**חלק ב' – פרויקט ברגרסיה לינארית**

קבוצה 15

מגישים:

יניב רוזנר ווקס – 316369792

אמיר יטיב – 207128513

נושא:

Cancer Mortality Rate



***תוכן עניינים***

1. [תקציר מנהלים](#תקציר_מנהלים)........................................................................................................3
2. [עיבוד מקדים](#עיבוד_מקדים)...........................................................................................................4
   1. [הסרה של משתנים](#הסרה_של_משתנים).............................................................................................4
   2. [התאמת משתנים](#התאמת_משתנים)...............................................................................................5
   3. [הגדרת משתני דמה](#הגדרת_משתני_דמה)............................................................................................6
   4. [הגדרת משתני אינטרקציה](#הגדרת_משתני_אינטרקציה)..................................................................................7
3. [התאמת מודל ובדיקת הנחות המודל](#התאמת_המודל_ובדיקת_הנחות)..........................................................................9
   1. [בחירת משתני המודל](#בחירת_משתני_המודל).........................................................................................9
   2. [בדיקת הנחות המודל](#בדיקת_הנחות_המודל).........................................................................................12
4. [שיפור המודל](#שיפור_המודל)..........................................................................................................14
5. [נספחים](#נספחים)..................................................................................................................16
6. תקציר מנהלים

במסגרת הפרויקט בחנו את שיעור התמותה במדינות שונות ממחלת הסרטן, ניתחנו את ההשפעות של מאפיינים שונים במדינות השונות ובאוכלוסייה המתגוררת בהן באמצעות מודלים של רגרסיה לינארית. המטרה שלנו היא יצירת מודל חיזוי טוב ככל הניתן.

בהתחלה, בדקנו האם אפשר לנפות משתנים שלא משפיעים באופן משמעותי מאוד על המשתנה המוסבר שלנו. עבור המשתנים הרציפים שלנו נעזרנו במתאם פירסון בין המשתנים, ולפיו החלטנו אילו משתנים רציפים יהיו חלק מהמודל ואילו לא. בעקבות כך, בחרנו להוריד את שני המשתנים המסבירים הבאים: " Percentage of residents in cities " ו " Depression rate ". עבור המשתנים קטגוריאליים נעזרנו בתרשימי פיזור שבאמצעותם אבחנו את הקשר ביניהם לבין המשתנה המוסבר. לאחר בדיקה זו החלטנו להשאיר את שני המשתנים הקטגוריאליים הבאי " Continent " ו- " Air pollution index".

עבור התאמת המשתנים הקטגוריאליים למודל נעזרנו בתרשימי הפיזור שביצענו והחלטנו לבצע את ההתאמות הבאות : שינוי המשתנה הרציף ”State development rate” למשתנה קטגוריאלי, ובנוסף במשתנה הקטגוריאלי " Air pollution index " בחרנו לבצע איחוד קטגוריאלי בין “High pollution” ו ”Extreme pollution”.

לצורך התאמה של המשתנים הקטגוריאליים למודל הרגרסיה התאמנו אליהם משתני דמה ומשתני אינטראקציה.

השתמשנו באלגוריתמים שונים על המודל המלא (לאחר השינויים בעיבוד המקדים) על מנת לבחור את המשתנים הרלוונטיים לבניית מודל הרגרסיה שלנו. האלגוריתמים שבהם השתמשנו: רגרסיה לפנים, רגרסיה לאחור ורגרסיה בצעדים לפי מדדי AIC ו BIC על מנת למצוא את המודל שנותן את ה Radj2 הטוב ביותר .המודל הטוב ביותר שקיבלנו בשלב זה מבין אלו שנבחנו התקבל מביצוע רגרסיה לאחור עפ"י מדד AIC .

לאחר מכן בחנו עבור מודל הרגרסיה החדש שקיבלנו את קיום הנחות המודל: בדיקת הנחות שוויון שונויות, בדיקת הנחת הנורמליות של השגיאות ובדיקת הנחת הליניאריות. את הנחות אלו בדקנו באמצעות תרשימים שונים - תרשימי פיזור שגיאות בהנחות שוויון שונויות ולינאריות, תרשים QQ-PLOT ותרשים היסטוגרמה. לאחר הבדיקות השונות שעשינו התגלה כי כלל הנחות המודל מתקיימות במודל שלנו.

במטרה לנסות לשפר את מודל הרגרסיה שהתקבל עד כה השתמשנו בטרנספורמציות ליניאריות שונות של המשתנה המוסבר וגילינו כי הטרנספורמציה אשר שיפרה את המודל בצורה המשמעותית ביותר ביחס לשאר היא טרנספורמציית ln(y) . לאחר שיפור קל ב R adj2 ומאחר וכל הנחות המודל מתקיימות, קיבלנו את המודל הסופי והטוב ביותר שבחרנו. [טבלת משתנים – (נספח 1.1 )](#טבלת_משתנים)

1. עיבוד מקדים

**2.1 הסרה של משתנים:**

בחלק זה נבחן את האופציה של הסרת משתנים עבור מודל הרגרסיה החדש שנבנה.

**Text

Description automatically generated with medium confidenceעבור המשתנים הרציפים:** נרצה לבדוק האם להסיר משתנה. לצורך כך אנחנו נשתמש במקדם המתאם של פירסון. מקדם המתאם של פירסון הוא מדד לקשר לינארי בין שני משתנים כמותיים המתקבלים במדגם, באמצעותו נבטא את מידת הקשר בין המשתנים בסקלה שבין (1-) ל1. כאשר 0 מציין חוסר קשר, 1 קשר חיובי מושלם ו(1-) קשר שלילי מושלם.

הנוסחה לחישוב מקדם המתאם של פירסון:

בטבלה הבאה מוצגים מקדמי ההתאמה של פירסון של כל משתנה מסביר רציף אל מול המשתנה המוסבר: [(נספח 2.1)](#מתאם_פרסון)

|  |  |
| --- | --- |
| שם משתנה | מקדם קורלציה |
| Obesity rate | 0.2300299 |
| Smoke rate | 0.2507967 |
| State development index | 0.8186998 |
| Median age | 0.7345901 |
| Average temperature | 0.5040913- |
| Percentage of residents in cities | 0.02034776- |
| Depression rate | 0.0843681 |

ניתן לראות בטבלה שהצגנו כי המשתנים המסבירים הרציפים :

"שיעור הדיכאון" ו"אחוז המתגוררים בעיר" הם בעלי מקדם פירסון נמוך מאד (השואף לאפס) למשתנה המוסבר (שיעור התמותה מסרטן) ולכן נבחר להוריד משתנים אלו. בנוסף, המשתנים המסבירים: "שיעור ההשמנה" ו"שיעור העישון" בעלי מקדם פירסון נמוך יחסית אך החלטנו להשאירם בכל זאת.

**משתנים קטגוריאליים :** נרצה לבדוק האם יש צורך להסיר משתנה קטגוריאלי באמצעות בדיקת קשר ביניהם לבין המשתנה המוסבר. נעשה זאת באמצעות תרשימי Scatterplot שבהם ניתן לראות האם יש קשר בין המשתנה המסביר למוסבר.

Chart, line chart, scatter chart

Description automatically generated

תרשים Scatterplot בין המשתנה הקטגוריאלי "יבשת" לבין המשתנה המוסבר "שיעור התמותה מסרטן". ניתן לראות בבירור כי ביבשת אירופה(0) נמצא שיעור התמותה הגבוה ביותר מסרטן, ביבשת אפריקה(5) נמצא שיעור התמותה הנמוך ביותר. נראה כי קיים קשר בין היבשת לבין שיעור התמותה מסרטן ולכן נבחר להשאיר משתנה זה.

מדד יבשת ביחס לתמותה

Chart, scatter chart

Description automatically generated

תרשים Scatterplot בין המשתנה הקטגוריאלי "מדד זיהום אוויר" לבין המשתנה המוסבר "שיעור התמותה מסרטן". ניתן לראות בבירור כאשר המדד שווה ל-0 שיעור התמותה מסרטן גבוהה יותר מאשר במדינות בהם המדד שווה ל3 או 4. נראה כי קיים קשר בין מדד זיהום האויר לבין שיעור התמותה מסרטן ולכן נבחר להשאיר משתנה זה.

מדד זיהום אוויר ביחס לתמותה

**2.2 התאמת משתנים:**

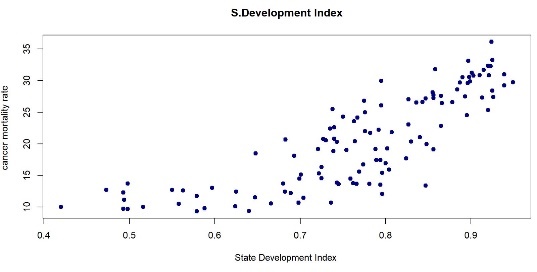
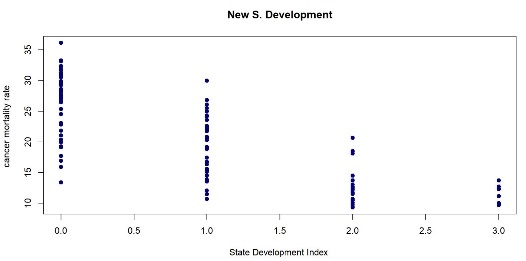
כדי לשפר את מודל הרגרסיה שלנו אנחנו נבצע את הפעולות הבאות :

**משתנה מדד פיתוח המדינה :**נבחר להפוך את המשתנה המסביר הרציף "מדד פיתוח המדינה" למשתנה הקטגוריאלי הבא:

|  |  |
| --- | --- |
| שם | המדד |
| גבוה מאד – 0 | 0.800 – 1.000 |
| גבוה – 1 | 0.700 – 0.799 |
| בינוני – 2 | 0.550 – 0.699 |
| נמוך – 3 | 0.350 – 0.549 |

אנו לא רואים משמעות לכך שהמשתנה הנ"ל הוא משתנה רציף, מה שמשנה לנו זה האם רמת הפיתוח היא גבוה מאד, גבוה, בינונית או נמוכה. ההשוואה המספרית בין מדד פיתוח מדינה של למשל 7.2 ל 7.4 אינה משפיעה באופן משמעותי על המשתנה המוסבר. יהיה משמעותי יותר לבצע השוואות במדד זה על פי הקטגוריות שהגדרנו לעיל הנכללות גם בהגדרת המדד, כפי שגילינו כאשר חקרנו את המשתנה המסביר הזה.

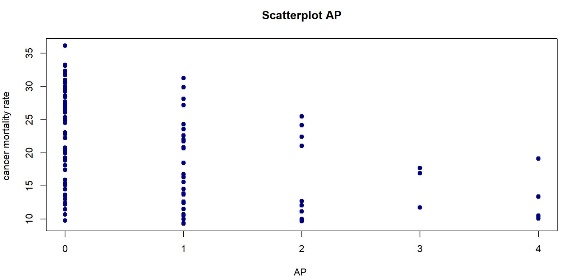
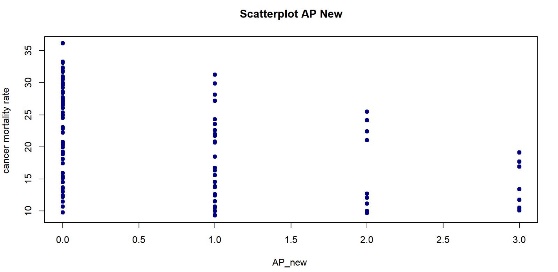
לפני השינוי: אחרי השינוי:



**משתנה אינדקס זיהום אויר :**

החלטנו לאחד את הקטגוריות "High pollution" ו"Extreme pollution" .ניתן לראות בתרשים הscatterplot של מדד זיהום האויר(לפני השינוי) שמספר התצפיות בקטגוריות אלו נמוך ביחס לקטגוריות האחרות. בנוסף, אין הבדל משמעותי בין ערך התצפיות ביחס למשתנה המוסבר. לכן נרצה לשלב ביניהן וליצור קטגוריה אחת שמכלילה את שתיהן. אנחנו מאמינים ששילוב של שתי קטגוריות אלו יכול לתרום להבנה של המשתנה המוסבר בצורה טובה יותר.

לפני השינוי: אחרי השינוי:



**לסיכום,** בחרנו להשאיר את כל הקטגוריות במשתנים כפי שהם מלבד איחוד קטגוריות במשתנה "מדד זיהום האויר" והפיכת המשתנה "מדד פיתוח המדינה" הרציף למשתנה קטגוריאלי.

**2.3 הגדרת משתנה דמה:**

עלמנת לתאר משתנים מסבירים קטגוריאליים נעזר במשתנה דמה. ניצור משתנה אינדיקטור המקבל את הערך 0 או 1 וכך נוכל לייצג קטגוריות שונות. בכל קטגוריה יהיה משתנה אחד אשר יהווה את קבוצת הבסיס (נהייה בה כאשר ערכי שאר המשתנים יהיו אפס), שאר המשתנים יביעו את התרומה השולית על החותך. עבור משתנים מסבירים בעלי יותר משתי קטגוריות, מספר משתני הדמה יהיה (1 ,(𝑛 − כאשר n מייצג את מספר הקטגוריות במשתנה.

המשתנים שעבורם ניצור משתני דמה יהיו :

מדד זיהום אויר-AP

משתנה קטגוריאלי ולו 4 חלופות – זיהום מינורי(0), זיהום נמוך(1), זיהום בינוני(2) , זיהום גבוה(3).

בשביל משתנה זה נצטרך 3 משתני דמה כאשר קבוצת הבסיס תהיה זיהום מינורי(0).

AP3 = 1, High pollution

0, other

AP2 = 1, Medium pollution

0, other

AP1 = 1, Low pollution

0, other

יבשת -Con

משתנה קטגוריאלי ולו 6 חלופות – אירופה(0), דרום אמריקה(1), צפון אמריקה(2) , אוסטרליה(3) , אסיה(4), אפריקה(5).

בשביל משתנה זה נצטרך 5 משתני דמה כאשר קבוצת הבסיס תהיה אירופה(0).

Con3 = 1, Australia

0, other

Con2 = 1, North America

0, other

Con5 = 1, Africa

0, other

Con4 = 1, Asia

0, other

Con1 = 1, South America

0, other

מדד פיתוח המדינה –Sd

בשביל משתנה זה נצטרך 3 משתני דמה כאשר קבוצת הבסיס תהיה Very high (0).

Sd3\*X1 = X1, Low

0, other

Sd2\*X1 = X1, Medium

0, other

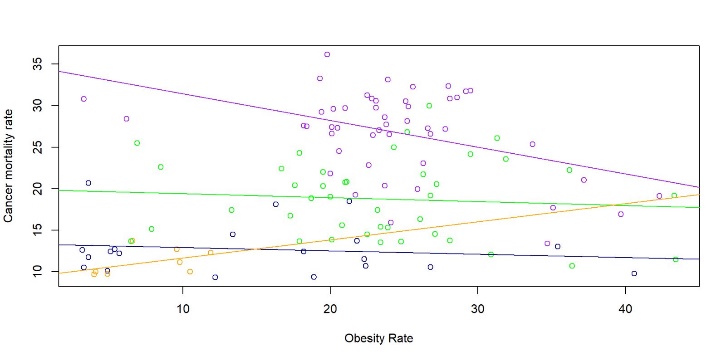
Sd1\*X1 = X1, High

0, other

* 1. **הגדרת משתנה אינטראקציה:**

משתני אינטראקציה מסייע להבין את השפעת המשתנים הקטגוריאליים על שיפוע קו הרגרסיה. רצינו לבחון את הקשר בין משתנה קטגוריאלי "מדד פיתוח המדינה" לבין משתנים רציפים המושפעים מרמות פיתוח שונות. בנוסף רצינו לבחון את הקשר בין משתנה קטגוריאלי "יבשת" לבין משתנים רציפים הקשורים המושפעים ממיקומם הגיאוגרפי. לאחר בחינת הקשרים הללו הצגנו שילובים אפשריים ובחנו את הקשר ביניהם.

**משתנה אינטראקציה 1 : משתנה קטגוריאלי "מדד פיתוח המדינה" ביחד  
 עם משתנה רציף "שיעור ההשמנה".**

ניתן לראות כי המשתנה המסביר הרציף "שיעור ההשמנה" משפיע באופן שונה על המשתנה המוסבר "שיעור התמותה מסרטן" בהתאם למדדי פיתוח המדינה השונים. את השינוי המשמעותי ביותר ניתן לראות עבור מדד פיתוח המדינה – "גבוה מאד" (0). שם יש את השיפוע החד ביותר. שינוי משמעותי נוסף ניתן לראות עבור מדד פיתוח המדינה – "נמוך" (3), בניגוד לשאר המדדים השיפוע שלו עולה. עבור השניים הנוספים, מדדי פיתוח המדינה "גבוה"(1) ו"בינוני" (2) קיבלנו שיפועים יחסית מתונים ודומים אחד לשני. בעקבות הבדלים אלו נרצה להכניס משתנה זה כמשתנה אינטראקציה.

מדד פיתוח המדינה ומדד השמנה

Sd1\*X1 = X1, High

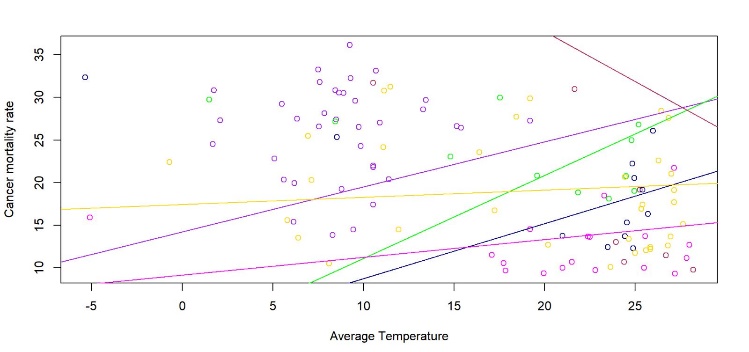
0, other

Sd2\*X1 = X1, Medium

0, other

Sd3\*X1 = X1, Low

0, other

**משתנה אינטראקציה 2 : משתנה קטגוריאלי "יבשת" ביחד עם משתנה רציף "טמפרטורה ממוצעת".**

ניתן לראות כי המשתנה המסביר הרציף "טמפרטורה ממוצעת" משפיע באופן שונה על המשתנה המוסבר "שיעור התמותה מסרטן" בהתאם ליבשות השונות. את השינוי המשמעותי ביותר ניתן לראות עבור יבשת – "אוסטרליה" (3). שם יש את השיפוע החד ביותר ובנוסף זהו השיפוע היחיד שיורד. שינוי משמעותי נוסף הינו עבור צפון אמריקה –(1) ודרום אמריקה(2) ויבשת "אירופה" (0) עבורם ניתן לראות שיפוע חד אך עולה. עבור השניים הנוספים, היבשות "אסיה"(4) ו"אפריקה" (5) קיבלנו שיפועים יחסית מתונים ודומים אחד לשני. בעקבות הבדלים אלו נרצה להכניס משתנה זה כמשתנה אינטראקציה.

מדד טמפרטורה ממוצעת ויבשת

1\*X6 = X6, South America

0, other

3\*X6 = X6, Australia

0, other

2\*X6 = X6, Asia

0, other

5\*X6 = X6, Africa

0, other

4\*X6 = X6, Asia

0, other

בחנו קומבינציה נוספת אפשרית (גיל חציוני עם מדד פיתוח מדינה) אך לא מצאנו הבדל משמעותי בין הקטגוריות השונות ועל כן לא הוספנו

אותם כמשתני אינטראקציה. [(נספח 2.4.1)](#חציוני)

**מודל הרגרסיה עד כה :**

++

+ 1\*X6+ 2\*X6+ 3\*X6+ 4\*X6+ 5\*X6

1. התאמת מודל ובדיקת הנחות המודל

**3.1 בחירת משתנה המודל:**

כדי לבחור את משתני המודל הטובים ביותר, יחד עם השאיפה למודל פשוט ככל האפשר, נשתמש בכמה אלגוריתמים למציאת המודלים הטובים ביותר ונשווה בין המודלים שהתקבלו על פי המדדים. – AIC, BIC, R2*adj.*

רגרסיה לפנים (Forward selection) –

באלגוריתם זה נתחיל במודל ללא משתנים ( החותך בלבד ) ובכל איטרציה נכניס רק משתנה אחד, כאשר המשתנה אשר יכנס למודל הוא המשתנה המובהק ביותר. לאחר מכן, נבדוק הוספה של משתנה נוסף, כאשר גם הוא בעל המובהקות הגבוהה ביותר.

רגרסיה לאחור (Backward elimination) -

באלגוריתם זה נתחיל במודל המלא הכולל את כל המשתנים ובכל איטרציה נסיר רק משתנה אחד, כאשר המשתנה בעל המובהקות הנמוכה ביותר הוא זה שייצא מהמודל. לאחר מכן, נבדוק הסרה של משתנה נוסף, שגם הוא יהיה בעל המובהקות הנמוכה ביותר.

רגרסיה בצעדים (Stepwise regression) –

זהו שילוב של שני האלגוריתמים שציינו לעיל. בכל שלב בודקים האם להכניס או להוציא משתנים מבין המשתנים שנוספו למודל בצעדים הקודמים. משתנה מסביר שנוסף בצעדים הקודמים עשוי להפוך למיותר עקב קשרים עם משתנים מסבירים אחרים שנוספו כעת למודל.

המדדים לבחינת טיב המודל :

מדד AIC – המדד מחושב בעזרת הנוסחה הבאה:



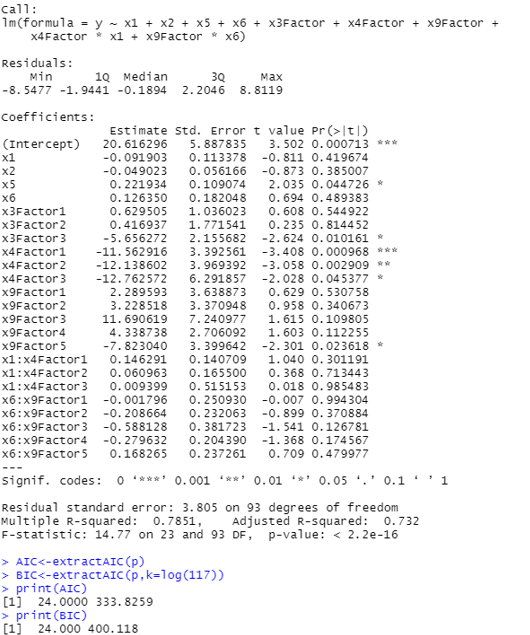
מדד BIC - המדד מחושב בעזרת הנוסחה הבאה:



מדד R2*adj*- אחוז השונות המוסברת במודל תוך כדי התחשבות במס' התצפיות ומס' המשתנים המסבירים. נרצה שהמדד יהיה כמה שיותר קרוב לערך 1. המדד מחושב ע"י הנוסחה הבאה:

בחרנו לבחון את המודל עפ"י המדדים AIC וBIC- כלומר עבור כל אלגוריתם הרצנו פעם אחת כך שהוא ממזער את מדד הAIC- ופעם נוספת כאשר הוא ממזער את מדד הBIC-. הרצנו עבור כל מדד רגרסיה לפנים, רגרסיה לאחור ורגרסיה בצעדים. לבסוף בחרנו את המודל שמקסם את המדד R2*adj.*

נבחן את המדדים עפ"י המודל המלא שהתקבל בסוף שלב עיבוד הנתונים :



AIC = 333.8259

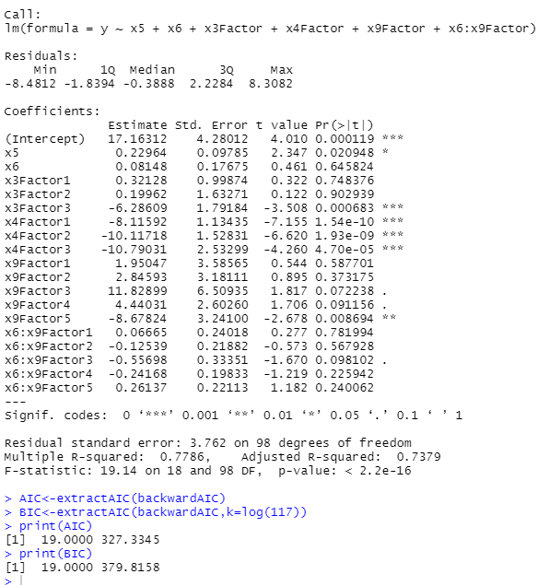
BIC = 400.118

R2*adj* = 0.732

לאחר הרצת האלגוריתמים [(נספח 3.1)](#אלג1) ניתן לראות כי בכל אלגוריתם התקבלו התוצאות הבאות :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | מזעור מדד AIC | | | מזעור מדד BIC | | |
| רגרסיה לפנים | רגרסיה לאחור | רגרסיה בצעדים | רגרסיה לפנים | רגרסיה לאחור | רגרסיה בצעדים |
| R2*adj* | 0.725 | 0.7379 | 0.725 | 0.702 | 0.702 | 0.702 |
| AIC | 328.7819 | 327.3345 | 328.7819 | 333.7366 | 333.7366 | 333.7366 |
| BIC | 367.4523 | 379.8158 | 367.4523 | 358.5962 | 358.5962 | 358.5962 |
| מספר המשתנים המסבירים | 13 | 18 | 13 | 8 | 8 | 8 |

ניתן לראות כי במזעור מדד AIC קיבלו ברגרסיה לפנים וברגרסיה בצעדים מודל זהה ובמזעור מדד BIC קיבלנו את אותו המודל על ידי כלל האלגוריתמים.   
נבחר במודל שהתקבל בו מדד ה R2*adj* הגבוה ביותר(מסומן בסגול בטבלה) שהתקבל במודל רגרסיה לאחור עפ"י מזעור מדד AIC.

פירוט תוצאות המודל לאחר האלגוריתם :

AIC = 327.3345

BIC = 379.8158

R2*adj* = 0.7379

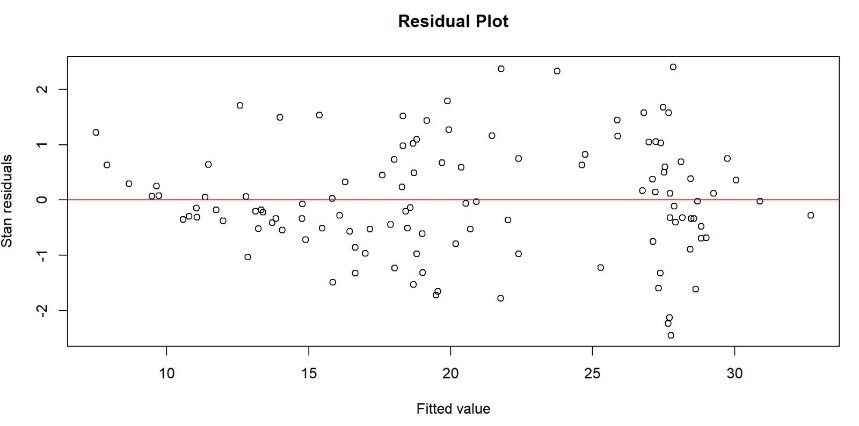
**המודל הסופי שהתקבל הוא:**

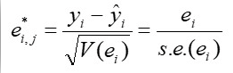
+ 1\*X6+

2\*X6+ 3\*X6+ 4\*X6+ 5\*X6

**3.2 בדיקת הנחות המודל:**

**בדיקת הנחות שוויון שונויות –**

כדי לבדוק האם המודל מקיים את הנחת שוויון השונויות, ראשית נסתכל על **תרשים הפיזור של השגיאות** המתוקננות עפ"י הנוסחה הבאה:



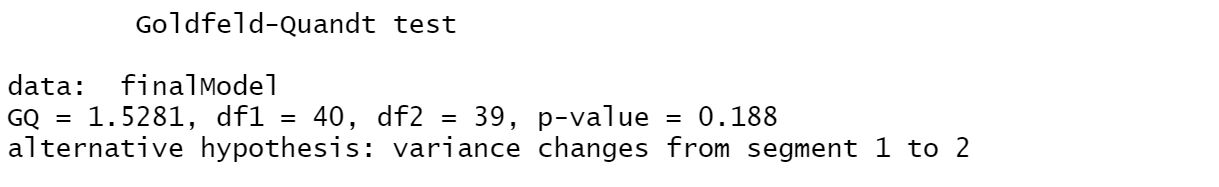
נסתכל על הפיזור סביב הקו y=0 , ניתן לראות ע"פ התרשים כי הפיזור נראה תקין מבחינת שוויון.

תרשים פיזור השגיאות

על מנת לבחון את ההנחה שלנו, נבצע מבחן סטטיסטי F לבדיקת שוויון שונויות בר"מ 5%. ההשערות שלנו יהיו:

:הנחת שוויון השונויות מתקיימת

:הנחת שוויון השוניות אינה מתקיימת



ניתן לראות כי >0.05 𝑝𝑣𝑎𝑙𝑢𝑒 ועל כן נקבל

את השערת האפס ונאמר כי הנחת שוויון

השונויות אכן מתקיימת.

**בדיקת הנחת הנורמאליות –**

על מנת לבחון את הנחת הנורמאליות של השגיאות, ניעזר בשני תרשימים : QQ-PLOT ותרשים היסטוגרמה.

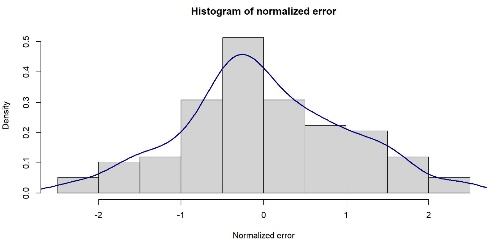
Chart, line chart

Description automatically generated

מתרשים הQQ-PLOT- ניתן לראות כי התצפיות מתלכדות בצורה די אחידה על הקו הלינארי (מלבד הקצוות) על כן, נניח עפ"י תרשים זה שהנחת הנורמאליות של השגיאות אכן מתקיימת.

qq - plot

נראה בתרשים ההיסטוגרמה כי הגרף המתקבל מזכיר את צורת ה"פעמון" של ההתפלגות הנורמלית. על כן גם בעקבות תרשים זה נוכל להניח שהנחת הנורמליות של השגיאות מתקיימת.

על מנת לבחון את ההנחה כי השגיאות מתפלגות נורמלית נבצע שני מבחנים סטטיסטיים בר"מ של 5%.

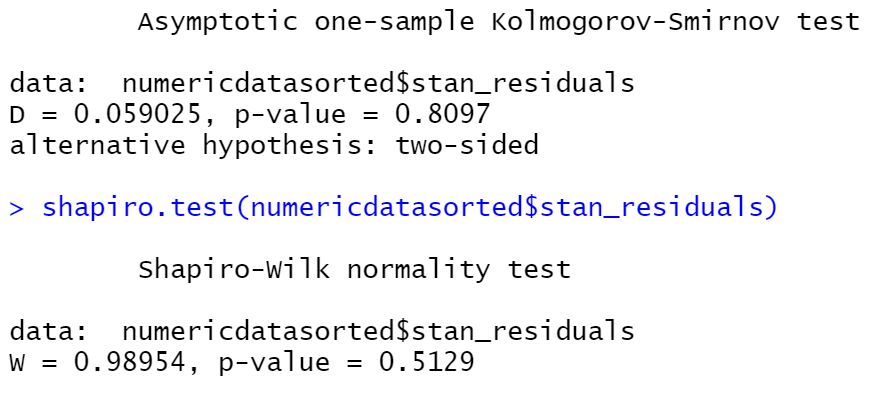
המבחנים שנבצע הם KS וSW-.

ההשערות שלנו יהיו:

היסטוגרמת טעיות מנורמלות

:הנחת שוויון השונויות מתקיימת

:הנחת שוויון השונויות אינה מתקיימת

מבחן KS :

ניתן לראות כי>0.05 ועל כן נקבל את השערת האפס ונאמר כי הנחת הנורמאליות של השגיאות אכן מתקיימת

מבחן SW:

ניתן לראות גם במבחן זה כי>0.05 ועל כן נקבל את השערת האפס ונאמר כי הנחת הנורמאליות של השגיאות אכן מתקיימת.

**לסיכום**, עפ"י שני המבחנים הללו נקבל את השערת האפס ונאמר כי בר"מ של 5% הנחת הנורמאליות של השגיאות מתקיימת בהתאמה להנחה הראשונית שלנו עפ"י התרשימים.

**בדיקת הנחת הליניאריות -**

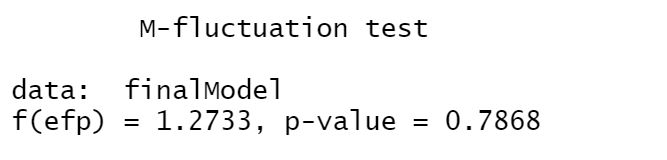
על מנת לבחון את הנחת הליניאריות , תחילה ניעזר בתרשים הפיזור של השאריות המתוקננות ונבחן האם ישנה מגמה בתצפיות.

ניתן לראות בגרף כי התצפיות מפוזרות באופן יחסית אחיד, מצב זה נראה תקין ולכן נניח כי הנחת הלינאריות מתקיימת.

Chart, scatter chart

Description automatically generated

טעויות

על מנת לבחון את השערה זו, נשתמש במבחן Chow test ברמת מובהקות של 5%. ניתן לראות כי במבחן זה כי>0.05 ועל כן נקבל את השערת האפס ונאמר כי הנחת הליניאריות אכן מתקיימת.

**לסיכום**, נאמר כי **כל** ההנחות מתקיימות.

שיפור המודל

בסעיף הקודם, ראינו כי כל הנחות המודל מתקיימות. נרצה לבצע טרנספורמציות שונות על מנת לשפר את

המודל עד כמה שניתן, נבחן מספר טרנספורמציות על המודל המלא בסיום שלב העיבוד המקדים. נראה

מתי מתקבל הערך המקסימלי של *adj*2R .

נרכז בטבלה הבאה את סוג הטרנספורמציה ואת ערך ה*adj*2R לכל אפשרות:

|  |  |
| --- | --- |
| סוג טרנספורמציה | *adj*2R |
| ללא שינוי | 0.7379 |
|  | 0.7483 |
| y2 | 0.6868 |
| ln(y) | 0.7594 |

ניתן לראות כי הערך הגבוה ביותר של *adj*2R מתקבל כאשר מדובר בטרנספורמציית ln(y) ולכן בחרנו

בטרנספורמציה זו לשיפור המודל. נבדוק את הערכים של המודל לאחר ביצוע הטרנספורמציה:

AIC = -370.3155

BIC = -304.0233

R2*adj* = 0.7594

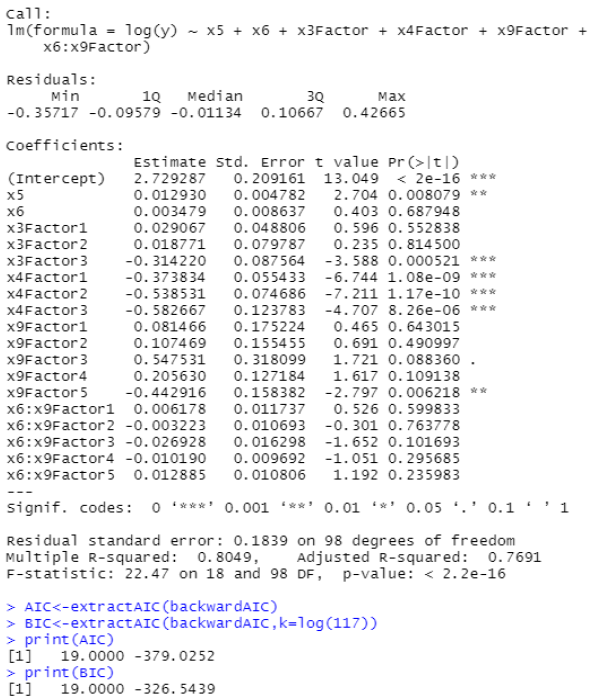
נריץ שוב את האלגוריתם למציאת המודל המיטבי כפי שביצענו בסעיפים הקודמים - עפ"י מזעור המדדים

AIC ו BIC.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | מזעור מדד AIC | | | מזעור מדד BIC | | |
| רגרסיה לפנים | רגרסיה לאחור | רגרסיה בצעדים | רגרסיה לפנים | רגרסיה לאחור | רגרסיה בצעדים |
| R2*adj* | 0.725 | 0.7691 | 0.725 | 0.702 | 0.6866 | 0.702 |
| AIC | 328.7819 | -379.0252 | 328.7819 | 333.7366 | -355.6521 | 333.7366 |
| BIC | 367.4523 | -326.5439 | 367.4523 | 358.5962 | -341.8413 | 358.5962 |
| מספר המשתנים המסבירים | 13 | 18 | 13 | 8 | 4 | 8 |

בחרנו את המזעור עפ"י מדד AIC שנתן את הערך הגבוה ביותר של R *2adj*2 *.* פירוט תוצאות המודל לאחר

האלגוריתם.

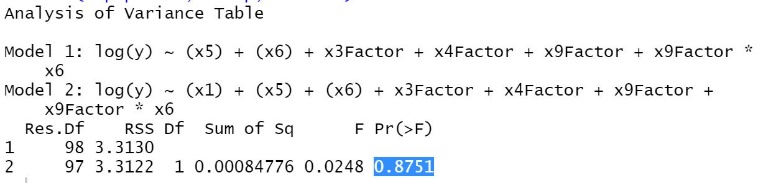


AIC = -379.0252

BIC = -326.5439

R2*adj* = 0.7691

לאחר ביצוע הטרנספורמציה אנו רואים כי לא התקבל שינוי משמעותי במיוחד ולכן נרצה לבדוק אופציה

נוספת לשיפורו. נרצה לשקול בחינה מחודשת של משתנה אשר הושמט מהמודל – “שיעור ההשמנה ”. אנחנו חושבים שמידע זה עשוי להוסיף מידע לגבי המשתנה המוסבר.

בחנו זאת באמצעותמבחן **ANOVA :**

ניתן לראות כי 0.8751 =𝑝𝑣𝑎𝑙𝑢𝑒 גדול מאד - מעל (0.05) ולכן לא ניתן לדחות את ההשערה ששני המודלים

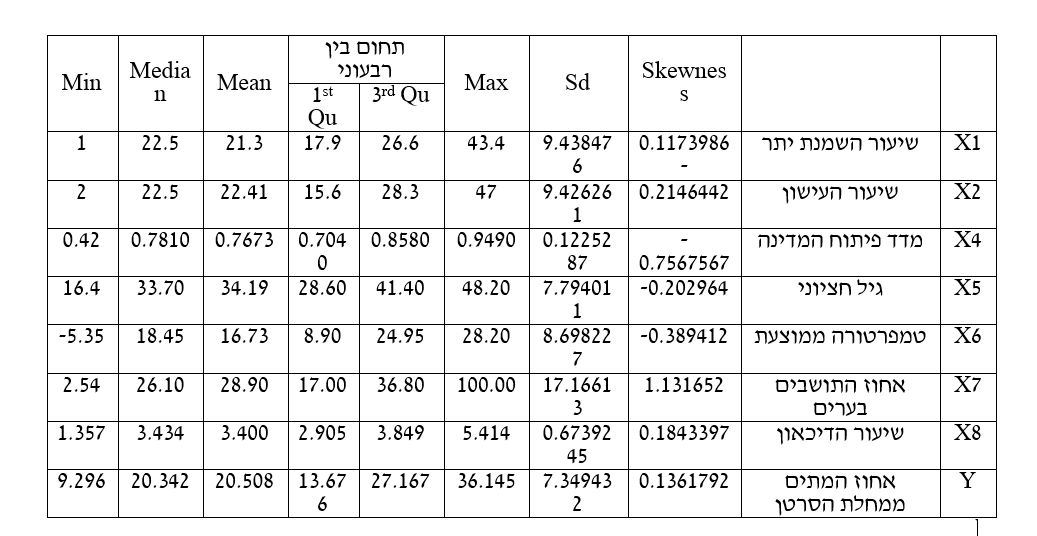
זהים ועל כן נבחר לא להוסיף משתנה זה למודל.

לסיכום, המודל המלא והכי טוב שלנו יהיה:

ln (𝑦)̂= 𝛽0̂+𝛽1̂𝑋5+𝛽2̂𝑋6+𝛽3̂AP1+𝛽4̂AP2+𝛽5̂AP3+𝛽6̂Sd1+𝛽7̂Sd2+𝛽8̂Sd3 +𝛽9̂Con1+𝛽10̂Con2+𝛽11̂Con3+𝛽12̂Con4+𝛽13̂Con5 +𝛽14̂ Con 1\*X6+𝛽15̂ Con 2\*X6+ 𝛽16̂Con 3\*X6+ 𝛽17̂Con 4\*X6+𝛽18̂ Con 5\*X6

נספחים

נספח 1.1 – שמות המשתנים



נספח 2.1– שימוש במתאם פירסון למציאת קורולציה בין המשתנים המסבירים הרציפים למוסבר

A picture containing text

Description automatically generated

Chart, scatter chart

Description automatically generatedנספח 2.2

משתנה רציף "הגיל החציוני" ביחד עם המשתנה

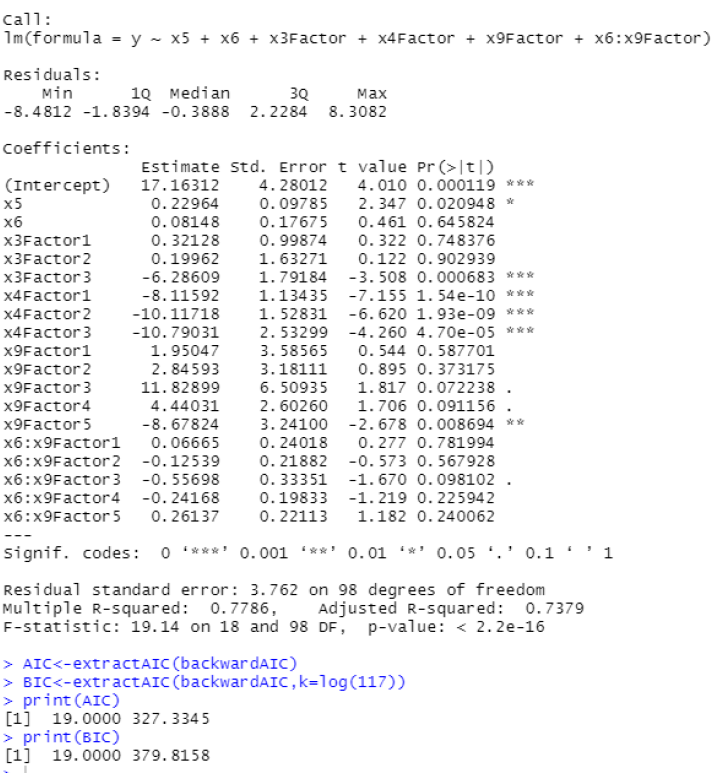
הקטגוריאלי "מדד פיתוח המדינה". ניתן לראות כי כל

השיפועים דיי מתונים ודומים אחד לשני ולכן נבחר לא

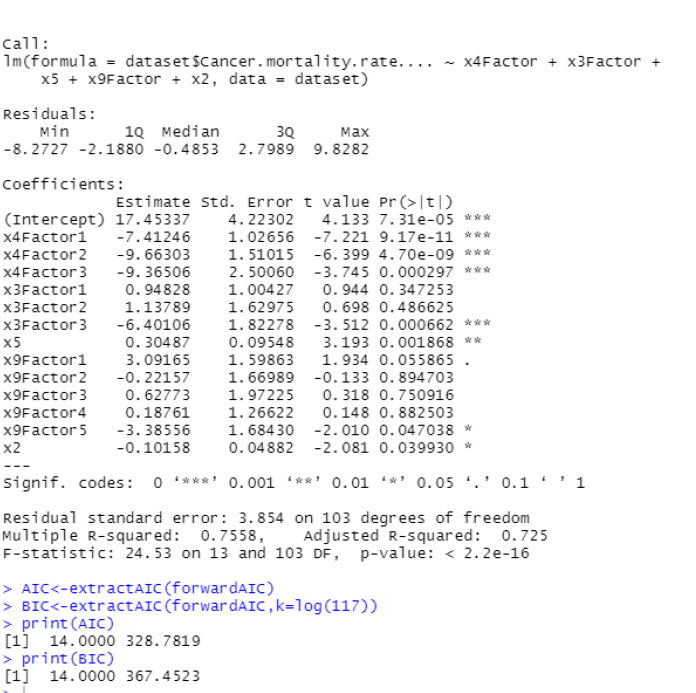
להוסיף אותו כמשתנה אינטראקציה.

נספח 3.1 – תוצאות הרצת הרלגוריטמים

אלגוריתם 1: תוצאות המודל לפי מזעור מדד AIC באמצעות BACKWORD ELIMINATION



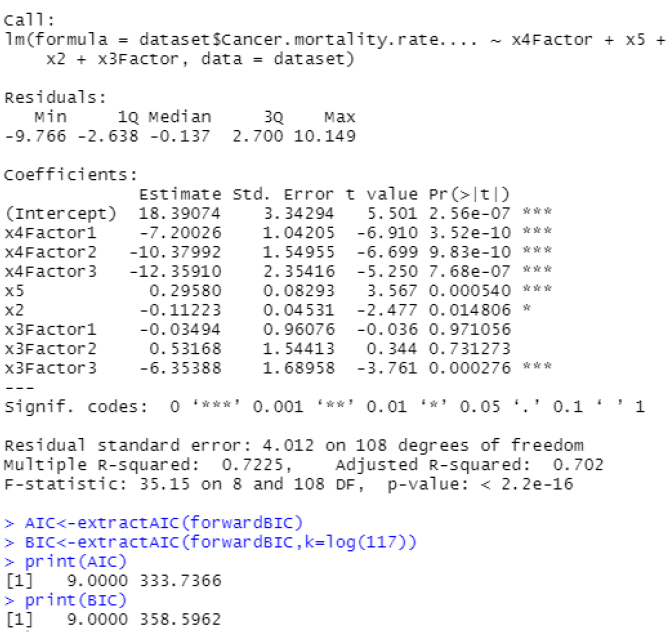
אלגוריתם 2: תוצאות המודל לפי מזעור מדד AICבאמצעות FORWORD ELIMINATION

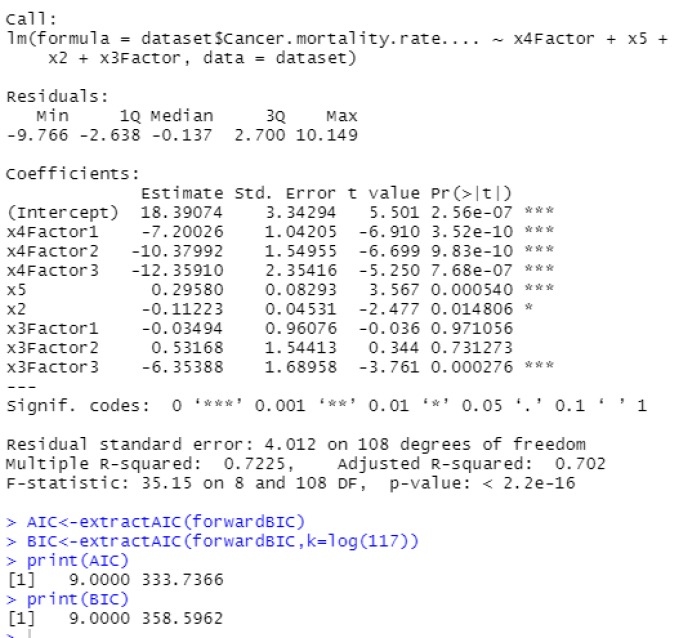


אלגוריתם 3: תוצאות המודל לפי מזעור מדד AIC באמצעות STEPWISE REGRESSION

**Table

Description automatically generated**

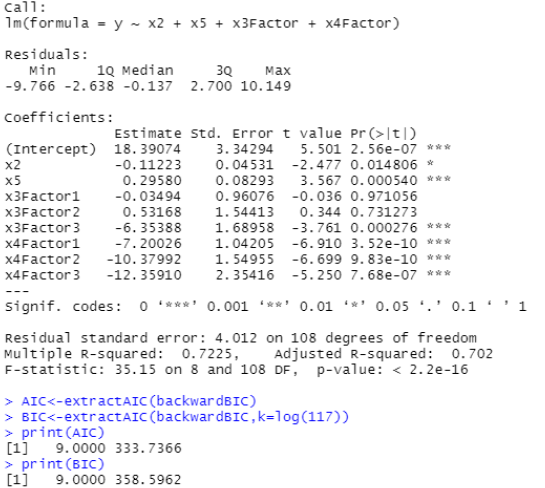
אלגוריתם 4: תוצאות המודל לפי מזעור מדד BIC באמצעות BACKWORD ELIMINATION****

אלגוריתם 5: תוצאות המודל לפי מזעור מדד BIC באמצעות FORWARD ELIMINATION 

אלגוריתם 6: תוצאות המודל לפי מזעור מדד BIC באמצעות STEPWISE REGRESSION **Text

Description automatically generated with medium confidence**

אלגוריתם 6: תוצאות המודל לפי מזעור מדד BICבאמצעות STEPWISE REGRESSION

****