Initial Project Report - MA5013

Applied Regression Analysis

The Hit Song Formula: A Regression Analysis of Spotify Song Popularity

Saranath P [DA25E003]
Shriprsad S [DA25E054]

Indian Institute of Technology Madras

September 12, 2025

Description of the Dataset

Layman's Description: This project aims to find a statistical "recipe" for hit songs by building a regression model to predict a song's Spotify popularity score based on its intrinsic audio characteristics.

Dataset Details:

- **Response Variable (Y):** `popularity`, a continuous numerical score (0-100) ideal for regression.
- **Predictors (X):** Key predictors include `danceability`, `energy`, `valence`, `acousticness`, `instrumentalness`, `loudness`, `tempo`, and `duration_ms` and 18 more.
- **Observations & Scope:** We will sample 5,000 to 10,000 recent songs from the full database for a manageable and relevant analysis.
- **Data Source Link:** Publicly available on Kaggle: https://www.kaggle.com/datasets/zaheenhamidani/ultimate-spotify-tracks-db

Research Questions

Our investigation will be guided by the following focused research questions:

1. **The "Big Five" of Audio Features:** Which fundamental features (`danceability`, `energy`, `valence`, `acousticness`, `instrumentalness`) have the most statistically and practically meaningful impact on song popularity?
2. **The "Goldilocks Zone" Hypothesis:** Do non-linear "sweet spots" exist for popularity? We will test for quadratic relationships for `tempo` and `duration_ms`.
3. **The "Sad Banger" Phenomenon:** Does the effect of `energy` on popularity depend on `valence`? We will test an interaction to see if high-energy, low-valence songs are disproportionately popular.
4. **The Acoustic Amplification Effect:** Does musical context alter emotional impact? We will test for an interaction between `acousticness` and `valence` to see if valence is a stronger predictor of popularity for highly acoustic tracks.

Possible Methods to be Applied

Our analysis will apply a range of techniques covered in MA 5013:

- **Baseline Model:** A Multiple Linear Regression (MLR) will form the foundation of our analysis.
- **Advanced Model Features:** We will test for non-linearity and context by fitting models with **polynomial (quadratic) terms** and **interaction terms** (e.g., `energy` * `valence`).
- **Multicollinearity Handling:** We will diagnose multicollinearity between predictors like `energy` and `loudness` using **Variance Inflation Factors (VIFs)** and apply **Ridge Regression** if necessary.
- **Model Selection:** To build a parsimonious model, we will use techniques like stepwise selection.
- **Validation and Interpretation:** We will analyze standardized beta coefficients to compare predictor importance and perform comprehensive residual analysis to check model assumptions and identify influential outliers.