Top\_song\_analysis.R

Apurva Sarode

2020-04-16

# split data into train and test  
set.seed(101)  
sample\_n(Data,10)

## X title artist  
## 1 434 Higher Carly Rae Jepsen  
## 2 96 Castle Walls (feat. Christina Aguilera) T.I.  
## 3 211 All of Me John Legend  
## 4 447 Treat You Better Shawn Mendes  
## 5 354 Yesterday (feat. Bebe Rexha) David Guetta  
## 6 318 Fireball (feat. John Ryan) Pitbull  
## 7 248 Can't Remember to Forget You (feat. Rihanna) Shakira  
## 8 132 Lights - Single Version Ellie Goulding  
## 9 526 These Days Rudimental  
## 10 355 Time of Our Lives Pitbull  
## Genre year bpm Duration Energy Dancebility Loudness Valence  
## 1 canadian pop 2016 114 234 87 65 77 44  
## 2 atl hip hop 2011 80 329 86 45 77 58  
## 3 neo mellow 2014 120 270 26 42 62 33  
## 4 canadian pop 2017 83 188 82 44 85 75  
## 5 dance pop 2015 128 243 78 57 85 28  
## 6 dance pop 2015 123 235 94 69 77 79  
## 7 colombian pop 2014 138 207 81 69 85 82  
## 8 dance pop 2012 120 211 80 68 69 78  
## 9 dance pop 2018 92 211 81 65 85 55  
## 10 dance pop 2015 124 229 80 72 69 72  
## Acoustiveness Popularity Rating  
## 1 1 46 Below Average  
## 2 7 49 Below Average  
## 3 92 86 Above Average  
## 4 11 84 Above Average  
## 5 2 46 Below Average  
## 6 9 67 Above Average  
## 7 12 62 Below Average  
## 8 3 65 Below Average  
## 9 19 80 Above Average  
## 10 9 45 Below Average

# Lets take a sample of 75/25 like before. Dplyr preserves class.  
training\_sample <- sample(c(TRUE, FALSE), nrow(Data), replace = T, prob = c(0.75,0.25))  
train <- Data[training\_sample, ]  
test <- Data[!training\_sample, ]  
  
fit<- lm(Popularity~Duration+Energy+Dancebility+Loudness+Valence+Acoustiveness,data = train)  
summary(fit)

##   
## Call:  
## lm(formula = Popularity ~ Duration + Energy + Dancebility + Loudness +   
## Valence + Acoustiveness, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -59.251 -6.735 2.563 8.765 28.476   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 77.30804 7.60943 10.159 < 2e-16 \*\*\*  
## Duration -0.04579 0.01992 -2.298 0.02201 \*   
## Energy -0.24190 0.06325 -3.825 0.00015 \*\*\*  
## Dancebility 0.06985 0.05660 1.234 0.21780   
## Loudness 0.17117 0.06683 2.561 0.01076 \*   
## Valence 0.01038 0.03569 0.291 0.77137   
## Acoustiveness -0.04579 0.03942 -1.162 0.24603   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 13.28 on 443 degrees of freedom  
## Multiple R-squared: 0.05339, Adjusted R-squared: 0.04057   
## F-statistic: 4.164 on 6 and 443 DF, p-value: 0.0004379

coefficients(fit)

## (Intercept) Duration Energy Dancebility Loudness   
## 77.30803737 -0.04579099 -0.24189821 0.06984747 0.17117232   
## Valence Acoustiveness   
## 0.01037873 -0.04579239

confint(fit,level=0.95)

## 2.5 % 97.5 %  
## (Intercept) 62.35296256 92.263112189  
## Duration -0.08494685 -0.006635123  
## Energy -0.36619886 -0.117597558  
## Dancebility -0.04138104 0.181075977  
## Loudness 0.03982101 0.302523625  
## Valence -0.05977312 0.080530568  
## Acoustiveness -0.12327104 0.031686265

# Predicted Values  
fitted(fit)

## 1 2 3 5 6 9 10 11   
## 65.03212 60.75252 69.16562 64.87399 65.36194 67.81960 69.23588 64.73262   
## 12 13 14 15 16 17 19 21   
## 63.19640 66.10702 65.41000 60.13229 63.53354 73.51360 60.80654 65.67944   
## 23 24 26 27 28 29 32 33   
## 70.06106 66.65067 63.45439 64.50433 64.94589 66.24588 67.10214 70.27231   
## 34 35 36 37 39 40 41 42   
## 67.32005 62.58122 72.39529 66.77912 60.59766 62.93427 64.98225 70.38174   
## 45 48 49 50 53 55 58 59   
## 65.70340 67.28001 58.04438 65.53398 63.13793 66.70695 64.96811 68.67403   
## 60 61 63 67 69 70 71 72   
## 67.75586 64.20582 59.57734 67.06304 62.98391 63.92387 63.75655 66.27112   
## 73 75 76 77 79 80 81 82   
## 66.37311 67.27358 60.21865 63.93898 70.70945 70.20433 63.25723 63.11920   
## 83 84 85 87 88 89 90 93   
## 64.44006 67.12414 68.97067 66.80702 67.48456 69.31305 65.97074 63.59041   
## 94 95 100 101 102 103 104 105   
## 68.53734 58.04438 59.88544 66.97356 64.47417 66.25058 65.71060 70.02836   
## 106 107 109 110 111 112 113 114   
## 63.99479 67.41618 71.51378 71.14027 73.39529 69.07568 68.96519 63.34190   
## 115 116 117 118 119 121 122 124   
## 66.57558 72.72257 68.54396 65.90522 67.58840 67.16670 62.32912 63.89161   
## 126 128 129 130 131 132 133 134   
## 62.91605 67.84408 64.99585 63.14437 65.52696 65.47455 67.32679 65.17172   
## 135 137 138 139 141 144 146 147   
## 67.40571 58.77959 64.64098 63.30665 67.79114 68.25660 60.45939 71.02708   
## 149 151 152 153 155 156 162 165   
## 70.69444 68.96519 69.63721 65.34434 66.08149 65.53852 61.27881 67.99856   
## 166 167 168 169 170 171 172 173   
## 67.18202 63.65691 70.15179 63.73902 65.47897 67.91944 67.54461 69.61840   
## 174 175 176 178 179 181 182 183   
## 67.99203 68.97210 64.28842 69.06685 68.28241 67.17132 66.00561 66.74653   
## 184 185 186 188 190 192 193 195   
## 66.74290 70.03682 60.89817 66.52200 64.06334 64.44808 62.63913 63.20804   
## 196 197 198 199 200 202 203 204   
## 64.05120 67.01580 66.83454 63.04535 68.59397 64.15926 67.94388 66.87029   
## 205 206 207 208 210 213 214 215   
## 66.07203 65.52374 66.97099 63.48534 71.45602 67.96138 66.34915 69.85942   
## 216 217 218 219 220 222 223 225   
## 64.12520 69.23432 72.46990 65.29850 67.18331 69.52988 61.67985 65.50235   
## 226 227 228 229 230 231 232 233   
## 66.02736 63.87692 63.50676 71.62120 70.00636 67.31782 68.50983 64.18375   
## 234 235 237 238 241 242 243 244   
## 69.35562 64.07675 64.95583 68.88792 69.92476 65.00829 69.38333 66.41935   
## 245 246 247 249 250 253 254 256   
## 64.91903 67.90622 66.69687 64.35974 66.71080 63.71533 70.38632 69.75881   
## 258 260 261 262 263 264 265 266   
## 66.98267 68.38841 67.33286 70.23092 67.78664 66.14992 65.70202 69.22865   
## 267 268 269 270 271 272 273 275   
## 69.15306 67.16891 64.85919 67.72767 64.99465 63.92701 68.86871 62.54164   
## 277 278 279 281 282 283 284 285   
## 67.38540 66.23538 69.23263 69.53922 64.70373 66.39748 66.04863 67.36237   
## 286 287 288 290 291 292 293 295   
## 62.93897 64.39058 69.41272 69.58496 63.01680 68.63433 66.82189 69.28824   
## 296 299 300 301 302 303 304 305   
## 65.72886 65.58391 70.21894 69.68819 70.24119 69.80723 62.13519 67.26833   
## 306 307 309 310 311 312 315 316   
## 65.06191 66.11904 72.14441 65.89408 68.23612 66.37531 62.21626 63.18731   
## 317 318 319 320 321 322 323 325   
## 61.85991 66.41487 67.99278 63.92701 62.43329 68.34119 65.45921 65.51665   
## 326 330 331 332 334 335 336 337   
## 64.63557 59.31671 69.08333 67.82801 65.66732 67.10816 66.33123 67.82406   
## 338 340 341 342 344 345 346 347   
## 67.91137 64.80121 66.85057 66.54911 66.25150 72.93186 67.94989 64.16263   
## 348 349 350 351 353 354 355 356   
## 69.32594 62.44551 60.38480 66.04274 67.25792 67.67927 64.55540 69.59674

residuals(fit)

## 1 2 3 5 6   
## 17.967877695 21.247476289 10.834383439 13.126013784 11.638064852   
## 9 10 11 12 13   
## 8.180400662 3.764117708 8.267376146 9.803599785 6.892976034   
## 14 15 16 17 19   
## 7.589995471 11.867714749 8.466461478 -2.513603969 8.193463098   
## 21 23 24 26 27   
## 2.320560224 -4.061056001 -1.650667140 1.545609486 -0.504334957   
## 28 29 32 33 34   
## -1.945888200 -3.245881509 -5.102143621 -8.272305751 -5.320052124   
## 35 36 37 39 40   
## -0.581221665 -11.395288844 -5.779117843 -1.597660284 -4.934273928   
## 41 42 45 48 49   
## -6.982250591 -13.381742736 -9.703403052 -15.280009046 -9.044380090   
## 50 53 55 58 59   
## -32.533983222 15.862072693 9.293045514 11.031890237 6.325973853   
## 60 61 63 67 69   
## 7.244139187 9.794176566 13.422655889 4.936962509 9.016085357   
## 70 71 72 73 75   
## 7.076125161 5.243449851 2.728881386 2.626889983 -0.273576088   
## 76 77 79 80 81   
## 6.781354767 2.061024319 -6.709454148 -6.204334556 0.742771637   
## 82 83 84 85 87   
## -0.119200154 -1.440061932 -4.124135892 -7.970671323 -6.807018524   
## 88 89 90 93 94   
## -8.484561795 -10.313047615 -7.970736263 -11.590408022 -18.537341518   
## 95 100 101 102 103   
## -9.044380090 -31.885439333 -39.973564772 -39.474170817 -59.250576685   
## 104 105 106 107 109   
## 14.289403160 8.971638183 15.005210351 11.583820362 5.486215789   
## 110 111 112 113 114   
## 4.859729737 2.604713720 6.924317371 6.034809160 10.658097145   
## 115 116 117 118 119   
## 7.424420896 0.277429370 4.456035638 7.094776701 4.411595913   
## 121 122 124 126 128   
## 4.833304637 9.670884240 7.108393216 5.083948997 -0.844076419   
## 129 130 131 132 133   
## 1.004149650 1.855628580 -0.526963284 -2.474548501 -6.326789109   
## 134 135 137 138 139   
## -6.171716900 -10.405710909 -20.779587327 23.359023534 21.693352913   
## 141 144 146 147 149   
## 13.208857625 9.743397962 16.540605166 5.972922552 5.305559829   
## 151 152 153 155 156   
## 6.034809160 5.362790046 9.655659199 7.918509332 8.461477936   
## 162 165 166 167 168   
## 9.721193218 2.001444848 2.817984366 6.343094806 -1.151786895   
## 169 170 171 172 173   
## 5.260975011 2.521029613 0.080562588 0.455393637 -2.618401681   
## 174 175 176 178 179   
## -0.992031303 -2.972099207 0.711584496 -6.066853669 -6.282408775   
#Anova Table  
anova(fit)

## Analysis of Variance Table  
##   
## Response: Popularity  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Duration 1 1082 1082.00 6.1371 0.013609 \*   
## Energy 1 1324 1323.69 7.5080 0.006391 \*\*  
## Dancebility 1 691 691.32 3.9212 0.048300 \*   
## Loudness 1 1064 1063.83 6.0341 0.014415 \*   
## Valence 1 6 6.33 0.0359 0.849762   
## Acoustiveness 1 238 237.88 1.3493 0.246033   
## Residuals 443 78103 176.30   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

vcov(fit)

## (Intercept) Duration Energy Dancebility  
## (Intercept) 57.90348536 -1.027372e-01 -1.359695e-01 -2.146779e-01  
## Duration -0.10273721 3.969366e-04 1.347027e-05 5.920213e-05  
## Energy -0.13596949 1.347027e-05 4.000132e-03 6.807489e-04  
## Dancebility -0.21467789 5.920213e-05 6.807489e-04 3.203021e-03  
## Loudness -0.14355265 3.413188e-05 -2.418697e-03 -1.652439e-04  
## Valence 0.02097914 1.302745e-04 -6.396639e-04 -9.569402e-04  
## Acoustiveness -0.10128607 -4.300423e-05 1.189534e-03 5.288583e-04  
## Loudness Valence Acoustiveness  
## (Intercept) -1.435526e-01 0.0209791364 -1.012861e-01  
## Duration 3.413188e-05 0.0001302745 -4.300423e-05  
## Energy -2.418697e-03 -0.0006396639 1.189534e-03  
## Dancebility -1.652439e-04 -0.0009569402 5.288583e-04  
## Loudness 4.466798e-03 -0.0001084142 -3.087668e-04  
## Valence -1.084142e-04 0.0012741064 -1.235214e-04  
## Acoustiveness -3.087668e-04 -0.0001235214 1.554145e-03

step <- stepAIC(fit, direction="both")

## Start: AIC=2334.44  
## Popularity ~ Duration + Energy + Dancebility + Loudness + Valence +   
## Acoustiveness  
##   
## Df Sum of Sq RSS AIC  
## - Valence 1 14.91 78118 2332.5  
## - Acoustiveness 1 237.88 78341 2333.8  
## - Dancebility 1 268.54 78371 2334.0  
## <none> 78103 2334.4  
## - Duration 1 931.32 79034 2337.8  
## - Loudness 1 1156.47 79259 2339.1  
## - Energy 1 2579.01 80682 2347.1  
##   
## Step: AIC=2332.53  
## Popularity ~ Duration + Energy + Dancebility + Loudness + Acoustiveness  
##   
## Df Sum of Sq RSS AIC  
## - Acoustiveness 1 229.31 78347 2331.8  
## <none> 78118 2332.5  
## - Dancebility 1 427.82 78545 2333.0  
## + Valence 1 14.91 78103 2334.4  
## - Duration 1 1008.85 79126 2336.3  
## - Loudness 1 1170.85 79288 2337.2  
## - Energy 1 2684.63 80802 2345.7  
##   
## Step: AIC=2331.84  
## Popularity ~ Duration + Energy + Dancebility + Loudness  
##   
## Df Sum of Sq RSS AIC  
## <none> 78347 2331.8  
## + Acoustiveness 1 229.31 78118 2332.5  
## - Dancebility 1 608.99 78956 2333.3  
## + Valence 1 6.33 78341 2333.8  
## - Duration 1 1048.83 79396 2335.8  
## - Loudness 1 1063.83 79411 2335.9  
## - Energy 1 2568.80 80916 2344.4

step$anova # display results

## Stepwise Model Path   
## Analysis of Deviance Table  
##   
## Initial Model:  
## Popularity ~ Duration + Energy + Dancebility + Loudness + Valence +   
## Acoustiveness  
##   
## Final Model:  
## Popularity ~ Duration + Energy + Dancebility + Loudness  
##   
##   
## Step Df Deviance Resid. Df Resid. Dev AIC  
## 1 443 78102.65 2334.439  
## 2 - Valence 1 14.90543 444 78117.55 2332.525  
## 3 - Acoustiveness 1 229.30739 445 78346.86 2331.844

fit6 <- lm(Popularity~Energy+Loudness+Duration+Dancebility,data = train)  
summary(fit6)

##   
## Call:  
## lm(formula = Popularity ~ Energy + Loudness + Duration + Dancebility,   
## data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -61.494 -6.552 2.380 9.000 27.419   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 74.25476 7.14840 10.388 < 2e-16 \*\*\*  
## Energy -0.20394 0.05339 -3.820 0.000153 \*\*\*  
## Loudness 0.16278 0.06622 2.458 0.014346 \*   
## Duration -0.04773 0.01956 -2.441 0.015046 \*   
## Dancebility 0.09031 0.04856 1.860 0.063568 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 13.27 on 445 degrees of freedom  
## Multiple R-squared: 0.05043, Adjusted R-squared: 0.04189   
## F-statistic: 5.908 on 4 and 445 DF, p-value: 0.0001221

attach(Data)

## The following object is masked \_by\_ .GlobalEnv:  
##   
## Rating

fc= predict.lm(fit6,data.frame(Duration=189,Energy=32,Loudness = 62,Dancebility=64))  
fc

## 1   
## 74.579

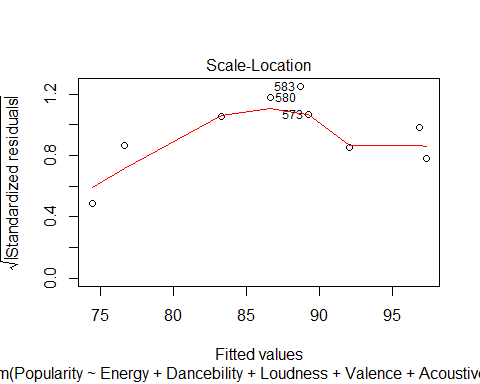
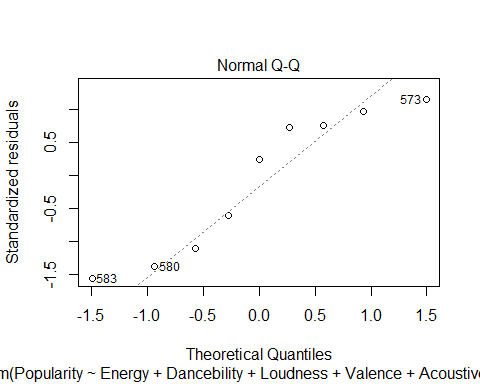
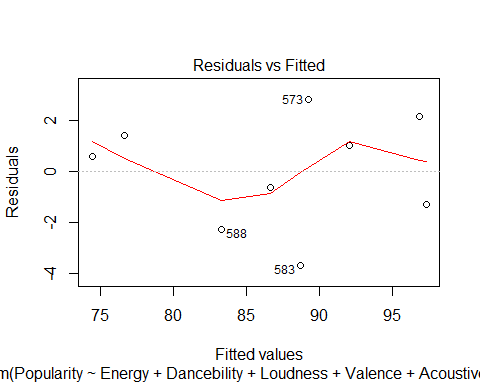
d\_g = Data[Data$Genre == 'pop' & Data$year == c(2019),c(4,7:13)]  
ft\_g = lm(Popularity~Energy+Dancebility+Loudness+Valence+Acoustiveness, data= d\_g)  
summary(ft\_g)

##   
## Call:  
## lm(formula = Popularity ~ Energy + Dancebility + Loudness + Valence +   
## Acoustiveness, data = d\_g)  
##   
## Residuals:  
## 568 570 572 573 580 583 588 591 595   
## 2.1587 -1.2913 0.9937 2.8030 -0.6407 -3.6889 -2.2917 1.3775 0.5796   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 141.77347 18.35008 7.726 0.00451 \*\*  
## Energy -0.87601 0.17643 -4.965 0.01569 \*   
## Dancebility -0.15562 0.11479 -1.356 0.26823   
## Loudness 0.27513 0.19323 1.424 0.24969   
## Valence 0.00154 0.09881 0.016 0.98854   
## Acoustiveness -0.26450 0.09281 -2.850 0.06511 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.496 on 3 degrees of freedom  
## Multiple R-squared: 0.9335, Adjusted R-squared: 0.8227   
## F-statistic: 8.423 on 5 and 3 DF, p-value: 0.05478

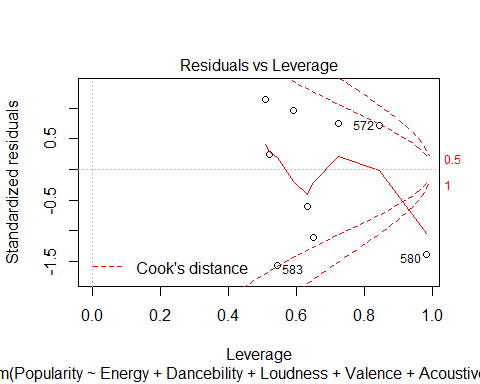
fc\_g= predict.lm(ft\_g,data.frame(Acoustiveness=32,Loudness = 62,Dancebility=64,Valence=59,Energy=70))  
fc\_g

## 1   
## 79.17777

#diagnostic plots  
plot(ft\_g)



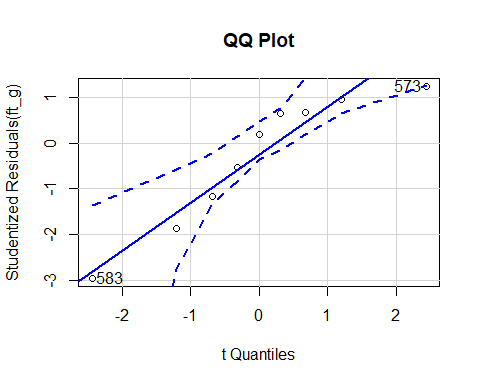
## Warning in sqrt(crit \* p \* (1 - hh)/hh): NaNs produced  
  
## Warning in sqrt(crit \* p \* (1 - hh)/hh): NaNs produced



# Assessing Outliers  
outlierTest(ft\_g)

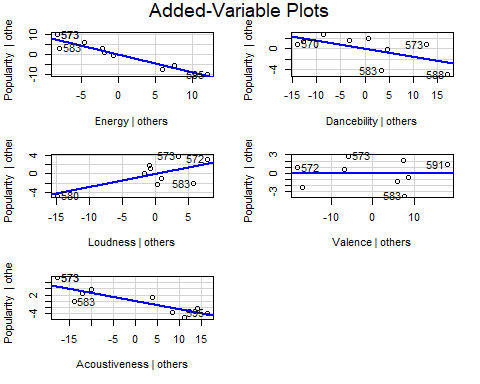
## No Studentized residuals with Bonferroni p < 0.05  
## Largest |rstudent|:  
## rstudent unadjusted p-value Bonferroni p  
## 583 -2.961341 0.097619 0.87857

qqPlot(ft\_g, main="QQ Plot")



## 573 583   
## 4 6

# added variable plots  
avPlots(ft\_g)



# distribution of studentized residuals  
sresid <- studres(ft\_g)  
hist(sresid, freq=FALSE,  
 main="Distribution of Studentized Residuals")  
xfit<-seq(min(sresid),max(sresid),length=40)  
yfit<-dnorm(xfit)  
lines(xfit, yfit)

