חלק ב' – פרויקט ברגרסיה לינארית

קבוצה 15

מגישים:

יניב רוזנר ווקס – 316369792

אמיר יטיב – 207128513

נושא:

Cancer Mortality Rate

תוכן עניינים

1. 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)

2. עיבוד מקדים

2.1 הסרה של משתנים:

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

עבור 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:

$$R = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

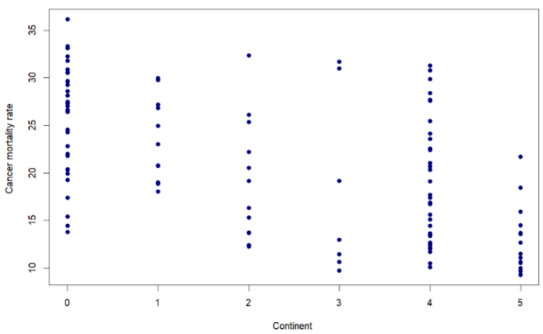
בטבלה הבאה מוצגים מקדמי ההתאמה של פירסון של כל משתנה מסביר רציף אל מול המשתנה המוסבר: (נספח 2.1)

ניתן לראות בטבלה שהצגנו כי המשתנים המסבירים הרציפים:

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

"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.

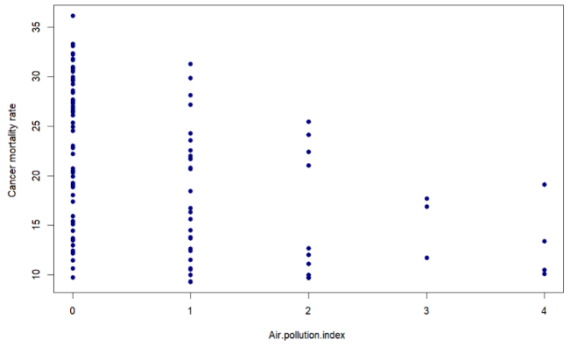
משתנים קטגוריאליים: נרצה לבדוק האם יש צורך להסיר משתנה קטגוריאלי באמצעות בדיקת קשר ביניהם לבין המשתנה המוסבר. נעשה זאת באמצעות תרשימי Scatterplot שבהם ניתן לראות האם יש קשר בין המשתנה המסביר למוסבר.



מדד יבשת ביחס לתמותה

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.

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 :

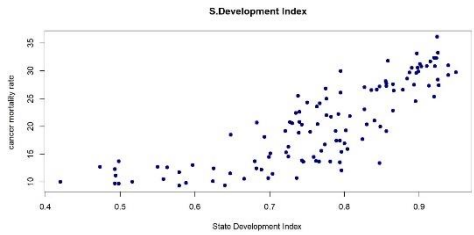
משתנה מדד פיתוח המדינה :

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

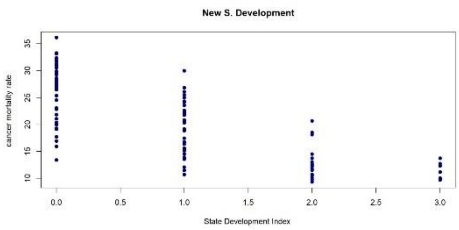
שם	המדד
גבוה מאד – 0	1.000 – 0.800
גבוה – 1	0.799 – 0.700
בינוני – 2	0.699 – 0.550
נמוך – 3	0.549 – 0.350

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 :



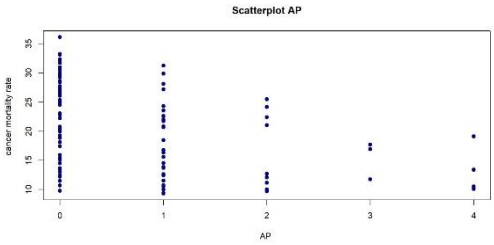
אחרי השינוי :



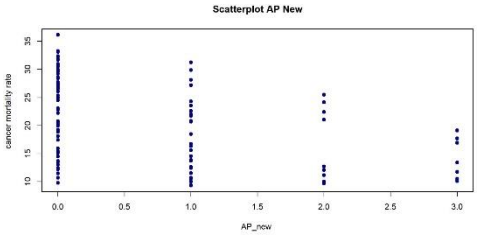
משתנה אינדקס זיהום אויר :

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 :

על מנת 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 $(n-1)$, 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).

$$AP1 = \begin{cases} 1, \text{Low pollution} \\ 0, \text{other} \end{cases} \qquad AP2 = \begin{cases} 1, \text{Medium pollution} \\ 0, \text{other} \end{cases} \qquad AP3 = \begin{cases} 1, \text{High pollution} \\ 0, \text{other} \end{cases}$$

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).

$$Con1 = \begin{cases} 1, \text{South America} \\ 0, \text{other} \end{cases} \qquad Con2 = \begin{cases} 1, \text{North America} \\ 0, \text{other} \end{cases} \qquad Con3 = \begin{cases} 1, \text{Australia} \\ 0, \text{other} \end{cases}$$

$$Con4 = \begin{cases} 1, \text{Asia} \\ 0, \text{other} \end{cases} \qquad Con5 = \begin{cases} 1, \text{Africa} \\ 0, \text{other} \end{cases}$$

מדד פיתוח המדינה –Sd

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

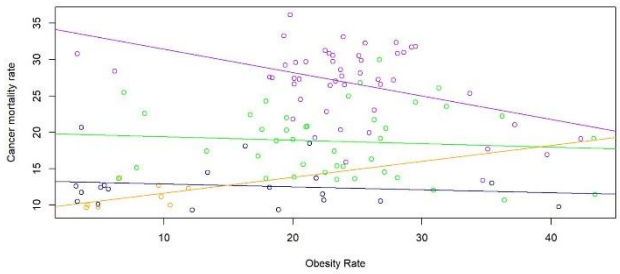
$$Sd1*X1 = \begin{cases} X1, \text{High} \\ 0, \text{other} \end{cases} \qquad Sd2*X1 = \begin{cases} X1, \text{Medium} \\ 0, \text{other} \end{cases} \qquad Sd3*X1 = \begin{cases} X1, \text{Low} \\ 0, \text{other} \end{cases}$$

2.4 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.

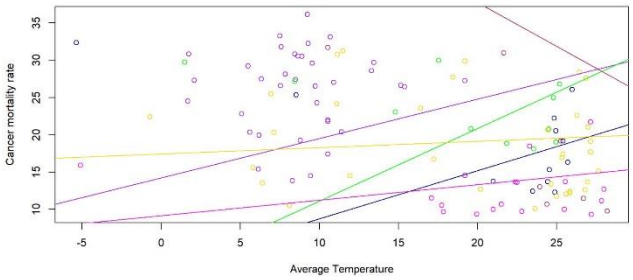


מדד פיתוח המדינה ומדד השמנה

$$Sd1*X1 = \begin{cases} X1, \text{High} \\ 0, \text{other} \end{cases} \qquad Sd2*X1 = \begin{cases} X1, \text{Medium} \\ 0, \text{other} \end{cases} \qquad Sd3*X1 = \begin{cases} X1, \text{Low} \\ 0, \text{other} \end{cases}$$

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

$$\text{Con 1*X6} = \begin{cases} \text{X6, South America} \\ \text{0, other} \end{cases}$$

$$\text{Con2*X6} = \begin{cases} \text{X6, Asia} \\ \text{0, other} \end{cases}$$

$$\text{Con 3*X6} = \begin{cases} \text{X6, Australia} \\ \text{0, other} \end{cases}$$

$$\text{Con 4*X6} = \begin{cases} \text{X6, Asia} \\ \text{0, other} \end{cases}$$

$$\text{Con 5*X6} = \begin{cases} \text{X6, Africa} \\ \text{0, other} \end{cases}$$

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)

מודל הרגרסיה עד כה :

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}X_1 + \widehat{\beta_2}X_2 + \widehat{\beta_3}X_5 + \widehat{\beta_4}X_6$$

$$+\widehat{\beta_5}\text{AP1} + \widehat{\beta_6}\text{AP2} + \widehat{\beta_7}\text{AP3}$$

$$+\widehat{\beta_8}\text{Con1} + \widehat{\beta_9}\text{Con2} + \widehat{\beta_{10}}\text{Con3} + \widehat{\beta_{11}}\text{Con4} + \widehat{\beta_{12}}\text{Con5}$$

$$+\widehat{\beta_{13}}\text{Sd1} + \widehat{\beta_{14}}\text{Sd2} + \widehat{\beta_{15}}\text{Sd3}$$

$$+\widehat{\beta_{16}}\text{Sd1} * \text{X1} + \widehat{\beta_{17}}\text{Sd2} * \text{X1} + \widehat{\beta_{18}}\text{Sd3} * \text{X1}$$

$$+\widehat{\beta_{19}}\ \text{Con 1*X6} + \widehat{\beta_{20}}\ \text{Con 2*X6} + \widehat{\beta_{21}}\text{Con 3*X6} + \widehat{\beta_{22}}\text{Con 4*X6} + \widehat{\beta_{23}}\ \text{Con 5*X6}$$

3. 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, R^2_{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:

$$AIC = n\log\left(\frac{SSE}{n}\right) + 2p$$

מדד BIC - המדד מחושב בעזרת הנוסחה הבאה:

$$BIC = n\log\left(\frac{SSE}{n}\right) + p\log(n)$$

מדד R^2_{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:

$$R^2_{adj} = 1 - \frac{SSE/(n-p)}{SST/(n-1)}$$

בחרנו לבחון את המודל עפ"י המדדים AIC ו-BIC כלומר עבור כל אלגוריתם הרצנו פעם אחת כך שהוא ממזער את מדד ה-AIC ופעם נוספת כאשר הוא ממזער את מדד ה-BIC. הרצנו עבור כל מדד רגרסיה לפנים, רגרסיה לאחור ורגרסיה בצעדים. לבסוף בחרנו את המודל שמקסם את המדד R^2_{adj}.

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

AIC = 333.8259

BIC = 400.118

$R^2_{adj} = 0.732$

```
Call:
lm(formula = y ~ x1 + x2 + x5 + x6 + x3Factor + x4Factor + x9Factor +
    x4Factor * x1 + x9Factor * x6)

Residuals:
    Min      1Q  Median      3Q     Max
-8.5477 -1.9441 -0.1894  2.2046  8.8119

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   20.616296   5.887835   3.502 0.000713 ***
x1            -0.091903   0.113378  -0.811 0.419674
x2            -0.049023   0.056166  -0.873 0.385007
x5             0.221934   0.109074   2.035 0.044726 *
x6             0.126350   0.182048   0.694 0.489383
x3Factor1      0.629505   1.036023   0.608 0.544922
x3Factor2      0.416937   1.771541   0.235 0.814452
x3Factor3     -5.656272   2.155682  -2.624 0.010161 *
x4Factor1    -11.562916   3.392561  -3.408 0.000968 ***
x4Factor2    -12.138602   3.969392  -3.058 0.002909 **
x4Factor3    -12.762572   6.291857  -2.028 0.045377 *
x9Factor1      2.289593   3.638873   0.629 0.530758
x9Factor2      3.228518   3.370948   0.958 0.340673
x9Factor3     11.690619   7.240977   1.615 0.109805
x9Factor4      4.338738   2.706092   1.603 0.112255
x9Factor5     -7.823040   3.399642  -2.301 0.023618 *
x1:x4Factor1   0.146291   0.140709   1.040 0.301191
x1:x4Factor2   0.060963   0.165500   0.368 0.713443
x1:x4Factor3   0.009399   0.515153   0.018 0.985483
x6:x9Factor1  -0.001796   0.250930  -0.007 0.994304
x6:x9Factor2  -0.208664   0.232063  -0.899 0.370884
x6:x9Factor3  -0.588128   0.381723  -1.541 0.126781
x6:x9Factor4  -0.279632   0.204390  -1.368 0.174567
x6:x9Factor5   0.168265   0.237261   0.709 0.479977
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.805 on 93 degrees of freedom
Multiple R-squared:  0.7851,    Adjusted R-squared:  0.732
F-statistic: 14.77 on 23 and 93 DF,  p-value: < 2.2e-16

> AIC<-extractAIC(p)
> BIC<-extractAIC(p,k=log(117))
> print(AIC)
[1]  24.0000 333.8259
> print(BIC)
[1]  24.000 400.118
```

לאחר הרצת האלגוריתמים (נספח 3.1) 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
R^2_{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 הR^2_{adj} הגבוה ביותר(מסומן בסגול בטבלה) שהתקבל במודל רגרסיה לאחור עפ"י מזעור מדד AIC.

פירוט תוצאות המודל לאחר האלגוריתם :

```
Call:
lm(formula = y ~ x5 + x6 + x3Factor + x4Factor + x9Factor + x6:x9Factor)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.4812 -1.8394 -0.3888  2.2284  8.3082 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     17.16312    4.28012   4.010 0.000119 ***
x5               0.22964    0.09785   2.347 0.020948 *  
x6               0.08148    0.17675   0.461 0.645824    
x3Factor1        0.32128    0.99874   0.322 0.748376    
x3Factor2        0.19962    1.63271   0.122 0.902939    
x3Factor3       -6.28609    1.79184  -3.508 0.000683 ***
x4Factor1       -8.11592    1.13435  -7.155 1.54e-10 ***
x4Factor2      -10.11718    1.52831  -6.620 1.93e-09 ***
x4Factor3      -10.79031    2.53299  -4.260 4.70e-05 ***
x9Factor1        1.95047    3.58565   0.544 0.587701    
x9Factor2        2.84593    3.18111   0.895 0.373175    
x9Factor3       11.82899    6.50935   1.817 0.072238 .  
x9Factor4        4.44031    2.60260   1.706 0.091156 .  
x9Factor5       -8.67824    3.24100  -2.678 0.008694 ** 
x6:x9Factor1     0.06665    0.24018   0.277 0.781994    
x6:x9Factor2    -0.12539    0.21882  -0.573 0.567928    
x6:x9Factor3    -0.55698    0.33351  -1.670 0.098102 .  
x6:x9Factor4    -0.24168    0.19833  -1.219 0.225942    
x6:x9Factor5     0.26137    0.22113   1.182 0.240062    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.762 on 98 degrees of freedom
Multiple R-squared:  0.7786,	Adjusted R-squared:  0.7379 
F-statistic: 19.14 on 18 and 98 DF,  p-value: < 2.2e-16

> AIC<-extractAIC(backwardAIC)
> BIC<-extractAIC(backwardAIC,k=log(117))
> print(AIC)
[1]  19.0000 327.3345
> print(BIC)
[1]  19.0000 379.8158
> 
```

AIC = 327.3345

BIC = 379.8158

R^2_{adj} = 0.7379

<u>המודל הסופי שהתקבל הוא:</u>

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}X_5 + \widehat{\beta_2}X_6 + \widehat{\beta_3}\text{AP1} + \widehat{\beta_4}\text{AP2} + \widehat{\beta_5}\text{AP3} + \widehat{\beta_6}\text{Sd1} + \widehat{\beta_7}\text{Sd2} + \widehat{\beta_8}\text{Sd3}$$

$$+\widehat{\beta_9}\text{Con1} + \widehat{\beta_{10}}\text{Con2} + \widehat{\beta_{11}}\text{Con3} + \widehat{\beta_{12}}\text{Con4} + \widehat{\beta_{13}}\text{Con5} + \widehat{\beta_{14}}\ \text{Con 1*X6} + \widehat{\beta_{15}}\ \text{Con}$$

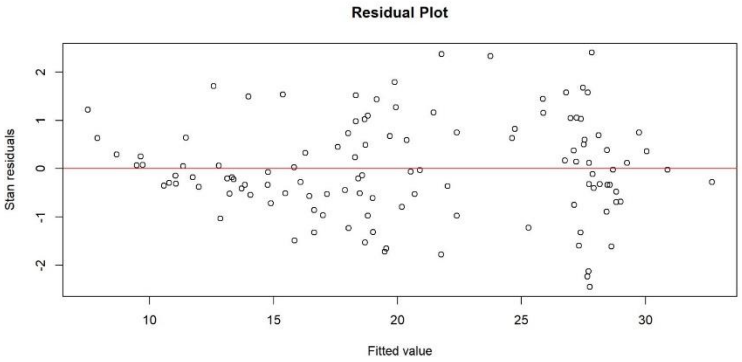
$$\text{2*X6} + \widehat{\beta_{16}}\text{Con 3*X6} + \widehat{\beta_{17}}\text{Con 4*X6} + \widehat{\beta_{18}}\ \text{Con 5*X6}$$

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:

$$e^*_{i,j} = \frac{y_i - \hat{y}_i}{\sqrt{V(e_i)}} = \frac{e_i}{s.e.(e_i)}$$



תרשים פיזור השגיאות

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

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

H_0: הנחת שוויון השונויות מתקיימת
H_1: הנחת שוויון השוניות אינה מתקיימת

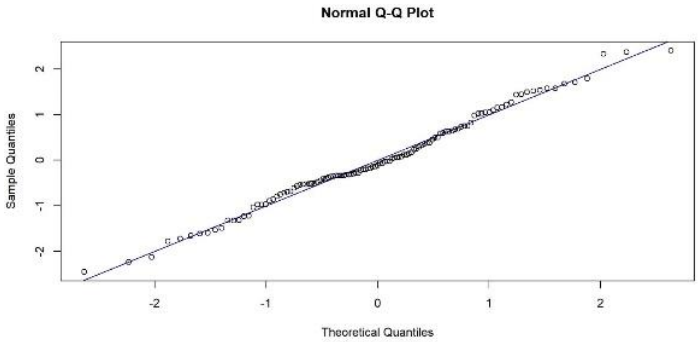
```
        Goldfeld-Quandt test

data:  finalModel
GQ = 1.5281, df1 = 40, df2 = 39, p-value = 0.188
alternative hypothesis: variance changes from segment 1 to 2
```

ניתן לראות כי p_{value} 0.05< ועל כן נקבל את השערת האפס ונאמר כי הנחת שוויון השונויות אכן מתקיימת.

בדיקת הנחת הנורמאליות –

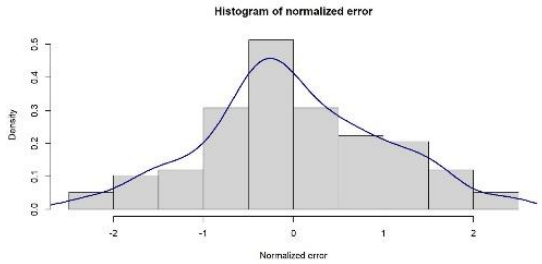
על מנת לבחון את הנחת הנורמאליות של השגיאות, ניעזר בשני תרשימים: QQ-PLOT ותרשים היסטוגרמה.



gq - plot

מתרשים ה-QQ-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.

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.



היסטוגרמת טעיות מנורמלות

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

המבחנים שנבצע הם KS ו-SW.

ההשערות שלנו יהיו:

הנחת שוויון השונויות מתקיימת: H_0

הנחת שוויון השונויות אינה מתקיימת: H_1

מבחן KS:

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  numericdatasorted$stan_residuals
D = 0.059025, p-value = 0.8097
alternative hypothesis: two-sided
```

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

מבחן SW:

```
> shapiro.test(numericdatasorted$stan_residuals)

        Shapiro-Wilk normality test

data:  numericdatasorted$stan_residuals
W = 0.98954, p-value = 0.5129
```

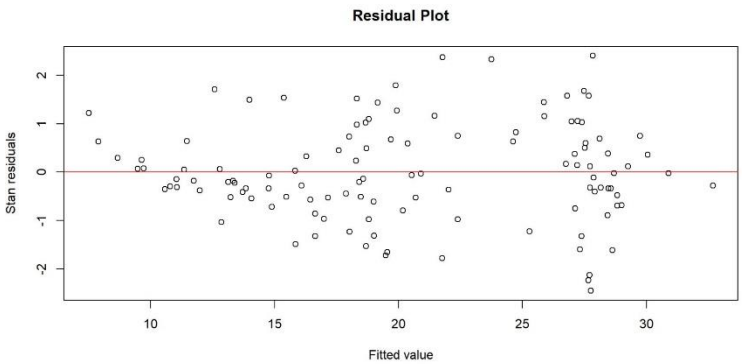
ניתן לראות גם במבחן זה כי p_{value} >0.05 ועל כן נקבל את השערת האפס ונאמר כי הנחת הנורמאליות של השגיאות אכן מתקיימת.

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.



טעויות

```
M-fluctuation test

data:  finalModel
f(efp) = 1.2733, p-value = 0.7868
```

על מנת לבחון את השערה זו, נשתמש במבחן Chow test ברמת מובהקות של 5%. ניתן לראות כי במבחן זה כי p_{value} >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 R^2_{adj} .
נרכז בטבלה הבאה את סוג הטרנספורמציה ואת ערך הR^2_{adj} לכל אפשרות:

סוג טרנספורמציה	R^2_{adj}
ללא שינוי	0.7379
$\sqrt{y}$	0.7483
y^2	0.6868
ln(y)	0.7594

ניתן לראות כי הערך הגבוה ביותר של R^2_{adj} 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

R^2_{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
R^2_{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^2_{adj2} . פירוט תוצאות המודל לאחר האלגוריתם.

```
Call:
lm(formula = log(y) ~ x5 + x6 + x3Factor + x4Factor + x9Factor +
    x6:x9Factor)

Residuals:
     Min       1Q   Median       3Q      Max
-0.35717 -0.09579 -0.01134  0.10667  0.42665

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.729287   0.209161  13.049  < 2e-16 ***
x5             0.012930   0.004782   2.704 0.008079 **
x6             0.003479   0.008637   0.403 0.687948
x3Factor1      0.029067   0.048806   0.596 0.552838
x3Factor2      0.018771   0.079787   0.235 0.814500
x3Factor3     -0.314220   0.087564  -3.588 0.000521 ***
x4Factor1     -0.373834   0.055433  -6.744 1.08e-09 ***
x4Factor2     -0.538531   0.074686  -7.211 1.17e-10 ***
x4Factor3     -0.582667   0.123783  -4.707 8.26e-06 ***
x9Factor1      0.081466   0.175224   0.465 0.643015
x9Factor2      0.107469   0.155455   0.691 0.490997
x9Factor3      0.547531   0.318099   1.721 0.088360 .
x9Factor4      0.205630   0.127184   1.617 0.109138
x9Factor5     -0.442916   0.158382  -2.797 0.006218 **
x6:x9Factor1   0.006178   0.011737   0.526 0.599833
x6:x9Factor2  -0.003223   0.010693  -0.301 0.763778
x6:x9Factor3  -0.026928   0.016298  -1.652 0.101693
x6:x9Factor4  -0.010190   0.009692  -1.051 0.295685
x6:x9Factor5   0.012885   0.010806   1.192 0.235983
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1839 on 98 degrees of freedom
Multiple R-squared:  0.8049,    Adjusted R-squared:  0.7691
F-statistic: 22.47 on 18 and 98 DF,  p-value: < 2.2e-16

> AIC<-extractAIC(backwardAIC)
> BIC<-extractAIC(backwardAIC,k=log(117))
> print(AIC)
[1]   19.0000 -379.0252
> print(BIC)
[1]   19.0000 -326.5439
```

AIC = -379.0252

BIC = -326.5439

$R^2_{adj} = 0.7691$

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 באמצעות מבחן: **ANOVA**

```
Analysis of Variance Table

Model 1: log(y) ~ (x5) + (x6) + x3Factor + x4Factor + x9Factor + x9Factor *
    x6
Model 2: log(y) ~ (x1) + (x5) + (x6) + x3Factor + x4Factor + x9Factor +
    x9Factor * x6
  Res.Df    RSS Df  Sum of Sq      F Pr(>F)
1     98 3.3130
2     97 3.3122  1 0.00084776 0.0248 0.8751
```

ניתן לראות כי $p_{value} = 0.8751$ 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:

$$\ln(\hat{y}) = \hat{\beta_0} + \hat{\beta_1}X_5 + \hat{\beta_2}X_6 + \hat{\beta_3}AP1 + \hat{\beta_4}AP2 + \hat{\beta_5}AP3 + \hat{\beta_6}Sd1 + \hat{\beta_7}Sd2 + \hat{\beta_8}Sd3 + \hat{\beta_9}Con1 + \hat{\beta_{10}}Con2 + \hat{\beta_{11}}Con3 + \hat{\beta_{12}}Con4 + \hat{\beta_{13}}Con5 + \hat{\beta_{14}}Con_{1}*X_6 + \hat{\beta_{15}}Con_{2}*X_6 + \hat{\beta_{16}}Con_{3}*X_6 + \hat{\beta_{17}}Con_{4}*X_6 + \hat{\beta_{18}}Con_{5}*X_6$$

נספחים

נספח 1.1 – שמות המשתנים

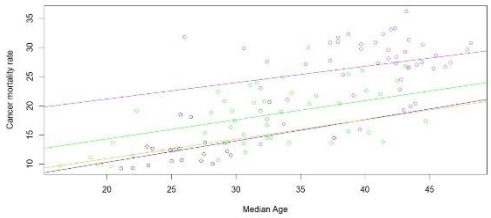
Min	Median	Mean	תחום בין רבעוני		Max	Sd	Skewness		
			1st Qu	3rd Qu					
1	22.5	21.3	17.9	26.6	43.4	9.438476	0.1173986-	שיעור השמנת יתר	X1
2	22.5	22.41	15.6	28.3	47	9.426261	0.2146442	שיעור העישון	X2
0.42	0.7810	0.7673	0.7040	0.8580	0.9490	0.1225287	-0.7567567	מדד פיתוח המדינה	X4
16.4	33.70	34.19	28.60	41.40	48.20	7.794011	-0.202964	גיל חציוני	X5
-5.35	18.45	16.73	8.90	24.95	28.20	8.698227	-0.389412	טמפרטורה ממוצעת	X6
2.54	26.10	28.90	17.00	36.80	100.00	17.16613	1.131652	אחוז התושבים בערים	X7
1.357	3.434	3.400	2.905	3.849	5.414	0.6739245	0.1843397	שיעור הדיכאון	X8
9.296	20.342	20.508	13.676	27.167	36.145	7.349432	0.1361792	אחוז המתים ממחלת הסרטן	Y

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

```
> cor(numericdatasorted$Obesity.rate, numericdatasorted$Cancer.mortality.rate.... , method = c("pearson"))
[1] 0.2300299
> cor(numericdatasorted$Smoke.rate, numericdatasorted$Cancer.mortality.rate.... , method = c("pearson"))
[1] 0.2507967
> cor(numericdatasorted$State.Development.Index..0.1., numericdatasorted$Cancer.mortality.rate.... , method =
 c("pearson"))
[1] 0.8186998
> cor(numericdatasorted$Median.age, numericdatasorted$Cancer.mortality.rate.... , method = c("pearson"))
[1] 0.7345901
> cor(numericdatasorted$Average.temperature, numericdatasorted$Cancer.mortality.rate.... , method = c("pearso
n"))
[1] -0.5040913
> cor(numericdatasorted$Percentage.of.residents.in.cities, numericdatasorted$Cancer.mortality.rate.... , metho
d = c("pearson"))
[1] -0.02034776
> cor(numericdatasorted$Depression.rate...., numericdatasorted$Cancer.mortality.rate.... , method = c("pearso
n"))
[1] 0.0843681
```

נספח 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

```
Call:
lm(formula = y ~ x5 + x6 + x3Factor + x4Factor + x9Factor + x6:x9Factor)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.4812 -1.8394 -0.3888  2.2284  8.3082 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   17.16312    4.28012   4.010 0.000119 ***
x5             0.22964    0.09785   2.347 0.020948 *  
x6             0.08148    0.17675   0.461 0.645824    
x3Factor1      0.32128    0.99874   0.322 0.748376    
x3Factor2      0.19962    1.63271   0.122 0.902939    
x3Factor3     -6.28609    1.79184  -3.508 0.000683 ***
x4Factor1     -8.11592    1.13435  -7.155 1.54e-10 ***
x4Factor2    -10.11718    1.52831  -6.620 1.93e-09 ***
x4Factor3    -10.79031    2.53299  -4.260 4.70e-05 ***
x9Factor1      1.95047    3.58565   0.544 0.587701    
x9Factor2      2.84593    3.18111   0.895 0.373175    
x9Factor3     11.82899    6.50935   1.817 0.072238 .  
x9Factor4      4.44031    2.60260   1.706 0.091156 .  
x9Factor5     -8.67824    3.24100  -2.678 0.008694 ** 
x6:x9Factor1   0.06665    0.24018   0.277 0.781994    
x6:x9Factor2  -0.12539    0.21882  -0.573 0.567928    
x6:x9Factor3  -0.55698    0.33351  -1.670 0.098102 .  
x6:x9Factor4  -0.24168    0.19833  -1.219 0.225942    
x6:x9Factor5   0.26137    0.22113   1.182 0.240062    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.762 on 98 degrees of freedom
Multiple R-squared:  0.7786,	Adjusted R-squared:  0.7379 
F-statistic: 19.14 on 18 and 98 DF,  p-value: < 2.2e-16

> AIC<-extractAIC(backwardAIC)
> BIC<-extractAIC(backwardAIC,k=log(117))
> print(AIC)
[1]  19.0000 327.3345
> print(BIC)
[1]  19.0000 379.8158
```

אלגוריתם 2: תוצאות המודל לפי מזעור מדד AIC באמצעות FORWORD ELIMINATION

```
Residuals:
    Min      1Q  Median      3Q     Max 
-8.2727 -2.1880 -0.4853  2.7989  9.8282 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 17.45337    4.22302   4.133 7.31e-05 ***
x4Factor1   -7.41246    1.02656  -7.221 9.17e-11 ***
x4Factor2   -9.66303    1.51015  -6.399 4.70e-09 ***
x4Factor3   -9.36506    2.50060  -3.745 0.000297 ***
x3Factor1    0.94828    1.00427   0.944 0.347253    
x3Factor2    1.13789    1.62975   0.698 0.486625    
x3Factor3   -6.40106    1.82278  -3.512 0.000662 ***
x5           0.30487    0.09548   3.193 0.001868 ** 
x9Factor1    3.09165    1.59863   1.934 0.055865 .  
x9Factor2   -0.22157    1.66989  -0.133 0.894703    
x9Factor3    0.62773    1.97225   0.318 0.750916    
x9Factor4    0.18761    1.26622   0.148 0.882503    
x9Factor5   -3.38556    1.68430  -2.010 0.047038 *  
x2          -0.10158    0.04882  -2.081 0.039930 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.854 on 103 degrees of freedom
Multiple R-squared:  0.7558,	Adjusted R-squared:  0.725 
F-statistic: 24.53 on 13 and 103 DF,  p-value: < 2.2e-16

> AIC<-extractAIC(forwardAIC)
> BIC<-extractAIC(forwardAIC,k=log(117))
> print(AIC)
[1]  14.0000 328.7819
> print(BIC)
[1]  14.0000 367.4523
```

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

```
Residuals:
    Min      1Q  Median      3Q     Max 
-8.2727 -2.1880 -0.4853  2.7989  9.8282 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 17.45337    4.22302   4.133 7.31e-05 ***
x4Factor1   -7.41246    1.02656  -7.221 9.17e-11 ***
x4Factor2   -9.66303    1.51015  -6.399 4.70e-09 ***
x4Factor3   -9.36506    2.50060  -3.745 0.000297 ***
x3Factor1    0.94828    1.00427   0.944 0.347253    
x3Factor2    1.13789    1.62975   0.698 0.486625    
x3Factor3   -6.40106    1.82278  -3.512 0.000662 ***
x5           0.30487    0.09548   3.193 0.001868 ** 
x9Factor1    3.09165    1.59863   1.934 0.055865 .  
x9Factor2   -0.22157    1.66989  -0.133 0.894703    
x9Factor3    0.62773    1.97225   0.318 0.750916    
x9Factor4    0.18761    1.26622   0.148 0.882503    
x9Factor5   -3.38556    1.68430  -2.010 0.047038 *  
x2          -0.10158    0.04882  -2.081 0.039930 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.854 on 103 degrees of freedom
Multiple R-squared:  0.7558,	Adjusted R-squared:  0.725 
F-statistic: 24.53 on 13 and 103 DF,  p-value: < 2.2e-16

> AIC<-extractAIC(twoSidedAIC)
> BIC<-extractAIC(twoSidedAIC,k=log(117))
> print(AIC)
[1]  14.0000 328.7819
> print(BIC)
[1]  14.0000 367.4523
>
```

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

```
Residuals:
   Min     1Q Median     3Q    Max 
-9.766 -2.638 -0.137  2.700 10.149 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  18.39074    3.34294   5.501 2.56e-07 ***
x4Factor1    -7.20026    1.04205  -6.910 3.52e-10 ***
x4Factor2   -10.37992    1.54955  -6.699 9.83e-10 ***
x4Factor3   -12.35910    2.35416  -5.250 7.68e-07 ***
x5            0.29580    0.08293   3.567 0.000540 ***
x2           -0.11223    0.04531  -2.477 0.014806 *  
x3Factor1    -0.03494    0.96076  -0.036 0.971056    
x3Factor2     0.53168    1.54413   0.344 0.731273    
x3Factor3    -6.35388    1.68958  -3.761 0.000276 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.012 on 108 degrees of freedom
Multiple R-squared:  0.7225,	Adjusted R-squared:  0.702 
F-statistic: 35.15 on 8 and 108 DF,  p-value: < 2.2e-16

> AIC<-extractAIC(forwardBIC)
> BIC<-extractAIC(forwardBIC,k=log(117))
> print(AIC)
[1]   9.0000 333.7366
> print(BIC)
[1]   9.0000 358.5962
```

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

```
Call:
lm(formula = dataset$Cancer.mortality.rate.... ~ x4Factor + x5 +
    x2 + x3Factor, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-9.766 -2.638 -0.137  2.700 10.149

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.39074    3.34294   5.501 2.56e-07 ***
x4Factor1    -7.20026    1.04205  -6.910 3.52e-10 ***
x4Factor2   -10.37992    1.54955  -6.699 9.83e-10 ***
x4Factor3   -12.35910    2.35416  -5.250 7.68e-07 ***
x5            0.29580    0.08293   3.567 0.000540 ***
x2           -0.11223    0.04531  -2.477 0.014806 *
x3Factor1    -0.03494    0.96076  -0.036 0.971056
x3Factor2     0.53168    1.54413   0.344 0.731273
x3Factor3    -6.35388    1.68958  -3.761 0.000276 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.012 on 108 degrees of freedom
Multiple R-squared:  0.7225,    Adjusted R-squared:  0.702
F-statistic: 35.15 on 8 and 108 DF,  p-value: < 2.2e-16

> AIC<-extractAIC(forwardBIC)
> BIC<-extractAIC(forwardBIC,k=log(117))
> print(AIC)
[1]   9.0000 333.7366
> print(BIC)
[1]   9.0000 358.5962
```

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

```
Residuals:
   Min     1Q Median     3Q    Max 
-9.766 -2.638 -0.137  2.700 10.149 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  18.39074    3.34294   5.501 2.56e-07 ***
x4Factor1    -7.20026    1.04205  -6.910 3.52e-10 ***
x4Factor2   -10.37992    1.54955  -6.699 9.83e-10 ***
x4Factor3   -12.35910    2.35416  -5.250 7.68e-07 ***
x5            0.29580    0.08293   3.567 0.000540 ***
x2           -0.11223    0.04531  -2.477 0.014806 *  
x3Factor1    -0.03494    0.96076  -0.036 0.971056    
x3Factor2     0.53168    1.54413   0.344 0.731273    
x3Factor3    -6.35388    1.68958  -3.761 0.000276 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.012 on 108 degrees of freedom
Multiple R-squared:  0.7225,	Adjusted R-squared:  0.702 
F-statistic: 35.15 on 8 and 108 DF,  p-value: < 2.2e-16

> AIC<-extractAIC(twoSidedBIC)
> BIC<-extractAIC(twoSidedBIC,k=log(117))
> print(AIC)
[1]   9.0000 333.7366
> print(BIC)
[1]   9.0000 358.5962
```

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

```
Call:
lm(formula = y ~ x2 + x5 + x3Factor + x4Factor)

Residuals:
   Min     1Q Median     3Q    Max 
-9.766 -2.638 -0.137  2.700 10.149 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  18.39074    3.34294   5.501 2.56e-07 ***
x2           -0.11223    0.04531  -2.477 0.014806 *  
x5            0.29580    0.08293   3.567 0.000540 ***
x3Factor1    -0.03494    0.96076  -0.036 0.971056    
x3Factor2     0.53168    1.54413   0.344 0.731273    
x3Factor3    -6.35388    1.68958  -3.761 0.000276 ***
x4Factor1    -7.20026    1.04205  -6.910 3.52e-10 ***
x4Factor2   -10.37992    1.54955  -6.699 9.83e-10 ***
x4Factor3   -12.35910    2.35416  -5.250 7.68e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.012 on 108 degrees of freedom
Multiple R-squared:  0.7225,	Adjusted R-squared:  0.702 
F-statistic: 35.15 on 8 and 108 DF,  p-value: < 2.2e-16

> AIC<-extractAIC(backwardBIC)
> BIC<-extractAIC(backwardBIC,k=log(117))
> print(AIC)
[1]   9.0000 333.7366
> print(BIC)
[1]   9.0000 358.5962
```