STA303 - Final Project Report

Yakun Wang

April 7, 2024

1 Introduction

Music is an integral part of human culture and has been celebrated for its profound impact on emotions, cognition, and behavior. There has been an increased focus on mental health over the past few decades. Thus, it would be meaningful to conduct some research on the relation between music and mental health. In this report, we will focus on the research question "Will listening to music have a positive effect on mental health of people?", or statistically, "Can we predict the impact of music on mental health of people based on their personal information, music preferences, and self-reported mental health?".

Seeking to understand the link between musical involvement and mental health, Wesseldijk et al. conducted a study on a substantial cohort of Swedish twins. They evaluated multiple variables, including musical engagement, self-reported mental health, and educational levels. The findings indicated a marginal uptick in mental health issues among those engaged in music, and playing an instrument was strongly linked to a higher incidence of schizotypal symptoms, depressive and burnout symptoms in the workplace (Wesseldijk et al., 2019). In separate research, Rebecchini synthesized results from a multitude of scholarly articles, concluding that the burgeoning field of music-based therapeutic services in healthcare owes much to the growing evidence of the benefits of musical activities. Rebecchini underscored the significant potential of music to enhance both the physical and mental health of individuals across all demographics (Rebecchini, 2021). Additionally, a different study detailed the experiences of six mental health service users with music therapy, gathered through individual interviews and analyzed using Interpretative Phenomenological Analysis. These narratives revealed the complexity of music therapy participation, with one participant notably expressing that it acknowledged him as a unique and individual person (McCaffrey & Edwards, 2016).

Compared with the above academic research, this study differs from them because apart from variables like music engagement and self-reported mental health, we will also use data that are directly related to people's music preferences to fit a logistic regression model.

2 Methods

2.1 Model Fitting

In this study, we focus on how to predict the impact of music on mental health based on multiple measures, where the response variable "Music Effect" is a binary categorical variable with two outcomes: "Improved" and "No effect". The binary outcome can be treated as 1 for "Improved" and 0 for "No effect" in this case. Therefore, a logistic regression model would be an appropriate statistical model in our study to obtain accurate prediction. The proposed model can be expressed in the form as follows:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n = \mathrm{X}\beta$$

where π is the probability of success, or the probability of the outcome “Improved”.

2.2 Variable Selection

Once we obtain a full model, we will employ AIC, BIC, and LASSO for variable selection. Both AIC and BIC are methods used for model selection in statistics. They are similar while BIC applies a larger penalty for models with more parameters, resulting in a simpler model. Generally, a lower AIC/BIC indicates a better quality of a model.

LASSO is another method to improve the prediction accuracy of a statistical model. It imposes a penalty on the absolute size of the regression coefficients, forcing some coefficients to shrink towards zero. This effectively removes those variables from the model and only significant variables are kept. In addition, LASSO is particularly helpful when the number of predictors is large, and the number of observations is small.

2.3 Model Diagnostics and Validation

Before fitting a model, we perform an Exploratory Data Analysis (EDA) to analyze our dataset and summarize its main characteristics, which also helps to identify potential issues with model assumptions. We will investigate different figures of response variable and predictors, including barplot and histogram, to check skewness, outliers, or other underlying characteristics. No skewness and no outliers are desired.

After we obtain a reduced model, Cross Validation and ROC curve are used to validate it and evaluate its prediction accuracy. Through the Calibration plot generated from Cross Validation method, we can interpret the performance of model by the criterion that the closer the points are to the diagonal line, the better the prediction accuracy is. In terms of ROC curve, a key metric derived from the ROC curve is the Area Under the Curve (AUC), which provides a single measure of overall performance of a predictive model. AUC ranges from 0 to 1, and that the closer the value of AUC is to 1 indicates a better discrimination ability.

After finishing the evaluation of the prediction accuracy of the final model, we will use our test dataset to further validate our model. Notably, we have split the cleaned data into two parts: training set (75%) for fitting a model, and test set (25%) for model validation. By calculating the prediction accuracy of the final model using test set, we will again evaluate the model performance.

3 Results

3.1 Description of Data

The dataset used in this study contains information collected by a survey on music preferences and self-reported mental health. The original dataset has 736 observations with 33 variables, where 1 numerical variable represents the personal information of respondents, 4 numerical variables represent self-reported mental health, 26 variables contain music-related information, and the rest 2 variables are irrelevant to this study. Table 1 shows the summary for numerical variables in the dataset (See Figure 3 & 4 in Appendix for additional plots generated from EDA).

Variables	*Min*	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max*
Age	10.00	18.00	21.00	24.79	27.00	89.00
Hours	0.000	2.000	3.000	3.702	5.000	24.000
BPM	0	100	120	1623500	144	999999999
Anxiety	0.000	4.000	6.000	5.884	8.000	10.000
Depression	0.000	2.000	5.000	4.894	7.000	10.000
Insomnia	0.000	1.000	3.000	3.801	6.000	10.000
OCD	0.000	0.000	2.000	2.659	5.000	10.000

Table 1: Summary of the numerical variables

In the original dataset, there are 120 rows containing missing values, which have been deleted during the process of data cleaning because if there are missing values in one row, then it cannot be used to fit a statistical model and this row won't give us valuable information on predicted outcomes.

3.2 Analysis Process

To fit a model with high prediction accuracy, we perform the following steps:

a) Use all relevant variables in the dataset to fit a full logistic model.
b) Select variables stepwise in both directions using AIC, BIC, or LASSO to get significant variables.
c) Fit a new model with only the significant variables selected by these methods to obtain three model candidates.
d) Compare model candidates and choose the best fit. In this case, model selected by AIC will be the final model with the best prediction accuracy.

Coefficients	**Estimate**	**Coefficients**	**Estimate**
Intercept	-0.53776	Gospel_freqSometimes	-0.38312
While_workingYes	-0.65350	Gospel_freqVery frequently	-15.33883
ComposerYes	-0.79736	Pop_freqRarely	1.66559
ExploratoryYes	-0.54906	Pop_freqSometimes	1.02154
EDM_freqRarely	-0.75051	Pop_freqVery frequently	0.89284
EDM_freqSometimes	0.04457	Anxiety	-0.23104
EDM_freqVery frequently	0.02413	Depression	0.12235
Gospel_freqRarely	-0.54391	Insomnia	0.08803

Table 2: Coefficients of the final reduced model

3.3 Model Evaluation

To evaluate the quality of our final model, we utilized two methods in this study, Cross Validation and ROC curve.

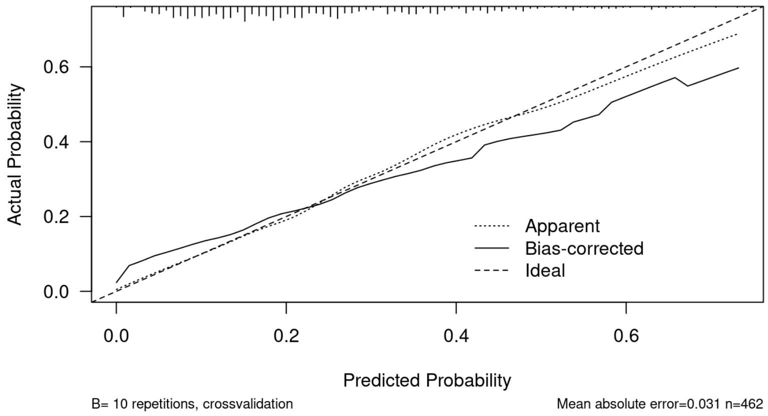

Figure 1: Calibration Plot generated from Cross Validation

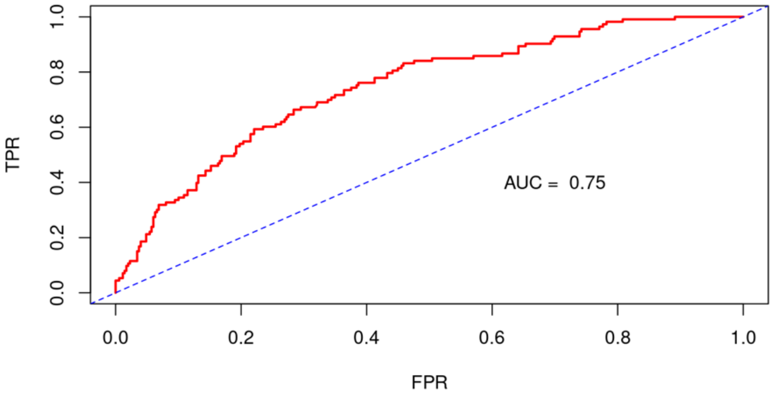

Figure 2: ROC curve

From Figure 1, we observe that biased-corrected line is relatively close to the diagonal line indicating the model's probability predictions are well-calibrated after bias correction. From Figure 2, AUC is 0.75, which means there is a 75% chance that the model will be able to distinguish between positive and negative classes. Generally, the final model has a good performance on prediction (See Figure 5 in Appendix for DFBETAS and Deviance Residuals).

4 Discussion

4.1 Conclusion

Given our final model, we can give a statistical interpretation. In terms of a numerical predictor, for example “Anxiety”, when self-reported anxiety increases by 1 while holding other predictors fixed, the odd ratio will be $e^{-0.23104} = 0.7937$. For a categorical variable, like “While working”, the odd ratio between the odd when “While working = Yes” and the odd when “While working = No", while holding other predictors fixed, is $e^{-0.6535} = 0.5202$. According to model validation and test set evaluation (calculated prediction accuracy is 91.56%), we conclude that the logistic regression model is an appropriate predictive model for our research and the impact of music on mental health of people can be statistically and relatively accurately predicted based on their personal information, music preferences, and self-reported mental health.

4.2 Limitations

The dataset used to fit a model has few observations, which may result in a low prediction accuracy when the final model is used to predict music effect with out-of-scope data. In addition, there may be a slight issue with Linearity Assumption since there exists skewness in the distribution of some predictors and there are some outliers in the dataset as well. Furthermore, there are some potential confounders that we did not consider when fitting the model. LASSO did not work on the full model with this dataset, where we may lose a potential model candidate to compare with other obtained models and choose the best fit. All these factors may cause our model to have a low prediction accuracy. Thus, it is required to fit a model with more data and the data should satisfy all the model assumptions.

Word count in main part: 1374 (excluding titles and captions)

Reference

McCaffrey, T., Edwards, Jane. (2016). “Music Therapy Helped Me Get Back Doing”: Perspectives of Music Therapy Participants in Mental Health Services. *Journal of Music Therapy, 53*(2), 121–148. doi:10.1093/jmt/thw002

Rebecchini, L. (2021). Music, mental health, and immunity. *Brain, Behavior, & Immunity – Health, Volume* 18. https://doi.org/10.1016/j.bbih.2021.100374

Wesseldijk, L.W., Ullén, F. & Mosing, M.A. (2019). The effects of playing music on mental health outcomes. *Sci Rep* 9, 12606. https://doi.org/10.1038/s41598-019-49099-9

Appendix

Additional plots generated when conducting an EDA on the data set

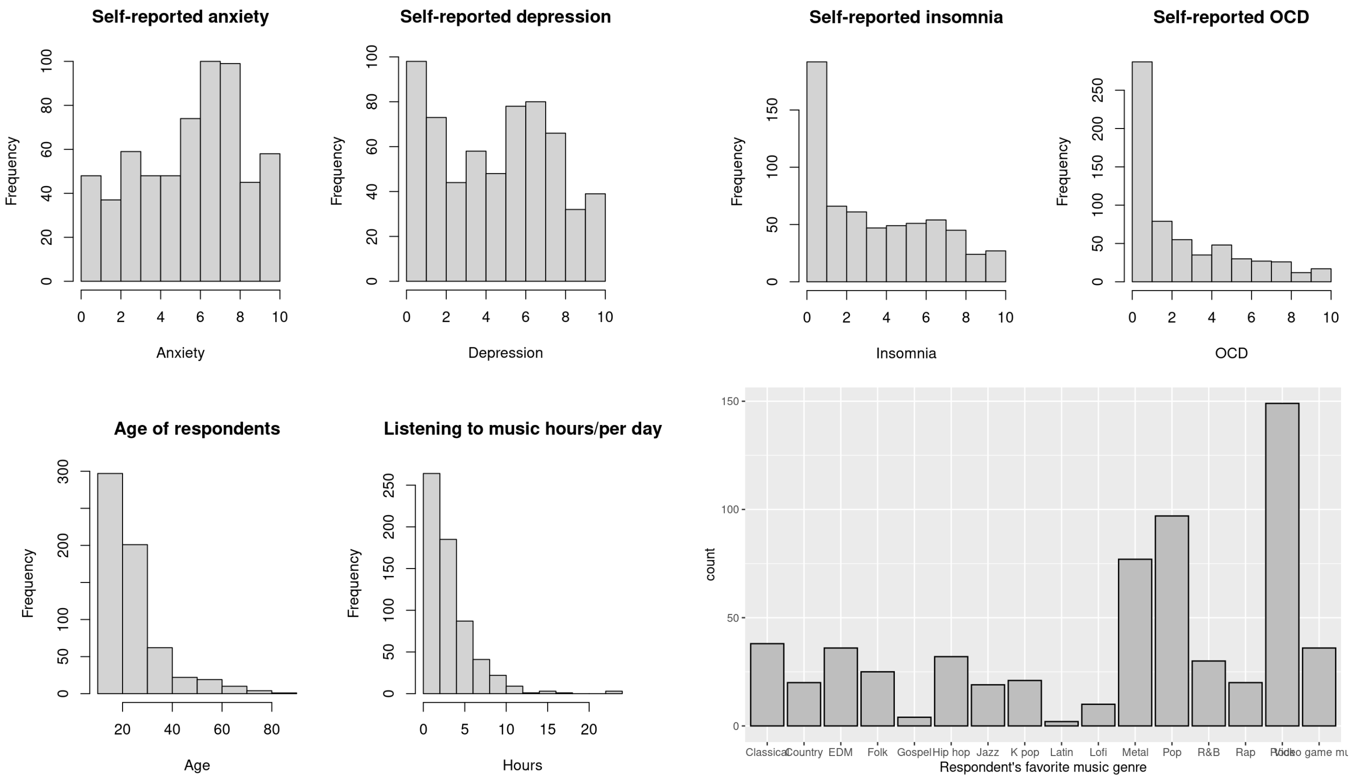

Figure 3: Histogram/Barplot of Variables in data set to conduct EDA

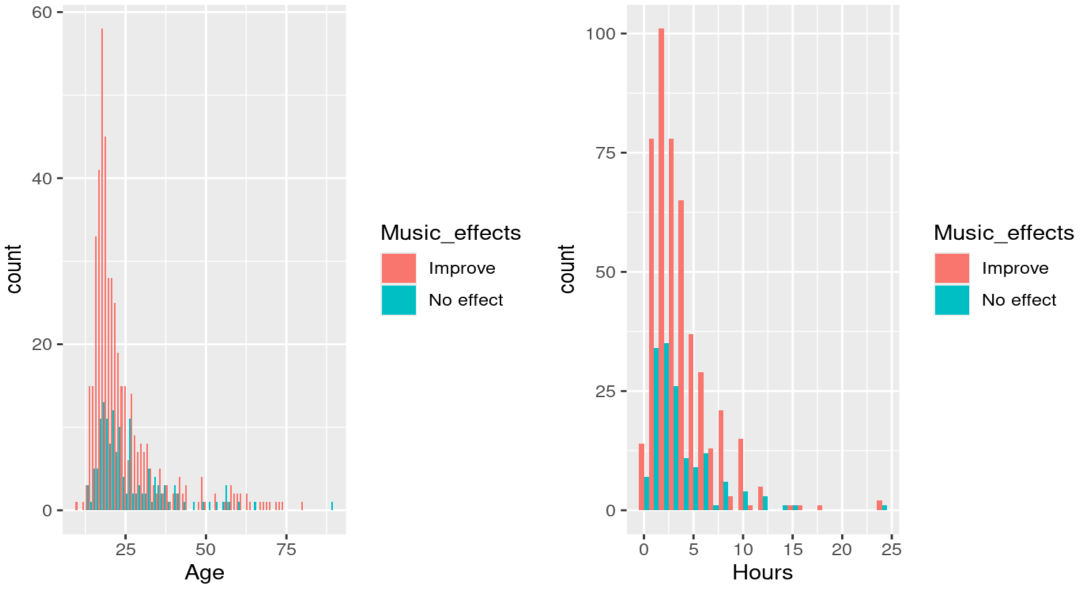

Figure 4: Histogram of the Relation between Numerical Variable and Response Variable

Additional plots generated when performing model validation

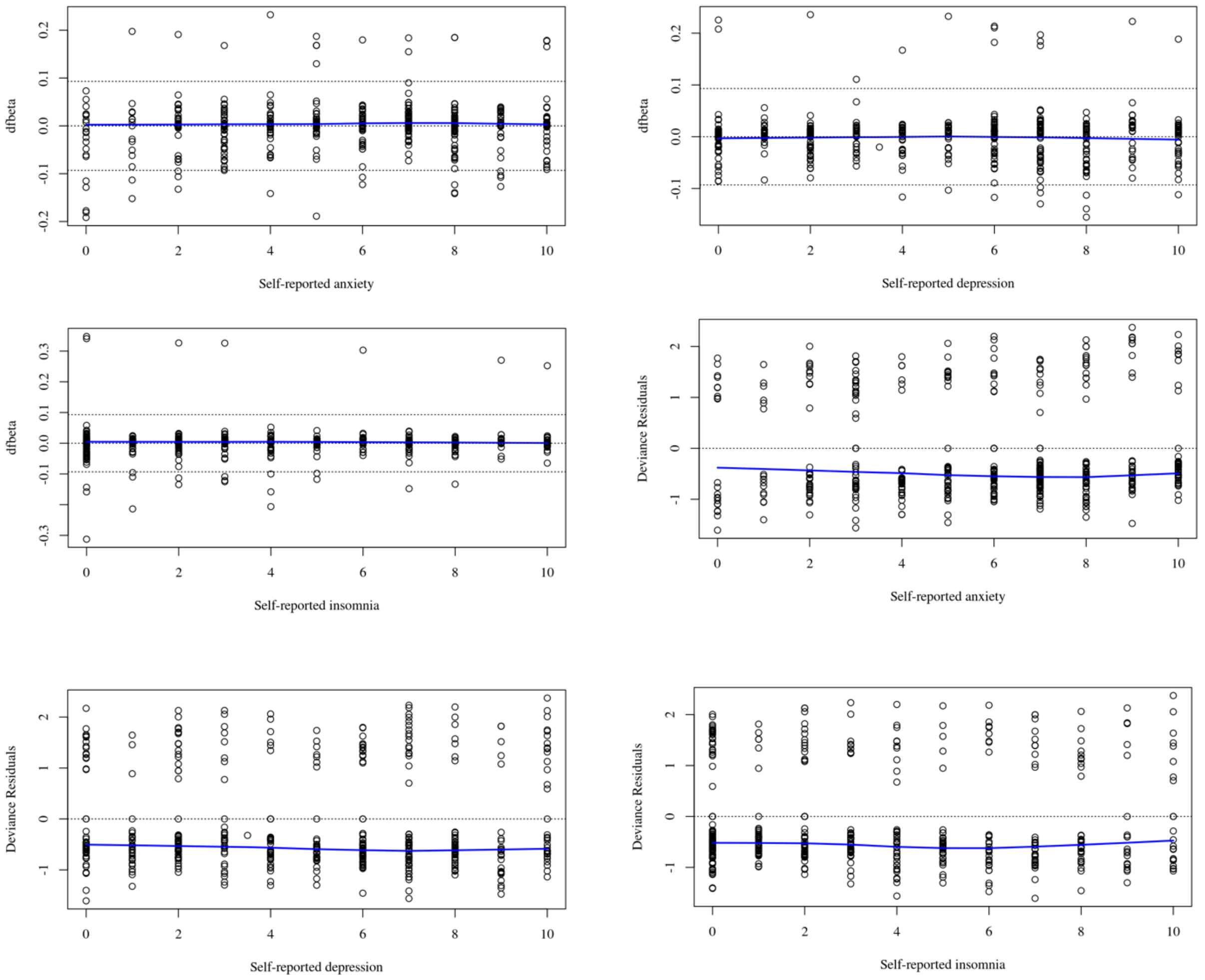

Figure 5: DFBETAS and Deviance Residuals

Interpretation of Figure 5:

From the DFBETAS plots (R1C1, R1C2, R2C1), there are some points exceeding the reference lines, which indicates these observations have a substantial influence on the parameter estimates. However, these influential points only accounts for a very small proportion and don’t significantly exceed the lines, so it is still acceptable. From the Deviance Residual plots (R2C2, R3C1, R3C2), no obvious patterns of residuals can be observed in these plots and these residuals are randomly distributed around zero within a reasonable range.