Musical Carbon Dating

A Regression Analysis of Music Evolution (1960–2020)

Audiofool

University Statistical Analysis Project

December 18, 2025

Table of Contents

1. Introduction
2. Methodology: Linear Foundations
3. Diagnostics: The Audit
4. The Structural Break (1999)
5. Model Selection
6. Commercial Applications
7. Conclusion

The Research Question

The Arrow of Time

"Can we define the 'Arrow of Time' for music using only audio signal properties?"

Objective: To quantify the evolution of musical production styles from 1960 to 2020 using Regression Analysis.

Why it matters:

- **Commercial**: Recommendation systems (Spotify) need to understand "vintage" vs "modern" aesthetics beyond just metadata.
- **Cultural**: Quantifying the impact of the Digital Revolution (1999).

The Dataset

Source: Spotify 600k Tracks Dataset.

Filtering Criteria:

- Year: $1960 \leq T \leq 2020$.
- Popularity > 30 (Focus on culturally relevant tracks).
- Cleaned $N = \mathbf{250,971}$.

Source: Yamac Eren Ay (Kaggle, 2021)

Features ($p = 13$):

- **Physical**: `loudness`, `tempo`, `duration`.
- **Perceptual**: `acousticness`, `energy`, `valence`.
- **Musical**: `key`, `mode`.

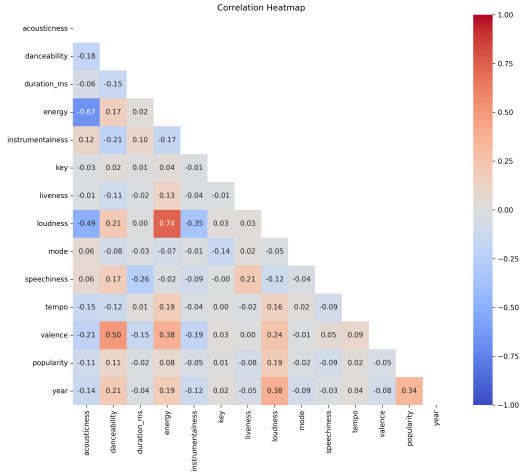


Figure: Feature Correlation Matrix (Note $r_{loud,energy} \approx 0.7$)

Phase II: Simple Linear Regression (The "Loudness War")

We started with a single predictor: **Loudness**.

$$y_i = \beta_0 + \beta_{loud} x_{i,loud} + \varepsilon_i$$

Result

- $R^2 = 0.144$.
- t-statistic $= 183.7$ ($p < 0.001$).
- **Interpretation**: Music has gotten significantly louder over time (+1.2 years per dB).

Phase III: Multiple Linear Regression (MLR)

We expanded to the Full Model ($p = 13$):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Model Performance:

- $R^2 = 0.296$.
- Test RMSE ≈ 12.06 years.

Key Insights:

- `Acousticness`: Strong negative trend ($\beta \approx -2.8$).
- `Danceability`: Positive trend ($\beta \approx +22$).

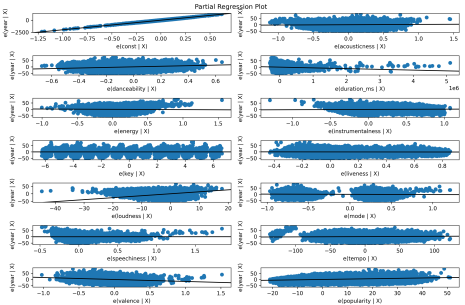


Figure: Partial Regression: Acousticness

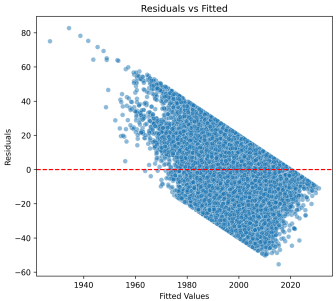
Phase IV: Diagnostics Overview

We rigorously tested the Gauss-Markov assumptions.

Assumption	Test Used	Outcome
Linearity	Partial F-Test (x^2_{dur})	Reject ($F = 304.7$)
Homoscedasticity	Breusch-Pagan	Reject ($LM = 19391$)
Multicollinearity	VIF Score	Pass (Max < 4.0)

Table: Diagnostic Summary

Residual Analysis: Boundedness & Heterogeneity



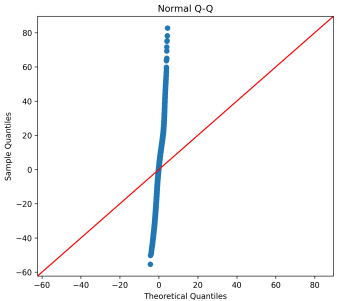


Figure: Residuals vs Fitted (Note the diagonal boundaries)

Boundedness Artifacts:

- $Y \in [1960, 2020]$ creates diagonal ceilings/floors ($e = y - \hat{y}$).
- Prediction space compressed at boundaries.

Heavy Tails (Q-Q Plot):

- Non-normal errors $\neq$ Model failure.
- Indicates **Stylistic Heterogeneity** (Retro/Futuristic outliers).

Remedy: Robustness via CLT ($N = 250k$). Non-normality justifies the **Nostalgia Index**.

Phase VI: The Digital Revolution

We hypothesized a structural break at $T = 1999$ (Napster & ProTools Era).

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \delta D_i + D_i(\mathbf{x}_i^\top \boldsymbol{\gamma}) + \varepsilon_i$$

where $D_i = \mathbb{I}(Year \geq 1999)$.

Explanatory Result

- R^2 **Variance Explained**: $0.296 \rightarrow \mathbf{0.734}$ (Explanatory).
- **Interpretation**: This reflects the model's ability to fit the data once the "Digital Era" is accounted for.
- **Conclusion**: The mechanism of music creation fundamentally changed in 1999.

The "Scissor Effect" (Interaction)

Acousticness Coefficient Flip:

- **Pre-1999**: $\beta \approx -2.8$ (Folk/Rock era).
- **Post-1999**: $\beta + \gamma \approx +0.9$.
- **Meaning**: In the digital era, acoustic elements became a stylistic choice (e.g., "Unplugged") rather than a technological limitation.

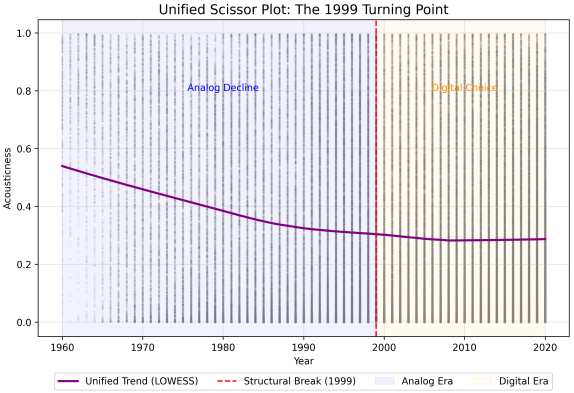


Figure: Unified Scissor Plot: LOWESS Trend

Phase V: Seeking Parsimony

We compared algorithms to select the optimal feature set.

Method	Criterion	Features	Verdict
Full Model	None	13	Baseline
LASSO	L_1 Penalty	12	**Preferred**
Stepwise	AIC	11	Too aggressive

Outcome: We retained complex features like `Instrumentalness` and `Speechiness` as they carry era-specific signal (e.g., Solos vs Rap).

The "Nostalgia Index"

We propose a **Retro Detector** metric:

$$\mathcal{N}_i = \hat{y}_{blind} - y_{actual}$$

- $\mathcal{N} \ll 0$: Song sounds older than it is ("Retro").
- $\mathcal{N} \gg 0$: Song sounds futuristic ("Avant-garde").

Song	**Actual**	**Pred**	**Δ (Yrs)**	**Vibe**
The Weeknd - *Echoes*	2012	1993.1	**-18.9**	90s R&B
Mark Ronson - *Uptown*	2015	2013.1	**-1.9**	Retro-ish
Dua Lipa - *Physical*	2020	2009.0	**-11.0**	80s Synth

Table: Verified Predictions on Test Set

Visualizing Prediction Accuracy

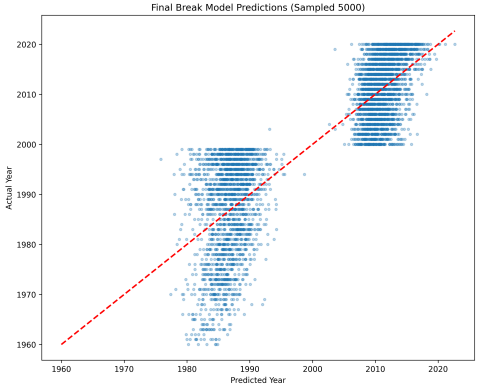


Figure: Actual vs Predicted (Test Set). Note the tighter fit post-1999.

Conclusion

1. **Success**: We can date music/audio with $RMSE \approx 7$ years (Era-Informed).
2. **Discovery**: 1999 was a structural singularity in music history.
3. **Utility**: The Nostalgia Index successfully identifies "Retro" hits.
4. **Curriculum Alignment**: Fully utilized SLR, MLR, Diagnostics, Ridge/Lasso, and Interaction effects.

Thank You

Questions?