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

Tiffanie Huang

Monday 3/6/23 at 11:59pm

```r
tswift <- read.csv("../data/tswift.csv")
#libraries
library(ggplot2)
library(estimatr)
#devtools::install_github("wilkelab/ungeviz")
library(ungeviz)
#install.packages("emmeans")
library(emmeans)
```

Final Report

The Taylor Swift dataset supplied me with a variety of audio feature measurements, something I found valuable since an artist's discography can usually only be described by qualitative reports when it comes to audio features like valence or energy (where energy might be described in a relative manner–low/medium/high and valence might just be one-word emotions like happy or sad) if it was just based on the human ear. Because Spotify broke down these features quantitatively, I wanted to investigate how other variables might inform me about energy, as energy is measured in a less intuitive way than variables like loudness, tempo, or acousticness. Based on the limited things I know about the DGP for this dataset, I formed my question: how can I explain the variation in energy observed in Taylor Swift's discography? Some things I kept in mind included -valence ranges from negative emotions to positive emotions rather than different emotions that can't be quantified on a 0-1 scale, -there is a possibility that some of these variables are fed into calculations for more complex variables (e.g. tempo into danceability), and -popularity is measured based on most recent streams, making it hard to analyze across different albums that were released at different times. In addition, we don't know enough about the DGP, as neither "treatment"/randomization occurred nor were the conditions set up in a way that variables could be treated "as if" random; as a result, all of my descriptions are of non-causal relationships, and I am looking at an association between my independent and dependent variables through patterns in variation.

When examining the distributions of energy and acousticness, I found that the distribution for energy was relatively normal, while the distribution for acousticness was extremely skewed to the right, suggesting a higher number of less acoustic tracks. I wanted to analyze the relationship between energy and acousticness, setting acousticness as my independent variable due to its more straightforward definition and how a track's timbre is measured in acousticness. I also thought that given Taylor Swift's evolution from country music star to pop star, it would be interesting to be able to analyze this relationship using a dataset of discography that fully covers that extensive of an acousticness range. Visualizing the joint distribution of (Energy, Acousticness) shows us a relatively negative relationship between the two (Figure 3), which we can see in the negative covariance value (-0.3949246). To assess their conditional relationship, OLS linear regression with robust standard errors was carried out; we found that as acousticness increases by 1 unit, energy generally decreases (by 0.3949246 units). While both estimates for my first model are highly statistically significant when we conduct hypothesis tests for slope or some other estimate = 0 at a significance level of 0.01, there might still be other independent variables that could be taken into account to increase the precision of my model and predictive power. I regressed energy on acousticness again (Model 2), this time holding tempo constant; however, my slope coefficient did not change much, meaning most of the relationship we observe between energy and acousticness is not explained by tempo – "taking out" the part that is explained by tempo did not change the coefficient on acousticness but much. For my second multivariate model, I wanted to observe how my original coefficient might change once we hold valence and loudness constant. This time, the my slope coefficient estimate for acousticness did show a noticeable change compared to Model 1: -0.1561469 vs -0.3949246 (both significant at $\alpha = 0.01$). Figure 5 below shows the slope coefficient on my main independent variable for all 3 models, where there is nearly no difference between 1 and 2, while the slope becomes noticeably "flatter" when holding loudness and valence constant. Here, when we take out the part of energy conditional on acousticness that shows variation associated with valence and loudness, the remaining part of the relationship shows a smaller (-0.1561469) change in energy as acousticness increases by 1 unit. My dataset encompasses every population unit, but if this had been a sample/if there were any songs missing, it would become easier to predict energy based on how it varies with acousticness and additional variables rather than just acousticness by itself, since more information has been added to the model.

Future Steps for Additional Analysis

While the coefficient estimates for my models turned out to be significant at a significance level of 0.01, it is still worth visualizing and reporting estimates that don't have p-values <= 0.01. Future steps for analysis could include looking at the same conditional relationship conditional on albums so we can see how it might change depending on the album, as audio features may be very different from album to album; there might be some interesting intra-album associations that might not be visible when we analyze all observations at once due to the sheer size of Taylor Swift's discography. For example, Figure 6 shows the mean energy for each album (energy conditional on album) as well as +/-1 SE bars and 95% CI density strips to help illustrate uncertainty. We see that the energy estimates for some albums like Red might be more precise from their narrower CIs vs wider ones like Lover. Figure 7 is just a rough visualization for one of the "future steps" I mentioned above when looking at albums as an additional alternate specification for the conditional relationship between energy and acousticness, but we can also carry out the same idea with other variables/models we're interested in.

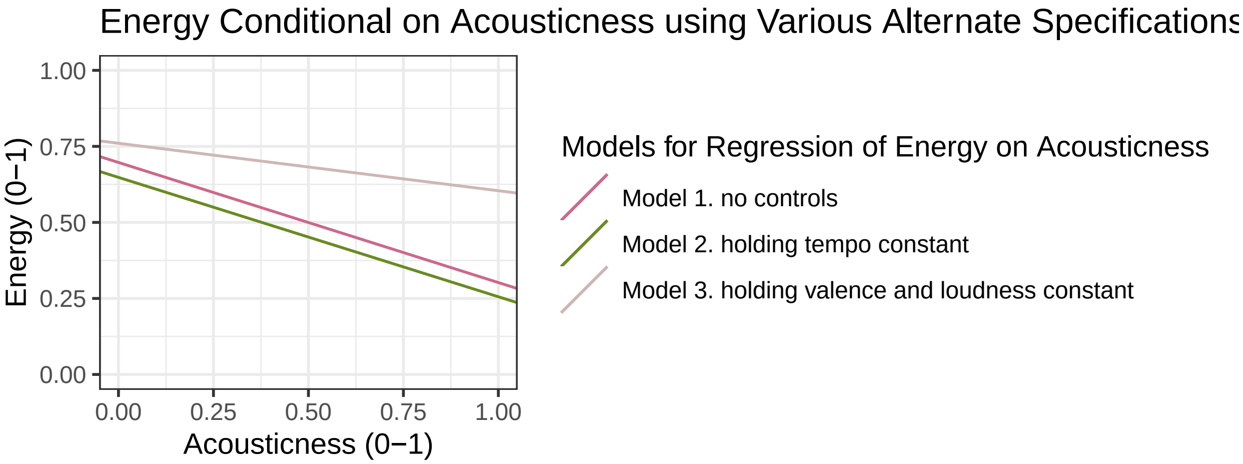
Fig. 5: Regressing energy on acousticness using 0, 1, and 2 controls

The plot below visualizes changes in the acousticness slope/intercept; technically they share different distributions and so it's not very accurate to compare them as whole models.

```
#Models 1-3 from part 2
bi_mod <- lm_robust(energy~acousticness,data=tswift) #no controls
#coef(bi_mod) #0.6971424-0.3949246X
multi_mod1 <- lm_robust(energy~acousticness+tempo,data=tswift) #1 control
#coef(multi_mod1) #0.6480846351-0.3925206030X1
multi_mod2 <- lm_robust(energy~acousticness+valence+loudness,data=tswift) #2 controls
coef(multi_mod2) #0.7602949-0.1561469X1
```

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

```
ggplot()+
  geom_abline(aes(intercept=0.6971424,slope=-0.3949246,color="Model 1. no controls"))+ #Model 1
  geom_abline(aes(intercept=0.6480846351,slope=-0.3925206030,color="Model 2. holding tempo constant"))+
  geom_abline(aes(intercept=0.7602949,slope=-0.1561469,color="Model 3. holding valence and loudness con
  xlim(0,1)+
  ylim(0,1)+
  theme_bw()+
  coord_fixed(ratio=0.8)+
  scale_color_manual(
    name="Models for Regression of Energy on Acousticness",
    values=c("Model 1. no controls"="palevioletred3",
             "Model 2. holding tempo constant"="olivedrab4",
             "Model 3. holding valence and loudness constant"="mistyrose3"))+
  labs(x="Acousticness (0-1)",
       y="Energy (0-1)",
       title="Energy Conditional on Acousticness using Various Alternate Specifications")
```



Additional Visualizations

```
#merge deluxe/regular albums
length(unique(tswift$album)) #20 unique albums; there should only be 10
```

```
## [1] 20
```

```
tswift$gen_albums <- sub(" \\(.*","",tswift$album)
length(unique(tswift$gen_albums)) #10 distinct albums now
```

```
## [1] 10
```

```
album_mod <- lm_robust(energy~gen_albums,data=tswift)
album_emm <- emmeans(album_mod,~gen_albums) #marginal means
album_df <- data.frame(album_emm)
album_df$gen_albums <- factor(album_df$gen_albums,levels=album_df$gen_albums[order(-album_df$emmean)],o
tswift_palette <- c("#CC99FF", #1989
                    "#09AFD6", #Taylor Swift
                    "#B67FED", #Speak Now
                    "#FACB2F", #Fearless
                    "#A8322D", #Red
                    "#000000", #reputation
                    "#FFADD5", #Lover
                    "#8A4300", #evermore
                    "#3F418F", #Midnights
                    "#ABABAB") #folklore
```

Fig. 6: CI strips for mean of energy for each album

```
ggplot(album_df, aes(x=emmean, y = gen_albums))+
  stat_confidence_density(aes(moe=SE), confidence = 0.95, fill = "#81A7D6", height = 0.7)+
  geom_errorbarh(aes(xmin=emmean-SE,xmax=emmean+SE),height=0.3)+
  geom_vpline(aes(x=emmean,color=factor(gen_albums)),height=0.7)+
  xlim(0,1)+
  scale_color_manual(values = tswift_palette)+
  theme_minimal()+
  labs(x="Mean Energy (0-1) ",
       y="Taylor Swift Albums",
       title="Marginal Mean of Energy and Uncertainty\n(95% CI Density)/Error (+/-1 SE) by Album")+
  theme(legend.position = "none")
```

```
## Warning: Using the ‘size‘ aesthetic in this geom was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ in the ‘default_aes‘ field and elsewhere instead.
```

```
## Warning: Removed 20 rows containing missing values (‘geom_tile()‘).
```

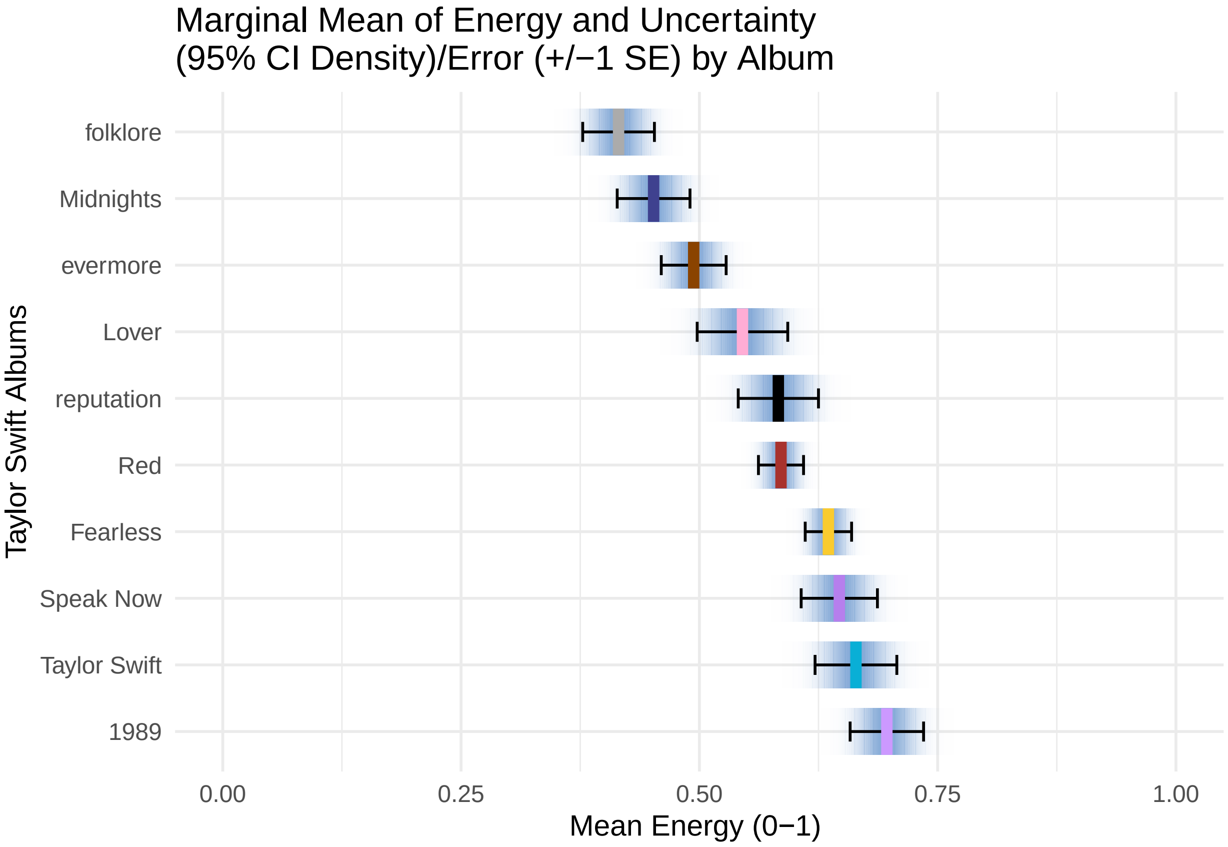
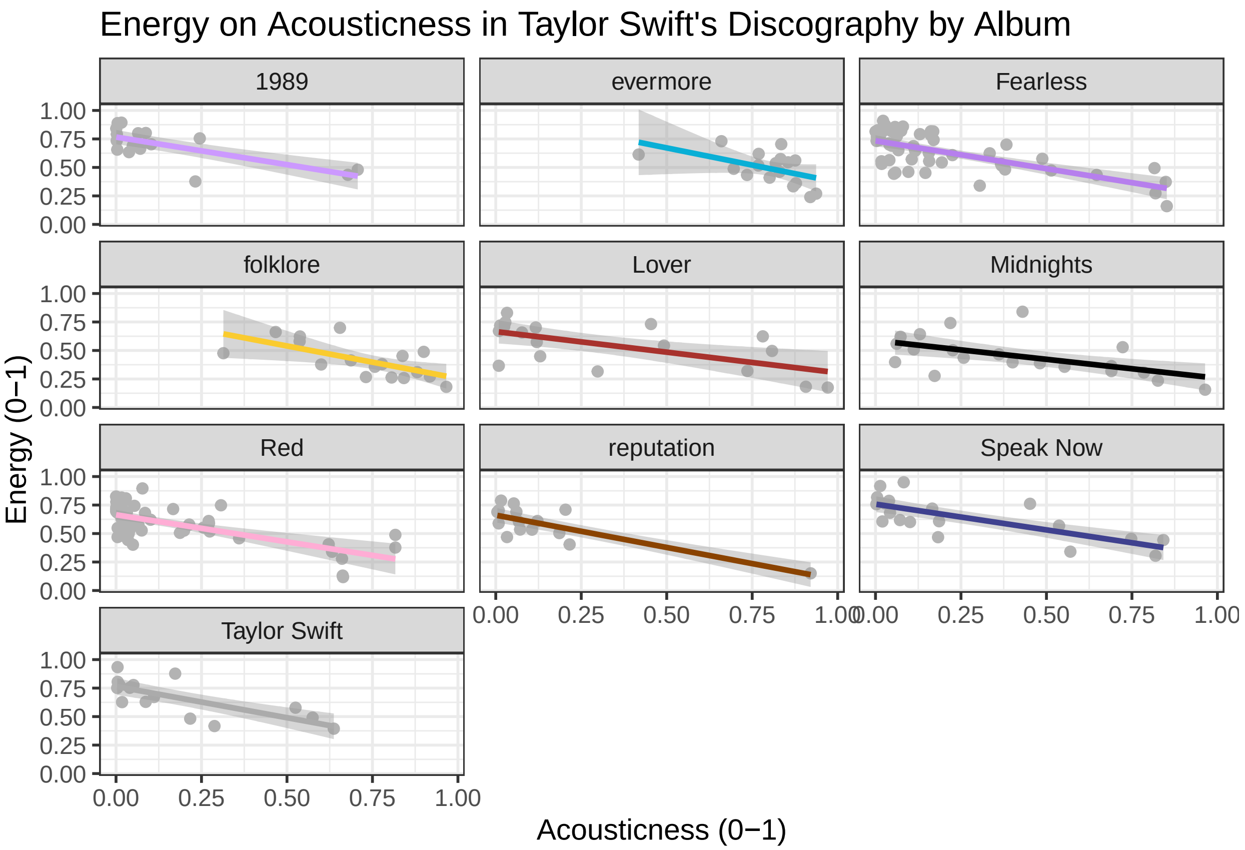


Fig. 7: Regressing energy on acousticness for each individual album

```
ggplot(tswift,aes(x=acousticness,y=energy))+
  geom_point(color="grey70")+
  geom_smooth(method="lm_robust",aes(color=gen_albums))+
  theme_bw()+
  labs(x="Acousticness (0-1)",
       y="Energy (0-1)",
       title="Energy on Acousticness in Taylor Swift's Discography by Album")+
  facet_wrap(~gen_albums,ncol=3)+
  scale_color_manual(values=tswift_palette)+
  theme(legend.position = "none")
```

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



Additional Notes

In Part 2, I forgot to include the definition for tempo: the pace of a track/overall beats per minute measurement. In addition, one motivation using valence as one of my alternate specifications in Model 3 despite it also being a more "complex" audio feature to measure along with energy would be if the emotion of a song might show some variation that we can associate with how acoustic the piece is and how energetic the piece is, since observed variation in acousticness associated with energy could have arose from something like valence that might also be associated with the acousticness of a piece.

References

1. "Miscellaneous geoms and stats." ungeviz. Accessed March 1, 2023. https://wilkelab.org/ungeviz/articles/misc-geoms-stats.html. #'
2. "emmeans: Estimated marginal means (Least-squares means)." RDocumentation. Accessed March 1, 2023. https://www.rdocumentation.org/packages/emmeans/versions/1.8.4-1/topics/emmeans. #'
3. Huntington-Klein, Nick, "The Effect." The Effect: An Introduction to Research Design and Causality. Accessed March 1, 2023. https://theeffectbook.net/ch-StatisticalAdjustment.html