Predicting a Song Becoming a Hit with Probabilistic Models

Maria Adshead

maria.adshead1@gmail.com

Estimating the probability of a song becoming a hit is both a scientifically intriguing issue and a relevant research topic for the music industry. The main aim of this paper is to find the relevant predictors of the probability of a song becoming a hit and to construct a probabilistic model based on them. However, it is also of interest to interpret the influence of various factors on this probability, so the direction and the character of the relationship between different variables and the dependent variable are examined as well. The sources reviewed for this work include some papers in the field of music. Some were studied to find out what parameters could potentially determine the fact of individuals finding the music pleasant and the song becoming popular eventually. For instance, the relationship between musical tempo and affect in some past works has been considered. The evidence is rather mixed; however, the pace of tempo and level of affect appear to form a U-shaped curve (Anand and Holbrook, 1985). To elaborate, it was discovered that as tempo increases, so does the level affect up to some point. Afterwards, as tempo decreases, the affect starts to decrease. The relationship between music sentiment and patterns of behaviour has been researched as well. The findings suggest that music conveying positive sentiment results in more optimistic mood in individuals and vice versa (Alpert and Alpert, 1986). This result is crucial for the research conducted in this paper as it directly suggests that music sentiment is a relevant determinant of preferences of individuals. Furthermore, it is anticipated that the decade in which a song was released is a relevant variable. Digital music has been widely introduced in the nineties, which has made music consumption more accessible (Fly, 2016). As a result, since it is easier to spread music, the probability of a song becoming a hit in more recent decades may have been affected. Another proposition is that genre is a relevant predictor of popularity because different genres have been popular throughout decades (Crawford and Anderson, 2005). However, it should be noted that the results in that paper only account for years 1982-2002, so our data may not replicate these results fully.

Based on previous findings, it is hypothesized that sentiment of the song, the decade in which it was released, and audio characteristics associated with the overall mood and energy of the music are going to be significant regressors. This paper focuses on various audio characteristics as regressors, including tempo and sentiment that were mentioned above. Other characteristics that are conceptually related to sentiment and tempo are also included, which will be described further. Variables are also introduced to account for the decade in which a song was released. The dataset used for the estimation of the model was extracted using Spotify Web API and obtained through Kaggle; it contains features on 'hit' and 'non-hit' songs from six decades: from 1960s to 2010s. Songs are classified into the two categories based on some criteria of the author of the dataset. The features are name of the song, name of the artist, danceability, energy, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, duration, time signature, chorus starting timestamp and number of sections. To elaborate, danceability is a variable in the range from 0 to 1 that reflects the probability of a song being appropriate for dancing. Energy variable indicates how likely the music is to be perceived as loud or intense. Mode is a binary variable, 1 standing for major and 0 for minor. Spechiness reflects the percentage of spoken (not sung) words present in the composition. Acousticness variable shows the probability of the song being acoustic. Instrumentalness measures the probability that the song consists entirely of instrument sound (has no vocals). Liveness indicates the probability that this song was recorded live. All the probabilities above are represented as shares from 0 to 1 in the dataset. Valence is an overall measure of the sentiment of a song, ranging from 0 to 1, where higher values correspond to more positive songs. Tempo is the tempo of a song in beats per minute. Duration is represented in milliseconds. Time signature is the number of beats in a bar. Chorus hit is the timestamp in milliseconds at which the chorus starts. Sections are different divisions of a song, such as verses and the choruses, represented as the overall number of them. Finally, the dependent binary variable is referred to as the target. It is equal to 1 if a song is a hit and 0 otherwise. The overall size of the sample is 41106.

It should be acknowledged that the data used in the research has some drawbacks that should be addressed in further research on the topic in order to improve the quality of the model that was estimated. As discussed above, different music genres were popular in different decades. For example, classic rock was 9 p.p. more popular in 1992 than in 1982 (Crawford and Anderson, 2005). Therefore, the effect of genres should be estimated and accounted for if this variable is indeed proven to be significant. However, this data was not available. Furthermore, time dynamics are captured in the models presented in this paper only through dummy variables indicating a specific decade. More precise and meaningful results may have been acquired if data on smaller periods of time was available.

All estimations and calculations are accomplished using R programming language (Appendix 12). This paper utilizes the approaches of linear probability model (LPM), logit and probit probabilistic models. LPM is fit through OLS method, while the latter two are fit through maximum likelihood estimation. We will choose the best specification in the framework of each method by conducting necessary tests. Then we will compare models obtained using each of three methods using prediction error rate.

The essence of maximum likehood estimation method is as follows. The values of the sample are plugged into the corresponding distribution function, the values are multiplied. This product of probabilities is then maximized, and parameters are found. The difference between the two models consists in the fact that logit model uses logistic distribution for the likelihood function, while in the probit model the cumulative distribution function of the standard normal distribution is used.

In the general form a likelihood function could be written as:

$$\text{L}=\prod_i p(target = target_i|X_i, \alpha, \beta)=\prod_{i:target=1} F(\alpha + \beta * X_i)*\prod_{i:target=0}(1 - F(\alpha + \beta * X_i)) \rightarrow \max_{\alpha,\beta}$$

where X_i is a vector of regressors in a given specification, α is the intercept and β is a vector of coefficients of the explanatory variables. To simplify the calculations often a log-likelihood function is maximized instead. This can be performed due to the fact that logarithmic transformation is a monotonic transformation, hence the optima of the original function are not affected. The new optimization problem will be as follows:

$$\text{logL}=\sum_i log(p(target = target_i|X_i, \alpha, \beta))=\sum_{i:target=1} log(F(\alpha + \beta * X_i))+\sum_{i:target=0} log((1 - F(\alpha + \beta * X_i))) \rightarrow \max_{\alpha,\beta}$$

For logit model $F(\alpha + \beta * X_i) = F(Z) = \frac{1}{1+e^{-(\alpha+\beta*X_i)}}$, for probit model $F(\alpha + \beta * X_i) = F(Z)$ - cumulative probability function for the standard normal distribution.

We start by estimating a linear probability model, initially including all listed variables and decade dummies in the specification. Dummies for 70s, 80s, 90s, 00s and 10s are denoted as "sev", "eig", "nin", "zer" and "ten" respectively. It should be noted that the sixties are chosen as the reference group. The results of this estimation can be found in Appendix 1.

Theoretical equation:

$target_i = \alpha + \beta_1 * danceability_i + \beta_2 * energy_i + \beta_3 * loudness_i + \beta_4 * mode_i + \beta_5 * speechiness_i + \beta_6 * acousticness_i + \beta_7 * instrumentalness_i + \beta_8 * liveness_i + \beta_9 * valence_i + \beta_{10} * tempo_i + \beta_{11} * durationms_i + \beta_{12} * timesignature_i + \beta_{13} * chorus_hit_i + \beta_{14} * sections_i + \beta_{15} * sev_i + \beta_{16} * eig_i + \beta_{17} * nin_i + \beta_{18} * zer_i + \beta_{19} * ten_i + u_i$

Estimated model:

$target_i = \underset{(0.029)}{0}.623 + \underset{(0.016)}{0}.653 * danceability_i - \underset{(0.017)}{0}.246 * energy_i + \underset{(0.0007)}{0}.013 * loudness_i + \underset{(0.005)}{0}.061 * mode_i - \underset{(0.025)}{0}.537 * speechiness_i - \underset{(0.010)}{0}.284 * acousticness_i - \underset{(0.007)}{0}.454 * instrumentalness_i - \underset{(0.013)}{0}.083 * liveness_i - \underset{(0.011)}{0}.026 * valence_i + \underset{(0.00008...)}{0}.0002 * tempo_i - \underset{(0.000...)}{0}.000... * durationms_i + \underset{(0.005)}{0}.022 * timesignature_i - \underset{(0.0001...)}{0}.0004 * chorus_hit_i - \underset{(0.001)}{0}.0006 * sections_i - \underset{(0.007)}{0}.050 * sev_i - \underset{(0.008)}{0}.116 * eig_i - \underset{(0.008)}{0}.112 * nin_i - \underset{(0.008)}{0}.127 * zer_i - \underset{(0.008)}{0}.140 * ten_i$ **(1)**

As can be seen from the output, the variables valence, duration, and sections are not significant at 1% significance level, since the p-values are higher than 0.01 (being equal to 0.023, 0.094 and 0.529 respectively). Let us look deeper into the reasons why the latter could be so. Intuitively, one would expect these two variables to be highly correlated, e.g., a song with more sections is likely to have a longer duration. In other words, the issue of multicollinearity may take place. A correlation matrix was obtained to scrutinize the data for the presence of this problem (Appendix 2, can be found in the Excel file attached). As was anticipated, the two variables – duration and sections – are indeed highly correlated with a correlation coefficient of 0.889. Valence is not highly correlated with either of duration or sections but is correlated with such variables as danceability and energy. This observation is also intuitively anticipated, because a more positive song is more likely to be appropriate for dancing and have high energy. One can note that for some other variables, for instance, loudness and energy, the correlation is also very high (0.773), however it has not influenced the significance of the variables.

Having taken the detected issue of multicollinearity into account, one could attempt to estimate the model first excluding each of the three variables and observe the results. Before this is done, one is ought to test the significance of the variables as a group to see whether any of them should be included at all using an F-test.

$H_0 : \beta_9 = \beta_{11} = \beta_{14} = 0$

H_1: at least one of the coefficients above is not equal to 0

$F_{st} = \frac{(R^2 - R^2 restricted)/m}{(1-R^2)/(n-p)} = \frac{(0.2688-0.2686)/3}{(1-R^2)/(1-0.2688)/(41106-20)} = 3.744$, where m - number of restrictions, n - sample size, p - number of parameters.

Under the null hypothesis this statistic follows the distribution F(3,41086), hence the critical value will be $F_{1\%}(3, 41086) = 3.79$. F_{st} is below the critical value, hence we fail to reject the null hypothesis at 1%. The restrictions are valid, e.g., these variables are not significant as a group, hence the specification in Appendix 3 is chosen as a result of implementing the linear probability model approach:

$target_i = \underset{(0.029)}{0}.625 + \underset{(0.013)}{0}.634 * danceability_i - \underset{(0.016)}{0}.259 * energy_i + \underset{(0.0007)}{0}.014 * loudness_i + \underset{(0.005)}{0}.061 * mode_i - \underset{(0.026)}{0}.531 * speechiness_i - \underset{(0.010)}{0}.286 * acousticness_i - \underset{(0.008)}{0}.453 * instrumentalness_i - \underset{(0.013)}{0}.083 * liveness_i + \underset{(0.00008...)}{0}.0002 * tempo_i + \underset{(0.005)}{0}.021 * timesignature_i - \underset{(0.0001...)}{0}.0004 * chorus_hit_i - \underset{(0.007)}{0}.051 * sev_i - \underset{(0.007)}{0}.116 * eig_i - \underset{(0.008)}{0}.111 * nin_i - \underset{(0.008)}{0}.126 * zer_i - \underset{(0.008)}{0}.137 * ten_i$ **(2)**

Having mentioned that it is relevant to study the character of influence of the variables on the probability of a song becoming a hit, it would be appropriate to elaborate on the marginal effects of the regressors. It the case of a linear probability model the marginal effect is the same for any particular item. Interpreting (1) to account for all variables present in the dataset, we get the following results:

- if danceability increases by 1 percentage point, e.g., by 0.01, the probability of becoming a hit increases on average by 0.623 p.p., or by 0.00623
- if energy increases by 1 percentage point, e.g., by 0.01, the probability of becoming a hit decreases on average by 0.246 p.p., or by 0.00246
- if loudness increases by 1 percentage point, e.g., by 0.01, the probability of becoming a hit increases on average by 0.013 p.p., or by 0.00013
- if mode changes from minor to major, the probability of becoming a hit increases on average by 0.061 or by 6.1 p.p.
- if speechiness increases by 1 percentage point, e.g., by 0.01, the probability of becoming a hit decreases on average by 0.537 p.p., or by 0.00537
- if acousticness increases by 1 percentage point, e.g., by 0.01, the probability of becoming a hit decreases on average by 0.284 p.p., or by 0.00284
- if instrumentalness increases by 1 percentage point, e.g., by 0.01, the probability of becoming a hit decreases on average by 0.454 p.p., or by 0.00454
- if liveness increases by 1 percentage point, e.g., by 0.01, the probability of becoming a hit decreases on average by 0.013 p.p., or by 0.00013

- if valence increases by 1 percentage point, e.g., by 0.01, the probability of becoming a hit decreases on average by 0.026 p.p., or by 0.00026
- if tempo increases by 1 beat per minute, the probability of becoming a hit increases on average by 0.0002 p.p., or by 0.000002
- if duration increases by 1 second, the probability of becoming a hit decreases on average by 0.065 p.p., or by 0.00065
- if time signature increases by 1 beat in a bar, the probability of becoming a hit increases on average by 0.022 p.p., or by 0.00022
- if chorus hit timestamp increases by 1 second, the probability of becoming a hit decreases on average by 0.380 p.p., or by 0.0038
- if the number of sections increases by 1, the probability of becoming a hit increases on average by 0.0006 p.p., or by 0.000006
- if the song was released in the 70s, the probability of becoming a hit decreases on average by 0.050, or by 5 p.p.
- if the song was released in the 80s, the probability of becoming a hit decreases on average by 0.116, or by 11.6 p.p.
- if the song was released in the 90s, the probability of becoming a hit decreases on average by 0.112, or by 11.2 p.p.
- if the song was released in the 00s, the probability of becoming a hit decreases on average by 0.127, or by 12.7 p.p.
- if the song was released in the 10s, the probability of becoming a hit decreases on average by 0.140, or by 14 p.p.

First of all, the results loosely suggest that compared to music released in the 60s, the probability of becoming a hit decreases with every decade. This may have to do with the perception of music: the music was less accessible in the past and less music was released, hence more songs were treated as hits, and the trend was down as music gradually became more accessible and disposable through time. Moreover, different factors may have contributed to becoming a hit in different decades, for which relevant variables can be missing. Another possibility is that the data is biased. It could be that because of the age gap only the most popular songs that are more likely to become hits 'survived' to the present date and were included in the dataset. However, this question is not the focus of this research. The fact that valence has a negative impact on probability contradicts previous findings: according to previous works, positive music is more likely to be pleasing to individuals, while our results suggest the opposite. Nevertheless, this variable was shown to be insignificant. The general finding is that songs that are more energetic, contain more spoken words than singing, are acoustic, consist mostly of instrument sounds, are recorded live, last longer and have a late chorus have a lower probability of becoming a hit. Conversely, songs that are more appropriate for dancing, have a major mode, have a higher tempo, have a higher time signature and more sections are more likely to become a hit.

We continue by estimating a logit model, including all listed variables and decade dummies in the specification as in the LPM case. The results of this estimation can be found in Appendix 4.

Theoretical equation:

$Z = \alpha + \beta_1 * danceability_i + \beta_2 * energy_i + \beta_3 * loudness_i + \beta_4 * mode_i + \beta_5 * speechiness_i + \beta_6 * acousticness_i + \beta_7 * instrumentalness_i + \beta_8 * liveness_i + \beta_9 * valence_i + \beta_{10} * tempo_i + \beta_{11} * durationms_i + \beta_{12} * timesignature_i + \beta_{13} * chorus_hit_i + \beta_{14} * sections_i + \beta_{15} * sev_i + \beta_{16} * eig_i + \beta_{17} * nin_i + \beta_{18} * zer_i + \beta_{19} * ten_i + u_i$

Estimated model:

$Z = \underset{(0.180)}{1}.106 + \underset{(0.099)}{4}.056 * danceability_i - \underset{(0.103)}{1}.826 * energy_i + \underset{(0.005)}{0}.133 * loudness_i + \underset{(0.026)}{0}.364 * mode_i - \underset{(0.156)}{2}.864 * speechiness_i - \underset{(0.056)}{1}.811 * acousticness_i - \underset{(0.069)}{3}.427 * instrumentalness_i - \underset{(0.070)}{0}.300 * liveness_i - \underset{(0.065)}{0}.273 * valence_i + \underset{(0.0004...)}{0}.003 * tempo_i + \underset{(0.000...)}{0}.000... * durationms_i + \underset{(0.033)}{0}.173 * timesignature_i - \underset{(0.000...)}{0}.002 * chorus_hit_i - \underset{(0.006)}{0}.016 * sections_i - \underset{(0.039)}{0}.326 * sev_i - \underset{(0.047)}{0}.777 * eig_i - \underset{(0.047)}{0}.842 * nin_i - \underset{(0.050)}{1}.099 * zer_i - \underset{(0.050)}{1}.224 * ten_i$ **(3)**

Before proceeding with further modifications, it should be tested whether the model is significant at all, e.g., test the restriction of all coefficients except the constant being equal to zero against the alternative that at least one coefficient by the variables is not equal to zero. To do so, we will conduct the likelihood ratio test. The test statistic, the rejection rule and the critical values will be as follows.

$LR_{st0}^{logit} = 2(LR_{ur} - LR_r) = 2(-21288.59 + 28492.51) = 14407.84$

Under the null hypothesis the statistic follows chi-squared distribution with the number of degrees of freedom equal to the number of restrictions imposed by the null hypothesis, e.g., nineteen. The rejection rule states that the null hypothesis is rejected if the calculated statistic is above the critical value. The critical value, $\chi^2_{1\%}(19) = 36.2$. Hence, we reject the null hypothesis that the restrictions are valid at 1% significance level and conclude that the model is indeed significant.

Assuming 1% significance level, it can be clearly seen from the estimation output that variables duration and sections are insignificant, with p-values being 0.300 and 0.014 respectively. As discussed above, this has occurred due to multicollinearity issues. One can proceed by estimating the model using either of the two variables in question and compare the performance of the models. The results for the two cases (excluding duration and excluding sections) can be viewed in Appendix 5 and Appendix 6 respectively. Excluding duration we have the theoretical equation and the estimated model as follows:

$$Z = \alpha + \beta_1 * danceability_i + \beta_2 * energy_i + \beta_3 * loudness_i + \beta_4 * mode_i + \beta_5 * speechiness_i + \beta_6 * acousticness_i + \beta_7 * instrumentalness_i + \beta_8 * liveness_i + \beta_9 * valence_i + \beta_{10} * tempo_i + \beta_{11} * timesignature_i + \beta_{12} * chorus_hit_i + \beta_{13} * sections_i + \beta_{14} * sev_i + \beta_{15} * eig_i + \beta_{16} * nin_i + \beta_{17} * zer_i + \beta_{18} * ten_i + u_i$$

Estimated model:

$$Z = \underset{(0.179)}{1.099} + \underset{(0.099)}{4.056} * danceability_i - \underset{(0.102)}{1.818} * energy_i + \underset{(0.005)}{0.132} * loudness_i + \underset{(0.026)}{0.363} * mode_i - \underset{(0.156)}{2.865} * speechiness_i - \underset{(0.056)}{1.813} * acousticness_i - \underset{(0.068)}{3.422} * instrumentalness_i - \underset{(0.070)}{0.299} * liveness_i - \underset{(0.065)}{0.280} * valence_i + \underset{(0.0004...)}{0.003} * tempo_i + \underset{(0.033)}{0.173} * timesignature_i - \underset{(0.001)}{0.002} * chorus_hit_i - \underset{(0.003)}{0.010} * sections_i - \underset{(0.039)}{0.322} * sev_i - \underset{(0.043)}{0.772} * eig_i - \underset{(0.047)}{0.837} * nin_i - \underset{(0.049)}{1.094} * zer_i - \underset{(0.050)}{1.220} * ten_i \quad \textbf{(4)}$$

For the case when sections variable is excluded we have:

$$Z = \alpha + \beta_1 * danceability_i + \beta_2 * energy_i + \beta_3 * loudness_i + \beta_4 * mode_i + \beta_5 * speechiness_i + \beta_6 * acousticness_i + \beta_7 * instrumentalness_i + \beta_8 * liveness_i + \beta_9 * valence_i + \beta_{10} * tempo_i + \beta_{11} * durationms_i + \beta_{12} * timesignature_i + \beta_{13} * chorus_hit_i + \beta_{14} * sev_i + \beta_{15} * eig_i + \beta_{16} * nin_i + \beta_{17} * zer_i + \beta_{18} * ten_i + u_i$$

Estimated model:

$$Z = \underset{(0.178)}{1.057} + \underset{(0.099)}{4.053} * danceability_i - \underset{(0.102)}{1.809} * energy_i + \underset{(0.005)}{0.132} * loudness_i + \underset{(0.026)}{0.364} * mode_i - \underset{(0.156)}{2.863} * speechiness_i - \underset{(0.056)}{1.813} * acousticness_i - \underset{(0.069)}{3.419} * instrumentalness_i - \underset{(0.070)}{0.299} * liveness_i - \underset{(0.065)}{0.281} * valence_i + \underset{(0.0004...)}{0.003} * tempo_i + \underset{(0.000...)}{0.000...} * durationms_i + \underset{(0.033)}{0.173} * timesignature_i - \underset{(0.000...)}{0.002} * chorus_hit_i - \underset{(0.039)}{0.324} * sev_i - \underset{(0.043)}{0.776} * eig_i - \underset{(0.047)}{0.838} * nin_i - \underset{(0.050)}{1.094} * zer_i - \underset{(0.050)}{1.220} * ten_i \quad \textbf{(5)}$$

By merely looking at the estimation outputs we can see that in the first case excluding the variable duration has led to all variables becoming significant at 1% significance level. At the same time in the second case the variable duration has remained insignificant at 1% significance level. However, tests need to be carried out to claim more confidently which specification is a better fit. To implement the latter, we will test the restrictions of coefficients of duration and sections being equal to zero separately and test the significance of these variables as a group. The test requires the estimation of the model without both duration and sections (Appendix 7).

Theoretical equation:

$$Z = \alpha + \beta_1 * danceability_i + \beta_2 * energy_i + \beta_3 * loudness_i + \beta_4 * mode_i + \beta_5 * speechiness_i + \beta_6 * acousticness_i + \beta_7 * instrumentalness_i + \beta_8 * liveness_i + \beta_9 * valence_i + \beta_{10} * tempo_i + \beta_{11} * durationms_i + \beta_{12} * timesignature_i + \beta_{13} * chorus_hit_i + \beta_{14} * sev_i + \beta_{15} * eig_i + \beta_{16} * nin_i + \beta_{17} * zer_i + \beta_{18} * ten_i + u_i$$

Estimated model:

$$Z = \underset{(0.177)}{1.006} + \underset{(0.099)}{4.047} * danceability_i - \underset{(0.102)}{1.818} * energy_i + \underset{(0.005)}{0.133} * loudness_i + \underset{(0.026)}{0.367} * mode_i - \underset{(0.156)}{2.857} * speechiness_i - \underset{(0.056)}{1.807} * acousticness_i - \underset{(0.068)}{3.429} * instrumentalness_i - \underset{(0.070)}{0.302} * liveness_i - \underset{(0.064)}{0.260} * valence_i + \underset{(0.0004...)}{0.003} * tempo_i + \underset{(0.033)}{0.172} * timesignature_i - \underset{(0.000...)}{0.002} * chorus_hit_i - \underset{(0.039)}{0.339} * sev_i - \underset{(0.043)}{0.795} * eig_i - \underset{(0.046)}{0.856} * nin_i - \underset{(0.049)}{1.112} * zer_i - \underset{(0.050)}{1.231} * ten_i \quad \textbf{(6)}$$

To test the significance of duration and sections as a group we will conduct the likelihood ratio test. The formula for the test statistic, the null and the alternative hypotheses and the rejection rule are similar to the ones in the previous test. We test the null hypothesis that the restrictions are valid (coefficients of duration and sections are equal to 0 and (6) should be used) against the alternative that the restrictions are invalid (at least one of coefficients of either duration or sections is not 0 and (3) should be used).

$$LR_{st1}^{logit} = 2(-21288.59 + 21294.27) = 11.36$$

$LR_{st} = 11.36$ is above $\chi^2_{1\%}(2) = 9.21$, therefore the null hypothesis is rejected at 1% significance level. The two variables are at least significant as a group, which means that out of the two models that were compared the unrestricted one should be used. It was discussed above that in (3) both duration and sections were shown to be insignificant as a result of a single coefficient significance test. However, after excluding the variables separately in the specification where duration was excluded sections became a significant variable, hence this may be the specification that is more correct. One could additionally test the restrictions of separate coefficients for duration and sections being equal to zero by conducting the likelihood ratio to confirm or deny this proposition. The critical value for both cases will be $\chi^2_{1\%}(1) = 6.64$. The likelihood ratio statistics for (4) and (5) respectively are:

$$LR_{st2}^{logit} = 2(-21288.59 + 21289.12) = 1.06$$

$$LR_{st3}^{logit} = 2(-21288.59 + 21291.55) = 5.92$$

The null hypothesis is however not rejected at 1% in both cases. At 1% significance level we conclude that in both cases the restrictions are valid. In fact, this result was rather anticipated because there most likely exists a multicollinearity issue between these two variables. Therefore, it is not contradictory that we conclude in each case that the coefficient of the other variable should be equal to zero. It merely means that one should leave only one of them in the specification. This conclusion can be even made based on the statistics calculated for the test above. The comparison of (3) and (4) clearly results in a lower statistic that the comparison of (3) and (5), therefore in the case of (4) the restriction is valid with a higher probability. Hence, out of all specifications one should select (4).

Here we will also demonsrate the marginal effects for the variables. It is a known fact that the marginal effects are not constant in logit and probit, hence it is impossible to state some facts that would apply to all subjects. Another issue is that based on the characteristics in the dataset the subjects could be categorized into many different groups, and it would be hardly possible to distinguish all of them. Therefore, for simplicity we are going to select a limited number of songs that differ from each other and calculate the marginal effects for them, which would provide us with general ideas about the direction of the influence of the variables. From each decade one hit and one non-hit song were selected. The songs and their characteristics are presented in Table 1, while marginal effects for the logit model (with all variables) are shown in Table 2. All tables can be found in the additionally attached Excel file.

As can be seen, the direction of influence (the signs) are the same as in the case of LPM except for duration, which was however proven to be insignificant. There is no trend visible from a table for danceability or energy, but marginal effects appear to be noticeably smaller for rock songs such as "Pagan Fears" or "Sometimes". For the same songs other marginal effects are smaller as well. The small marginal effects of decade dummies for these songs imply that this type of music probably has a small chance of being popular in general. The marginal effects of tempo and chorus hit are overall the smallest, which indicates that these variables have a small influence. Overall, the largest effects in absolute value can be observes from such variables as danceability, speechiness and instrumentalness.

We proceed with estimating a probit model similarly by including all available variables in the specification. The sixties are likewise used as a reference group. The details of the output can be found in Appendix 8.

Theoretical equation:

$$Z = \alpha + \beta_1 * danceability_i + \beta_2 * energy_i + \beta_3 * loudness_i + \beta_4 * mode_i + \beta_5 * speechiness_i + \beta_6 * acousticness_i + \beta_7 * instrumentalness_i + \beta_8 * liveness_i + \beta_9 * valence_i + \beta_{10} * tempo_i + \beta_{11} * durationms_i + \beta_{12} * timesignature_i + \beta_{13} * chorus_h it_i + \beta_{14} * sections_i + \beta_{15} * sev_i + \beta_{16} * eig_i + \beta_{17} * nin_i + \beta_{18} * zer_i + \beta_{19} * ten_i + u_i$$

Esimated model:

$$Z = \underset{(0.106)}{0.623} + \underset{(0.058)}{2.387} * danceability_i - \underset{(0.060)}{1.041} * energy_i + \underset{(0.003)}{0.077} * loudness_i + \underset{(0.016)}{0.215} * mode_i - \underset{(0.092)}{1.794} * speechiness_i - \underset{(0.033)}{1.049} * acousticness_i - \underset{(0.039)}{1.938} * instrumentalness_i - \underset{(0.042)}{0.184} * liveness_i - \underset{(0.039)}{0.152} * valence_i + \underset{(0.0002...)}{0.002} * tempo_i + \underset{(0.000...)}{0.000...} * durationms_i + \underset{(0.019)}{0.102} * timesignature_i - \underset{(0.000...)}{0.001} * chorus_h it_i - \underset{(0.004)}{0.009} * sections_i - \underset{(0.023)}{0.184} * sev_i - \underset{(0.026)}{0.456} * eig_i - \underset{(0.028)}{0.487} * nin_i - \underset{(0.029)}{0.637} * zer_i - \underset{(0.029)}{0.713} * ten_i \quad \textbf{(7)}$$

As in the case of the logit model, let us first determine whether the model is significant at all, e.g., test the validity of restrictions that all variables except the constant are 0.

$$LR^{probit}_{st0} = 2(-21320.77 + 28492.51) = 14343.48$$

This statistic is higher than the critical value $\chi^2_{1\%}(19) = 36.2$, hence the null hypothesis which stated that the restrictions are valid is rejected at 1%, meaning the model is significant. Therefore, we proceed by observing the implications of the output. It can be easily seen that similarly to the case of the logit model, duration and sections are not statistically significant at 1% significance level (p-values are 0.417 and 0.020 respectively. Similarly as we did in the case of logit model, we can estimate the model excluding each of these two variables (statistics 1 and 2) and then both (statistic 3) and compare the results with the original model. The outputs for corresponding regressions are shown in Appendices 9,10,11. Below I present the estimated models. The theoretical equations for the Z value will be the same as in the corresponding cases in the logit model analysis.

$$Z = \underset{(0.106)}{0.620} + \underset{(0.058)}{2.387} * danceability_i - \underset{(0.058)}{1.037} * energy_i + \underset{(0.003)}{0.077} * loudness_i + \underset{(0.016)}{0.215} * mode_i - \underset{(0.092)}{1.704} * speechiness_i - \underset{(0.033)}{1.050} * acousticness_i - \underset{(0.036)}{1.936} * instrumentalness_i - \underset{(0.042)}{0.183} * liveness_i - \underset{(0.038)}{0.155} * valence_i + \underset{(0.0003...)}{0.002} * tempo_i + \underset{(0.019)}{0.102} * timesignature_i - \underset{(0.000...)}{0.001} * chorus_h it_i - \underset{(0.002)}{0.006} * sections_i - \underset{(0.023)}{0.182} * sev_i - \underset{(0.026)}{0.453} * eig_i - \underset{(0.027)}{0.485} * nin_i - \underset{(0.029)}{0.634} * zer_i - \underset{(0.029)}{0.711} * ten_i \quad \textbf{(8)}$$

$$Z = \underset{(0.106)}{0.596} + \underset{(0.058)}{2.385} * danceability_i - \underset{(0.060)}{1.031} * energy_i + \underset{(0.003)}{0.077} * loudness_i + \underset{(0.016)}{0.215} * mode_i - \underset{(0.092)}{1.703} * speechiness_i - \underset{(0.033)}{1.050} * acousticness_i - \underset{(0.036)}{1.934} * instrumentalness_i - \underset{(0.042)}{0.183} * liveness_i - \underset{(0.039)}{0.156} * valence_i + \underset{(0.0003...)}{0.003} * tempo_i + \underset{(0.000...)}{0.000...} * durationms_i + \underset{(0.019)}{0.102} * timesignature_i - \underset{(0.000...)}{0.001} * chorus_h it_i - \underset{(0.023)}{0.183} * sev_i - \underset{(0.026)}{0.455} * eig_i - \underset{(0.028)}{0.485} * nin_i - \underset{(0.029)}{0.634} * zer_i - \underset{(0.029)}{0.711} * ten_i \quad \textbf{(9)}$$

$$Z = \underset{(0.104)}{0.562} + \underset{(0.057)}{2.381} * danceability_i - \underset{(0.060)}{1.037} * energy_i + \underset{(0.003)}{0.078} * loudness_i + \underset{(0.016)}{0.217} * mode_i - \underset{(0.092)}{1.700} * speechiness_i - \underset{(0.033)}{1.047} * acousticness_i - \underset{(0.036)}{1.939} * instrumentalness_i - \underset{(0.042)}{0.185} * liveness_i - \underset{(0.038)}{0.143} * valence_i + \underset{(0.0003...)}{0.002} * tempo_i + \underset{(0.019)}{0.102} * timesignature_i - \underset{(0.000...)}{0.001} * chorus_h it_i - \underset{(0.023)}{0.193} * sev_i - \underset{(0.025)}{0.467} * eig_i - \underset{(0.027)}{0.496} * nin_i - \underset{(0.029)}{0.646} * zer_i -$$

$\underset{(0.029)}{0}.719 * ten_i$ **(10)**

$LR^{probit}_{st1} = 2(-21320.77 + 21321.09) = 0.64$

$LR^{probit}_{st2} = 2(-21320.77 + 21323.37) = 5.2$

$LR^{probit}_{st3} = 2(-21320.77 + 21326.59) = 11.64$

The critical value for the first two cases that test the significance of duration and sections respectively is $\chi^2_{1\%}(1) = 6.64$, therefore the null hypothesis is not rejected at 1% in both of the first two cases, the restrictions on coefficients are concluded to be valid. For the third case where the significance of these values as a group is tested the critical value is $\chi^2_{1\%}(2) = 9.21$, hence here the null hypothesis is rejected at the same significance level, e.g., here the restrictions are not valid and the two variables are significant as a group. The results outlined above are very similar to the ones that were obtained in the analysis of the logit model: variables should not be included together. Likewise in the case where duration was excluded the restriction is valid with a lower statistic, so one would choose specification (8).

The marginal effects for the probit model including all of the variables are presented in Table 3 (can be found in the attached Excel file). Results are similar to the ones obtained under the logit model. The directions of influence are the same as in the LPM and logit. It can be likewise observes that for the songs "Sometimes" and "Pagan Fears" all of the marginal effects are smaller than for other ones. These songs are both non-hits, but there are other non-hit songs in the sample; the difference could be explaines by the fact that both these songs were released by non-popular artists as well. Here danceability, instrumentalness and speechiness also demonstrate the biggest marginal effects by absolute value, but they are smaller than in the case of logit.

The final choice of the model falls to the comparison of the LPM, logit and probit models that were chosen as a result of previous analysis. We start by estimating the error rate of the models in prediction, which is calculated as the number of cases where $target_i \neq prediction_i$ divided by the total sample size. Probabilities from the three models were obtained and the prediction is categorized as 1 is the value is above 0.5. The threshold is a parameter of choice, but here is was decided to assume that probability above a half could be considered a hit. The figures for this error for LPM, logit and probit respectively were 0.266, 0.279 and 0.328. In other words, linear probability model has performed the best on the data. The final choice for the probabilistic model to predict the probability of a song becoming a hit is the LPM specification denoted by (2).

To explain this result we could additionally test the data for heteroskedasticity. This is relevant because if the data is homogenous, then the marginal effects will be the same on average and LPM indeed performs better. If not, then a more complex model is needed, and a lower error rate was obtained merely due to a large sample size or because neither probit nor logit were the right choice for this type of data. One can start by exploring the residual plots for the LPM, logit and probit (Figure 1, 2 and 3 respectively).

Figure 1

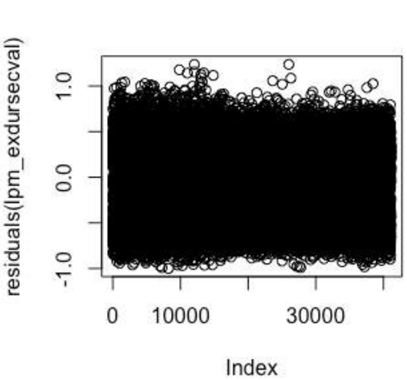


Figure 2

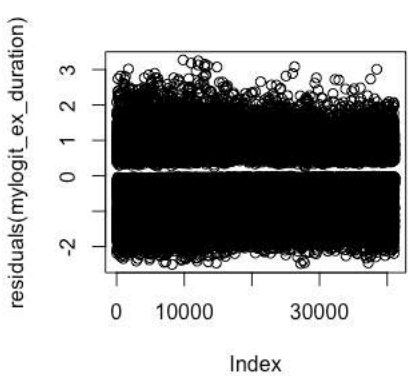
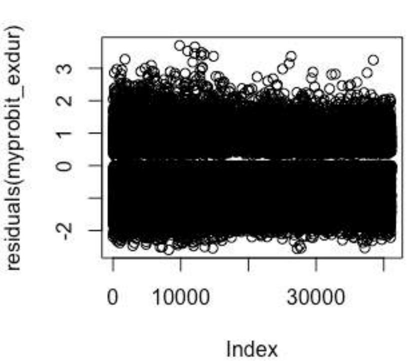


Figure 3



It is not clearly visible from the graphs whether the errors are heteroskedastic, because for some parts of the sample the plot is rather homogenous, while for some it is not. Additional tests need to be carried out in order to claim confidently whether heteroskedasticity is present. White test using the auxiliary equation without interactions was implemented, obtaining R-squared of 6.55e-31. We test the null hypothesis of homoskedasticity against the alternative of heterskedasticity of any type by calculating the following statistic:

$$\chi^2_{st} = n * R^2_{aux} = 41106 * 6.55e - 31 = 2.69 * 10e - 26$$

Under the null hypothesis the statistic follows the distribution $\chi^2(26)$, where 26 is the number of regressors in the White auxiliary equation. The critical value is $\chi^2_{1\%}$=45.6, hence we fail to reject the null hypothesis at 1% significance level. There is no heteroskedasticity, hence the marginal effects are the same on average. Therefore, more sophisticated models are not required and LPM performs better.

In conclusion, the data used clearly has disadvantages because is misses some key factors that explain the probability of becoming a hit depending on a particular decade that may differ, the specification that was constructed could be improved by including this data. There is large room for improvement in terms of the error rate, however the model can be used in some cases since its accuracy is 73.4%. It means that many significant variables, such as danceability, mode and others have been captured. However, the main concern which consists in omitting such features as genre remains and should be mitigated in future works.

Sources

1. Alpert, J., Alpert, M., Background Music as an Influence in Consumer Mood and Advertising Responses. Advances in Consumer Research, Vol. 16, 1988, 485-491

2. Alpert, J., Alpert, M., The Effects of Music in Advertising on Mood and Purchase Intentions (unpublished working paper 85/86). 1986

3. Anand, P., Holbrook, M., Chasing the Wundt Curve: An Adventure in Consumer Aesthetics. Advances in Consumer Research, Vol. 12, 1985, 655-657

4. Author's own calculations

5. Fly, B., How Does Music Consumption Impact the Music Industry and Benefit Artists? Accounting Undergraduate Honours Theses, 2016

6. Kaggle, The Spotify Hit Predictor Dataset (1960-2019), https://www.kaggle.com/theoverman/the-spotify-hit-predictor-dataset

7. Mizell, L., Crawford, B., Anderson, C., Music Preferences in the U.S.: 1982-2002. National Endowment for the Arts, 2005

Annexes

Appendix 1

```
                   Estimate  Std. Error   t value  Pr(>|t|)
(Intercept)       6.2343e-01  2.8637e-02   21.7700 < 2.2e-16 ***
danceability      6.5284e-01  1.6092e-02   40.5689 < 2.2e-16 ***
energy           -2.4563e-01  1.7401e-02  -14.1160 < 2.2e-16 ***
loudness          1.3362e-02  6.6279e-04   20.1601 < 2.2e-16 ***
mode              6.1089e-02  4.5717e-03   13.3623 < 2.2e-16 ***
speechiness      -5.3662e-01  2.4742e-02  -21.6888 < 2.2e-16 ***
acousticness     -2.8426e-01  9.6660e-03  -29.4081 < 2.2e-16 ***
instrumentalness -4.5376e-01  6.6290e-03  -68.4508 < 2.2e-16 ***
liveness         -8.3100e-02  1.2823e-02   -6.4808 9.229e-11 ***
valence          -2.5680e-02  1.1294e-02   -2.2737 0.0229876 *
tempo             2.3113e-04  7.6705e-05    3.0132 0.0025870 **
duration_ms      -6.5004e-08  3.8796e-08   -1.6755 0.0938356 .
time_signature    2.1764e-02  4.8486e-03    4.4887 7.187e-06 ***
chorus_hit       -3.8037e-04  1.1467e-04   -3.3170 0.0009108 ***
sections          5.9391e-04  9.4246e-04    0.6302 0.5285839
sev              -4.9578e-02  7.1883e-03   -6.8970 5.387e-12 ***
eig              -1.1562e-01  7.6091e-03  -15.1948 < 2.2e-16 ***
nin              -1.1169e-01  7.9076e-03  -14.1248 < 2.2e-16 ***
zer              -1.2707e-01  8.0564e-03  -15.7731 < 2.2e-16 ***
ten              -1.4031e-01  8.2204e-03  -17.0680 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix 3

```
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.252e-01  2.898e-02  21.576  < 2e-16 ***
danceability      6.338e-01  1.345e-02  47.115  < 2e-16 ***
energy           -2.588e-01  1.639e-02 -15.788  < 2e-16 ***
loudness          1.358e-02  7.071e-04  19.204  < 2e-16 ***
mode              6.133e-02  4.623e-03  13.267  < 2e-16 ***
speechiness      -5.306e-01  2.578e-02 -20.583  < 2e-16 ***
acousticness     -2.858e-01  9.541e-03 -29.953  < 2e-16 ***
instrumentalness -4.530e-01  7.963e-03 -56.883  < 2e-16 ***
liveness         -8.344e-02  1.275e-02  -6.542 6.13e-11 ***
tempo             2.185e-04  7.538e-05   2.898 0.003754 **
time_signature    2.121e-02  5.159e-03   4.111 3.94e-05 ***
chorus_hit       -4.187e-04  1.116e-04  -3.752 0.000176 ***
sev              -5.116e-02  6.857e-03  -7.460 8.80e-14 ***
eig              -1.161e-01  7.384e-03 -15.726  < 2e-16 ***
nin              -1.114e-01  7.776e-03 -14.325  < 2e-16 ***
zer              -1.256e-01  7.887e-03 -15.919  < 2e-16 ***
ten              -1.367e-01  7.755e-03 -17.628  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix 4

```
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.106e+00  1.796e-01   6.159 7.31e-10 ***
danceability      4.056e+00  9.869e-02  41.102  < 2e-16 ***
energy           -1.826e+00  1.025e-01 -17.825  < 2e-16 ***
loudness          1.326e-01  4.893e-03  27.102  < 2e-16 ***
mode              3.637e-01  2.617e-02  13.897  < 2e-16 ***
speechiness      -2.864e+00  1.563e-01 -18.321  < 2e-16 ***
acousticness     -1.811e+00  5.627e-02 -32.176  < 2e-16 ***
instrumentalness -3.427e+00  6.861e-02 -49.948  < 2e-16 ***
liveness         -3.001e-01  7.026e-02  -4.271 1.95e-05 ***
valence          -2.729e-01  6.519e-02  -4.187 2.83e-05 ***
tempo             3.111e-03  4.327e-04   7.189 6.51e-13 ***
duration_ms       2.810e-07  2.713e-07   1.036 0.300238
time_signature    1.726e-01  3.254e-02   5.304 1.13e-07 ***
chorus_hit       -2.403e-03  7.086e-04  -3.390 0.000698 ***
sections         -1.573e-02  6.426e-03  -2.448 0.014350 *
sev              -3.255e-01  3.925e-02  -8.293  < 2e-16 ***
eig              -7.768e-01  4.335e-02 -17.922  < 2e-16 ***
nin              -8.421e-01  4.681e-02 -17.989  < 2e-16 ***
zer              -1.099e+00  4.967e-02 -22.126  < 2e-16 ***
ten              -1.224e+00  4.996e-02 -24.492  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix 5

```
                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)       1.0990507  0.1794446   6.125 9.08e-10 ***
danceability      4.0563721  0.0986937  41.101  < 2e-16 ***
energy           -1.8180053  0.1021269 -17.801  < 2e-16 ***
loudness          0.1323399  0.0048852  27.090  < 2e-16 ***
mode              0.3629106  0.0261608  13.872  < 2e-16 ***
speechiness      -2.8649709  0.1563772 -18.321  < 2e-16 ***
acousticness     -1.8128913  0.0562359 -32.237  < 2e-16 ***
instrumentalness -3.4220148  0.0684208 -50.014  < 2e-16 ***
liveness         -0.2990507  0.0702625  -4.256 2.08e-05 ***
valence          -0.2800585  0.0648227  -4.320 1.56e-05 ***
tempo             0.0030561  0.0004294   7.117 1.11e-12 ***
time_signature    0.1729615  0.0325392   5.315 1.06e-07 ***
chorus_hit       -0.0021271  0.0006569  -3.238  0.00120 ** 
sections         -0.0099214  0.0031325  -3.167  0.00154 ** 
sev              -0.3215955  0.0390682  -8.232  < 2e-16 ***
eig              -0.7723185  0.0431252 -17.909  < 2e-16 ***
nin              -0.8368020  0.0465273 -17.985  < 2e-16 ***
zer              -1.0936160  0.0493931 -22.141  < 2e-16 ***
ten              -1.2196753  0.0498125 -24.485  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Appendix 6

```
                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)       1.057e+00  1.784e-01   5.926 3.11e-09 ***
danceability      4.053e+00  9.866e-02  41.074  < 2e-16 ***
energy           -1.809e+00  1.022e-01 -17.704  < 2e-16 ***
loudness          1.323e-01  4.890e-03  27.059  < 2e-16 ***
mode              3.637e-01  2.617e-02  13.897  < 2e-16 ***
speechiness      -2.863e+00  1.564e-01 -18.298  < 2e-16 ***
acousticness     -1.813e+00  5.626e-02 -32.228  < 2e-16 ***
instrumentalness -3.419e+00  6.851e-02 -49.913  < 2e-16 ***
liveness         -2.991e-01  7.026e-02  -4.257 2.07e-05 ***
valence          -2.805e-01  6.511e-02  -4.308 1.65e-05 ***
tempo             2.970e-03  4.288e-04   6.927 4.30e-12 ***
duration_ms      -3.036e-07  1.316e-07  -2.308  0.02100 *  
time_signature    1.731e-01  3.255e-02   5.319 1.04e-07 ***
chorus_hit       -1.707e-03  6.492e-04  -2.629  0.00856 ** 
sev              -3.236e-01  3.924e-02  -8.248  < 2e-16 ***
eig              -7.756e-01  4.334e-02 -17.897  < 2e-16 ***
nin              -8.381e-01  4.677e-02 -17.919  < 2e-16 ***
zer              -1.094e+00  4.963e-02 -22.051  < 2e-16 ***
ten              -1.220e+00  4.992e-02 -24.428  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Appendix 7

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.0058164	0.1770734	5.680	1.35e-08	***
danceability	4.0466142	0.0986044	41.039	< 2e-16	***
energy	-1.8184631	0.1021438	-17.803	< 2e-16	***
loudness	0.1330698	0.0048828	27.253	< 2e-16	***
mode	0.3674979	0.0261169	14.071	< 2e-16	***
speechiness	-2.8570800	0.1563840	-18.270	< 2e-16	***
acousticness	-1.8073350	0.0561908	-32.164	< 2e-16	***
instrumentalness	-3.4288092	0.0684907	-50.062	< 2e-16	***
liveness	-0.3024655	0.0702072	-4.308	1.65e-05	***
valence	-0.2602273	0.0645156	-4.034	5.49e-05	***
tempo	0.0029836	0.0004287	6.959	3.43e-12	***
time_signature	0.1722544	0.0325638	5.290	1.22e-07	***
chorus_hit	-0.0017942	0.0006479	-2.769	0.00562	**
sev	-0.3388446	0.0386854	-8.759	< 2e-16	***
eig	-0.7945841	0.0425651	-18.668	< 2e-16	***
nin	-0.8560016	0.0461450	-18.550	< 2e-16	***
zer	-1.1117450	0.0490558	-22.663	< 2e-16	***
ten	-1.2309298	0.0496815	-24.776	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Appendix 8

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	6.228e-01	1.063e-01	5.859	4.66e-09	***
danceability	2.387e+00	5.758e-02	41.448	< 2e-16	***
energy	-1.041e+00	6.044e-02	-17.223	< 2e-16	***
loudness	7.723e-02	2.861e-03	26.992	< 2e-16	***
mode	2.150e-01	1.556e-02	13.819	< 2e-16	***
speechiness	-1.704e+00	9.150e-02	-18.623	< 2e-16	***
acousticness	-1.049e+00	3.308e-02	-31.711	< 2e-16	***
instrumentalness	-1.938e+00	3.646e-02	-53.153	< 2e-16	***
liveness	-1.838e-01	4.194e-02	-4.381	1.18e-05	***
valence	-1.521e-01	3.857e-02	-3.942	8.07e-05	***
tempo	1.787e-03	2.585e-04	6.912	4.77e-12	***
duration_ms	1.290e-07	1.589e-07	0.812	0.416917	
time_signature	1.020e-01	1.926e-02	5.298	1.17e-07	***
chorus_hit	-1.460e-03	4.191e-04	-3.483	0.000496	***
sections	-8.755e-03	3.775e-03	-2.319	0.020370	*
sev	-1.841e-01	2.328e-02	-7.905	2.68e-15	***
eig	-4.555e-01	2.566e-02	-17.753	< 2e-16	***
nin	-4.870e-01	2.761e-02	-17.641	< 2e-16	***
zer	-6.370e-01	2.923e-02	-21.792	< 2e-16	***
ten	-7.133e-01	2.941e-02	-24.255	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Appendix 9

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	0.6195091	0.1062206	5.832	5.47e-09	***
danceability	2.3865834	0.0575799	41.448	< 2e-16	***
energy	-1.0370237	0.0602436	-17.214	< 2e-16	***
loudness	0.0771084	0.0028569	26.990	< 2e-16	***
mode	0.2145705	0.0155489	13.800	< 2e-16	***
speechiness	-1.7042421	0.0915179	-18.622	< 2e-16	***
acousticness	-1.0501199	0.0330625	-31.762	< 2e-16	***
instrumentalness	-1.9361885	0.0363690	-53.237	< 2e-16	***
liveness	-0.1834125	0.0419438	-4.373	1.23e-05	***
valence	-0.1553084	0.0383620	-4.049	5.15e-05	***
tempo	0.0017619	0.0002567	6.865	6.65e-12	***
time_signature	0.1021889	0.0192546	5.307	1.11e-07	***
chorus_hit	-0.0013333	0.0003889	-3.428	0.000608	***
sections	-0.0060780	0.0018252	-3.330	0.000868	***
sev	-0.1822872	0.0231770	-7.865	3.69e-15	***
eig	-0.4534652	0.0255343	-17.759	< 2e-16	***
nin	-0.4846265	0.0274463	-17.657	< 2e-16	***
zer	-0.6344900	0.0290689	-21.827	< 2e-16	***
ten	-0.7114343	0.0293213	-24.263	< 2e-16	***

Appendix 10

```
                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)       5.955e-01  1.056e-01   5.637 1.73e-08 ***
danceability      2.385e+00  5.756e-02  41.425  < 2e-16 ***
energy           -1.031e+00  6.029e-02 -17.108  < 2e-16 ***
loudness          7.708e-02  2.860e-03  26.953  < 2e-16 ***
mode              2.149e-01  1.556e-02  13.814  < 2e-16 ***
speechiness      -1.703e+00  9.155e-02 -18.600  < 2e-16 ***
acousticness     -1.050e+00  3.308e-02 -31.756  < 2e-16 ***
instrumentalness -1.934e+00  3.641e-02 -53.134  < 2e-16 ***
liveness         -1.834e-01  4.194e-02  -4.372 1.23e-05 ***
valence          -1.562e-01  3.853e-02  -4.054 5.04e-05 ***
tempo             1.709e-03  2.563e-04   6.669 2.57e-11 ***
duration_ms      -1.958e-07  7.672e-08  -2.552  0.01070 *  
time_signature    1.023e-01  1.926e-02   5.313 1.08e-07 ***
chorus_hit       -1.075e-03  3.850e-04  -2.792  0.00524 ** 
sev              -1.832e-01  2.328e-02  -7.870 3.54e-15 ***
eig              -4.549e-01  2.566e-02 -17.729  < 2e-16 ***
nin              -4.848e-01  2.759e-02 -17.573  < 2e-16 ***
zer              -6.343e-01  2.921e-02 -21.717  < 2e-16 ***
ten              -7.110e-01  2.939e-02 -24.194  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix 11

```
                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)       0.5622471  0.1048935   5.360 8.31e-08 ***
danceability      2.3810563  0.0575339  41.385  < 2e-16 ***
energy           -1.0376246  0.0602555 -17.220  < 2e-16 ***
loudness          0.0775595  0.0028553  27.163  < 2e-16 ***
mode              0.2173547  0.0155240  14.001  < 2e-16 ***
speechiness      -1.6996746  0.0915208 -18.571  < 2e-16 ***
acousticness     -1.0465792  0.0330409 -31.675  < 2e-16 ***
instrumentalness -1.9392384  0.0363975 -53.279  < 2e-16 ***
liveness         -0.1851827  0.0419173  -4.418 9.97e-06 ***
valence          -0.1431024  0.0381859  -3.748 0.000179 ***
tempo             0.0017186  0.0002563   6.706 2.00e-11 ***
time_signature    0.1017855  0.0192684   5.283 1.27e-07 ***
chorus_hit       -0.0011329  0.0003841  -2.949 0.003184 ** 
sev              -0.1931079  0.0229501  -8.414  < 2e-16 ***
eig              -0.4672087  0.0252098 -18.533  < 2e-16 ***
nin              -0.4963268  0.0272153 -18.237  < 2e-16 ***
zer              -0.6457550  0.0288653 -22.371  < 2e-16 ***
ten              -0.7185000  0.0292405 -24.572  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```