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")
```

```
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))
```

```
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.