Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in

STATISTICA PER LE TECNOLOGIE E LE SCIENZE

Hit Song Science: il modello di regressione Beta per l'analisi dei gusti musicali nel mondo

Relatrice: prof.ssa Laura Ventura
Dipartimento di Scienze Statistiche

Laureando: Angelica Spada
Matricola n. 1224387

Anno Accademico 2022/23

Alle mie nonne
Ai miei amici e a mio fratello

Indice

Introduzione

Nella nostra società la musica è onnipresente in ogni aspetto della vita, in molte forme. Negli ultimi anni si parla spesso anche di musica sotto forma di dati. Ogni giorno vengono tracciati infatti milioni di bit di dati su Spotify e Apple Music, oltre a quelli forniti da altre piattaforme, quali i social media Instagram o TikTok. Nell'odierna economia dello streaming, infatti, il successo dell'artista o del brano musicale sulle piattaforme di streaming musicali e sui social è uno dei principali indicatori del successo. Ecco perché la comprensione dei dati diventa sempre più importante per le etichette discografiche, per pianificare le uscite degli album e la pubblicazione di video e post sui social media.

Lo scopo di questo lavoro è indagare la relazione tra le caratteristiche audio di un brano, ottenute dal database di Spotify, e la popolarità di questo, misurata dal numero di stream, e confrontare come differisce tale rapporto da paese a paese.

La motivazione che mi ha spinto ad approfondire la connessione tra musica e dati ha una duplice natura. In primo luogo, il mio interesse per la musica e i miei precedenti studi al Conservatorio hanno acceso in me il desiderio di conoscere meglio le dinamiche interne ed economiche dell'industria musicale. D'altra parte la scelta di sviluppare un'analisi per paese è stata influenzata dalla mia esperienza Erasmus di circa un anno a Madrid, durante la quale sono entrata in contatto con realtà musicali differenti da quella a cui ero abituata.

La relazione è articolata in cinque capitoli. I capitoli 1 e 2 forniscono un'introduzione alle caratteristiche dell'app Spotify, nonché una descrizione dei dati raccolti e delle relative Audio Features; il terzo capitolo si occupa di selezionare un modello che possa prevedere la popolarità di un brano. Nel quarto capitolo viene presentata un'analisi dei dati, che cerca di individuare i fattori di variazione e le eventuali correlazioni tra essi, soprattutto in merito all'influenza del paese, del genere musicale e delle audio features. Nell'ultimo capitolo, infine, si commentano i risultati ottenuti dall'analisi, discutendo brevemente le applicazioni pratiche al settore musicale.

Capitolo 1

Spotify e R

Spotify (SpotifyWebAPI, 2019) è una delle app musicali più famose al mondo, un programma di ultima generazione, cresciuto esponenzialmente negli ultimi anni. La piattaforma permette di ascoltare in streaming musica su computer, smartphone e tablet, scegliendo tra oltre 30 milioni di brani, vecchi e nuovi, delle principali case discografiche internazionali, senza dover acquistare legalmente singoli brani o album.
I dati studiati in questo elaborato sono stati estratti dalle classifiche settimanali di Spotify, chiamate *Spotify Weekly Charts*. Queste sono elenchi aggiornati settimanalmente delle 200 canzoni più ascoltate in streaming in 64 paesi, incluse le classifiche a livello globale (*Weekly Top Songs - Global)*.
Viene costruito quindi un dataset con 7 colonne che rappresentano i seguenti attributi: nome del brano (*track_name*), artista (*artist_name*), URL del brano (*track_url*), numero di ascolti (*streams*), posizione del brano tra i primi 200 (*track_position*), paese (*country* e *country_code*) e settimana (*week*). Il dataset comincia dalla settimana del 04/02/2021 a quella del 14/07/2022 con un totale di 15000 brani (Yelexa, 2022).

track_name	**artist_name**	**track_id**	**streams**	**track_ position**	**country**	**country_ code**	**week**
As it Was	Harry Styles	4LRPiXqCikLl N15c3yImP7	38,742,484	2	Global		Sep 2 - 8 2022
MI FAI IMPAZZIRE	BLANCO, Sfera Ebbasta	1x3Qb8np6S1 UvpSLthwEJN	4,197,515	1	Italy	IT	Aug 13 - 19 2021
Pink Venom	BLACKPINK	0skYUMpS0Ac bpjcGsAbRGj	291,759	4	South Korea	KR	Sep 2 - 8 2022
Running Up That Hill (A Deal With God)	Kate Bush	75FEaRjZTKL hTrFGsfMUXR	944,397	1	Australia	AU	Jul 29 - 4 2022

Tabella 1.1 Esempio di quattro brani e delle caratteristiche relative alle classifiche settimanali (*13/09/2022).*

Il passo successivo consiste poi nel raccogliere altre informazioni, quali la popolarità del brano (*track_popularity,* con un punteggio compreso tra 0 e 100), l'ID dell'artista (*artist_id*), il genere dell'artista (*artist_genre*), la popolarità dell'artista (*artist_popularity,* sempre con un punteggio da 0 a 100) e i followers dell'artista (*artist_followers*).

track_ name	**track_ popularity**	**artist_ name**	**artist_id**	**artist_ followers**	**artist_ genre**	**artist_ popularity**
As it Was	96	Harry Styles	6KImCVD70vtIoJWnq6nGn3	23167128	pop'	91
MI FAI IMPAZZIRE	72	BLANCO, Sfera Ebbasta	1MRiIeZbc0cRuxOafDUCtH	1168704	italian pop'	71
Pink Venom	92	BLACKPINK	41MozSoPIsD1dJM0CLPjZF	32089651	k-pop', 'k-pop girl group'	84
Running Up That Hill (A Deal With God)	93	Kate Bush	1aSxMhuvixZ8h9dK9jIDwL	1430718	art pop', 'art rock', 'baroque pop', 'new wave pop', 'permanent wave', 'piano rock', 'singer-songwriter'	79

Tabella 1.2 Esempio di quattro brani e delle caratteristiche relative alla popolarità (*15/09/2022).*

1.1 Accedere Spotify API utilizzando R

1.1.1 Configurazione di un Developer Account

L'accesso all'API richiede la creazione di un Developer Account, che, di conseguenza, richiede l'esistenza di un account Spotify. È necessario quindi accedere alla dashboard per sviluppatore e configurare un'applicazione cliccando *Crea un ID cliente*. Verranno chieste alcune informazioni di base sull'app, quali il nome, una breve descrizione e il tipo di app che viene creata. Una volta configurata l'app, vengono forniti il Client ID e e il Client Secret, le due chiavi private da utilizzare per autenticare la connessione all'API (Stochasticism, 2022).

1.1.2 Connessione all'API

L'accesso all'API Spotify necessita una richiesta HTTP sicura, per le quali c'è bisogno della libreria *httr*. Una richiesta HTTP avviene in genere in due passaggi: una richiesta e una risposta. Deve essere quindi contattato il server con una domanda, inviando tutti i dati necessari, che in questo caso sono le chiavi di autenticazione (il Client ID e il Client Secret), e il server, in cambio, fornirà una risposta. La risposta contiene i dati richiesti e, tramite un ID univoco chiamato token di accesso, autorizza il server a rispondere alle richieste dell'Account Developer.

```
> library(httr)

> clientID = '##############################'
> secret = '##################################'

> response = POST('https://accounts.spotify.com/api/token', accept_json(),
    authenticate(clientID, secret), body=list(grant_type='client_credentials'),
    encode='form', verbose())

> mytoken = content(response)$access_token
> HeaderValue = paste0('Bearer', mytoken)
```

1.1.3 Download delle informazioni su un artista

La documentazione fornita da Spotify elenca una serie di funzioni disponibili da utilizzare sull'API.
Un esempio è quello di ottenere alcune informazioni di base su un artista specifico (ad esempio, il numero di follower, i generi musicali, il punteggio di popolarità). Ad ogni artista viene assegnata una stringa ID univoca, che può essere recuperata visualizzando la pagina di un artista e copiando la stringa di lettere e numeri compresa tra *https://open.spotify.com/artist/* e il punto interrogativo. Ad esempio per il cantante Harry Styles recuperiamo il codice ID *6KImCVD70vtIoJWnq6nGn3* dall'URL *https://open.spotify.com/artist/6KImCVD70vtIoJWnq6nGn3?si=rCjb8A4tQnK1GFy-RAYSgA*. La richiesta HTTP per ottenere informazioni sull'artista si basa quindi sull'URL dell'endpoint *https://api.spotify.com/v1/artists/{id}* dove {id} deve essere sostituito dall'ID dell'artista.

```
> artistID = "6KImCVD70vtIoJWnq6nGn3"
> URI = paste0('https://api.spotify.com/v1/artists/', artistID)
> response2 = GET(url = URI, add_headers(Authorization = HeaderValue))
> Artist = content(response2)
```

Otteniamo un oggetto stringa Artist che contiene le informazioni relative a quell'artista. Ad esempio, il numero di follower si ottiene con:

```
> Artist$followers$total
  23167128
```

Le informazioni che possiamo recuperare sono le seguenti:

```
> names(Artist)
  "external_urls"     "followers"       "genres"          "href"      "id"
  "images"            "name"            "popularity"      "type"      "uri"
```

1.1.4 Download delle informazioni sugli album dell'artista

È possibile ottenere un elenco degli album dell'artista seguendo un endpoint diverso all'indirizzo *https://api.spotify.com/v1/artists/{id}/albums*.

```
> URI = paste0('https://api.spotify.com/v1/artists/', artistID,'/albums')
> response2 = GET(url = URI, add_headers(Authorization = HeaderValue))
> Albums = content(response2)
```

L'elenco degli album si trova con `Albums$items`.
Inoltre è possibile ottenere i dettagli per ogni album. Per il primo album, ad esempio, possiamo recuperare i seguenti dati:

```
> Albums$items[[1]]$id
  "5r36AJ6VOJtp00oxSkBZ5h"
> Albums$items[[1]]$name
  "Harry's House"
```

```
> Albums$items[[1]]$release_date
  "2022-05-20"
> Albums$items[[1]]$total_tracks
  13
```

Come per gli artisti, anche ad ogni album nel database di Spotify viene assegnato un ID stringa univoco e questo ID può essere utilizzato per ottenere ulteriori dati relativi all'album.

```
> names(Albums$items[[1]])
  "album_group"        "album_type"       "artists"          "available_markets"
  "external_urls"      "href"             "id"               "images"
  "name"               "release_date"     "release_date_precision"
  "total_tracks"       "type"             "uri"
```

1.1.5 Download delle informazioni su un album e le sue tracce

Il download di qualsiasi altro dato dall'API segue praticamente lo stesso schema delle richieste HTTP. Dato un ID album specifico, in questo caso l'album "Harry's House" di Harry Styles, i comandi sono i seguenti:

```
> albumID = "5r36AJ6VOJtp00oxSkBZ5h"
> track_URI = paste0('https://api.spotify.com/v1/albums/', albumID,'/tracks')
> track_response = GET(url=track_URI, add_headers(Authorization=HeaderValue))
> tracks = content(track_response)
```

che ancora una volta forniscono un oggetto stringa album, da cui è possibile ottenere le informazioni richieste. Per ogni brano dell'album è possibile ottenere i seguenti dati:

```
> names(Albums$items[[1]])
  "name"               "artist"           "disc_number"
  "track_number"       "duration_ms"
```

Ad esempio, sempre per l'album Harry's House, otteniamo:

```
  "Music For a Sushi Restaurant"   "Harry Styles"    1
  1                                193813
```

Capitolo 2

Le Audio Features

Spotify Web consente agli utenti di estrarre diverse caratteristiche audio dei brani, le cosiddette *Audio Features*. Dopo aver ottenuto le informazioni su ciascun brano e artista, allo stesso modo, si procede con l'estrazione delle caratteristiche audio, che andranno poi, assieme alla popolarità, a comporre il dataset da analizzare. Le caratteristiche audio analizzate in questa relazione sono elencate nella Tabella 2.1.

Nome	**Tipo**	**Descrizione**
acousticness	float	Una misura da 0 a 1 relativa al fatto che la traccia sia acustica. 1 rappresenta un'elevata sicurezza.
analysis_url	string	Un URL per accedere all'analisi audio completa di questa traccia. Da tenere presente che per accedere a questi dati è necessario un token di accesso.
danceability	float	La ballabilità descrive quanto sia adatto un brano per ballare in base a una combinazione di elementi musicali tra cui tempo, stabilità del ritmo, forza del ritmo e regolarità generale. Il valore di 0 indica meno ballabile mentre 1 più ballabile.
duration_ms	integer	La durata della traccia in millisecondi.
energy	float	L'energia è una misura da 0 a 1 e rappresenta una misura percettiva di intensità e attività. In genere, le tracce energiche sono veloci e rumorose. Ad esempio, il death metal ha un'energia elevata, mentre un preludio di Bach ha un punteggio basso sulla scala. Le caratteristiche percettive che contribuiscono a questo attributo includono la gamma dinamica, il volume percepito, il timbro e l'entropia generale.
id	string	L'ID Spotify per la traccia.
instrumentalness	float	Indica se una traccia contiene o non contiene parti cantate. I suoni "ooh" e "aah" sono trattati come strumentali in questo contesto. Le tracce rap o parlate sono chiaramente "vocali". Più il valore della strumentalità è vicino a 1, maggiore è la probabilità che la traccia non contenga contenuto vocale. I valori superiori a 0.5 rappresentano in genere tracce strumentali, ma la sicurezza è maggiore quando il valore si avvicina a 1.
key	integer	La chiave in cui si trova la traccia. I numeri interi vengono mappati alle altezze utilizzando la notazione standard. Per esempio. 0 = C, 1 = C♯/RE♭, 2 = RE, e così via. Se non è stata rilevata alcuna chiave, il valore è -1.

liveness	float	Rileva la presenza di un pubblico nella registrazione. Valori di liveness più elevati rappresentano una maggiore probabilità che la traccia sia stata eseguita dal vivo. Un valore superiore a 0.8 fornisce una forte probabilità che la traccia sia in live.
loudness	float	Il volume complessivo di una traccia in decibel (dB). I valori di sonorità vengono mediati sull'intera traccia e sono utili per confrontare l'intensità sonora relativa delle tracce. I valori in genere sono compresi tra -60 e 0 db.
mode	integer	Mode indica la modalità (maggiore o minore) di un brano, il tipo di scala da cui deriva il suo contenuto melodico. Il maggiore è rappresentato da 1 e il minore è 0.
speechiness	float	La speechiness rileva la presenza di parole pronunciate in una traccia. Più la registrazione è esclusivamente simile al parlato (ad es. talk show, audiolibri, poesie), più il valore dell'attributo si avvicina a 1. I valori superiori a 0.66 descrivono tracce che probabilmente sono composte interamente da parole pronunciate. I valori compresi tra 0.33 e 0.66 descrivono tracce che possono contenere sia musica che parlato, in sezioni o a strati, inclusi casi come la musica rap. I valori inferiori a 0.33 molto probabilmente rappresentano musica e altre tracce non vocali.
tempo	float	Il tempo complessivo stimato di una traccia in battiti al minuto (BPM). Nella terminologia musicale, il tempo è la velocità o il ritmo di un dato brano e deriva direttamente dalla durata media del battito.
time_signature	integer	Un tempo in chiave stimato. Il tempo in chiave (metro) è una convenzione di notazione per specificare quante battute ci sono in ogni misura. Il tempo in chiave varia da 3 a 7 indicando i tempi in chiave da "3/4", a "7/4".
track_href	string	Un collegamento all'endpoint dell'API Web che fornisce tutti i dettagli della traccia.
type	string	Il tipo di oggetto.
uri	string	L'URI di Spotify per la traccia.
valence	float	Una misura da 0 a 1 che descrive la positività musicale veicolata da un brano. I brani con valenza alta suonano più positivi (ad es. più allegri, euforici), mentre i brani con valenza bassa suonano più negativi (ad es. più tristi, depressi, arrabbiati).

Tabella 2.1 Descrizione delle *Audio Features* di Spotify.

Lo scopo di questa prima parte della relazione è indagare se le caratteristiche audio di Spotify possano essere considerate determinanti nella popolarità delle canzoni.

2.1 Download delle funzionalità di traccia

L'univoco ID di un brano può essere utilizzato nell'API per estrarre le funzionalità della traccia audio. Similmente al procedimento utilizzato per l'artista e per l'album è possibile recuperare i valori delle audio features in questo modo:

```
> track_id = “1aSxMhuvixZ8h9dK9jIDwL”
> track_URI2 = paste0('https://api.spotify.com/v1/audio-features/', track_id)
> track_response2 = GET(url=track_URI2, add_headers(Authorization=HeaderValue))
> tracks2 = content(track_response2)
```

Questo comando fornisce quindi le Audio Features elencate nella Tabella 2.1. La popolarità d’altra parte può essere ottenuta tramite il seguente comando, sostituendo audio features con tracks e quindi richiedendo le caratteristiche generali del brano invece che quelle audio:

```
> track_URI3 = paste0('https://api.spotify.com/v1/tracks/', track_id)
> track_response3 = GET(url=track_URI3, add_headers(Authorization=HeaderValue))
> tracks3 = content(track_response3)
```

Ad esempio, la popolarità del brano “Pink Venom” del gruppo BLACKPINK si ottiene con il comando:

```
> tracks3$popularity
  91
```

Di seguito un esempio:

track_ name	dance ability	energy	loudness	speechiness	acoustic ness	instrumental ness	liveness	valence	tempo
As it Was	0,52	0,731	-5,338	0,0557	0,342	0,00101	0,311	0,662	173,93
Pink Venom	0,798	0,697	-7,139	0,0891	0,0202	0	0,259	0,745	90,031
MI FAI IMPAZZIRE	0,477	0,844	-4,87	0,0458	0,00273	0	0,137	0,203	169,941
Running Up That Hill (A Deal With God)	0,629	0,547	-13,123	0,055	0,72	0,00314	0,0604	0,197	108,375

Tabella 2.2 Esempio di quattro brani e delle caratteristiche audio (*15/09/2022).*

2.2 La popolarità dell'API

Tra tutte le caratteristiche analizzate da Spotify, la popolarità di un brano gioca un ruolo fondamentale. La popolarità di una traccia è un valore compreso tra 0 e 100, dove 100 indica più popolare. Il valore numerico è calcolato da un algoritmo e si basa, nella maggior parte dei casi, sul numero totale di riproduzioni della traccia, tenendo conto di quanto siano recenti tali riproduzioni. In generale, le canzoni che vengono ascoltate molto al momento avranno una popolarità maggiore rispetto alle canzoni che sono state ascoltate molto in passato. Le tracce duplicate (come ad esempio la stessa traccia di un singolo e di un album) vengono valutate in modo indipendente. La popolarità dell'artista e dell'album deriva matematicamente dalla popolarità del brano. È da tenere presente che il valore della popolarità non viene aggiornato in tempo reale e la popolarità effettiva può essere fornita solo dopo alcuni giorni. La popolarità delle canzoni è una questione fondamentale per l'industria musicale, soprattuto per quanto riguarda le conseguenze economiche che ne derivano. Nel 2021 l'industria musicale ha generato 14,99 miliardi di dollari solo negli Stati Uniti e grazie ai servizi di streaming in crescita (Spotify, Apple Music, ecc.) continua a prosperare. I numeri arrivano dalla RIAA, la Recording Industry Association of America, che ha pubblicato i suoi dati a Marzo 2022, segnati da una crescita clamorosa rispetto al 2020, il 23% in più, pari a 2,85 miliardi di dollari (Ingham, 2022).

Naturalmente i fattori che vanno ad incidere sulla popolarità di un brano sono stati studiati in precedenza con vari gradi di successo, come riporta lo scritto *Predict-the-Hit: Prediction of Hit Songs based on Multimodal Data* di Samyak, Parth e Sarthak (2022).

Pachet e Roy (2008) hanno utilizzato caratteristiche esterne estratte dall'ecosistema musicale, come la presenza sui social media, e caratteristiche interne relative all'audio: hanno scelto 632 funzionalità etichettate manualmente per ogni brano per incapsulare tutte le funzionalità interne ed esterne, ma non sono stati in grado di sviluppare un modello accurato e hanno concluso che non sarebbe stato possibile farlo con tecniche di apprendimento automatico. Ni et al. (2011) si sono concentrati invece solo sull'utilizzo di funzionalità interne per prevedere la popolarità di una canzone. Tramite un'algoritmo di apprendimento che sfruttava anche la variabile tempo sono stati in grado di ottenere una precisione del 60%. Tuttavia l'algoritmo era limitato alle classifiche del Regno Unito e non era in grado di generalizzare correttamente. Yang et al. (2017) hanno poi sperimentato modelli di deep learning e i loro esperimenti hanno prodotto risultati promettenti su diversi set di dati. Infine Singhi e Brown (2014) hanno utilizzato 31 caratteristiche di rima, sillaba e metro. Tuttavia per sviluppare il loro modello bayesiano, che ha dato una precisione del 21,4%, utilizzarono un set di dati sbilanciato con circa il 7% delle canzoni totali considerate hit e il resto

non hit. Studi precedenti simili, che consideravano il testo di un brano per prevederne la popolarità, avevano un successo limitato. Per questo motivo, e per la complessità dell'argomento, in questo elaborato ci limiteremo ad un'analisi delle caratteristiche musicali tralasciando totalmente l'analisi testuale.

Nel prossimo capitolo si farà un breve richiamo teorico alla distribuzione Beta e al suo uso inferenziale per poi procedere, nel Capitolo 4, alla sua applicazione ai dati.

Capitolo 3

Determinare la popolarità di un brano dalle Audio Features

Lo scopo di questo capitolo è l'identificazione delle caratteristiche che vanno ad influenzare la popolarità di un brano. In particolare, vogliamo indagare la possibile relazione tra le Audio Features del dataset Spotify (quali energy, loudness, ecc...) e la popolarità del brano, anch'essa disponibile nel dataset. L'individuazione di un modello in grado di descrivere questa relazione e la determinazione all'interno del set di caratteristiche di quelle ritenute più significative nel rendere un brano popolare sono argomenti molto interessanti per coloro che mirano a prevedere il successo di nuovi prodotti musicali. In mercati culturali come quello musicale, la modellazione a scopo previsionale è molto complessa. Studi in questo campo chiamato, Hit Song Science (HSS), interessano le case discografiche, ma anche Spotify e i consumatori stessi.

I precedenti tentativi in questa direzione, come visto nel Capitolo 2, hanno fatto riferimento a una serie di modelli differenti. In questo capitolo, rielaborazione dal libro *Ho perso le parole: come ritrovarle con la sentiment analysis* di Sciandra e Spera (2020) e de *Il modello di regressione con variabile risposta Beta* di Magro (2011), si preferisce l'applicazione di una regressione Beta, particolarmente adatta per modellare variabili risposta con dominio (0,1).

3.1 Il modello di regressione Beta

La distribuzione Beta è una distribuzione di probabilità continua definita nell'intervallo unitario con una funzione di densità data da:

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1},$$

con $\mu \in (0,1)$ e $\phi > 0$, dove $\Gamma(\,.\,)$ indica la funzione Gamma.

Il parametro μ indica il valore atteso di Y, cioè $E(Y) = \mu$. Il parametro ϕ soddisfa la definizione di un parametro di precisione poiché, per μ fisso, maggiore è il valore di ϕ, minore è la varianza della variabile dipendente μ. In particolare, si ha $Var(Y) = \frac{\mu(1-\mu)}{1+\phi}$.

Nei modelli di regressione Beta, il parametro che indica la media $\mu \in (0,1)$ viene espresso in funzione delle covariate, mentre il parametro di precisione $\phi \in R^+$ viene trattato come un parametro di disturbo. Per fare in modo che il predittore lineare assuma valori nello spazio dato dal supporto della variabile dipendente, il link logit rappresenta la funzione legame più comunemente scelta, ossia si assume

$$g(\mu_i) = log\left(\frac{\mu_i}{1-\mu_i}\right) = x_i^T \beta,$$

dove x_i^T indica un vettore di n variabili esplicative, e β è il al vettore dei coefficienti di regressione, con $i = 1, \ldots n$.

La distribuzione Beta è definita solo sull'intervallo unitario aperto. Se si osservano uno e zero esatti, questi valori devono essere trasformati in modo da garantire la natura del supporto della distribuzione Beta. La trasformazione applicata più frequentemente è data da

$$Y^* = [Y(n-1) + 0.5]/n,$$

dove Y^* è la trasformata e Y è la variabile dipendente non trasformata. In alternativa, è stato suggerito di aggiungere una piccola quantità ε, ad esempio $0{,}005$ o $0{,}01$ al limite inferiore, e di sottrarre lo stesso valore dal limite superiore.

3.1.1 Interpretazione dei β

Assumendo la funzione di legame logit, si suppone che il valore del j-esimo regressore venga aumentato da una costante c e che le variabili esplicative rimangano invariate. Allora varrà la relazione:

$$e^{c\beta_j} \simeq \frac{\mu^+/(1-\mu^+)}{\mu(1-\mu)},$$

dove con μ^+ viene indicata la media sotto le nuove condizioni e con μ la media sotto le condizioni iniziali (con le covariate originali). Il j-esimo regressore moltiplicato per la costante c é pari al logaritmo tra il rapporto tra quote (odds ratio, O.R.) sotto le nuove condizioni e quello sotto le condizioni originarie.

3.2 Stime di β e ϕ e inferenza

3.2.1. Stime di β e ϕ

La log-verosimiglianza per (β, ϕ) basata sul campione di n osservazioni indipendenti è

$$l(\beta, \phi) = \sum_{t=1}^{n} l_i(\mu_i, \phi),$$

dove

$$l_i(\mu_i, \phi) = log\Gamma(\phi) - log\Gamma(\mu_i\phi) - log\Gamma((1-\mu_i)\phi) +$$
$$+(\mu_i\phi - 1)log\, y_i + ((1-\mu_i)\phi - 1)log(1-y_i),$$

con μ_i come definito nel paragrafo 3.1.

La funzione score, detta anche funzione di punteggio, è data da

$$\nabla l(\beta, \phi) = \begin{bmatrix} l_\beta \\ l_\phi \end{bmatrix},$$

con $l_\beta = \dfrac{\partial l(\beta, \phi)}{\partial \beta}$ e $l_\phi = \dfrac{\partial l(\beta, \phi)}{\partial \phi}$.

Le stime di massima verosimiglianza per β e per ϕ sono ottenibili dalle equazioni di verosimiglianza, che sono le funzioni di punteggio poste uguali a zero. Nel caso specifico di questo tipo di modelli, queste equazioni non sono risolvibili analiticamente e si deve ricorrere a metodi numerici per risolverle quali, ad esempio, l'algoritmo di Newton - Raphson.

La regola di aggiornamento dell'algoritmo di Newton - Raphson al passo s per il vettore $(k+1)$-dimensionale $\theta = (\beta_1, \ldots, \beta_k, \phi)$ è data da

$$\hat{\theta}_{s+1} = \hat{\theta}_s - \left(\frac{\partial^2 l(\hat{\theta}_s)}{\partial\theta\partial\theta^T}\right)^{-1} \frac{\partial l(\hat{\theta}_s)}{\partial\theta}, \quad s = 0,1,...,$$

dove $\hat{\theta}_s = (\hat{\beta}_{1s}, \ldots, \hat{\beta}_{ks}, \hat{\phi}_s)$. È noto che i metodi numerici necessitano di valori iniziali per i parametri dai quali far partire la prima iterazione dell'algoritmo, dati da β_0 e ϕ_0. Questi valori possono essere scelti arbitrariamente, ma certe scelte risultano migliori e più leggere computazionalmente di altre. In particolare, Ferrari e Cribari-Neto (2004) suggeriscono la seguente soluzione. Per il vettore β, una buona scelta è data da

$$\beta_0 = (X^T X)^{-1} X^T z,$$

dove $z = (g(y_1), \ldots, g(y_n))^T$. Questa soluzione deriva dall'idea di effettuare una regressione lineare utilizzando come variabili risposta le $g(y_1), \ldots, g(y_n)$, e di stimare i parametri β con il metodo dei minimi quadrati.

Per la stima iniziale di ϕ, si ricorda che $Var(Y_i) = \dfrac{\mu_i(1-\mu_i)}{1+\phi}$, da cui è facile ricavare che

$$\phi = \frac{\mu_i(1-\mu_i)}{Var(Y_i)} - 1, \; i = 1,...,n.$$

Utilizzando i primi due termini dello sviluppo di Taylor di $g(y_i)$ in μ_i, si ha

$$Var(g(y_i)) \simeq Var(g(\mu_i) + (y_i - \mu_i)g'(y_i)) = Var(y_i)[g'(\mu_i)]^2$$

e quindi che

$$Var(y_i) \simeq \frac{Var(g(y_i))}{[g'(\mu_i)]^2}, \quad i = 1,...,n.$$

Di conseguenza, il valore iniziale per ϕ è

$$\phi_0 = \frac{1}{n}\sum_{i=1}^{n} \frac{\breve{\mu}_i(1-\breve{\mu}_i)}{\breve{\sigma}_i^2} - 1,$$

dove $\breve{\mu}_i$ è ottenuto applicando $g^{-1}(\cdot)$ all'i-esimo valore stimato dal modello di regressione lineare di $g(y_1), \ldots, g(y_n)$ su X, ovvero

$$\breve{\mu}_i = g - 1(x_i^T(X^TX)^{-1}X^Tz), \quad i = 1,...,n,$$

mentre

$$\breve{\sigma}_i^2 = \frac{\breve{e}^T\breve{e}}{(n-k)[g^{-1}(\breve{\mu}_i)]^2},$$

con $\breve{e}$ vettore dei residui empirici della regressione di z su X, ossia $\breve{e} = z - X(X^TX)^{-1}X^Tz$.

3.2.2 Inferenza

Per determinare la distribuzione asintotica degli stimatori di massima verosimiglianza risulta utile la matrice d'informazione attesa ricavata al paragrafo 3.2.1, in quanto la sua inversa fornisce una stima della matrice di covarianza asintotica di $(\hat{\beta}, \hat{\phi})$ sotto (β, ϕ). Sotto condizioni di regolarità, vale

$$\begin{pmatrix} \hat{\beta} \\ \hat{\phi} \end{pmatrix} \sim N_{k+1}\left(\begin{pmatrix} \beta \\ \phi \end{pmatrix}, i(\hat{\beta}, \hat{\phi})^{-1}\right),$$

dove $\hat{\beta}$ e $\hat{\phi}$ sono gli stimatori di massima verosimiglianza di β e di ϕ.

È utile ottenere l'espressione dell'inversa della matrice d'informazione attesa. Per le proprietà delle matrici a blocchi, si ha che

$$i(\beta,\phi)^{-1} = \begin{pmatrix} i^{\beta\beta} & i^{\beta\phi} \\ i^{\phi\beta} & i^{\phi\phi} \end{pmatrix},$$

dove

$$i^{\beta\beta} = \frac{1}{\phi}(X^TWX)^{-1}\left(I_k + \frac{X^TTcc^TT^TX(X^TWX)-1}{\gamma^{\phi}}\right),$$

$$i^{\beta\phi} = (i^{\phi\beta})^T = -\frac{1}{\gamma\phi}(X^TWX)^{-1}X^TTc$$

e

$$i^{\phi\phi} = \frac{1}{\gamma},$$

con $\gamma = tr(D) - \phi^{-1}c^TT^TX(X^TWX)^{-1}X^TTc$ e I_k è la matrice identità di dimensione $k \times k$.

È possibile quindi ricavare anche stime intervallari approssimate per i parametri. Posto $z_{1-\alpha/2}$ il quantile $1-\frac{\alpha}{2}$ della $N(0,1)$ e posto $i^{rr}(\hat{\beta},\hat{\phi})$ l'r-esima componente della diagonale dell'inversa della matrice dell'informazione attesa valutata in $(\hat{\beta},\hat{\phi})$, con $r = 1,...,k+1$, si ha che

$$[\hat{\beta}_r - z_{1-\alpha/2}(i^{rr}(\hat{\beta},\hat{\phi}))^{1/2}, \hat{\beta}_r + z_{1-\alpha/2}(i^{rr}(\hat{\beta},\hat{\phi}))^{1/2}]$$

e

$$[\hat{\phi} - z_{1-\alpha/2}(i^{(k+1)(k+1)}(\hat{\beta},\hat{\phi}))^{1/2}, \hat{\phi} + z_{1-\alpha/2}(i^{(k+1)(k+1)}(\hat{\beta},\hat{\phi}))^{1/2}]$$

rappresentano, rispettivamente, intervalli di confidenza per β_r e per ϕ di livello approssimato $1-\alpha$.

Un intervallo di confidenza di livello approssimato $1-\alpha$ per la media μ è dato da

$$[g^{-1}(\hat{\eta} - z_{1-\alpha/2}se(\hat{\eta})), g^{-1}(\hat{\eta} + z_{1-\alpha/2}se(\hat{\eta}))],$$

dove $\hat{\eta} = x^T\hat{\beta}$ e $se(\hat{\eta}) = \sqrt{x^Ti^{\beta\beta}(\hat{\beta},\hat{\phi})x}$, con x vettore di covariate fissato e $i^{\beta\beta}(\hat{\beta},\hat{\phi})$ componente (β,β) dell'inversa dell'informazione di Fisher valutata con in $(\hat{\beta},\hat{\phi})$. Si noti che questo intervallo di confidenza è valido solo se la funzione di legame è strettamente crescente.

Supponiamo ora di essere interessati ad effettuare una verifica d'ipotesi del tipo

$$H_0 : \beta_1 = \beta_1^{(0)} \text{ vs } H_1 : \beta_1 \neq \beta_1^{(0)},$$

con $\beta_1 = (\beta_1,\ldots,\beta_m)^T$, $\beta_1^{(0)} = (\beta_1^{(0)},\ldots,\beta_m^{(0)})^T$, per $m \leq k$, e $\beta_1^{(0)}$ vettore di costanti note e fissate. Per tale problema, si può far riferimento alla statistica test log-rapporto di verosimiglianza (Pace e Salvan, 2001), data da

$$W_{lr} = 2(l(\hat{\beta},\hat{\phi}) - (\tilde{\beta},\tilde{\phi})),$$

dove $(\tilde{\beta}, \tilde{\phi})$ sono le stime di massima verosimiglianza di (β, ϕ) sotto l'ipotesi nulla H_0. Sotto condizioni di regolarità e sotto H_0, si ha che $W_{lr} \to^d \chi^2_m$. L'ipotesi nulla viene rifiuta per valori alti della statistica W_{lr}.
Un altro test che si può utilizzare è il test alla Wald, dato da

$$W_w = (\hat{\beta}_1 - \beta_1^{(0)})^T (i_{11}^{\beta\beta}(\hat{\beta}, \hat{\phi}))^{-1} (\hat{\beta}_1 - \beta_1^{(0)}),$$

dove $i_{11}^{\beta\beta}(\hat{\beta}, \hat{\phi})$ equivale a $i_{11}^{\beta\beta}$ valutata nelle stime di massima verosimiglianza. Con $i_{11}^{\beta\beta}$ si intende il blocco $i^{\beta\beta}$ privato delle righe e delle colonne in cui compaiono gli elementi diagonali legati ai parametri non testati. Il test W_w è asintoticamente equivalente al test W_{lr}. In particolare, per testare se il *j*-esimo parametro di regressione β_j $(j = 1,...,k)$ significativo, si può utilizzare la statistica test di Wald $\hat{\beta}_j / \sqrt{i_{jj}^{\beta\beta}(\hat{\beta}, \hat{\phi})}$ che si distribuisce asintoticamente come una normale standard.
I test W_{lr} e W_w possono essere anche utilizzati per confrontare modelli annidati.

3.3 Bontà di adattamento

Una volta stimato un modello, è importante effettuare un'analisi diagnostica per valutare la bontà dell'adattamento.
Una misura globale della varianza spiegata, e quindi dell'adattamento del modello ai dati, può essere ottenuta calcolando l'indice pseudo-R^2, denotato con R_p^2. Posto $\hat{\eta} = (\hat{\eta}_1, \ldots, \hat{\eta}_n)$ e $g(y) = (g(y_1), \ldots, g(y_n))$, l'indice pseudo-$R^2$ è definito come il quadrato del coefficiente di correlazione calcolato tra $\hat{\eta}$ e $g(y)$, per cui $0 \leq R_p^2 \leq 1$. Per quanto riguarda l'interpretazione di tale indice si può dire che è analoga a quella dell'R^2 per i modelli lineari normali. Un'ultima cosa da notare sull'indice pseudo-R^2 è che, in caso di perfetto accordo tra $\hat{\eta}$ e $g(y)$, che equivale ad un accordo perfetto tra $\hat{\mu}$ e y, esso assume il valore 1.
Un'altra valutazione del modello può essere ottenuta a partire dai residui standardizzati, o di Pearson, dati da

$$r_i = \frac{y_i - \hat{\mu}}{\sqrt{\hat{var}(Y_i)}}, \quad i = 1,...,n,$$

dove $\hat{\mu}_i = g^{-1}(x_i^T \hat{\beta})$ e $\hat{var}(Y_i) = [\hat{\mu}_i(1 - \hat{\mu}_i)]/(1 + \hat{\phi})$. Questi residui possono essere utilizzati per costruire diagrammi di dispersione che li mettono a confronto con gli indici delle osservazioni *i*, formando le coppie di punti (i, r_i), oppure con i valori $\hat{\eta}_i$, formando le coppie $(\hat{\eta}_i, r_i)$, $i = 1,...,n$. La presenza in questi grafici di andamenti sistematici indica che il modello adottato non è adeguato.

Un altro tipo di residui che si può considerare sono i residui di devianza, definiti come

$$r_i^d = sgn(y_i - \hat{\mu}_i)\sqrt{2(l_i(\ddot{\mu}_i, \hat{\phi}) - l_i(\hat{\mu}_i, \hat{\phi}))}, \quad i = 1,...,n,$$

dove $\ddot{\mu}_i$ è il valore di μ_i che risolve l'equazione $\partial l_i / \partial \mu_i = 0$, ossia $\phi(y_i^* - \mu_i^*) = 0$. Questi residui derivano dalla quantità

$$D(y; \hat{\mu}, \hat{\phi}) = \sum_{i=1}^{n} 2(l_i(\ddot{\mu}_i, \hat{\phi}) - l_i(\hat{\mu}_i, \hat{\phi})),$$

che è detta devianza del modello. È facile notare che la relazione che lega le quantità r_i^d e $D(y; \hat{\mu}, \hat{\phi})$ è data da $D(y; \hat{\mu}, \hat{\phi}) = \sum_{i=1}^{n} (r_i^d)^2$. Si ha quindi che più il valore di r_i^d è grande, più l'i-esima osservazione contribuisce alla devianza del modello, e viceversa. I residui di devianza vengono analizzati con gli stessi grafici dei residui standard e ci si aspetta, come nei primi, che se il modello è buono non ci siano andamenti sistematici.

L'ultima misura diagnostica che può essere considerata è la distanza di Cook che misura l'influenza di una singola osservazione sulle stime dei parametri di regressione, nel momento in cui viene tolta dal singolo processo di stima. Nel caso in cui la distanza di Cook assuma valori elevati (solitamente maggiori di 1) si può affermare che l'osservazione è molto influente, e si può quindi scegliere di ignorarla, se si ritiene che essa alteri in maniera scorretta le stime. La distanza di Cook è definita come

$$Cook_i = \frac{(\hat{\beta} - \hat{\beta}_{(i)})^T X^T W X (\hat{\beta} - \hat{\beta}_{(i)})}{k}, \quad i = 1,...,n,$$

dove $\hat{\beta}_{(i)}$ è la stima di massima verosimiglianza del parametro β effettuata senza l'i-esima osservazione. Si noti che tale quantità rappresenta una distanza tra $\hat{\beta}_{(i)}$ e $\hat{\beta}$. Infine, si osserva che, per evitare di dover stimare il modello per $n + 1$ volte, e quindi affrontare un algoritmo computazionalmente pesante, si può usare una comoda approssimazione della distanza di Cook, data da

$$C_i = \frac{h_{ii} - r_i^2}{k(1 - h_{ii})^2},$$

dove h_{ii} indica l'i-esimo elemento della diagonale della matrice di proiezione

$$H = W^{1/2} X (X^T W X)^{-1} X^T W^{1/2}$$

e r_i indica l'i-esimo residuo standard.

Capitolo 4

Analisi dei gusti musicali in 10 paesi

Si considera in questo capitolo l'applicazione ai dati del modello appena descritto. In particolare, si propone di analizzare le preferenze musicali in differenti paesi. È stato scelto un campione di 9 stati che completiamo con i dati delle classifiche globali. I paesi selezionati sono 3 europei (Italia, Regno Unito e Spagna), 2 asiatici (Corea del Sud, Giappone), Australia, Sud Africa e infine 2 americani (Brasile e Stati Uniti). Abbiamo quindi 10 dataset, ognuno ricavato da 52 classifiche settimanali.

Dopo aver eliminato i duplicati nei brani nella classifica Spotify Weekly Charts sono stati considerati i totali di brani unici divisi tra i paesi nel seguente modo: 1848 per l'Italia, 1633 per il Regno Unito, 1310 per la Spagna, 1735 per la Corea del Sud, 1772 per il Sud Africa, 933 in Giappone, 1305 per l'Australia, 1277 per il Brasile, 1829 per gli Stati Uniti.

Poiché il totale di brani nei dataset di ogni paese è lo stesso in partenza, appare evidente che, maggiore è il numero di brani che rimangono dopo l'eliminazione dei duplicati, più rapidamente i brani entrano ed escono dalle classifiche settimanali.

L'obiettivo dunque è individuare le caratteristiche che influenzano in modo significativo la popolarità o, in altri termini, che caratterizzano le canzoni più popolari per ogni paese e confrontare i risultati fra loro.

Se osserviamo la Tabella 4.1 su 14557 brani, 6798 sono presenti in un singolo paese.

	Stati Uniti	**Australia**	**Regno Unito**	**Corea del Sud**	**Spagna**
Brani unici presenti unicamente nelle classifiche di quel paese	530	151	494	700	998
Brani unici presenti nelle classifiche del paese	1829	1305	1633	1742	1310

Sudafrica	**Giappone**	**Italia**	**Brasile**	**TOT**
700	697	1473	1055	6798
1772	1848	1848	1277	14564

Tabella 4.1 Totale brani unici e totale brani unici presenti solamente nelle classifiche dello stato.

La Figura 4.1 ci indica invece il numero di brani comuni a 2 o più classifiche. I brani presenti quindi in tutte e 9 le classifiche sono solamente 72.

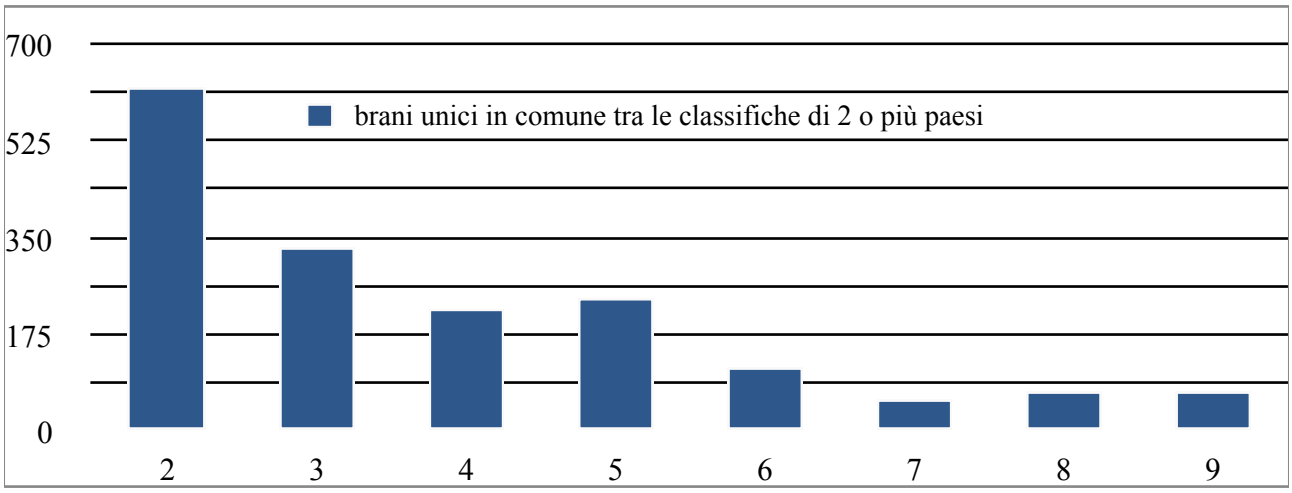

Figura 4.1 Brani presenti in 2 o più classifiche tra i 9 stati scelti.

4.1 Audio Features vs Genere e vs Paese

Prima di tutto si considerano le classifiche relative ai 5 generi più ascoltati in ogni paese, per poter avere un'idea generale delle differenze di ascolto tra i 10 paesi del nostro campione (Tabella 4.2) (Yelexa, 2022).

GLOBAL	**Spagna**	**Corea del Sud**	**Giappone**	**Sud Africa**
pop	trap latino	k-pop	j-pop	pop
trap latino	latin	pop	japanese teen pop	amapiano
latin	reggaeton	k-pop girl group	anime	afro soul
reggaeton	spanish pop	k-pop boy group	k-pop	south african pop
arrocha	urbano espanol	k-rap	j-rock	rap

Italia	**Regno Unito**	**Stati Uniti**	**Brasile**	**Australia**
italian hip hop	pop	pop	arrocha	pop
italian pop	uk hip hop	rap	sertanejo universitario	dance pop

trap italiana	uk pop	chicago rap	sertanejo pop	uk pop
italian adult pop	dance pop	dance pop	sertanejo	rap
pop	rap	hip pop	funk carioca	australian hip hop

Tabella 4.2 I 5 generi più ascoltati per ogni paese.

Si approfondisce un'analisi sulle caratteristiche audio dei brani più ascoltati in ogni paese. Nelle seguenti Heatmaps vengono rappresentati i valori medi delle caratteristiche audio di tutti i brani dopo la standardizzazione. La Figura 4.2 ci permette di comprendere la relazione tra i generi più ascoltati e le relative Audio Features che li caratterizzano, mentre la Figura 4.3 presenta le preferenze musicali specifiche per paese, sempre relativamente alle Audio Features dei brani.

	danceability	energy	speechiness	acousticness	instrumentalness	liveness	valence	duration	loudness	tempo
pop	0,6678	0,6595	0,1007	0,2401	0,0238	0,1702	0,5205	208915	-5,730	121,2
trap latino	0,6966	0,6271	0,1714	0,1895	0,0238	0,1849	0,4816	207581	-6,642	120,1
k-pop	0,6675	0,6132	0,1189	0,2139	0,0426	0,1687	0,4755	230559	-7,177	119,9
j-pop	0,6301	0,7358	0,0789	0,1619	0,0153	0,1962	0,5881	232126	-4,947	123,2
italian hip hop	0,6930	0,6717	0,1722	0,2093	0,0137	0,1724	0,5011	183062	-6,552	120,3
arrocha	0,6140	0,6042	0,1038	0,2427	0,0190	0,1590	0,4626	218242	-6,954	120,9
raeggeton	0,6374	0,5872	0,1190	0,2968	0,0441	0,1927	0,4742	201981	-7,299	120,5
trap	0,6800	0,6434	0,1521	0,1939	0,0019	0,1899	0,4552	211639	-6,377	122,2
latin	0,6959	0,6292	0,1273	0,2974	0,0309	0,1711	0,4981	214363	-6,831	119,8
hip hop	0,6696	0,7217	0,1009	0,2944	0,0163	0,2465	0,6024	185274	-5,900	119,1
sertanejo	0,6532	0,5755	0,1039	0,2409	0,0265	0,1591	0,4331	216255	-7,323	125,7
rock	0,6167	0,5694	0,1455	0,2836	0,0542	0,1854	0,4073	203739	-7,496	129,6
house	0,6350	0,7005	0,1316	0,2062	0,0016	0,1719	0,4528	185442	-6,314	122,1
rap	0,6643	0,6406	0,1011	0,2296	0,0236	0,1800	0,4974	207704	-5,943	121,7
afro soul	0,6222	0,5863	0,0764	0,3320	0,0376	0,1545	0,4573	211070	-7,053	119,4

Figura 4.2 Audio Feature vs Genere Heatmap.

Il colore verde indica un valore medio della caratteristica alto, mentre il bianco indica un valore più basso della relativa Audio Feature. Ad esempio, il genere rock presenta un valore molto alto per la strumentalità, mentre un valore molto più basso di ballabilità.

	danceability	energy	speechiness	acousticness	instrumentalness	liveness	valence	duration	loudness	tempo
Australia	0,6411	0,6140	0,1094	0,2589	0,0269	0,1759	0,4845	206058	-6,964	119,1
Brazil	0,6807	0,6960	0,1137	0,3483	0,0167	0,2548	0,6448	185260	-3,169	128,0
Italy	0,6688	0,6767	0,1445	0,2159	0,0159	0,1753	0,4983	186595	-6,346	120,9
Japan	0,6145	0,7322	0,0698	0,1730	0,0181	0,1955	0,5823	232203	-5,003	124,0
Korea	0,6445	0,6484	0,0948	0,2617	0,0233	0,1727	0,5091	206386	-5,770	119,8
United Kingdor	0,6484	0,6277	0,1316	0,2560	0,0292	0,1791	0,4981	204311	-6,974	121,2
United States	0,6575	0,6015	0,1310	0,2538	0,0314	0,1830	0,4705	199837	-7,123	122,3
Spain	0,7224	0,6697	0,1193	0,2629	0,0246	0,1622	0,5928	202322	-5,637	122,7
South Africa	0,6892	0,5824	0,1264	0,2249	0,0587	0,1597	0,4482	244672	-8,191	117,8
GLOBAL	0,6657	0,6311	0,1110	0,2600	0,0250	0,1790	0,5098	202733	-6,449	121,4

Figura 4.3 Audio Feature vs Paese Heatmap.

Dai grafici si osserva che l'Italia è uno tra i paesi con il numero più alto di brani presenti esclusivamente nelle classifiche italiane. Gli ascoltatori italiani preferiscono quindi musica in italiano rispetto a canzoni in lingua straniera e sembra infatti venga data molta importanza al testo, sia per i valori elevati di *speechiness*, sia per la presenza del trap tra i generi più ascoltati.

Con i valori *speechiness* più bassi troviamo invece il Giappone. Il mercato giapponese, secondo mercato musicale più grande al mondo, preferisce infatti musica energica e ad alto volume, focalizzata più sul ritmo che sul testo, come comprova la classifica dei generi più ascoltati che include j-pop, k-pop e j-rock.

Altro paese che ama il ritmo è la Spagna. Si trova infatti al primo posto per strumentalità e ha valori alti sia per valenza che per ballabilità. Il reggaeton e il latin sono infatti i due generi più ascoltati in questo paese e, se osserviamo la Figura 4.2, entrambi hanno una valenza e una ballabilità elevata, mentre il reggaeton ha una strumentalità molto alta.

Passiamo poi a Stati Uniti, Australia e Gran Bretagna, che presentano valori molto simili tra loro e la presenza di pop, hip hop e rap tra i generi preferiti. Notiamo che il numero di brani presenti unicamente nelle classifiche australiane, Tabella 4.2, è molto basso e deduciamo che gli utenti australiani ascoltino principalmente brani conosciuti anche a livello globale (in particolare possiamo presumere brani statunitensi e britannici) e che, anche per questo motivo, presenti valori simili agli altri due paesi.

È stato deciso di includere in questo campione anche Brasile e Sud Africa in quanto mercati musicali che ad oggi si stanno espandendo rapidamente e che prevedono una ulteriore crescita in futuro.

In Brasile, sia il sertanejo (una controparte brasiliana della musica country statunitense), che l'arrocha (caratterizzato dalla presenza di percussioni e chitarra elettrica) presentano alti valori in *acousticness*, affiancati da valori elevati in volume e tempo.

Il Sud Africa invece mostra valori elevati in strumentalità e durata mentre valori molto bassi in energia. Sia il genere afro soul che il genere amapiano (un ibrido tra deep house, jazz e lounge), infatti, sono caratterizzati dalla presenza di parti strumentali senza testo e ritmi più rilassanti.
Infine la Corea del Sud presenta valori molto equilibrati, caratterizzanti del pop; i generi più ascoltati sono infatti k-pop (maschile e femminile) e pop.

4.2 Analisi della correlazione delle caratteristiche audio

Per completezza si considera l'analisi delle correlazioni tra le funzionalità audio che classifichiamo utilizzando il coefficiente di correlazione di Pearson.
Nella Figura 4.4 la Heatmap che presenta le correlazioni tra le funzionalità audio.
Una correlazione positiva è indicata dal colore rosso mentre il bianco indica una correlazione negativa.
Notiamo che volume e energia hanno una forte correlazione positiva mentre *acousticness* ed *energy* hanno una correlazione negativa più forte rispetto ad altre coppie. La *valence* inoltre ha una leggera correlazione positiva con ballabilità, energia e volume.

	danceability	energy	speechiness	acousticness	instrumentalness	liveness	valence	duration	loudness	tempo
danceability	1,000	0,102	0,169	-0,246	-0,018	-0,129	0,346	-0,099	0,102	-0,121
energy	0,102	1,000	-0,008	-0,519	-0,065	0,155	0,392	-0,047	0,695	0,083
speechiness	0,169	-0,008	1,000	-0,029	-0,086	0,037	0,028	-0,113	-0,070	0,069
acousticness	-0,246	-0,519	-0,029	1,000	-0,009	0,007	-0,107	-0,064	-0,379	-0,060
instrumentalness	-0,018	-0,065	-0,086	-0,009	1,000	-0,028	-0,075	0,128	-0,182	0,014
liveness	-0,129	0,155	0,037	0,007	-0,028	1,000	0,057	-0,078	0,091	-0,001
valence	0,346	0,392	0,028	-0,107	-0,075	0,057	1,000	-0,179	0,258	0,049
duration	-0,099	-0,047	-0,113	-0,064	0,128	-0,078	-0,179	1,000	-0,152	-0,016
loudness	0,102	0,695	-0,070	-0,379	-0,182	0,091	0,258	-0,152	1,000	0,044
tempo	-0,121	0,083	0,069	-0,060	0,014	-0,001	0,049	-0,016	0,044	1,000

Figura 4.4 Heatmap per la correlazione tra le caratteristiche audio.

4.3 Regressione Beta per l'Italia

Procedo ora nell'utilizzare le spiegazioni teoriche del Capitolo 3 per le analisi di questo capitolo. Come viene influenzata la formula che abbiamo ricavato in precedenza? Il gusto musicale di ogni paese ha un'incidenza rilevante o avremmo potuto analizzare la sola classifica *Global* e ottenere comunque dei risultati più che soddisfacenti?

4.4.1 Analisi esplorative

Il primo passo consiste nel riportare gli scatterplot in cui vengono messe in relazione le singole caratteristiche audio con l'indice di popolarità dei brani.

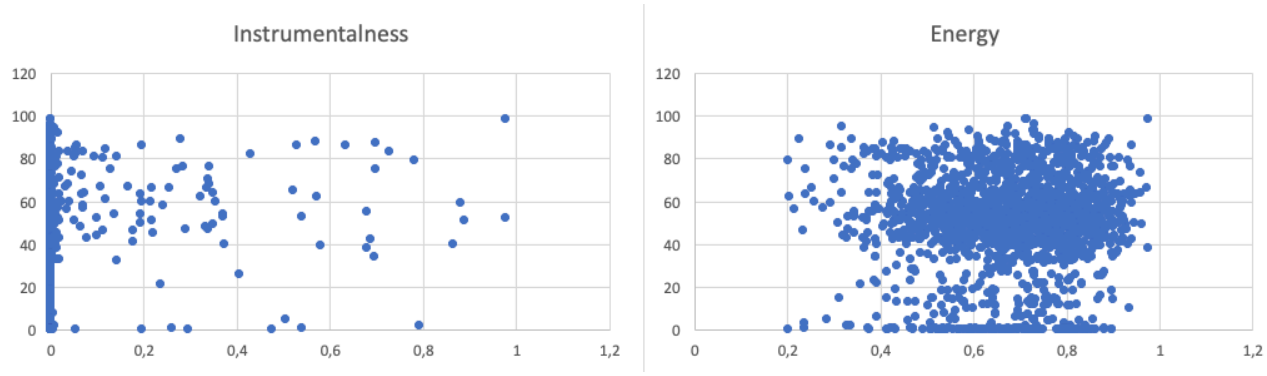

Figura 4.4 Scatterplot della Popolarità vs Instrumentalness e Energy.

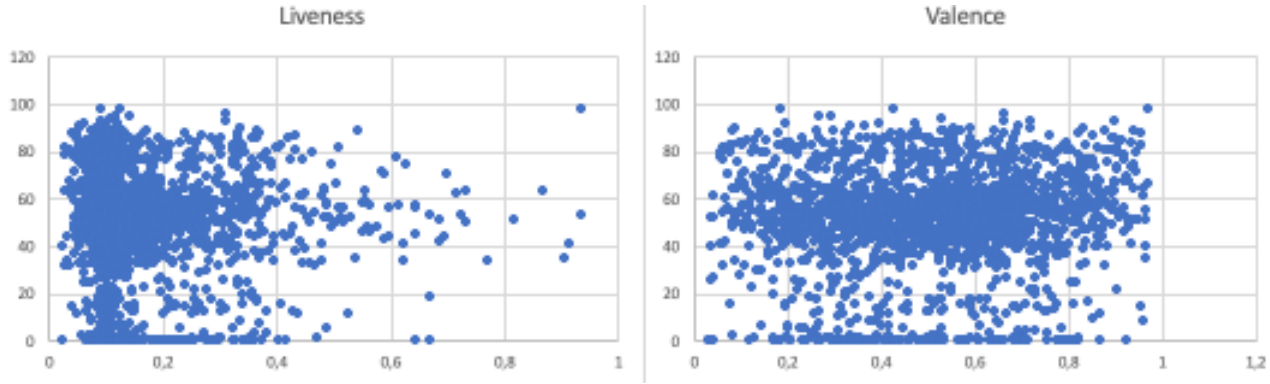

Figura 4.5 Scatterplot della Popolarità vs Liveness e Valence.

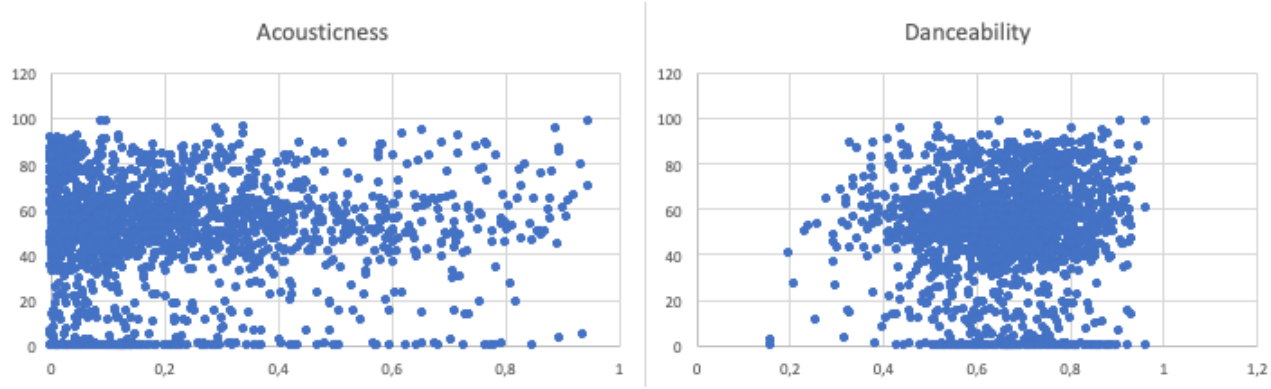

Figura 4.6 Scatterplot della Popolarità vs Acousticness e Danceability.

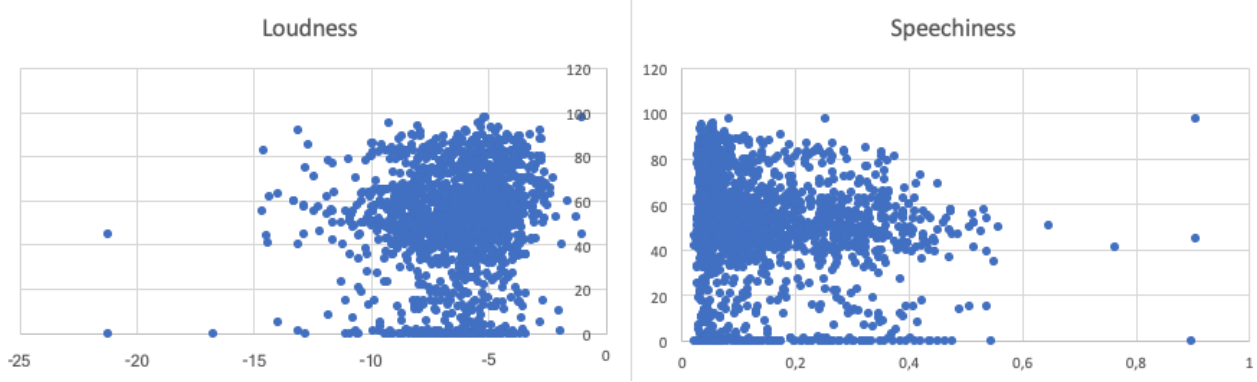

Figura 4.7 Scatterplot della Popolarità vs Loudness e Speechiness.

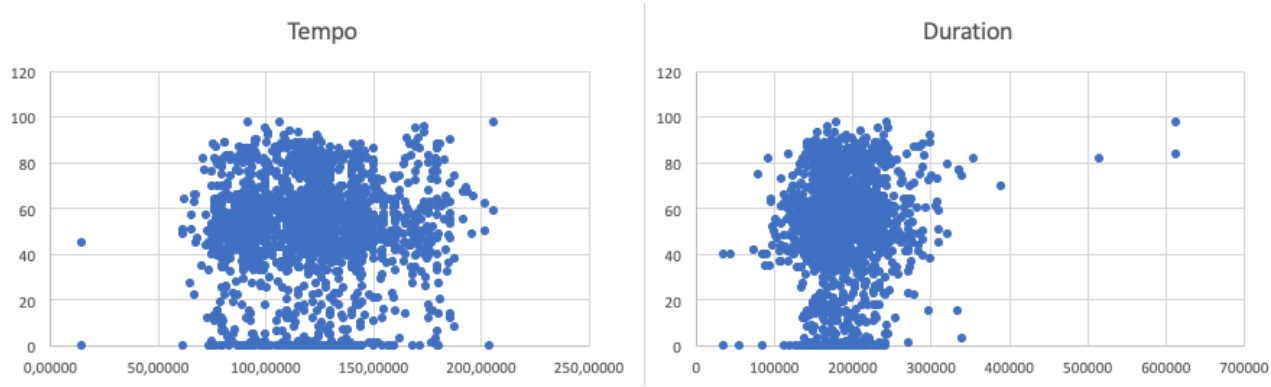

Figura 4.8 Scatterplot della Popolarità vs Tempo e Duration.

I grafici evidenziano che la maggior parte dei brani delle classifiche nelle Top Weekly Charts sono poco loquaci (*speechiness*), poco live (*liveness*), con molta energia e presentano valori di volume (*loudness*) e strumentalità (*strumentalness*) prossimi allo zero. In particolare, si osserva un andamento decrescente della popolarità rispetto a loquacità, acustica e live, ovvero all'aumentare di queste caratteristiche audio diminuisce la popolarità della canzone. I grafici della durata, tempo, valenza e strumentalità non sembrano

mostrare alcun andamento particolare, mentre energia, ballabilità e volume sembrano avere un andamento positivo.

Al fine di individuare quali caratteristiche musicali influenzino la popolarità e se questa relazione sia simile per tutti i paesi si è stimato un modello Beta in cui l'indice di popolarità è una variabile continua e limitata in un intervallo (0,1); è stato inoltre necessario trasformare tale indice, data la presenza di zeri esatti, in modo che i valori siano contenuti all'interno dell'intervallo.

4.4.2 Selezione del modello Beta

Si inizia considerando un modello che include tutte le caratteristiche audio. La funzione legame scelta è quella canonica, la funzione logit. La stima del modello `m1` fornisce i risultati riportati qui sotto.

```
> library("betareg")
> m1 <- betareg(formula = italy$popularity ~ italy$danceability + italy$energy +
   italy$speechiness + italy$acousticness + italy$instrumentalness + italy$valence +
   italy$liveness + italy$duration + italy$tempo + italy$loudness +
   italy$energy:italy$valence, data = italy)

> summary(m1)
   Call:
   betareg(formula = italy$popularity ~ italy$danceability + italy$energy +
   italy$speechiness + italy$acousticness + italy$instrumentalness + italy$valence +
       italy$liveness + italy$duration + italy$tempo + italy$loudness +
   italy$energy:italy$valence, data = italy)

   Standardized weighted residuals 2:
       Min      1Q  Median      3Q     Max
   -4.9764  0.0021  0.3219  0.5668  2.5599

   Coefficients (mean model with logit link):
                              Estimate Std. Error z value Pr(>|z|)
   (Intercept)                2.473e-01  4.555e-01   0.543 0.587149
   italy$danceability         2.538e-01  2.216e-01   1.145 0.252173
   italy$energy              -1.665e+00  5.055e-01  -3.294 0.000987 ***
   italy$speechiness         -9.615e-01  2.255e-01  -4.264 2.01e-05 ***
   italy$acousticness         7.441e-02  1.413e-01   0.527 0.598415
   italy$instrumentalness     4.501e-01  2.964e-01   1.518 0.128905
   italy$valence             -1.090e+00  6.035e-01  -1.806 0.070853 .
   italy$liveness             7.022e-02  2.127e-01   0.330 0.741357
   italy$duration             3.027e-06  6.960e-07   4.349 1.37e-05 ***
   italy$tempo                1.005e-03  9.413e-04   1.067 0.285819
   italy$loudness             5.776e-02  1.789e-02   3.230 0.001240 **
   italy$energy:italy$valence 2.198e+00  8.662e-01   2.538 0.011153 *
```

```
Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)  2.19881    0.06329   34.74   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 94.88 on 13 Df
Pseudo R-squared: 0.0293
Number of iterations: 24 (BFGS) + 3 (Fisher scoring)
```

Nell'output ci sono alcune statistiche di sintesi dei residui pesati standard, le stime di massima verosimiglianza del vettore β e del parametro ϕ (con relativi standard error) e i valori della statistica test di Wald, per testare la significatività di ogni coefficiente, con relativo di p-value. Sono quindi riportati il valore della log-verosimiglianza massimizzata con i relativi gradi di libertà $(k + h)$, il valore dell'indice pseudo-R^2 e il numero di iterazioni che sono state necessarie per stimare i parametri. Si nota che il modello ha i coefficienti *acousticness, liveness e tempo* non significativi.

Per selezionare il modello più adatto utilizziamo quindi la funzione StepBeta(·), la quale esegue un algoritmo per definire il miglior predittore lineare in base a un criterio definito dall'utente, di default il Akaike Information Criterion, noto anche come AIC. Otteniamo quindi il modello `m2` il cui adattamento è riportato in qui sotto.

```
> StepBeta(m1)
```

```
> StepBeta(m1)
[1] "100 % of the process"

Call:
"betareg(formula =  italy$popularity ~ italy$speechiness + italy$duration +
italy$valence + italy$loudness + italy$energy + italy$instrumentalness +
italy$energy:italy$valence data = italy )"

Coefficients (mean model with logit link):
(Intercept)              italy$speechiness              italy$duration
5.646e-01                  -9.268e-01                   2.936e-06
italy$valence              italy$loudness                italy$energy
-9.966e-01                  5.809e-02                   -1.677e+00
italy$instrumentalness  italy$valence:italy$energy
4.562e-01                   2.135e+00

Phi coefficients (precision model with identity link):
(phi)
2.196
```

```
> m2 <- betareg(formula = italy$popularity ~ italy$duration + italy$speechiness +
   italy$valence + italy$danceability + italy$loudness + italy$energy +
   italy$instrumentalness + italy$energy:italy$valence, data = italy )

> summary(m2)
   Call:
   betareg(formula = italy$popularity ~ italy$speechiness + italy$duration +
   italy$valence + italy$loudness + italy$energy + italy$instrumentalness +
   italy$energy:italy$valence,
       data = italy)

   Standardized weighted residuals 2:
       Min      1Q  Median      3Q     Max
   -4.9393  0.0107  0.3253  0.5704  2.5502

   Coefficients (mean model with logit link):
                               Estimate Std. Error z value Pr(>|z|)
   (Intercept)                 5.646e-01  3.968e-01   1.423 0.154768
   italy$speechiness          -9.268e-01  2.245e-01  -4.129 3.65e-05 ***
   italy$duration              2.936e-06  6.910e-07   4.249 2.14e-05 ***
   italy$valence              -9.966e-01  5.939e-01  -1.678 0.093354 .
   italy$loudness              5.809e-02  1.783e-02   3.258 0.001123 **
   italy$energy               -1.677e+00  4.897e-01  -3.426 0.000613 ***
   italy$instrumentalness      4.562e-01  2.961e-01   1.541 0.123387
   italy$valence:italy$energy  2.135e+00  8.595e-01   2.484 0.012974 *

   Phi coefficients (precision model with identity link):
         Estimate Std. Error z value Pr(>|z|)
   (phi)  2.19586    0.06319   34.75   <2e-16 ***
   ---
   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

   Type of estimator: ML (maximum likelihood)
   Log-likelihood: 93.78 on 9 Df
   Pseudo R-squared: 0.0286
   Number of iterations: 18 (BFGS) + 2 (Fisher scoring)
```

Nel caso italiano i risultati del evidenziano che *speechiness, valence, energy e danceability* sono le caratteristiche che influenzano negativamente l'indice di popolarità, mentre *duration, loudness, instrumentalness* sono quelle che lo influenzano positivamente. Si nota inoltre che l'interazione tra positività ed energia ha un effetto positivo molto alto.

La Tabella 4.4 riporta le l'Odds Radio dell'indice di popolarità per le caratteristiche significative, il quale, moltiplicato per 100, esprime la variazione in termini percentuali dell'indice di popolarità delle canzoni, quando le caratteristiche audio passano dal valore 0 al valore 1.

Caratteristiche	Odds Ratio
duration	0.3958
speechiness	1.0000
valence	0.3691
loudness	1.0598
energy	0.1869
instrumentalness	1.5780
valence:energy	8.4570

Tabella 4.4 Odds Ratio dell'indice di popolarità per le Audio Features significative.

Inoltre per una verifica della bontà di adattamento del modello, si esegue una analisi grafica dei residui (Figura 4.9)

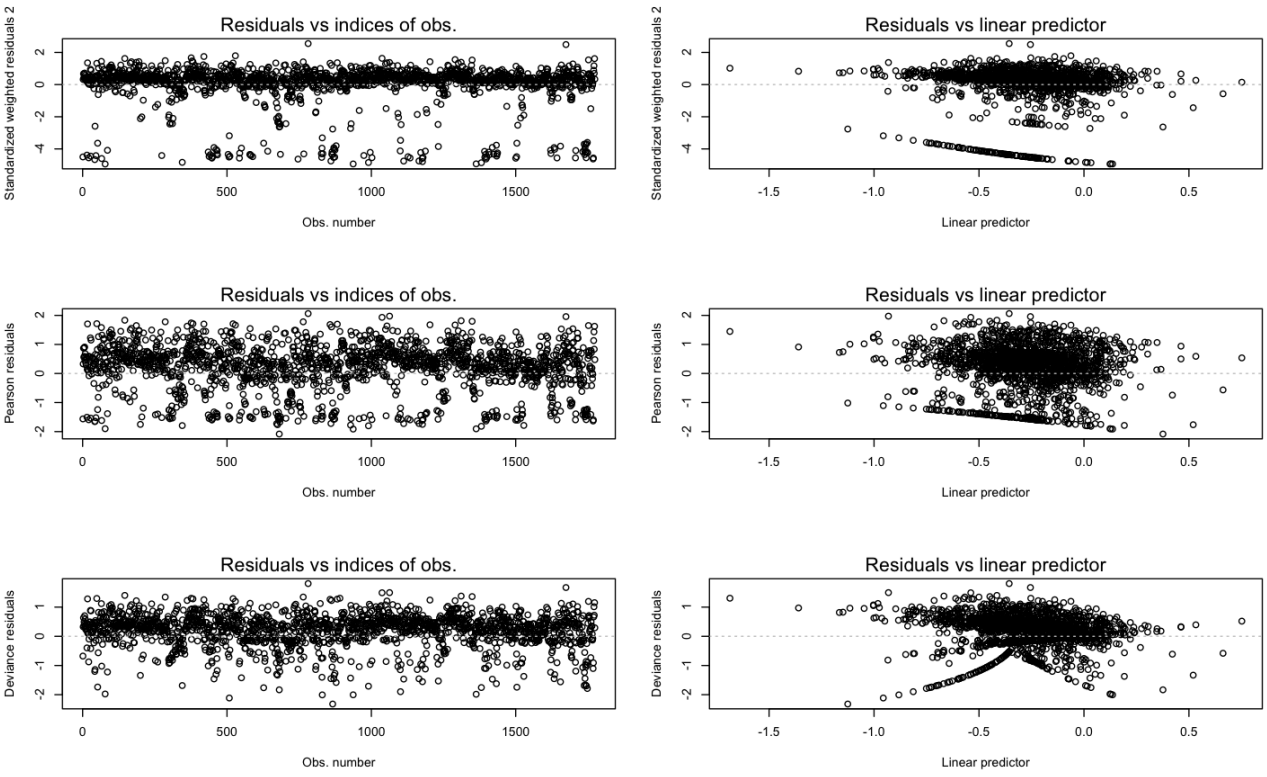

Figura 4.9 Residui `m2`.

4.4 Regressione Beta per i singoli stati

Riportiamo di seguito i modelli selezionati per gli altri paesi.

Modello Beta scelto per l'Australia:

```
> summary(australia)

  Call:
  betareg(formula = popularity ~ loudness + liveness + danceability + tempo +
  acousticness, data = australia)

  Standardized weighted residuals 2:
      Min      1Q  Median      3Q     Max
  -5.4892 -0.0678  0.2592  0.5219  1.7059

  Coefficients (mean model with logit link):
               Estimate Std. Error z value Pr(>|z|)
  (Intercept)   0.307658   0.222721   1.381   0.1672
  loudness      0.054142   0.012923   4.189  2.8e-05 ***
  liveness     -0.543305   0.221437  -2.454   0.0141 *
  danceability  0.407886   0.199924   2.040   0.0413 *
  tempo         0.001905   0.001021   1.866   0.0620 .
  acousticness  0.181136   0.123916   1.462   0.1438

  Phi coefficients (precision model with identity link):
        Estimate Std. Error z value Pr(>|z|)
  (phi)  2.66694    0.09399   28.38   <2e-16 ***
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  Type of estimator: ML (maximum likelihood)
  Log-likelihood: 97.78 on 7 Df
  Pseudo R-squared: 0.02159
  Number of iterations: 23 (BFGS) + 3 (Fisher scoring)
```

Modello Beta scelto per gli Stati Uniti:

```
> summary(usa)

  Call:
  betareg(formula = popularity ~ loudness + speechiness + duration + liveness +
  danceability + acousticness, data = usa)
```

```
Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-4.7012  0.0625  0.2984  0.5266  1.7584

Coefficients (mean model with logit link):
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.742e-02  2.070e-01   0.084 0.932932
loudness      4.657e-02  1.144e-02   4.070 4.71e-05 ***
speechiness  -7.713e-01  2.231e-01  -3.457 0.000545 ***
duration      1.512e-06  4.920e-07   3.073 0.002120 **
liveness     -3.416e-01  1.903e-01  -1.795 0.072655 .
danceability  4.185e-01  1.908e-01   2.193 0.028304 *
acousticness  2.450e-01  1.167e-01   2.100 0.035733 *

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)  2.00327    0.05704   35.12   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 43.58 on 8 Df
Pseudo R-squared: 0.02098
Number of iterations: 17 (BFGS) + 1 (Fisher scoring)
```

Modello Beta scelto per il Regno Unito:

```
> summary(uk)
Call:
betareg(formula = popularity ~ loudness + acousticness + speechiness + valence +
duration + liveness, data = uk)

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-5.7147 -0.0924  0.2608  0.5470  1.9391

Coefficients (mean model with logit link):
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.415e-01  1.504e-01   1.606 0.108327
loudness      5.604e-02  1.112e-02   5.039 4.68e-07 ***
acousticness  3.986e-01  1.113e-01   3.582 0.000341 ***
speechiness  -5.102e-01  2.071e-01  -2.464 0.013744 *
valence       3.443e-01  1.168e-01   2.948 0.003197 **
duration      1.297e-06  4.570e-07   2.838 0.004535 **
liveness     -3.819e-01  1.935e-01  -1.973 0.048469 *
```

```
Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)  2.71531    0.08482   32.01   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 94.72 on 8 Df
Pseudo R-squared: 0.02592
Number of iterations: 29 (BFGS) + 1 (Fisher scoring)
```

I panorama musicali in Australia, negli Stati Uniti e nel Regno Unito sono molto simili tra loro. In tutti e tre i generi più diffusi sono il pop, le sue varianti dance pop e hip pop, e il rap. Le variabili più significative sono infatti volume e ballabilità per l'Australia, *speechiness e loudness,* cosi come ballabilità e acustica per gli Stati Uniti e acustica e volume per il Regno Unito.

Modello Beta scelto per il Giappone:

```
> summary(japan)
Call:
betareg(formula = popularity ~ tempo + danceability, data = japan)

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-5.4241  0.0199  0.2762  0.5170  2.3863

Coefficients (mean model with logit link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.882212   0.250192  -3.526 0.000422 ***
tempo         0.004111   0.001152   3.568 0.000359 ***
danceability  0.517095   0.270185   1.914 0.055638 .

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)   2.8493     0.1198   23.79   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 37.37 on 4 Df
Pseudo R-squared: 0.009886
Number of iterations: 18 (BFGS) + 1 (Fisher scoring)
```

Modello Beta scelto per la Corea del Sud:

```
> summary(korea)

Call:
betareg(formula = popularity ~ acousticness + duration + danceability + liveness, data
= korea)

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-4.5121 -0.0421  0.3192  0.6358  1.9236

Coefficients (mean model with logit link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.089e-01  2.218e-01  -0.491 0.623580
acousticness  6.209e-01  1.097e-01   5.660 1.52e-08 ***
duration     -2.328e-06  6.106e-07  -3.814 0.000137 ***
danceability  5.906e-01  2.130e-01   2.772 0.005566 **
liveness     -4.940e-01  2.246e-01  -2.200 0.027841 *

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)   1.9095     0.0555   34.41   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 46.68 on 6 Df
Pseudo R-squared: 0.02533
Number of iterations: 14 (BFGS) + 2 (Fisher scoring)
```

Il modello Beta per il Giappone, i cui generi più ascoltati includono j-pop, k-pop e j-rock, mostra dei valori significativi solo per le variabili *danceability* e *tempo*. Dall'inizio degli anni '90 il termine *j-pop*, o "Japanese pop", viene comunemente utilizzato in Occidente per indicare la musica pop giapponese, la quale comprende diversi generi, dai più soft, pop, r&b e jazz, al j-rock, j-metal e visual key. Diversamente il k-pop tende a sonorità più lontane dal rock e metal e più riconducibili al pop e all'hip-hop. Inoltre, solitamente i brani k-pop sono accompagnati da coreografie complesse. La ballabilità e l'acustica sono infatti variabili significative per il modello Beta scelto per la Corea del Sud.

Modello Beta scelto per il Sud Africa:

```
> summary(africa)
```

```
Call:
betareg(formula = popularity ~ duration + acousticness + loudness + energy +
danceability + energy:valence, data = africa)

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-5.2151 -0.0603  0.3133  0.5920  1.9849

Coefficients (mean model with logit link):
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.441e+00  2.616e-01   5.509 3.61e-08 ***
duration       -1.764e-06  3.209e-07  -5.498 3.83e-08 ***
acousticness    3.808e-01  1.328e-01   2.868  0.00413 **
loudness        5.838e-02  1.168e-02   4.998 5.79e-07 ***
energy         -8.332e-01  2.876e-01  -2.897  0.00376 **
danceability   -4.028e-01  1.962e-01  -2.054  0.04001 *
energy:valence  3.639e-01  2.183e-01   1.667  0.09557 .

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)   2.3414     0.0692   33.84   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 88.56 on 8 Df
Pseudo R-squared: 0.06955
Number of iterations: 30 (BFGS) + 1 (Fisher scoring)
```

Il Sudafrica ha una vibrante scena musicale popolata da un'ampia varietà di generi e stili. Nel corso degli anni, l'ambiente politico del paese ha avuto una grande influenza sulla sua musica, portando alla nascita di generi originali come kwaito, che deriva da una fusione di musica house, afropop ed elementi hip hop, jazz africano e mbube, forma di musica vocale sudafricana tradizionalmente eseguita a cappella. Le variabili *speechiness* e *instrumentalness* non influenzano in modo significativo l'indice popolarità, vengono ascoltati infatti sia brani vocali sia strumentali. Le variabili più influenti sono invece il volume, l'energia e l'acustica, caratteristiche fondamentali per tutti i generi più ascoltati nel paese.

Modello Beta scelto per la Spagna:

```
> summary(spain)
```

```
Call:
betareg(formula = popularity ~ danceability + loudness + speechiness + liveness, data
= spain)

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-6.6362 -0.1484  0.1375  0.4817  3.1103

Coefficients (mean model with logit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.91068    0.20254   4.496 6.91e-06 ***
danceability -0.61237    0.22807  -2.685  0.00725 **
loudness      0.02459    0.01248   1.970  0.04882 *
speechiness  -0.52293    0.26224  -1.994  0.04614 *
liveness      0.35087    0.22359   1.569  0.11658

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)   3.5704     0.1285   27.79   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 151.6 on 6 Df
Pseudo R-squared: 0.01071
Number of iterations: 20 (BFGS) + 2 (Fisher scoring)
```

Modello Beta scelto per il Brasile:

```
> summary(brazil)

Call:
betareg(formula = popularity ~ valence + acousticness + duration + tempo, data =
brazil)

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-4.9365  0.0165  0.3098  0.5673  1.7773

Coefficients (mean model with logit link):
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   8.445e-01  2.031e-01   4.158 3.21e-05 ***
valence      -1.005e+00  1.494e-01  -6.727 1.73e-11 ***
acousticness -4.503e-01  1.382e-01  -3.257  0.00112 **
duration      8.997e-07  5.088e-07   1.768  0.07703 .
tempo        -1.821e-03  1.165e-03  -1.563  0.11795
```

```
Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)  2.29012    0.07992   28.66   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 39.79 on 6 Df
Pseudo R-squared: 0.03988
Number of iterations: 16 (BFGS) + 2 (Fisher scoring)
```

Nel dicembre del 2019, Billboard ha riferito che la musica latina era il quarto genere più ascoltato negli Stati Uniti su DSP, come Spotify, e il terzo per lo streaming video su YouTube. Dal reggaeton e la vibrante salsa al messicano regionale, dalla samba brasiliana e dalla rilassata bossanova all'appassionato tango argentino, la musica latina comprende i generi più differenti (Audio Network UK, 2020).

Nel panorama brasiliano individuiamo come generi più diffusi: la samba, icona dell'identità nazionale brasiliana, composta principalmente da strumenti a percussione che suonano ritmi sincopati e la bossanova, influenzate dal jazz e concentrate su emozioni personali, quali l'amore, il desiderio e la natura.

Il Brasile presenta infatti valori significativi per l'acustica e per la positività, mentre la ballabilità risulta significativa nel panorama musicale spagnolo e non in quello brasiliano, il quale include generi sia ballabili (samba) sia rilassati (bossanova).

Capitolo 5

L'industria musicale e i Big Data

Il seguente capitolo nasce da una riflessione sull'effetto dei big data nell'industria musicale. L'enorme mole di dati ricavabile da social e streaming sta offrendo all'industria una serie di nuovi tool per promuovere nuove tracce o generi musicali tramite modelli basati sullo studio di utenti delle piattaforme streaming, frequentatori dei concerti e più in generale tutti i consumatori di musica. In seguito si riporta un esempio che mostra come le aziende moderne dell'industria musicale implementino l'uso dei big data per garantire il loro successo nel mercato e come possiamo utilizzare le analisi dei capitoli precedenti per comprenderlo al meglio.

5.1 La *Korean Wave* e l'occidentalizzazione del K-Pop

La *Korean Wave* (o *Hallyu* in coreano), "onda coreana", è un neologismo che indica l'incremento della diffusione globale della cultura di massa sudcoreana, un fenomeno socio-culturale che ha permesso alla Corea del Sud di diventare, negli ultimi anni, uno degli epicentri della cultura pop a livello mondiale.
Tra tutti i gruppi idol sudcoreani che hanno debuttato negli Stati Uniti, i BTS hanno senza dubbio avuto il maggior successo nell'irrompere nell'America mainstream. Lo Hyundai Research Institute ha stimato che la band ha raccolto più di 3,6 miliardi di dollari ogni anno per l'economia sudcoreana, equivalente al contributo di 26 aziende di medie dimensioni. Nel 2017 quasi il 7% di tutti i visitatori registrati in Corea del Sud ha affermato che i BTS erano la motivazione principale per visitare il paese. Secondo la Billboard Hot 100, sono inoltre il primo gruppo, dopo i Beatles, con sei canzoni al primo posto nella Hot 100 in poco più di un anno (Bartlett, 2022).
Tuttavia sono in molti ad affermare che per poter raggiungere un tale successo globale la band sia dovuta scendere a compromessi sul proprio stile musicale. L'uscita di "Dynamite" nell'Agosto 2020, la prima canzone dei BTS completamente in inglese, fece inizialmente pensare ad un modo strategico per la band di pubblicizzarsi ad un pubblico più ampio, oltre la Corea del Sud e i territori asiatici. La nuova traccia, infatti,

raggiunse immediatamente il primo posto della Billboard Hot 100 nel Settembre 2020. Tuttavia, l'uscita di due nuovi singoli esclusivamente in inglese, "Butter", nel Maggio 2021, e "Permission to Dance", pubblicato il Luglio dello stesso anno, scatenò reazioni contrastanti tra i fans degli *idol* coreani (Chang, 2022). Alcuni definirono la band "troppo occidentalizzata", mentre altri compresero l'efficacia della tecnica di marketing vantaggiosa per soddisfare un pubblico internazionale. I testi in inglese hanno indubbiamente contribuito alla occidentalizzazione della band, ma il principale oggetto di critiche per gli ultimi brani è la musica stessa. II BTS sono emersi come un gruppo incentrato sull'hip hop coreano, ma negli ultimi anni hanno lentamente modificato la loro musica fino a produrre quasi esclusivamente pop mainstream.

Se si guarda alla Figura 4.3 si notano subito le principali differenze nei valori delle Audio Features tra Stati Uniti e Corea del Sud. Di seguito i valori dei tre brani più recenti in inglese, "Dynamite" (2020), "Butter" (2021) e "Permission to Dance "(2021) a confronto con tre tra i più popolari degli anni precedenti al successo globale, "Dope" (2015), "Blood, Sweat and Tears" (2016), "Fake Love" (2018).

track_ name	**dancea bility**	**energy**	**loudness**	**speechi ness**	**acoustic ness**	**instrumen talness**	**liveness**	**valence**	**tempo**	**popul arity**
"Dynamite"	0,746	0,765	-4,41	0,0993	0,0112	0	0,0936	0,737	114,044	86
"Butter"	0,759	0,459	-5,187	0,0948	0,00323	0	0,0906	0,695	109,997	79
"Permission to Dance"	0,702	0,741	-5,33	0,0427	0,00544	0	0,337	0,646	124,925	81
"Dope"	0,595	0,89	-3,109	0,245	0,0486	0	0,32	0,62	154,071	68
"Blood, Sweat and Tears"	0,584	0,885	-3,571	0,104	0,0244	0	0,357	0,605	92,9	75
"Fake Love"	0,557	0,719	-4,515	0,0371	0,00267	0	0,306	0,345	77,502	77

Tabella 5.1 Audio Features di sei brani dei BTS (*13/10/2022).*

	danceability	energy	speechiness	acousticness	instrumentalness	liveness	valence	duration	loudness	tempo
Korea	0,64446	0,64844	0,09477	0,26170	0,02333	0,17269	0,50909	206385,87186	-5,77037	119,84837
United States	0,65753	0,60150	0,13101	0,25376	0,03144	0,18296	0,47047	199837,22823	-7,12303	122,25655

Figura 5.1 Valori medi Audio Features Stati Uniti e Corea del Sud.

Osserviamo come le differenze nei valori audio delle canzoni meno e più recenti riflettano rispettivamente i valori medi associati ai brani più popolari in Corea del Sud e negli Stati Uniti.

5.2 Regressione Beta che include variabile dicotomica "paese"

Per comprendere al meglio le differenze tra le preferenze musicali di questi due paesi andiamo a costruire un modello beta a cui aggiungiamo la variabile dicotomica paese. Quest'ultima è una variabile concomitante che può assumere esclusivamente due valori, in questo caso "South Korea" e "United States", riconducibile ai valori 0 e 1. Riportiamo il modello Beta che include i brani sia degli Stati Uniti che della Corea del Sud e osserviamo come il coefficiente della variabile `country_dicotomica` presenta un valore alto che va quindi senza dubbio a influenzare la popolarità.

```
> summary(korea_usamodel)

Call:
"betareg(formula =  popularity ~ country_dicotomica + speechiness + danceability +
acousticness + loudness + liveness data = korea_usa)"

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-4.7586  0.0281  0.3099  0.5862  1.9057

Coefficients (mean model with logit link):
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -0.34299    0.11806  -2.905 0.003671 **
country_dicotomica  0.35163    0.04088   8.601  < 2e-16 ***
speechiness        -0.67339    0.18409  -3.658 0.000254 ***
danceability        0.54532    0.14217   3.836 0.000125 ***
acousticness        0.41616    0.08551   4.867 1.14e-06 ***
loudness            0.02428    0.00870   2.790 0.005264 **
liveness           -0.39614    0.14788  -2.679 0.007388 **

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)  1.84489    0.03723   49.55   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 74.57 on 8 Df
Pseudo R-squared: 0.02345
Number of iterations: 17 (BFGS) + 1 (Fisher scoring)
```

Inoltre costruiamo un modello in cui la variabile risposta è la variabile dicotomica utilizzando un modello logistico tramite la funzione glm(·).

```
> summary(mlogit)
Call:
glm(formula = country_dicotomica ~ danceability + energy + speechiness +
    acousticness + liveness + tempo + loudness + energy:valence +
    popularity, family = binomial, data = ku)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6921  -1.0290   0.4777   1.0377   2.2401

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -4.633117   0.497632  -9.310  < 2e-16 ***
danceability    1.042491   0.297815   3.500 0.000464 ***
energy          1.307516   0.452886   2.887 0.003888 **
speechiness     2.483245   0.374124   6.637 3.19e-11 ***
acousticness   -0.984105   0.189017  -5.206 1.93e-07 ***
liveness        0.951446   0.288902   3.293 0.000990 ***
tempo           0.003466   0.001295   2.677 0.007425 **
loudness       -0.310459   0.023395 -13.270  < 2e-16 ***
popularity      1.670488   0.158873  10.515  < 2e-16 ***
energy:valence -1.245918   0.307327  -4.054 5.03e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4748.6  on 3426  degrees of freedom
Residual deviance: 4204.6  on 3417  degrees of freedom
AIC: 4224.6

Number of Fisher Scoring iterations: 4
```

Osserviamo che, come si poteva intuire dalla Figura 5.1, le caratteristiche audio che influenzano maggiormente la popolarità di un brano per entrambi i paesi Corea del Sud e Stati Uniti sono l'energia, la loquacità, seguite dalla ballabilità, dalla *liveness* e dall'acustica. Le prime due presentano valori maggiori per i brani degli Stati Uniti in cui sono popolari brani più energici e parlati, così come più ballabili, mentre in Corea del Sud si preferiscono brani con valori più elevati in acustica e volume.

5.3 Le criticità nell'uso dei Big Data per le industrie creative

Sembra che in questo caso, invece che adattare i dati alla musica tramite la scelta di una strategia di marketing ottimale, sia stata la musica ad adattarsi ai dati. Proprio per questo motivo oggigiorno viene spesso criticata l'importanza che viene attribuita ai Big Data nelle industrie creative e in particolare in quella musicale. Tuttavia il successo della band coreana negli Stati Uniti ha permesso loro di condividere la loro musica e i loro messaggi con un pubblico sempre più ampio. Al di là dell'enorme impatto economico, i BTS hanno infatti anche svolto un ruolo importante nel proiettare un'immagine positiva della Corea del Sud all'estero, interagendo con i principali leader mondiali. Nel 2018 la band si è unita al Segretario generale delle Nazioni Unite, Antonio Guterres, e ad altri leader mondiali per lanciare una partnership volta a responsabilizzare i giovani e per promuovere la non violenza all'estero. Nel 2022, la Casa Bianca ha invitato i BTS a parlare con il presidente degli Stati Uniti, Joe Biden, e la vicepresidente, Kamala Harris, sui crimini d'odio anti-asiatici, l'inclusione asiatica e la diversità (Bartlett, 2022). Sicuramente un impatto positivo che non sarebbe stato possibile senza un'analisi accurata del mercato e una studiata strategia di marketing basata sui dati.

Da un lato è certamente un fatto innegabile che lo scenario mainstream si sia drasticamente appiattito negli ultimi 10-15 anni e che il motivo stia in buona parte nell'adozione dei modelli di marketing offerti dall'analisi dei Big Data; tuttavia gli antidoti all'omologazione che minaccerebbe di lasciare artisti dotati nell'ombra ci sono. Lo sfruttamento dei dati e degli algoritmi può essere infatti utilizzato per scovare esecutori con un piccolo seguito, ma un grande potenziale – modello adottato dalla Snafu Records e da altre case discografiche (Heitner, 2018). Inoltre poiché i dati relativi all'industria musicali sono disponibili gratuitamente, l'analisi dei dati non deve necessariamente andare a influenzare la creatività dell'artista, ma semplicemente consentire agli artisti stessi di costruire strategicamente il loro pubblico. Quando i Brooklyn's Cigarettes After Sex hanno scoperto tramite i dati di Spotify che la loro musica veniva trasmessa in streaming in città asiatiche come Seoul, in Corea del Sud, la band è andata in tournée in quei territori con promozioni su misura per i singoli mercati. Questo approccio li ha aiutati a creare una base di fans globale di oltre 4 milioni di ascoltatori Spotify mensili (Mccabe, 2019).

Conclusione

La musica ha da sempre avuto un ruolo fondamentale nella vita dell'uomo. Ogni cultura utilizza la musica in modi differenti, ma questa rimane un mezzo per comunicare, condividere un'emozione, una storia, un messaggio, nella speranza di raggiungere più persone possibili e abbattere qualsiasi barriera comunicativa. Appare indiscutibile dover considerare quindi l'importanza e la necessità di avere a disposizione dati e informazioni per poter analizzare e comprendere al meglio la realtà musicale in cui viviamo, non solo per motivi di interesse economico, ma anche sociale.

Perché ciò sia possibile sono necessari i mezzi adatti e, personalmente, credo che i Big Data possano essere una risorsa conoscitiva fondamentale. Imparare a leggere, interpretare e dare valore ai dati è diventato oggigiorno un requisito fondamentale per lavorare nell'industria musicale, e in generale per qualsiasi industria nel campo dell'intrattenimento.

L'analisi dei Big Data per il settore musicale, infatti, non si limita certamente ai soli dati ricavati dai servizi di streaming musicali, quali Spotify e Apple Music, ma anche a quelli ricavabili da piattaforme strettamente connesse con questo settore, primi fra tutti i Social Network, in particolare YouTube, Instagram e TikTok. L'ascesa di pop star "self-made", quali Doja Cat e Lil Nas X, che abbiamo potuto osservare negli ultimi anni, è stata infatti possibile grazie allo studio dei social media e a strategie di marketing legate ai trend del momento. Lil Nas X, ad esempio, ha sfruttato in maniera esemplare il boom su YouTube dei video beat Lo-fi. Quando ha infatti pubblicato il video musicale per la sua canzone "Montero", ne ha pubblicato sei versioni, inclusa una versione beat Lo-fi su cui studiare, che ha ricevuto milioni di visualizzazioni e ha contribuito a mantenere il brano e il suo messaggio in primo piano, facendo salire "Montero" in cima alle classifiche musicali di YouTube in 20 nazioni (Avallone, 2018).

Non da sottovalutare anche l'influenza che dati ricavati da altre tipologie di media, quali serie tv e film, hanno sul mercato musicale. La hit «Running Up That Hill» uscita nel 1985, è stata utilizzata come colonna sonora di Stranger Things e grazie alla popolarità della serie tv, il brano è tornato prepotentemente in vetta alle classifiche, diventando la canzone più ascoltata in streaming in tutto il mondo e facendo guadagnare all'artista circa 2 milioni e trecentomila dollari.

Queste nuove strategie di marketing, basate sui dati e l'analisi dei numeri, hanno promosso una crescita enorme del mercato musicale. Il 22 marzo l'IFPI (International Federation of Phonographic Industry), l'organizzazione che rappresenta gli interessi dell'industria discografica mondiale, ha pubblicato il suo

annuale “Global Music Report”. La versione di quest′anno del rapporto sembra particolarmente promettente in quanto, secondo l′IFPI, nel 2021 i ricavi della musica registrata hanno segnato il loro settimo anno consecutivo di crescita, raggiungendo il livello di 25,9 miliardi di dollari, corrispondente a un aumento del 18,5% rispetto al 2020 e il più alto livelli di fatturato mai raggiunti dal mercato globale della musica registrata in questo millennio (Anote Music, 2022). Particolarmente rilevante è come l’analisi dei Big Data abbia permesso a paesi con mercati musicali piccoli, quali l’Africa, di avviare la loro crescita e di farsi conoscere nel resto dell’Europa. Come ha affermato Iseunife Ajayi, specialista in comunicazioni creative, “La rapida ascesa della scena musicale nigeriana, negli ultimi tempi, è stata il risultato di un marketing intenzionale e basato sui dati nel corso degli anni”. Da dati ricavati dai servizi di streaming musicale quali Spotify ai dati ricavati dai Social Media, i Big Data hanno sicuramente contribuito aumento delle vendite di musica nel 2021 del 9,6% che ha registrato l’Africa sub sahariana e a quello del 35% che la regione MENA (Middle East and North Africa) ha registrato, rendendola la regione con il tasso di crescita più rapido a livello globale (IFPI Global Music Report, 2022).

In conclusione molte sono le opinioni sul rapporto tra musica e Big Data, ma io credo che finché questo strumento permette ai musicisti di tutto il mondo di raggiungere il giusto target, comunicare il proprio messaggio, e permettere alla musica, di qualsiasi paese, di diffondersi in tutto il mondo, non può che essere uno strumento fondamentale e positivo.

Bibliografia

Magro E. *Il modello di regressione con variabile risposta Beta*, (2011).

Samyak J., Parth C., Sarthak J. *Predict-the-Hit: Prediction of Hit Songs based on Multimodal Data* in "International Journal of Scientific and Research Publications", volume 12, Issue 9, (2022, Settembre).

Sciandra M, Spera I. C. *Ho perso le parole: come ritrovarle con la sentiment analysis,* (2020, Giugno).

Sitiografia

Avallone A., *TikTok sta salvando o distruggendo l'industria musicale?,* (2018, Maggio). https://kmagazine.it/it/trend/tik-tok-audio-industria-musicale/ [consultato il 17 Gennaio 2023]

Bartlett J., *Domestic and Global Political Impacts of K-Pop: BoA, BTS, and Beyon,* The Diplomat, (2022, Giugno). https://thediplomat.com/2022/06/domestic-and-global-political-impacts-of-k-pop-boa-bts-and-beyond/#:~:text=The%20Hyundai%20Research%20Institute%20estimated,of%2026%20mid%2Dsized%20companies [consultato il 17 Gennaio 2023]

Chang D., *BTS' new music becomes 'too Westernized' as recent singles are all in English,* (2022, Settembre). https://shhsaccolade.com/8356/ae/bts-new-music-becomes-too-westernized-as-recent-singles-are-all-in-english/ [consultato il 14 Dicembre 2022]

Different Types of Latin Music Genres You Need To Know, Audio Network UK, (2020, Maggio). https://www.audionetwork.com/content/the-edit/inspiration/different-types-latin-music-genres [consultato il 5 Aprile 2023]

Heitner D., *Big Data Is Revolutionizing the Music Industry. Here Are the Lessons for Your Business,* (2018, Maggio). https://www.inc.com/darren-heitner/big-data-is-revolutionizing-music-industry-here-are-lessons-for-your-business.html [consultato il 7 Marzo 2023]

IFPI Global Music Report 2022, IFPI, (2022, Maggio). https://www.ifpi.org/resources/ [consultato il 17 Gennaio 2023]

Ingham T., *With $15bn in revenue, 2021 was the US record industry's biggest EVER year (kind of...),* Music Business Worldwide, (2022, Marzo). https://www.musicbusinessworldwide.com/with-15bn-revenue-2021-was-the-us-record-industrys-biggest-ever-year-kind-of/ [consultato il 25 Febbraio 2023]

Mccabe A., *Why Big Data Has Been (Mostly) Good for Music,* (2019, Novembre). https://www.wired.com/story/big-data-music/ [consultato il 25 Febbraio 2023]

Spotify for developers. Get Track's Audio Features. https://developer.spotify.com/documentation/web-api/reference/get-audio-features [consultato il 17 Dicembre 2022]

Stochasticism, *Accessing Spotify's API Using R,* (2020, Maggio). https://medium.com/swlh/accessing-spotifys-api-using-r-1a8eef0507c [consultato il 15 Dicembre 2022]

The global music industry reaching new peaks - Everything you need to know, Anote Music, (2022, Aprile). https://blog.anotemusic.com/the-global-music-industry-reaching-new-peaks-everything-you-need-to-know-at-a-glance [consultato il 16 Dicembre 2022]

Yelexa, *Spotify Weekly Top 200 Songs Streaming Data,* (2022, Luglio). https://www.kaggle.com/datasets/yelexa/spotify200?select=final.csv [consultato il 27 Novembre 2022]

Yelexa, *The Modern A&R Experience: Which Global Artist Will You Sign?,* (2022, Agosto). https://public.tableau.com/app/profile/yejielee/viz/TheModernARExperienceWhichGlobalArtistWillYouSign/ar_dashboard [consultato il 27 Novembre 2022]