Lab 07: What makes a song more positive?

The A Team: Naomi Rubin, Jwalin Patel, Annie Sawers, Alex Williams, JM Stroh

3-23-21

Load packages & data

```
library(tidyverse)
library(broom)
library(knitr)
library(rms)
```

```
spotify <- read_csv("data/spotify-popular.csv") %>%
  mutate(key = factor(key),
         mode = factor(mode))
```

Exercise 1

```
full_model <- lm(valence ~ danceability + energy + key + loudness +
                   mode + speechiness + acousticness + instrumentalness +
                   liveness + tempo + duration_ms + playlist_genre,
                 data=spotify)

tidy(full_model)%>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-0.398	0.149	-2.678	0.008
danceability	0.700	0.071	9.871	0.000
energy	0.685	0.082	8.335	0.000
key1	0.006	0.035	0.164	0.869
key2	0.054	0.040	1.357	0.176
key3	0.043	0.051	0.847	0.398
key4	-0.019	0.040	-0.471	0.638
key5	0.038	0.038	0.983	0.326
key6	0.036	0.040	0.911	0.363
key7	-0.004	0.039	-0.102	0.919
key8	0.009	0.040	0.237	0.813
key9	0.018	0.039	0.463	0.643
key10	0.034	0.040	0.865	0.388
key11	0.047	0.038	1.243	0.214
loudness	-0.004	0.005	-0.861	0.389
mode1	0.015	0.018	0.850	0.396
speechiness	-0.047	0.086	-0.544	0.586
acousticness	0.130	0.043	3.007	0.003
instrumentalness	-0.132	0.158	-0.839	0.402

term	estimate	std.error	statistic	p.value
liveness	-0.052	0.068	-0.761	0.447
tempo	0.000	0.000	1.217	0.224
duration_ms	0.000	0.000	-0.416	0.677
playlist_genrelatin	-0.075	0.084	-0.893	0.372
playlist_genrepop	-0.110	0.081	-1.350	0.178
playlist_genrer&b	-0.124	0.085	-1.449	0.148
playlist_genrerap	-0.156	0.082	-1.896	0.059
playlist_genrerock	-0.045	0.090	-0.502	0.616

```
int_only_model <- lm(valence ~ 1, data = spotify)
tidy(int_only_model)%>%
  kable(digits=3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.51	0.01	51.272	0

Exercise 2

```
backward_aic <- step(full_model, direction="backward", results = "hide")
```

```
tidy(backward_aic)%>%
  kable(digits=3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-0.314	0.103	-3.058	0.002
danceability	0.673	0.066	10.232	0.000
energy	0.655	0.059	11.133	0.000
acousticness	0.134	0.042	3.198	0.001
playlist_genrelatin	-0.052	0.077	-0.674	0.500
playlist_genrepop	-0.089	0.075	-1.182	0.238
playlist_genrer&b	-0.108	0.080	-1.359	0.175
playlist_genrerap	-0.135	0.077	-1.753	0.080
playlist_genrerock	-0.028	0.083	-0.339	0.735

Exercise 3

```
## number of observations
n <- nrow(spotify)
```

```
backward_bic <- step(full_model, direction="backward", k=log(n), results = "hide")
```

```
tidy(backward_bic)%>%
  kable(digits=3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-0.413	0.063	-6.514	0
danceability	0.643	0.063	10.283	0
energy	0.697	0.057	12.173	0

term	estimate	std.error	statistic	p.value
acousticness	0.146	0.041	3.574	0

Exercise 4

The models do not have the same number of predictors. The model using AIC has all of the same predictors as the model using BIC plus an additional five predictor variables.

This is the model we would expect to have more predictors because for data with more than eight observations, like the spotify data which has 508 observations, the penalty for BIC is larger than that of AIC. This means that BIC tends to favor more parsimonious models (i.e. models with fewer terms). Therefore, we would expect the model using BIC to have fewer predictors, and this is in fact the case.

Exercise 5

```
forward_aic <- step(int_only_model,formula(full_model),direction = "forward",results = "hide")
```

```
## Start:  AIC=-1518.87
## valence ~ 1
##
##                     Df Sum of Sq    RSS     AIC
## + energy             1    5.1432 20.305 -1631.6
## + danceability       1    3.5393 21.909 -1592.9
## + loudness           1    3.1647 22.284 -1584.3
## + playlist_genre     5    1.5512 23.897 -1540.8
## + acousticness       1    0.7567 24.692 -1532.2
## <none>                            25.448 -1518.9
## + instrumentalness   1    0.0851 25.363 -1518.6
## + mode               1    0.0267 25.422 -1517.4
## + tempo              1    0.0153 25.433 -1517.2
## + duration_ms        1    0.0056 25.443 -1517.0
## + liveness           1    0.0012 25.447 -1516.9
## + speechiness        1    0.0003 25.448 -1516.9
## + key               11    0.8048 24.643 -1513.2
##
## Step:  AIC=-1631.56
## valence ~ energy
##
##                     Df Sum of Sq    RSS     AIC
## + danceability       1    3.1853 17.120 -1716.2
## + playlist_genre     5    0.5398 19.765 -1635.2
## + acousticness       1    0.1057 20.200 -1632.2
## <none>                            20.305 -1631.6
## + liveness           1    0.0773 20.228 -1631.5
## + loudness           1    0.0329 20.272 -1630.4
## + tempo              1    0.0275 20.278 -1630.2
## + speechiness        1    0.0216 20.284 -1630.1
## + mode               1    0.0190 20.286 -1630.0
## + duration_ms        1    0.0159 20.289 -1630.0
## + instrumentalness   1    0.0005 20.305 -1629.6
## + key               11    0.4022 19.903 -1619.7
##
## Step:  AIC=-1716.24
```

```
## valence ~ energy + danceability
##
##                     Df Sum of Sq    RSS     AIC
## + acousticness       1   0.42326 16.697 -1727.0
## + playlist_genre     5   0.51363 16.606 -1721.7
## <none>                           17.120 -1716.2
## + speechiness        1   0.05359 17.066 -1715.8
## + mode               1   0.03486 17.085 -1715.3
## + duration_ms        1   0.01556 17.104 -1714.7
## + liveness           1   0.01397 17.106 -1714.7
## + tempo              1   0.01278 17.107 -1714.6
## + loudness           1   0.00543 17.114 -1714.4
## + instrumentalness   1   0.00236 17.117 -1714.3
## + key               11   0.27197 16.848 -1702.4
##
## Step:  AIC=-1726.96
## valence ~ energy + danceability + acousticness
##
##                     Df Sum of Sq    RSS     AIC
## + playlist_genre     5   0.42384 16.273 -1730.0
## <none>                           16.697 -1727.0
## + mode               1   0.03500 16.662 -1726.0
## + speechiness        1   0.03225 16.664 -1725.9
## + tempo              1   0.01952 16.677 -1725.5
## + instrumentalness   1   0.01216 16.684 -1725.3
## + duration_ms        1   0.01098 16.686 -1725.3
## + liveness           1   0.01026 16.686 -1725.3
## + loudness           1   0.00610 16.691 -1725.2
## + key               11   0.23407 16.463 -1712.1
##
## Step:  AIC=-1730.02
## valence ~ energy + danceability + acousticness + playlist_genre
##
##                     Df Sum of Sq    RSS     AIC
## <none>                           16.273 -1730.0
## + tempo              1  0.039168 16.234 -1729.2
## + liveness           1  0.028872 16.244 -1728.9
## + mode               1  0.021451 16.251 -1728.7
## + instrumentalness   1  0.020691 16.252 -1728.7
## + loudness           1  0.003130 16.270 -1728.1
## + speechiness        1  0.002164 16.271 -1728.1
## + duration_ms        1  0.000167 16.273 -1728.0
## + key               11  0.234349 16.038 -1715.4
```

```
tidy(forward_aic)%>%
  kable(digits=3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-0.314	0.103	-3.058	0.002
energy	0.655	0.059	11.133	0.000
danceability	0.673	0.066	10.232	0.000
acousticness	0.134	0.042	3.198	0.001
playlist_genrelatin	-0.052	0.077	-0.674	0.500
playlist_genrepop	-0.089	0.075	-1.182	0.238

term	estimate	std.error	statistic	p.value
playlist_genrer&b	-0.108	0.080	-1.359	0.175
playlist_genrerap	-0.135	0.077	-1.753	0.080
playlist_genrerock	-0.028	0.083	-0.339	0.735

Exercise 6

```
forward_bic <- step(int_only_model,formula(full_model),direction="forward",k=log(n),results="hide")
```

```
## Start:  AIC=-1514.64
## valence ~ 1
##
##                    Df Sum of Sq    RSS     AIC
## + energy            1    5.1432 20.305 -1623.1
## + danceability      1    3.5393 21.909 -1584.5
## + loudness          1    3.1647 22.284 -1575.9
## + acousticness      1    0.7567 24.692 -1523.7
## + playlist_genre    5    1.5512 23.897 -1515.4
## <none>                          25.448 -1514.6
## + instrumentalness  1    0.0851 25.363 -1510.1
## + mode              1    0.0267 25.422 -1508.9
## + tempo             1    0.0153 25.433 -1508.7
## + duration_ms       1    0.0056 25.443 -1508.5
## + liveness          1    0.0012 25.447 -1508.4
## + speechiness       1    0.0003 25.448 -1508.4
## + key              11    0.8048 24.643 -1462.4
##
## Step:  AIC=-1623.1
## valence ~ energy
##
##                    Df Sum of Sq    RSS     AIC
## + danceability      1    3.1853 17.120 -1703.5
## <none>                          20.305 -1623.1
## + acousticness      1    0.1057 20.200 -1619.5
## + liveness          1    0.0773 20.228 -1618.8
## + loudness          1    0.0329 20.272 -1617.7
## + tempo             1    0.0275 20.278 -1617.6
## + speechiness       1    0.0216 20.284 -1617.4
## + mode              1    0.0190 20.286 -1617.3
## + duration_ms       1    0.0159 20.289 -1617.3
## + instrumentalness  1    0.0005 20.305 -1616.9
## + playlist_genre    5    0.5398 19.765 -1605.6
## + key              11    0.4022 19.903 -1564.7
##
## Step:  AIC=-1703.55
## valence ~ energy + danceability
##
##                    Df Sum of Sq    RSS     AIC
## + acousticness      1   0.42326 16.697 -1710.0
## <none>                          17.120 -1703.5
## + speechiness       1   0.05359 17.066 -1698.9
## + mode              1   0.03486 17.085 -1698.4
## + duration_ms       1   0.01556 17.104 -1697.8
```

```
## + liveness          1   0.01397 17.106 -1697.7
## + tempo             1   0.01278 17.107 -1697.7
## + loudness          1   0.00543 17.114 -1697.5
## + instrumentalness  1   0.00236 17.117 -1697.4
## + playlist_genre    5   0.51363 16.606 -1687.9
## + key              11   0.27197 16.848 -1643.2
##
## Step:  AIC=-1710.04
## valence ~ energy + danceability + acousticness
##
##                     Df Sum of Sq    RSS     AIC
## <none>                           16.697 -1710.0
## + mode              1   0.03500 16.662 -1704.9
## + speechiness       1   0.03225 16.664 -1704.8
## + tempo             1   0.01952 16.677 -1704.4
## + instrumentalness  1   0.01216 16.684 -1704.2
## + duration_ms       1   0.01098 16.686 -1704.1
## + liveness          1   0.01026 16.686 -1704.1
## + loudness          1   0.00610 16.691 -1704.0
## + playlist_genre    5   0.42384 16.273 -1692.0
## + key              11   0.23407 16.463 -1648.7
```

```
tidy(forward_bic)%>%
  kable(digits=3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-0.413	0.063	-6.514	0
energy	0.697	0.057	12.173	0
danceability	0.643	0.063	10.283	0
acousticness	0.146	0.041	3.574	0

Exercise 7

```
glance(forward_aic) %>% select(r.squared, adj.r.squared)
```

```
## # A tibble: 1 x 2
##   r.squared adj.r.squared
##       <dbl>         <dbl>
## 1     0.361         0.350
```

```
glance(forward_bic) %>% select(r.squared, adj.r.squared)
```

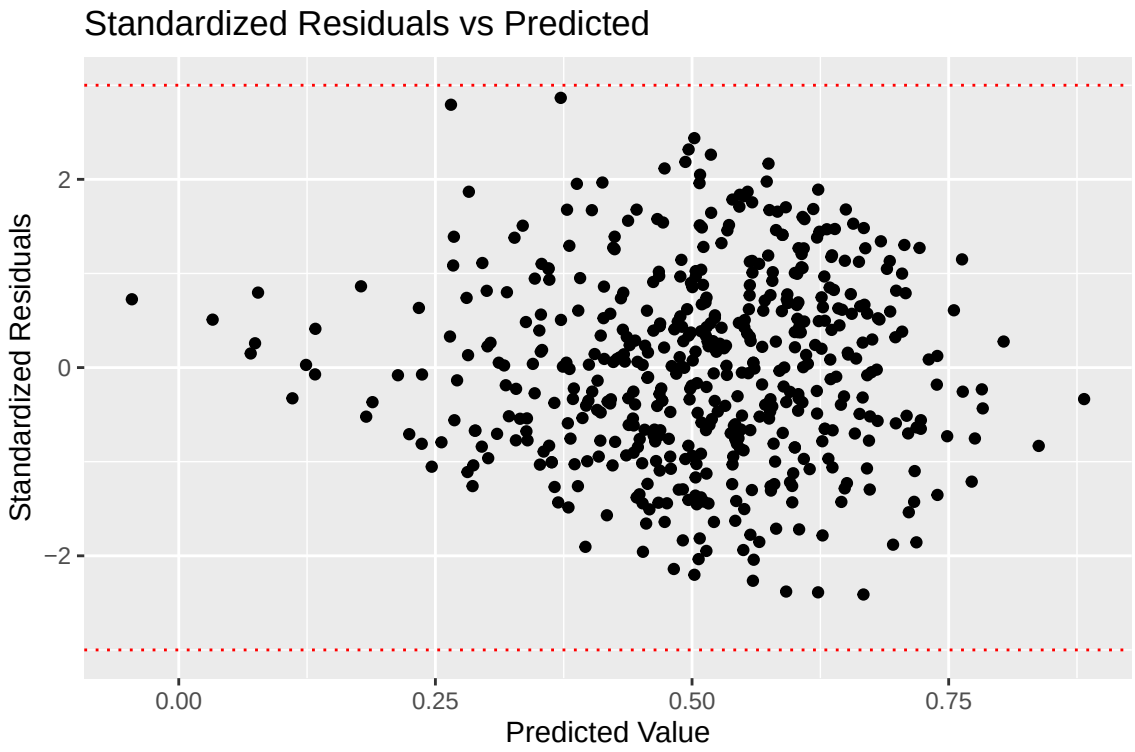
```
## # A tibble: 1 x 2
##   r.squared adj.r.squared
##       <dbl>         <dbl>
## 1     0.344         0.340
```

In general, we choose the model with the highest adjusted R^2 value, i.e. we choose the regression model that explains the most amount of variation in the response, given the correction adjusted R^2 performs for the number of predictor variables. Thus, we would choose the forward_aic model.

Exercise 8

```
selected_aug<- augment(forward_aic)
selected_aug <- selected_aug %>%
  mutate(obs_num = 1:nrow(selected_aug))

ggplot(data = selected_aug, aes(x = .fitted, y = .std.resid)) + geom_point() +
  labs(x = "Predicted Value", y = "Standardized Residuals",
       title = "Standardized Residuals vs Predicted") +
  geom_hline(yintercept = -3,color = "red",linetype = "dotted") +
  geom_hline(yintercept = 3,color = "red",linetype = "dotted")
```



Based on the above plot of standardized plot vs residuals, I am not sure that the linearity assumption is met. residuals appear to be centered and most varied at about predicted value .5, and then get fewer and smaller as you move away from that value in either direction.