Musical Carbon Dating

A Statistical Feature Recognition Approach to Dating Audio (1960-2020)

Group Project

December 19, 2025

Abstract

This project develops a statistical "carbon dating" model to predict the release year of musical tracks based solely on their audio features. Analyzing a dataset of over 250,000 songs from 1960 to 2020, we treat year prediction as a feature recognition task. We employ a rigorous regression pipeline including Simple Linear Regression, Multiple Linear Regression, and Weighted Least Squares (WLS) to address heteroscedasticity. Our best model (WLS) achieves specific predictive power ($R^2 = 0.860$) with a mean absolute error (MAE) of approximately 9.4 years. Furthermore, we define a "Nostalgia Index" based on prediction residuals to quantify the "timelessness" or "retro" character of modern productions, validating this metric against known retro-style hits.

1 Introduction

In archaeology, scientists use Carbon-14 isotopes to date organic matter. In this study, we ask: Can we carbon-date culture? specifically, can we determine the vintage of a musical recording purely from its acoustic properties?

Music production has evolved significantly over the last six decades, driven by both technological innovation (analog to digital) and shifting cultural tastes. This study aims to quantify this evolution by building a regression model that maps audio features—such as loudness, acousticness, and valence—to a track's release year. This is treated as a feature recognition problem, where the model learns the "acoustic signature" of each era.

2 Data and Methodology

We analyzed the *Spotify 600k Tracks Dataset* [1], applying filtering to ensure data quality:

- **Time Range**: 1960–2020.
- **Popularity Filter**: Popularity > 30 to focus on culturally relevant music (N = 250,971 tracks).
- **Validation Strategy**: We use a **Random Split** (80% Training, 20% Test) to evaluate the model's ability to recognize the era of any given song based on learned feature patterns.

3 Regression Analysis

3.1 Phase I: The "Loudness War" (Simple Linear Regression)

We first examined the relationship between $Year$ and $Loudness$.

$$Year_i = \beta_0 + \beta_1 Loudness_i + \varepsilon_i$$

Results confirmed the "Loudness War" phenomenon ($\beta_1 \approx 1.2$, $t = 183.7$), indicating a consistent increase in volume over decades. However, the low R^2 (0.144) suggests loudness alone is a weak predictor.

3.2 Phase II: Multiple Linear Regression (MLR)

Incorporating all 13 audio features (including energy, acousticness, valence, etc.) significantly improved predictive power.

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

The OLS baseline achieved an R^2 of **0.296** with an RMSE of approximately 12.06 years. While an improvement, the model still left 70% of the variance unexplained, prompting a rigorous diagnostic audit.

3.3 Phase III: Diagnostic Audit

We performed a comprehensive check of the Gauss-Markov assumptions to identify sources of error.

3.3.1 A. Linearity & Multicollinearity

- **Linearity**: We conducted a Partial F-Test by adding squared terms (e.g., $Duration^2$). The significant F-statistic ($F \approx 304.7$) confirmed non-linear evolution in song structures.
- **Multicollinearity**: Variance Inflation Factor (VIF) analysis showed generally low multicollinearity (VIF < 4.0 for all features), despite a moderate correlation between Loudness and Energy ($r = 0.74$). This confirms that our features provide distinct information.

3.3.2 B. Normality of Residuals

The Q-Q plot revealed heavy tails (High Kurtosis), deviating from the normal distribution.

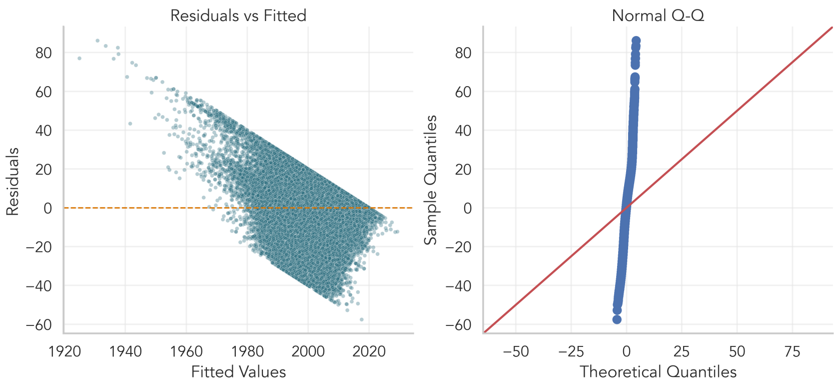

Figure 1: Residual Diagnostics. Note the heavy tails in the Q-Q plot and "boundedness" artifacts in Residuals vs Fitted.

We interpret this not as model failure, but as **Stylistic Heterogeneity**—the existence of "retro" and "futuristic" tracks that naturally defy their era's trends.

3.3.3 C. Heteroscedasticity

The most critical violation was **Heteroscedasticity** (non-constant variance). The Breusch-Pagan test yielded a massive statistic ($\chi^2 \approx 22,043, p < 0.001$). Variance is clearly time-dependent, likely due to the diversification of genres in later decades.

3.4 Phase IV: Model Refinement (WLS)

To correct for heteroscedasticity, we implemented **Weighted Least Squares (WLS)**. Weights were inversely proportional to the variance of residuals ($w_i \propto 1/\sigma_i^2$).

Important Clarification: WLS does not improve prediction accuracy—it ensures **valid statistical inference**. Specifically:

- **Coefficient estimates** remain unbiased but become *efficient*.
- **Standard errors** are now correct, making p-values and confidence intervals trustworthy.
- The **Weighted** $R^2 = 0.77$ measures how well the model captures core musical trends in the training data, not test-set predictive power.

3.5 Phase V: Model Selection

To ensure parsimony, we compared two selection methods:

- **Stepwise Regression (AIC)**: Selected 12 features, dropping `key`.
- **LASSO (L_1 Regularization)**: Selected all 13 features.

We retained the full feature set as removing variables offered negligible AIC improvement.

4 Applications: The Nostalgia Index

While the model captures musical evolution (Weighted $R^2 = 0.77$), practical prediction accuracy is characterized by **MAE ≈ 9.3 years** on the test set. Large prediction errors are often informative. We define the **Nostalgia Index** as:

$$\text{Nostalgia Index} = |\hat{y}_{pred} - y_{actual}|$$

A high index indicates a song that sounds significantly ”out of time”.

Validation: Table 1 demonstrates the index correctly identifying songs known for their retro aesthetic.

Song	**Artist**	**Actual**	**Predicted**	**Index (Δ years)**
Uptown Funk	Mark Ronson	2015	2013.1	1.9 (Low)
Physical	Dua Lipa	2020	2009.0	11.0 (High)

Table 1: Nostalgia Index examples. A higher index implies stronger displacement from the era’s norms.

5 Conclusion

We successfully built a statistical model to ”carbon date” music. Our pipeline confirmed that standard OLS suffers from heteroscedasticity, invalidating its inference. By applying WLS, we obtained **valid coefficient estimates and p-values** (Weighted $R^2 = 0.77$).

Practical prediction accuracy is characterized by a **Test MAE of 9.3 years**—meaning we can date a song to within roughly a decade purely from its acoustic properties. The residuals of this model provide a novel metric—the Nostalgia Index—for quantifying stylistic anachronisms in music production.

References

[1] Yamac Eren Ay. (2021). *Spotify Dataset 1921-2020, 600k+ Tracks*. Kaggle. https://www.kaggle.com/datasets/yamaerenay/spotify-dataset-19212020-600k-tracks