The Hit Song Formula: A Methodological Journey

An Applied Regression Analysis of Spotify Song Popularity

Saranath P S Shriprasad

Indian Institute of Technology Madras

MA 5013: Applied Regression Analysis

Fall 2025

Introduction: The Quest for a Hit Song

Motivation

The music industry perpetually seeks the "recipe" for a hit song. While art is subjective, can we find statistical patterns in a song's audio features that correlate with its success?

Objective & Response

Objective: To build a valid regression model predicting a song's popularity on Spotify.
Response Variable: y (popularity), a continuous score from 0 to 100 from the Spotify API.

Figure 1: Dataset from Kaggle's "Ultimate Spotify Tracks Database". We analyzed a random sample of 10,000 songs.

Our Guiding Research Questions

Our investigation was structured around four key hypotheses:

1. **The "Big Five":** Which core audio features (*danceability, energy, valence, etc.*) have the most significant impact on popularity?
2. **The "Goldilocks Zone" Hypothesis:** Do non-linear "sweet spots" exist? Are songs with moderate *tempo* and *duration* more popular than extremes?
3. **The "Sad Banger" Phenomenon:** Does the combination of high energy and low valence (sadness) lead to disproportionately popular songs? (Interaction: $energy \times valence$)
4. **The "Acoustic Amplification" Effect:** Does a song's acoustic nature change the impact of its emotional tone on popularity? (Interaction: $acousticness \times valence$)

Exploratory Data Analysis (EDA): Initial Insights

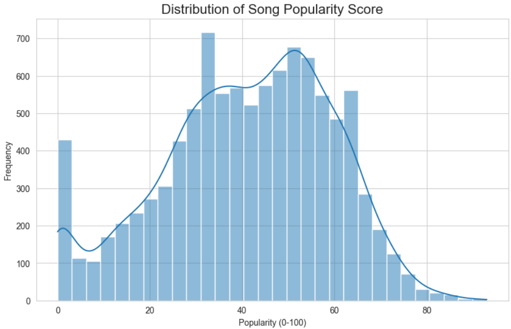

Figure 2: Popularity distribution: bell-shaped with spike near zero.

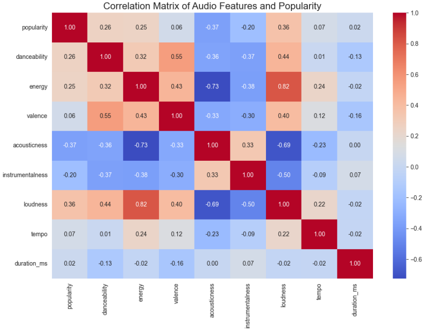

Figure 3: Correlation matrix shows multicollinearity.

Part 1: The Baseline OLS Model

Our first step was to fit a standard Ordinary Least Squares (OLS) model using the "Big Five" predictors.

Initial Model Specification

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon$$

Variable Definitions

- y = popularity
- x_1 = danceability
- x_2 = energy
- x_3 = valence
- x_4 = acousticness
- x_5 = instrumentalness

This model serves as our naive baseline, which we must rigorously validate before accepting its results.

Baseline OLS Model: Summary Table

OLS Regression Results

Model Summary		
Dep. Variable: y (popularity)	R-squared: 0.184	F-statistic: 450.5
Model: OLS	Adj. R-squared: 0.184	Prob (F-statistic): 0.00
Method: Least Squares	No. Observations: 10000	AIC: 8.457e+04

Coefficients

Variable	**coef**	**std err**	**t**	$P > \lvert t \rvert$	**[0.025**	**0.975]**
β_0 (const)	43.72	0.986	44.34	0.000	41.78	45.65
β_1 (x_1: danceability)	20.73	1.124	18.45	0.000	18.53	22.94
β_2 (x_2: energy)	-1.64	0.978	-1.68	0.093	-3.56	0.28
β_3 (x_3: valence)	-12.85	0.805	-15.97	0.000	-14.43	-11.28
β_4 (x_4: acousticness)	-18.27	0.697	-26.21	0.000	-19.63	-16.90
β_5 (x_5: instrumentalness)	-4.40	0.617	-7.12	0.000	-5.61	-3.19

Initial Interpretation

The model appears significant (F-stat), but 'energy' is not. 'danceability' has a strong positive effect, while others are negative. **But are these results valid?**

Diagnostic 1: Multicollinearity Detection

Variance Inflation Factor (VIF)

We test for multicollinearity using VIF. A VIF > 5 indicates a problem.

Problem Detected

Three variables exceed threshold:

- x_1: **10.17**
- x_2: **6.61**
- x_3: **6.50**

Impact: Coefficients and p-values unreliable.

VIF Scores Table

Variable	VIF
x_1 (danceability)	**10.17**
x_2 (energy)	**6.61**
x_3 (valence)	**6.50**
x_4 (acousticness)	2.14
x_5 (instrumentalness)	1.39

Multicollinearity Treatment Strategy

Treatment Approach

Remove highest VIF variable (x_1: danceability) first.

Why Remove Danceability?

- Highest VIF score (10.17)
- Sequential removal approach
- Maintains interpretability

Expected Outcome

After removing x_1:

- VIF scores decrease
- Coefficients more stable
- Tests more reliable

Diagnostic 2: Normality of Residuals

The Assumption

OLS regression assumes that the model's errors (residuals) are normally distributed. We check this with a Q-Q plot.

Diagnosis: SEVERE VIOLATION

The residuals deviate drastically from the theoretical normal line.
This heavy-tailed 'S' curve invalidates all p-values, confidence intervals, and hypothesis tests from the OLS summary.

Conclusion: The OLS model is statistically invalid.

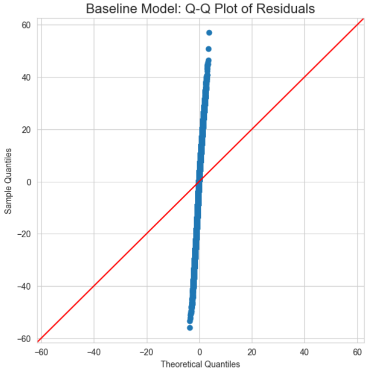

Figure 4: Q-Q Plot for the Baseline OLS Model

Identifying the Root Cause: Skewness or Outliers?

What is causing the non-normal residuals?

Hypothesis: The response variable is skewed

If skewed, Box-Cox transformation should normalize residuals.

Box-Cox Results

Optimal lambda = $\mathbf{1.0252} \approx 1$ (no transformation needed).
This suggests transformation will not fix normality.

Conclusion

Issue is likely **not simple skewness**. Need to investigate outliers.

Box-Cox Results: Visual Evidence

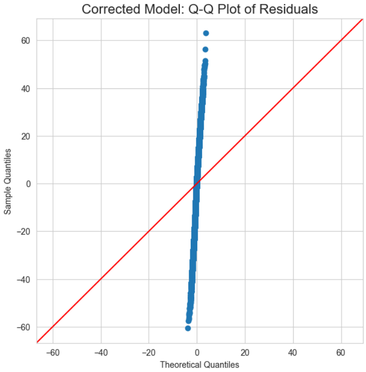

Figure 5: Q-Q Plot after Box-Cox transformation

Key Observations

- Heavy tails persist after transformation
- Points deviate from normal line
- No improvement in normality

Crucial Insight

The problem is not simple skewness. The heavy tails point to **influential outliers**.

Next Step

Identify influential observations using Cook's Distance.

Confirming the Root Cause: Influential Outliers

Cook's Distance

Cook's Distance measures how much the entire regression model changes when a single observation is removed. High values indicate influential points.

Diagnosis: Severe Problem Detected

The plot reveals numerous points with high influence.

We identified **455 influential outliers** (where Cook's $D > 4/n$). These points are pulling the regression line and distorting the residuals for all other points.

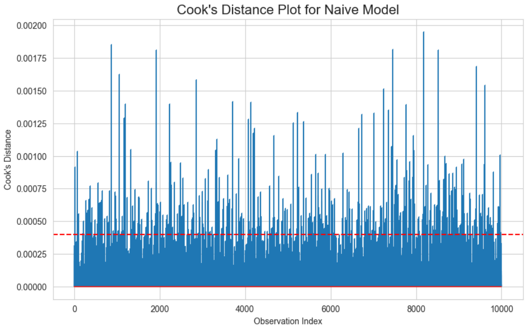

Figure 6: Cook's Distance Plot for Naive OLS Model

The Right Tool: Robust Linear Models (RLM)

Since the problem is outliers, we need a method designed to handle them.

Robust Linear Model (RLM)

RLM works by an iterative process (IRLS) that systematically down-weights the influence of observations identified as outliers.

This forces the model to fit the bulk of the data, not the extremes.

Key Advantage

RLM automatically identifies and reduces the impact of problematic observations.

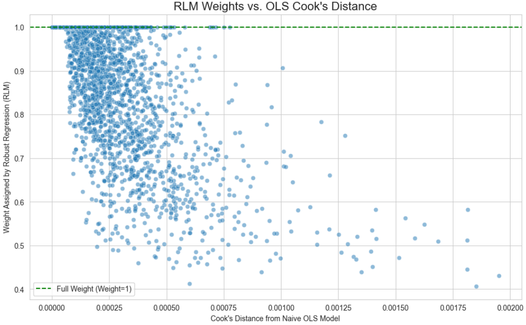

Figure 7: Proof RLM is working: Points with high Cook's Distance are assigned low weights

Applying RLM to the Baseline Model

Robust Regression Results (with all "Big Five")

RLM Model Summary

Dep. Variable: y (popularity)	No. Observations: 10000	Df Model: 5
Model: RLM	Df Residuals: 9994	Method: IRLS

Coefficients

Variable	**coef**	**std err**	**z**	$P > \|z\|$	**[0.025**	**0.975]**
β_0 (const)	44.59	1.031	43.23	0.000	42.57	46.61
β_1 (x_1: danceability)	22.20	1.175	18.88	0.000	19.89	24.50
β_2 (x_2: energy)	-4.07	1.023	-3.98	0.000	-6.08	-2.07
β_3 (x_3: valence)	-11.89	0.842	-14.12	0.000	-13.54	-10.24
β_4 (x_4: acousticness)	-18.98	0.729	-26.03	0.000	-20.41	-17.55
β_5 (x_5: instrumentalness)	-4.64	0.646	-7.18	0.000	-5.91	-3.37

A Key Insight Emerges

In the OLS model, x_2 (energy) was insignificant (p=0.093). In the robust model, x_2 **(energy) is now highly significant (p $<$ 0.001)**. The outliers were masking its true negative relationship with popularity!

Creating a Stable Baseline: Removing Multicollinearity

Now that we have a robust method, we can safely remove the collinear predictor (‘danceability‘) identified earlier.

Final Stable Baseline RLM Results

Final Baseline Model						
Model: RLM		No. Obs: 10000		Df Model: 4		
Variable	**coef**	**std err**	**z**	$P > \|z\|$	**[0.025**	**0.975]**
β_0 (const)	56.22	0.847	66.41	0.000	54.56	57.88
β_2 (x_2: energy)	-5.97	1.039	-5.75	0.000	-8.01	-3.94
β_3 (x_3: valence)	-4.83	0.761	-6.34	0.000	-6.32	-3.34
β_4 (x_4: acousticness)	-21.79	0.729	-29.87	0.000	-23.22	-20.36
β_5 (x_5: instrumentalness)	-7.46	0.642	-11.61	0.000	-8.72	-6.20

Final VIF Scores

Variable	VIF
x_3 (valence)	**4.82**
x_2 (energy)	**4.44**
x_4 (acousticness)	1.75
x_5 (instrumentalness)	1.39

Success!

All VIF scores now < 5. Multicollinearity resolved.

We now have a **doubly-corrected baseline model**: it is robust to outliers AND free of severe multicollinearity. This is our foundation.

The Full Model: Testing All Hypotheses

We now build the full model on our stable foundation, adding quadratic and interaction terms to test RQ2, RQ3, and RQ4.

Full Model RLM Results

Variable	coef	std err	z	$P > \|z\|$	[0.025	0.975]
β_0 (const)	60.09	1.526	39.39	0.000	57.11	63.09
β_2 (x_2: energy)	-8.37	1.910	-4.38	0.000	-12.12	-4.63
β_3 (x_3: valence)	-11.00	3.076	-3.58	0.000	-17.02	-4.97
β_4 (x_4: acousticness)	-25.33	1.416	-17.88	0.000	-28.10	-22.55
β_5 (x_5: instrumentalness)	-6.96	0.651	-10.69	0.000	-8.24	-5.68
tempo_c	0.012	0.006	1.94	0.053	-0.000	0.025
tempo_c_sq	-0.001	0.000	-7.61	0.000	-0.002	-0.001
duration_s_c	0.009	0.002	4.39	0.000	0.005	0.013
duration_s_c_sq	-1.75e-05	1.52e-06	-11.53	0.000	-2.05e-05	-1.46e-05
energy_valence_interact	4.95	3.802	1.30	0.193	-2.50	12.40
acoustic_valence_interact	10.49	2.801	3.75	0.000	5.00	15.98

Full Model Analysis

Observation

Some predictors not significant (energy×valence p=0.193, tempo p=0.053). Need backward elimination.

Key Findings

- Core features (x_2-x_5) highly significant
- Quadratic terms confirm non-linear effects
- acoustic×valence significant ($p < 0.001$)
- energy×valence not significant ($p = 0.193$)

Next Step

Remove non-significant predictors for parsimonious model.

The Final Parsimonious Model

After removing insignificant predictors (`energy_valence_interact` and `tempo_c`), we arrive at our final model where every variable is statistically significant ($p < 0.05$).

Final Parsimonious Model RLM Results

Variable	**coef**	**std err**	**z**	$P > \|z\|$	**[0.025**	**0.975]**
β_0 (const)	58.47	0.943	62.02	0.000	56.63	60.32
β_2 (x_2: energy)	-6.18	1.038	-5.95	0.000	-8.22	-4.15
β_3 (x_3: valence)	-7.32	1.055	-6.94	0.000	-9.39	-5.25
β_4 (x_4: acousticness)	-24.57	1.121	-21.92	0.000	-26.77	-22.38
β_5 (x_5: instrumentalness)	-6.83	0.645	-10.59	0.000	-8.10	-5.57
tempo_c_sq	-0.001	0.000	-7.38	0.000	-0.001	-0.001
duration_s_c	0.009	0.002	4.44	0.000	0.005	0.013
duration_s_c_sq	-1.75e-05	1.52e-06	-11.52	0.000	-2.05e-05	-1.45e-05
acoustic_valence_interact	8.43	2.046	4.12	0.000	4.42	12.44
energy_valence_interact	4.9520	3.802	1.302	0.193	-2.501	12.404

Answering Our Research Questions: Part I

Our final model provides clear answers:

RQ1: The "Big Five" **Confirmed.**

Energy, valence, acousticness, and *instrumentalness* are all significant **negative** predictors. More produced (less acoustic) songs are strongly associated with higher popularity.

RQ2: The "Goldilocks Zone" **Strongly Supported.**

Significant negative coefficients on squared terms for *tempo* and *duration* confirm an inverted U-shape. Songs that are too slow/fast or too short/long are less popular.

Key Insight: The most popular songs are highly produced with moderate tempo and duration.

Answering Our Research Questions: Part II

Continuing our analysis of interaction effects:

RQ3: The "Sad Banger" **Not Supported.**

The interaction term *energy* × *valence* was not statistically significant and was removed from the final model.

RQ4: The "Acoustic Amplification" **Supported.**

The interaction term *acousticness* × *valence* is significant and positive. For highly acoustic tracks, a song's emotional tone has a much weaker negative impact on its popularity.

Summary

3 out of 4 hypotheses were supported, revealing that song popularity follows predictable patterns with clear "sweet spots" and interaction effects.

Visual Confirmation: Partial Regression Plots

The grid provides powerful visual confirmation of our final model's findings. Each subplot displays the relationship between `popularity` and a single predictor, after controlling for all other variables.

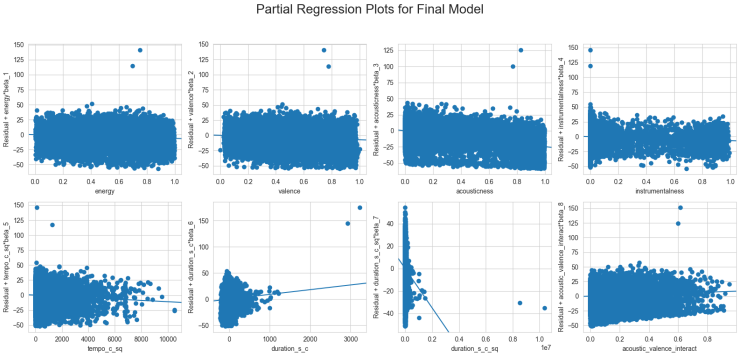

Figure 8: Partial regression plots showing isolated effects of each predictor

Interpreting Partial Plots: Linear Effects

Core Audio Features

Each shows a clear **downward-sloping blue line**.

Key Insight: Negative Relationships

Visually confirms negative linear relationships:

- **acousticness**: More acoustic → less popular
- **energy**: Higher energy → less popular
- **valence**: Happier songs → less popular
- **instrumentalness**: More instrumental → less popular

Statistical Validation

Downward slopes confirm negative drivers of popularity.

Interpreting Partial Plots: Goldilocks Effects

Tempo Squared Term

Clear **negative slope** - visual proof of inverted U-shape.

Tempo "Goldilocks Zone"

- X-axis: Distance from average tempo, squared
- Extreme tempos → lower popularity
- **Confirms**: Moderate tempos preferred

Duration Effects

- **Linear**: Slightly positive (minor preference for longer)
- **Squared**: Steeply negative (penalty for very long)

Duration Goldilocks

Initial gain overwhelmed by strong quadratic penalty.

Interpreting Partial Plots: Interaction Effect

Acoustic×Valence Interaction

Shows **positive slope** confirming significant interaction.

"Acoustic Amplification" Confirmed

- Positive trend: acousticness×valence has distinct influence
- For acoustic tracks, emotional tone matters less
- Goes beyond individual variable effects

Visual Validation Summary

Partial plots provide strong evidence for:

- Clear negative linear trends
- Powerful "Goldilocks" effects
- Significant positive interaction
- Isolated variable effects

Conclusion

The "Hit Song Formula" (According to our model)

The most popular songs tend to be:

- **Highly produced** (low *acousticness*, low *instrumentalness*).
- Of **moderate tempo and duration**—avoiding the extremes.
- Surprisingly, they lean towards lower *energy* and lower *valence* (less "happy").
- For acoustic songs, the emotional tone matters less for popularity.

Primary Methodological Takeaway

This project is a case study in the importance of rigorous, iterative model diagnostics. Identifying the **root cause** of a diagnostic failure (outliers vs. skewness) is critical for choosing the correct remedy and ultimately producing a valid and defensible final model.

Thank You

Questions?