Spotify Dataset - Analisi delle tracce musicali e previsione della popolarità

Riccardo Cervero[1]

Sommario
Durante l'ultimo ventennio, grazie allo sviluppo di applicazioni *web* per l'ascolto sempre più efficienti, la fruizione di contenuti musicali ha mostrato un'importante e rapida evoluzione, permettendo all'utente di accedere a qualsiasi brano - o qualunque versione dello stesso - in brevissimo tempo e, nella maggior parte dei casi, gestire autonomamente un archivio di tracce e artisti preferiti. L'industria discografica ha saputo sfruttare tale progresso, non soltanto incrementando il volume di distribuzione e di campagne promozionali, ma anche approfondendo quantitativamente e qualitativamente le tendenze d'ascolto di un pubblico catalogato. Moderne piattaforme offrono, infatti, la possibilità di estrarre ed analizzare una vasta varietà di parametri acustici e non, cosicché il produttore possa dedurre quali sono le caratteristiche più ricercate dal pubblico in un determinato momento e aumentare così la popolarità delle proprie canzoni. È il caso, ad esempio, di *Spotify*. Questo progetto, pertanto, ha come obiettivo l'analisi statistica delle caratteristiche qualitative e quantitative registrate da *Spotify* in un *database* di tracce disponibili all'interno del proprio servizio di *streaming musicale*. Più precisamente, nella seconda e terza sezione verranno esaminate le variabili, testata la loro reciproca connessione - di tipo lineare o non - e calcolate stime intervallari per le rispettive medie o proporzioni fra le modalità. Infine, nelle due successive sezioni, verranno esposti i risultati dei modelli di regressione lineare - sia semplice che multivariata - per la previsione della popolarità del brano e della positività emotiva da esso trasmessa.

Keywords
Inferenza - Regressione lineare

[1] *794126, Dipartimento di Informatica, Sistemistica e Comunicazione*

Indice

1. Caso di studio

I dati esaminati nell'ambito di questo progetto provengono dal database di *Kaggle* "Spotify Dataset"[1], che raccoglie un totale di 19 caratteristiche acustiche e qualitative relative a 169909 tracce, rese disponibili sulla piattaforma *Spotify for Developers*.[2]

Le variabili oggetto di studio[3] corrispondono alle seguenti grandezze:

- *Acousticness* - o "acusticità" , è misura numerica - fra 0 e 1 - della confidenza con cui è possibile definire "acustica" una traccia: se 1, indica massima certezza nell'affermare che il brano sia stato prodotto senza strumenti elettronici
- *Danceability* è il grado di ballabilità calcolato fra 0 e 1, come combinazione di vari elementi musicali, fra cui la stabilità del ritmo e il tempo
- *Duration* è la durata del brano in millisecondi
- *Energy* rappresenta la percentuale di intensità della traccia, sulla base di elementi percettivi quali sonorità e timbro
- *Explicit* è variabile binaria che rileva la presenza o meno di contenuto esplicito

[1] Il link da cui è possibile scaricare il database "Spotify Dataset" è `/kaggle/spotify-dataset-160k-tracks`.

[2] La documentazione della piattaforma *Spotify for Developers* è disponibile al link `developer.spotify.com`.

[3] La descrizione ufficiale delle variabili è disponibile ai link `developer.spotify/audio-features` e `developer.spotify/get-track`.

- *Instrumentalness* misura, fra 0 e 1, l'assenza di contenuto vocale: più il valore si avvicina a 1, maggiore è la confidenza con cui si definisce "strumentale" la traccia
- *Key* registra la stima della chiave complessiva della traccia, codificata in un valore numerico compreso fra 0 (Do) e 11 (Si)
- *Liveness* rileva la presenza di un pubblico udibile nella registrazione, definendone la probabilità: se superiore a 0,8, fornisce una forte evidenza che la traccia sia stata registrata dal vivo
- *Loudness* è il volume complessivo in decibel, fra -60 e circa 4
- *Mode* è la variabile binaria circa la tonalità del brano: 1 se maggiore, 0 se minore
- *Speechiness* indica la presenza di parlato: maggiore la somiglianza tra la traccia e un discorso - come nel caso del genere *rap* -, maggiore la vicinanza del valore a 1
- *Tempo* è il ritmo misurato in battiti al minuto (bpm)
- *Valence* definisce il grado di positività emotiva trasmessa dal brano, tra 0 e 1
- *Year* è l'anno di uscita del brano, dal 1921 al 2020
- *Popularity* è una variabile intera compresa fra 0 e 100, che esprime il grado di popolarità del brano, calcolato a partire dal numero totale di riproduzioni su *Spotify* e da quanto sono recenti tali ascolti: in generale, tracce riprodotte molto frequente durante l'anno corrente avranno un valore di popolarità molto elevato

Tali colonne verranno presentate singolarmente nei diversi paragrafi delle Sezioni 2 e 3. In particolare, la popolarità verrà considerata come principale variabile *target* nell'ambito della definizione di modelli di regressione.

1.1 Pre-processing

Il dataset originale non presentava alcun valore mancante, e le uniche operazioni di *pre-processing* hanno implicato il raggruppamento delle modalità dell'anno di pubblicazione in classi di decennio (Sezione 2) e la rimozione delle variabili poco interessanti per l'obiettivo di analisi statistica: l'identificatore primario della traccia, il titolo, la lista di artisti accreditati e la data di pubblicazione. In particolare, la data di pubblicazione *Release_date* si manifestava in formato YYYY-MM-DD soltanto nel 70% dei casi, mentre la restante parte riportava soltanto l'anno, così come la colonna *Year*. Non potendo risalire al mese e al giorno, si è preferito eliminare *Release_date*.

2. Analisi dei caratteri qualitativi

Variabile *Year*

Le osservazioni del database "Spotify Dataset" sono distribuite con una certa coerenza fra i singoli anni - dal 1921 al 2020: le frequenze relative delle modalità di *Year* sono identiche o comunque molto simili a 0.012 - che corrisponde alla frequenza relativa massima - in 73 anni su 100. Osservazioni meno numerose sono comprensibilmente relative agni anni precedenti al 1947. Perciò, la distribuzione di questa variabile qualitativa ordinale, oltre ad essere multimodale, può anche essere definita estremamente eterogenea: infatti, la mutabilità del fenomeno, misurata dall'indice normalizzato di Gini, mostra un valore di 0.999.
Poichè è noto che le tendenze musicali possono essere meglio definite nell'arco di decadi, piuttosto che di singole stagioni, si è deciso di aggregare le osservazioni per decenni. Le rispettive frequenze delle decadi sono intuibili in Figura 1.

Figura 1. Aerogramma per la visualizzazione delle frequenze relative dei decenni.

Si è poi proceduto a verificare l'esistenza di una connessione fra questa nuova variabile e il grado di popolarità dei brani, eseguendo un test Chi-quadrato, cui ipotesi nulla

$$H_0 : \chi^2 = 0 \tag{1}$$

coincide con l'indipendenza statistica fra i due fenomeni. L'indice di associazione χ^2 è dunque calcolato con $(k_1 - 1)(k_2 - 1) = 990$ gradi di libertà, con $k_1 = 11$ e $k_2 = 100$ i rispettivi conteggi delle modalità. Fissato un livello di significatività $\alpha = 0.05$, è stato ottenuto un valore *pvalue* considerevolmente inferiore ad α[4], portando a rifiutare, con confidenza al 95%, l'ipotesi nulla di indipendenza fra il decennio di uscita della traccia e il proprio grado di popolarità. Ciò significa che le distribuzioni condizionate della popolarità variano in base alla decade considerata, come visibile in Figura 2.

[4] Il *pvalue* in questione è inferiore alla soglia $2.2 \bullet 10^{-16}$.

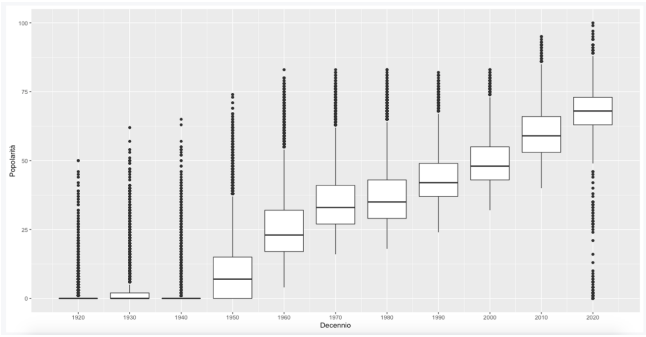

Figura 2. *BoxPlot* condizionati per la visualizzazione delle distribuzioni condizionate di popolarità in base al decennio esaminato.

La natura di questa dipendenza statistica viene approfondita mediante test ANOVA, basato sull'ipotesi nulla di uguaglianza di tutte le popolarità medie, ovvero sull'ipotesi che le osservazioni di popolarità relative a ciascun decennio provengano da popolazioni che seguono una distribuzione normale con varianza e media pari. Si punta pertanto a verificare l'ipotesi alternativa, cioè che la media di almeno un gruppo differisca dalle altre, dimostrando che il fattore condizionante della decade ha un'influenza sulla popolarità. La statistica test è associata a 10 gradi di libertà e, fissato un livello di significatività $\alpha = 0.05$, è stato ottenuto un valore *pvalue* ancora una volta nettamente inferiore ad α, causando il rifiuto dell'ipotesi nulla di uguaglianza delle popolarità medie per decennio e confermando, con una confidenza al 95%, la presenza di una forte relazione fra queste due variabili. Osservando la Figura, è possibile riscontrare, almeno visivamente, una tendenza comprensibile: la preferenza degli utenti a livello aggregato appare quasi perfettamente ordinata per "novità" del brano.

Variabile *Explicit*

Un fenomeno, al contrario, distribuito in maniera poco equilibrata è la presenza di contenuto esplicito nel testo: la frequenza percentuale delle tracce non esplicite è pari al 91.51% (Figura 3) e il suo indice di Gini normalizzato conferma la bassa mutabilità: $\sim$0.311.

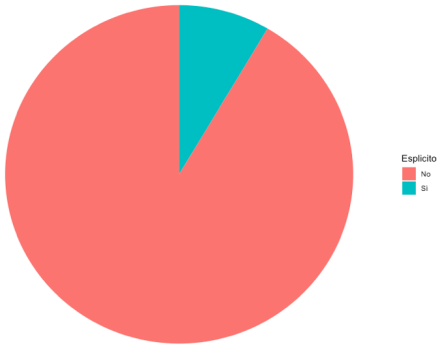

Figura 3. Aerogramma per la visualizzazione delle rispettive frequenze relative di contenuto esplicito e non esplicito.

Nonostante l'elevata numerosità del campione studiato, rimane più utile e affidabile la pratica di stima intervallare del parametro di proporzione p, che deriva dallo studio della distribuzione campionaria della frequenza relativa di una determinata classe nella popolazione. L'intervallo di confidenza ottenuto, avrà pertanto i seguenti estremi:

$$(p - z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}; p + z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}) \qquad (2)$$

In questo caso, fissato un livello di significatività $\alpha = 0.05$, i rispettivi intervalli di confidenza delle modalità "esplicito" e "non esplicito" - con estremi approssimati al terzo decimale - sono:

$$(0.084, 0.086) \qquad (0.914, 0.916) \qquad (3)$$

Anche in questo caso, viene verificata l'esistenza di una connessione fra il grado di popolarità e la presenza di contenuto esplicito: ad un livello di confidenza del 95% e con 99 gradi di libertà, l'ipotesi di indipendenza fra le due variabili viene rifiutata poiché il *pvalue* è calcolato nettamente inferiore alla soglia di significatività.
Inoltre, eseguendo un test ANOVA sull'uguaglianza delle due popolarità medie in corrispondenza di contenuto esplicito e non esplicito, si ottiene un *pvalue* molto basso, rifiutando nuovamente l'ipotesi nulla con un livello di confidenza al 95%.

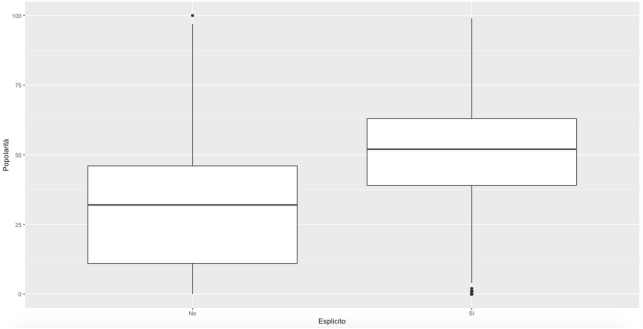

Figura 4. *BoxPlot* condizionati per la visualizzazione delle distribuzioni condizionate di popolarità in base alla presenza o meno di contenuto esplicito.

Variabile *Mode*

La distribuzione delle due tonalità si presenta sbilanciata: la frequenza percentuale della modalità "maggiore" è circa il 70.9%, contro il 29.1% della "minore" (Figura 5).

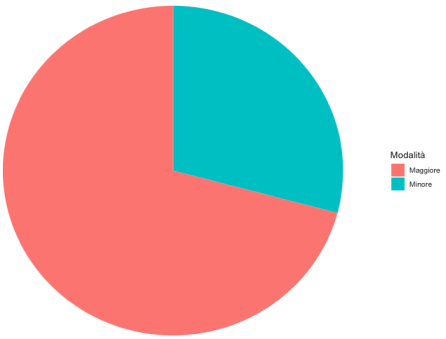

Figura 5. Aerogramma per la visualizzazione delle frequenze relative delle due tonalità.

L'indice di Gini normalizzato viene calcolato pari a 0.83. Si è proceduto dunque ad estrarre un'informazione più accurata e affidabile tramite stima intervallare delle proporzioni di tracce in tonalità minore e maggiore nella popolazione: ad un livello di confidenza del 95%, le rispettive probabilità sono contenute nei seguenti intervalli di confidenza - con estremi approssimati al terzo decimale:

$$(0.289, 0.294) \qquad (0.706, 0.711) \tag{4}$$

Viene rifiutata, ad un livello di confidenza del 95%, l'ipotesi che l'indice di connessione χ^2 per l'individuazione di una dipendenza statistica fra la tonalità del brano e la sua popolarità, calcolato con 99 gradi di libertà, sia nullo: il *pvalue* ha un valore estremamente basso. Per lo stesso motivo, si rifiuta, tramite test ANOVA con $\alpha = 0.05$, l'uguaglianza delle popolarità medie delle tracce condizionate alle due diverse tonalità.

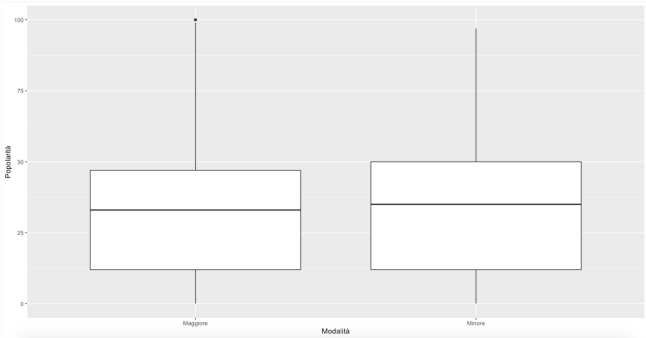

Figura 6. *BoxPlot* condizionati per la visualizzazione delle distribuzioni condizionate di popolarità in base alla tonalità.

Variabile *Key*

L'ultimo carattere qualitativo, riferito alla chiave globale della traccia, mostra un'elevatissima eterogeneità: l'indice normalizzato di Gini è infatti pari a 0.991. È possibile, fra le 11 modalità, individuarne la moda: la chiave di "Do", con una frequenza percentuale del 12.65%. Tuttavia, per poter affermare con più affidabilità che la chiave più frequente fra i brani musicali sia effettivamente "Do", è necessario confrontare le rispettive stime intervallari delle frequenze relative di tutte le modalità di *Key*. Il risultato di ogni stima è riassunto in Figura 7. Poiché gli estremi superiori degli intervalli di confidenza delle altre chiavi sono sempre minori rispetto al limite inferiore dell'intervallo di confidenza di "Do", è possibile affermare con un livello di confidenza del 95% che essa sia, in effetti, la chiave più usata per comporre una traccia. Allo stesso modo, "Re diesis" può essere considerata come la chiave meno utilizzata.

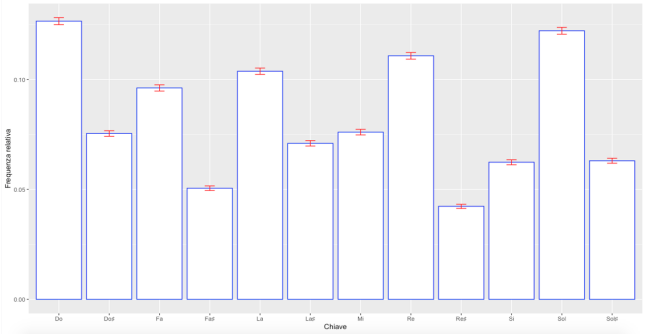

Figura 7. *BarPlot* delle frequenze relative di ciascuna chiave nella popolazione. I segmenti rossi verticali indicano gli intervalli di confidenza calcolati con $(1-\alpha) = 0.95$. Dall'osservazione del grafico, è possibile dedurre che "Do" sia la chiave più frequente.

Viene poi rifiutata, con un livello di significatività $\alpha = 0.05$, l'ipotesi di indipendenza statistica fra la popolarità e la chiave. Infine, eseguendo un test ANOVA con livello di confidenza del 95%, si rifiuta anche l'ipotesi di uguaglianza fra le popolarità medie condizionate alla chiave globale con cui è stata composta la traccia. Ciò significa che la scelta della chiave, così come la selezione di una determinata tonalità e l'introduzione di contenuto esplicito nel testo, ha un'influenza sul livello di popolarità che raggiungerà il brano.
In Figura 8, è possibile osservare le distribuzioni della popolarità condizionata alla chiave.

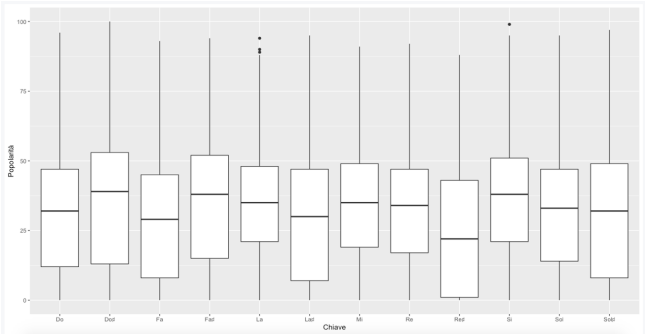

Figura 8. *BoxPlot* condizionati per la visualizzazione delle distribuzioni condizionate di popolarità in base alla chiave globale.

3. Analisi dei caratteri quantitativi

Tutte le rimanenti 11 variabili quantitative sono state esaminate in varie fasi:

- calcolo delle statistiche descrittive di base, tra cui i principali momenti e quantili
- analisi della forma della distribuzione e confronto con una Normale
- stima intervallare della media della variabile - sempre con un livello di significatività $\alpha = 0.05$ -, poiché, come già esposto per quanto riguarda la proporzione campionaria delle modalità qualitative, la sola stima puntuale derivata dal campione è dominata da una certa incertezza e mostra scarsa utilità
- analisi della significatività delle reciproche correlazioni lineari

Per evitare una descrizione prolissa del lavoro svolto, in questa Sezione verranno esposti soltanto i risultati riscontrati sulle variabili più interessanti.

Variabile *Acousticness*

Le statistiche descrittive sul livello di "acusticità" della traccia sono riassunte nella Tabella 1:

Minimo	0
1o Quartile	0.0945
Mediana	0.492
Media	0.493
3o Quartile	0.888
Massimo	0.996
Moda	0.995
Coeff. di Variazione	0.764

Tabella 1. Statistiche descrittive calcolate sul livello di acusticità.

È possibile dedurre che la distribuzione sia asimmetrica. Tuttavia, la differenza quasi trascurabile fra media e mediana è indizio di uno scarso grado di deviazione dalla condizione di simmetria. Queste deduzioni sono confermate dal valore - positivo e basso - dell'indice di asimmetria: $\frac{m_3}{m_2^{3/2}} = 0.009$.
La distribuzione di *Acousticness* è, inoltre, platicurtica, perchè la propria curtosi è inferiore al valore dell'indice in condizione di normalità, ovvero 3: $\frac{m_4}{m_2^2} = 1.386$. Questa caratteristica implica generalmente un inspessimento delle code ed un appiattimento della campana gaussiana. Tali peculiarità vengono ritrovate visualizzando la distribuzione dell'acusticità (Figura 9).

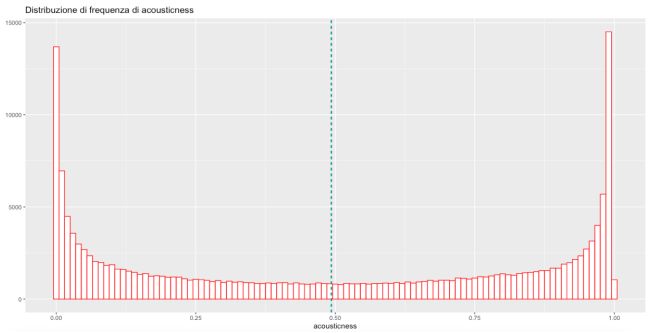

Figura 9. Istogramma dell'acusticità. Le linee verticali tratteggiate corrispondono alla media (in blue) e alla mediana (in verde).

La forte deviazione dalla condizione Normale è suggerita anche dal grafico *Quantile-Quantile* (Figura 10), specialmente per quanto riguarda le code: quanto minore la somiglianza fra l'andamento dei punti corrispondenti ai quantili empirici e la retta associata alla distribuzione dei quantili teorici normali, tanto maggiore la differenza fra la variabile esaminata e la Gaussiana.

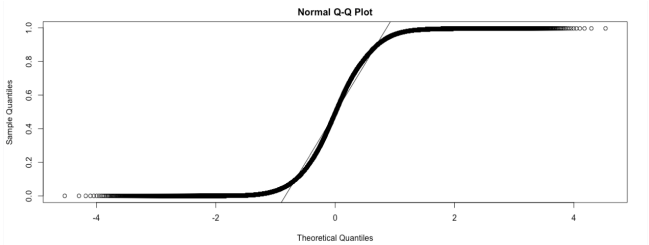

Figura 10. Quantile-Quantile Plot per il confronto fra i quantili empirici e i quantili teorici.

Infine, con un livello di confidenza al 95%, è possibile affermare che l'acusticità media sia contenuta nell'intervallo:

$$(0.491, 0.495) \tag{5}$$

Ciò significa che, generalizzando con confidenza al 95%, le tracce musicali hanno in media una probabilità di essere completamente acustiche inferiore rispetto alla probabilità di includere strumenti elettronici, poichè l'intervallo di confidenza di "acousticness" ha estremo superiore minore di 0.5.

Variabile *Danceability*

Le statistiche descrittive sul livello di ballabilità della traccia sono riassunte nella Tabella 2:

Minimo	0
1o Quartile	0.417
Mediana	0.548
Media	0.538
3o Quartile	0.667
Massimo	0.988
Moda	0.565
Coeff. di Variazione	0.326

Tabella 2. Statistiche descrittive calcolate sul livello di ballabilità.

La ballabilità dei brani rivela una forma asimmetrica negativa, con un valore dell'indice di asimmetria pari a -0.213. Tale risultato indica che le frequenze più elevate della distribuzione tendono a disporsi su valori elevati di ballabilità, com'è ragionevole supporre, dato che le tracce musicali sono in gran parte concepite per essere accompagnate da qualche forma di ballo. L'allontanamento dalla normalità è molto lieve, di tipo platicurtico (come riscontrato in Figura 11): la curtosi è, infatti, pari a 2.575.

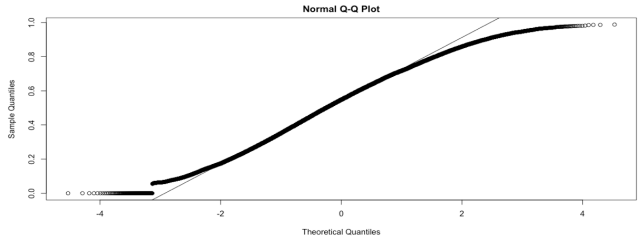

Figura 11. Quantile-Quantile Plot per il confronto fra i quantili empirici e i quantili teorici.

In Figura 12 è visibile la forma della distribuzione della ballabilità.

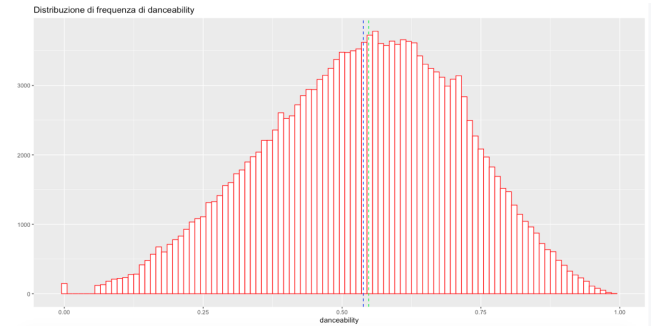

Figura 12. Istogramma della ballabilità. Le linee verticali tratteggiate corrispondono alla media (in blue) e alla mediana (in verde).

La ballabilità media generale è stimata, con un livello di confidenza del 95%, nel seguente intervallo:

$$(0.537, 0.539) \qquad (6) \qquad .$$

Si conclude che l'industria discografica preferisce produrre, in media, basi musicali che possano essere più facilmente accompagnate da una danza, ovvero con un livello di *danceability* ben oltre lo 0.5.

Variabile *Instrumentalness*

Il grado di confidenza con cui si definisce "strumentale" la traccia è, innanzitutto, descritto dalle grandezze in Tabella 3:

Minimo	0
1o Quartile	0
Mediana	0
Media	0.162
3o Quartile	0.087
Massimo	1
Moda	0
Coeff. di Variazione	1.91

Tabella 3. Statistiche descrittive calcolate sul livello di strumentalità.

Esso presenta, pertanto, una variabilità nettamente superiore rispetto alle variabili precedentemente esaminate, ed è caratterizzato da una forte asimmetria positiva - misurata pari a 1.682. La forma della distribuzione, descritta da un indice di curtosi superiore a 3 (4.119), è caratterizzata da un allontanamento dalla normalità di tipo leptocurtico: le code tendono ad assottigliarsi, spingengo verso l'alto le frequenze più elevate. Infatti, in Figura 13, soltanto la coda di destra - per la forte asimmetria positiva - si presenta molto lunga ed estremamente sottile. Questa condizione viene riscontrata anche osservando il grafico *Quantile-Quantile* in Figura 14.

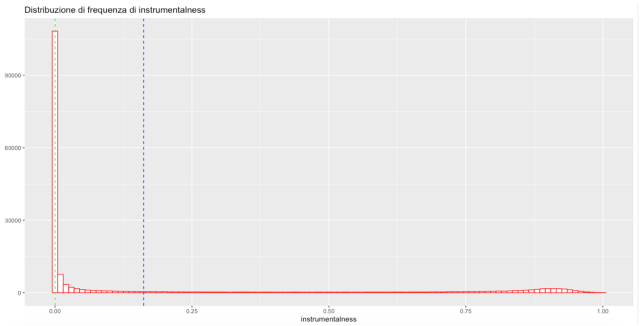

Figura 13. Istogramma della *instrumentalness*. Le linee verticali tratteggiate corrispondono alla media (in blue) e alla mediana (in verde).

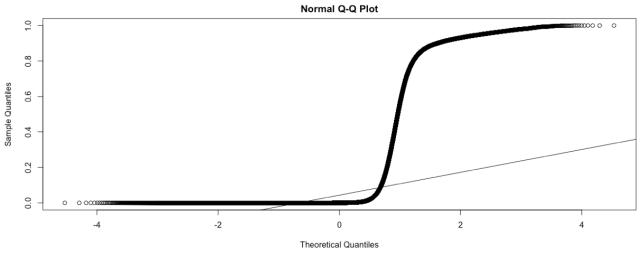

Figura 14. Quantile-Quantile Plot per il confronto fra i quantili empirici e i quantili teorici.

Infine, la stima della quantità media di contenuto esclusivamente strumentale all'interno delle tracce, con un livello di confidenza del 95%, è compreso nel seguente intervallo di confidenza:

$$(0.16, 0.163) \tag{7}$$

È quindi ragionevole concludere - con confidenza al 95% - che, in media, si tenda a produrre brani che contengono tra il 16 e il 16.3% di parti strumentali.

Variabile *Speechiness*

Poichè nel paragrafo precedente si è dimostrato come la voce formi elemento tendenzialmente essenziale nella produzione di una traccia, è utile approfondire la variabile che distingue gli elementi vocali in "cantato" e "parlato", con un grado di *speechiness* compreso fra 0 e 1. Le statistiche descrittive di base sono mostrate nella Tabella 4:

Minimo	0
1o Quartile	0.0349
Mediana	0.045
Media	0.094
3o Quartile	0.075
Massimo	0.969
Moda	0.0347
Coeff. di Variazione	1.594

Tabella 4. Statistiche descrittive calcolate sul livello di *speechiness*.

La variabilità tende ad essere ancora una volta elevata, e la distribuzione presenta un'asimmetria positiva di gran lunga superiore alle variabili precedenti: 4.236. L'elevatissimo valore di curtosi (22.375) indica che la forma della distribuzione è di tipo leptocurtico, come rilevato riguardo alla "strumentalità", ma con un ulteriore assottigliamento della coda destra (Figura 15 e 16).

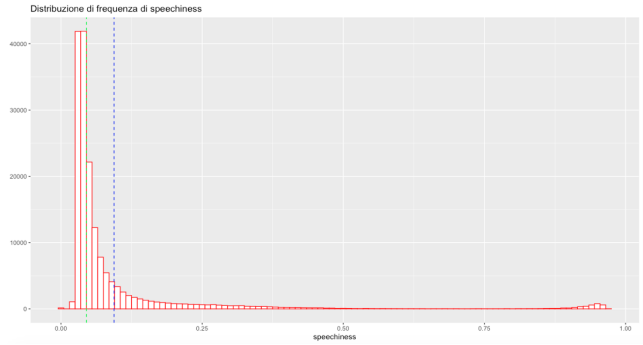

Figura 15. Istogramma della *speechiness*. Le linee verticali tratteggiate corrispondono alla media (in blue) e alla mediana (in verde).

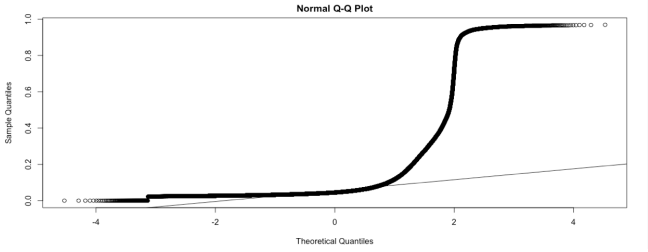

Figura 16. Quantile-Quantile Plot per il confronto fra i quantili empirici e i quantili teorici.

Queste caratteristiche sono, in altre parole, manifestazione di una generale tendenza a preferire una base vocale cantata, piuttosto che simile ad un discorso, nonostante la grande diffusione di generi musicali come il *rap*. Infatti, la probabilità media di rilevare del parlato all'interno delle canzoni è, comprensibilmente, molto bassa, stimata - con un livello di confidenza del 95% - all'interno del seguente intervallo:

$$(0.093, 0.095) \tag{8}$$

ovvero tra il 9.3% e il 9.5%.

Variabile *Liveness*

Un altro carattere quantitativo interessante consiste nella probabilità che la traccia sia stata registrata durante un'esibizione dal vivo. Le statistiche descrittive sono riportate in Tabella 5:

Minimo	0
1o Quartile	0.0984
Mediana	0.135
Media	0.207
3o Quartile	0.263
Massimo	1
Moda	0.111
Coeff. di Variazione	0.855

Tabella 5. Statistiche descrittive calcolate sul livello di *liveness*.

La distribuzione è positivamente asimmetrica, con un indice pari a 2.146, e leptocurtica ($\frac{m_4}{m_2^2} = 7.916$). La sua variabilità è moderata. In Figura 17, è possibile notare la sottile coda di destra, specialmente per valori maggiori di 0.5, e l'elevata frequenza delle probabilità più ridotte, in un intervallo compreso fra 0 e, circa, 0.375. L'allontanamento dalla condizione di normalità è perfettamente descritto nel grafico *Quantile-Quantile* (Figura 18): anche qui, come nei due casi precedenti, i quantili empirici relativi alla coda destra sono estremamente distanti da quelli teorici.

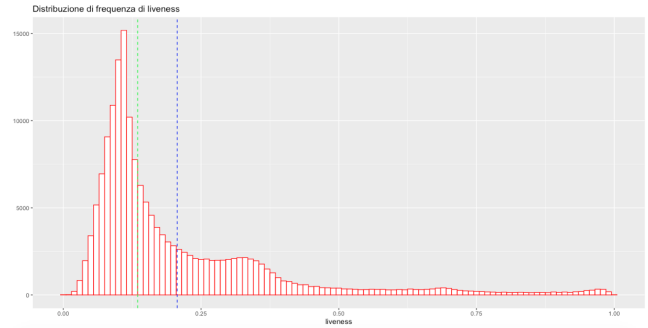

Figura 17. Istogramma della *liveness*. Le linee verticali tratteggiate corrispondono alla media (in blue) e alla mediana (in verde).

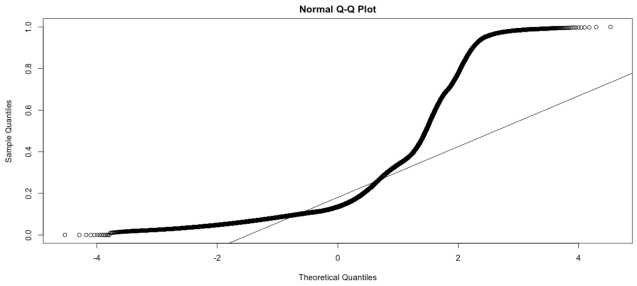

Figura 18. Quantile-Quantile Plot per il confronto fra i quantili empirici e i quantili teorici.

Infine, è possibile affermare, con confidenza al 95%, che la probabilità media che la traccia sia stata registrata dal vivo sia compresa nell'intervallo

$$(0.206, 0.208) \tag{9}$$

Variabile *Valence*

La variabile che registra la positività percepita durante l'ascolto del brano è caratterizzata dalle seguenti statistiche di base (Tabella 6):

Minimo	0
1o Quartile	0.322
Mediana	0.544
Media	0.532
3o Quartile	0.749
Massimo	1
Moda	0.961
Coeff. di Variazione	0.493

Tabella 6. Statistiche descrittive calcolate sul livello di positività.

La distribuzione presenta una leggera asimmetria negativa (-0.124) e una forma platicurtica - la curtosi è 1.949 -, come osservato in Figura 19.

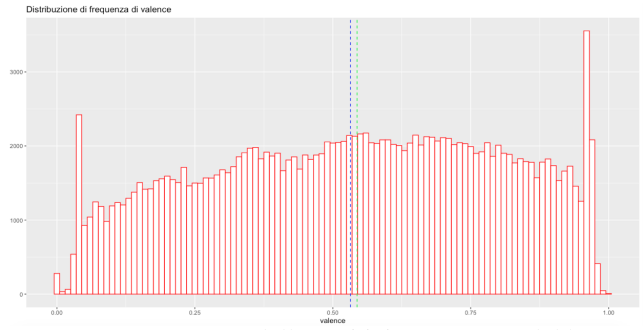

Figura 19. Istogramma della positività trasmessa dal brano. Le linee verticali tratteggiate corrispondono alla media (in blue) e alla mediana (in verde).

Le code spesse deformano la tipica forma a campana verso una densità visibilmente più piatta. Questa anomalia è riscontrabile in Figura 20.

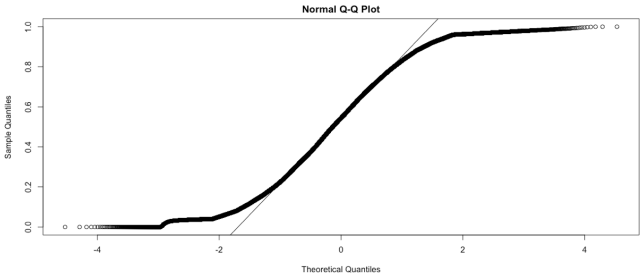

Figura 20. Quantile-Quantile Plot per il confronto fra i quantili empirici e i quantili teorici.

Il livello emotivo, pertanto, si distribuisce in maniera bilanciata: la produzione musicale, nelle proprie numerose declinazioni di genere, è in grado di esprime qualsiasi tipo di stato d'animo, dal più negativo - quando *valence* è nulla - al più positivo - quando la colonna assume valore 1, senza essere dominata da nessuna delle due polarità. Tuttavia, con confidenza al 95%, è possibile affermare che l'industria discografica propenda leggermente verso un sentimento più positivo, poiché il livello emotivo medio è stimato all'interno

dell'intervallo

$$(0.531, 0.533) \tag{10}$$

Si è proceduto, poi, a verificare l'esistenza di qualche tipo di connessione statistica fra il livello emotivo e altre variabili qualitative. Innanzitutto, si è deciso di studiare il legame fra il decennio di pubblicazione e la positività trasmessa, per estrarre evidenze sul fatto che i contenuti musicali risentano del periodo storico non solo da un punto di vista stilistico, ma anche da un punto di vista dei sentimenti espressi. Pertanto, è stato eseguito un test del Chi-quadrato sulla relazione fra *Valence* e *Decade*, con 1 grado di libertà, supponendo una reciproca indipendenza fra le varie decadi. Fissato un livello di significatività del 5%, s'ottiene un *pvalue* inferiore alla soglia, rifiutando l'ipotesi di indipendenza. Viene anche rifiutata l'ipotesi di uguaglianza dei livelli emotivi medi per ogni decennio, tramite test ANOVA con $\alpha = 0.05$ e 10 gradi di libertà, confermando ulteriormente l'esistenza di una relazione fra i due fenomeni. In Figura 21, è possibile distinguere le differenti distribuzioni della positività per gruppo.

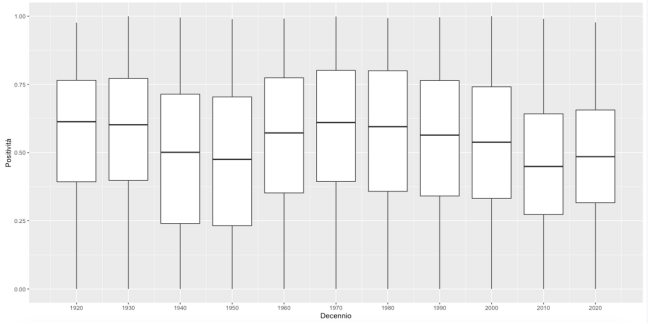

Figura 21. *BoxPlot* condizionati per la visualizzazione delle distribuzioni condizionate di positività in base al decennio esaminato.

Nella stessa maniera, vengono esaminate le dipendenze della positività rispetto alla tonalità e alla chiave del brano, ipotizzando che gli elementi tecnici di base possano influenzare il sentimento della traccia finale. In entrambi i casi, l'ipotesi di indipendenza viene rifiutata tramite calcolo dell'indice di connessione χ^2 con livello di significatività 0.05, e così anche l'ipotesi di uguaglianza delle medie di *Valence* rispetto a entrambi i fattori condizionanti - test ANOVA -, sempre con confidenza al 95%.

Variabile *Popularity*

La popolarità è associata alle seguenti statistiche (Tabella 6):

Minimo	0
1o Quartile	12
Mediana	33
Media	31.56
3o Quartile	48
Massimo	100
Moda	0
Coeff. di Variazione	0.684

Tabella 7. Statistiche descrittive calcolate sul livello di popolarità.

La moda viene individuata in corrispondenza di 0: il grado di popolarità su *Spotify* più diffuso all'interno del campione esaminato è quello nullo, coprendo addirittura fino al 16.1% delle osservazioni totali. La variabilità è bassa e la distribuzione rivela un discreto indice di asimmetria negativo, pari a -0.021, e un forma moderatamente platicurtica (Figura 23), con curtosi uguale a 1.985. Osservando la Figura 22, è possibile notare, senza considerare il valore di moda, come le misure più frequenti siano concentrate nell'intervallo 25-50.

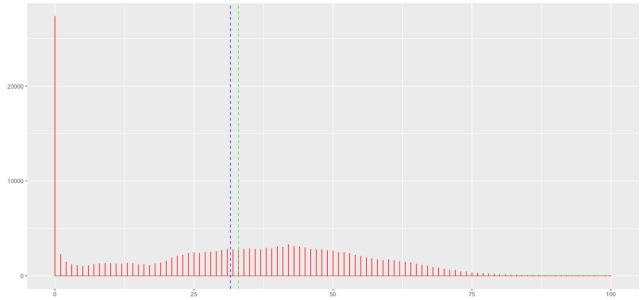

Figura 22. Istogramma della popolarità. Le linee verticali tratteggiate corrispondono alla media (in blue) e alla mediana (in verde).

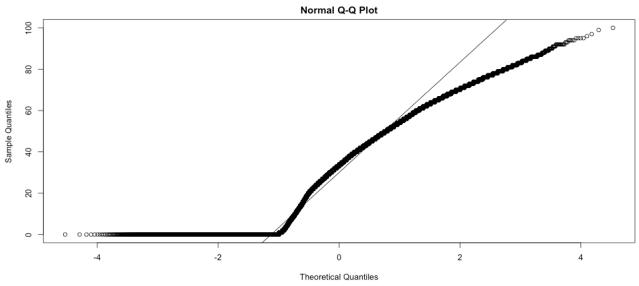

Figura 23. Quantile-Quantile Plot per il confronto fra i quantili empirici e i quantili teorici.

3.1 Correlazioni lineari

Successivamente, sono stati stimati i coefficienti di correlazione di Pearson per tutte le possibili coppie di variabili quantitative, effettuando il corrispettivo test con livello di confidenza fissato al 95%, per valutare la significatività statistica dei valori ρ. Il grado di reciproca dipendenza lineare fra i vari caratteri è visibile in Figura 24.

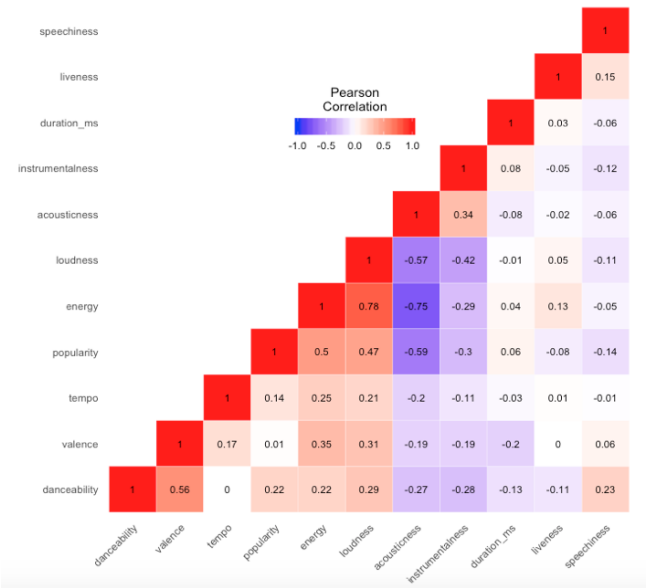

Figura 24. *Heatmap* per la visualizzazione dei coefficienti di correlazione lineare di Pearson fra tutte le variabili quantitative.

L'unica correlazione lineare statisticamente non significativa ad un livello di confidenza del 95% viene riscontrata fra *liveness* e positività: il *pvalue* è 0.86.
I valori ρ più rilevanti possono essere riscontrati fra le seguenti coppie di fenomeni:

- energia e volume: $\rho = 0.78$
- energia e "acusticità": $\rho = -0.75$
- popolarità e "acusticità: $\rho = -0.59$, per cui i brani con strumenti elettronici attrarrebbero maggiori ascolti
- volume e "acusticità": $\rho = -0.57$, per cui all'assenza di strumenti elettronici si associa un minor livello di decibel
- ballabilità e *valence*: $\rho = 0.56$, per cui maggiore la positività percepita dall'ascolto di un brano, maggiore la facilità ad associargli una danza, e viceversa.

In fase di costruzione di un modello di regressione lineare, sarà necessario valutare queste collinearità fra le colonne della matrice del disegno, poiché potrebbero penalizzare la corretta stima dei coefficienti di regressione.

4. Regressione lineare semplice

Appurata l'esistenza di legami lineari statisticamente significativi fra il grado di popolarità di una traccia - considerato in questa fase come variabile dipendente - e molte delle altre colonne disponbili - variabili indipendenti -, si pone un vincolo di linearità sulla forma funzionale del modello che possa descrivere tali relazioni, ricercando, dunque, la retta ottimale in termini di Errore Quadratico Medio (MSE).

In particolare, il modello di regressione lineare semplice è specificato dalla seguente relazione:

$$y_i = f(x_i;\beta) + \in_i \tag{11}$$

dove $\in_i$ è l'errore casuale e $f(x_i;\beta)$ è la funzione che combina i valori delle variabili indipendenti e i rispettivi coefficienti di regressione per determinare l'osservazione y_i. In questa Sezione, verranno stimati modelli univariati caratterizzati da una funzione $f(x_i;\beta)$ di primo grado.
È opportuno specificare che l'eccessiva differenza di scala fra alcune variabili ha reso necessaria una mappatura di tutti i caratteri quantitativi fra 0 e 1, in quanto, altrimenti, le stime dei coefficienti di regressione non sarebbero state confrontabili, dunque inutili per l'obiettivo del progetto. Questa operazione ha implicato l'uso di una trasformazione in grado di mantenere inalterata la forma della distribuzione e gestire valori negativi:

$$X_i' = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}} \tag{12}$$

con $X_{\min}$ e $X_{\max}$ rispettivamente il massimo e il minimo della colonna X. Si è inoltre deciso di convertire le variabili qualitative in tipo `factor`, in modo che ogni loro modalità potesse essere considerata come una colonna *dummy*, e potesse essere quindi possibile estrarre l'influenza di ciascuna classe al netto delle altre.
La significatività statistica dei coefficienti di regressione di ciascun modello è stata verificata mediante test t con un livello di significatività $\alpha = 0.05$, mentre la bontà di adattamento è stata valutata mediante coefficiente di determinazione R^2. I risultati di ciascuna prova sono riassunti come segue.

Popolarità ~ Positività

Data P, variabile del grado di popolarità, e V, positività trasmessa, il modello stimato è

$$P = 0.311484 + 0.007671 \bullet V \tag{13}$$

Il test t rifiuta l'ipotesi nulla per quanto riguarda sia l'intercetta che il coefficiente angolare, producendo un *pvalue* estremamente basso nel primo caso e moderatamente basso nel secondo. Il coefficiente angolare β_1 mostra che, ad un aumento di un'unità percentuale del livello di positività trasmessa, la popolarità cresce, al netto dell'intercetta, di circa uno 0.008%. Nonostante la significatività statistica del coefficiente di regressione, la capacità di tale modello di descrivere la variabilità della popolarità rimane molto scarsa: $R^2 < 0.001$.

Popolarità ~ Chiave

Data P, variabile del grado di popolarità, il modello ottenuto include i seguenti parametri β - con stime approssimate al secondo decimale per esigenze di leggibilità:

	Parametro	PValue
Intercetta	0.31	0
Do#	0.04	0
Re	0.01	0
Re#	-0.06	0
Mi	0.02	0
Fa	-0.02	0
Fa#	0.04	0
Sol	0.005	0.02
Sol#	-0.002	0.37
La	0.02	0
La#	-0.016	0
Si	0.04	0

La stima del coefficiente di regressione della chiave di "Sol" è positiva ma associata ad una bassa significatività statistica: il *pvalue* è di poco inferiore ad α. Il coefficiente di regressione della chiave di "Sol#" è, invece, non significativo: non vi è evidenza di alcuna relazione lineare fra l'uso di tale chiave in fase di composizione del brano e il livello di popolarità. Tutte le altre stime sono altamente significative - il *pvalue* tende ad essere nullo -, ma è possibile operare una distinzione. Infatti, le chiavi di "Do#", "Re", "Mi", "Fa#", "La" e "Si" hanno un'influenza positiva sulla popolarità: il loro uso permette di incrementare il grado della variabile dipendente, da una variazione minima dell'1% (con "Re") ad un incremento massimo del 4% ("Do#" e "Fa#"). Al contrario, le chiavi di "Re#", "Fa" e "La#" si rivelano fattori penalizzanti, con un decremento di popolarità che va dall'1% al 6%. In generale, la bontà di adattamento del modello è ancora scarsa: $R^2 = 0.013$.

Popolarità ∼ Decennio

I coefficienti di regressione associati a ciascun decennio sono, ad un livello di confidenza del 95%, tutti altamente significativi, tranne il parametro relativo agli anni '40. È possibile confermare quanto ipotizzato nella Sezione 2: l'influenza esercitata dalla modalità di decennio sulla popolarità è sempre positiva - i β sono tutti maggiori di 0 -, ma la dimensione dell'incremento - ovvero il valore di tali parametri - cresce in ordine cronologico (Figura 25).

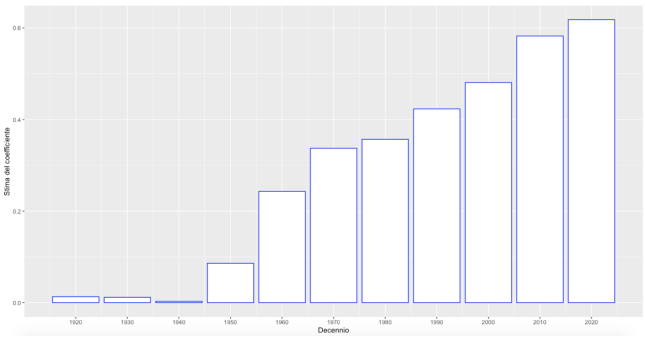

Figura 25. Le stime dei coefficienti di regressione per decennio seguono un ordine cronologico crescente.

Popolarità ∼ Tonalità

Data P, popolarità del brano, il modello stimato è:

$$P = 0.3266221 - 0.0156036 \bullet \text{Tonalità maggiore} \quad (14)$$

Perciò, la scelta di una tonalità maggiore penalizza il valore della variabile dipendente, abbassando di circa l'1.6% il grado di popolarità del brano.

5. Regressione lineare multipla

I modelli multivariati includono, invece, due o più variabili indipendenti per prevedere il valore della Y. I coefficienti di regressione rappresentano, quindi, le stime delle variazioni attese della variabile dipendente associate ad una variazione unitaria della singola x_j, mantenendo fisso tutto il resto del modello, quindi al netto degli altri predittori.
I successivi paragrafi esporranno i risultati ottenuti dalla costruzione di modelli lineari multivariati per la previsione del grado di popolarità del brano e del livello di positività da esso trasmesso. La significatività statistica della relazione fra le variabili dipendenti e l'insieme dei predittori è verificata mediante test F, cui ipotesi nulla nega l'esistenza di tale dipendenza lineare e ipotesi alternativa conferma il legame con almeno un regressore. La bontà di adattamento è stata valutata mediante il coefficiente di determinazione multiplo R^2 e la sua versione corretta, per evitare di sovrastimare la capacità del modello.

5.1 Previsione di *Popularity*

Innanzitutto, viene regredita la variabile *popularity* sull'insieme totale di caratteri quantitativi. I coefficienti di regressione vengono stimati, con livello di confidenza al 95%, tutti altamente significativi, eccetto il parametro associato alla durata del brano, che non appare pertanto utile per prevedere il grado di diffusione di un brano. Gli unici coefficienti con segno negativo sono relativi ad acusticità (-0.23), presenza di parti esclusivamente strumentali (-0.07), *liveness* (-0.08), presenza di parlato *speechiness* (-0.25) e positività (-0.23). Pertanto, sono da ritenersi più popolari le tracce che contengono strumenti elettronici - hanno dunque bassa *acousticness* -, prive di parti strumentali, registrate in studio, prive di "parlato" e con un sentimento tendente verso una polarità negativa. Al contrario, i fenomeni con il maggior impatto positivo sulla Y sono la ballabilità (0.27), il volume complessivo (0.21) e l'energia (0.11). Eseguendo un test F con $\alpha = 0.05$, viene rifiutata l'ipotesi di incorrelazione fra la popolarità e il set di caratteri quantitativi. Il coefficiente di determinazione multiplo è abbastanza buono (0.46), ed è pari alla propria versione corretta, indice del fatto che la capacità del modello non viene sopravvalutata.
Come visto nella Sezione 3.1, molte delle variabili incluse nella matrice del disegno sono risultate fortemente correlate. È dunque necessario verificare che il modello appena stimato non sia affetto da problemi di multicollinearità, che possano penalizzarne la correttezza. Questa operazione è possibile grazie all'indice *Variance Inflation Factor (VIF)*, il quale fornisce,

per ciascun regressore, l'incremento subito dalla varianza della stima di β rispetto alla condizione di perfetta incorrelazione fra le variabili indipendenti. Considerando una soglia pari a 4, la variabile "energia" risulta affetta da un moderato problema di collinearità. Risultava, infatti, essere il fenomeno associato ai più elevati valori di correlazione. L'energia viene pertanto rimossa dal set di candidati esplicativi.
Dopo la rimozione, il coefficiente di regressione associato al volume raggiunge un valore quasi doppio rispetto alla stima precedente (0.38), confermandosi come fattore quantitativo di maggior influenza sulla popolarità della traccia. L'R^2, invece, rimane quasi invariato.
Tale modello è stato infine integrato con le variabili qualitative della chiave, tonalità, presenza di contenuto esplicito e decennio. Le modalità di quest'ultima variabile sono associate a coefficienti molto elevati e tutti altamente significativi. Soltanto due delle classi della chiave, invece, vengono ritenute significativamente correlate alla variabile dipendente. I caratteri quantitativi, esclusa l'energia a priori, sono altamente significativi, ma le stime sono nettamente inferiori rispetto a quelle relative ai decenni. L'introduzione di *Decade* ha pertanto ridimensionato la relativa importanza di tutti gli altri regressori, elevando, però, il coefficiente di determinazione multiplo: $R^2 = 0.79$, pari alla propria versione corretta. Il test F rifiuta l'ipotesi nulla con un livello di confidenza del 95%, ribadendo l'evidenza di un forte legame lineare fra *popularity* e il set di predittori. Infine, il calcolo del *VIF* verifica l'assenza di problemi di multicollinearità nella matrice del disegno.
Tuttavia, in Figura **??** e 27 è possibile individuare due caratteristiche negative dei residui di regressione ottenuti dall'ultimo modello stimato: l'eteroschedasticità - dato che, come si osserva, le previsioni sui valori più elevati sono associate a residui più variabili - e la distribuzione dei quantili abbastanza differente rispetto alla Normale.

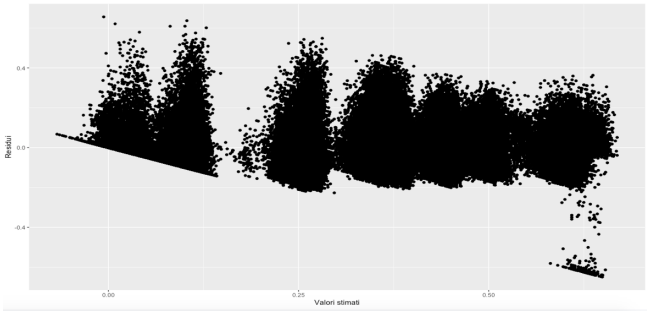

Figura 26. *Scatter plot* dei residui sui valori stimati dal modello.

5.2 Previsione di *Valence*

Allo stesso modo, è stato stimato un modello che potesse prevedere il livello di positività trasmesso dal brano tramite il volume, l'energia, il tempo, la ballabilità, il decennio di pubblicazione, la tonalità e la chiave. Tramite test F, si rifiuta, con $\alpha = 0.05$, l'ipotesi di incorrelazione fra la variabile dipendente e l'insieme dei regressori. Il volume risulta, tuttavia poco significativo - il *pvalue* è 0.03 -, così come alcune modalità della chiave. I decenni sono associati a parametri negativi, eccezion fatta per gli anni '30, e ordinati, stavolta, in ordine cronologico decrescente. Ciò implica che, col passare del tempo, il sentimento tendenziale sia stato progressivamente attratto da una polarità negativa, specialmente nel nuovo millennio.
Il coefficiente di determinazione multiplo è, come la propria versione corretta, pari a 0.4897. Infine, i valori VIF non segnalano alcun problema di multicollinearità.

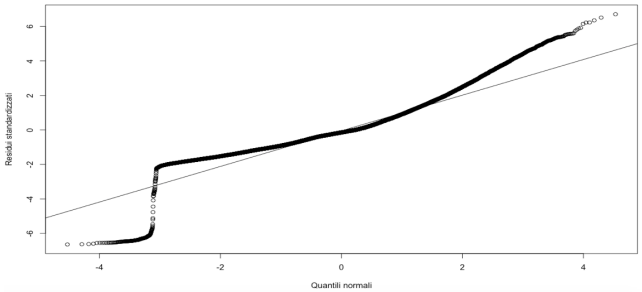

Figura 27. Grafico *Quantile-Quantile* per la valutazione della normalità dei residui.

6. Conclusioni e Discussione

Dalle analisi è evidente come il grado di popolarità su *Spotify* e di positività di un brano dipendano da una serie di grandezze relative sia all'aspetto stilistico - nell'ambito tecnico della composizione e della scelta dei diversi elementi che vanno a comporre il risultato finale -, sia al decennio di pubblicazione. In particolare, quest'ultimo carattere si dimostra predominante: la sua inclusione nei modelli di regressione lineare causa un forte incremento delle misure di bontà di adattamento R^2 ed R^2 corretto. Appare quindi necessario, per futuri approfondimenti, l'utilizzo di modelli di regressione più complessi: di tipo non lineare e multilivello. In particolare, i secondi potrebbero estrarre utilissime ulteriori informazioni sui fattori più popolari per ciascun decennio, stimando una diversa retta di regressione per ogni decade.

Codice

L'intero codice, implementato con linguaggio `R`, è disponibile al link: `github.com/RCrvro/codice.R`.