
## year2016         10.45041    0.31002   33.71  < 2e-16 ***
## year2017         11.75520    0.28644   41.04  < 2e-16 ***
## year2018         12.58689    0.28501   44.16  < 2e-16 ***
## year2019         13.95107    0.28610   48.76  < 2e-16 ***
## year2020         14.78045    0.28825   51.28  < 2e-16 ***
## year2021         16.08653    0.29086   55.31  < 2e-16 ***
## year2022         19.25554    0.29016   66.36  < 2e-16 ***
## year2023          8.35803    0.31303   26.70  < 2e-16 ***
## danceability      6.69596    0.29283   22.87  < 2e-16 ***
```

```
## energy           -8.31881    0.32990  -25.22  < 2e-16 ***
## key1              0.36704    0.18568    1.98   0.0481 *
## key2              0.04308    0.17968    0.24   0.8105
## key3              1.08019    0.26801    4.03  5.6e-05 ***
## key4              0.19228    0.19754    0.97   0.3304
## key5              0.27772    0.19579    1.42   0.1561
## key6              0.91342    0.20870    4.38  1.2e-05 ***
## key7             -0.52750    0.17419   -3.03   0.0025 **
## key8              1.19971    0.21160    5.67  1.4e-08 ***
## key9             -0.31040    0.18204   -1.71   0.0882 .
## key10             0.28971    0.20905    1.39   0.1658
## key11             0.41509    0.19979    2.08   0.0377 *
## loudness          0.25912    0.01365   18.98  < 2e-16 ***
## modeminor         0.41555    0.09141    4.55  5.5e-06 ***
## speechiness      -5.27570    0.36633  -14.40  < 2e-16 ***
## acousticness     -3.54697    0.19216  -18.46  < 2e-16 ***
## instrumentalness -6.38454    0.13834  -46.15  < 2e-16 ***
## liveness         -2.46093    0.22767  -10.81  < 2e-16 ***
## valence          -5.12328    0.20354  -25.17  < 2e-16 ***
## tempo            -0.00246    0.00149   -1.65   0.0992 .
## time_signature1  -2.16548    1.39958   -1.55   0.1218
## time_signature3  -1.67811    1.35176   -1.24   0.2145
## time_signature4  -0.60535    1.34765   -0.45   0.6533
## time_signature5  -2.02204    1.37712   -1.47   0.1420
## duration_m       -0.66555    0.02145  -31.03  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.3 on 115855 degrees of freedom
## Multiple R-squared:  0.194,  Adjusted R-squared:  0.194
## F-statistic:  571 on 49 and 115855 DF,  p-value: <2e-16
```

Let's do some first tests.

```
check_model(model_add_noyear, data_tst)
```

```
##     shapiro bptest high_leverage_count outliers_count influence_count
## BP 1.63e-30      0                3567           5565            5449
##    high_vif_count loocv_rmse adjusted_r_squared
## BP              4      11.91             0.4391
```

```
check_model(model_add_noyear_nogenre, data_tst)
```

```
##      shapiro bptest high_leverage_count outliers_count influence_count
## BP 1.643e-36      0                6052           4584            4292
##    high_vif_count loocv_rmse adjusted_r_squared
## BP              1      15.28            0.07651
```

```
check_model(model_add_nogenre, data_tst)
```

```
##      shapiro bptest high_leverage_count outliers_count influence_count
## BP 8.318e-30      0                2890           4989            4601
##    high_vif_count loocv_rmse adjusted_r_squared
## BP              2      14.27             0.1941
```

‘‘‘