

פרויקט רגרסיה – קבוצה 10

**316468586, 318321627**

Spotify

תוכן עניינים

1. [תקציר מנהלים 3](#_תקציר_מנהלים)

[2. עיבוד מקדים 6](#_2._עיבוד_מקדים)

[2.1 הסרה של משתנים 6](#_Toc125194353)

[2.2 התאמת משתנים 8](#_Toc125194354)

[2.3 הגדרת משתנה דמה 8](#_Toc125194355)

[2.4 הגדרת משתני אינטראקציה 8](#_Toc125194356)

[3. התאמת המודל ובדיקת הנחות המודל 10](#_Toc125194357)

[3.1 בחירת משתני המודל 10](#_Toc125194358)

[3.2 בדיקת הנחות המודל 11](#_Toc125194359)

[**3.2.1** בדיקת הנחת הלינאריות 11](#_Toc125194360)

[**3.2.2** בדיקת הנחת שוויון שוניות 11](#_Toc125194361)

[**3.2.3** בדיקת הנחת הנורמליות של השגיאות 12](#_Toc125194362)

[4. שיפור המודל 14](#_Toc125194363)

[נספחים 15](#_Toc125194364)

# **תקציר מנהלים**

**בפרויקט הנ"ל התבקשנו לבצע ניתוח סטטיסטי של מאגר נתונים התואם את נתוני השירים שחופשו באפליקציה ספוטיפיי. מטרתנו העיקרית היא לבנות מודל רגרסיה מדויק בצורה מקסימלית אשר באמצעותו נוכל לחזות בעתיד מה גורם לאנשים לחפש ולאחר מכן להאזין לשירים באפליקציה. במסמך זה נפרט את הפעולות שביצענו על מנת לבנות את המודל שלנו, המסקנות ואת נקודת מבטנו על הנושא.**

**ראשית, רצינו לבדוק אילו נתונים מסט הנתונים שקיבלנו האם קטגוריאליים ואילו רציפים. לאחר סיווג זה בדקנו אילו משתנים כדאי להסיר מהמודל – בשל קורלציה נמוכה יחסית בין המשתנה המוסבר למסביר. שוב, תמיד המטרה היא לשפר את המודל. ביצענו מתאם פירסון בין המשתנים הרציפים שלנו למשתנה המוסבר במטרה לוודא בדרך נוספת אילו משתנים כדאי להסיר. יתרה מכך, ביצענו רגרסיה לאחור בשביל להיות בטוחים במלוא האחוזים שאנחנו שומרים על אמינות בסיס הנתונים שלנו. לאחר שהיינו בטוחים במלוא מובן המילה שכך יש לפעול – הסרנו את המשתנים הבאים:** singer gender, crowd, acousticness, energy**.**

**לאחר הסרת המשתנים, רצינו להבין איזה התאמות יש לבצע במשתנים שלא הוסרו. לכן, ניגשנו לעבור ולבדוק האם יש לאחד משתנים, או לבצע איחוד קטגוריות במשתנים קטגוריאליים. קודם כל, המשתנים הקטגוריאליים שלנו הם בינאריים לחלוטין ( 0 או 1 ) ולכן איחוד קטגוריות עבור משתנים אלו היה לא רלוונטי כאן. לאחר מכן, פנינו לבחון האם יש הצדקה למשתני דמה ומשתני אינטראקציה. בחרנו להוסיף משתנה אינטראקציה אחד והוא:** mode𝑖 ∗ loudness. **לאחר מכן, ביצענו השוואות בין ארבעת המודלים השונים שנלמדו בכיתה – רגרסיה לאחור, רגרסיה לפנים, ורגרסיה בצעדים. את המודל השוונו באמצעות המדדים:** AIC, Bic, **. קיבלנו כי ברגרסיה לאחור, לפנים, ובצעדים קיבלנו את אותם הנתונים ובחרנו לעבוד לפי רגרסיה לאחור. בחרנו לעבוד עם מודל זה מכיוון שמדד ה-**BIC **היה נמוך יותר ממדד זה במודל המלא.   
לאחר בחירת המודל, ביצענו את 3 בדיקות הנחות המודל שהן: בדיקות הנחות שוויון שוניות, הנחת הנורמאליות של השגיאות, ובדיקת הנחת הלינאריות. מצאנו כי המודל עומד בכל שלושת ההנחות.  
לבסוף, רצינו לשפר את המודל על ידי ביצוע מספר טרנספורמציות על המשתנה המוסבר וקיבלנו כי משפר את המודל מאחר ומתקבל מדד** BIC **נמוך באופן חד משמעי מהמודל שהתקבל תחילה.**

**לתקציר זה, מצורף טבלת המשתנים מסט הנתונים המקורי.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| סוג המשתנה -מוסבר/מסביר | סימון במודל | יחידת מידה | סוג המשתנה – רציף / קטגוריאלי | הסבר קצר על המשתנה |
| Spotify **– מוסבר** |  | **מיליונים** | **רציף** | **כמות החיפושים של השיר באפליקציה (במיליונים)** |
| youtube views **- מסביר** |  | **מיליונים** | **רציף** | **מספר הצפיות באתר Youtube במיליונים** |
| Energy **- מסביר** |  | **לא ידוע** | **רציף** | **מדד תפיסתי של העוצמה** |
| loudness **– מסביר** |  | **דציבלים** | **רציף** | **קולניות הרעש הכללי של רצועת השיר בדציבלים** |
| mode **– מסביר** |  | **0 – מנור**  **1 – מיג'ור** | **קטגוריאלי** | **באיזה סולם השיר מנוגן** |
| speechiness **– מסביר** |  | **אחוזים** | **רציף** | **אחוז המילים המדוברות בשיר** |
| acousticness **– מסביר** |  | **לא ידוע** | **רציף** | **מידת האקוסטיות של השיר** |
| crowd **– מסביר** |  | **אחוזים** | **רציף** | **נוכחות הקהל בהקלטת השיר** |
| valence **- מסביר** |  | **לא ידוע** | **רציף** | **האם השיר חיובי** |
| tempo **– מסביר** |  | BPM | **רציף** | **מקצב השיר** |
| duration **– מסביר** |  | Milliseconds | **רציף** | **כמה זמן השיר נמשך** |
| singer\_gender **– מסביר** |  | **0 – נקבה**  **1 – זכר** | **קטגוריאלי** | **מין הזמר המבצע – הנחנו ש-0 זה נקבה ו-1 זכר.** |

# **2. עיבוד מקדים**

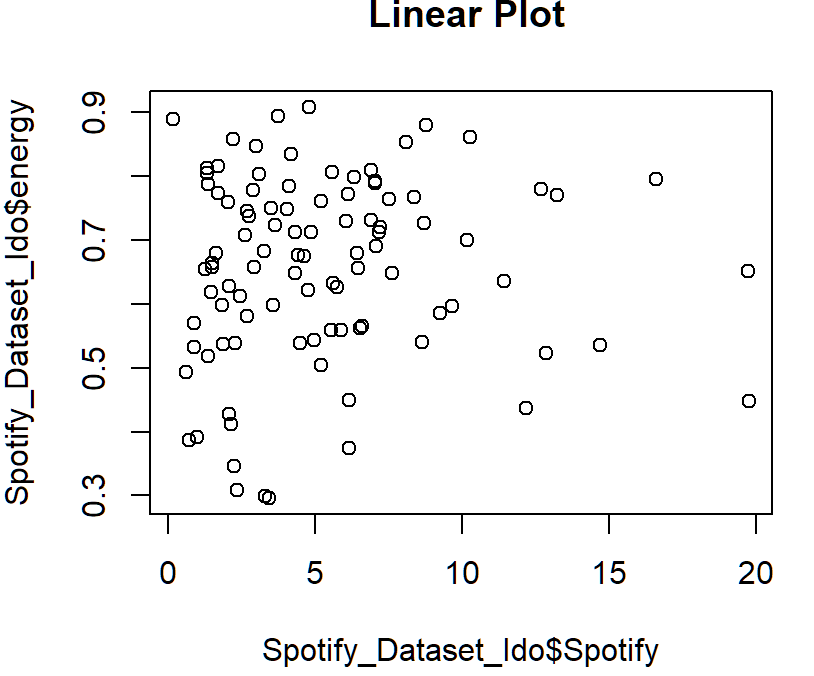
## [2.1 הסרה של משתנים](#_2.1)

**לפי חלק א', אין משתנים שהסרנו. לכן נקדיש חשיבות רבה לחלק זה על מנת לבחון אילו משתנים רלוונטים בשביל להסביר את המשתנה המוסבר שלנו.**

**בניסיון לבחון אילו משתנים נצטרך להסיר מהמודל, נבחן את מתאם פירסון בין המשתנים הרציפים למשתנה המוסבר ונבחן את תרשימי הפיזור. היות וקשר לינארי חלש מוגדר בין 0.4, -0.4 במקרה של הנתונים שלנו נצטרך להסיר את כל המשתנים המסבירים. לכן, על מנת לעמוד עדיין עם המודל בחרנו לצמצם את לינאריות המודל להפרש בין 0.1, -0.1 .**

**טבלת המתאם המתקבלת בהתאם למתאם פירסון:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Youtube Views | Energy | Loudness | Crowd | Speechiness | Acousticness | Tempo | Valence | Duration |
| 0.3993559 | 0.05428806 | 0.1940468 | -0.01017211 | -0.171322 | -0.02510019 | -0.1204119 | -0.1516683 | 0.1954884 |

**המשתנים המסומנים בצהוב מעידים על מתאם חלש בהתאם להגדרה החדשה שלנו. משתנים אלו חשודים כך שלא קיים קשר לינארי בינם לבין המשתנה המוסבר ועל מנת לבחון את השערה זו לעומק, נבדוק זאת על ידי תרשימי הפיזור.**

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

**לפי תרשימי הפיזור, ניתן לראות שלא קיים קשר לינארי כללי בין המשתנים המוסברים הרציפים למסבירים השונים.**

**נעבור לבדוק את המשתנים הקטגוריאלים ביחס למסביר, ע"פ מתאם ספירמן.**

|  |  |  |
| --- | --- | --- |
| **משתנה** | Mode | Singer gender |
| **קורלציה** | **-0.106** | **-0.022** |

**לפי הרצת מבחן רגרסיה לאחור לפי מדד** BIC **, נרצה להסיר את המשתנים הבאים:** duration, energy, crowd**,** acousticness, singer gender, Speechiness tempo loudness, mode,

**נסתכל לפי המשתנים שעלה צורך להסירם בחלק הקודם וגם על פי מבחן** BIC **ונבצע את הסרתם של המשתנים הבאים:** singer gender, crowd, acousticness, energy

**מכאן שאת המשתנים:** Speechiness, duration, loudness, mode, tempo **לא נסיר מהמודל.**

## [2.2 התאמת משתנים](#_2.2)

בהתאם לסעיף קודם ניתן לראות שקשה לראות קשר לינארי מובהק בין המשתנים הרציפים המסבירים לבין המשתנה המוסבר. נרצה לבצע דיסקרטזיציה כאשר כן ניתן לראות קשר כזה. לכן נבחר שלא לבצע תהליך זה עבור המשתנים הקטגוריאלים על מנת לא לפגוע באמינות בסיס הנתונים שלנו.

בנוסף, נבדוק את הנתונים שלנו במטרה לאחד קטגוריות בין משתנים קטגוריאלים. המשתנים הקטגוריאלים במדגם שלנו הם בינארים (0 או 1). כלומר – אחד מהם מייצג את סולם השיר ( mode ) – כאשר 1 מיוחס לסולם מג'ורי ו - 0 לסולם מינור. השני מייצג את חיוביות השיר ( valence ) כאשר 1 עבור שיר חיובי ו – 0 עבור לא חיובי. בשל העובדה שהקטגוריות הן ניגודיות לחלוטין ( שייך לקטגוריה מסוימת או לא ) בחרנו שלא לאחד קטגוריות אלו, שוב במטרה לשמור על אמינות הנתונים שלנו.

## [2.3 הגדרת משתנה דמה](#_2.3)

בשביל לשלב בבניית מודל הרגרסיה שלנו את המשתנה הקטגוריאלי הנותר , Mode, נרצה להגדיר משתני דמה אשר יצדיקו את היותו משתנה בינארי במודל. המשתנה נע בין 0 ל1 ובהתאם האם השיר חיובי או לא. נגדיר זאת כך:

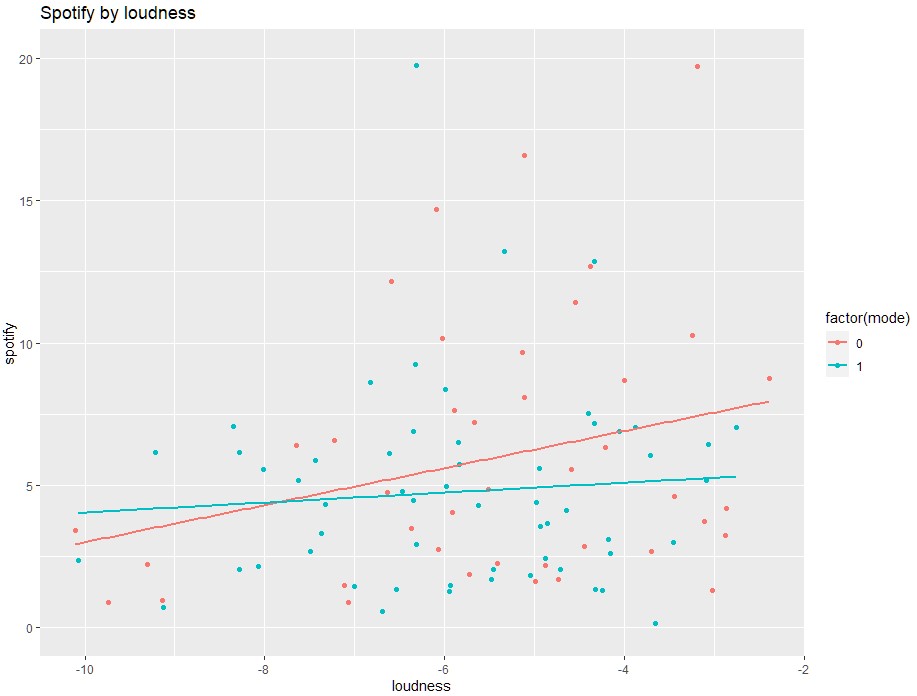
## [2.4 הגדרת משתני אינטראקציה](#_2.4)

**ע"י הגדרת משתנה אינטראקציה ננסה לאמוד את התרומה היחסית של כל משתנה בנפרד עבור שיפוע קו הרגרסיה. נבדוק את האינטראקציה בין המשתנים הרציפים למשתנה הדמה שלנו, האם הן מועילות למודל שלנו או לא. נבדוק זאת באמצעות תרשים פיזור ( על מנת לא להעמיס על הדוח, תרשימי פיזור עבור משתנים שלא בחרנו להוסיף להם משתני אינטראקציה מצורפים בנספחים).**

1. **כמות החיפושים של השיר באפליקציה מול קולניות הרעש הכללי בדציבלים**

**נגדיר משתנה אינטראקציה כך:** mode𝑖 ∗ loudness

**נבחן את השפעתם של שני המשתנים המסבירים, האם השיר מנוגן במאג'ור או במינור אל מול מדד קולניות הרעש. ננסה לראות אם ניתן להסביר כך בצורה טובה יותר את המשתנה המוסבר שלנו (**Spotify**):**



**מן ההבנה שלנו במוזיקה, אנו מאמינים כי יש קשר בין סולם השיר (** mode **) לבין מדד הקולניות בשיר (**(Loudness **. כיודע, ניתן לסווג שירים על פי סוג הסולם שלהם – יותר שמחים / עצובים. לכן, נאמין שגם במקרה שלנו יש קשר מסוים לסולם השיר בהתאם למדד הקולניות שלו. על פי התרשים ניכר כי בשני המקרים יש השפעה חיובית אך ניתן להבין שעבור המינורי (0), השיפוע גדול יותר מהמאז'ורי (1). לכן, נבחר להוסיף לצמד הזה משתנה אינטראקציה.**

# **3. התאמת המודל ובדיקת הנחות המודל**

## [3.1 בחירת משתני המודל](#_3.1)

**בכדי לבחור את המודל המיטבי עבורנו, נבצע השוואה בין המודל המלא לבין המודלים אשר התקבלו באמצעות האלגוריתמים השונים שלמדנו בהרצאה:**

**•** רגרסיה לפנים **- המודל מתחיל ללא משתנים, הכנסת המשתנים תעשה בשלבים. כל איטרציה נכניס למודל את המשתנה המובהק ביותר לפי** AIC**.**

**•** רגרסיה לאחור - **המודל מתחיל כמודל מלא, הורדת המשתנים מתבצעת בשלבים. כל איטרציה נוריד מהמודל את המשתנה הכי פחות מובהק ולפי מדד** AIC **.**

**•** רגרסיה בצעדים - **מדובר בשילוב של 2 השיטות שמצויינת למעלה (רגרסיה לפנים ולאחור) בכל איטרציה נבדוק האם כדאי להכניס או להוציא משתנה מהמודל לפי ערך ה**AIC**.**

**על מנת לבחון את החלופות שהתקבלו בכל אחד מהאלגוריתמים השונים, נשווה באמצעות מדד** BIC**. מדד זה בוחן את טיב ההתאמה של המודל על פי מדד הנראות, נרצה למזער ערכים אלו. כלומר - המודל בעל מדד ה** BIC**הקטן ביותר הוא הטוב ביותר. טבלת סיכום המדדים:**

**( כל הפלטים מגובים בנספחים )**

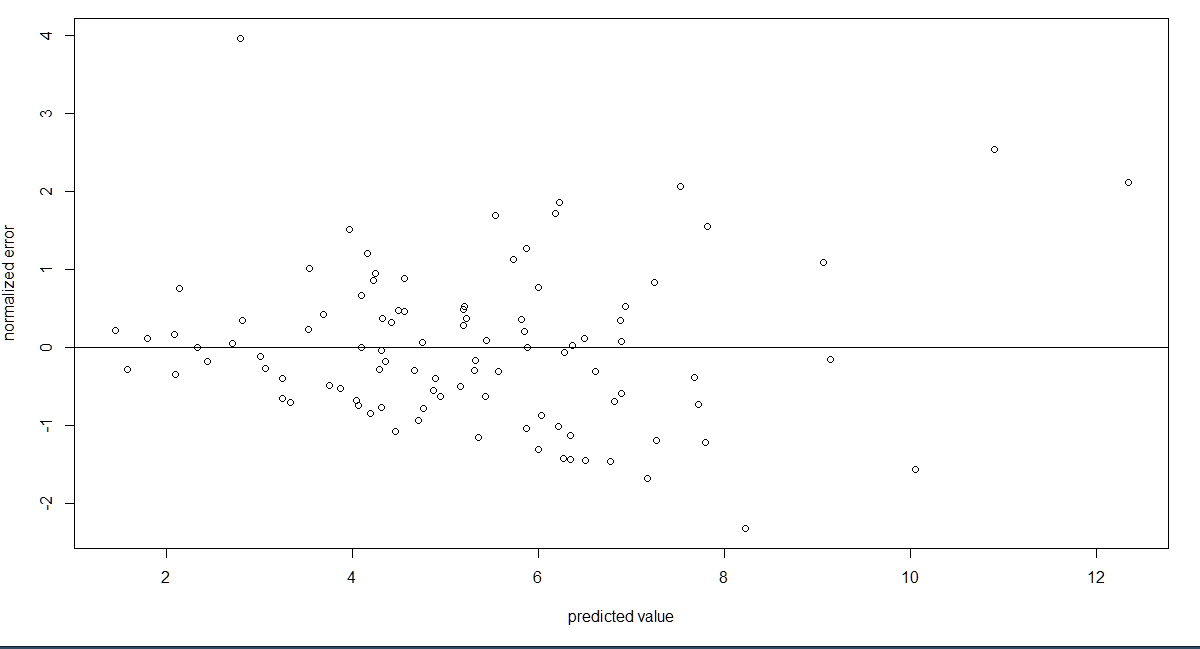
|  |  |  |  |
| --- | --- | --- | --- |
| אלגוריתם | מדד | מדד AIC סופי | מדד BIC סופי |
| מודל מלא | 0.2192 | 288.08 | **567.1224** |
| [רגרסיה לאחור](#_Backward_Regression) | 0.2175 | 267.07 | 549.7197 |
| [רגרסיה לפנים](#_Regression_forward) | 0.2175 | 267.07 | 549.7197 |
| [רגרסיה בצעדים](#_Regression_with_steps) | 0.2175 | 267.07 | 549.7197 |

במקרה הנוכחי קיבלנו כי הן הרגרסיה לאחור, רגרסיה לפנים, ורגרסיה בצעדים הן בעלי אותן מדדים. ניתן לראות כי המודלים הללו עדיפים לפי מדדי AIC ו-BIC על פני המודל המלא ובחרנו לעסוק ברגרסיה לאחור.

## 3.2 בדיקת הנחות המודל

### **3.2.1** בדיקת הנחת הלינאריות

**על מנת לבצע בדיקה זו, נבחן את תרשים פיזור השגיאות המתוקננות ביחס למשתנה המוסבר (**Spotify**). ניתן לראות את המגמה – תצורה של משפך רחב – כלומר, כך התצפיות מתפזרות. במטרה לוודא את השערתנו נבצע מבחן**  [CHOW](#_Chow_test) **לבחינת ההנחה. במבחן זה התקבל ולכן נקבל את השערת האפס ונאמר שהנחת הלינאריות אכן מתקיימת.**

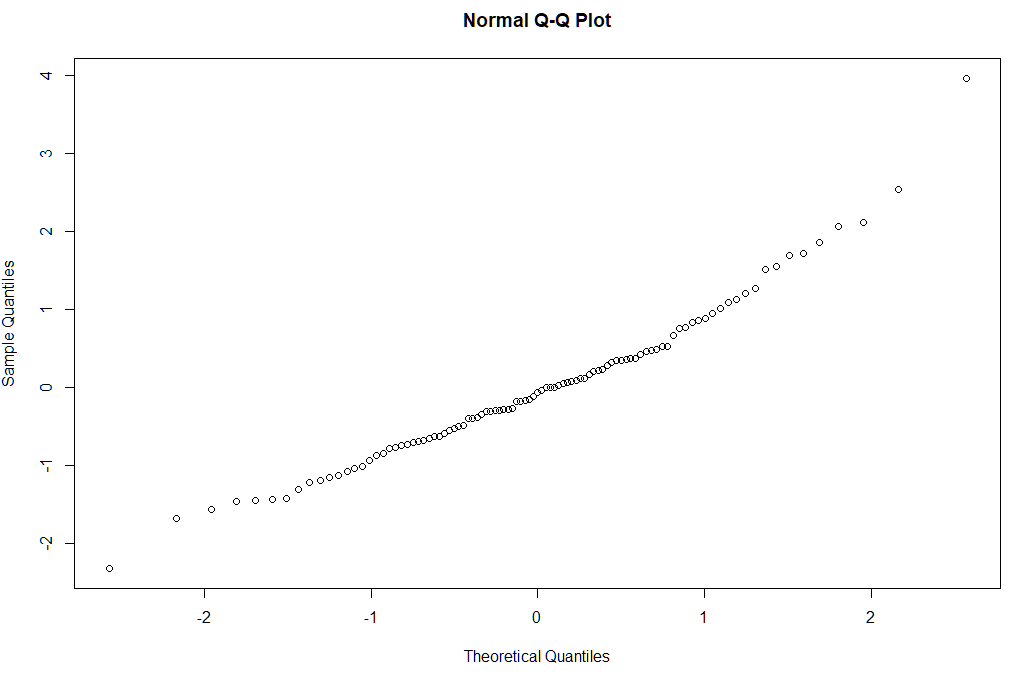
****

### **3.2.2** בדיקת הנחת שוויון שוניות

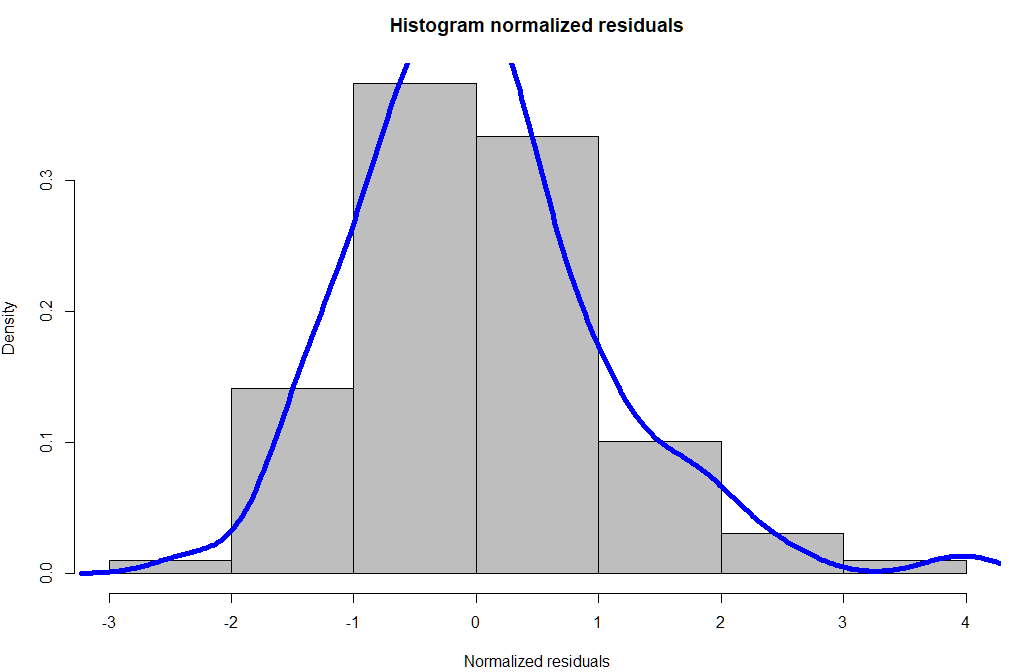
**על מנת לבצע בדיקה זו, נבחן שוב את תרשים פיזור השגיאות המתוקננות ביחס למשתנה המוסבר. על מנת לבחון את שוויון השונויות לעומק,** [**נבצע מבחן סטטיסטי F.**](#_3.2.3) **חילקנו את ערכי הערך המוסבר לשליש עליון ותחתון. עם חלוקה זו, ביצענו את המבחן ורצינו לבדוק את יחס השונויות בר"מ של 5%. ע"פ התוצאות שיצאו – נצטרך לדחות את השערת האפס בר"מ של 5% ונאמר כי לא מתקיים שוויון שונויות ( מצורף נספח המציג את המבחן שבוצע )**

### **3.2.3** בדיקת הנחת הנורמליות של השגיאות

**על מנת לבדוק האם השגיאות במודל מתפלגות נורמלית, נבחן זאת באמצעות תרשים** PLOT**-** QQ**. נעשה זאת במטרה להשוות בין הנתונים הקיימים לבין הערכים שהיינו מתקבלים במידה והם בהכרח היו מתפלגים נורמלית.**

****

**בתרשים ניתן לראות שהתצפיות כן תופסות קו של 45 מעלות אך לא בצורה וודאית**. **על מנת לוודא כי אכן השערתנו להנחת הנורמליות מתקיימת נבחן גם תרשים היסטוגרמה ונצפה לראות צורת פעמון (בדיקת ההנחה מצורפת בקוד)**

****

**כצפוי, לראות עינינו ניתן לראות כן בבירור את צורת הפעמון. על מנת לבצע ווידוא נוסף, נבצע את מבחן** [KS](#_Ks_test) **ברמת מובהקות של 5%. ע"פ המבחן קיבלנו**  **ולכן בר"מ של 5% לא נדחה את השערת האפס ונאמר שהשאריות אכן מתפלגות נורמלית.**

# **[4. שיפור המודל](#_,𝑦.)**

**בסעיף זה נרצה לשפר את המודל שלנו על ידי טרנספורמציות שונות על המשתנה המוסבר. נבחן את הטרנספורמציות השונות של המשתנה המוסבר בעזרת מדד** BIC **ונרצה למזער אותו. נבחן שלושה טרנספורמציות שונות והן:** **, , . כזכור, עבור המודל הנוכחי, ערך מדד ה**BIC **היה** 549.7197 **לפני ביצוע הטרנספורמציות.**

|  |  |
| --- | --- |
| אלגוריתם | מדד BIC סופי |
|  | 241.2274 |
|  | 1111.029 |
|  | 321.9508 |

**ניתן לראות שגם המדד**  **וגם המדד** ***משפרות את ערך המדד*** BIC***. כאשר נרצה להחליט מי מבין המדד עדיף לנו, כמובן שעדיף לבחור את המדד בעל ה-***BIC ***המינימלי אשר* מתקבל לפי מדד**  ***והוא 241.2274* שבאופן חד משמעי עדיף מערך המדד הקודם (549.7197).**

# נספחים

## **2.1**

Text

Description automatically generated

## 2.2

**Text

Description automatically generated**

Text

Description automatically generated

## 2.3

A black screen with white text

Description automatically generated with low confidence

Graphical user interface, text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated with medium confidence Text

Description automatically generated with medium confidence

Graphical user interface, text

Description automatically generated

Graphical user interface, text

Description automatically generated

Text

Description automatically generated

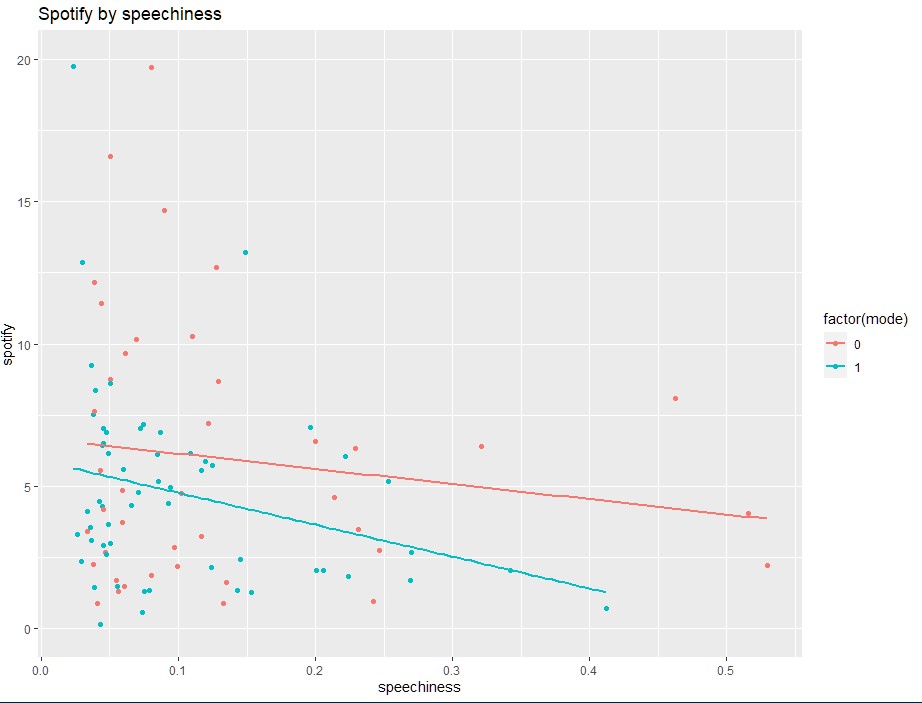
Text

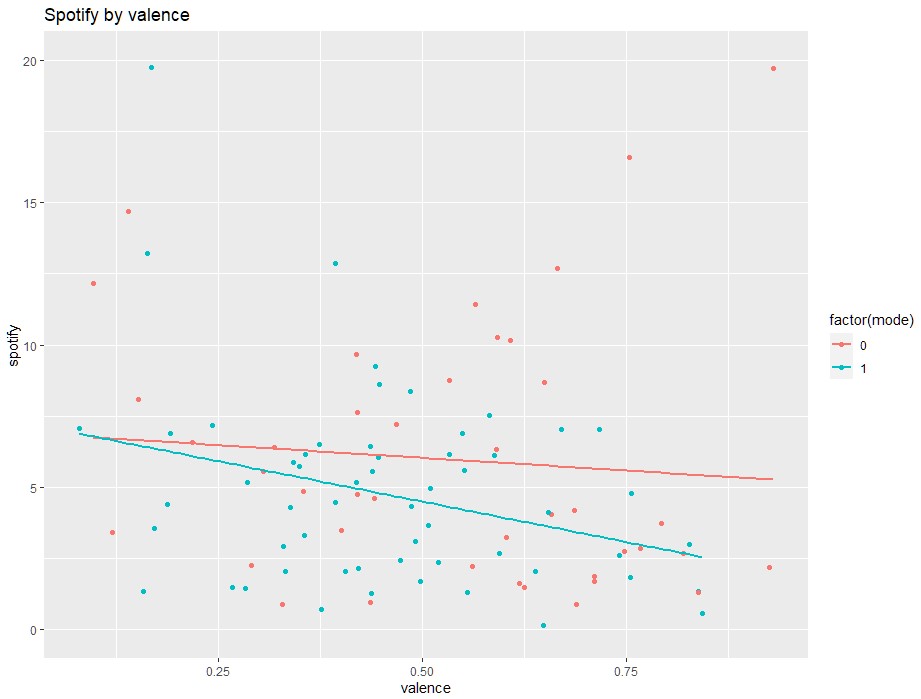
Description automatically generated

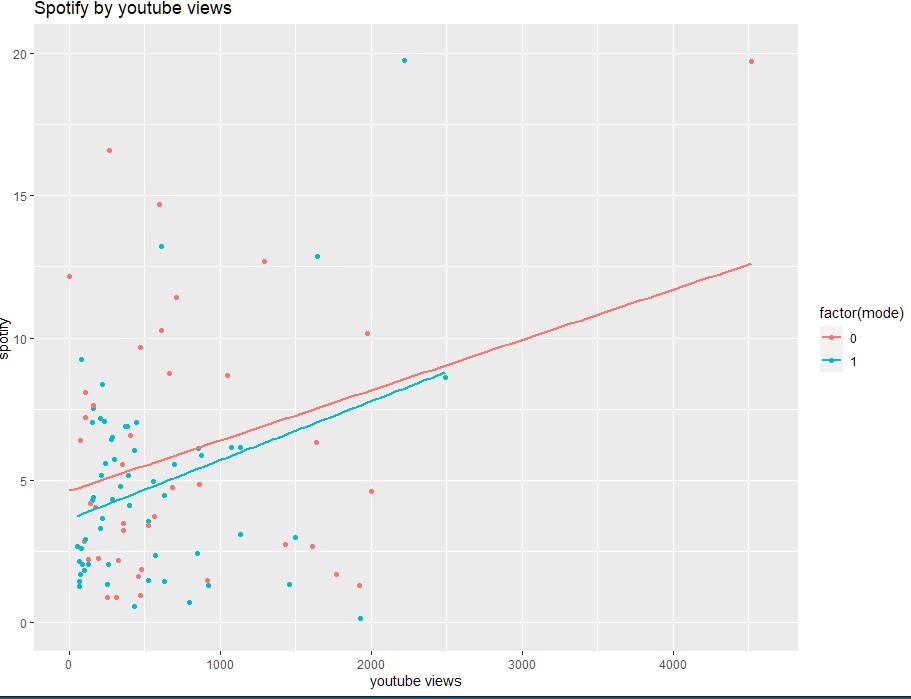
Text

Description automatically generated

## 2.4

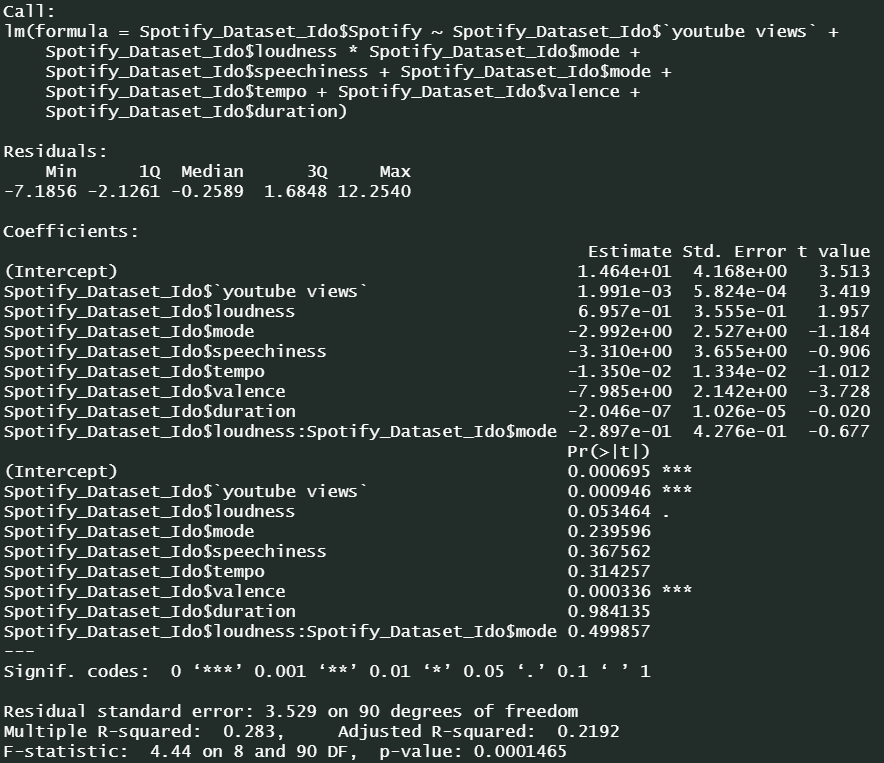






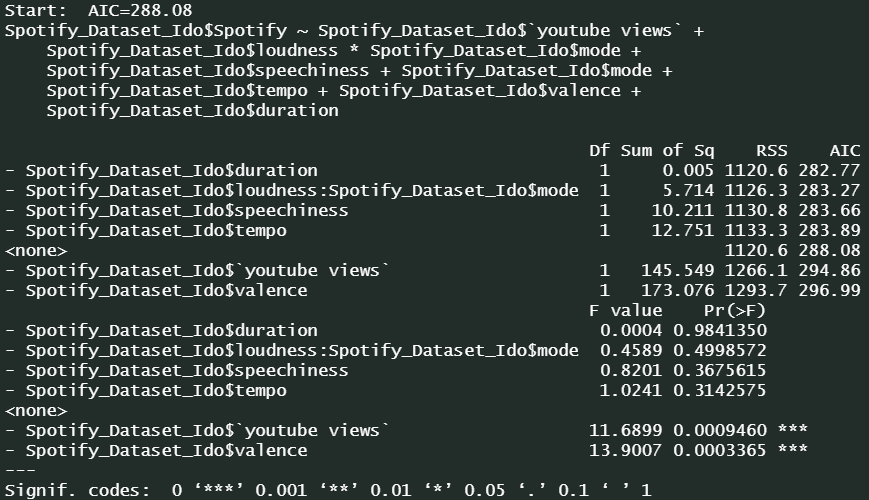
## 3.1

Model summary



## Backward Regression

first step full model



Second step full model

Text

Description automatically generated

Third step full model

Text

Description automatically generated

Fourth step full model

Text

Description automatically generated

Fith step full model

Text

Description automatically generated

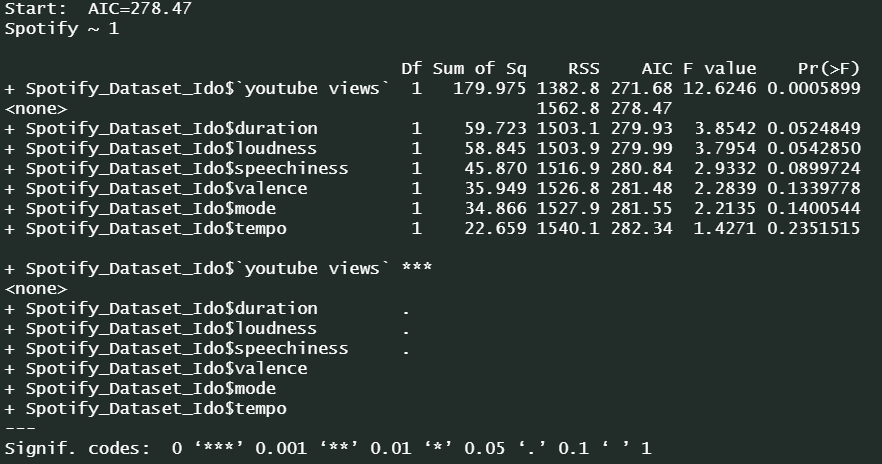
Sixth step full model

Text

Description automatically generated

## Regression forward

First step



Second step

A picture containing text

Description automatically generated

Third step

Text

Description automatically generated

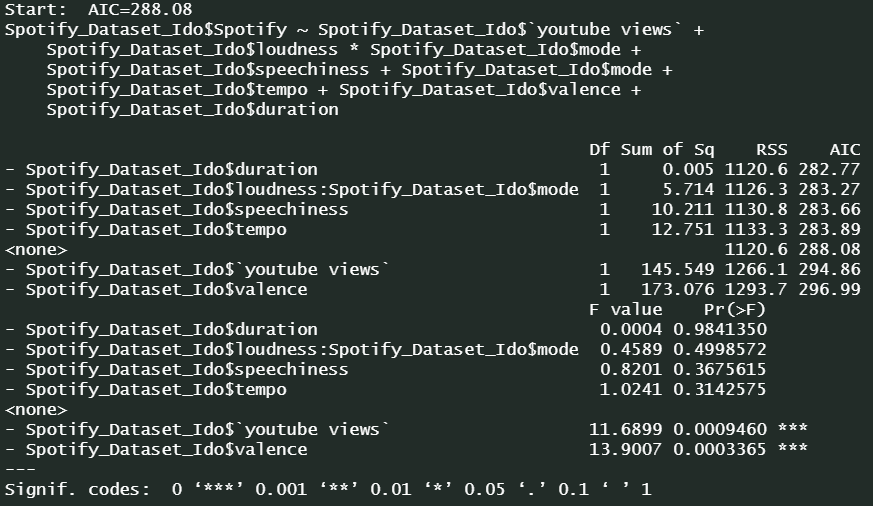
Fourth step

Text

Description automatically generated

## Regression with steps

Firs step



Second step

Text

Description automatically generated

Third step

Text

Description automatically generated

Fourth step

Text

Description automatically generated

Fifth step

Text

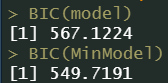
Description automatically generated

Sixth step

Text

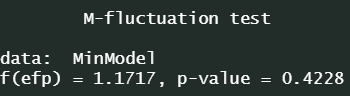
Description automatically generated

## BIC Check

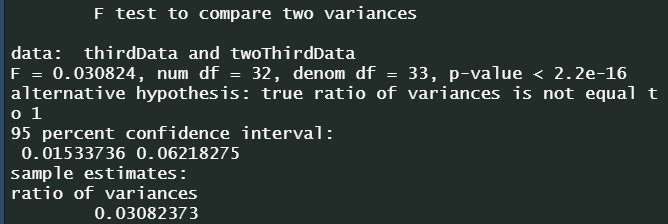


* All different models were the same as MinModel

## Chow test

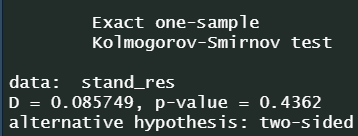


## 3.2.3

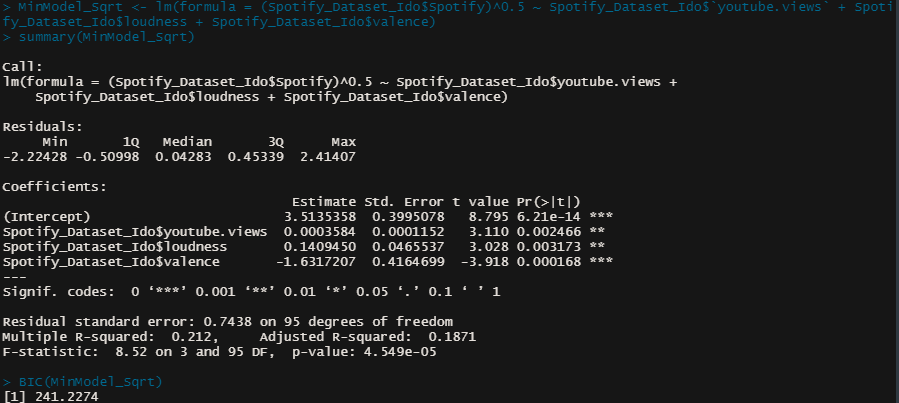


## 3.2.4

## Ks test



## 



תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי