Final 210 Project

Sean Villoresi and Ellie Kang

Read in the data

```
library(tidyverse)
library(tidymodels)
library(broom)
library(leaps)
library(MASS)
library(caret)
library(glmnet)
library(Stat2Data)
library(nnet)
library(lme4)
music <- read_csv("data/Spotify_Youtube.csv")
```

##Cleaning the data

```
music$Uri=NULL
music$Url_youtube=NULL
music$Url_spotify=NULL

music <- music[complete.cases(music$Stream), ]
music <- music[complete.cases(music$Danceability), ]
music <- music[complete.cases(music$Licensed), ]
sum(is.na(music$Licensed))
```

```
[1] 0
```

Variable Selection

```r
indices <- sample(1:5000, size = 5000 * 0.8, replace = F)
train.data  <- music %>%
  slice(indices)
test.data <- music %>%
  slice(-indices)

lm_none <- lm(Stream ~ 1 , data = train.data)

lm_all <- lm(Stream ~ Danceability + Energy + Key + Loudness + Speechiness +
               Acousticness + Instrumentalness + Liveness +  Valence + Tempo +
               Duration_ms + official_video, data = train.data)
```

```r
stepAIC(lm_all,
        scope = list(lower = lm_none, upper = lm_all),
        data = music, direction = "both")
```

```
Start:  AIC=153126.6
Stream ~ Danceability + Energy + Key + Loudness + Speechiness +
    Acousticness + Instrumentalness + Liveness + Valence + Tempo +
    Duration_ms + official_video

                   Df  Sum of Sq        RSS    AIC
- Tempo             1 2.9136e+15 1.6779e+20 153125
- Valence           1 2.0427e+16 1.6780e+20 153125
- Instrumentalness  1 4.4143e+16 1.6783e+20 153126
- Danceability      1 6.1816e+16 1.6784e+20 153126
<none>                           1.6778e+20 153127
- Energy            1 1.1600e+17 1.6790e+20 153127
- Key               1 1.3149e+17 1.6791e+20 153128
- Duration_ms       1 1.3882e+17 1.6792e+20 153128
- Liveness          1 1.5710e+17 1.6794e+20 153128
- Speechiness       1 3.4363e+17 1.6813e+20 153133
- Loudness          1 4.5537e+17 1.6824e+20 153135
- official_video    1 1.2530e+18 1.6904e+20 153154
- Acousticness      1 1.3751e+18 1.6916e+20 153157

Step:  AIC=153124.7
Stream ~ Danceability + Energy + Key + Loudness + Speechiness +
    Acousticness + Instrumentalness + Liveness + Valence + Duration_ms +
```

```
    official_video

                    Df  Sum of Sq        RSS    AIC
- Valence            1 1.8832e+16 1.6780e+20 153123
- Instrumentalness   1 4.5161e+16 1.6783e+20 153124
- Danceability       1 5.8904e+16 1.6784e+20 153124
<none>                            1.6779e+20 153125
- Energy             1 1.1576e+17 1.6790e+20 153125
- Key                1 1.3085e+17 1.6792e+20 153126
- Duration_ms        1 1.3995e+17 1.6793e+20 153126
- Liveness           1 1.5918e+17 1.6794e+20 153126
+ Tempo              1 2.9136e+15 1.6778e+20 153127
- Speechiness        1 3.4300e+17 1.6813e+20 153131
- Loudness           1 4.6063e+17 1.6825e+20 153134
- official_video     1 1.2553e+18 1.6904e+20 153152
- Acousticness       1 1.3855e+18 1.6917e+20 153156

Step:  AIC=153123.1
Stream ~ Danceability + Energy + Key + Loudness + Speechiness +
    Acousticness + Instrumentalness + Liveness + Duration_ms +
    official_video

                    Df  Sum of Sq        RSS    AIC
- Instrumentalness   1 3.5661e+16 1.6784e+20 153122
- Danceability       1 4.0539e+16 1.6784e+20 153122
<none>                            1.6780e+20 153123
- Duration_ms        1 1.3395e+17 1.6794e+20 153124
- Key                1 1.3530e+17 1.6794e+20 153124
+ Valence            1 1.8832e+16 1.6779e+20 153125
- Liveness           1 1.5758e+17 1.6796e+20 153125
+ Tempo              1 1.3180e+15 1.6780e+20 153125
- Energy             1 1.7243e+17 1.6798e+20 153125
- Speechiness        1 3.3023e+17 1.6813e+20 153129
- Loudness           1 5.1983e+17 1.6832e+20 153133
- official_video     1 1.2704e+18 1.6907e+20 153151
- Acousticness       1 1.4794e+18 1.6928e+20 153156

Step:  AIC=153122
Stream ~ Danceability + Energy + Key + Loudness + Speechiness +
    Acousticness + Liveness + Duration_ms + official_video

                    Df  Sum of Sq        RSS    AIC
- Danceability       1 5.5766e+16 1.6790e+20 153121
```

```
<none>                          1.6784e+20 153122
+ Instrumentalness  1 3.5661e+16 1.6780e+20 153123
- Duration_ms       1 1.3313e+17 1.6797e+20 153123
- Key               1 1.3964e+17 1.6798e+20 153123
- Liveness          1 1.4658e+17 1.6799e+20 153123
+ Valence           1 9.3319e+15 1.6783e+20 153124
+ Tempo             1 2.3956e+15 1.6784e+20 153124
- Energy            1 2.0453e+17 1.6804e+20 153125
- Speechiness       1 3.1807e+17 1.6816e+20 153128
- Loudness          1 8.1932e+17 1.6866e+20 153139
- official_video    1 1.2597e+18 1.6910e+20 153150
- Acousticness      1 1.5196e+18 1.6936e+20 153156

Step:  AIC=153121.3
Stream ~ Energy + Key + Loudness + Speechiness + Acousticness +
    Liveness + Duration_ms + official_video

                  Df  Sum of Sq        RSS    AIC
<none>                          1.6790e+20 153121
+ Danceability      1 5.5766e+16 1.6784e+20 153122
+ Instrumentalness  1 5.0888e+16 1.6784e+20 153122
- Key               1 1.3255e+17 1.6803e+20 153122
- Duration_ms       1 1.4419e+17 1.6804e+20 153123
+ Valence           1 1.3398e+15 1.6789e+20 153123
+ Tempo             1 8.5860e+13 1.6790e+20 153123
- Liveness          1 1.7511e+17 1.6807e+20 153123
- Energy            1 2.1658e+17 1.6811e+20 153124
- Speechiness       1 2.7774e+17 1.6817e+20 153126
- Loudness          1 9.4740e+17 1.6884e+20 153142
- official_video    1 1.2596e+18 1.6916e+20 153149
- Acousticness      1 1.6726e+18 1.6957e+20 153159


Call:
lm(formula = Stream ~ Energy + Key + Loudness + Speechiness +
    Acousticness + Liveness + Duration_ms + official_video, data = train.data)

Coefficients:
      (Intercept)             Energy                Key           Loudness
        2.464e+08         -5.920e+07         -1.624e+06          4.875e+06
      Speechiness       Acousticness           Liveness        Duration_ms
       -7.314e+07         -9.563e+07         -3.914e+07         -4.348e+01
```

```
official_videoTRUE
         4.165e+07
```

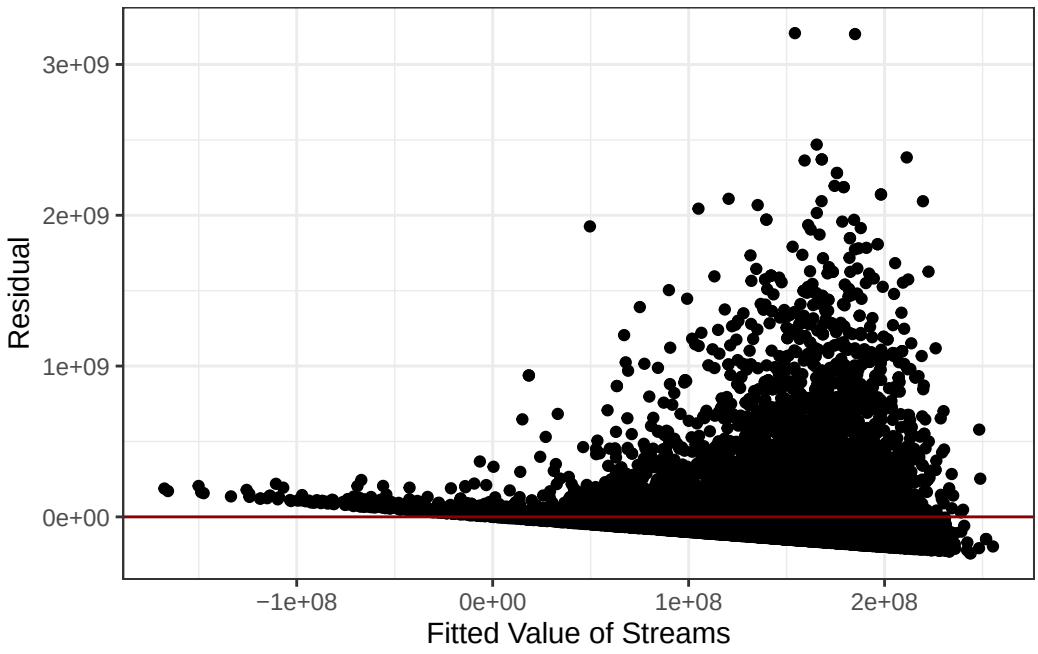
```
y <- music$Stream
x <- model.matrix(Stream ~ Danceability + Energy + Key + Loudness + Speechiness +
                Acousticness + Instrumentalness + Liveness +  Valence + Tempo +
                Duration_ms + official_video,
                   data = music)

m_lasso_cv <- cv.glmnet(x, y, alpha = 1)


best_lambda <- m_lasso_cv$lambda.min
m_best <- glmnet(x, y, alpha = 1, lambda = best_lambda)
m_best$beta
```

```
13 x 1 sparse Matrix of class "dgCMatrix"
                              s0
(Intercept)         .
Danceability         5.820487e+07
Energy              -1.328335e+08
Key                 -6.234273e+05
Loudness             6.927111e+06
Speechiness         -6.226337e+07
Acousticness        -8.358566e+07
Instrumentalness    -3.809595e+07
Liveness            -4.327010e+07
Valence             -5.213674e+07
Tempo               -4.972216e+04
Duration_ms         -2.663998e+01
official_videoTRUE   4.992768e+07
```

```
## CHANGE LATER AFTER WE ACTUALLY CHOOSE MODELS FROM ABOVE< CURRENTLY A PLACE HOLDER.
pretransform_model <- lm(Stream ~ Danceability + Energy + Key + Loudness + Speechiness +
                Acousticness + Instrumentalness + Liveness +  Valence + Tempo +
                Duration_ms + official_video, data = music)
```

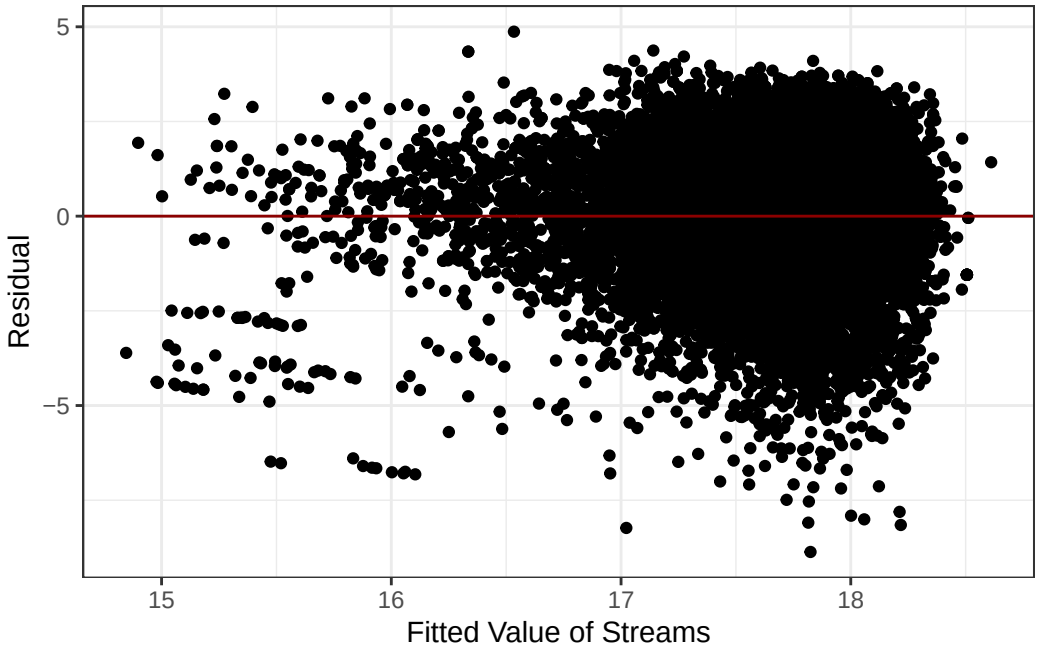
Linearity Assumptions and Checks for Transformations

```
ptmodel_aug <- augment(pretransform_model)

transform_model <- lm(log(Stream) ~ Danceability + Energy + Key + Loudness +
                        Speechiness +Acousticness + Instrumentalness + Liveness
                      +  Valence + Tempo + Duration_ms + official_video,
                      data = music)

tmodel_aug <- augment(transform_model)


ggplot(ptmodel_aug, aes(x = .fitted, y=.resid)) +
  geom_point() +
  geom_hline(yintercept=0, color ="darkred") +
  labs(x ="Fitted Value of Streams", y = "Residual") +
  theme_bw()
```

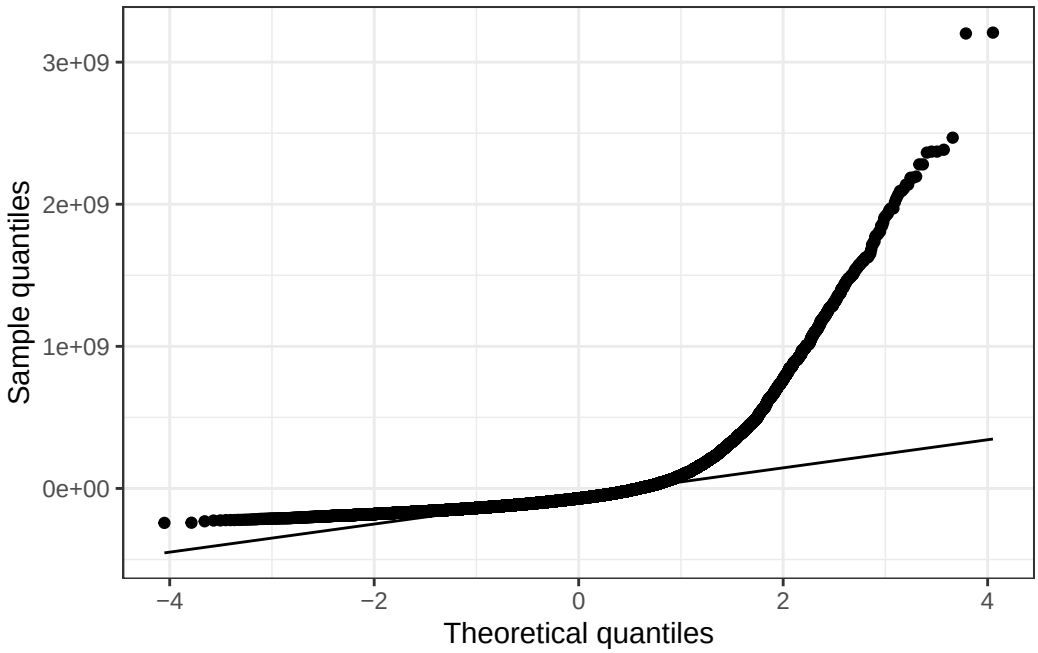



```
ggplot(tmodel_aug, aes(x = .fitted, y=.resid)) +
  geom_point() +
  geom_hline(yintercept=0, color ="darkred") +
```

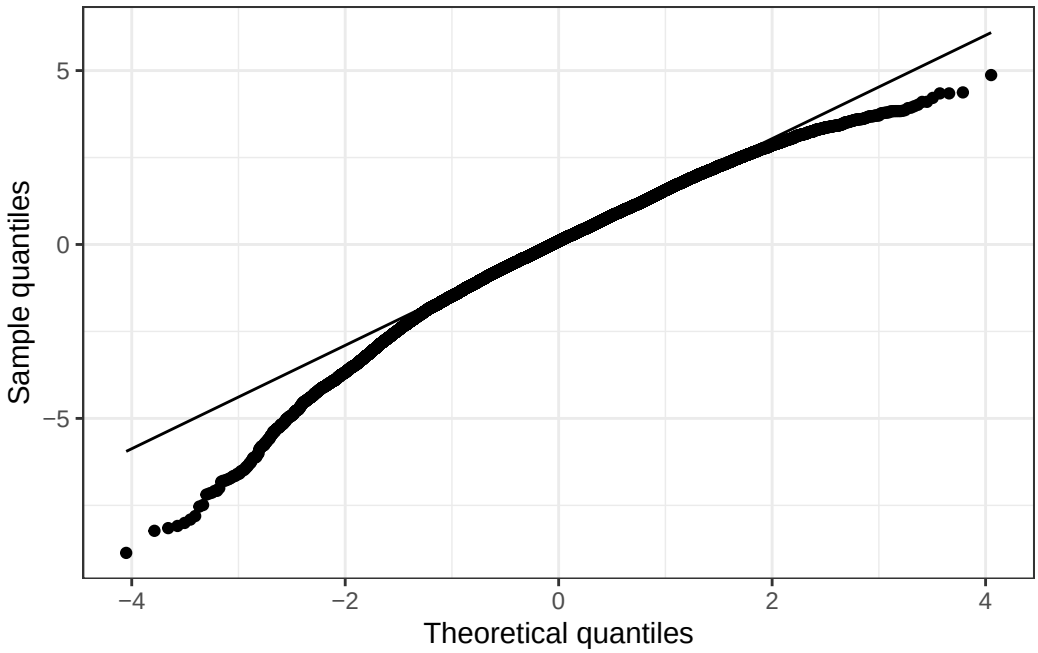
```
  labs(x ="Fitted Value of Streams", y = "Residual") +
  theme_bw()
```




```
ggplot(ptmodel_aug, aes(sample = .resid)) +
  stat_qq() +
  stat_qq_line() +
  theme_bw() +
  labs(x = "Theoretical quantiles",
       y = "Sample quantiles")
```




```
ggplot(tmodel_aug, aes(sample = .resid)) +
  stat_qq() +
  stat_qq_line() +
  theme_bw() +
  labs(x = "Theoretical quantiles",
       y = "Sample quantiles")
```


Sample quantiles
5
0
−5
−4
−2
0
2
4
Theoretical quantiles