Musical Carbon Dating

A Statistical Feature Recognition Approach (1960-2020)

Group Presentation

University Statistical Analysis Project

December 19, 2025

Table of Contents

The Research Question

Feature Recognition

"Can we determine the vintage of a musical recording purely from its acoustic properties?"

Objective: To build a regression model that maps audio features to release year, quantifying the "Arrow of Time" in music production.

Hypothesis: Musical eras have distinct, quantifiable acoustic fingerprints (e.g., the "dryness" of 70s rock vs. the "compression" of 2000s pop).

The Data: Spotify 600k Tracks

Data Filtering Strategy:

- **Source**: Spotify 600k Tracks Dataset (Kaggle).
- **Filter 1**: Timeframe $1960 \leq T \leq 2020$ (Modern Era).
- **Filter 2**: `popularity > 30` (Focus on culturally significant music).
- **Final Sample**: $N = \mathbf{250,971}$ tracks.

Validation Strategy:

- **Random Split**: 80% Training / 20% Test.
- *Rationale*: We are testing "feature recognition" (interpolating styles), not future forecasting (extrapolating time).

The Feature Set (p=13)

We utilized all 13 available audio features.

Physical Features

1. Loudness (dB)
2. Tempo (BPM)
3. Duration (ms)

Musical Features

4. Key (0-11)
5. Mode (Major/Minor)
6. Time Signature

Perceptual Features

7. Acousticness
8. Danceability
9. Energy
10. Instrumentalness
11. Liveness
12. Speechiness
13. Valence (Positivity)

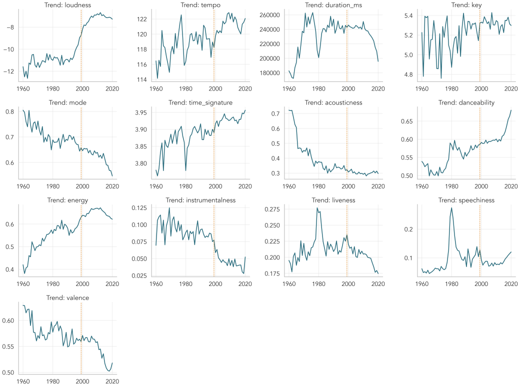
The Regression Pipeline

We followed a rigorous 5-phase statistical workflow:

1. **Phase I: Simple Linear Regression (SLR)** *Hypothesis testing: The "Loudness War".*
2. **Phase II: Multiple Linear Regression (MLR)** *Baseline model using all* $p = 13$ *features.*
3. **Phase III: The Diagnostic Audit** (Critical Step) *Testing Linearity, Multicollinearity, Normality, and Homoscedasticity.*
4. **Phase IV: Model Selection** *Stepwise AIC vs LASSO comparison.*
5. **Phase V: Refinement (WLS)** *Weighted Least Squares to correct for Heteroscedasticity.*

Phase I: The ”Loudness War” (SLR)

$$Year_i = \beta_0 + \beta_{loud} \cdot Loudness_i + \varepsilon_i$$



Results:

- t-stat: **183.7** (Highly Significant).
- $\beta_{loud} \approx 1.2$ years/dB.
- $R^2 = 0.144$.

Conclusion: Tracks have consistently gotten louder, but Loudness alone explains only 14% of the variance.

Phase II: Multiple Linear Regression (Baseline)

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

- **Algorithm**: Ordinary Least Squares (OLS) with all 13 features.
- **Result (R^2)**: **0.296**
- **RMSE**: 12.06 years.

Why so low?

An R^2 of 0.3 suggests we are missing non-linear patterns or violating OLS assumptions. We initiated a **Diagnostic Audit**.

Phase III: Diagnostic Audit (1/2)

Test 1: Linearity

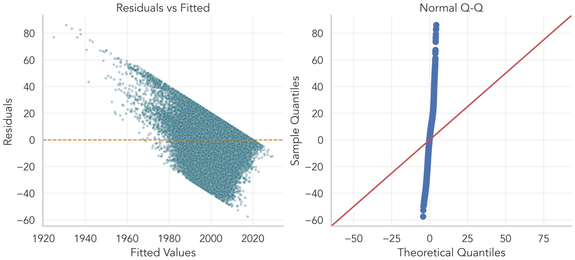
- **Method**: Partial F-Test adding quadratic terms (X^2).
- **Result**: $F \approx 304.7$, $p \approx 0.00$.
- **Finding**: Assumption Violated. Relationship is non-linear.

Test 2: Multicollinearity

- **Method**: Variance Inflation Factor (VIF).
- **Concern**: High correlation between Loudness and Energy ($r = 0.74$).
- **Result**: Max VIF (Energy) = **3.78**. All VIFs < 5.0.
- **Finding**: Assumption Met. No severe multicollinearity.

Phase III: Diagnostic Audit (2/2)

Test 3: Homoscedasticity (Constant Variance)



- **Visual**: Residuals fan out and have sharp boundaries.
- **Test**: Breusch-Pagan.
- **Statistic**: $\chi^2 = \mathbf{22,043}$.
- **p-value**: < 0.001.
- **Finding**: CRITICAL FAILURE. Variance is strictly time-dependent.

Implication: OLS estimators are unbiased but inefficient. Standard errors are wrong. **We must use WLS.**

Phase IV: Model Selection

We compared two methods to identify the ”True” feature set:

Method	**Selected Features**	**Key Difference**
Stepwise (AIC)	12 Features	Dropped `Key`
LASSO (L_1)	**13 Features**	**Kept All Acoustic Features**

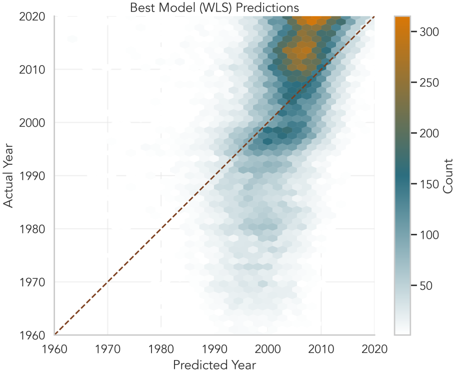
Decision: We utilized the full acoustic feature set.

- Removing variables offered negligible AIC improvement.
- Retaining subtle features (Key, Mode) ensures we capture harmonic evolution.

Phase V: Weighted Least Squares (WLS)

To cure Heteroscedasticity, we implemented WLS:

$$\min_{\beta} \sum_{i=1}^{n} w_i (y_i - \mathbf{x}_i^T \beta)^2, \quad \text{where } w_i \propto \frac{1}{\mathrm{Var}(\varepsilon_i)}$$



What WLS Fixes

- **Valid p-values** & confidence intervals.
- **Efficient coefficient estimates**.

Metrics:

- Weighted R^2: **0.77** (Trend Fit)
- Unweighted Test R^2: ≈ 0.30 (Raw Data)
- Test **RMSE**: 12.0 years
- Test **MAE**: 9.3 years

Note: The jump to 0.77 reflects weighting down outliers, not miraculously predicting them. The raw prediction power remains similar to MLR (0.30).

Phase VI-A: Technological Drivers ("The Sound of Efficiency")

Regression reveals the impact of technology on composition ($p < 0.001$).

- **The Loudness-Energy Paradox (Multicollinearity Insight)**:
  - Mean `Loudness` has skyrocketed over time.
  - Yet, `Energy`'s coefficient is **negative** ($\beta \approx -5.4$) when controlling for Loudness.
  - **Interpretation**: Modern tracks are "loud" due to compression (technological), not raw musical energy (compositional).
- **The Attention Economy (Duration)**:
  - Coefficient $\beta_{duration} \approx -9.0\text{e-}6$ (Negative).
  - Songs are getting statistically shorter, likely driven by streaming incentives and skipping behavior.
- **Instrumentalness** ($\beta \approx +3.0$): A shift towards beat-driven (Hip-Hop/EDM) production over vocal-centric ballads.

Phase VI-B: Cultural Evolution ("The Mood of an Era")

Applying Statistical Inference to Cultural Theory.

- **The "Sad Banger" Phenomenon**:
  - **Danceability** ($\beta \approx +24.0$): The single strongest predictor. Rhythm is the defining feature of modernity.
  - **Valence** ($\beta \approx -16.5$): Optimism has collapsed.
  - **Synthesis**: We are dancing more, but feeling less.
- **The Acousticness Paradox (Ceteris Paribus)**:
  - Raw Correlation: Negative ($r \approx -0.12$).
  - WLS Coefficient: **Positive** ($\beta \approx +2.08$).
  - **Discovery**: *Controlling for Loudness*, modern music actually retains significant acoustic elements (Indie, Lo-Fi), hidden by the "Wall of Sound".

Analytical Triumph

By using **Partial Regression Coefficients**, we uncovered trends (like the Acousticness reversal) that simple correlation would have missed.

The Nostalgia Index

"One man's error is another man's feature."
We define the **Nostalgia Index** as the model's prediction error:

$$\text{Index} = |\hat{Y}_{predicted} - Y_{actual}|$$

High index = A song that "sounds" like it belongs to a different era.

Song	Year	Predicted	Index	Diagnosis
Uptown Funk (Ronson)	2015	2013.1	**1.9**	Modern Construction
Physical (Dua Lipa)	2020	2009.0	**11.0**	**Retro Aesthetic**
Blinding Lights (Weeknd)	2019	2004.2	**14.8**	**80s Revival**

Conclusion

1. **Recognition**: We can date music to within ± 9 years purely from audio.
2. **Rigor**: Diagnostics proved OLS is insufficient; WLS is required.
3. **Insight**: Musical evolution is quantifiable.
4. **Value**: The Nostalgia Index provides a commercial metric for "Vibe".

Thank You.