# 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$Url_spotify=NULL
music$Description=NULL

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

```
[1] 0
```

## Variable Selection

```
indices <- sample(1:19691, size = 15000 * 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 + factor(Key) + Loudness + Speechiness +
               Acousticness + Instrumentalness + Liveness +  Valence + Tempo +
               Duration_ms + official_video, data = train.data)
```

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

```
Start:  AIC=463561.8
Stream ~ Danceability + Energy + factor(Key) + Loudness + Speechiness +
    Acousticness + Instrumentalness + Liveness + Valence + Tempo +
    Duration_ms + official_video

                   Df  Sum of Sq        RSS    AIC
- factor(Key)      11 6.1467e+17 7.1573e+20 463550
- Tempo             1 5.4401e+16 7.1517e+20 463561
- Duration_ms       1 6.6584e+16 7.1518e+20 463561
<none>                           7.1512e+20 463562
- Instrumentalness  1 3.4769e+17 7.1546e+20 463566
- Speechiness       1 3.7362e+17 7.1549e+20 463566
- Danceability      1 5.0646e+17 7.1562e+20 463568
- Liveness          1 5.9823e+17 7.1571e+20 463570
- Valence           1 7.7068e+17 7.1589e+20 463573
- Energy            1 3.0572e+18 7.1817e+20 463611
- Acousticness      1 3.2460e+18 7.1836e+20 463614
- Loudness          1 4.5314e+18 7.1965e+20 463636
- official_video    1 4.9499e+18 7.2007e+20 463643

Step:  AIC=463550.1
Stream ~ Danceability + Energy + Loudness + Speechiness + Acousticness +
    Instrumentalness + Liveness + Valence + Tempo + Duration_ms +
```

```
    official_video

                   Df  Sum of Sq        RSS    AIC
- Tempo             1 5.5549e+16 7.1579e+20 463549
- Duration_ms       1 7.4040e+16 7.1580e+20 463549
<none>                           7.1573e+20 463550
- Speechiness       1 3.0522e+17 7.1604e+20 463553
- Instrumentalness  1 3.5241e+17 7.1608e+20 463554
- Danceability      1 5.5060e+17 7.1628e+20 463557
- Liveness          1 6.1740e+17 7.1635e+20 463558
- Valence           1 7.9588e+17 7.1653e+20 463561
+ factor(Key)      11 6.1467e+17 7.1512e+20 463562
- Energy            1 3.1099e+18 7.1884e+20 463600
- Acousticness      1 3.3687e+18 7.1910e+20 463604
- Loudness          1 4.6068e+18 7.2034e+20 463625
- official_video    1 4.9643e+18 7.2069e+20 463631

Step:  AIC=463549
Stream ~ Danceability + Energy + Loudness + Speechiness + Acousticness +
    Instrumentalness + Liveness + Valence + Duration_ms + official_video

                   Df  Sum of Sq        RSS    AIC
- Duration_ms       1 6.8473e+16 7.1585e+20 463548
<none>                           7.1579e+20 463549
+ Tempo             1 5.5549e+16 7.1573e+20 463550
- Speechiness       1 3.2694e+17 7.1611e+20 463552
- Instrumentalness  1 3.5115e+17 7.1614e+20 463553
- Liveness          1 6.0295e+17 7.1639e+20 463557
- Danceability      1 6.3426e+17 7.1642e+20 463558
+ factor(Key)      11 6.1582e+17 7.1517e+20 463561
- Valence           1 8.5048e+17 7.1664e+20 463561
- Energy            1 3.1188e+18 7.1890e+20 463599
- Acousticness      1 3.3314e+18 7.1912e+20 463603
- Loudness          1 4.5583e+18 7.2034e+20 463623
- official_video    1 4.9587e+18 7.2074e+20 463630

Step:  AIC=463548.2
Stream ~ Danceability + Energy + Loudness + Speechiness + Acousticness +
    Instrumentalness + Liveness + Valence + official_video

                   Df  Sum of Sq        RSS    AIC
<none>                           7.1585e+20 463548
+ Duration_ms       1 6.8473e+16 7.1579e+20 463549
```

```
+ Tempo             1 4.9983e+16 7.1580e+20 463549
- Speechiness       1 3.1675e+17 7.1617e+20 463551
- Instrumentalness  1 3.4271e+17 7.1620e+20 463552
- Liveness          1 5.9437e+17 7.1645e+20 463556
- Danceability      1 6.7182e+17 7.1653e+20 463557
+ factor(Key)      11 6.2291e+17 7.1523e+20 463560
- Valence           1 8.3565e+17 7.1669e+20 463560
- Energy            1 3.1436e+18 7.1900e+20 463599
- Acousticness      1 3.3101e+18 7.1916e+20 463602
- Loudness          1 4.5686e+18 7.2042e+20 463623
- official_video    1 4.9583e+18 7.2081e+20 463629


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

Coefficients:
       (Intercept)        Danceability              Energy            Loudness
         270556720             56223432         -140929899             7511210
       Speechiness        Acousticness    Instrumentalness            Liveness
          -49832363            -79766063           -34054767           -44227741
           Valence  official_videoTRUE
          -41787432             49535932
```

```r
y <- music$Stream
x <- model.matrix(Stream ~ Danceability + Energy + factor(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
```

```
23 x 1 sparse Matrix of class "dgCMatrix"
                              s0
```

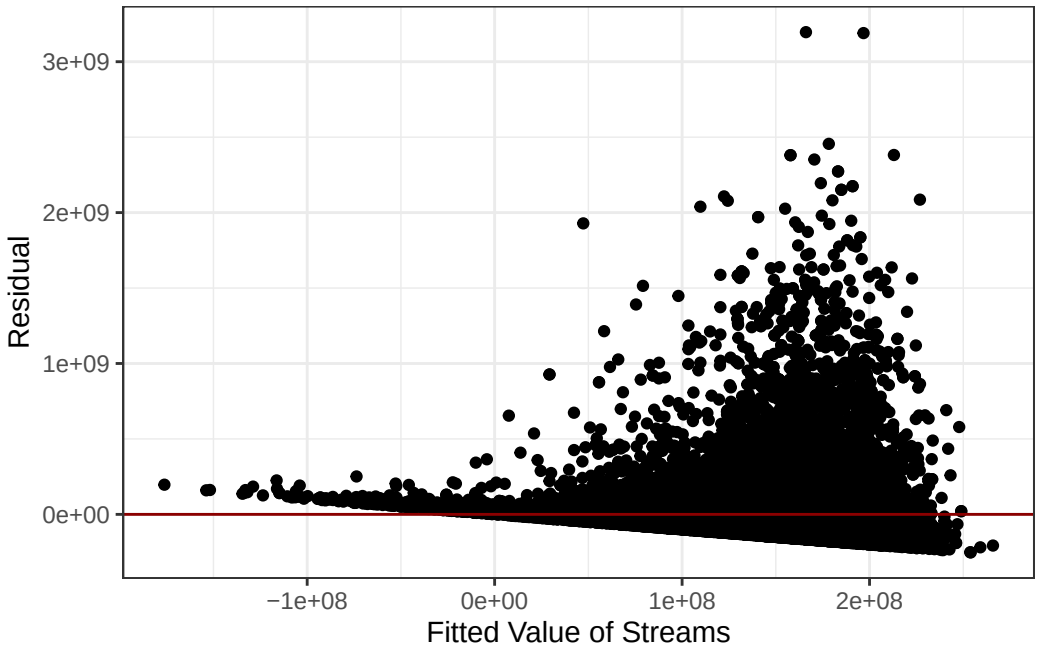
```
(Intercept)          .
Danceability         5.523128e+07
Energy              -1.292150e+08
factor(Key)1         1.229817e+07
factor(Key)2        -9.150133e+06
factor(Key)3        -3.856918e+06
factor(Key)4        -9.275403e+05
factor(Key)5         .
factor(Key)6         2.212346e+06
factor(Key)7        -1.103897e+07
factor(Key)8         2.950704e+06
factor(Key)9        -1.642045e+07
factor(Key)10       -3.280228e+06
factor(Key)11        6.424310e+06
Loudness             6.733644e+06
Speechiness         -6.772129e+07
Acousticness        -8.108656e+07
Instrumentalness    -3.814151e+07
Liveness            -4.189008e+07
Valence             -5.066748e+07
Tempo               -4.344071e+04
Duration_ms         -2.384572e+01
official_videoTRUE   4.956963e+07
```

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

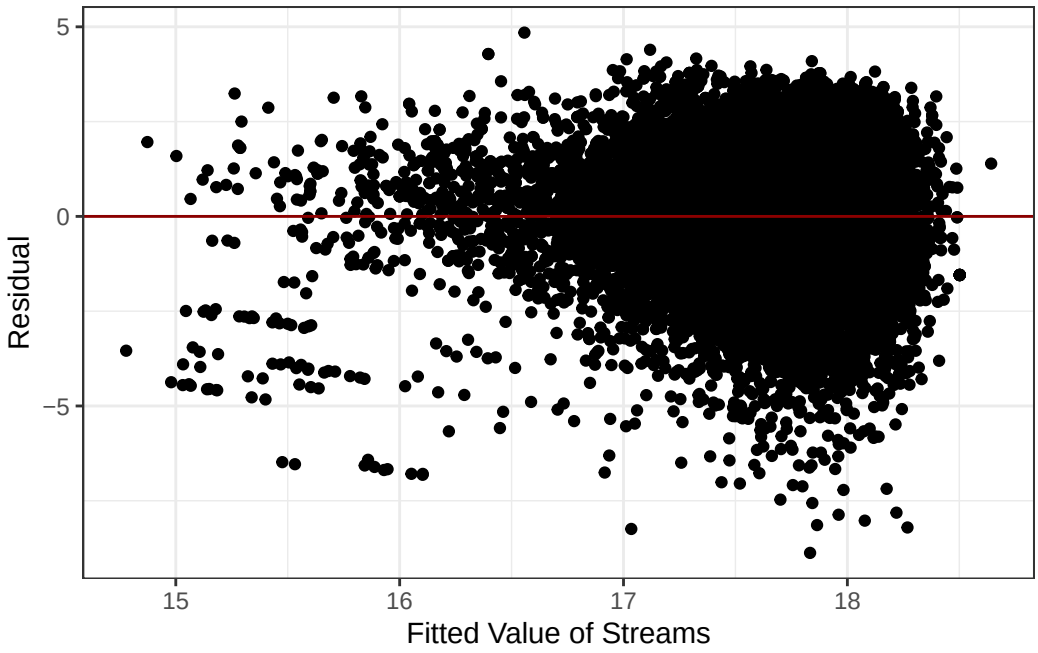
### Linearity Assumptions and Checks for Transformations

```
ptmodel_aug <- augment(pretransform_model)

transform_model <- lm(log(Stream) ~ Danceability + Energy + factor(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()
```

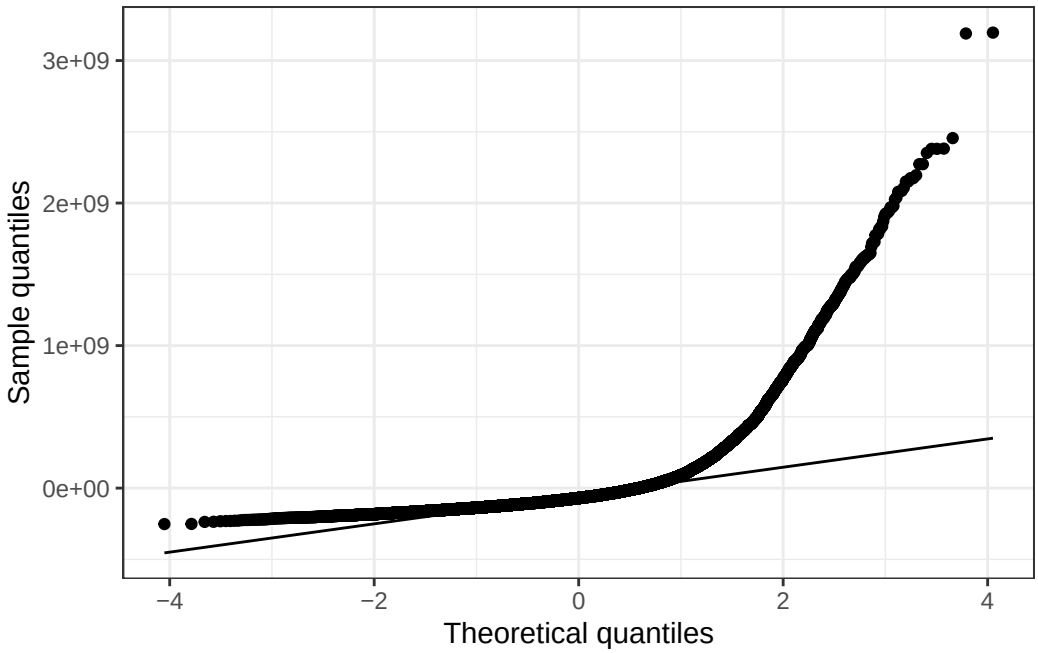



```
ggplot(tmodel_aug, aes(x = .fitted, y=.resid)) +
  geom_point() +
  geom_hline(yintercept=0, color ="darkred") +
  labs(x ="Fitted Value of Streams", y = "Residual") +
  theme_bw()
```

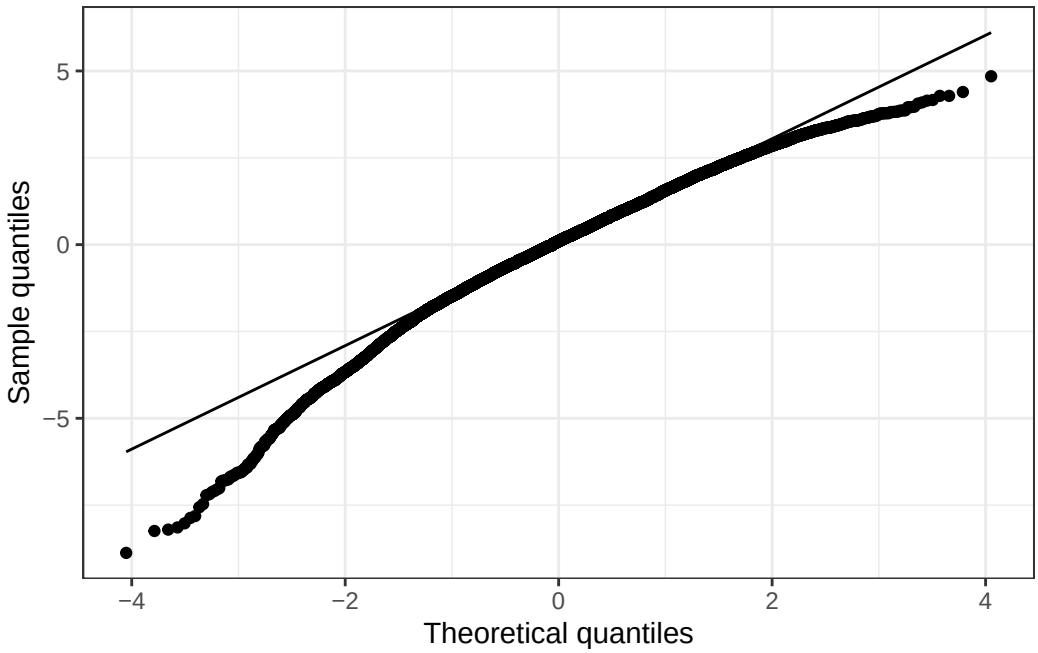



Looking at the visualizations above, we can see that the transformed model gives us a much better spread on the residual split around our red line then our untransformed model. As such, the residuals appear roughly symmetrical along the horizontal axis for our transformed plot, so we feel it safe to assume approximate linearity.

```
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