Final Project Part 2, Social Science Inquiry II (SOSC13200-W22-3)

Tiffanie Huang

Friday 2/24/23 at 5pm

```
ts <- read.csv("../data/taylor_swift_spotify.csv")
#libraries
library(ggplot2)
library(estimatr)
library(gridExtra)
library(grid)
```

Selecting for data of interest

```
#remove certain albums (karaoke, live recordings, demos, voice memos, everything but regular/deluxe alb
tswift <- ts[ts$album %in% unique(ts$album)[-c(7,11,13,14,15,17,18,19,24,25,26,27,28,30,31,32,34,35,37,3
tswift <- tswift[!grepl("Karaoke",tswift$name) & !grepl("Voice Memo",tswift$name) & !grepl("Original De
#deciding between remaining duplicates based on popularity -- keep more widely streamed version
tswift <- tswift[order(-tswift$popularity),]
tswift <- tswift[!duplicated(tswift$name),-7] #remove duplicates + url col
#table(is.na(tswift)) #4046 false (no NA values)
#save as csv for part 3:
#write.csv(tswift,file="../data/tswift.csv",row.names=FALSE)
```

Describing main variables of interest: energy, acousticness

Both energy and acousticness for each track are audio features measured on a scale of 0.0-1.0, with 0 being completely no energy or acousticness and 1 being full energy or acousticness. Energy describes how intense/active the audio in a track is and takes into account timbre and entropy of noise, while "acousticness" describes how acoustic (non-electrically amplified) it is.

```
#mean + SE of energy
print(c(mean(tswift$energy),sd(tswift$energy)/sqrt(nrow(tswift)))) #mean: 0.58047, SE: #0.01190
```

```
## [1] 0.58047479 0.01190366
```

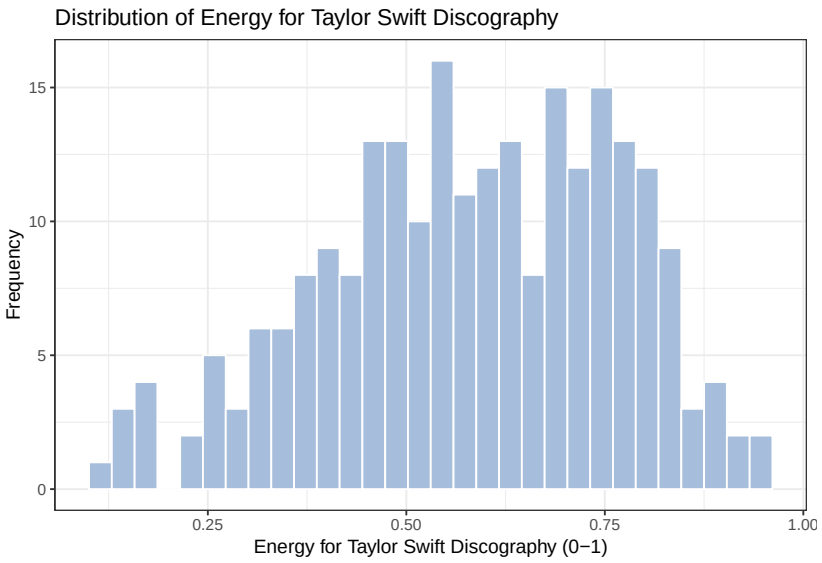
```
#mean + SE of acousticness
print(c(mean(tswift$acousticness),sd(tswift$acousticness)/sqrt(nrow(tswift)))) #mean: 0.29542, SE: 0.02
```

```
## [1] 0.29541736 0.02053328
```

Fig 1. Track energy distribution

```
#range(tswift$energy) #0.118 0.950
```

```
ggplot(tswift,
       aes(energy))+
  geom_histogram(bins=30,color="white",fill="#a6bddb")+
  theme_bw()+
  labs(x="Energy for Taylor Swift Discography (0-1)",
       y="Frequency",
       title="Distribution of Energy for Taylor Swift Discography")
```



The distribution of energy is only slightly skewed to the left but otherwise not too distorted by outliers.

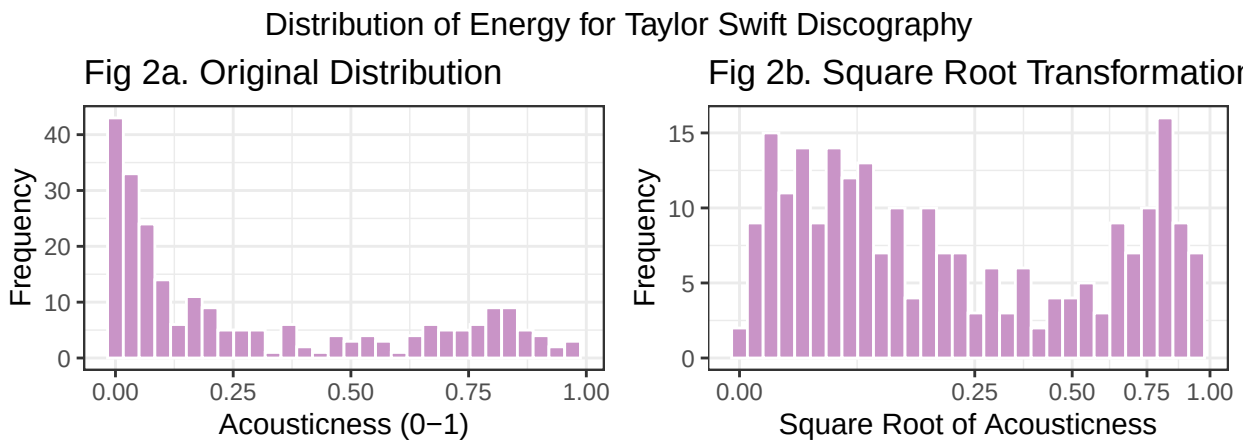
Fig 2a. Track acousticness distribution

```
g_reg <- ggplot(tswift,
       aes(acousticness))+
  geom_histogram(bins=30,color="white",fill="#c994c7")+
  theme_bw()+
  labs(x="Acousticness (0-1)",
       y="Frequency",
       title="Fig 2a. Original Distribution")
```

Fig 2b. Square root transformation of acousticness distribution

```
g_log <- ggplot(tswift,
       aes(acousticness))+
  geom_histogram(bins=30,color="white",fill="#c994c7")+
  theme_bw()+
  scale_x_sqrt()+
  labs(x="Square Root of Acousticness",
       y="Frequency",
       title="Fig 2b. Square Root Transformation")
```

```
grid.arrange(g_reg, g_log, ncol=2,heights=c(1,1),top=textGrob("Distribution of Energy for Taylor Swift D
```



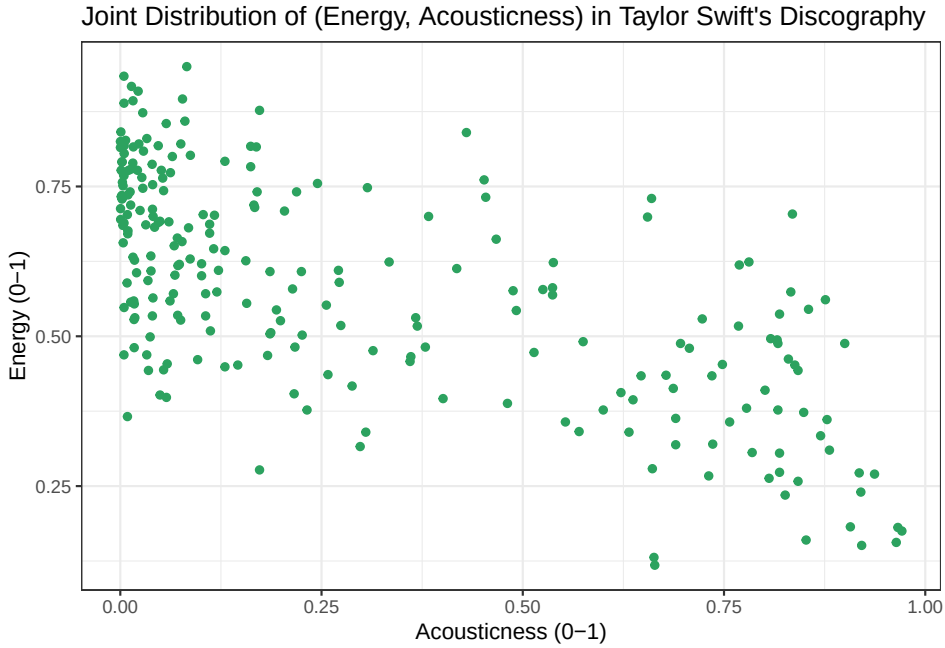
The original distribution is very skewed to the right, so we can also look at the median and square root transformation (looks more bimodal).

```
#median(tswift$acousticness) #0.138
```

Fig 3. Joint distribution of acousticness and energy

```
cov_xy <- cov(x=tswift$acousticness,y=tswift$energy) #Cov[Acousticness, Energy]= -0.03962852
```

```
ggplot(tswift,aes(x=acousticness,y=energy))+
  geom_point(color="#2ca25f")+
  theme_bw()+
  labs(x="Acousticness (0-1)",
       y="Energy (0-1)",
       title="Joint Distribution of (Energy, Acousticness) in Taylor Swift's Discography")
```



Regressing energy on acousticness and other variables

Energy, unlike acousticness, can't be measured simply by a track audio's timbre/instrumentals present. Spotify's definition for the "energy" of a song can potentially be determined by a lot of factors, such as how upbeat a song is, how loud it is, or what tones are present. I'm taking energy as my dependent variable because I want to explore how energy might change as some other variable changes. I set acousticness as my independent variable since it is a more "known/fixed" quality for a song compared to energy as an audio feature and I am not interested in seeing how it might change conditional on energy. My goal is to observe variation in energy and find out how to get a more high-energy or low-energy track.

Model 1: Regressing energy on acousticness

- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ where Y=energy and X=acousticness

```
bi_mod <- lm_robust(energy~acousticness,data=tswift)
coef(bi_mod) #Y = 0.6971-0.3949X
```

```
##  (Intercept) acousticness
##    0.6971424   -0.3949246
```

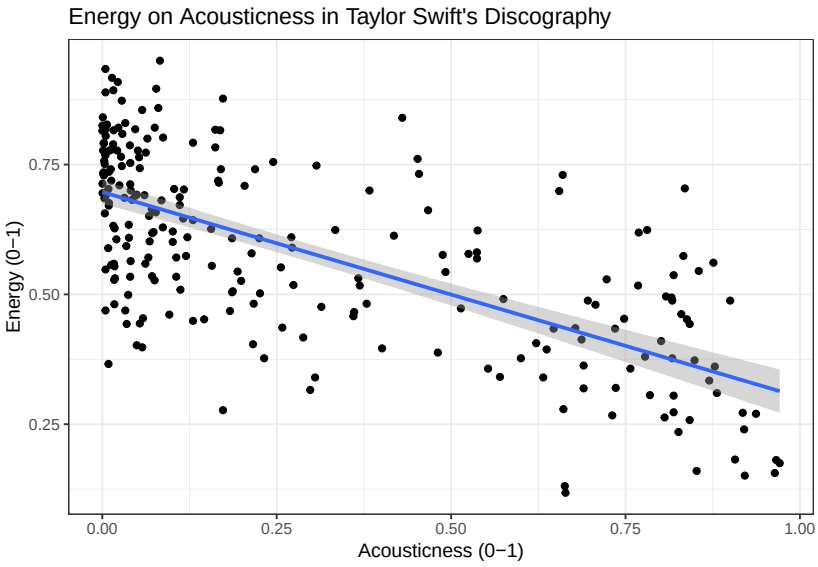
```
#summary(bi_mod)
#cov_xy/var(tswift$acousticness) #-0.3949246
```

This linear model describes the conditional relationship between energy and acousticness, specifically an approximation of the expectation function for energy given acousticness in Taylor Swift's discography. I am producing estimates for my parameters of interest (coefficients in the model) using my sample of the population joint distribution of my random variables energy and acousticness. My intercept ($\hat{\beta}_0$) is 0.6971, meaning I would predict (estimate) an energy value of 0.6971 if acousticness is fixed at 0. My slope ($\hat{\beta}_1$) is -0.3949, so when acousticness increases by 1 unit, my prediction for energy decreases by 0.3949. This relationship is not causal, as we don't know from the DGP how the "treatment" (acousticness) was assigned – there doesn't seem to be a temporal succession with the "cause" and "effect" in this relationship, and it's also not a study design where acousticness is as-if random. However, the relationship can still inform us about how a decrease in our dependent variable is associated with a 1-unit increase in the independent variable in the context of Taylor Swift's discography. If we plot it the model shows the relationship as a shape.

Fig 4. Energy on acousticness

```
ggplot(tswift,aes(x=acousticness,y=energy))+
  geom_point()+
  geom_smooth(method="lm_robust")+
  theme_bw()+
  labs(x="Acousticness (0-1)",
       y="Energy (0-1)",
       title="Energy on Acousticness in Taylor Swift's Discography")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



For the intercept, my standard error is 0.01156, the p-value is 2.413e-145, and the confidence interval is between 0.6744 and 0.7199. For the intercept, my standard error is 0.02767, the p-value is 9.629e-34, and the confidence interval is between -0.4494 and-0.3404. I can reject the null hypothesis that the slope is equal to 0 at both p=0.05 and p=0.01, and even at p=0.001 because the p-value is less than all those values. If I wasn't able to reject the null hypothesis, it would mean that under the null hypothesis (slope=0), my estimate for the slope coefficient is not very unlikely/"statistically significant." In other words, the decrease in energy conditional on acousticness I observed could be just from randomness and not very extreme compared to a slope of 0.

Model 2: Regressing energy on acousticness, holding tempo constant

- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$ where Y=energy, X_1=acousticness, and X_2=tempo

```
multi_mod1 <- lm_robust(energy~acousticness+tempo,data=tswift)
coef(multi_mod1) #Y = 0.64808-0.3925X1+0.000392X2
```

```
##   (Intercept)  acousticness         tempo
##  0.6480846351 -0.3925206030  0.0003924814
```

```
#summary(multi_mod1)
```

Model 2 regresses energy on acousticness while controlling for tempo – now the coefficient for my independent variable acousticness is -0.3925206, which is how much energy changes (decreasing) with a 1-unit increase in acousticness, holding tempo constant. For the coefficient for tempo, 0.0003925, we can interpret it as a 0.0003925 increase in energy for a 1-unit change in tempo, holding acousticness constant. However, this has a p-value of around 0.2 and is not statistically significant. The slope coefficient on acousticness does not show any meaningful change, and its p-value for the independent variable shows it is still highly significant despite increasing by ~2 factors of 10. There is no drastic change in my slope coefficient estimate, which suggests that accounting for variation of tempo with variation in both acousticness and energy does not really impact the change in energy conditional on acousticness and that the first model does not mistakenly include how tempo explains energy/acousticness as part of the conditional relationship between energy and acousticness.

Model 3: Regressing energy on acousticness, controlling for valence and loudness

- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}$ where Y=energy, X_1=acousticness, X_2=valence, and X_3=loudness

```
multi_mod2 <- lm_robust(energy~acousticness+valence+loudness,data=tswift)
coef(multi_mod2) #Y = 0.76029-0.1561X1+0.2642X2+0.03315X3
```

```
##  (Intercept) acousticness      valence     loudness
##    0.7602949   -0.1561469    0.2642305    0.0331528
```

```
#summary(multi_mod2)
```

Model 3 regresses energy on acousticness while controlling for valence and loudness. The coefficient for acousticness is -0.15615, which is how much energy changes (decreasing) with a 1-unit increase in acousticness, holding valence and loudness constant. Also included are the coefficients 0.26423 for increase in energy with a 1-unit increase in valence when holding acousticness and loudness constant and 0.03315 for increase in energy with a 1-unit increase in loudness when holding acousticness and valence constant. #' Unlike Model 2, controlling for valence and loudness drastically changes the coefficient on acousticness, so now energy decreases by a lower amount for every 1-unit increase in acousticness. The p-value for this coefficient is highly significant but increases by a lot more than the increase observed in Model 1; meanwhile, the p-values for the other 2 coefficients are slightly smaller. Because we see the conditional relationship between energy and acousticness while holding valence and loudness constant becoming more positive (-0.156 compared to -0.395 from the bivariate model), the slope for that relationship gets more "flattened" due to controlling for the part of valence and loudness that "explain" both acousticness and energy; this part might have been hidden in the standard error term of the simple bivariate model. This suggests that valence and loudness also help explain variation in energy along with acousticness (both indicate a positive change in energy for every one-unit increase in valence or loudness), and so including them in this model will increase predictive power for energy if we were to try to predict something like the energy index of a newly released track.

References

1. "The R Graph Gallery." Help and inspiration for R Charts. Accessed February 25, 2023. https://r-graph-gallery.com/index.html. #'
2. "Web Api." Get Tracks' Audio Features. Accessed February 24, 2023. https://developer.spotify.com/documentation/web-api/reference/#/operations/get-several-audio-features.