Group A6
Tamzeed Quazi
Aronn Laurel
Frances Shen

Regression Model for Spotify Streams

Introduction

Motivated by our group's curiosity and common interest in music, our research aims to explore and build a suitable model to examine the relationship between a song's Spotify streaming number with its music and popularity classifications. Determining a song's popularity by its stream volume, various research on this topic has found that songs tend to be more successful if they exhibit a more 'party-like' or 'sad' dynamic in comparison to songs that display 'brightness' or 'happiness' (Interiano et al, 2018). Similarly, our research will explore the following research question:

"Can we find an optimal model for predicting the number of Spotify streams of a song based on different song attributes?"

In this exploratory study, we will focus on data extracted from Spotify in 2023 and apply a linear regression model for Spotify streams to attempt to determine what makes a song succeed in streaming volume during 2023. Our findings could be used to explain broader questions about the music industry and consumption.

Data Collection

The data is sourced from Spotify, which is a music streaming platform but was collected and found on Kaggle (2023). It comprises some of the most streamed songs on Spotify up to the year 2023 with columns containing track name, artist(s) name, release date of the song, song BPM, and etc. Below is a snippet of our dataframe.

track_nam	artist(s)_na	artist_cour	released_y	released_r	released_c	in_spotify_	in_spotify_	streams	in_apple_p	in_apple_c	in_deezer_	in_deezer_	in_shazam	bpm	key	mode	danceabili	valence_%	energy_%	acousticne	instrument	liveness_%	speechiness_
Flowers	Miley Cyrus	1	2023	1	12	12211	115	1.32E+09	300	215	745	58	1,021	118		Major	71	65	68	6	0	3	7
Ella Baila S	Eslabon Ar	2	2023	3	16	3090	50	7.26E+08	34	222	43	13	418	148	F	Minor	67	83	76	48	0	8	3
Shakira: Bz	Shakira, Bi	2	2023	1	11	5724	44	7.22E+08	119	108	254	29	22	122	D	Minor	78	50	63	27	0	9	5
TQG	Karol G, Sh	2	2023	2	23	4284	49	6.19E+08	115	123	184	18	354	180	E	Minor	72	61	63	67	0	9	28
La Bebe - P	Peso Plum	2	2023	3	17	2953	44	5.54E+08	49	110	66	13	339	170	D	Minor	81	56	48	21	0	8	33
Die For Yo	Ariana Gra	2	2023	2	24	3408	47	5.19E+08	87	86	74	1	16	67	C#	Minor	53	50	53	23	0	44	7
un x100to	Bad Bunny	2	2023	4	17	2876	40	5.06E+08	41	205	54	12	251	83	F#	Minor	57	56	72	23	0	27	5
Cupid - Tw	Fifty Fifty	1	2023	2	24	2942	77	4.97E+08	91	212	78	6	0	120	B	Minor	78	76	59	43	0	34	3
PRC	Natanael C	2	2023	1	23	961	26	4.36E+08	19	143	10	6	15	138	G	Minor	78	89	83	10	0	12	5
OMG	NewJeans	1	2023	1	2	1783	27	4.31E+08	26	124	15	1	22	127	A	Minor	80	74	77	36	0	11	4
Last Night	Morgan Wa	1	2023	1	31	2420	19	4.3E+08	52	107	15	1	325	204	F#	Major	52	52	68	46	0	15	4
Daylight	David Kush	1	2023	4	14	3528	98	3.88E+08	80	156	182	24	1,281	130	D	Minor	51	32	43	83	0	9	3
Like Crazy	Jimin	1	2023	3	24	596	68	3.63E+08	8	104	23	2	29	120	G	Major	63	36	73	0	0	36	4
BESO	Rauw Alejar	2	2023	3	24	4053	50	3.58E+08	82	121	182	12	171	95	F	Minor	77	53	64	74	0	17	14
El Azul	Junior H, P	2	2023	2	10	692	25	3.54E+08	10	107	6	3	62	144	A	Minor	56	84	65	23	0	10	6
Classy 101	Feid, Youn	2	2023	3	31	2610	40	3.35E+08	43	100	54	14	187	100	B	Major	86	67	66	14	0	12	16
Fin de Sem	Oscar May	2	2023	1	13	592	14	3.07E+08	11	84	6	1	30	98		Major	70	37	54	6	0	9	8
WHERE SI	Bad Bunny	1	2023	5	18	3133	50	3.03E+08	84	133	87	15	425	144	A	Minor	65	23	80	14	63	11	6
X SI VOLVE	Karol G, Ro	2	2023	2	2	2127	33	2.67E+08	45	80	53	8	4	178	C#	Minor	79	58	78	34	0	11	25
Moonlight	Kali Uchis	1	2023	2	24	2649	42	2.56E+08	67	79	57	1	615	137	G	Minor	64	88	72	51	0	17	5
El Gordo Tr	Chino Paca	1	2023	1	27	539	21	2.56E+08	7	71	4	2	13	140	G	Minor	74	96	80	18	0	5	5

In line with our project's goal, we will be filtering the data into columns as follows:

Description of Variables		
Variable	**Variable Type (Categorical/Continuous)**	**Description (Units)**
streams	Response Variable Continuous	Total number of streams on Spotify **(# streams)** as of Sep 2023. *Range: [2762, 3703895074]* *Mean: 566856949*
artist_count	Explanatory Variable Continuous	Number of artists contributing to the song **(# artists)** *Range: [1, 8]* *Mean: 1.56*
bpm	Explanatory Variable Continuous	Beats per minute, a measure of song tempo **(beats per minute)** *Range: [65, 206]* *Mean: 122.55*
danceability	Explanatory Variable Continuous	Percentage indicating how suitable the song is for dancing **(%)** *Range: [0, 100]* *Mean: 66.98*
energy	Explanatory Variable Continuous	Energy is a measurement representing a perceptual measure of intensity and activity. **(%)** *Range: [0, 100]* *Mean: 64.27*
acoustic-ness	Explanatory Variable Continuous	Amount of acoustic sound in the song **(%)** *Range: [0, 100]* *Mean: 27.08*

liveness	Explanatory Variable Continuous	Presence of live performance elements **(%)** *Range: [0, 100]* *Mean: 18.21*
speech-ness	Explanatory Variable Continuous	Amount of spoken words in the song **(%)** *Range: [0, 100]* *Mean: 10.14*

Analysis

Data Exploration

Our analysis begins with some preliminary visualizations of the data we are working with. Our plots include a histogram of each of the variables we are interested in to get an idea of the scale and plan for our methodology.

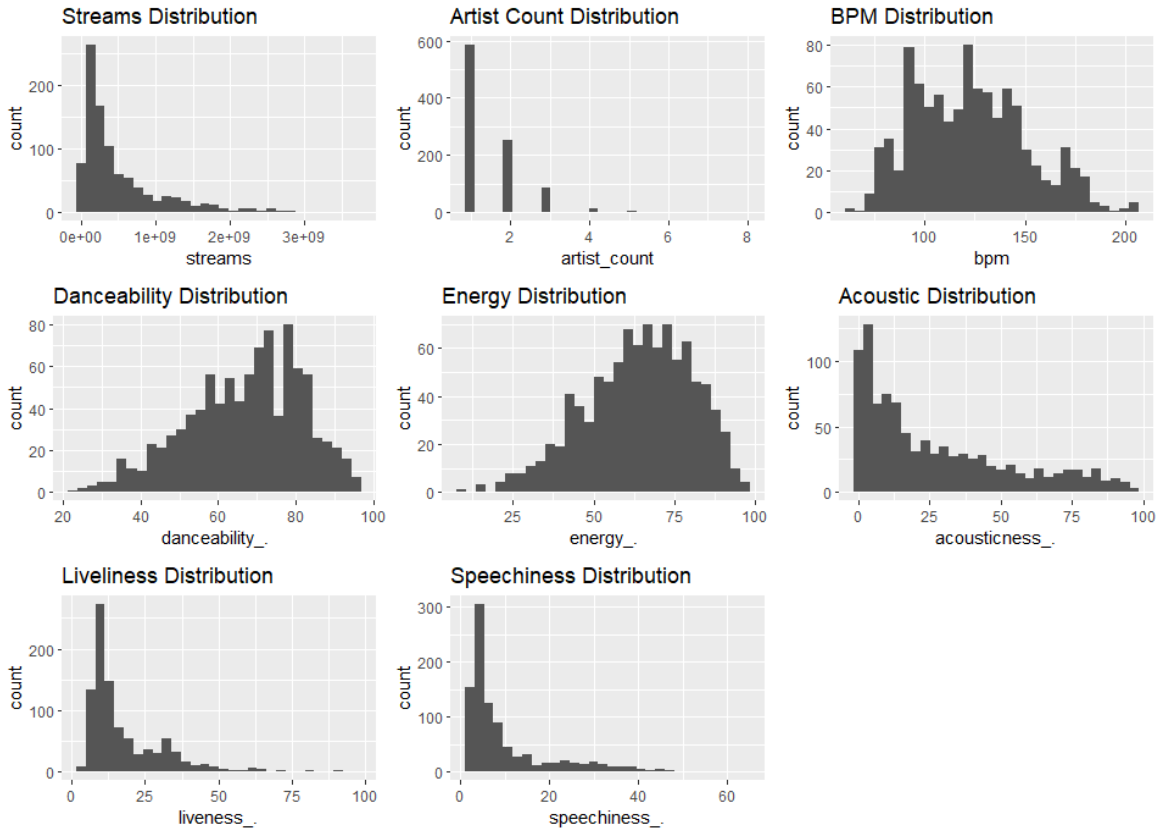

Figure 1.1 - Variable Histograms

The distributions of each variable above indicate different characteristics. In terms of the magnitude, we see that most of our variables follow a very similar scale so not much scaling is required to help with interpretation.

In terms of the shape of the underlying distribution of each variable, which fundamentally affects the goodness of fit of the fitted linear regression model, we noticed that ‘stream’, our response variable has a highly right-skewed distribution, hence, the variability of the response variable may not be the same across all levels of predictor variables, which leads to concerns for heteroscedasticity, a violation of the constant residual variance assumption. Moreover, the skewness can also impact the normality of residuals, which is another assumption for linear regression models.

Regarding collinearity between variables, we created a correlation matrix plot to make sure we do not run into issues of collinearity between our predictor variables.

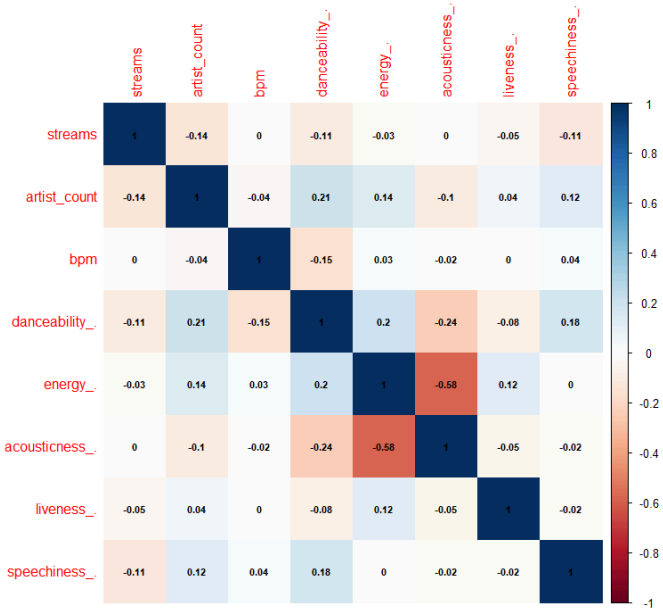

Figure 1.2 - Variable Correlation Matrix

Analyzing the correlation coefficient between streams and other predictor variables, the correlation matrix plot has displayed a weak and negative correlation for artist count, danceability, energy, liveliness, and speech-ness whereas beat per minute and acoustic shows a weak positive correlation. These findings suggest that the predictor variables are not strongly related to the number of streams individually, however, there remains a possibility for a stronger correlation if we integrate different factors into the model.

Methodology

In the ‘Data Exploration’ part, we discovered that the distribution of ‘streams’ for a song is highly right-skewed and is on a much larger scale than our other variables. To make sure the linear model fitted to the dataset has adequate predictive power, we will experiment with performing transformations on ‘streams’, which is our response

variable. From our correlation plot, we also see that some variables have very minor correlations with our response variable, so we will define a full model that includes all variables and reduce the model to a simpler and more effective one. To land on an optimal model, we will take a **backward elimination** and **exhaustive search** approach to find which variables offer the best model and compare the features from both approaches to find an optimal model.

The Model

By fitting our full model with the streams as is and creating a residual plot (Figure 2.1), our original data of fitted stream values displays a concern for unbiasedness and heteroscedasticity. Hence, we will rescale our response variables (streams) by a square root and logarithm to determine which has a much more desirable residual plot with a constant variance.

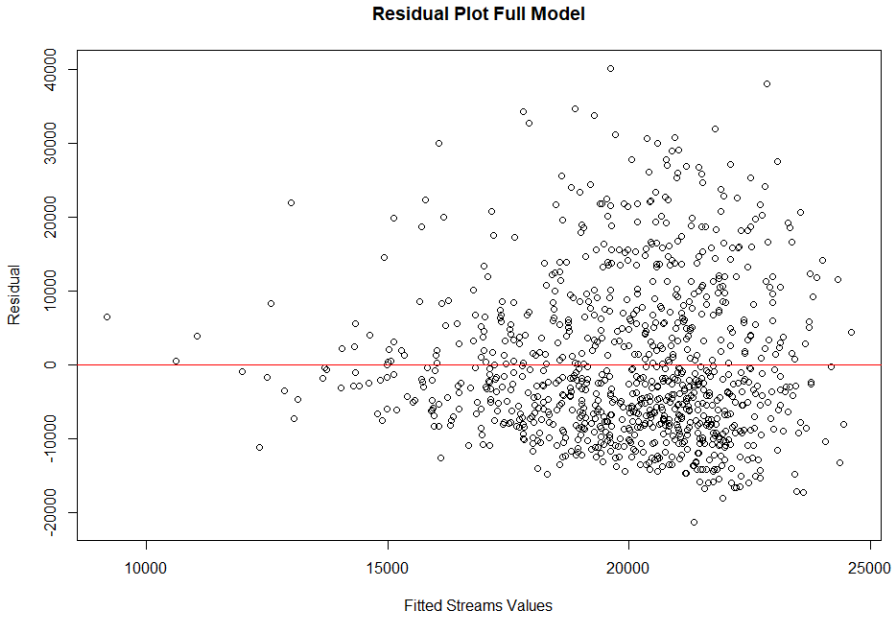

Figure 2.1 - Residual plot of full model and unscaled streams

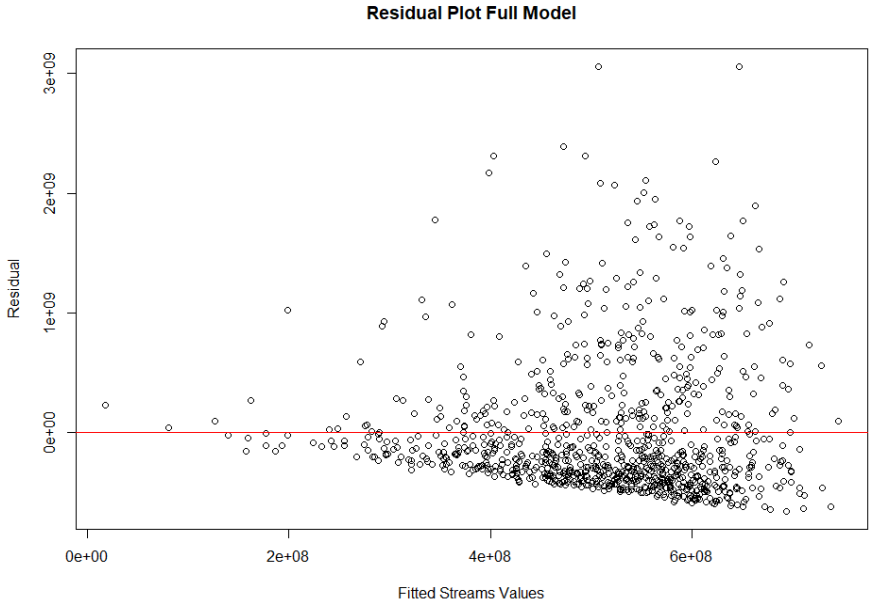

Figure 2.2 - Residual plot of full model and square rooted streams

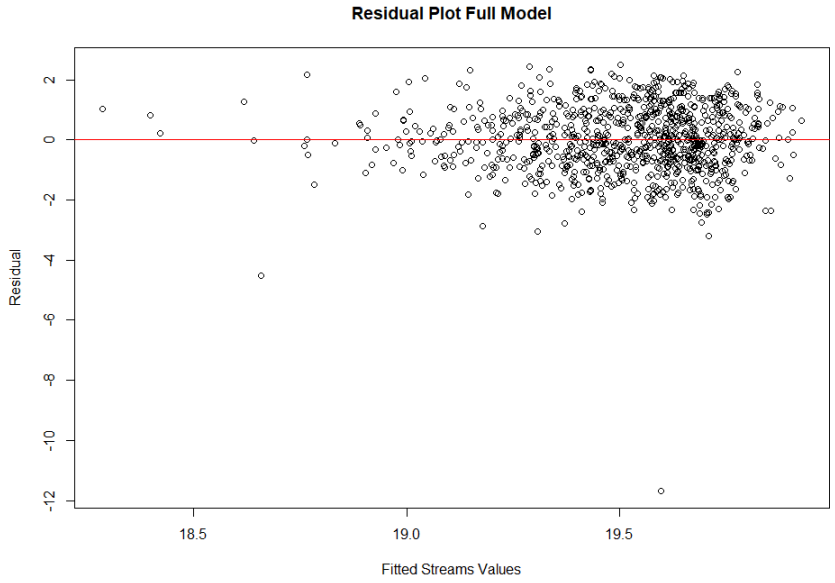

Figure 2.3 - Residual plot of full model and logged streams

The following is the model summary of our defined full model for logged streams:

```
> summary(full_model)

Call:
lm(formula = log(streams) ~ artist_count + bpm + danceability_. +
    energy_. + acousticness_. + liveness_. + speechiness_., data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-11.6725  -0.7015  -0.0167   0.7931   2.4919

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    20.4216241  0.3287486  62.119  < 2e-16 ***
artist_count   -0.1821298  0.0422854  -4.307 1.83e-05 ***
bpm             0.0000732  0.0013240   0.055   0.9559
danceability_. -0.0030256  0.0027154  -1.114   0.2655
energy_.       -0.0028162  0.0027450  -1.026   0.3052
acousticness_. -0.0034923  0.0017441  -2.002   0.0455 *
liveness_.     -0.0037173  0.0027030  -1.375   0.1694
speechiness_.  -0.0093176  0.0037818  -2.464   0.0139 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.128 on 944 degrees of freedom
Multiple R-squared:  0.03883,   Adjusted R-squared:  0.0317
F-statistic: 5.448 on 7 and 944 DF,  p-value: 3.795e-06
```

The 'Pr (> |t|)' column indicates that while predictors like '**artist count'**, '**acoustic-ness'**, and '**speech-ness'** show significant relationships with the log of streams, the overall model explains a small portion of the variance in the response variable, as reflected by an R-squared value of approximately 0.0388.

After performing backward elimination to simplify the model, we achieved the model

```
> summary(be_model)

Call:
lm(formula = log(streams) ~ artist_count + acousticness_. + speechiness_.,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-11.5716  -0.7188  -0.0316   0.8179   2.4187

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    19.966780   0.091327 218.629  < 2e-16 ***
artist_count   -0.197115   0.041456  -4.755  2.3e-06 ***
acousticness_. -0.002012   0.001414  -1.423  0.15518
speechiness_.  -0.009744   0.003717  -2.622  0.00889 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.128 on 948 degrees of freedom
Multiple R-squared:  0.03428,   Adjusted R-squared:  0.03122
F-statistic: 11.22 on 3 and 948 DF,  p-value: 3.096e-07
```

The model produced by backward elimination seems less ideal than our full model for a couple reasons. The R-squared values for the reduced model are slightly lower than the full model and the residual plot produced by the model are slightly shifted and less random.

Our next approach to creating a better model was to run an exhaustive search from our full model and compute Mallow's CP for each to find an optimal model from there.

```
> # Summary of the method
> summary(model_selection)$which
  (Intercept) artist_count   bpm danceability_. energy_. acousticness_. liveness_. speechiness_.
1        TRUE         TRUE FALSE          FALSE    FALSE          FALSE      FALSE         FALSE
2        TRUE         TRUE FALSE          FALSE    FALSE          FALSE      FALSE          TRUE
3        TRUE         TRUE FALSE          FALSE    FALSE           TRUE      FALSE          TRUE
4        TRUE         TRUE FALSE          FALSE    FALSE           TRUE       TRUE          TRUE
5        TRUE         TRUE FALSE           TRUE    FALSE           TRUE       TRUE          TRUE
6        TRUE         TRUE FALSE           TRUE     TRUE           TRUE       TRUE          TRUE
7        TRUE         TRUE  TRUE           TRUE     TRUE           TRUE       TRUE          TRUE
> # CP values of models
```

Similarly, using the function **regsubset** has presented the same model as we found using the backward elimination for a model with 4 variables. Next, we will plot the CP values for each number of parameters and we get the following:

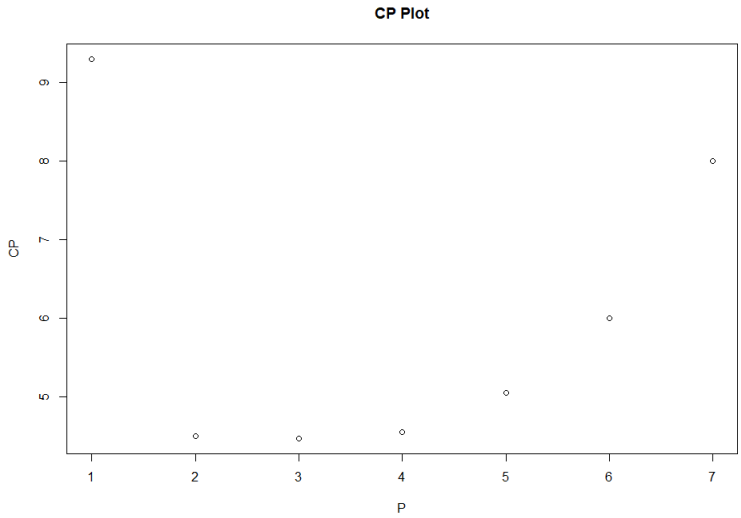

Figure 3 - Mallow's CP plot for exhaustive search

Models with 4, 5, and 6 parameters seem the most promising since C_p is closest to P, which is the number of parameters in the model. From our previous search, we already know the result of model 4 and have yet to explore models 5 and 6. Therefore, we decided to not pursue model 6 because it would introduce energy and acoustic-ness both into the model since energy and acoustic-ness displayed a moderate correlation from our preliminary search. Instead, we will use model 5 to avoid multicollinearity in our regression model which is also similar to the model produced from backward elimination but adds danceability as a predictor variable.

```
> summary(final_model)

Call:
lm(formula = log(streams) ~ artist_count + danceability_. + acousticness_. +
    liveness_. + speechiness_., data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-11.6648  -0.7128  -0.0186   0.8206   2.4750

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    20.248426   0.203123  99.686  < 2e-16 ***
artist_count   -0.185707   0.042115  -4.410 1.15e-05 ***
danceability_. -0.003265   0.002669  -1.223   0.2215
acousticness_. -0.002510   0.001453  -1.727   0.0845 .
liveness_.     -0.004030   0.002684  -1.501   0.1336
speechiness_.  -0.009134   0.003766  -2.425   0.0155 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.127 on 946 degrees of freedom
Multiple R-squared:  0.03776,   Adjusted R-squared:  0.03267
F-statistic: 7.424 on 5 and 946 DF,  p-value: 7.653e-07

> # Residual plot of final model
> plot(x=final_model$fitted.values, y=final_model$residuals, xlab = "Fitted Log Stre
```

This model is simpler than our full model and it improves our adjusted R-squared value and slightly reduces our residual standard error. There are also no apparent issues with the residual plot and looks very similar to the full model.

The values in ‘Pr(>|t|)’ column as well as the significance code indicate that the coefficients have a decent impact on the response variable. The residual plot does not show any clear pattern, which implies that there is no obvious violation of linear relationship assumption. Moreover, residual variance seems constant across the range of fitted values.

Conclusion

By ‘brute forcing’ the best regression model and finding the optimum number of parameters with Cp Mallow, we have finalized our regression model as follows:

$$Y = 20.248426 + (-0.185707)(\text{artist count}) + (-0.003265)(\text{danceability}) + (-0.002510)(\text{Acousticness}) + (-0.004030)(\text{Live-ness}) + (-0.009134)(\text{speech-ness})$$

Y: logged song streams

Note: since we are using a log(stream_values), we will have to do 2^(Y) to compute the actual number of streams.

Discussions & Concerns

As reflected in our model, the process of predicting streams is inherently complex as song popularity is influenced by a multitude of factors beyond the scope of the current model. For future research, expanding the dataset to include additional variables, such as genre, playlist inclusion, artist popularity, and social media presence could potentially improve the model's explanatory power. Further exploring more complex models or machine learning algorithms that can capture non-linear relationships and interactions may yield better predictive performance as well. Our current analysis highlights the complexity of predicting song stream volumes on one streaming platform. For instance, certain attributes like a lower artist count and reduced speech-ness are associated with more streams, yet these factors alone cannot robustly predict the popularity of a song.

Since we have created a regression model using Spotify data as of 31st December 2023, any prediction from the model would be reasonable up to that year. However, it is important to consider that our 'best' predictive variables for the model could vary for the year 2024 due to non-statistical factors such as changes in popular culture. Additionally, our regression model limits us from predicting within the range of observed data. Any predictions beyond our range would cause the problem of extrapolation and our predictions would be unreliable.

Furthermore, it is important to also acknowledge unmeasurable factors that may influence the number of streams of a song. Some potential confounding factors could include the degree of marketing and promotion efforts that were put into a song, the existing popularity of the artists in the song, the demographic of the song type, and the frequency of song usage across social media, etc.

Reference

Interiano, M., Kazemi, K., Wang, L., Yang, J., Yu, Z., & Komarova, N. L. (2018). Musical trends and predictability of success in contemporary songs in and out of the top charts. In Royal Society Open Science (Vol. 5, Issue 5, p. 171274). The Royal Society. https://doi.org/10.1098/rsos.171274