Using Music Listening Behaviors to Predict Levels of Depression

Emma Hughes (eah9921) & Sarvar Khamidov (sk10583)

Final Report – Regression I: Linear Regression and Modeling

GPH-GU 2353, Section 002

Spring 2025

Abstract

This study explored the relationship between music-listening behaviors and depression scores using data from the Music & Mental Health (MxMH) Survey. A multiple linear regression model was built to predict depression based on music engagement habits and demographic variables. The model explained 18.2% of the variance ($R^2 = 0.182$) and generalized reasonably well to new data (RMSE: 3.09 train, 2.74 test). Age was negatively associated with depression, while genre-specific patterns emerged: selecting Lofi as a favorite genre was linked to higher depression scores, but frequent listening to Lofi was associated with lower depression. Classical music frequency was positively associated with depression, while Country music showed a protective effect. Model diagnostics confirmed assumptions were largely satisfied. Limitations included missing potential confounders (e.g., gender, race/ethnicity), a cross-sectional design preventing causal inference, and a young, non-representative sample. Future research should use longitudinal designs and broader covariates to better understand how music engagement influences a range of mental health outcomes.

Introduction

Music has been recognized as a powerful tool for emotional expression and psychological well-being. Music therapy is a clinical practice that utilizes music to improve the health and well-being of individuals. It can also help people to reduce stress, change mood, and enhance their mental health. Previous studies have found a positive association between music interventions and health-related quality of life (McCrary, 2022). The findings from this project may help improve the knowledge of psychologically beneficial music genres, ultimately elevating the validity and effectiveness of music therapy for mental health issues like depression.

This analysis aims to predict depression levels using various music-listening behaviors (i.e. how frequent one listens to a specific genre) while adjusting for accompanied demographic data. By using linear regression, the association between a wide range of genres and self-reported depression levels can be explored in order to improve upon current music therapy methodologies. However, utilizing this approach presents potential challenges of multicollinearity, non-linearity, and heteroscedasticity, all which can affect model validity, interpretation, and prediction accuracy. To assess these common issues, diagnostic tools will be employed.

Data for this project was sourced from the *Music & Mental Health (MxMH) Survey* on Kaggle (Rasgaitis, 2020). The survey was advertised, distributed, and completed online by respondents. Survey links were posted on various social media platforms and through QR codes on flyers in libraries and public parks. Because the survey was primarily posted on social media platforms, the sample age was skewed right (Appendix A, Table 1) which may limit the generalizability to other age groups. Additionally, as the survey was conducted during the height of the COVID-19 pandemic, it must be acknowledged that self-reported depression levels may also be skewed, but to the left, due to heightened stress and lack of social interaction during that time period.

The following sections provide detailed information on this analysis, including data preprocessing steps, key results and interpretations, and implications of the findings.

Methods

Data Description

This project uses data from the *MxMH Survey*, which includes responses from 736 individuals. The dataset, as it was mentioned before, was collected as part of a cross-sectional study examining the relationship between music and mental health with each row corresponding to a unique participant.

The final analytic sample included 728 participants after removing observations with missing responses. Since there was only a small fraction of responses that had missing values we decided to just drop. Also, each response value for frequency of listening to particular music genres was transformed into continuous variables. The values had quantitative ordered connotations Never - 0, Rarely - 1, Sometimes - 2, Very frequently - 3. Also, the rest of the categorical variables were factorized in R and used directly with linear regression function since the function automatically transforms them into one hot encoding.

The final dataset included 37 variables, both continuous and categorical: self-reported depression score, hours of listening to music per day, whether they play a musical instrument, if they compose any music, favorite music genre, if they like to explore new music, do they listen to any foreign music and frequency of listening to each music genre (Classic, Country, EDM, Folk, Gospel, Hiphop, Jazz, Kpop, Latin, Lofi, Metal, Pop, RnB, Rap, Rock, Video Game Music).

Descriptive Analysis

To explore the music survey data, descriptive research was conducted. Table for continuous variables (Appendix A, Table 1) was created summarizing using means, standard deviation, median and IQR. Categorical variables were summarized (Appendix A, Table 2) using frequency tables with proportions.

Linear Regression Modeling

Multiple Linear Regression was used for prediction with depression score as the outcome and remaining variables as predictors. The data was split into 80/20 train-test split for model validation. To assess the relationship between depression and the predictors, parameter coefficients were estimated in the model output along with its 95% confidence interval and p-value. All analyses employed an alpha-level of 0.05 to identify statistically significant variables. In turn, the coefficients allow for interpretation of the association between depression and relevant predictors in terms of direction and magnitude. The model was evaluated using R-squared, adjusted R-squared, and RMSE metrics.

To ensure validity and accuracy of the model, linear regression assumptions were assessed by constructing multiple diagnostic plots (Appendix C).

- Linearity – residuals vs. fitted values
- Normality of residuals – Q-Q plot
- Homoscedasticity – scale location plot
- Outliers influential points – residuals vs. leverage residuals; Cook’s distance

Multicollinearity was assessed using Variance Inflation Factors (VIF). A VIF threshold of 5 was used to identify potential multicollinearity concerns among predictors. Variables exceeding this threshold were reviewed to ensure model stability and interpretability.

Results

Descriptive Analysis

Across the 728 observations in the final analytic sample, participants had a mean depression score of 4.796 (SD = 3.03) with a median of 5 and interquartile range (IQR) of 2-7, representing a relatively normal distribution. Conversely, the median age of participants was 25.21 (SD = 12.05) years old with a median of 21 (IQR = 18-27), suggesting a right-skewed distribution with a higher proportion of younger adults. Similarly, the number of hours a person listened to music per day was right-skewed (mean = 3.57 [SD = 3.03]; median = 3 [IQR = 2-5]) (Appendix A, Table 1).

Regarding additional music-listening habits and behaviors captured dichotomously and/or categorically, 31.93% of the sample identified as someone who plays an instrument, 17.12% as a composer or someone who writes music, 71.33% explore new genres, and 54.89% listen to music in a language other than English. In terms of favorite genres of music, a diverse selection was reported with the largest proportion (25.54%) listing Rock as their favorite genre, followed by Pop (15.49%), and Metal (11.96%) (Appendix A, Table 2). Another measurement collected was how often the respondent listened to each of the 16 genres, Rock music has the largest proportion of listening time with 44.84% having reported listening “Very Frequently”–other genres people report as listening to “Very Frequently” include Pop (37.64%), Metal (19.84%), Rap (17.11%), Hip Hop (16.71%), Video Game Music (15.896%), R&B (15.76%), and Classical (14.67%) (Appendix A, Table 3).

Parameter Estimation, Inference, and Interpretation

From the multiple linear regression model, at an alpha-level of 0.05, four of the 37 variables were found to be statistically significant predictors of depression (Appendix B, Table 4):

1. Age was negatively associated with depression score, inferring that as an individual’s age increases by year, their depression score decreases by 0.03 units (β = -0.03; 95% CI = [-0.06, -0.01]; p-value = 0.018) per year.
2. Of the Favorite Genre responses, Lofi was the only genre with a statistically significant relationship to depression, showing that those who selected Lofi as

their favorite genre experienced 2.98 unit increase in depression level on average, compared to those who selected Classical (reference group).

    - While not statistically significant, it should be acknowledged that for frequent lofi listening showed as a protective factor against depression (β = -0.27, 95% CI = [-0.58, 0.04], p-value = 0.087)

- Of the Frequency responses, Classical and Country were the two genres with significant relationships to depression.

3. The model suggests that the more frequently someone listens to Classical music, their depression increases by 0.33 units.

4. Conversely, the findings show those who listened to country music more frequently had depression scores decrease by 0.48 units.

Model Performance

The multiple linear regression model predicting depression score had an R-squared of 0.182, indicating that approximately 18.2% of the variance in depression scores is explained by the predictors in the model. The RMSE values were 3.09 for the training set and 2.74 for the test set, suggesting the model generalizes reasonably well to unseen data. The relatively low R-squared, however, implies that a substantial portion of variability in depression scores remains unexplained by the current predictors.

Model Diagnostics

Diagnostic checks were conducted to verify the assumptions of linear regression. Figure 4 represents the Residual vs Fitted plot which was used to check the assumption of linearity and homoscedasticity. The red line represents a smooth fit of residuals very slightly curved which indicates a little non-linearity.

Figure 5 (Appendix C) shows the Q-Q plot which is used to assess the assumption of normality in a linear regression. It compared the standardized residuals to a normal distribution. Most points fall fairly close to the diagonal, indicating that the residuals are approximately normally distributed. Slight deviations from the line are visible in both tails (especially the upper-right), suggesting mild skewness or presence of a few outliers.

Scale-Location plot Figure 6 (Appendix C) was used to assess the assumption of homoscedasticity, whether the residuals have constant variance across fitted values. We can see that the plot has a red smooth line of square root of standardized residuals on the y-axis versus fitted values on the x-axis. The line is relatively flat, but there is a

small bump around value of 4. However, homoscedasticity assumption is mostly satisfied, though non severe indication of non-constant variance.

Figure 7 (Appendix C) Residuals vs. Leverage plot helps us to identify the influential observations in the multiple linear regression model, those points might have an impact on the regression coefficients. We could see that most points are clustered in lower leverage regions suggesting typical influence. There are few points like 568, 683 and 171 that have higher leverage, which could be influential points. However, none of them break out of Cook's distance.

Lastly, we ran Variance Inflation Factor function to examine the multicollinearity among predictor variables. Most variables were below 2, indicating low multicollinearity. Interestingly, the variable favourite genre with 15 levels, had high raw GVIF (~55.2), but its adjusted value (GVIF^(1/(2*Df)) = 1.14) was well within acceptable limits. The highest values were identified for FrequencyHipHop 1.8 and FrequencyRap 1.78 which were also below the threshold of 5. So these would suggest that multicollinearity wasn't an issue for our model.

Discussion

The research's objective was to build a predictive model using multiple linear regression models and investigate the results by finding the most important variables for questions. The model that was built on 728 responders demonstrated fair predictive results on the train and test data, with relatively low RMSE. Additionally, findings like negatively associated age to the depression score and positive relationship of frequency of listening to lofi music with depression was important. Moreover, the study offers valuable insights into the complex and contradictory role that music plays in mental health. While age has a robust and clear relationship with depression, the music-listening-related predictors posed contradictory findings. As the model presents lofi as a favorite genre associated with higher levels of depression, frequent listening of lofi was associated with lower depression scores. The contradicting parameter coefficients suggest that individuals with depression may be using lofi as a coping mechanism and because of frequent listening, may also identify lofi as their favorite genre. The positive correlation between classical music and depression poses a similar idea that individuals with depression are listening to it more frequently as a coping mechanism. On the other hand, country music had a negative association which could suggest mood-lifting benefits and effects of the genre.

The model posits nuance and complexity of the relationship between music-listening habits and depression. The role music can play in mental health is multifaceted and must be further explored to investigate how music may be used for mental health interventions.

This project included few limitations. First, the dataset didn't include important confounders affecting the research question. Variables like gender, race/ethnicity, genetics, diet etc. are essential to influence mental health outcomes. There could have been inaccurate results and lead to incorrect conclusions about the relationship between music listening preferences and depression score. Also, data we worked with was cross-sectional which didn't allow us to draw causal inferences between predicting variables and depression score. Additionally, there could have been self-reported bias that led to inconsistent and incorrect conclusions. Responders due to various factors including social desirability, memory limitation, misinterpretation of questions might have not answered questions correctly. Finally, as we have seen in the descriptive analysis and Table 1 (Appendix A) our data wasn't representative of the whole population. The mean age of our participants was around 25 years. This would lead to a generalizability issue of findings to the older population.

First of all, future research should address the limitations of this project by using longitudinal study design. That would allow us to make causal inferences and better assess changes in mental health over time with different genres of music. Also, the mentioned above limitation - lack of covariates, should be addressed by adding more relevant variables, including demographic and mental health related. Inclusion of these variables would strengthen the validity of future models. Additionally, in the future researchers should consider other types of mental health conditions besides depression, such as anxiety, insomnia, and obsessive-compulsive disorder (OCD) which would allow better understanding of the broader impacts of music engagement on psychological well-being.

Appendix A

Table 1. Continuous variables from MxMH Dataset

	Mean (SD) Median (IQR)
Depression Level	4.796 (3.03) 5 (2-7)
Age	25.21 (12.05) 21 (18-27)
Hours listened to per day	3.57 (3.03) 3 (2-5)

Table 2. Categorical variables from MxMH Dataset

	n (%)
Instrumentalist	
Yes	235 (31.93%)
No	497 (67.53%)
Composer	
Yes	126 (17.12%)
No	609 (82.74%)
Explores new genres	
Yes	525 (71.33%)
No	211 (28.67%)
Listens in foreign language	
Yes	404 (54.89%)
No	328 (44.57%)
Favorite genre	
Classical	53 (7.20%)
Country	25 (3.396%)
EDM	37 (5.03%)
Folk	30 (4.08%)
Gospel	6 (0.82%)
Hip hop	35 (4.76%)
Jazz	20 (3.53%)
K pop	26 (3.53%)
Latin	3 (0.41%)
Lofi	10 (1.36%)
Metal	88 (11.96%)
Pop	114 (15.49%)
R&B	35 (4.76%)
Rap	22 (2.989%)
Rock	188 (25.54%)
Video game music	44 (5.97%)

Table 3. Frequency of [Genre] Predictor from MxMH Dataset

	n (%)			
	Never	**Rarely**	**Sometimes**	**Very Frequently**
Classical	169 (22.96%)	259 (35.19%)	200 (27.17%)	108 (14.67%)
Country	343 (46.60%)	233 (31.66%)	111 (15.08%)	49 (6.66%)
EDM	307 (41.71%)	194 (26.36%)	146 (19.84%)	89 (12.09%)
Folk	292 (39.67%)	221 (30.02%)	145 (19.70%)	78 (10.597%)
Gospel	535 (72.69%)	135 (18.34%)	52 (7.07%)	14 (1.90%)
Hip hop	181 (24.59%)	214 (29.08%)	218 (29.62%)	123 (16.71%)
Jazz	261 (35.46%)	247 (33.56%)	175 (23.78%)	53 (7.20%)
K pop	416 (56.5%)	176 (23.91%)	67 (9.10%)	77 (10.46%)
Latin	443 (60.19%)	172 (23.37%)	88 (11.96%)	33 (4.48%)
Lofi	280 (38.04%)	211 (28.67%)	160 (21.74%)	85 (11.55%)
Metal	264 (35.87%)	192 (26.09%)	134 (18.21%)	146 (19.84%)
Pop	56 (7.61%)	142 (19.29%)	261 (35.46%)	277 (37.64%)
R&B	225 (30.57%)	211 (28.67%)	184 (25.00%)	116 (15.76%)
Rap	200 (27.17%)	215 (29.21%)	195 (26.49%)	126 (17.11%)
Rock	91 (12.36%)	96 (13.04%)	219 (29.76%)	330 (44.84%)
Video game music	236 (32.07%)	197 (26.77%)	186 (25.27%)	117 (15.896%)

Figure 1. Pie Chart of Favorite Genres (n = 736) from MxMH Dataset

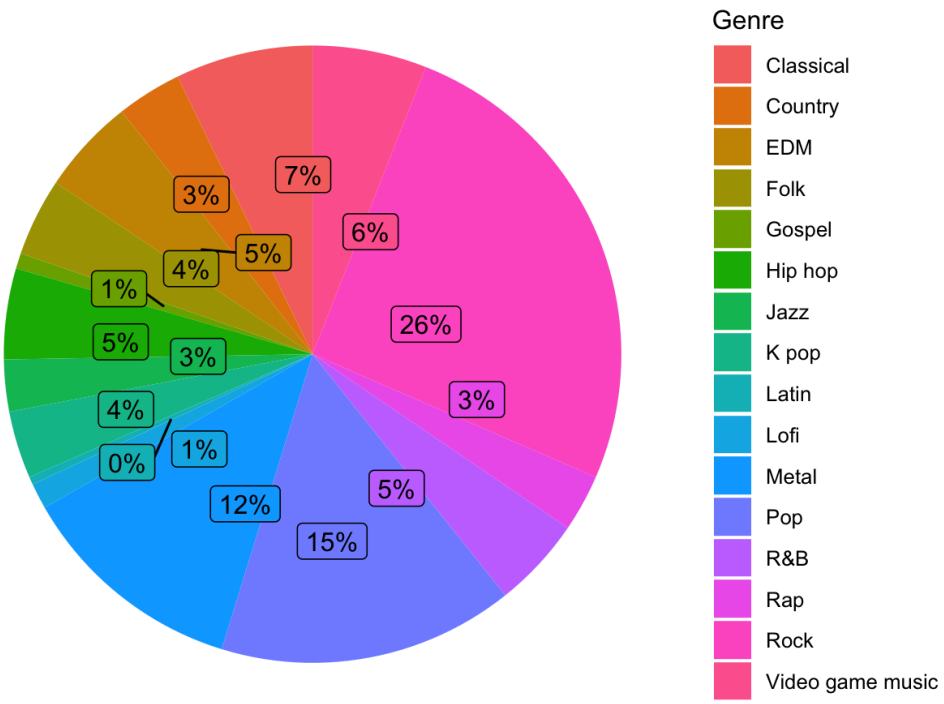
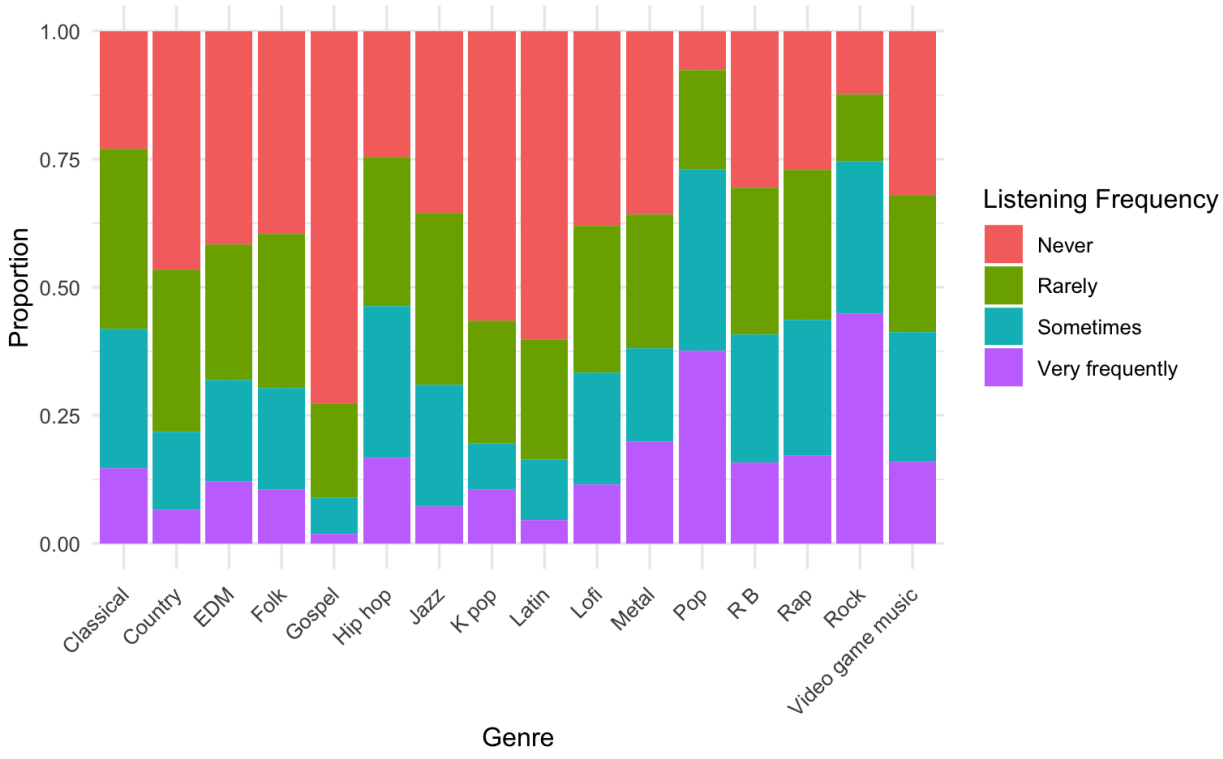


Figure 2. Proportions of Response Type to Listening Frequency Predictor per Genre (n =736) from MxMH Dataset

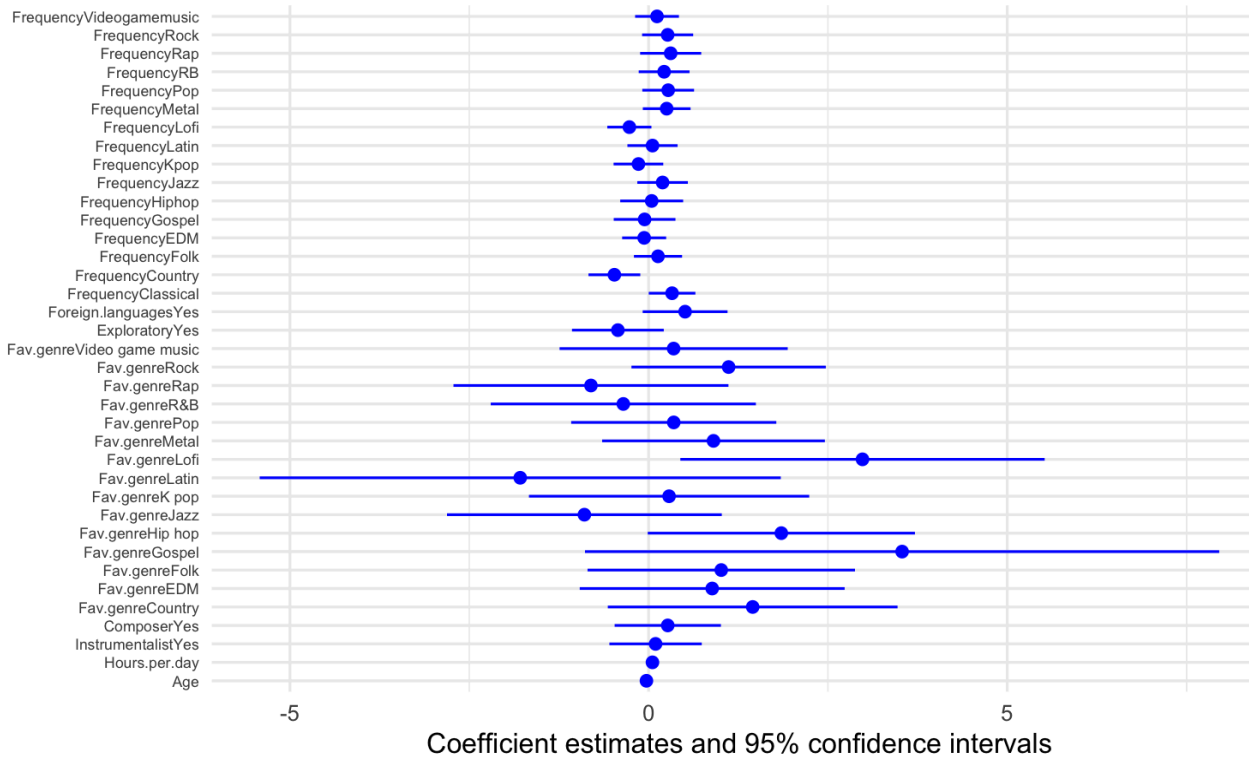


Appendix B

Table 4. Linear Regression Model Output Summary, Predicting Depression

	Coefficient	95% CI	p-value
Age	**-0.03**	**(-0.06, -0.01)**	**0.018**
Hours listened to per day	0.05	(-0.03, 0.14)	0.232
Instrumentalist	0.10	(-0.55, 0.74)	0.767
Composer	0.27	(-0.47, 1.01)	0.479
Explores new genres	-0.43	(-1.07, 0.21)	0.189
Listens in foreign language	0.51	(-0.08, 1.10)	0.092
Favorite genre - Country	1.45	(-0.57, 3.47)	0.159
Favorite genre - EDM	0.89	(-0.96, 2.73)	0.346
Favorite genre - Folk	1.01	(-0.85, 2.88)	0.286
Favorite genre - Gospel	3.53	(-0.89, 7.96)	0.117
Favorite genre - Hip hop	1.85	(-0.01, 3.71)	0.052
Favorite genre - Jazz	-0.90	(-2.81, 1.02)	0.359
Favorite genre - K pop	0.29	(-1.67, 2.24)	0.774
Favorite genre - Latin	-1.79	(-5.42, 1.84)	0.333
Favorite genre - Lofi	**2.98**	**(0.44, 5.52)**	**0.021**
Favorite genre - Metal	0.90	(-0.65, 2.46)	0.253
Favorite genre - Pop	0.35	(-1.08, 1.78)	0.630
Favorite genre - R&B	-0.35	(-2.20, 1.495)	0.708
Favorite genre - Rap	-0.80	(-2.72, 1.11)	0.410
Favorite genre - Rock	1.12	(-0.24, 2.47)	0.106
Favorite genre - Video game music	0.35	(-1.24, 1.94)	0.667
Frequency - Classical	**0.33**	**(0.001, 0.65)**	**0.049**
Frequency - Country	**-0.48**	**(-0.84, -0.12)**	**0.010**
Frequency - EDM	0.13	(-0.20, 0.47)	0.443
Frequency - Folk	-0.06	(-0.37, 0.24)	0.692
Frequency - Gospel	-0.06	(-0.49, 0.37)	0.799
Frequency - Hip hop	0.04	(-0.396, 0.48)	0.847
Frequency - Jazz	0.19	(-0.16, 0.55)	0.277
Frequency - K pop	-0.14	(-0.49, 0.20)	0.422
Frequency - Latin	0.05	(-0.295, 0.40)	0.760
Frequency - Lofi	-0.27	(-0.58, 0.04)	0.087
Frequency - Metal	0.25	(-0.08, 0.58)	0.137
Frequency - Pop	0.27	(-0.09, 0.63)	0.137
Frequency - R&B	0.22	(-0.14, 0.57)	0.231
Frequency - Rap	0.31	(-0.12, 0.73)	0.156
Frequency - Rock	0.27	(-0.09, 0.62)	0.143
Frequency - Video game music	0.12	(-0.19, 0.42)	0.451

Figure 3. β Coefficients for Predicting Depression from Music-Listening Behaviors of MxMH Dataset (Depression ~ All Predictors)



Appendix C

Figure 4. Residuals vs. Fitted Plot

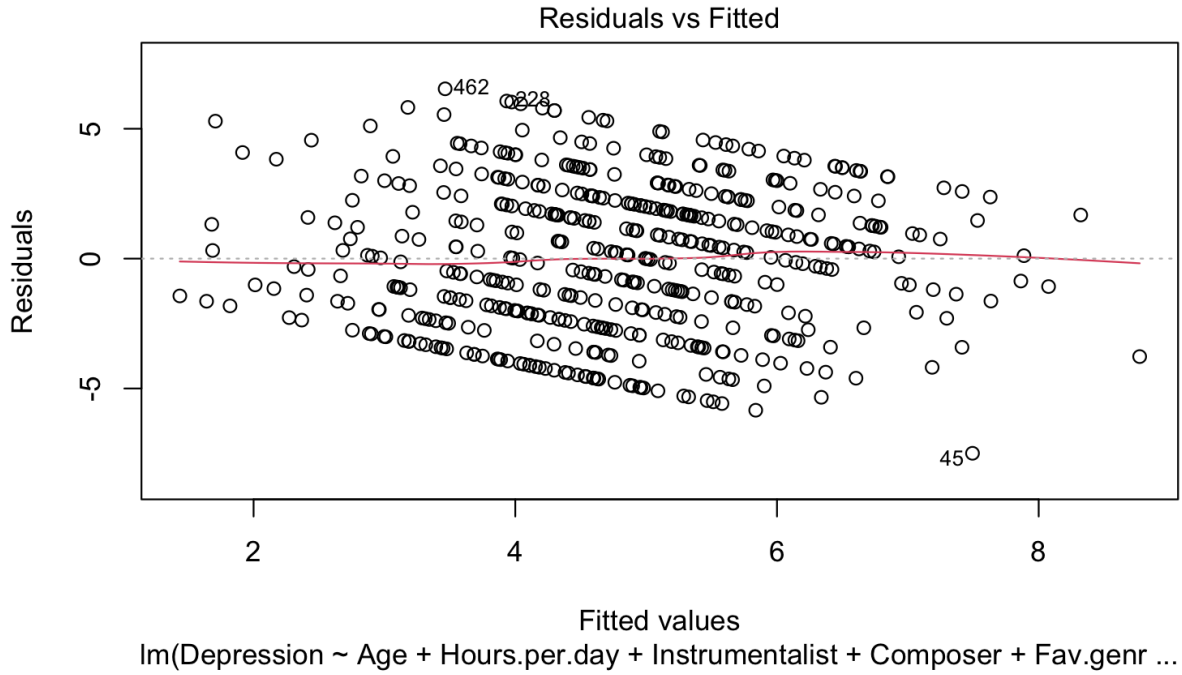


Figure 5. Q-Q Residuals Plot

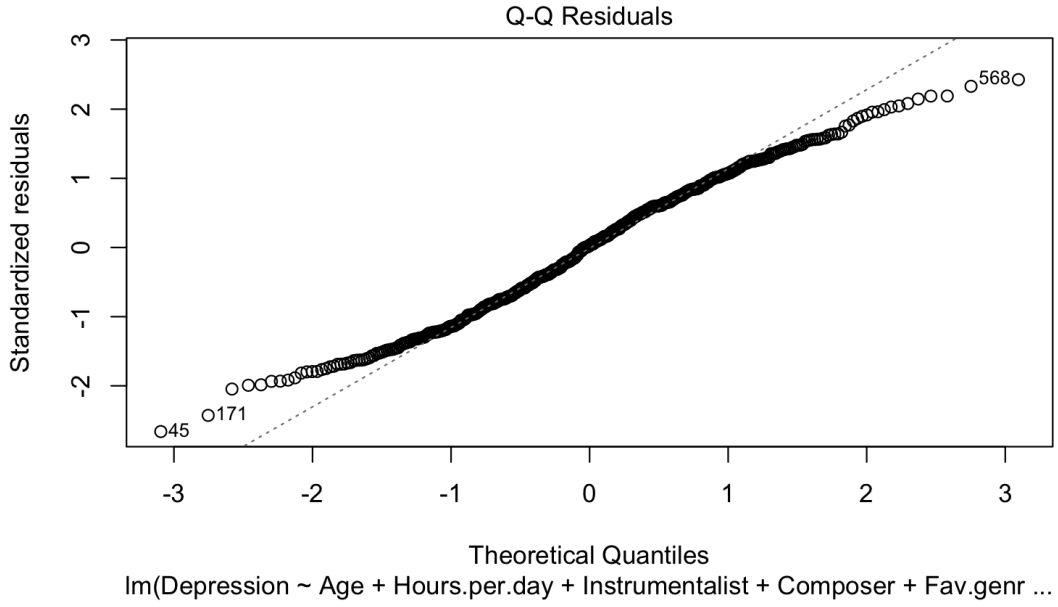


Figure 6. Scale-Location Plot

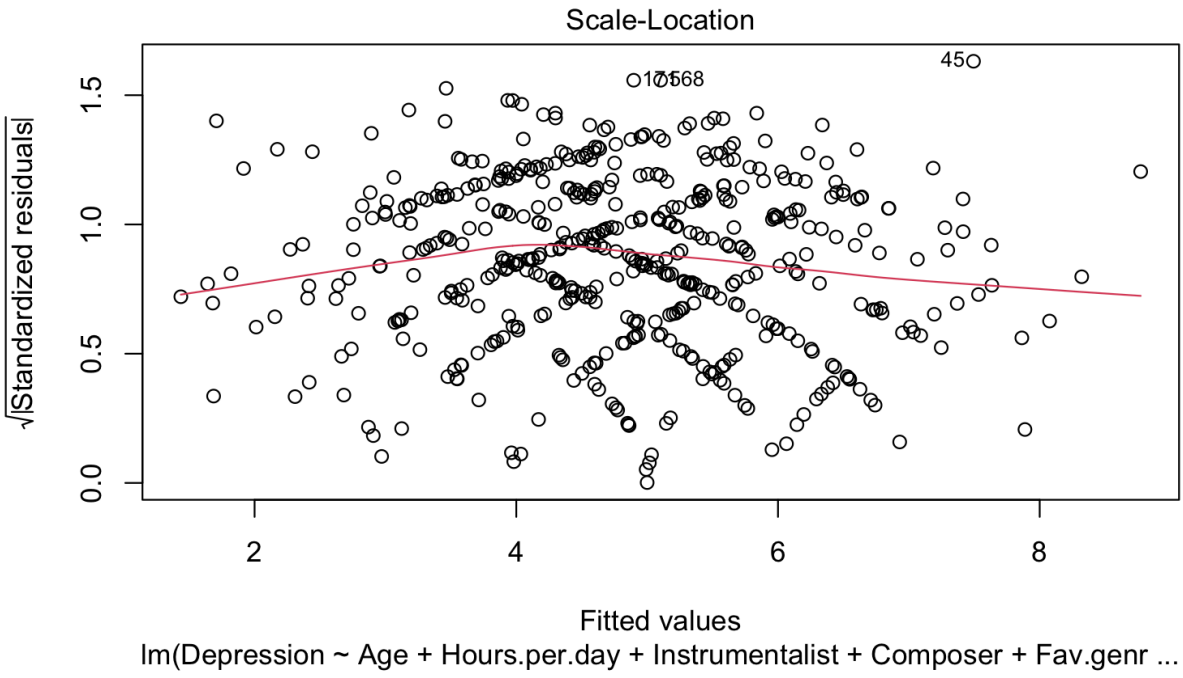
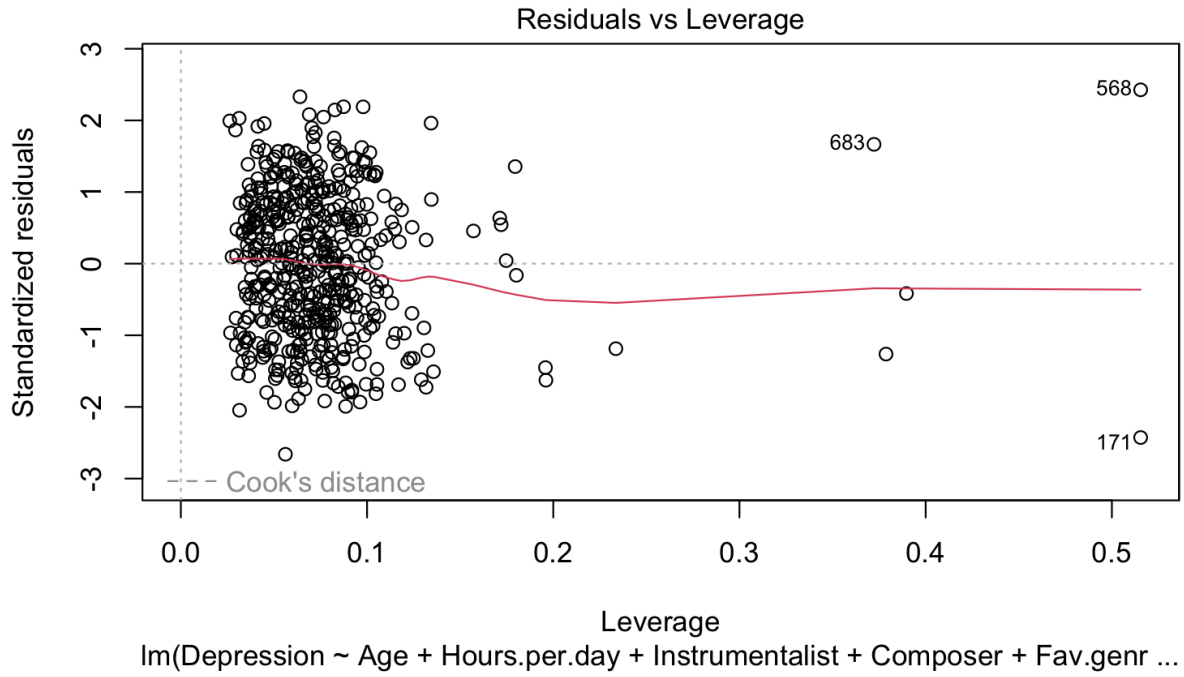


Figure 7. Residuals vs. Leverage Plot



References

McCrary, J. M., Altenmüller, E., Kretschmer, C., & Scholz, D. S. (2022). Association of Music Interventions With Health-Related Quality of Life: A Systematic Review and Meta-analysis. *JAMA Network Open*, *5*(3), e223236. https://doi.org/10.1001/jamanetworkopen.2022.3236

Rasgaitis, C. (2022, July 27). *Music & Mental Health Survey Results*. Kaggle. https://www.kaggle.com/datasets/catherinerasgaitis/mxmh-survey-results