# Лабораторна робота №3

“Регресійний аналіз”

*Підготували:*

* *Руслан Дюбакін*
* *Петриченко Нікіта*
* *Пріхно Ілля*
* *Резниченко Єлизавета*

*Група: КМ-01*

## Вступ

У рамках даної лабораторної роботи була побудована регресійна модель для дослідження причинно-наслідкового зв’язку між популярністю та трьох характеристик треків: energy, danceability, explicit.

# Зміст

[**Лабораторна робота №3 1**](#_zdokrnwpxg11)

[**Вступ 1**](#_nigdwc97siog)

[**Зміст 2**](#_fe3ltv5do0ky)

[**Датасет з наявністю нульової популярності 3**](#_kwb9cifie8l3)

[Росподіли 3](#_q90lzdbjovra)

[Моделі 5](#_ey4bgjr3elsj)

[Додавання контрольованих змінних 5](#_ben4wmwsznsq)

[Додавання поліномів вищого порядку 6](#_86dlg6otkthd)

[Додавання факторів взаємодії 9](#_6nwxzby8sd8k)

[Тести на статистичну значущість груп коефіцієнтів 10](#_6t58htzbhqxj)

[**Датасет без нульової популярності 15**](#_17zjukc5n5y5)

[Розподіли 15](#_4z8ine360kn5)

[Моделі 18](#_8szs4ptufns7)

[Додавання контрольованих змінних 18](#_yjymf2qglfla)

[Додавання поліномів вищого порядку 19](#_yozoday3mz5m)

[Додавання факторів взаємодії 22](#_li0li27mciz)

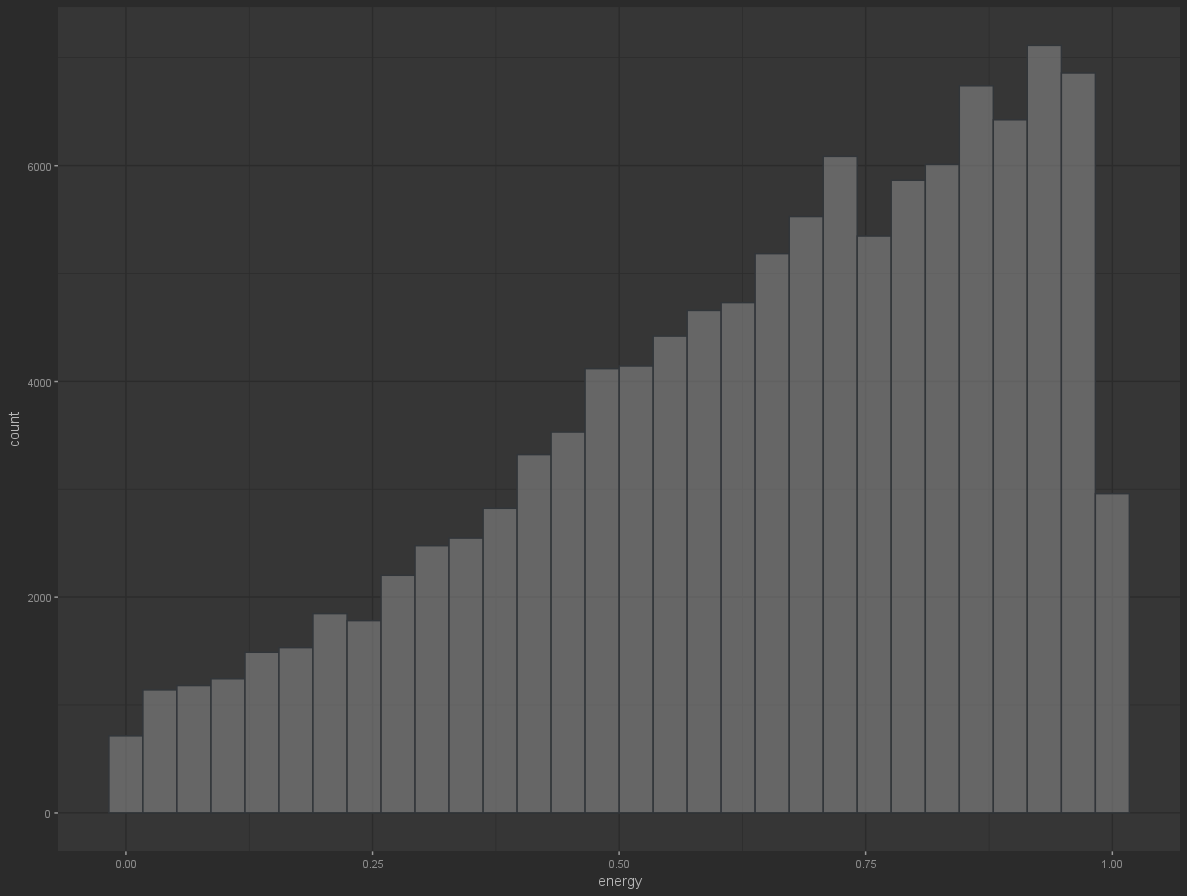
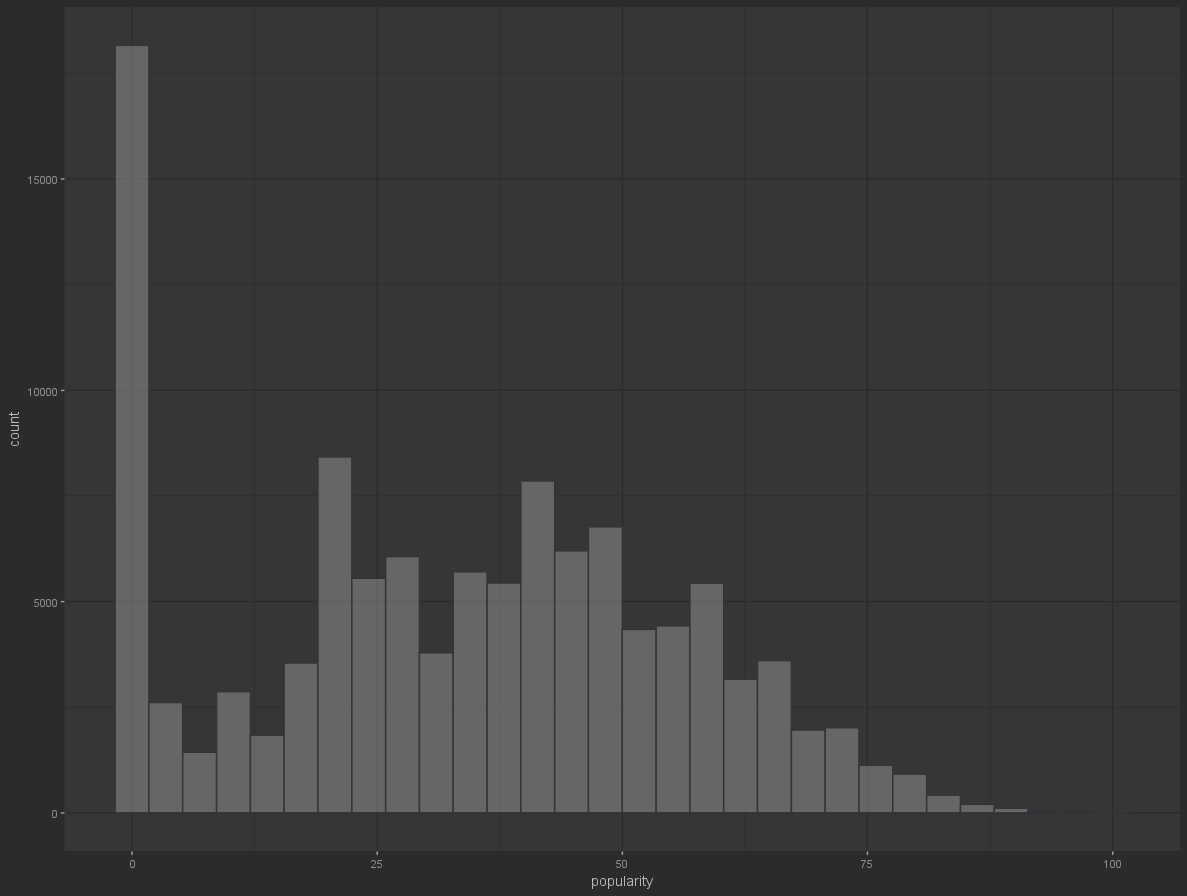
[Тести на статистичну значущість груп коефіцієнтів 23](#_ficxwshvi94e)

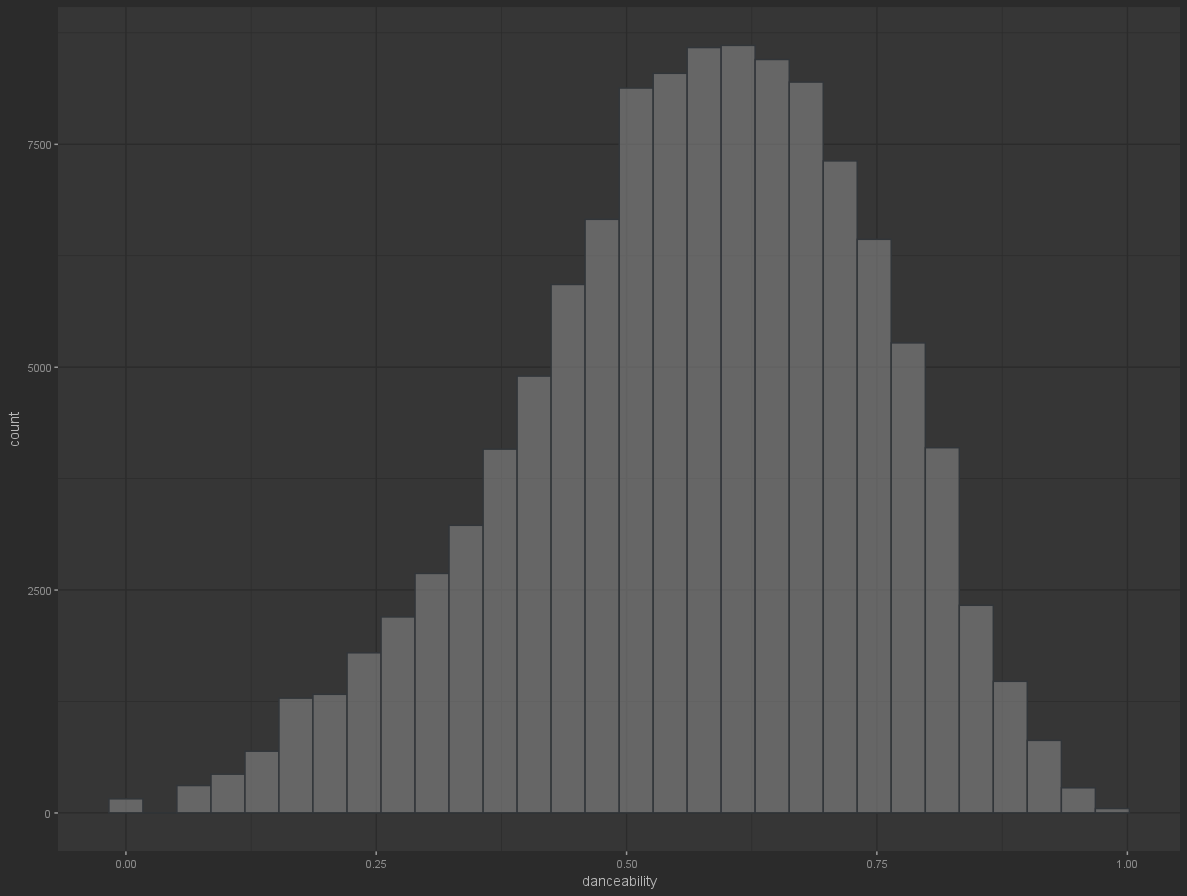
[**Висновки 27**](#_3r0by51w1klj)

[**Список літератури 28**](#_ahcfu55s60cx)

# Датасет з наявністю нульової популярності

## Росподіли

Змінні, які ми обрали для аналізу, мають наступні розподіли:



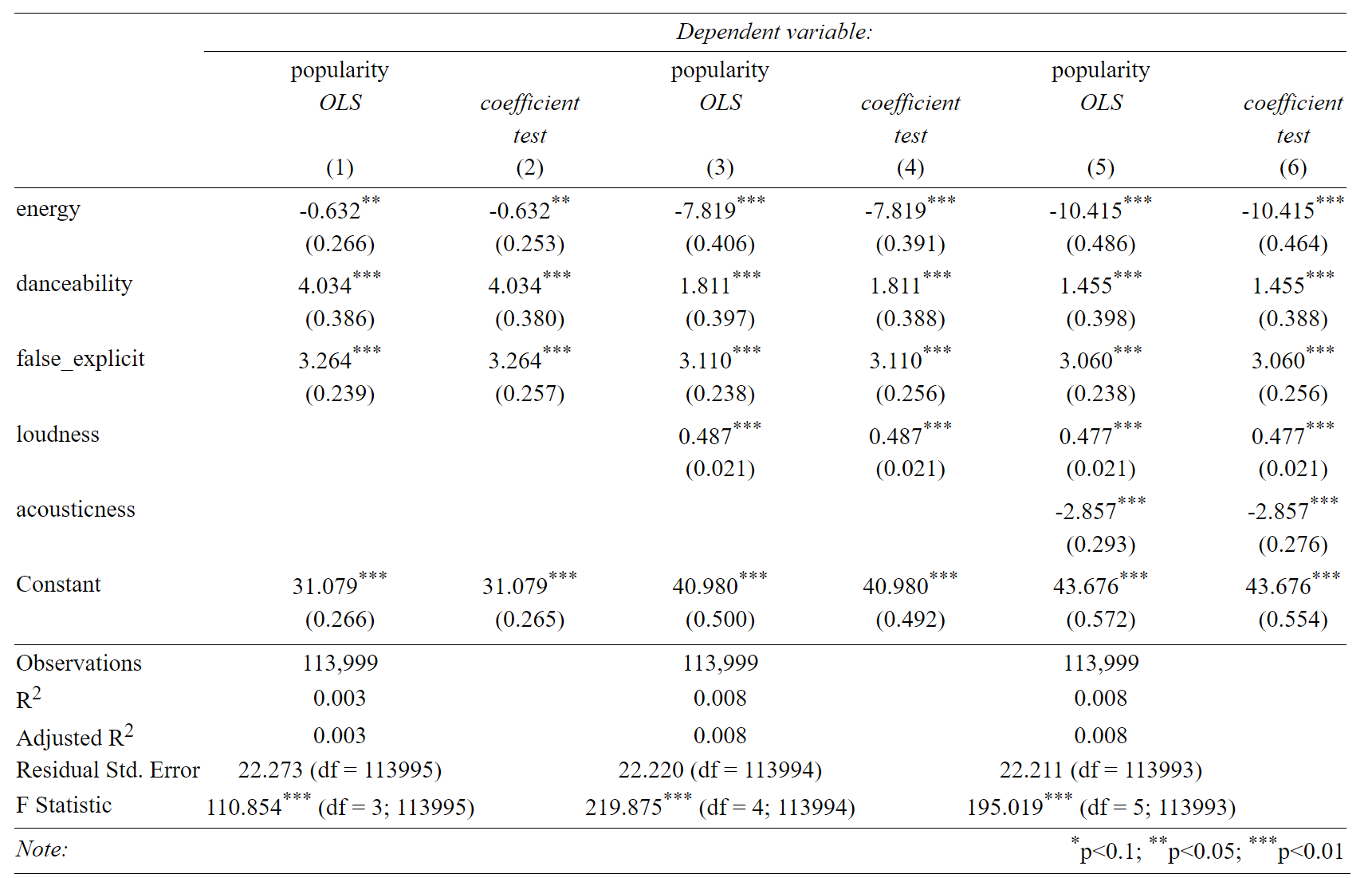
За графіком кореляції між змінними можна побачити, що взяті нами змінні до дослідження мають доволі високі кореляції з деякими іншими змінними:

* energy - loudness
* energy - acousticness
* danceability - valence

Це означає, що при неврахуванні цих змінних зміщення (OVB) можуть призвести до порушення умови про нульове умовне сподівання похибки.

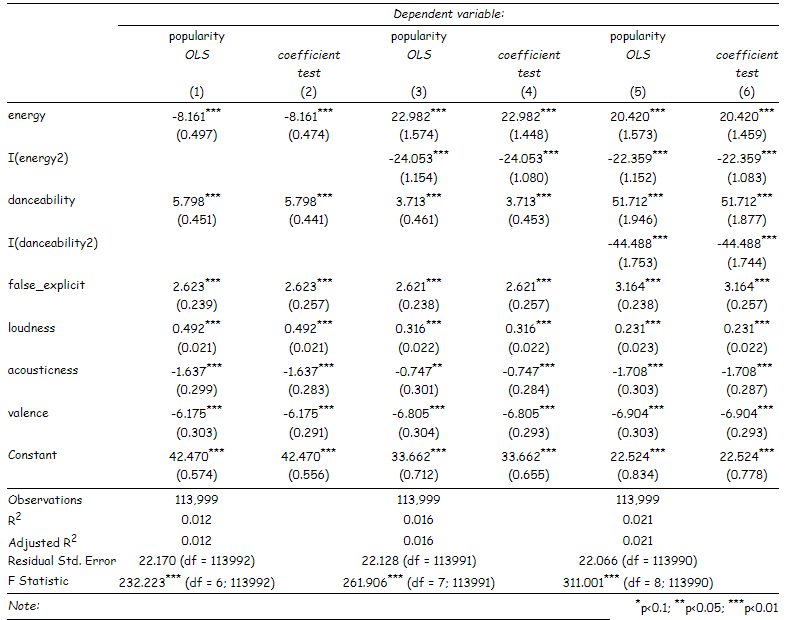
## Моделі

### Додавання контрольованих змінних

За таблицею розглянемо побудовані модели. Бачимо, що всі коефіцієнти є статистично значущими. За першою базовою моделлю бачимо, що коефіцієнти при досліджуваних змінних є досить малими, мають досить малий вплив на популярність.

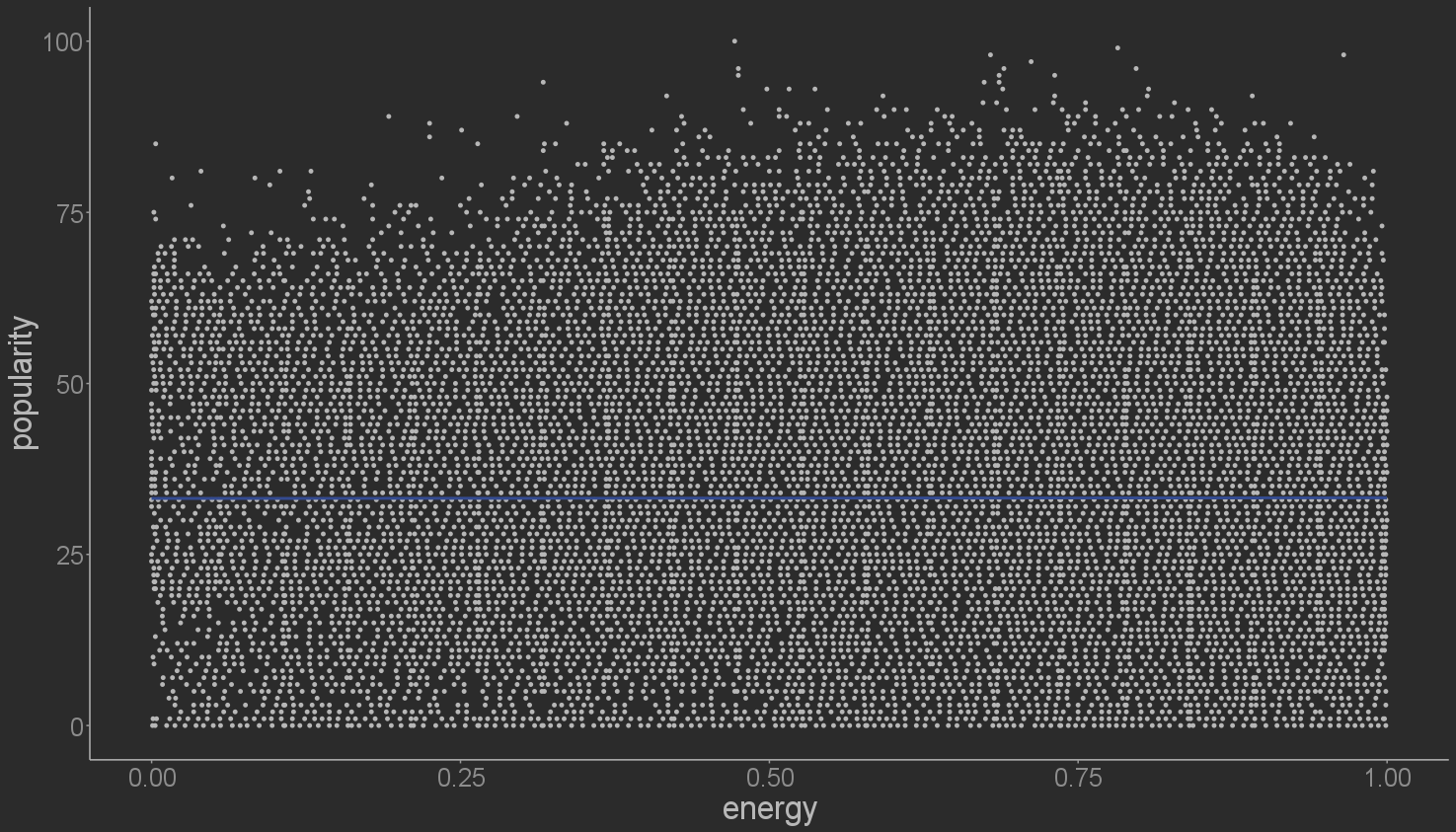
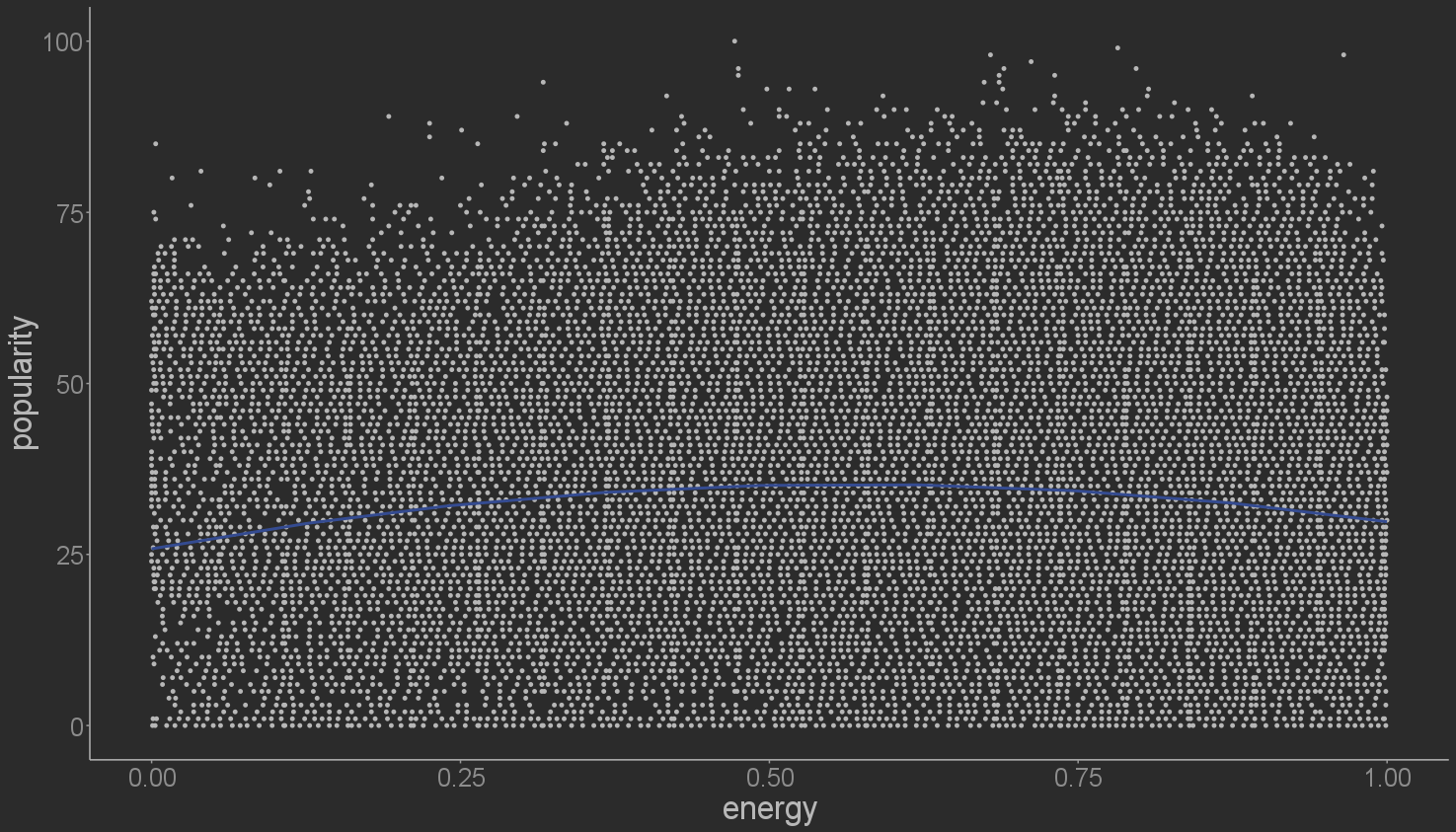
При додаванні контрольованих змінних, що корелюють з energy, бачимо зростання впливу energy на popularity.

### Додавання поліномів вищого порядку

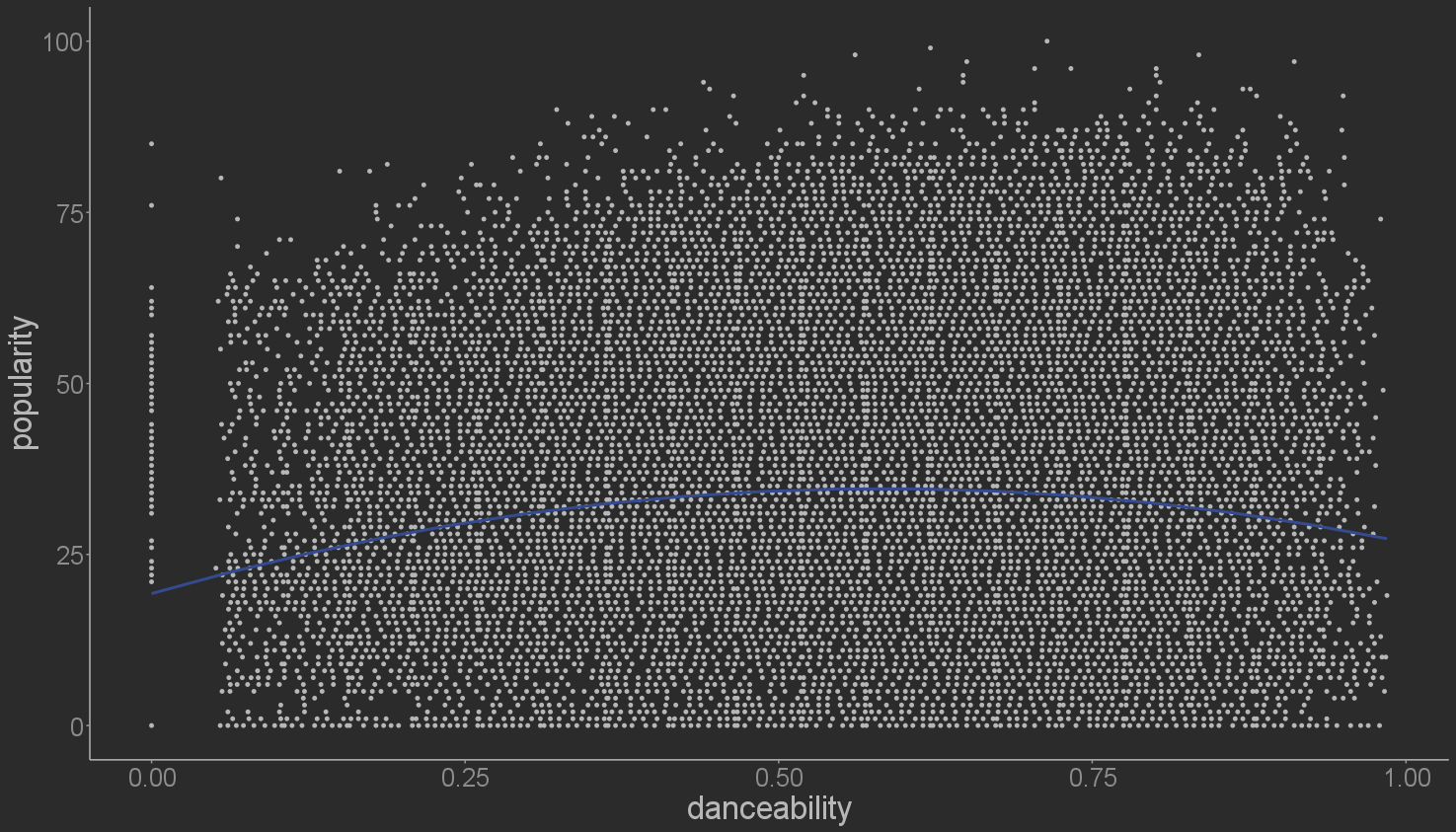
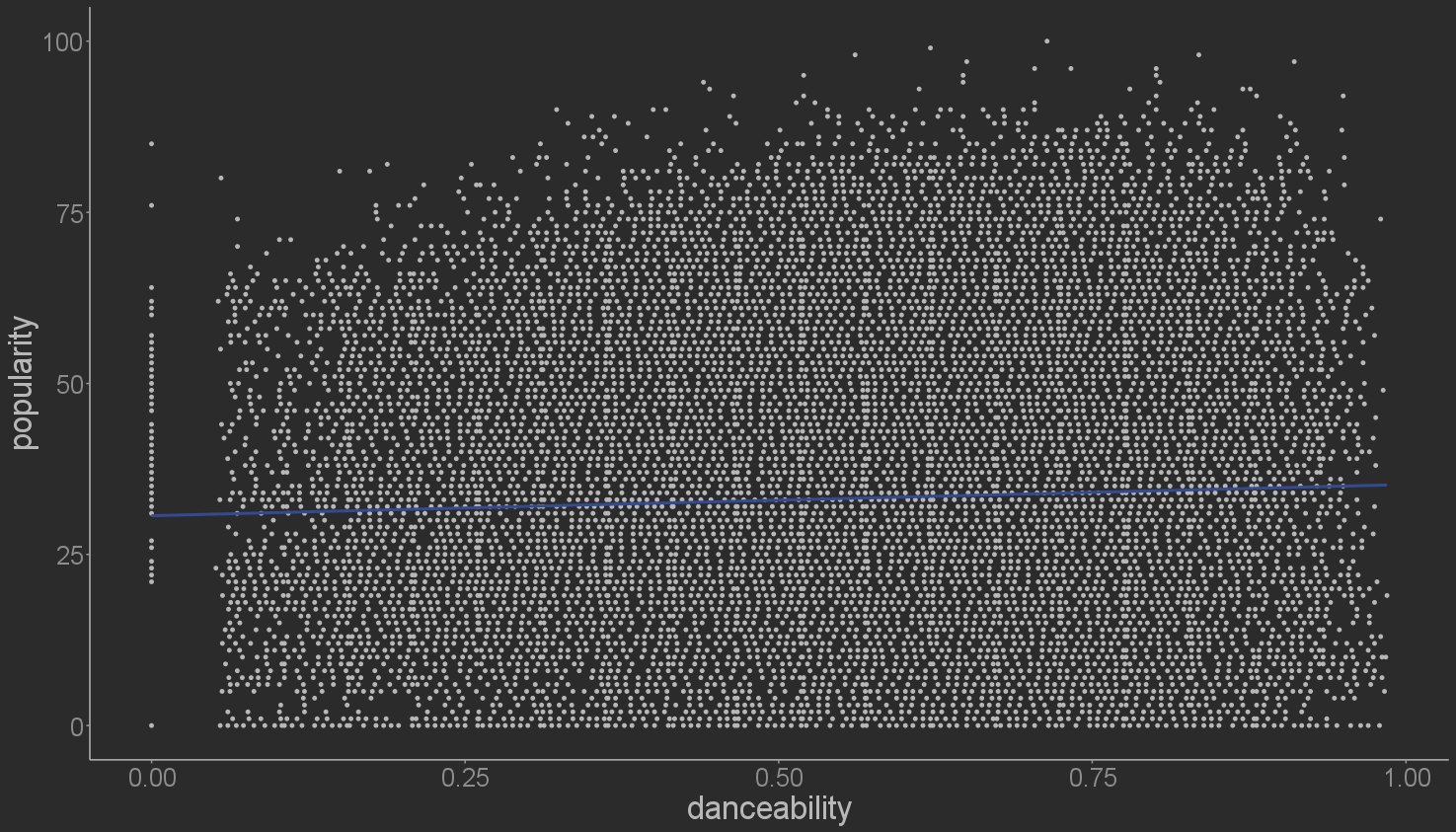


При додаванні контрольованих змінних, що корелюють з danceability, спостерігаємо підвищення впливу danceability на популярність.

Розглянемо графічно модель popuplarity ~ energy та popularity ~ energy + energy^2.

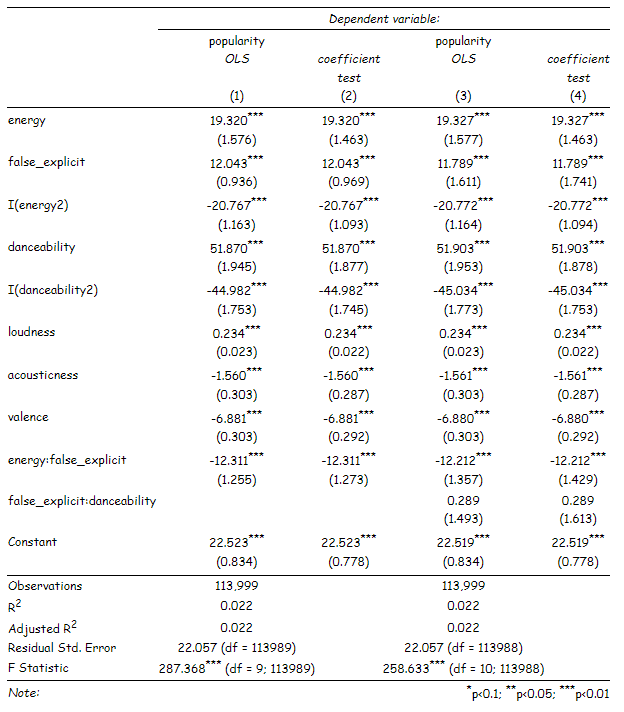


Можемо додати energy^2 до нашої модели. Бачимо, що вплив energy доволі зріс. Усі коефіцієнти біля доданих змінних зберігають свою статистичну значущість.

Розглянемо графічно модель popularity ~ danceability та popularity ~ danceability + danceability^2.

Можемо додати danceabillity^2 до нашої модели. Бачимо, що вплив danceabillity доволі зріс. Усі коефіцієнти біля доданих змінних зберігають свою статистичну значущість. При цьому коефіцієнти при energy та false\_explicit мають несуттєві зміни.

### Додавання факторів взаємодії



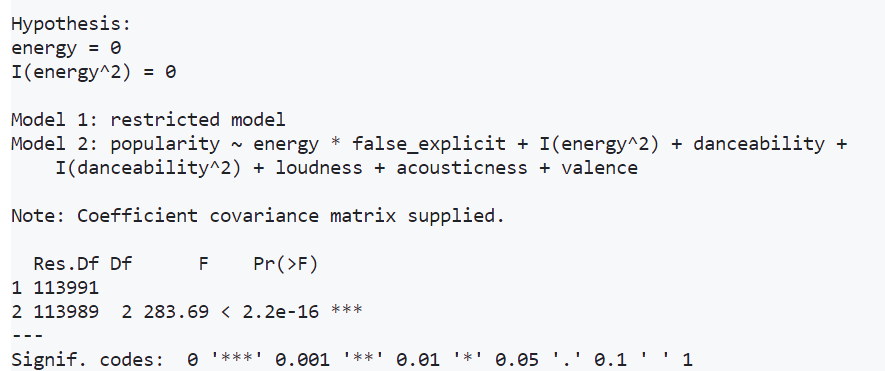
Спробуємо додати до модели фактор взаємодії між false\_explicit та energy. У таблиці вище бачимо, що вплив false\_explicit доволі зріс, при чому інші змінні не мають занадто суттєвих змін.

Після додавання фактору взаємодії між false\_explicit та danceabillity суттєвих змін у коефіцієнтах не спостерігається, а коефіцієнт біля самого фактору взаємодії є статистично незначущим. Отже, немає особливого сенсу додавати цей фактор до модели.

Отже, після сьомої модифікації модели бачимо, що коефіцієнти біля ключових змінних не мають суттєвих змін, тобто можна вважати модель стійкою.

### Тести на статистичну значущість груп коефіцієнтів

Проведемо тести на статистичну значущість груп коефіцієнтів.



Гіпотези, що перевіряються, виглядають наступним чином:

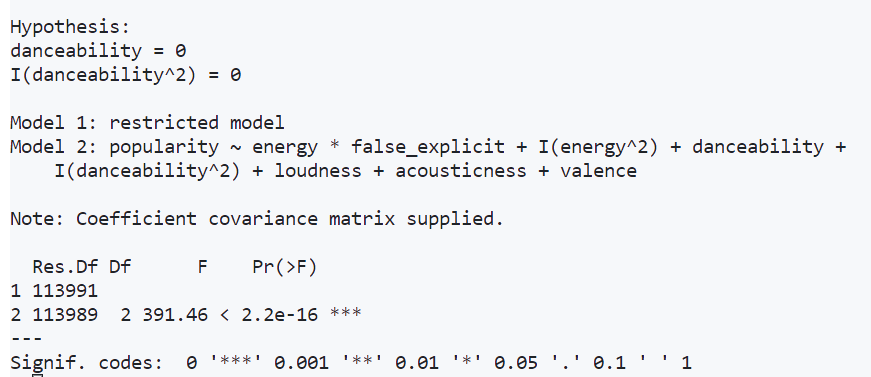
H0: energy = 0

H0: I(energy^2) = 0

Тобто перша гіпотеза стверджує, що коефіцієнт при змінній "energy" дорівнює нулю, а друга гіпотеза стверджує, що коефіцієнт при квадраті змінної "energy" також дорівнює нулю.

Результати тесту показують, що для обох гіпотез було обрано альтернативну модель (Model 2). Це означає, що є статистично значущі докази проти нульових гіпотез. Значення F-статистики дорівнює 283.69, а p-значення менше за 2.2e-16, що є дуже малим.

Отже, можна вважати, що є статистично значущий вплив змінних "energy" і "I(energy^2)" на модель, яка включає також інші регресори, такі як "false\_explicit", "danceability", "I(danceability^2)", "loudness", "acousticness" і "valence".



Даний результат представляє результати тесту лінійних гіпотез для моделі. Тест виконує порівняння двох моделей: обмеженої моделі (Model 1) і альтернативної моделі (Model 2).

Гіпотези, що перевіряються, виглядають наступним чином:

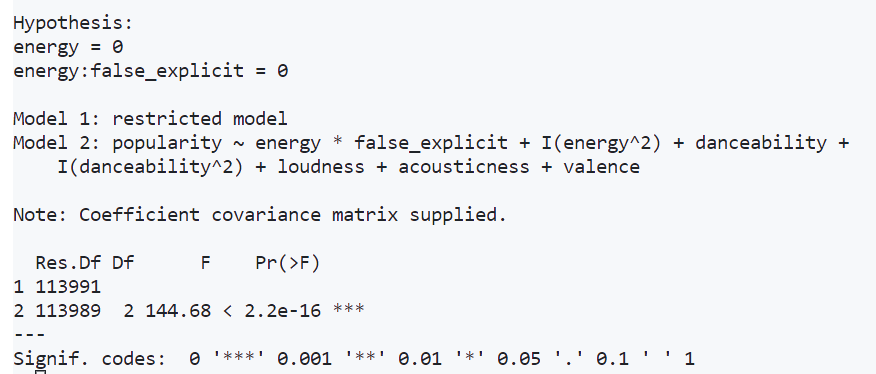
H0: danceability = 0

H0: I(danceability^2) = 0

Тобто перша гіпотеза стверджує, що коефіцієнт при змінній "danceability" дорівнює нулю, а друга гіпотеза стверджує, що коефіцієнт при квадраті змінної "danceability" також дорівнює нулю.

Результати тесту показують, що для обох гіпотез було обрано альтернативну модель (Model 2). Це означає, що є статистично значущі докази проти нульових гіпотез. Значення F-статистики дорівнює 391.46, а p-значення менше за 2.2e-16, що є дуже малим.

Отже, можна вважати, що є статистично значущий вплив змінних "danceability" і "I(danceability^2)" на модель, яка включає також інші регресори, такі як "energy", "false\_explicit", "I(energy^2)", "loudness", "acousticness" і "valence".



Даний результат представляє результати тесту лінійних гіпотез для моделі. Тест виконує порівняння двох моделей: обмеженої моделі (Model 1) і альтернативної моделі (Model 2).

Гіпотези, що перевіряються, виглядають наступним чином:

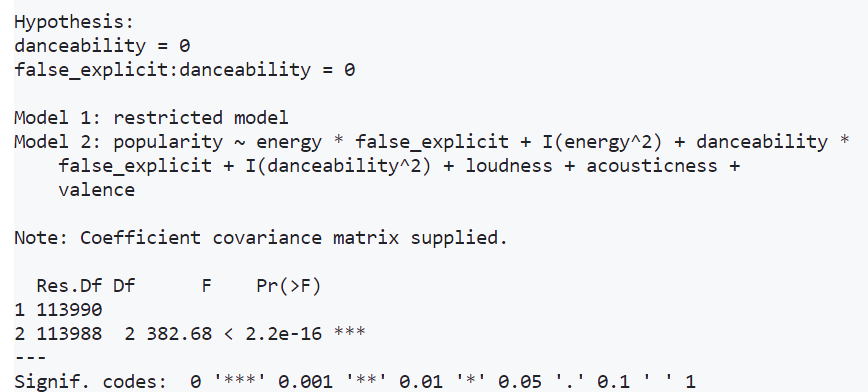
H0: energy = 0

H0: energy:false\_explicit = 0

Тобто перша гіпотеза стверджує, що коефіцієнт при змінній "energy" дорівнює нулю, а друга гіпотеза стверджує, що коефіцієнт при взаємодії між "energy" і "false\_explicit" дорівнює нулю.

Результати тесту показують, що для обох гіпотез було обрано альтернативну модель (Model 2). Це означає, що є статистично значущі докази проти нульових гіпотез. Значення F-статистики дорівнює 144.68, а p-значення менше за 2.2e-16, що є дуже малим.

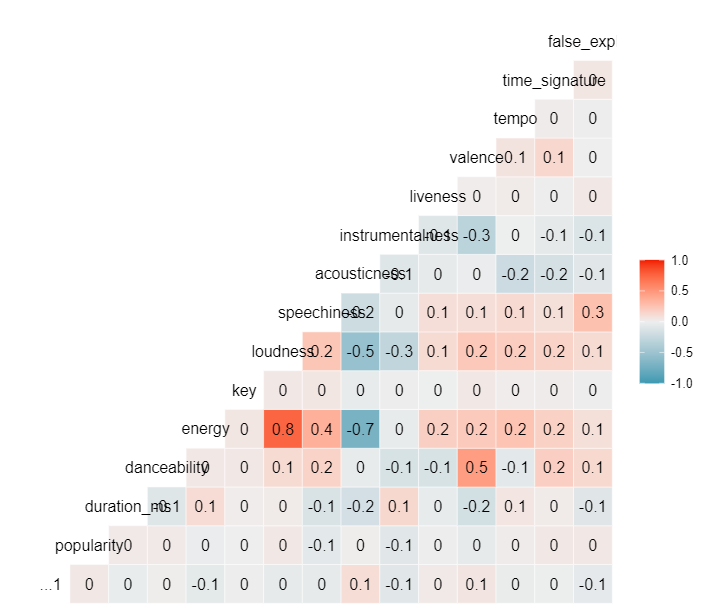
Отже, можна вважати, що є статистично значущий вплив змінних "energy" і "energy:false\_explicit" на модель, яка включає також інші регресори, такі як "false\_explicit", "danceability", "I(danceability^2)", "loudness", "acousticness" і "valence".



Результати тесту на лінійну гіпотезу показують, що обидва перевіряні коефіцієнти - "danceability" та "false\_explicit:danceability" - мають значущий вплив на популярність пісень.

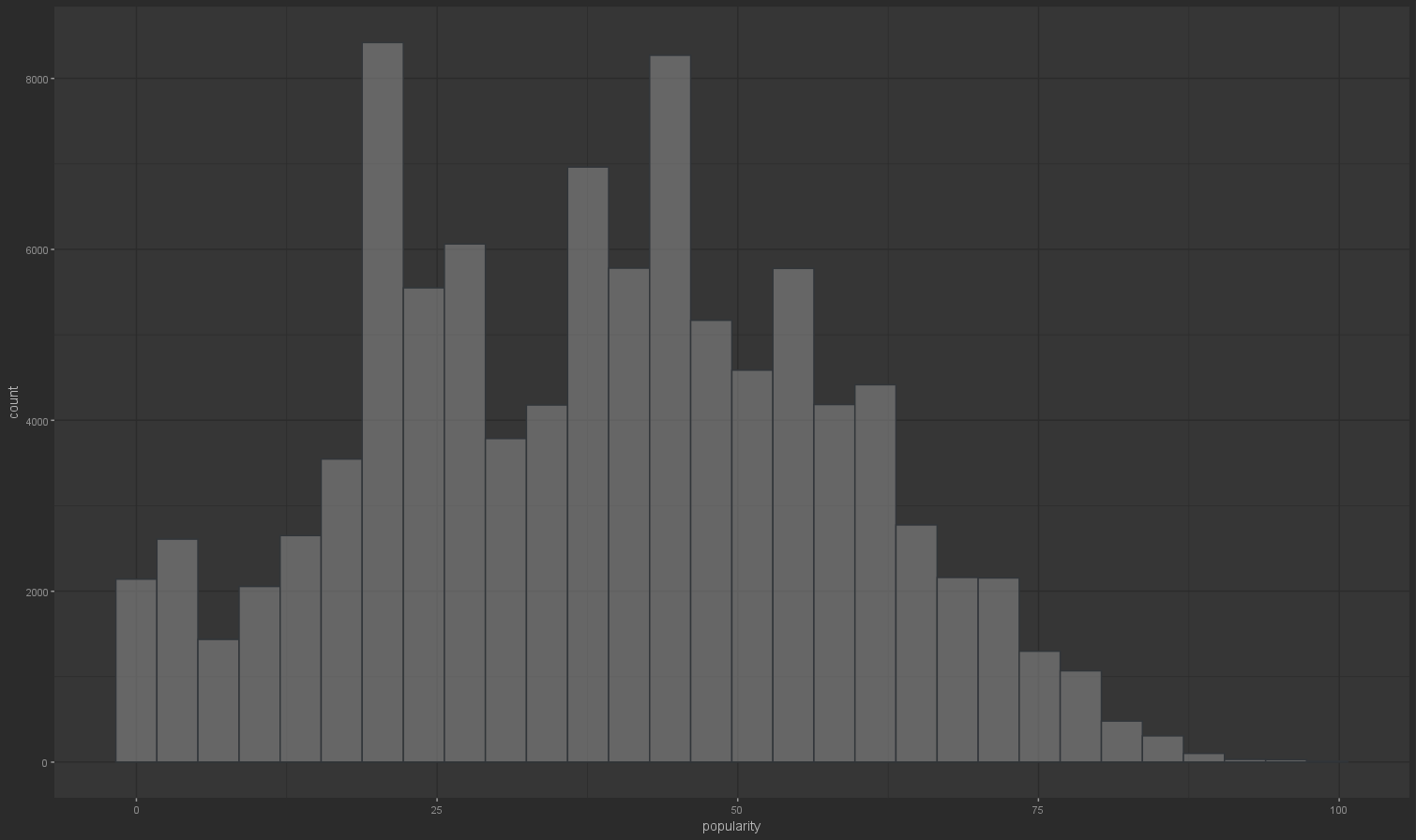
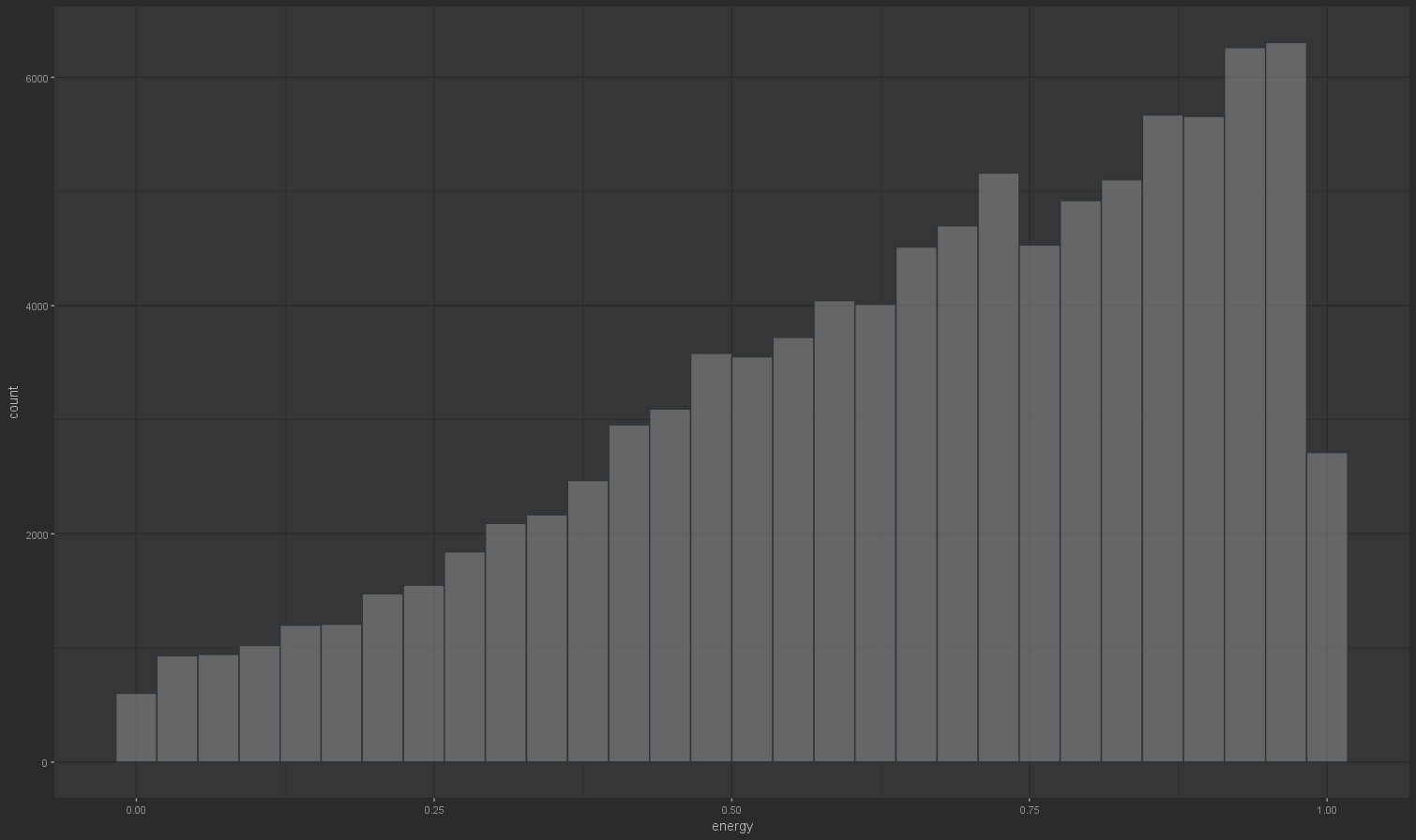
F-статистика тесту дорівнює 382.68, що є високим значенням. Значення p-значення менше за рівень значущості 0.05, що свідчить про статистичну значущість обох коефіцієнтів.

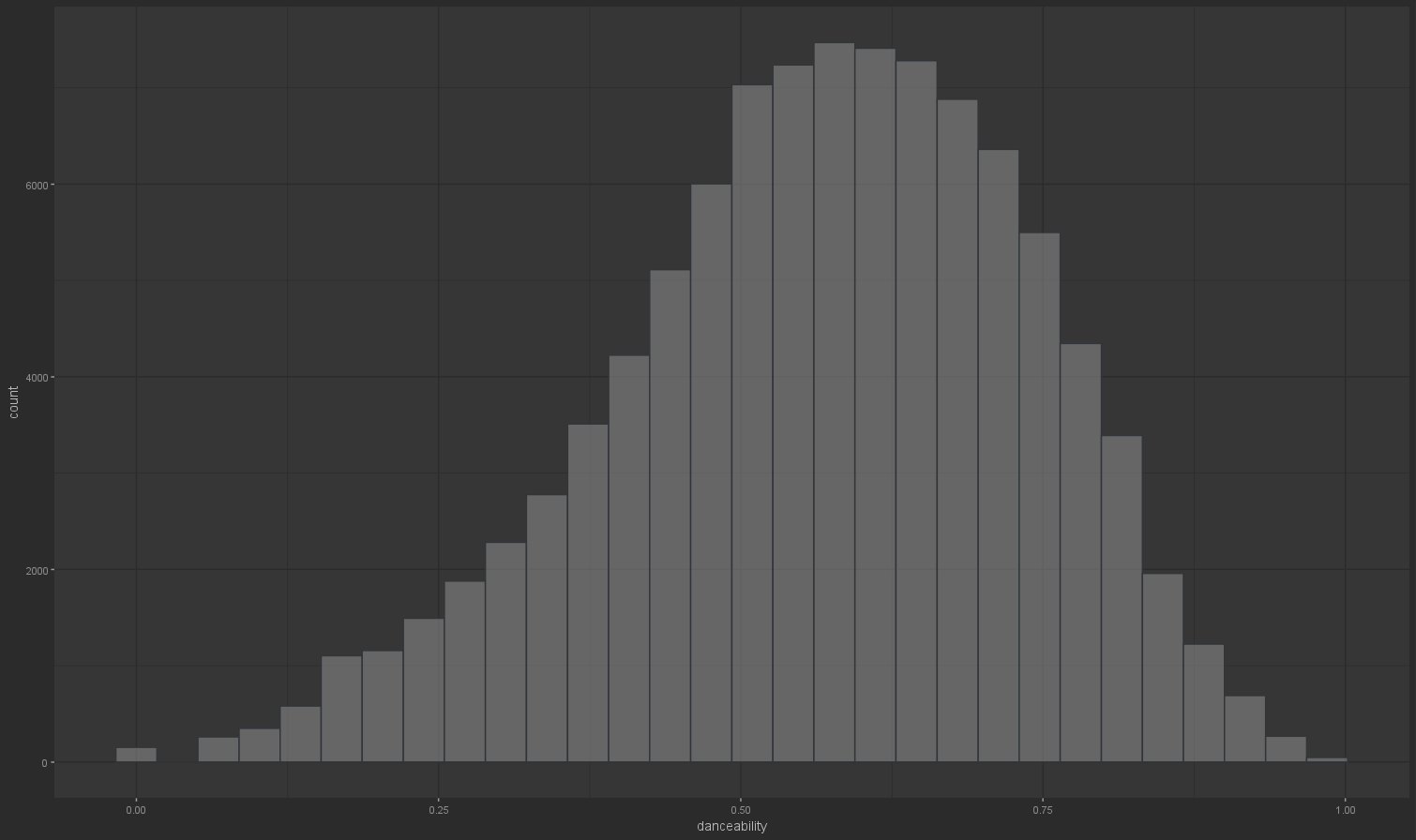
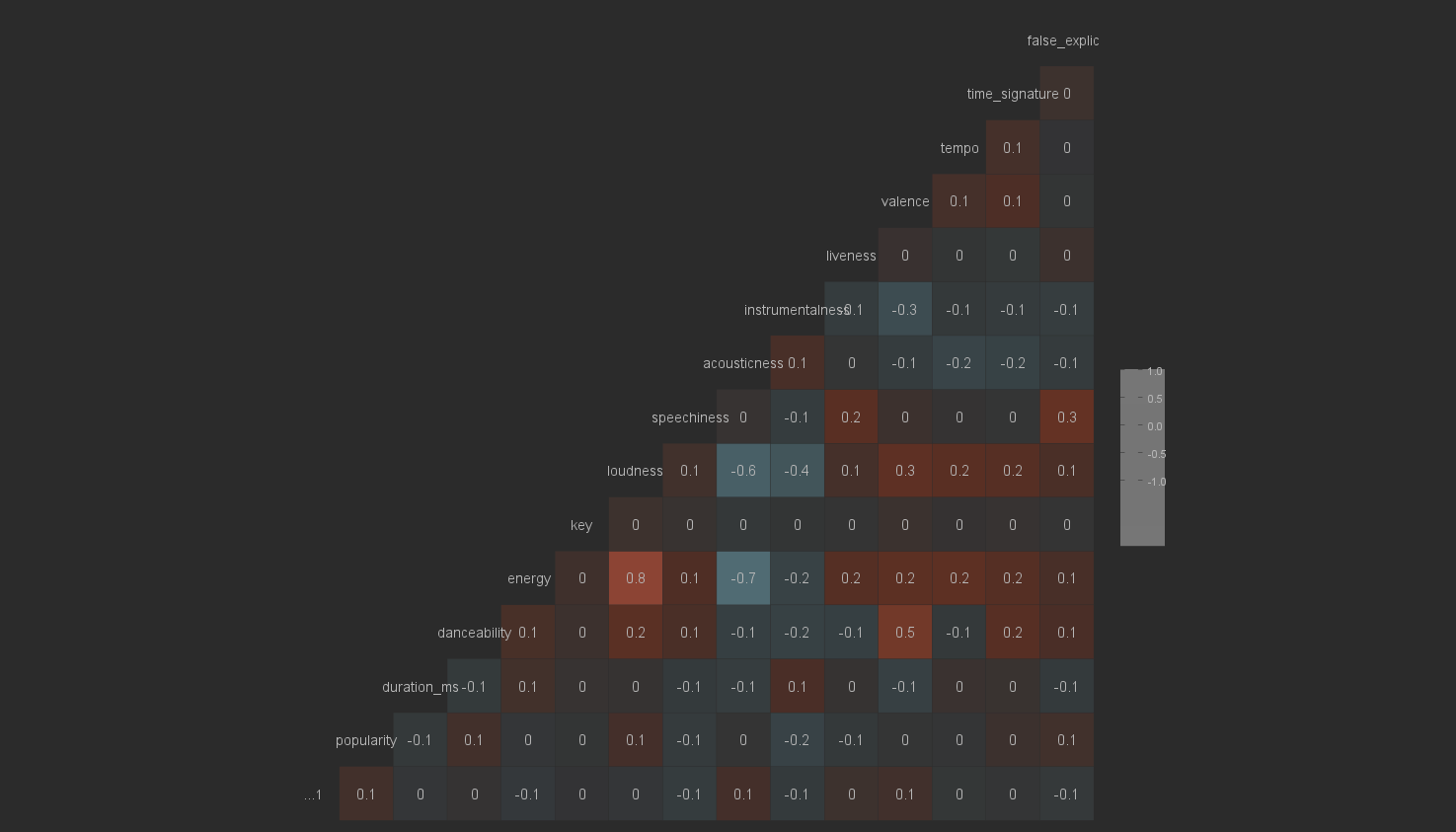
Отже, на основі цих результатів ми можемо стверджувати, що як "danceability", так і взаємодія між "false\_explicit" і "danceability" мають статистично значущий вплив на популярність пісень.



# Датасет без нульової популярності

## Розподіли

Змінні, які ми обрали для аналізу, мають наступні розподіли:



Після видалення нульової популярності розподіли змінних змінилися несуттєво. Кореляції між змінними також відрізняються несуттєво.

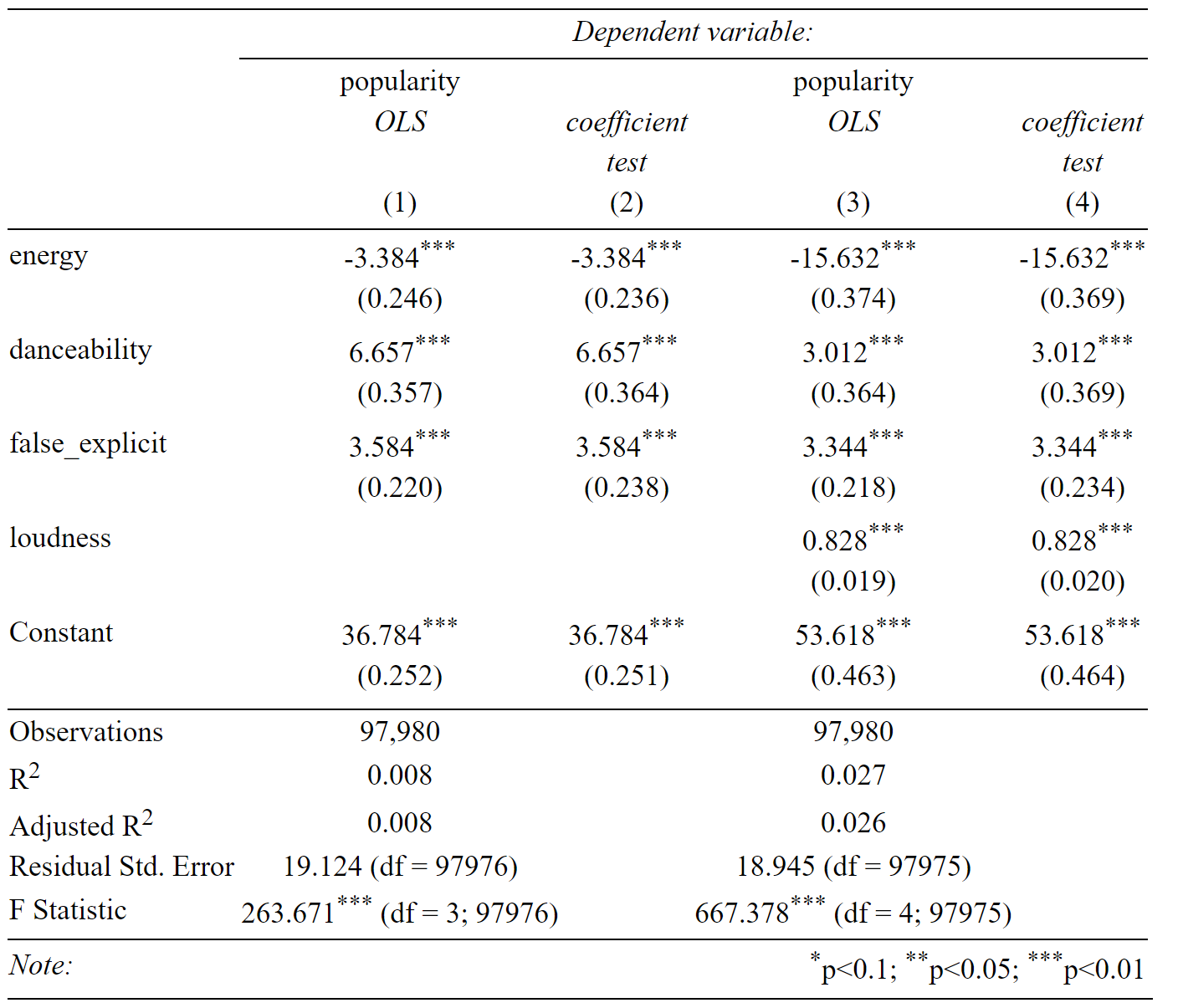
За графіком кореляції між змінними можна побачити, що взяті нами змінні до дослідження мають доволі високі кореляції з деякими іншими змінними:

* energy - loudness
* energy - acousticness
* danceability - valence

Це означає, що при неврахуванні цих змінних зміщення (OVB) можуть призвести до порушення умови про нульове умовне сподівання похибки.

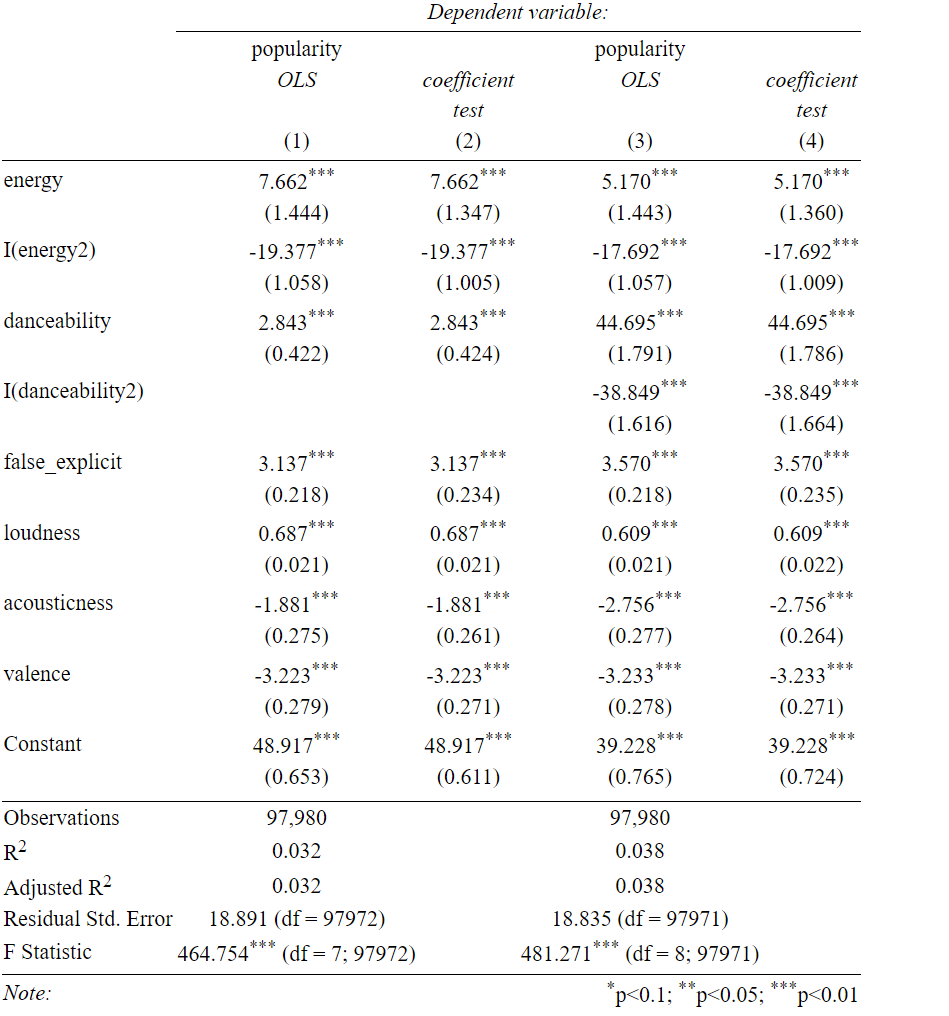
## Моделі

### Додавання контрольованих змінних



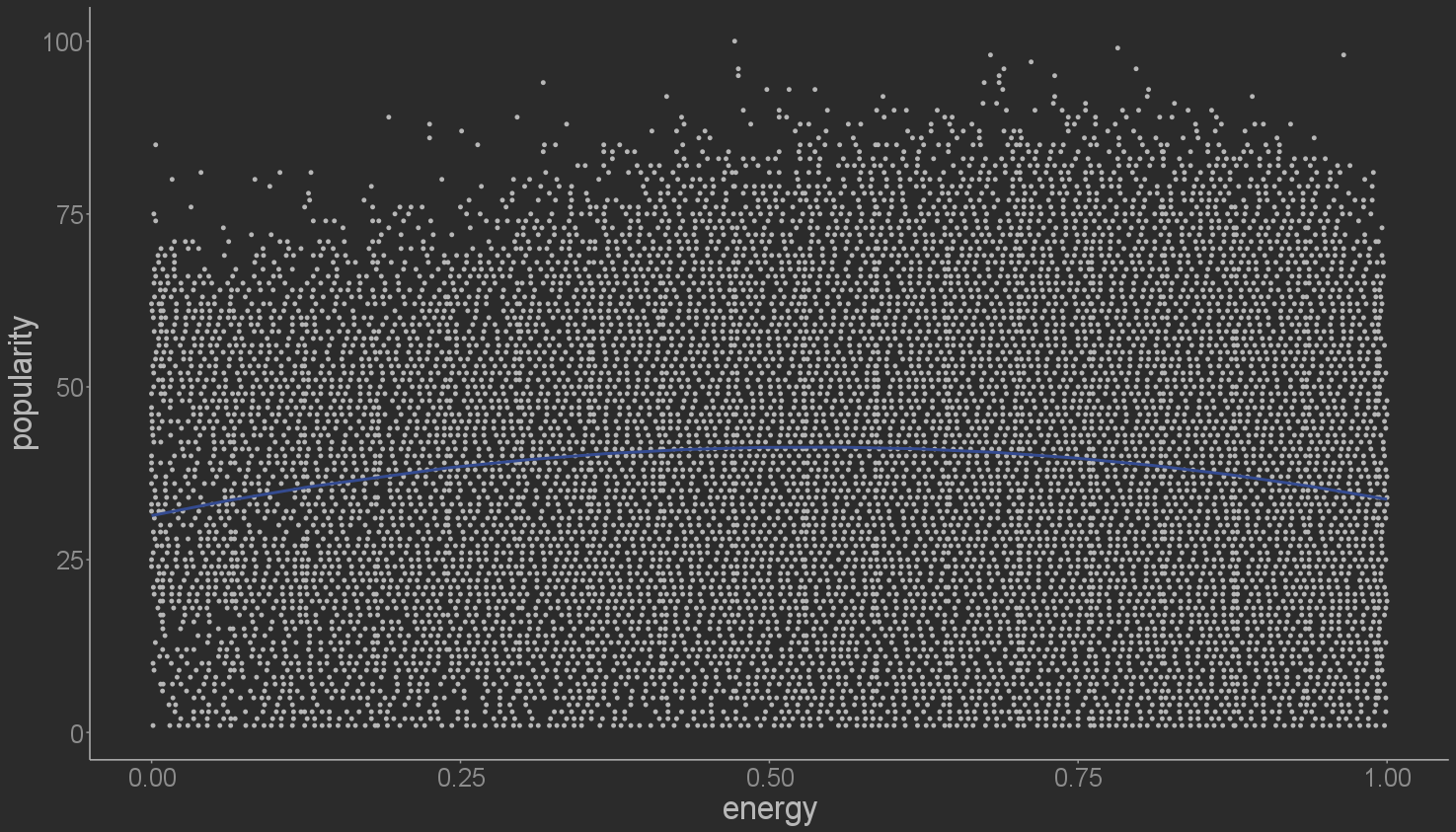
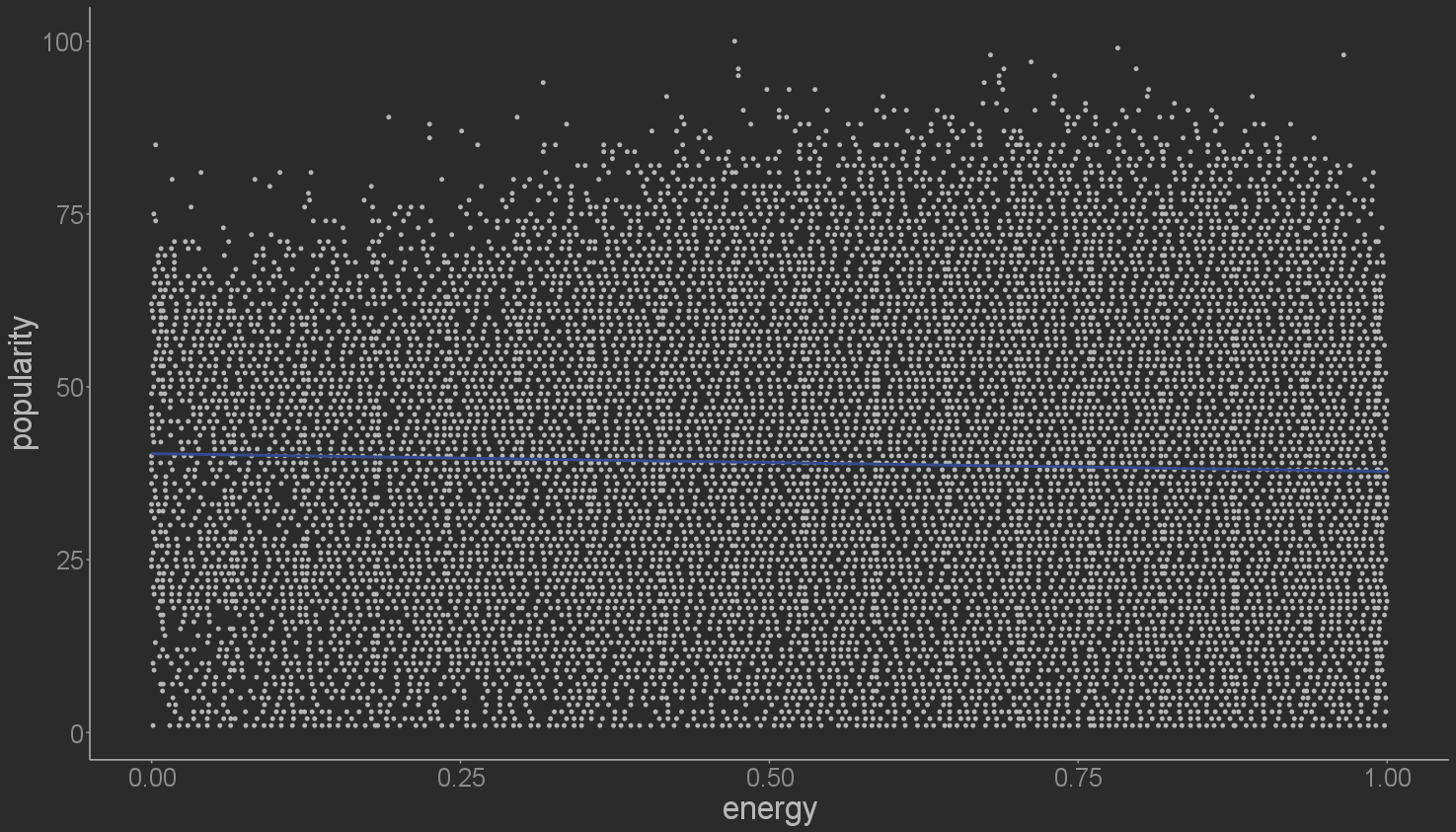
За таблицею розглянемо побудовані модели. Бачимо, що всі коефіцієнти є статистично значущими. За першою базовою моделлю бачимо, що коефіцієнти при досліджуваних змінних є досить малими, мають досить малий вплив на популярність.

### Додавання поліномів вищого порядку

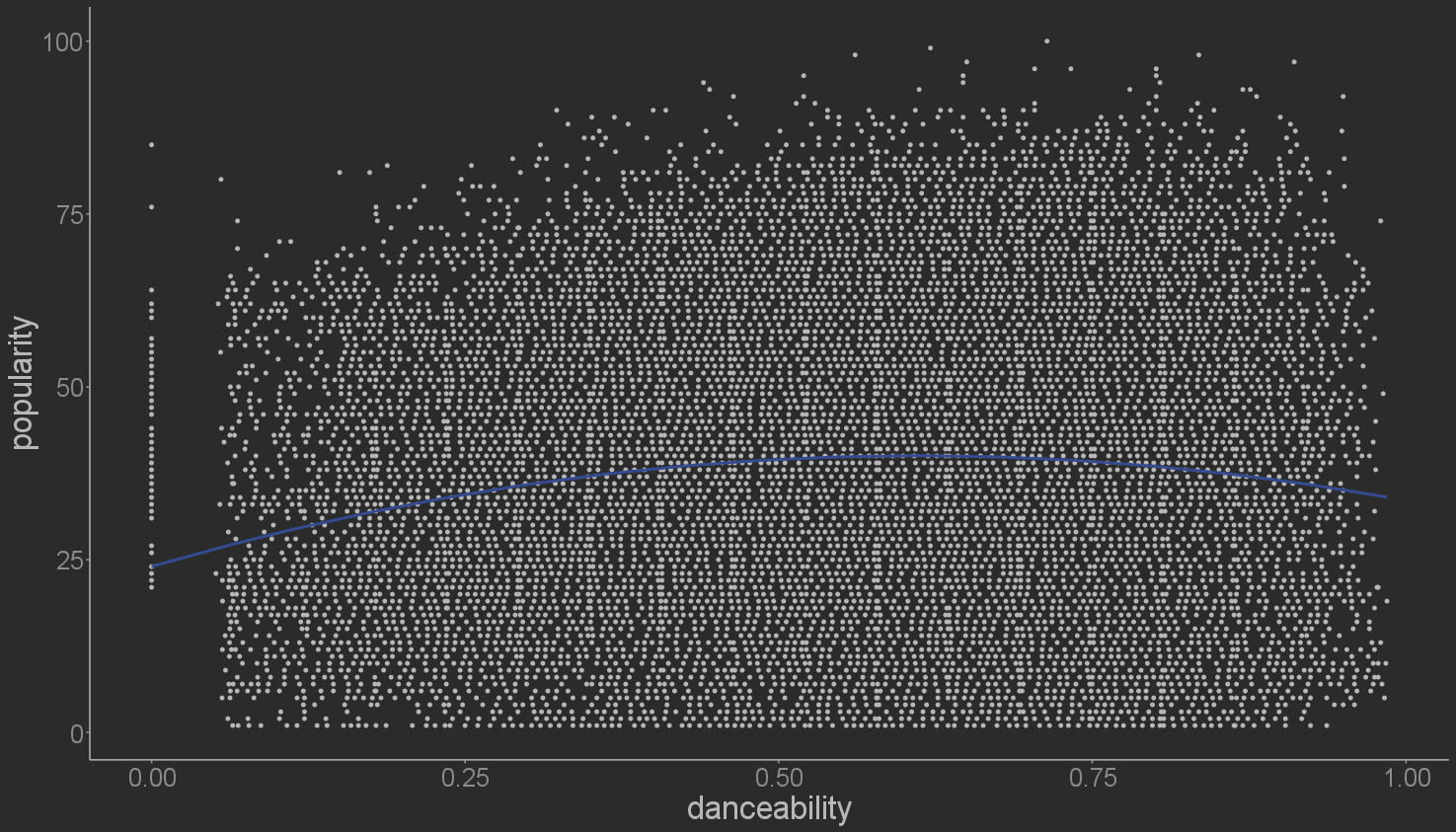
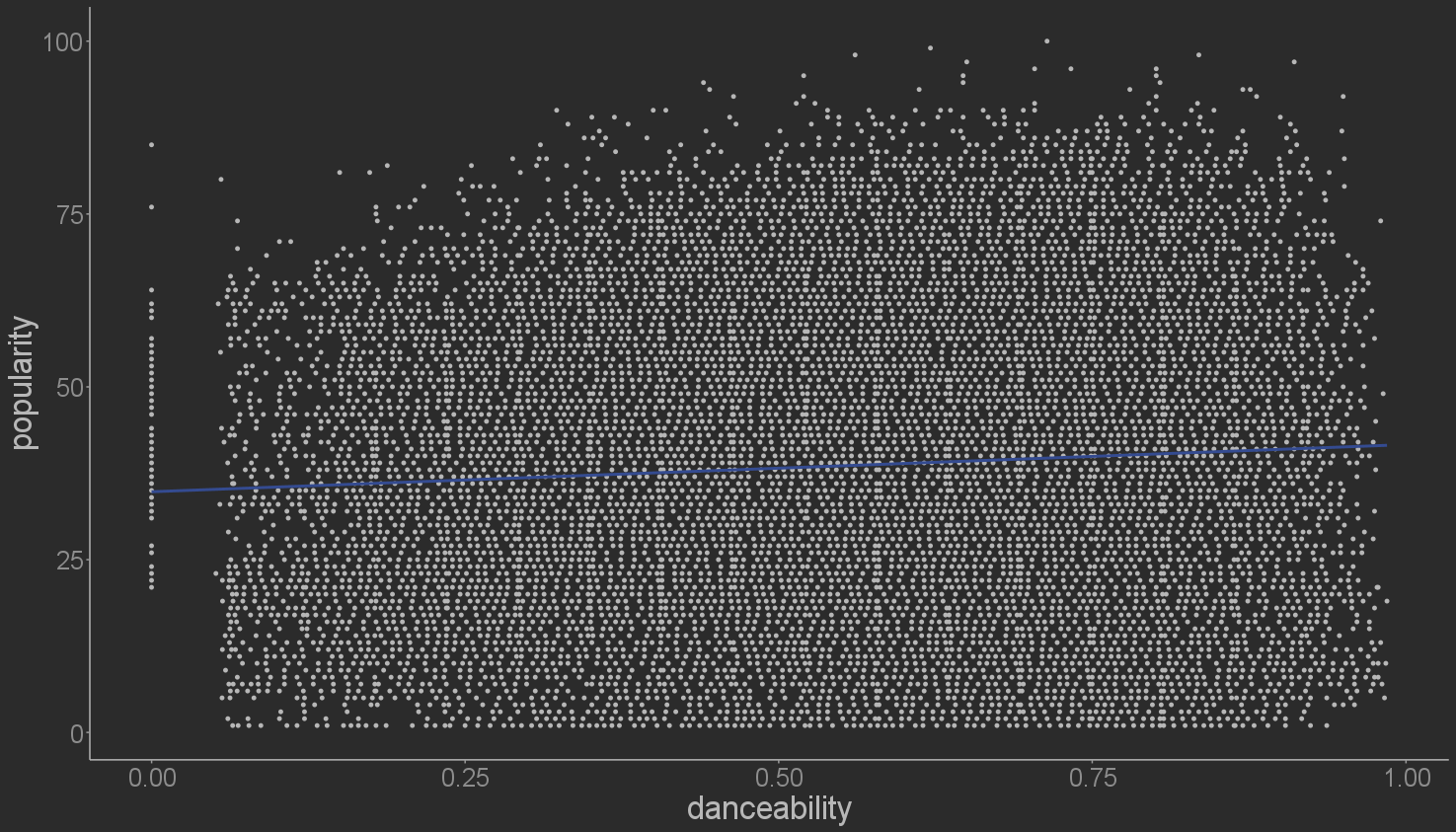


При додаванні контрольованих змінних, що корелюють з energy, бачимо зростання впливу energy на popularity.

При додаванні контрольованих змінних, що корелюють з danceability, спостерігаємо підвищення впливу danceability на популярність.

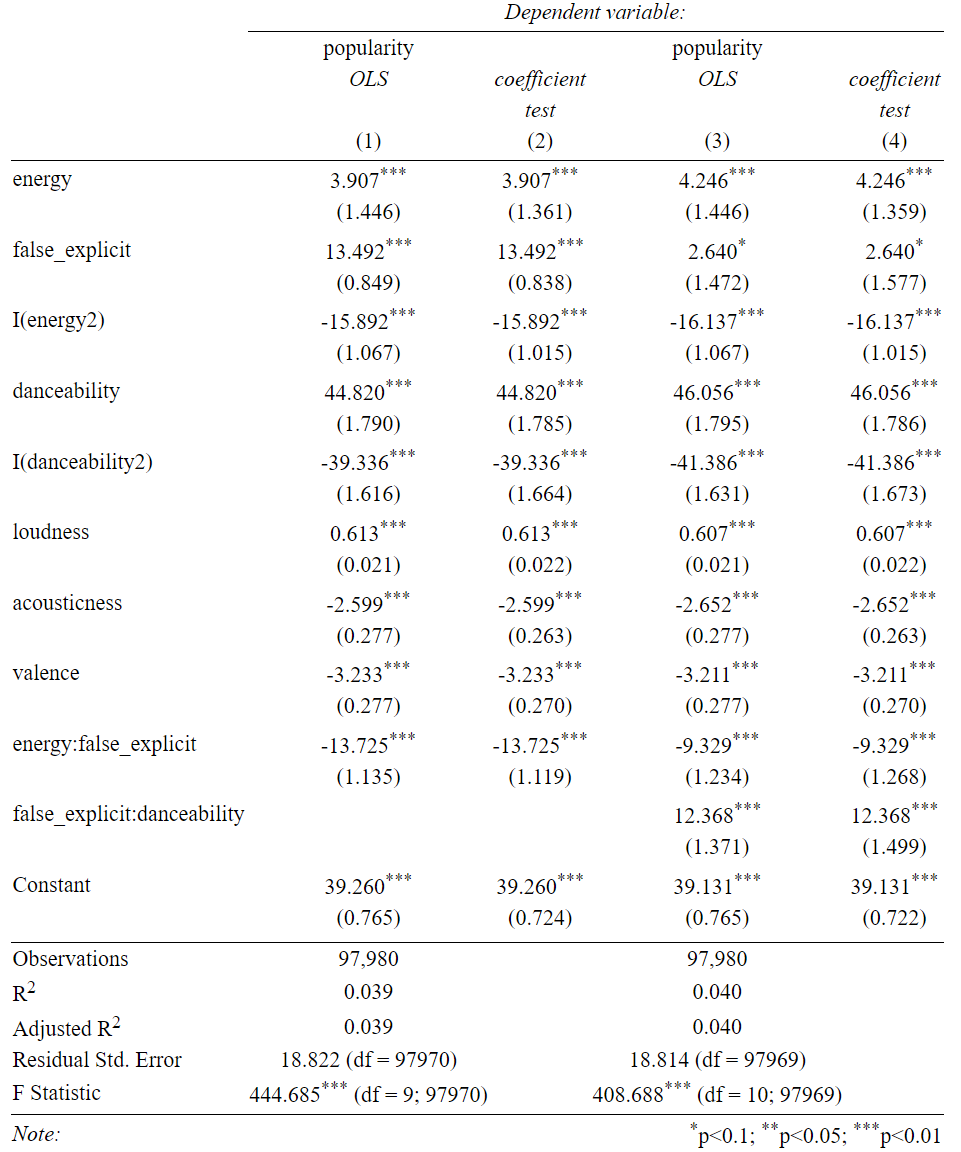
Розглянемо графічно модель popuplarity ~ energy та popularity ~ energy + energy^2.

Можемо додати energy^2 до нашої модели. Бачимо, що вплив energy доволі зріс. Усі коефіцієнти біля доданих змінних зберігають свою статистичну значущість.

Розглянемо графічно модель popularity ~ danceability та popularity ~ danceability + danceability^2.

Можемо додати danceabillity^2 до нашої модели. Бачимо, що вплив danceabillity доволі зріс. Усі коефіцієнти біля доданих змінних зберігають свою статистичну значущість. При цьому коефіцієнти при energy та false\_explicit мають несуттєві зміни.

### Додавання факторів взаємодії



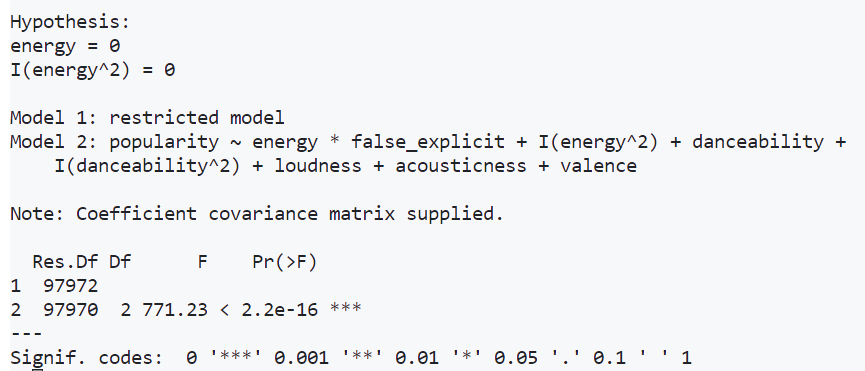
Спробуємо додати до модели фактор взаємодії між false\_explicit та energy. У таблиці вище бачимо, що вплив false\_explicit доволі зріс, при чому інші змінні не мають занадто суттєвих змін.

Після додавання фактору взаємодії між false\_explicit та danceabillity спостерігаються суттєві зміни, а коефіцієнт біля самого фактору взаємодії є статистично значущим. Отже, є сенс додавати цей фактор до модели.

Отже, після восьмої модифікації модели бачимо, що коефіцієнти біля ключових змінних не мають суттєвих змін, тобто можна вважати модель стійкою.

### Тести на статистичну значущість груп коефіцієнтів

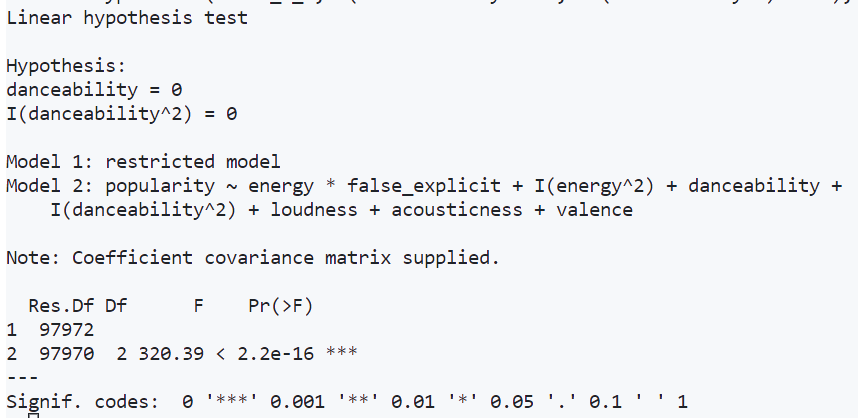
Проведемо тести на статистичну значущість груп коефіцієнтів.



Результати тесту на лінійну гіпотезу показують, що як коефіцієнт "energy", так і коефіцієнт "I(energy^2)" мають статистично значущий вплив на популярність пісень.

F-статистика тесту дорівнює 771.23, що є дуже високим значенням. Значення p-значення менше за рівень значущості 0.05, що свідчить про статистичну значущість обох коефіцієнтів.

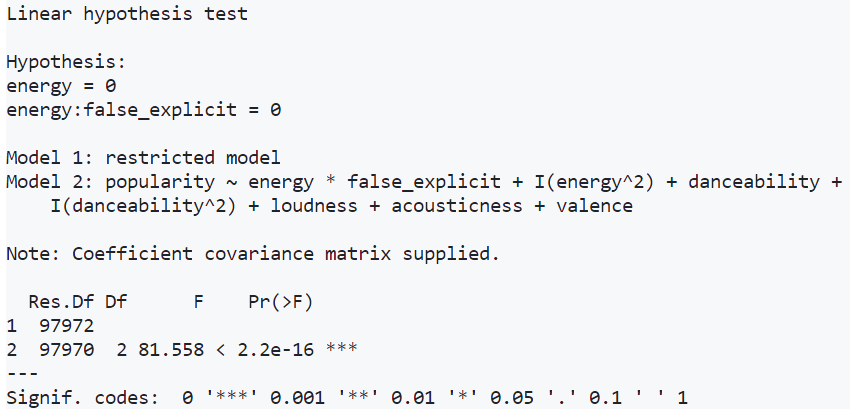
Отже, на основі цих результатів можна зробити висновок, що як лінійний коефіцієнт "energy", так і квадратичний коефіцієнт "I(energy^2)" мають статистично значущий вплив на популярність пісень.



Результати тесту на лінійну гіпотезу показують, що як коефіцієнт "danceability", так і коефіцієнт "I(danceability^2)" мають статистично значущий вплив на популярність пісень.

F-статистика тесту дорівнює 320.39, що є високим значенням. Значення p-значення менше за рівень значущості 0.05, що свідчить про статистичну значущість обох коефіцієнтів.

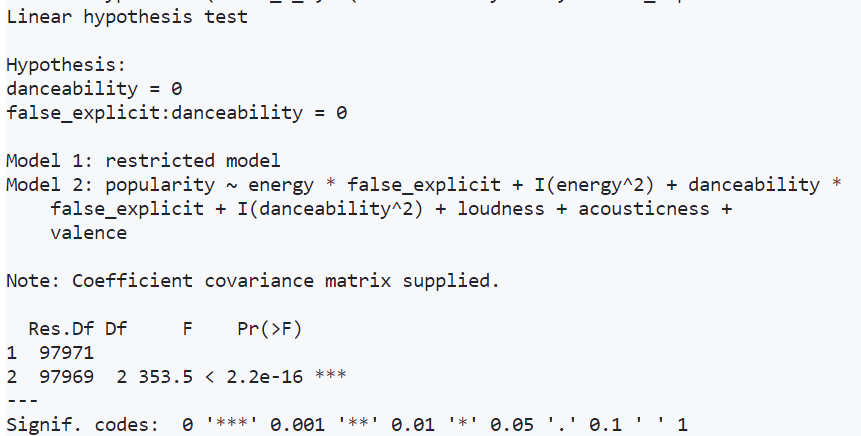
Отже, на основі цих результатів можна зробити висновок, що як лінійний коефіцієнт "danceability", так і квадратичний коефіцієнт "I(danceability^2)" мають статистично значущий вплив на популярність пісень.



Результати тесту на лінійну гіпотезу показують, що як коефіцієнт "energy", так і коефіцієнт "energy:false\_explicit" мають статистично значущий вплив на популярність пісень.

F-статистика тесту дорівнює 81.558, що є високим значенням. Значення p-значення менше за рівень значущості 0.05, що свідчить про статистичну значущість обох коефіцієнтів.

Отже, на основі цих результатів можна зробити висновок, що як лінійний коефіцієнт "energy", так і взаємодіючий коефіцієнт "energy:false\_explicit" мають статистично значущий вплив на популярність пісень.



Результати тесту на лінійну гіпотезу показують, що як коефіцієнт "danceability", так і взаємодіючий коефіцієнт "false\_explicit:danceability" мають статистично значущий вплив на популярність пісень.

F-статистика тесту дорівнює 353.5, що є високим значенням. Значення p-значення менше за рівень значущості 0.05, що свідчить про статистичну значущість обох коефіцієнтів.

Отже, на основі цих результатів можна зробити висновок, що як лінійний коефіцієнт "danceability", так і взаємодіючий коефіцієнт "false\_explicit:danceability" мають статистично значущий вплив на популярність пісень.

# Висновки

# Після модифікацій моделі при дослідження впливу на популярність з нульовими значеннями коефіцієнти біля ключових змінних мали несуттєві зміни, а отже, модель можна вважати стійкою. Але враховуючи достатньо мале значення R^2 не можна точно стверджувати, що саме від цих обраних змінних повноцінно залежить популярність треку.

# Після видалення нульової популярності з датасету особливості даних суттєво не змінилися і після модифікацій моделі коефіцієнти біля ключових змінних мали несуттєві зміни, а отже, модель можна вважати стійкою. Але додавання фактору взаємодії false\_explicit:danceability стало статистично значущим та має значний вплив на популярність треку. Але враховуючи достатньо мале значення R^2 не можна точно стверджувати, що саме від цих обраних змінних повноцінно залежить популярність треку.

# 

# Список літератури

* Конспекти лекцій з предмету “Аналіз даних” [Лекції – Google Диск](https://drive.google.com/drive/folders/1LsWpozxahLgv51DOZKrIhKdIvhYMfEbX)
* Записи лекцій з предмету “Аналіз даних” [Записи лекцій – Google Диск](https://drive.google.com/drive/folders/1A9hTKNKyxRqC62QM7EN6FBaGJT7aV8tI)