ST344 Lab Report 5

1729346

04/11/2019

Task 1: Introduction, Loading Packages and Reading Data

This lab report studies the Spotify dataset which consists of albums from 1960s to 2010s and aims to investigate the relationship between a track's valence and danceability. The motivation behind studying this relationship comes from the conjecture that tracks with a higher valence should sound more positive, and that these tracks are more suitable for dancing. Furthermore, the report tries to distinguish this relationship for rap albums and non-rap albums by using the speechiness of the tracks. Variables considered:

- **TrackDanceability**: describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
- **TrackValence**: describes the musical positiveness conveyed by a track, measured from 0.0 to 1.0 . Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).
- **TrackSpeechiness**: Speechiness detects the presence of spoken words in a track. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music.

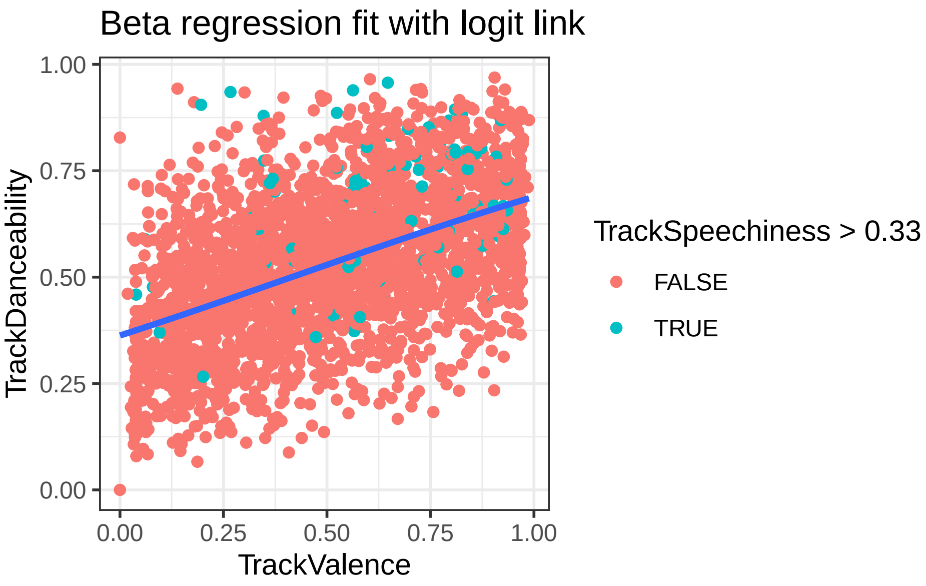
Following are the packages used in this report and the dataset imported:

```r
library(ggplot2)
library(readxl)
library(dplyr)
library(betareg)

Spotify <- read_excel("edited_spotify.xlsx")
```

Task 2: EDA

To investigate our initial conjecture, we fit a regression model with TrackDanceability as our response variable. We use a Generalised Linear Model which uses a beta distribution to model our response along with a logit link function. In order to use the betareg library, we need to replace the 0s in our response variable to a very small negligible number. In order to visualise the difference between rap albums and non-rap albums, we will colour code the data points depending on their speechiness.



As seen through the plot, there is a weak correlation implying tracks with high valence tend to have higher danceability. Note that there is no separation between rap and non-rap albums in this model and the colours are only for visualisation purposes.

Task 3: Valence vs Danceability in Rap Albums

We will create a nested sequence of 3 multiplicative Generalised Linear Models to exlpore the relationship between danceability and valence:

- **model1** is a simple glm which uses TrackDanceability as the response variable and TrackValence as the predictor variable. TrackSpeechiness is ignored here.
- **model2** uses the same response and predictor as model1, but has a different intercept for tracks with speechiness less than and more than 0.33.
- **model3** has different parameters in the linear predictor for speechy and non-speechy tracks.

The error distribution for each model is a log link function.

```
## Analysis of Deviance Table
##
## Model 1: TrackDanceability ~ TrackValence
## Model 2: TrackDanceability ~ TrackValence + factor(TrackSpeechiness >
##     0.33)
## Model 3: TrackDanceability ~ -1 + TrackValence + factor(TrackSpeechiness >
##     0.33) + factor(TrackSpeechiness > 0.33):TrackValence
##   Resid. Df Resid. Dev Df Deviance       F   Pr(>F)
## 1      2565     295.62
## 2      2564     292.56  1   3.0566 32.8887 1.09e-08 ***
## 3      2563     292.23  1   0.3362  3.6175  0.05729 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table gives us a variance analysis of our models and performs an F test to test whether our complex models are significantly better at capturing the data or not. The table includes p-values (row labelled 'Pr(>F)') which can be used to answer our question. To test our hypothesis at a 95% confidence, the p-value of our models must be <0.05. Clearly, model2 is a significant improvement on fit than model1 since the p-value is significantly small. However, model3 is not a significant improvement and fails the F-test.