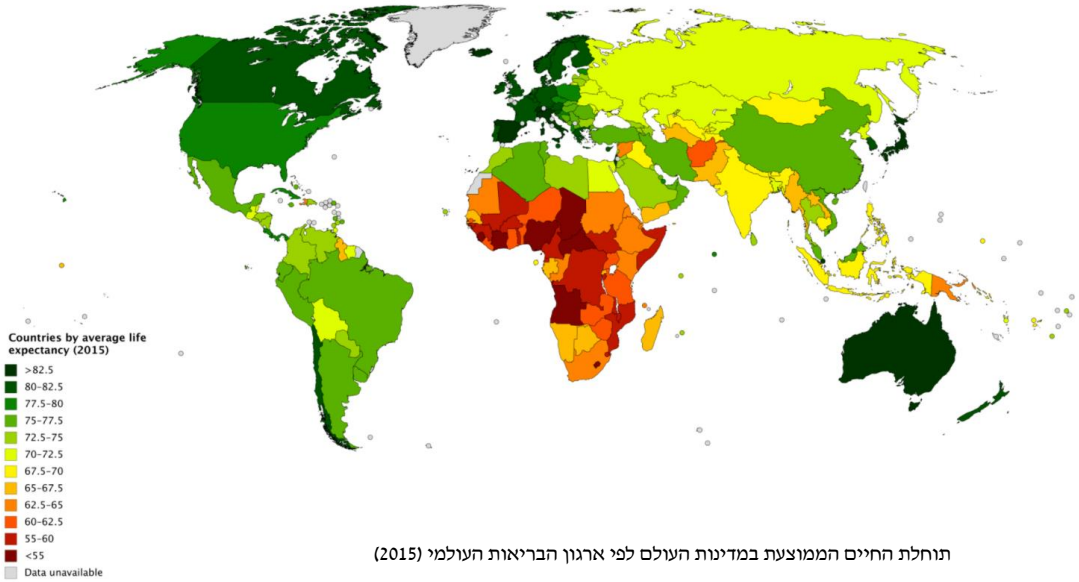
פרויקט רגרסיה ליניארית - חלק ב'



תוחלת החיים הממוצעת במדינות העולם לפי ארגון הבריאות העולמי (2015)

קורס: מודלים של רגרסיה לינארית

תאריך: 11/01/2022

מרצה: הלל בר גרא

מגישים :

קבוצה 14

315695643

316161694

תוכן עניינים

1. תקציר מנהלים

מטרתנו היא לבנות את המודל שמסביר בצורה הטובה ביותר את תוחלת החיים במדינות. כלומר, נרצה לדעת מיהם המשתנים המסבירים התורמים לחיזוי המשתנה המוסבר.

בתחילת התהליך בחנו את המשתנים אשר נרצה להכניס למודל הראשוני. על מנת לעשות זאת בחרנו אילו משתנים ניתן להסיר מהמודל עקב חוסר מתאם בינם לבין המשתנה המוסבר, איחדנו קטגוריות בעלות מאפיינים דומים. כמו כן, בחנו את הצורך במשתני דמה ומשתני אינטראקציה אותם נכניס למודל המלא. עבור כל משתנה קטגוריאלי יצרנו מספר משתני דמה ולאחר מכן יצרנו משתנה אינטראקציה בין משתנה הדמה למשתנה רציף כאשר ישנה השפעה שונה עבור כל אחת מהקבוצות של המשתנה הקטגוריאלי על המשתנה המוסבר.

בהמשך התהליך בכדי לבחור את המודל הסופי ביצענו מספר אלגוריתמים שמטרתם לאתר את קבוצת המשתנים המסבירים אשר שילובם יחד יוביל למודל הטוב ביותר. השוואנו את המודלים שקיבלנו עפ״י מדדי טיב התאמה ובחרנו את המודל הבא:

$$\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1}\mathrm{X_{i2}} + \widehat{\beta_2}\mathrm{X_{i3}} + \widehat{\beta_3}\mathrm{X_{i4}} + \widehat{\beta_4}\mathrm{I_{i1}} + \widehat{\beta_5}\mathrm{I_{i2}} + \widehat{\beta_6}\mathrm{I_{i3}} + \widehat{\beta_7}X_{i2}I_{i1} + \widehat{\beta_8}X_{i2}I_{i2}$$

$$+\widehat{\beta_9}X_{i2}I_{i3} + \widehat{\beta_{10}}X_{i3}I_{i1} + \widehat{\beta_{11}}X_{i3}I_{i2} + \widehat{\beta_{12}}X_{i3}I_{i3}$$

לאחר מכן, בחנו האם מתקיימות הנחות מודל רגרסיה לינארית: הנחת הליניאריות, שיוויון שונויות של השגיאות, והנחת נורמליות השגיאות. לאחר יצירת תרשימים המתארים את הנתונים ומבחנים סטטיסטים עבור כל הנחה מצאנו כי המודל מקיים את שלושת ההנחות.

כמו כן, בשלב האחרון בחרנו לנסות ולשפר את המודל באמצעות טרנספורמציה על המשתנה המוסבר. לאחר ביצוע הטרנספורמציה, וביצוע השוואה בין המודל המקורי למודל לאחר הטרנספורמציה קיבלנו R^2_{adj} גדול יותר המעיד על מודל בו הקשר בין המשתנים המסבירים למשתנה המוסבר טוב יותר. לכן, המודל החדש והסופי הינו:

$$\widehat{Y_i}^3 = \widehat{\beta_0} + \widehat{\beta_1}\mathrm{X_{i2}} + \widehat{\beta_2}\mathrm{X_{i3}} + \widehat{\beta_3}\mathrm{X_{i4}} + \widehat{\beta_4}\mathrm{I_{i1}} + \widehat{\beta_5}\mathrm{I_{i2}} + \widehat{\beta_6}\mathrm{I_{i3}} + \widehat{\beta_7}X_{i2}I_{i1} + \widehat{\beta_8}X_{i2}I_{i2}$$
$$+\widehat{\beta_9}X_{i2}I_{i3} + \widehat{\beta_{10}}X_{i3}I_{i1} + \widehat{\beta_{11}}X_{i3}I_{i2} + \widehat{\beta_{12}}X_{i3}I_{i3}$$

משתנה	סוג המשתנה - מוסבר/מסביר	סימון	יחידת מידה	סוג המשתנה – רציף / קטגוריאלי	הסבר קצר על המשתנה
Outdoor air pollution (%)	מסביר	X1	אחוז (%)	רציף	אחוז זיהום האויר במדינה הנבדקת המייצג את אחוז הגזים המסוכנים באוויר.
HIV - Estimated number of people that have been infected	מסביר	X2	איש	בדיד	מספר האנשים במדינה הנבדקת שנדבקו בנגיף ה-HIV מתוך כלל האוכלוסייה.
malaria - Estimated number of people that have been infected	מסביר	X3	איש	בדיד	מספר האנשים במדינה הנבדקת החולים במלריה מתוך כלל האוכלוסייה.
Average income per person ($)	מסביר	X4	דולר ($)	רציף	הכנסה שנתית ממוצעת לאדם בדולרים - ההכנסה הממוצעת המחושבת ע"י סך הכנסות של כל אזרחי המדינה לחלק בכמות האנשים בה.
Alcohol consumption per person (liters, year)	מסביר	X5	ליטר (L)	רציף	ממוצע צריכת האלכוהול השנתית לאדם בליטר - מחושב ע"י כמות האלכוהול הנצרך בשנה במדינה הנבדקת לחלק כמות האנשים בה.
density per square (km)	מסביר	X6	איש	רציף	צפיפות אוכלוסין לקילומטר מרובע - מייצג את כמות האנשים המתגוררים בקילומטר מרובע, מחושב ע"י סך האנשים במדינה לחלק בשטח שלה.
Cigarette consumption (%)	מסביר	X7	אחוז (%)	רציף	אחוז צרכני הסיגריות - מספר האנשים המעשנים במדינה ביחס לכלל האוכלוסייה.
Continent	מסביר	X8	-	קטגוריאלי	היבשת בה ממוקמת המדינה - היבשות מיוצגות ע"י מספרים חד ערכיים, כך שכל יבשת מיוצגת ע"י מספר שונה. 1-אסיה, 2- אפריקה והמפרץ הפרסי, 3- דרום אמריקה, 4- אירופה, 5- מרכז אמריקה.
Member of OECD	מסביר	X9	-	קטגוריאלי	שייכות לארגון ה-OECD - האם המדינה חברה בארגון,כך שמיוצג ע"י משתנה בינארי (1-כן, 0-לא).
Life expectancy (year)	מוסבר	y	שנה	רציף	תוחלת החיים במדינה - אומדן למספר השנים הממוצע שבני אדם חיים במדינה הנבדקת.

2. עיבוד מקדים:

2.1 הסרה של משתנים מסבירים:

	Y
Y	1.00000000
X1	-0.20166610
X2	-0.21760221
X3	-0.34217309
X4	0.57810201
X5	0.01830823
X6	0.08092025
X7	0.15723031
X8	0.21507312
X9	0.19013586

על מנת לבחון אילו משתנים נרצה להסיר מהמודל, נסתכל על שלושת המשתנים בעלי ערך הקורלציה הנמוך ביותר (בערך מוחלט). עבור משתנים אלו, נבחן במבחן סטטיסטי האם יש קשר ביניהם לבין המשתנה המוסבר. בהתאם לתוצאת P-value נחליט האם נקבל או נדחה את ההשערה, ובהתאמה האם להשאיר או להסיר את המשתנה הנבדק מהמודל.

כפי שניתן לראות, שלושת המשתנים עם הקורלציה הנמוכה ביותר הם המשתנים המייצגים את צריכת האלכוהול (X5), צפיפות האוכלוסין (X6), ואחוז צריכת הסיגריות (X7). ייתכן, ומשתנים אלו אינם מסבירים באופן מובהק את המשתנה המוסבר, ולכן ונבחן האם להסיר אותם מהמודל.

עבור כל אחד משלושת המשתנים נבחן את ההשערה הבאה:

$$\begin{cases} H_0: \beta_i = 0 \\ H_1: \beta_i \neq 0 \end{cases}$$

בשלושת המבחנים נמצא כי P-value גדול מ-5%, ובהתאמה רווחה הסמך מכיל את הערך 0. לכן, נקבל את השערת האפס ונסיק שאין קשר בין משתנים אלו למשתנה המוסבר בר״מ של 5%. על כן, נחליט להסיר את משתנים אלו מהמודל.

```
>   cor.test(Dataset$X7, Dataset$Y,alternative = "two.sided",method = "pearson")

        Pearson's product-moment correlation

data:  Dataset$X7 and Dataset$Y
t = 1.5761, df = 98, p-value = 0.1182
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.0404366  0.3430539
sample estimates:
      cor
0.1572303
```

```
>   cor.test(Dataset$X5, Dataset$Y,alternative = "two.sided",method = "pearson")

        Pearson's product-moment correlation

data:  Dataset$X5 and Dataset$Y
t = 0.18127, df = 98, p-value = 0.8565
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1787527  0.2139569
sample estimates:
       cor
0.01830823
```

```
>   cor.test(Dataset$X6, Dataset$Y,alternative = "two.sided",method = "pearson")

        Pearson's product-moment correlation

data:  Dataset$X6 and Dataset$Y
t = 0.80371, df = 98, p-value = 0.4235
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1173633  0.2729993
sample estimates:
       cor
0.08092025
```

2.2 התאמת משתנים

איחוד קטגוריות X8:

במדגם קיימות שתי רשומות בלבד מתוך מאה מדינות שנבדקו השייכות לקבוצה 4 במשתנה המייצג את היבשת. לכן, נרצה לאחד קטגוריה זו עם קטגוריה אחרת הדומה לה במאפייניה. עפ"י התרשים ניתן לראות שהחציון והטווח הבין רבעוני של קבוצה 3 דומים. בהתאם לכך, נאחד את שתי קבוצות אלו. כלומר, מעתה נתייחס ליבשות אירופה ודרום אמריקה כקטגוריה אחת המיוצגת ע"י הערך 3.

לאחר שיצרנו תרשים פיזור עבור כלל המשתנים הרציפים והבדידים (מצורף בנספחים) לא נרצה לבצע איחוד לקטגוריות. עפ"י הפיזור ניתן לראות שטווח ערכים של המשתנים הינו גדול, ולכן נרצה לנתח את כלל הממצאים כפי שהם, ולא להכניסם לקטגוריות בהם לא ניתן לראות את הערכים המדויקים.

2.3 הגדרת משתנה דמה

- **משתנה קטגוריאלי יבשת** (X8):

לאחר שביצענו איחוד בין הקטגוריות 3 ו-4, כעת ישנן ארבע קטגוריות: אסיה (1), אפריקה והמפרץ הפרסי (2), אירופה ודרום אמריקה (3), מרכז אמריקה (5). קבוצת הבסיס תהיה $X_8 = 5$, ולכן נגדיר:

$$I_{i1} = \begin{cases} 1 & X_8 = 1 \\ 0 & \text{else} \end{cases}, \qquad I_{i2} = \begin{cases} 1 & X_8 = 2 \\ 0 & \text{else} \end{cases}, \qquad I_{i3} = \begin{cases} 1 & X_8 = 3 \\ 0 & \text{else} \end{cases}$$

- **משתנה קטגוריאלי חברות בארגון ה-OECD** (X9), ולכן נגדיר:

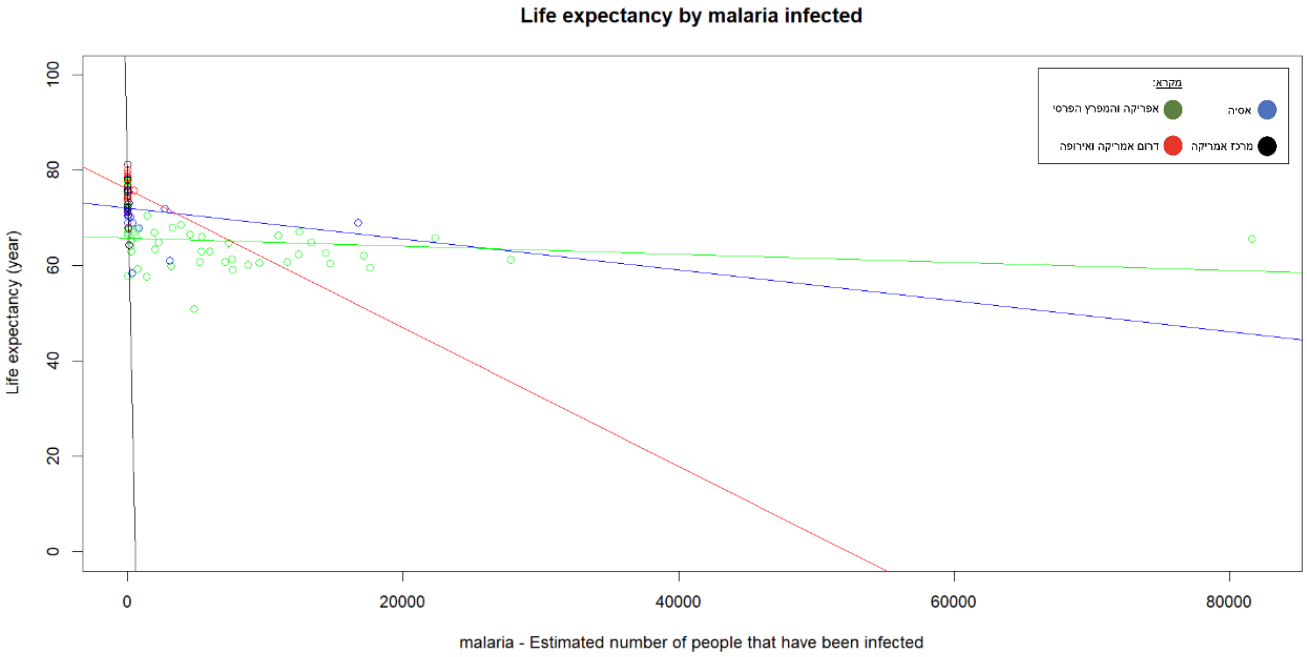
$$L_i = \begin{cases} 1 & X_9 = 1 \\ 0 & X_9 = 0 \end{cases}$$

2.4 משתני אינטראקציה

על מנת להחליט אילו משתני אינטראקציה נוסיף למודל, נייצר תרשימי פיזור עבור כל אחת מהמשתנים הרציפים בהשפעתם על המשתנה המוסבר, עפ״י חלוקה לקטגוריות של המשתנה הקטגוריאלי. עבור כל קטגוריה נבנה קו רגרסיה פרטני, במידה ונראה הבדלים משמעותיים בין קווי הרגרסיה (בחותך או בשיפוע) נרצה להכניס משתנה אינטראקציה למודל.

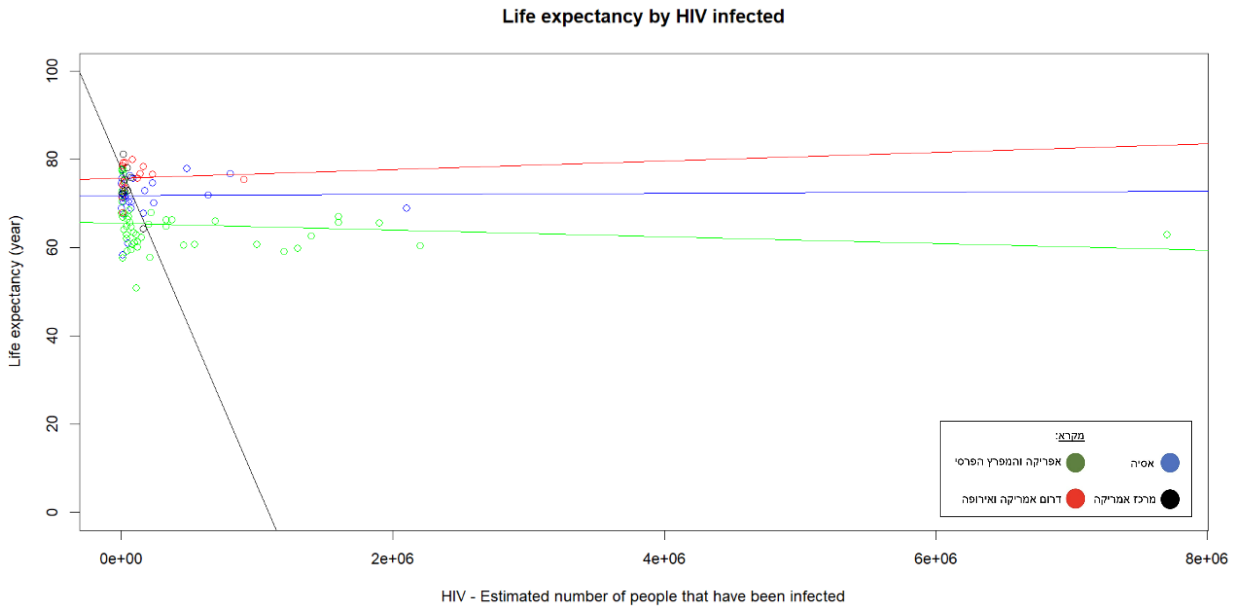
לאחר בחינת כל תרשימי הפיזור ובדיקת מדדים אלו בחרנו להכניס את שלושת משתני האינטראקציה הבאים:

- להלן המשתנים עבורם נייצר משתני אינטראקציה:

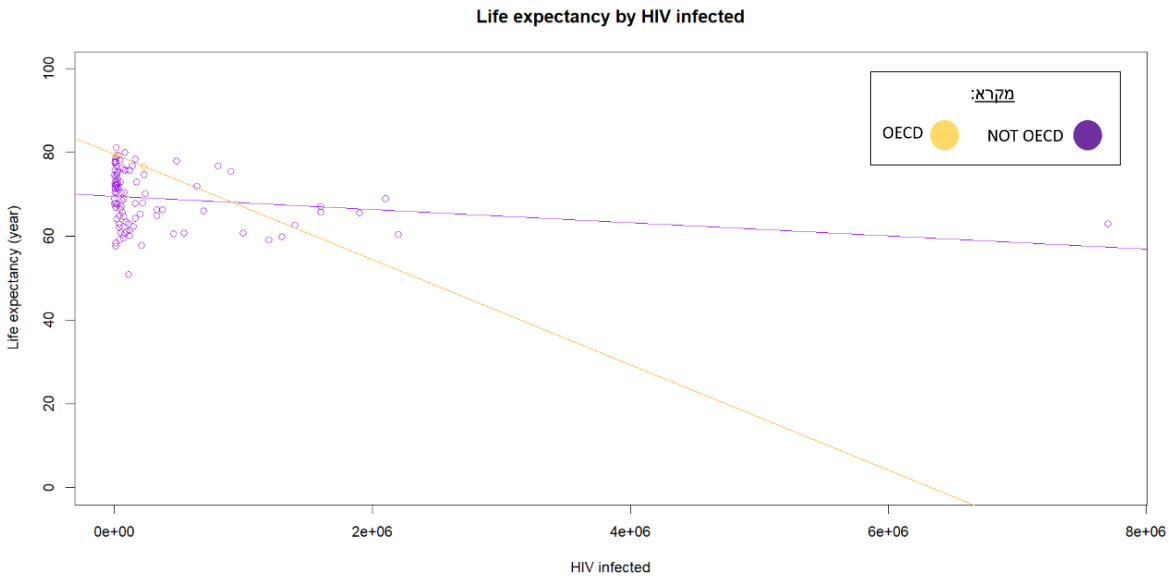


כפי שניתן לראות, קיימים הבדלים משמעותיים בקווי הרגרסיה בין יבשות שונות. כלומר, ביבשות שונות יש השפעה שונה של כמות החולים במלריה על תוחלת החיים. לדוגמה, ביבשת אפריקה והמפרץ הפרסי תוחלת החיים אינה משתנה באופן משמעותי כאשר מספר החולים במדינה גדל, זאת בניגוד ליבשות דרום אמריקה ואירופה בהן כמות חולים רבה במדינה גורמת לקטנת תוחלת החיים בה. כמו כן, ביבשת מרכז אמריקה תוחלת החיים משתנה עבור מדינות עם מספר חולים זהה.

ייתכן ועקב מערכת בריאות חלשה במדינות יבשת דרום אמריקה החולים במדינות אינם מקבלים טיפול רפואי כנדרש וכתוצאה מכך תוחלת החיים קטנה יותר ממדינות הממוקמות ביבשות אחרות. בנוסף, ייתכן ובמדינות מרכז אמריקה איכות מערכת הבריאות היא שונה עבור כל אחת מהמדינות הממוקמות בה, ולכן תוחלת החיים היא שונה.



כפי שניתן לראות, קיימים הבדלים משמעותיים בקווי הרגרסיה בין יבשות שונות. כלומר, ביבשות שונות יש השפעה שונה לכמות הנדבקים ב-HIV על תוחלת החיים. לדוגמה, במרכז אמריקה תוחלת החיים קטנה באופן משמעותי ככל שמספר הנדבקים ב-HIV עולה לעומת שאר היבשות. בנוסף, בתרשים הפיזור התצפיות המייצגות את מדינות דרום אמריקה ובאירופה גבוהה יותר מיתר התצפיות, ולכן ניתן להסיק שתוחלת החיים של מדינות אלו גבוהה משל מדינות הנמצאות ביבשות אפריקה והמפרץ הפרסי ואסיה ללא תלות במספר החולים במדינות. ייתכן במדינות דרום אמריקה ואירופה מערכות הבריאות מדינה מעניקות טיפול טוב יותר לנבדקים ב-HIV ולכן תוחלת החיים בהם גבוהה יותר.



כפי שניתן לראות, קיימים הבדלים משמעותיים בקווי הרגרסיה אשר מתארים את ההשפעה של כמות החולים ב-HIV על תוחלת החיים. ניתן לראות שתוחלת החיים של מדינות החברות בארגון היא גבוהה ביחס לשאר המדינות שאינן חברות בארגון. כלומר, תוחלת החיים במדינות היא שונה

כאשר מחלקים אותם עפ"י שייכותם לארגון ה-OECD. ייתכן ומשום שבארגון חברות מדינות עמידות עם מערכות בריאות מתקדמות הן נותנות טיפול טוב יותר לנבדקים ב-HIV, ולכן תוחלת החיים בהם גבוהה ביחס לשאר המדינות.

3. התאמת המודל ובדיקת הנחות המודל:

3.1 בחירת משתני המודל

המודל המלא הינו:

$$\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1}X_{i1} + \widehat{\beta_2}X_{i2} + \widehat{\beta_3}X_{i3} + \widehat{\beta_4}X_{i4} + \widehat{\beta_5}I_{i1} + \widehat{\beta_6}I_{i2} + \widehat{\beta_7}I_{i3} + \widehat{\beta_8}L_i + \widehat{\beta_9}X_{i2}I_{i1} + \widehat{\beta_{10}}X_{i2}I_{i2} + \widehat{\beta_{11}}X_{i2}I_{i3} + \widehat{\beta_{12}}X_{i2}L_{i1} + \widehat{\beta_{13}}X_{i3}I_{i1} + \widehat{\beta_{14}}X_{i3}I_{i2} + \widehat{\beta_{15}}X_{i3}I_{i3}$$

כאשר נרצה לבצע השוואה בין מודלים שונים, נרצה להשוות בין ערכי המדדים הבאים:

1. מדד AIC – מדד המאפיין את טיב ההתאמה של המודל בהתבסס על פונקציית הנראות – ככל שערך פונקציית הנראות גדול יותר כך המדד יקטן. בעזרת מודל זה נוכל להשוות בין מודלים עם כמות שונה של פרמטרים. בעוד שלערך המדד אין משמעות בפני עצמו בעת ביצוע השוואה בין שני מודלים, נעדיף את המודל בעל מדד קטן יותר.
2. מדד R^2_{adj} - מדד המייצג את אחוז השונות המוסברת במודל. בעזרת מדד זה נוכל להשוות בין מודלים עם כמות פרמטרים שונה, כך שככל שערך המדד גבוה יותר כך המשתנים המסבירים מסבירים טוב יותר את המשתנה המוסבר ולכן נעדיף מודל עם R^2_{adj} גדול יותר.

נרצה לבדוק אילו משתנים מהמודל המלא שיצרנו יכנסו אל המודל הסופי. נעשה זאת בעזרת שלושה אלגוריתמים שונים לבניית מודל רגרסיה:

1. Forward Selection– נתחיל ממודל ריק ללא משתנים ובכל איטרציה נבחן את השפעת ההכנסה של כל אחד משאר המשתנים אל המודל ונבחר להכניס את המשתנה אשר ישפר את המדד הנבדק. בעת ביצוע האלגוריתם בתוכנת R המדד הנבדק הינו מדד ה-AIC ולכן נרצה להכניס משתנה אשר ימזער את המדד של המודל החלקי. האלגוריתם יסתיים כאשר הוספת כל אחד מהמשתנים הנותרים יגדיל או לא ישפיע על ערך המדד.

   המודל שהתקבל הוא: $\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1}X_{i4} + \widehat{\beta_2}I_{i1} + \widehat{\beta_3}I_{i2} + \widehat{\beta_4}I_{i3}$
2. Backward Elimination- נתחיל מהמודל המלא המכיל את כל המשתנים ובכל איטרציה נבחן את השפעת הוצאת אחד המשתנים. נבחר להוציא משתנה אשר ישפר את המדד הנבדק ולכן בעת ביצוע האלגוריתם בתוכנת ה-R נרצה להכניס את המשתנה שימזער את מדד ה-AIC . האלגוריתם יסתיים כאשר כל אחד מהמשתנים הנותרים יגדיל או לא ישפיע על ערך המדד.

   המודל שהתקבל הוא: $\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1}X_{i2} + \widehat{\beta_2}X_{i3} + \widehat{\beta_3}X_{i4} + \widehat{\beta_4}I_{i1} + \widehat{\beta_5}I_{i2} + \widehat{\beta_6}I_{i3} + \widehat{\beta_7}X_{i2}I_{i1} + \widehat{\beta_8}X_{i2}I_{i2} + \widehat{\beta_9}X_{i2}I_{i3} + \widehat{\beta_{10}}X_{i3}I_{i1} + \widehat{\beta_{11}}X_{i3}I_{i2} + \widehat{\beta_{12}}X_{i3}I_{i3}$

3. Stepwise Regression- אלגוריתם המשלב את שני האלגוריתמים שהצגנו קודם. נתחיל ממודל ריק כאשר בכל איטרציה נרצה לבדוק הן הכנסת משתנה שאינו נמצא כעת במודל והן הוצאת משתנה אשר נמצא ממודל. כמו כן, משתנה מסביר שהוספנו באיטרציה קודמת יכול להפוך למיותר באיטרציה הנוכחית בעקבות קשרים עם משתנים מסבירים אחרים. גם באלגוריתם זה נרצה למזער את מדד ה-AIC והוא יסתיים כאשר כל אחת מהפעולות האפשריות תגדיל או לא תשפיע על ערך המדד.

המודל שהתקבל הוא: $\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1}\mathrm{X_{i4}} + \widehat{\beta_2}\mathrm{I_{i1}} + \widehat{\beta_3}\mathrm{I_{i2}} + \widehat{\beta_4}\mathrm{I_{i3}}$

השוואה בין המודלים שהתקבלו:

	מודל מלא	רגרסיה לפנים	רגרסיה לאחור	רגרסיה בצעדים
AIC	332.54	285.29	288.16	285.29
R^2_{adj}	0.6384	0.6343	0.6497	0.6343

כפי שניתן לראות מדד ה-AIC המינימלי מתקבל עבור המודל שהתקבל מהרגרסיה לפנים, בעוד שמדד ה-R^2_{adj} המקסימלי מתקבל עבור המודל שהתקבל מהרגרסיה לאחור. נבחר במודל שהתקבל מהרגרסיה לאחור כמודל הסופי מכיוון ו- R^2_{adj} מייצג בצורה טובה יותר את רמת ההתאמה של המשתנים המסבירים למשתנה המוסבר. כמו כן ניתן לראות שביחס למודל המלא, המודל החלקי הנ"ל מקטין את ערך ה-AIC ומגדיל את ערך ה-R^2_{adj}.

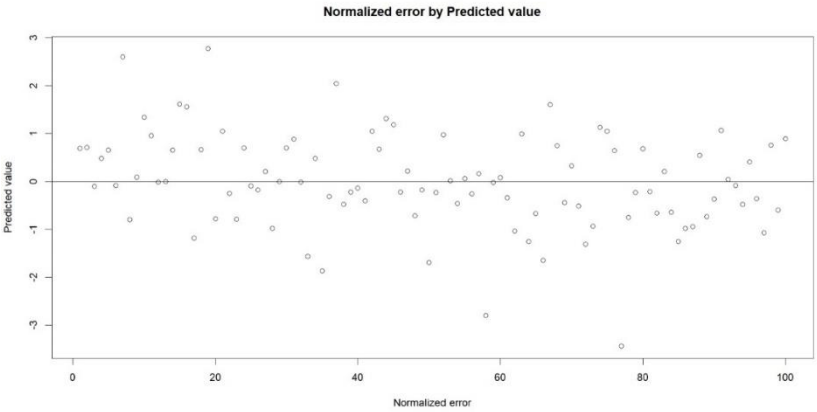
על פי מודל הרגרסיה שחושב קודם לכן, המודל הסופי הינו:

$$\widehat{Y_i} = 68.47 + 8.622 * 10^{-6}\mathrm{X_{i2}} - 1.255 * 10^{-3}\mathrm{X_{i3}} + 2.823 * 10^{-4}\mathrm{X_{i4}} - 5.04\mathrm{I_{i1}} + 3.643\mathrm{I_{i2}} + 4.499\mathrm{I_{i3}} - 9.198 * 10^{-6}X_{i2}I_{i1} - 8.511 * 10^{-6}X_{i2}I_{i2} - 1.296 * 10^{-5}X_{i2}I_{i3} + 1.236 * 10^{-3}X_{i3}I_{i1} - 5.104 * 10^{-4}X_{i3}I_{i2} - 0.1036X_{i3}I_{i3}$$

3.2 בדיקת הנחות המודל:

בכדי לקבל אינדיקציה על קיומן של הנחות מודל הרגרסיה נשתמש בסטטיסטיקה תיאורית.

עבור הנחת הליניאריות והנחת שוויון השונויות במודל, נסתכל על תרשים השאריות:



כפי שניתן לראות בתרשים, ישנן תצפיות מעל ומתחת לקו האפס לכל אורך התחום של צריך ה-X, והפיזור של התצפיות מעל ומתחת לקו האפס ברובו אחיד. לכן, ניתן לשער שהנחת הלינאריות והנחת שוויון השונויות מתקיימות אך עלינו לבדוק זאת גם במבחן סטטיסטי.

מבחן לינאריות - chow

$$\begin{cases} H_0: & \text{מודל לינארי} \\ H_1: & \text{אחרת} \end{cases}$$

```
>   sctest(FinalModel,type="Chow")

        M-fluctuation test

data:  FinalModel
f(efp) = 1.7171, p-value = 0.06914
```

במבחן chow קיבלנו P-value > 0.05 , ולכן ברמת מובהקת של 5% לא נדחה את השערת האפס.
כלומר, הנחת הליניאריות של המודל מתקיימת.

מבחן שוויון שונויות - Goldfeld-Quandt

$$\begin{cases} H_0: & \text{שוויון שונויות} \\ H_1: & \text{אחרת} \end{cases}$$

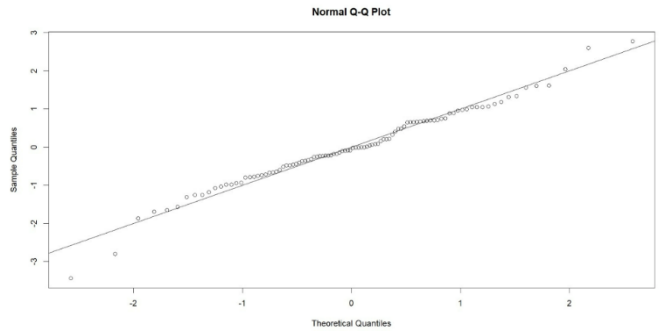
```
>   gqtest(FinalModel,alternative = "two.sided",fraction=30, data = DatasetNew)

        Goldfeld-Quandt test

data:  FinalModel
GQ = 1.6151, df1 = 22, df2 = 22, p-value = 0.2686
alternative hypothesis: variance changes from segment 1 to 2
```

במבחן Goldfels-Quandt קיבלנו P-value > 0.05, ולכן ברמת מובהקת של 5% לא נדחה את השערת האפס. כלומר, הנחת שיווין השונויות של שגיאות המודל מתקיימת.

עבור הנחת הנורמליות של השגיאות במודל נבחן את היסטוגרמת תדירות השגיאות ו- QQ PLOT:



כפי שניתן לראות בהיסטוגרמה פונקציית הצפיפות מזכירה את פונקציית הצפיפות של התפלגות נורמלית, ובתרשים ה-QQ PLOT התצפיות קרובות לקו המתואר. ייתכן והנתונים מגיעים מההתפלגות הנורמלית, אך לא ניתן לקבוע זאת באופן מוחלט ולכן נבצע שני מבחנים- מבחן קולמוגורב-סמירנוף (KS) ומבחן שפירו-ווילקס (SW):

$$\begin{cases} H_0: \text{התפלגות } \textit{נורמלית} \\ H_1: \text{אחרת} \end{cases}$$

```
>   ks.test(x= Dataset$stan_residuals, y="pnorm", alternative = "two.sided", exact = NULL)

        One-sample Kolmogorov-Smirnov test

data:  Dataset$stan_residuals
D = 0.062093, p-value = 0.8354
alternative hypothesis: two-sided

>   shapiro.test(Dataset$stan_residuals)

        Shapiro-Wilk normality test

data:  Dataset$stan_residuals
W = 0.97991, p-value = 0.1307
```

בשני המבחנים קיבלנו ש P-value > 0.05 , ולכן ניתן לומר בהתבסס על שני המבחנים הסטטיסטים שלא נדחה את השערת האפס ברמת מובהקות של 5%, כלומר שגיאות המודל מתאימות להתפלגות נורמלית.

3.3 דוגמה לשימוש במודל חיזוי

המודל שבנינו משמש לחיזוי תוחלת החיים במדינה, לכן בכדי לבחון את המודל נבחן מדינה שאינה קיימת במדגם ונאסוף עבורה את כלל המשתנים המסבירים (X_i) הקיימים במודל כפי שנאספו עבור שאר המדינות שבמדגם.

את הנתונים אספנו עבור מדינת ישראל והינם:

מספר הנדבקים ב-HIV: $X_2 = 9294$

מספר החולים במלריה: $X_3 = 105$

ההכנסה השנתית הממוצעת לאדם בדולר: $X_4 = 45{,}291$.

יבשת: אסיה – $X_8 = 1$

תוחלת החיים: $Y = 82$

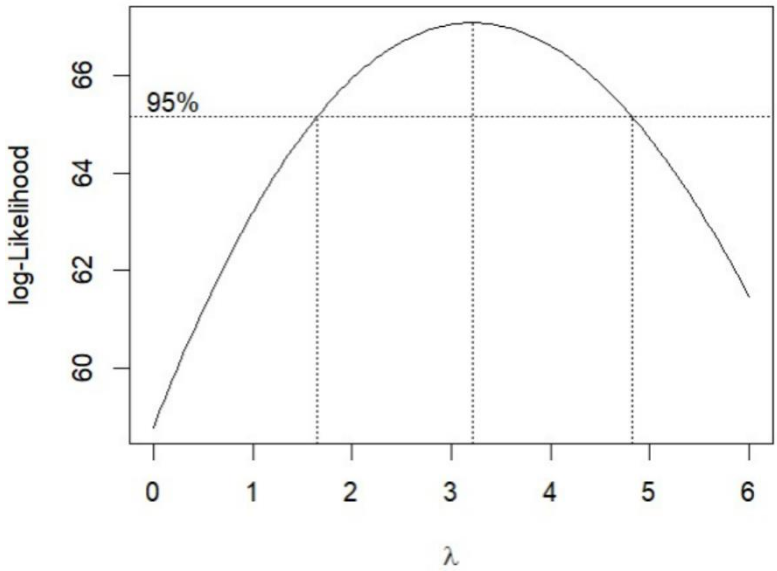
נבדוק מהו אומדן תוחלת החיים בהתאם למודל הרגרסיה שהצגנו ע"י הצבת הערכים במשוואת המודל:

$$\hat{Y}_{israel} = 68.47 + 8.622 * 10^{-6} * 9294 - 1.255 * 10^{-3} * 105 + 2.823 * 10^{-4} * 45291$$
$$-5.04 * 1 + 3.643 * 0 + 4.499 * 0 - 9.198 * 10^{-6} * 9294 * 1 - 8.511 * 10^{-6} * 9294 * 0$$
$$-1.296 * 10^{-5} * 9294 * 0 + 1.236 * 10^{-3} * 105 * 1 - 5.104 * 10^{-4} * 105 * 0$$
$$-0.1036 * 105 * 0 = 76.2$$

קיים פער של כ-7% בין תוחלת החיים בישראל לבין האומד על פי המודל. הסבר אפשרי הוא שמדינת ישראל אינה דומה במאפיינה התרבותיים והכלכליים לרוב מדינות אסיה שבמדגם. ייתכן ובמידה ונגדיל את המדגם ונאסוף נתונים עם מאפיינים זהים יהיה ניתן לשפר את המודל.

4. <u>שיפור המודל</u>

לאחר שבדקנו את כלל הנחות המודל קיבלנו עומד בשלושת הנחות המודל - הנחת הליניאריות, הנחת שוויון השונויות של השגיאות והנחת הנורמליות. נרצה לבדוק האם ניתן לבצע שיפור במודל על ידי ביצוע טרנספורמציות ובכך לנסות לשפר את מדד טיב ההתאמה R^2_{adj}. נבחר לבצע טרנספורמציה על המשתנה המוסבר (Y), בעזרת שימוש בBox-Cox.



כפי שניתן לראות בתרשים קיבלנו שערכה של λ נמצא ברמת ביטחון של 95% בטווח (1.7,4.8), ומקבלת ערך מקסימלי עבור $\lambda \cong 3.2$. בכדי לפשט את החישובים נבחר $\lambda = 3$ ונבצע את הטרנספורמציה על המשתנה המוסבר, Y^{λ}. המודל שקיבלנו לאחר הטרנספורמציה:

$$\widehat{Y_i}^3 = \widehat{\beta_0} + \widehat{\beta_1}\mathrm{X_{i2}} + \widehat{\beta_2}\mathrm{X_{i3}} + \widehat{\beta_3}\mathrm{X_{i4}} + \widehat{\beta_4}\mathrm{I_{i1}} + \widehat{\beta_5}\mathrm{I_{i2}} + \widehat{\beta_6}\mathrm{I_{i3}} + \widehat{\beta_7}X_{i2}I_{i1} + \widehat{\beta_8}X_{i2}I_{i2}$$
$$+\widehat{\beta_9}X_{i2}I_{i3} + \widehat{\beta_{10}}X_{i3}I_{i1} + \widehat{\beta_{11}}X_{i3}I_{i2} + \widehat{\beta_{12}}X_{i3}I_{i3}$$

המודל לאחר הטרנספורמציה מקיים את כל הנחות המודל הליניארי (מופיע בנספחים) ולכן ניתן להשוות בין המודלים בעזרת מדד R^2_{adj}. בעוד שבמודל לפני הטרנספורמציה קיבלנו $R^2_{adj} = 0.6497$ ולאחר הטרנספורמציה קיבלנו $R^2_{adj} = 0.6745$. כפי שניתן לראות לאחר הטרנספורמציה קיבלנו שטיב ההתאמה השתפר ולכן המודל הסופי שלנו הוא המודל שלאחר הטרנספורמציה.

$$\widehat{Y_i}^3 = 3.261 * 10^5 + 1.255 * 10^{-1} * \mathrm{X_{i2}} - 6.758 * 10^4 * \mathrm{X_{i3}} + 5.806 * 10^4 * \mathrm{X_{i4}}$$

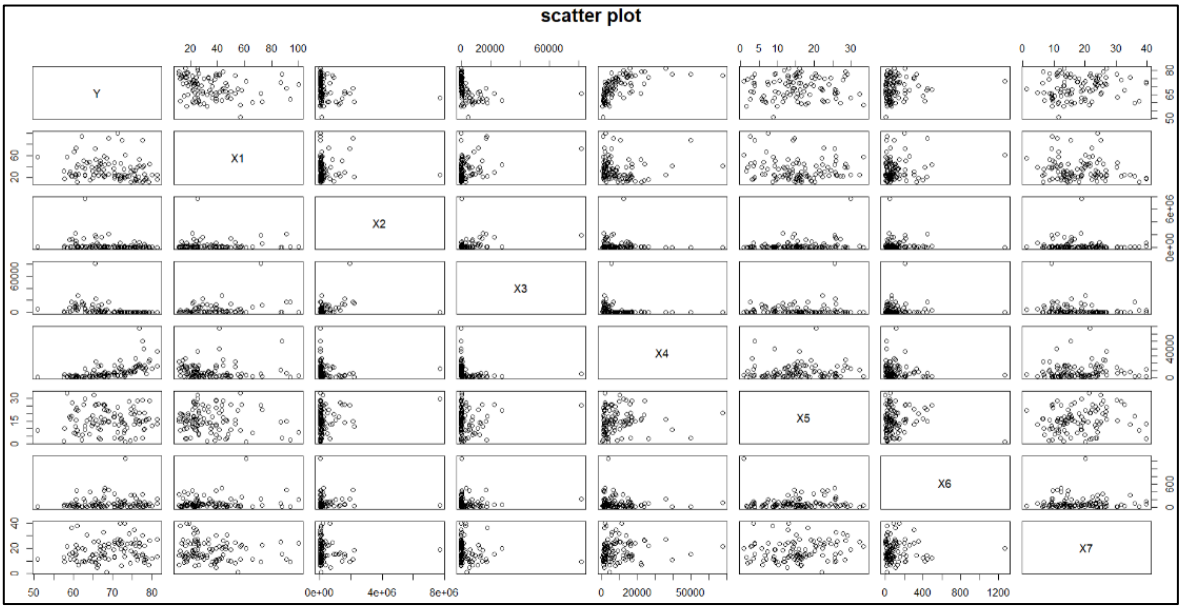
$$+7.025 * 10^4\mathrm{I_{i1}} - 1.853 * 10 * \mathrm{I_{i2}} + 4.159 * \mathrm{I_{i3}} - 1.341 * 10^{-1} * X_{i2}I_{i1}$$
$$-1.257 * X_{i2}I_{i2} - 2.227 * 10^{-1} * X_{i2}I_{i3} + 1.824 * 10 * X_{i3}I_{i1} - 1.499 * 10 * X_{i3}I_{i2}$$
$$- 1.497 * 10^3 X_{i3}I_{i3}$$

נספחים:

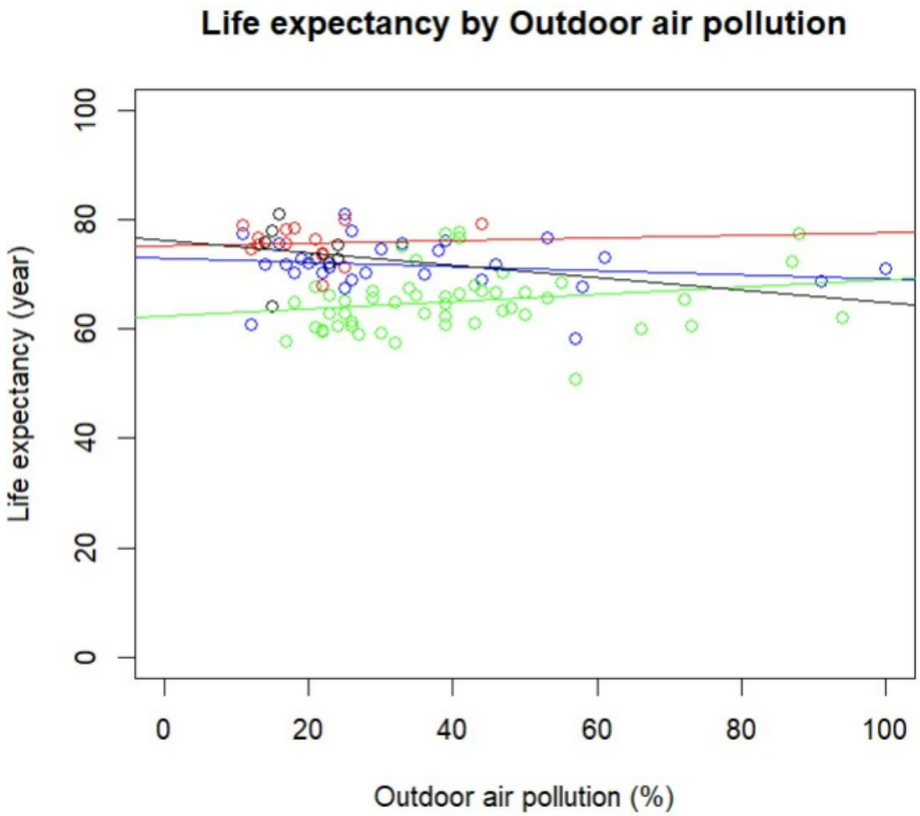
2.1 טבלת קורלציה בין כל המשתנים:

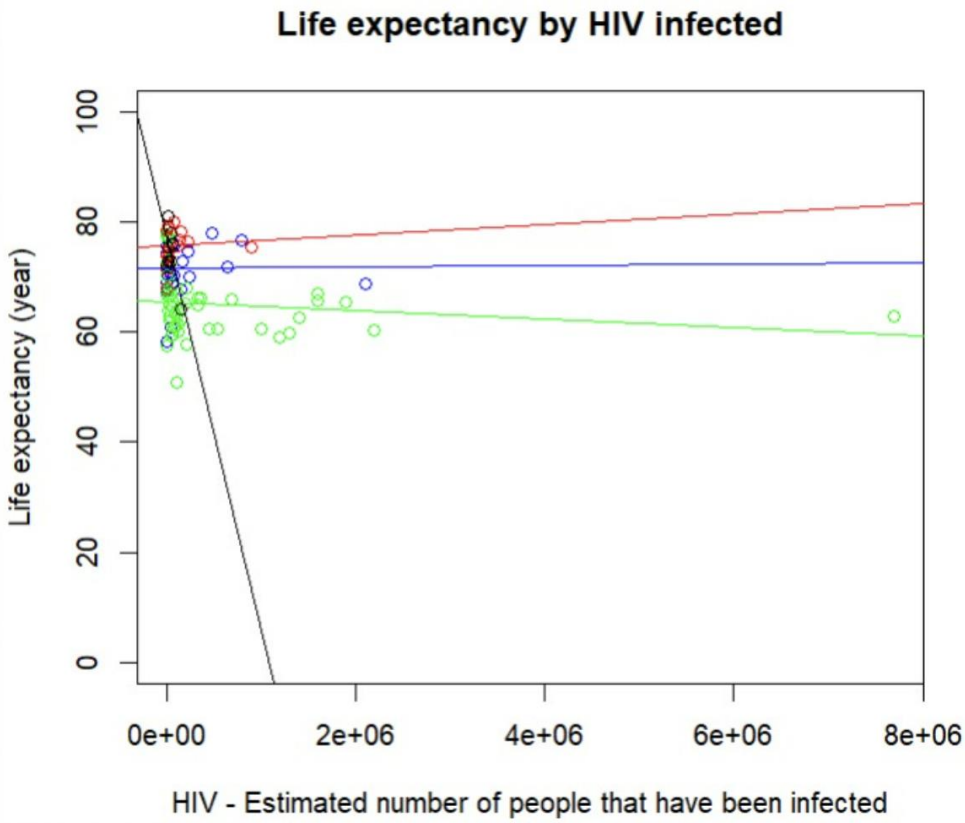
	Y	X1	X2	X3	X4	X5	X6	X7	X8	X9
Y	1.00000000	-0.20166610	-0.217602214	-0.34217309	0.57810201	0.01830823	0.080920248	0.157230312	0.21507312	0.19013586
X1	-0.20166610	1.00000000	0.022800681	0.29952317	-0.02940095	-0.11135116	0.151569035	-0.135506262	-0.28460705	-0.01466409
X2	-0.21760221	0.02280068	1.000000000	0.27890999	-0.06411276	0.24551050	0.006548176	-0.066677843	-0.06612567	-0.03270504
X3	-0.34217309	0.29952317	0.278909992	1.00000000	-0.21762686	0.09225196	0.037236252	-0.181003902	-0.06672715	-0.05721802
X4	0.57810201	-0.02940095	-0.064112757	-0.21762686	1.00000000	0.02973430	-0.118289354	0.079093841	0.04546723	0.15979990
X5	0.01830823	-0.11135116	0.245510495	0.09225196	0.02973430	1.00000000	-0.026133786	0.155007300	-0.01418909	0.12012489
X6	0.08092025	0.15156903	0.006548176	0.03723625	-0.11828935	-0.02613379	1.000000000	-0.000435957	-0.08467849	-0.02324887
X7	0.15723031	-0.13550626	-0.066677843	-0.18100390	0.07909384	0.15500730	-0.000435957	1.000000000	-0.21870411	0.03209051
X8	0.21507312	-0.28460705	-0.066125666	-0.06672715	0.04546723	-0.01418909	-0.084678487	-0.218704110	1.00000000	0.18105560
X9	0.19013586	-0.01466409	-0.032705037	-0.05721802	0.15979990	0.12012489	-0.023248868	0.032090513	0.18105560	1.00000000

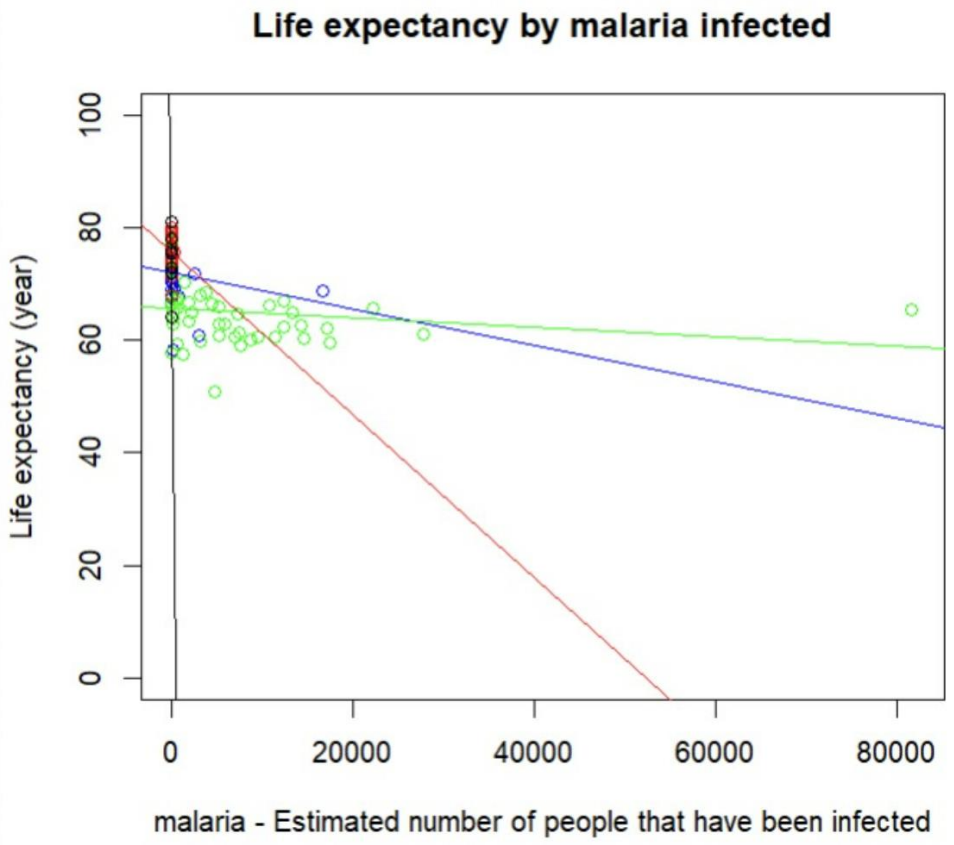
תרשים פיזור:

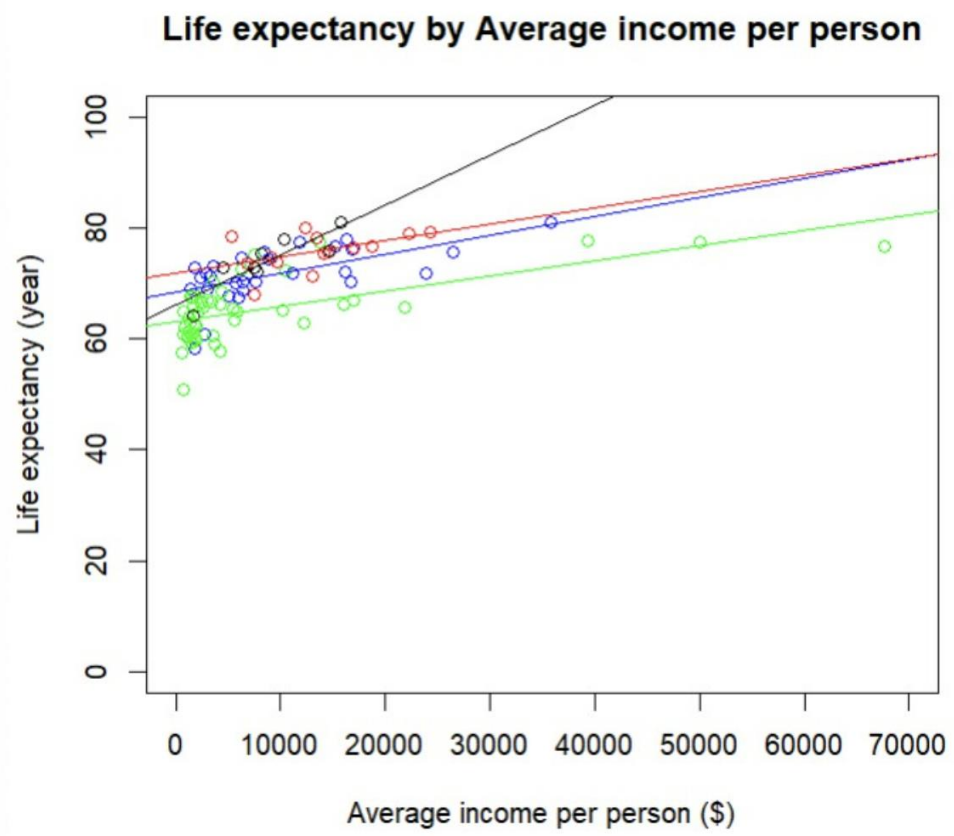


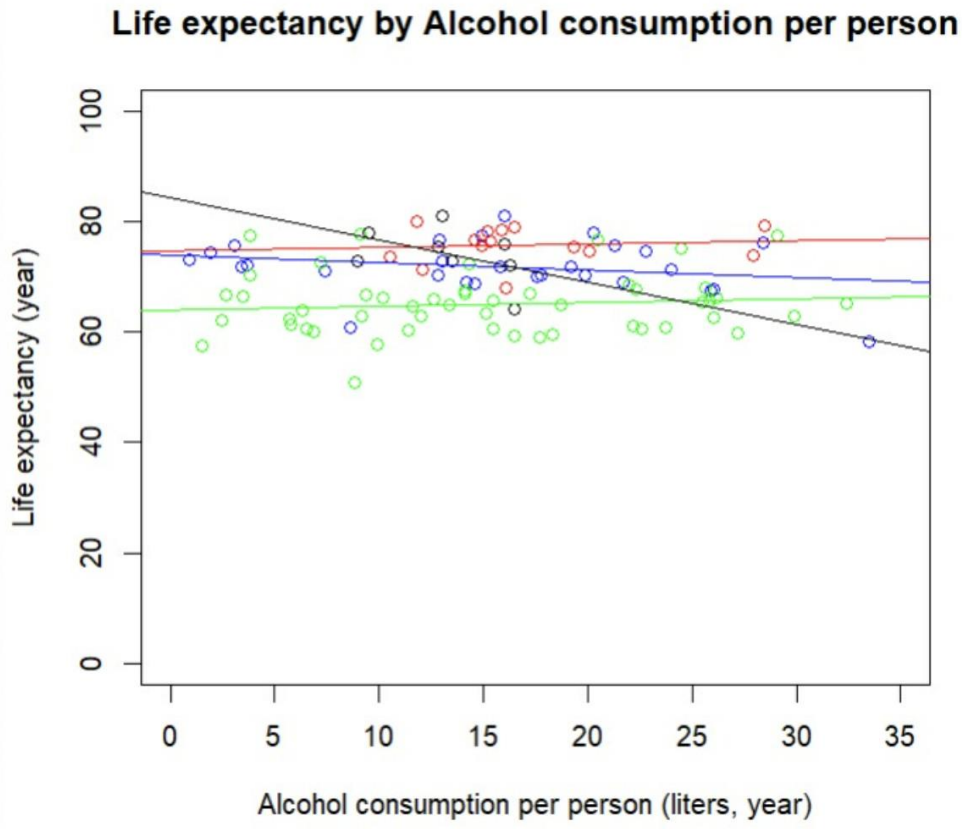
2.4 משתני אינטראקציה

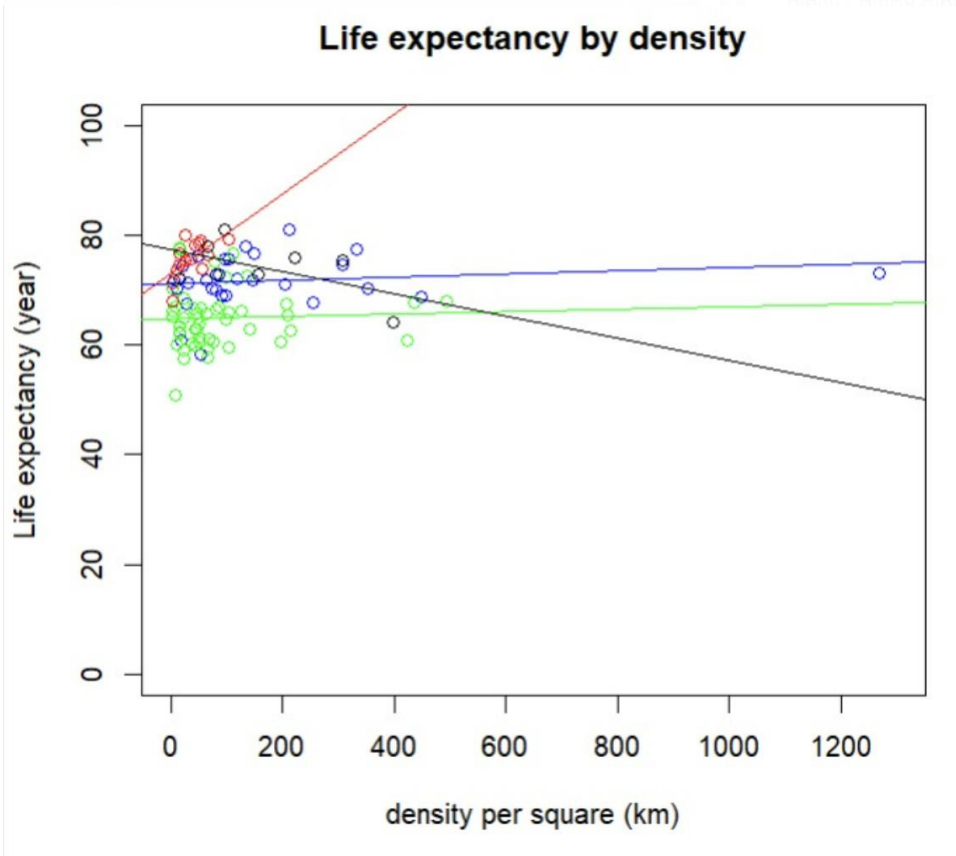


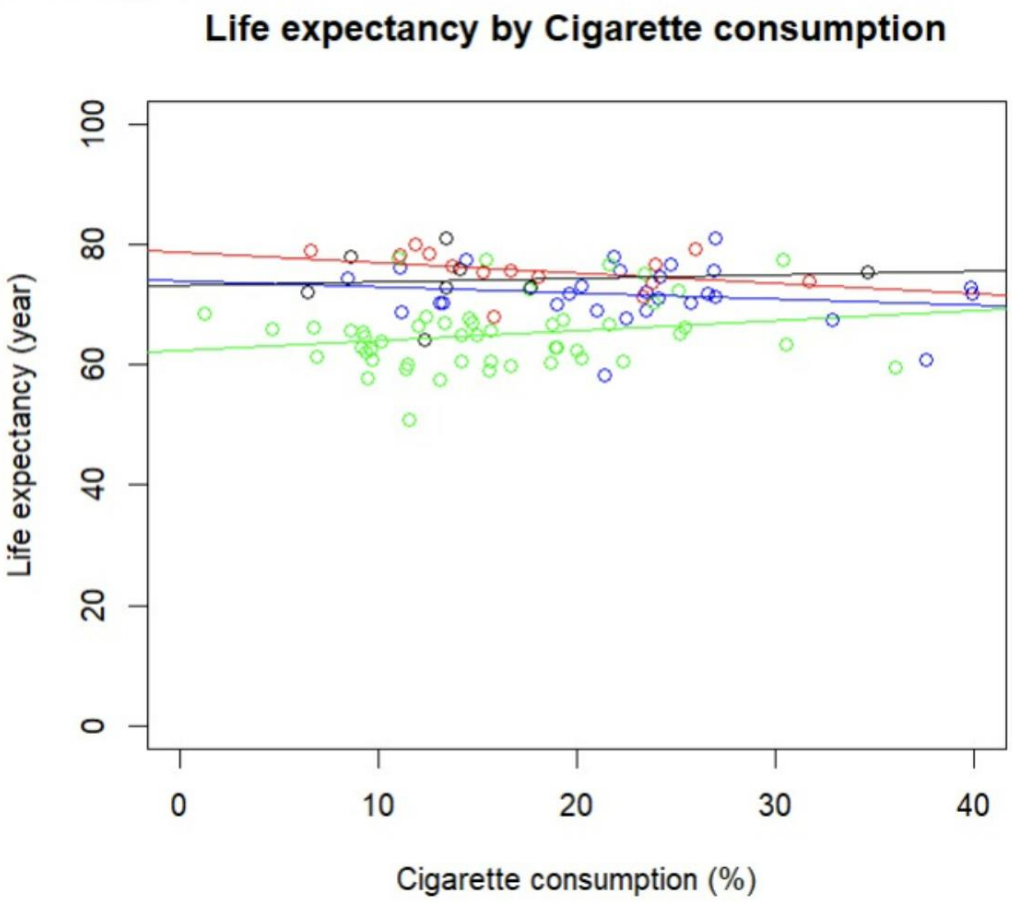


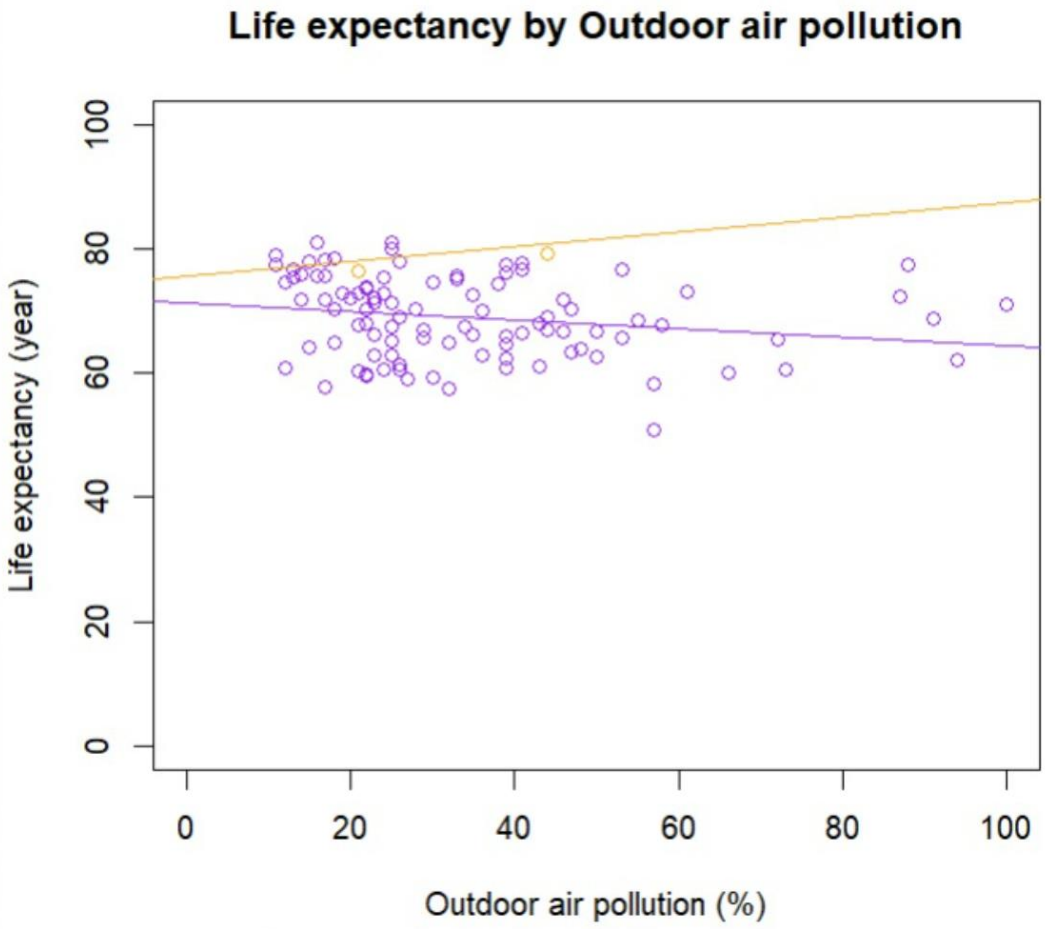

Life expectancy by malaria infected
Life expectancy (year)
0
20
40
60
80
100
0
20000
40000
60000
80000
malaria - Estimated number of people that have been infected

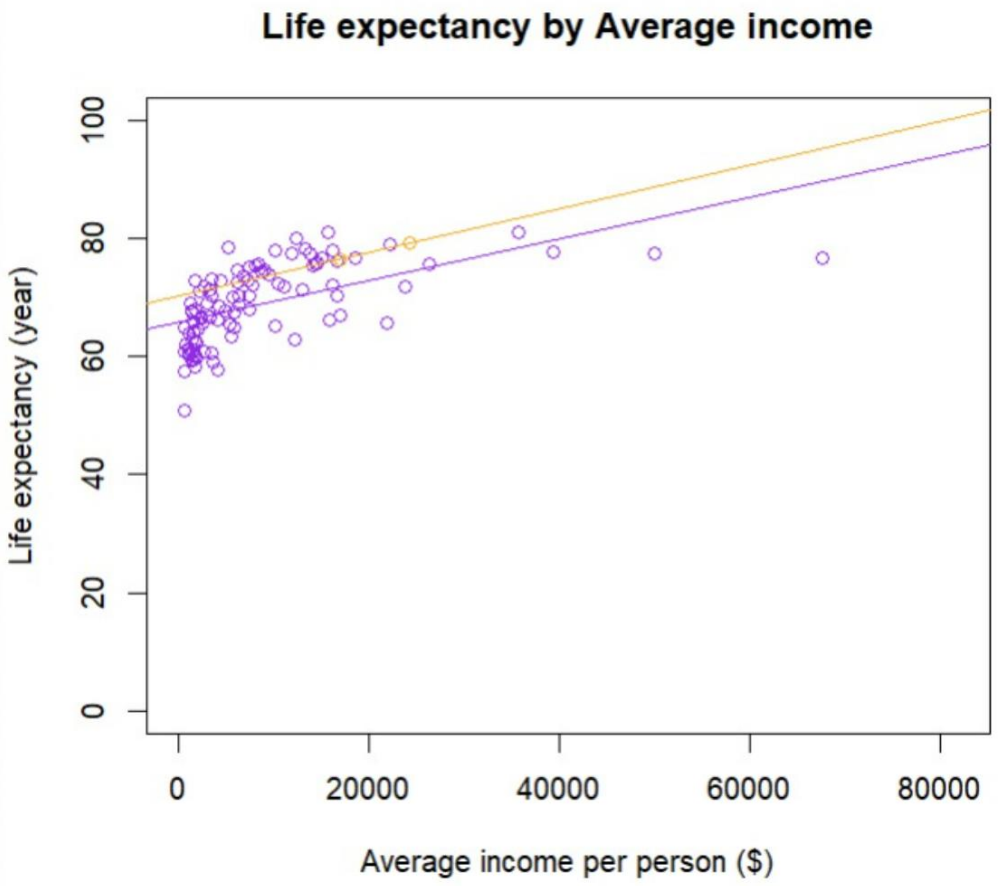

Life expectancy by Average income per person
Life expectancy (year)
0
20
40
60
80
100
0
10000
20000
30000
40000
50000
60000
70000
Average income per person ($)

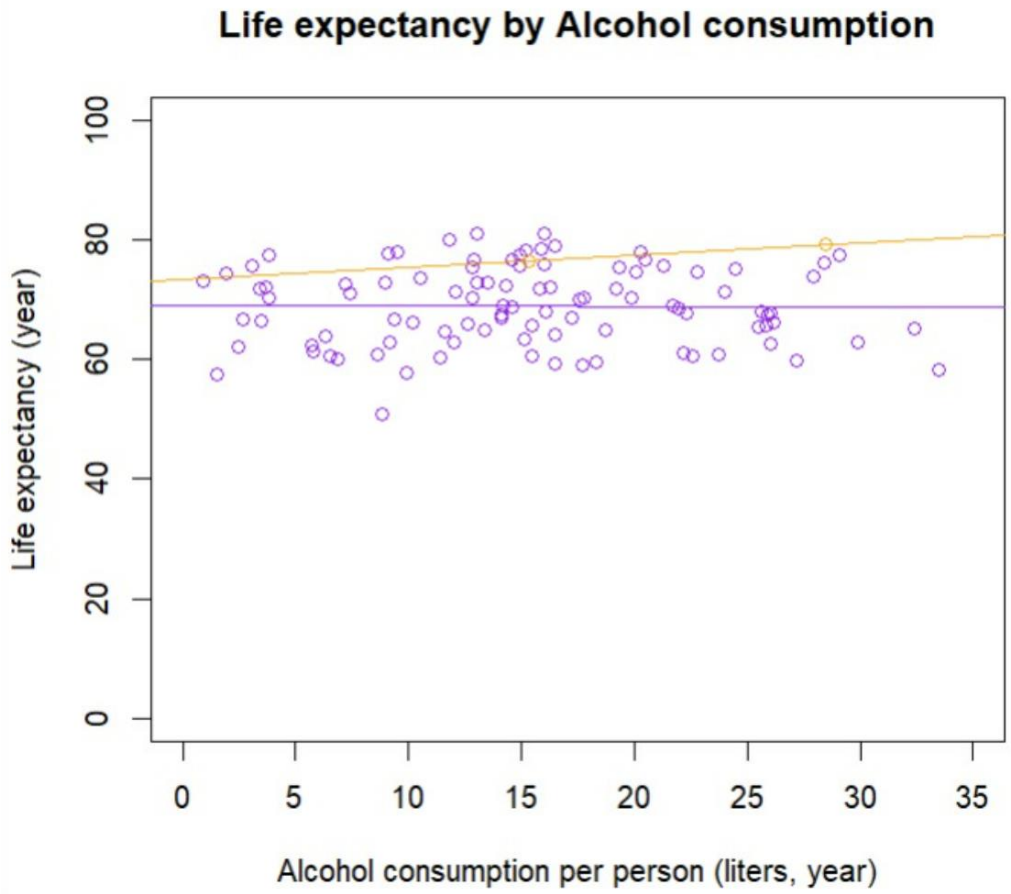

Life expectancy by Alcohol consumption per person
Life expectancy (year)
0
20
40
60
80
100
0
5
10
15
20
25
30
35
Alcohol consumption per person (liters, year)

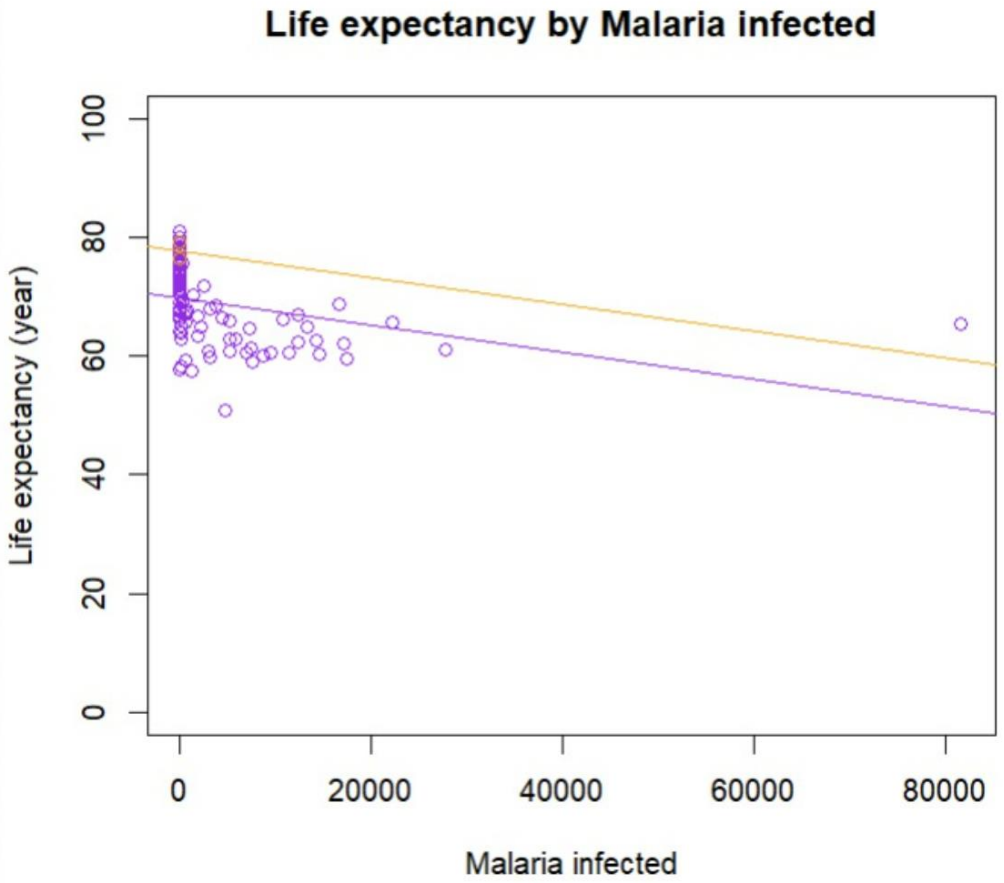

Life expectancy by density
Life expectancy (year)
0
20
40
60
80
100
0
200
400
600
800
1000
1200
density per square (km)

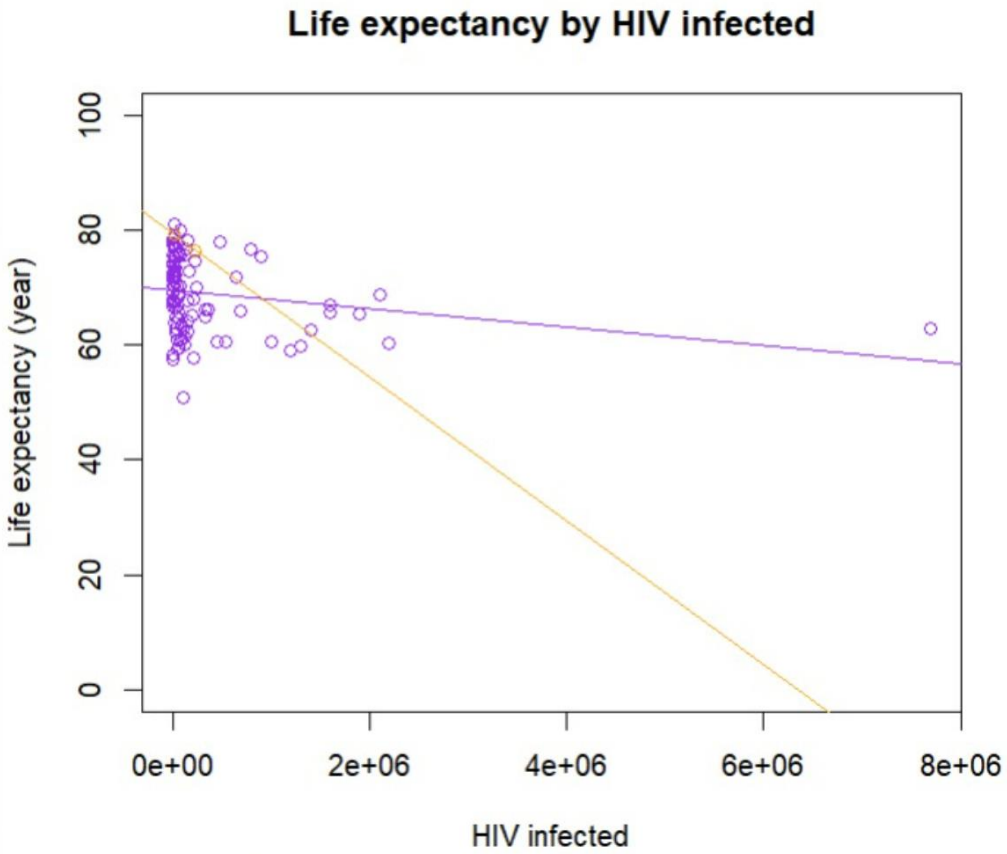

Life expectancy by Cigarette consumption
Life expectancy (year)
0
20
40
60
80
100
0
10
20
30
40
Cigarette consumption (%)

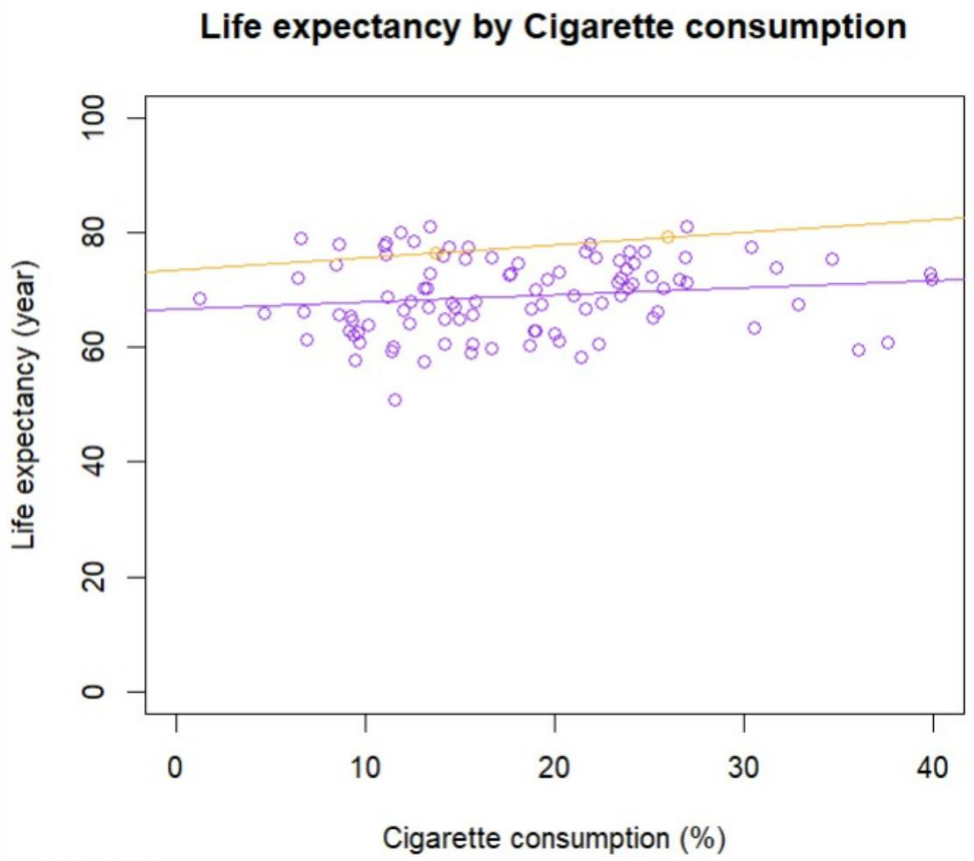

Life expectancy by Outdoor air pollution
Life expectancy (year)
0
20
40
60
80
100
0
20
40
60
80
100
Outdoor air pollution (%)

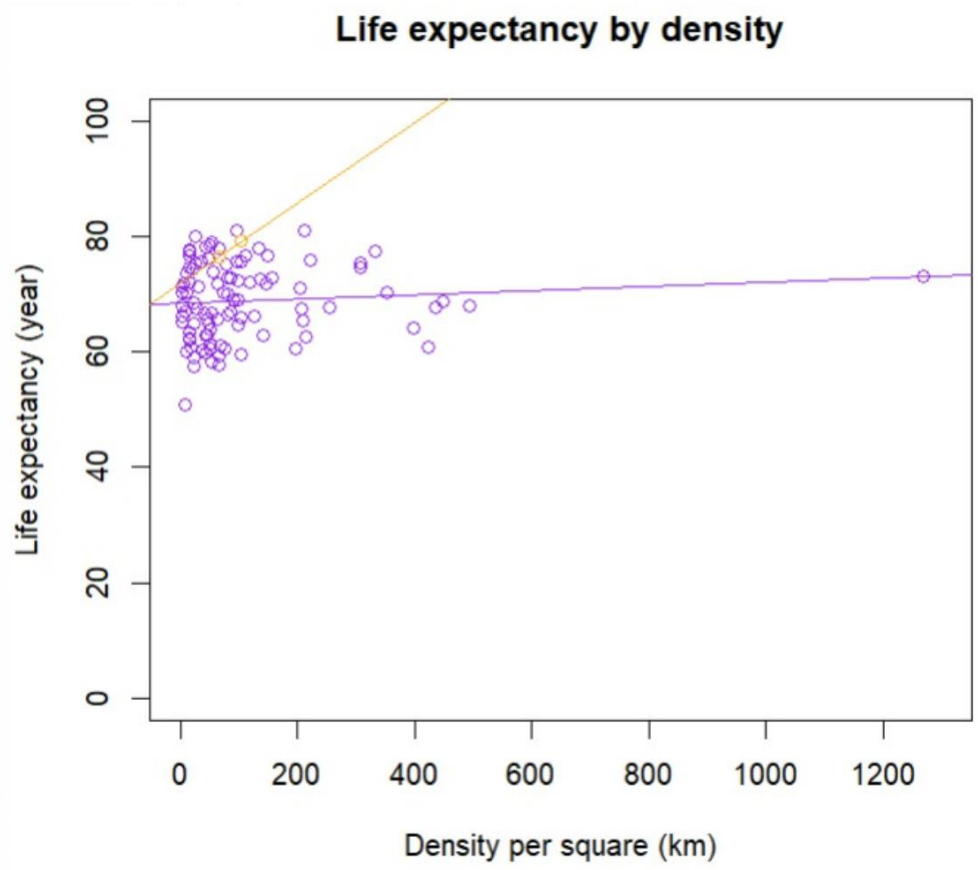

Life expectancy by Average income
Life expectancy (year)
0
20
40
60
80
100
0
20000
40000
60000
80000
Average income per person ($)


Life expectancy by Alcohol consumption
Life expectancy (year)
0
20
40
60
80
100
0
5
10
15
20
25
30
35
Alcohol consumption per person (liters, year)


Life expectancy by Malaria infected
Life expectancy (year)
0
20
40
60
80
100
0
20000
40000
60000
80000
Malaria infected


Life expectancy by HIV infected
Life expectancy (year)
0
20
40
60
80
100
0e+00
2e+06
4e+06
6e+06
8e+06
HIV infected


Life expectancy by Cigarette consumption
Life expectancy (year)
0
20
40
60
80
100
0
10
20
30
40
Cigarette consumption (%)


Life expectancy by density
Life expectancy (year)
0
20
40
60
80
100
0
200
400
600
800
1000
1200
Density per square (km)

3.1 מודל מלא

```
> NewModel <- lm(DatasetNew$Y ~ DatasetNew$X1 + DatasetNew$X2 * factor(DatasetNew$X8) +
+                DatasetNew$X2 * factor(DatasetNew$X9)+  DatasetNew$X3 * factor(DatasetNew$X8)+
+                  DatasetNew$X3 + DatasetNew$X4 )
> summary(NewModel)

Call:
lm(formula = DatasetNew$Y ~ DatasetNew$X1 + DatasetNew$X2 * factor(DatasetNew$X8) +
    DatasetNew$X2 * factor(DatasetNew$X9) + DatasetNew$X3 * factor(DatasetNew$X8) +
    DatasetNew$X3 + DatasetNew$X4)

Residuals:
     Min       1Q   Median       3Q      Max
-12.7652  -2.0791  -0.3697   2.4974  10.3010

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                          6.810e+01  1.206e+00  56.453  < 2e-16 ***
DatasetNew$X1                        1.209e-02  2.499e-02   0.484   0.6297
DatasetNew$X2                        8.483e-06  3.971e-06   2.136   0.0355 *
factor(DatasetNew$X8)2              -5.139e+00  1.107e+00  -4.642 1.24e-05 ***
factor(DatasetNew$X8)3               3.785e+00  1.637e+00   2.312   0.0232 *
factor(DatasetNew$X8)5               4.604e+00  2.966e+00   1.552   0.1244
factor(DatasetNew$X9)1               8.691e-02  4.550e+00   0.019   0.9848
DatasetNew$X3                       -1.275e-03  5.271e-04  -2.419   0.0177 *
DatasetNew$X4                        2.807e-04  4.010e-05   7.000 5.56e-10 ***
DatasetNew$X2:factor(DatasetNew$X8)2 -9.015e-06  4.009e-06  -2.249   0.0271 *
DatasetNew$X2:factor(DatasetNew$X8)3 -8.171e-06  6.237e-06  -1.310   0.1937
DatasetNew$X2:factor(DatasetNew$X8)5 -1.154e-05  7.504e-05  -0.154   0.8781
DatasetNew$X2:factor(DatasetNew$X9)1 -2.044e-06  2.676e-05  -0.076   0.9393
factor(DatasetNew$X8)2:DatasetNew$X3  1.251e-03  5.278e-04   2.371   0.0200 *
factor(DatasetNew$X8)3:DatasetNew$X3 -4.717e-04  9.273e-03  -0.051   0.9596
factor(DatasetNew$X8)5:DatasetNew$X3 -1.051e-01  1.303e-01  -0.807   0.4222
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.985 on 85 degrees of freedom
Multiple R-squared:  0.6927,    Adjusted R-squared:  0.6384
F-statistic: 12.77 on 15 and 85 DF,  p-value: 5.828e-16
```

רגרסיה לפנים:

```
> fwd.model <-  step(Emp, direction='forward', scope= ~ DatasetNew$Y ~ DatasetNew$X1 + DatasetNew$X2 * factor(DatasetNew$X8) +
+                     DatasetNew$X2 * factor(DatasetNew$X9)+  DatasetNew$X3 * factor(DatasetNew$X8)+
+                     DatasetNew$X3 + DatasetNew$X4)
Start:  AIC=383.02
Y ~ 1

                       Df Sum of Sq    RSS    AIC
+ factor(DatasetNew$X8)  3   1894.12 2498.2 332.03
+ DatasetNew$X4          1   1480.53 2911.8 343.50
+ DatasetNew$X3          1    514.01 3878.3 372.45
+ DatasetNew$X2          1    200.27 4192.0 380.31
+ DatasetNew$X1          1    189.09 4203.2 380.58
+ factor(DatasetNew$X9)  1    159.05 4233.2 381.30
<none>                              4392.3 383.02

Step:  AIC=332.03
Y ~ factor(DatasetNew$X8)

                       Df Sum of Sq    RSS    AIC
+ DatasetNew$X4          1    956.28 1541.9 285.29
+ DatasetNew$X3          1     68.90 2429.3 331.20
<none>                              2498.2 332.03
+ DatasetNew$X2          1     33.37 2464.8 332.67
+ factor(DatasetNew$X9)  1     10.29 2487.9 333.61
+ DatasetNew$X1          1      9.55 2488.6 333.64

Step:  AIC=285.29
Y ~ factor(DatasetNew$X8) + DatasetNew$X4

                       Df Sum of Sq    RSS    AIC
<none>                              1541.9 285.29
+ DatasetNew$X2          1   23.3191 1518.6 285.75
+ DatasetNew$X3          1    9.1559 1532.7 286.69
+ DatasetNew$X1          1    3.1690 1538.7 287.08
+ factor(DatasetNew$X9)  1    0.0120 1541.9 287.29


> summary(fwd.model)

Call:
lm(formula = Y ~ factor(DatasetNew$X8) + DatasetNew$X4, data = DatasetNew)

Residuals:
     Min       1Q   Median       3Q      Max
-12.1766  -2.2877  -0.3221   2.8807  10.5686

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             6.885e+01  8.483e-01  81.165  < 2e-16 ***
factor(DatasetNew$X8)2 -5.973e+00  9.484e-01  -6.299 9.04e-09 ***
factor(DatasetNew$X8)3  2.954e+00  1.320e+00   2.238   0.0275 *
factor(DatasetNew$X8)5  2.641e+00  1.607e+00   1.643   0.1036
DatasetNew$X4           3.012e-04  3.903e-05   7.716 1.13e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.008 on 96 degrees of freedom
Multiple R-squared:  0.649,     Adjusted R-squared:  0.6343
F-statistic: 44.37 on 4 and 96 DF,  p-value: < 2.2e-16
```

רגרסיה לאחור

```
> bw.model <-  step(Full, direction='backward', scope= ~1)
Start:  AIC=293.86
DatasetNew$Y ~ DatasetNew$X1 + DatasetNew$X2 * factor(DatasetNew$X8) +
    DatasetNew$X2 * factor(DatasetNew$X9) + DatasetNew$X3 * factor(DatasetNew$X8) +
    DatasetNew$X3 + DatasetNew$X4

                                      Df Sum of Sq    RSS    AIC
- DatasetNew$X2:factor(DatasetNew$X9)  1      0.09 1350.0 291.87
- DatasetNew$X1                        1      3.72 1353.7 292.14
- DatasetNew$X2:factor(DatasetNew$X8)  3     80.67 1430.6 293.72
<none>                                             1349.9 293.86
- factor(DatasetNew$X8):DatasetNew$X3  3    100.03 1450.0 295.08
- DatasetNew$X4                        1    778.17 2128.1 337.83

Step:  AIC=291.87
DatasetNew$Y ~ DatasetNew$X1 + DatasetNew$X2 + factor(DatasetNew$X8) +
    factor(DatasetNew$X9) + DatasetNew$X3 + DatasetNew$X4 + DatasetNew$X2:factor(DatasetNew$X8) +
    factor(DatasetNew$X8):DatasetNew$X3

                                      Df Sum of Sq    RSS    AIC
- factor(DatasetNew$X9)                1      0.05 1350.1 289.87
- DatasetNew$X1                        1      3.85 1353.9 290.16
- DatasetNew$X2:factor(DatasetNew$X8)  3     80.59 1430.6 291.72
<none>                                             1350.0 291.87
- factor(DatasetNew$X8):DatasetNew$X3  3    100.03 1450.1 293.09
- DatasetNew$X4                        1    781.05 2131.1 335.98

Step:  AIC=289.87
DatasetNew$Y ~ DatasetNew$X1 + DatasetNew$X2 + factor(DatasetNew$X8) +
    DatasetNew$X3 + DatasetNew$X4 + DatasetNew$X2:factor(DatasetNew$X8) +
    factor(DatasetNew$X8):DatasetNew$X3

                                      Df Sum of Sq    RSS    AIC
- DatasetNew$X1                        1      3.81 1353.9 288.16
- DatasetNew$X2:factor(DatasetNew$X8)  3     80.69 1430.8 289.73
<none>                                             1350.1 289.87
- factor(DatasetNew$X8):DatasetNew$X3  3    100.01 1450.1 291.09
- DatasetNew$X4                        1    788.09 2138.2 334.31

Step:  AIC=288.16
DatasetNew$Y ~ DatasetNew$X2 + factor(DatasetNew$X8) + DatasetNew$X3 +
    DatasetNew$X4 + DatasetNew$X2:factor(DatasetNew$X8) + factor(DatasetNew$X8):DatasetNew$X3

                                      Df Sum of Sq    RSS    AIC
<none>                                             1353.9 288.16
- DatasetNew$X2:factor(DatasetNew$X8)  3     84.63 1438.5 288.28
- factor(DatasetNew$X8):DatasetNew$X3  3     97.95 1451.8 289.21
- DatasetNew$X4                        1    803.67 2157.6 333.22
```

```
> summary(bw.model)

Call:
lm(formula = DatasetNew$Y ~ DatasetNew$X2 + factor(DatasetNew$X8) +
    DatasetNew$X3 + DatasetNew$X4 + DatasetNew$X2:factor(DatasetNew$X8) +
    factor(DatasetNew$X8):DatasetNew$X3)

Residuals:
     Min       1Q   Median       3Q      Max
-12.5614  -2.2646  -0.3114   2.5176  10.2859

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                              6.847e+01  9.194e-01  74.467  < 2e-16 ***
DatasetNew$X2                            8.622e-06  3.897e-06   2.212   0.0295 *
factor(DatasetNew$X8)2                  -5.040e+00  1.071e+00  -4.706 9.32e-06 ***
factor(DatasetNew$X8)3                   3.643e+00  1.508e+00   2.416   0.0178 *
factor(DatasetNew$X8)5                   4.499e+00  2.912e+00   1.545   0.1259
DatasetNew$X3                           -1.255e-03  5.172e-04  -2.426   0.0173 *
DatasetNew$X4                            2.823e-04  3.906e-05   7.228 1.70e-10 ***
DatasetNew$X2:factor(DatasetNew$X8)2 -9.198e-06  3.927e-06  -2.342   0.0214 *
DatasetNew$X2:factor(DatasetNew$X8)3 -8.511e-06  6.061e-06  -1.404   0.1638
DatasetNew$X2:factor(DatasetNew$X8)5 -1.296e-05  7.381e-05  -0.176   0.8611
factor(DatasetNew$X8)2:DatasetNew$X3  1.236e-03  5.185e-04   2.383   0.0193 *
factor(DatasetNew$X8)3:DatasetNew$X3 -5.104e-04  9.040e-03  -0.056   0.9551
factor(DatasetNew$X8)5:DatasetNew$X3 -1.036e-01  1.282e-01  -0.808   0.4212
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.922 on 88 degrees of freedom
Multiple R-squared:  0.6918,    Adjusted R-squared:  0.6497
F-statistic: 16.46 on 12 and 88 DF,  p-value: < 2.2e-16
```

רגרסיה בצעדים

```
> sw.model <-  step(Emp, direction='both', scope= ~ DatasetNew$Y ~ DatasetNew$X1 + DatasetNew$X2 * factor(DatasetNew$X8) +
+                   DatasetNew$X2 * factor(DatasetNew$X9)+  DatasetNew$X3 * factor(DatasetNew$X8)+
+                   DatasetNew$X3 + DatasetNew$X4)
Start:  AIC=383.02
Y ~ 1

                       Df Sum of Sq    RSS    AIC
+ factor(DatasetNew$X8)  3   1894.12 2498.2 332.03
+ DatasetNew$X4          1   1480.53 2911.8 343.50
+ DatasetNew$X3          1    514.01 3878.3 372.45
+ DatasetNew$X2          1    200.27 4192.0 380.31
+ DatasetNew$X1          1    189.09 4203.2 380.58
+ factor(DatasetNew$X9)  1    159.05 4233.2 381.30
<none>                              4392.3 383.02

Step:  AIC=332.03
Y ~ factor(DatasetNew$X8)

                       Df Sum of Sq    RSS    AIC
+ DatasetNew$X4          1    956.28 1541.9 285.29
+ DatasetNew$X3          1     68.90 2429.3 331.20
<none>                              2498.2 332.03
+ DatasetNew$X2          1     33.37 2464.8 332.67
+ factor(DatasetNew$X9)  1     10.29 2487.9 333.61
+ DatasetNew$X1          1      9.55 2488.6 333.64
- factor(DatasetNew$X8)  3   1894.12 4392.3 383.02


Step:  AIC=285.29
Y ~ factor(DatasetNew$X8) + DatasetNew$X4

                       Df Sum of Sq    RSS    AIC
<none>                              1541.9 285.29
+ DatasetNew$X2          1     23.32 1518.6 285.75
+ DatasetNew$X3          1      9.16 1532.7 286.69
+ DatasetNew$X1          1      3.17 1538.7 287.08
+ factor(DatasetNew$X9)  1      0.01 1541.9 287.29
- DatasetNew$X4          1    956.28 2498.2 332.03
- factor(DatasetNew$X8)  3   1369.88 2911.8 343.50
```

```
> summary(sw.model)

Call:
lm(formula = Y ~ factor(DatasetNew$X8) + DatasetNew$X4, data = DatasetNew)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.1766  -2.2877  -0.3221   2.8807  10.5686 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)             6.885e+01  8.483e-01  81.165  < 2e-16 ***
factor(DatasetNew$X8)2 -5.973e+00  9.484e-01  -6.299 9.04e-09 ***
factor(DatasetNew$X8)3  2.954e+00  1.320e+00   2.238   0.0275 *  
factor(DatasetNew$X8)5  2.641e+00  1.607e+00   1.643   0.1036    
DatasetNew$X4           3.012e-04  3.903e-05   7.716 1.13e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.008 on 96 degrees of freedom
Multiple R-squared:  0.649,	Adjusted R-squared:  0.6343 
F-statistic: 44.37 on 4 and 96 DF,  p-value: < 2.2e-16
```

5. שיפור המודל – לאחר הטרנספורמציה

```
Call:
lm(formula = (DatasetNew$Y)^3 ~ DatasetNew$X2 + factor(DatasetNew$X8) + 
    DatasetNew$X3 + DatasetNew$X4 + DatasetNew$X2:factor(DatasetNew$X8) + 
    factor(DatasetNew$X8):DatasetNew$X3)

Residuals:
    Min      1Q  Median      3Q     Max 
-129759  -30711   -5700   34000  151542 

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)                            3.261e+05  1.261e+04  25.854  < 2e-16 ***
DatasetNew$X2                          1.255e-01  5.345e-02   2.348  0.02115 *  
factor(DatasetNew$X8)2                -6.758e+04  1.473e+04  -4.587 1.50e-05 ***
factor(DatasetNew$X8)3                 5.806e+04  2.068e+04   2.808  0.00615 ** 
factor(DatasetNew$X8)5                 7.025e+04  3.993e+04   1.759  0.08204 .  
DatasetNew$X3                         -1.853e+01  7.092e+00  -2.613  0.01058 *  
DatasetNew$X4                          4.159e+00  5.365e-01   7.753 1.57e-11 ***
DatasetNew$X2:factor(DatasetNew$X8)2 -1.341e-01  5.386e-02  -2.490  0.01468 *  
DatasetNew$X2:factor(DatasetNew$X8)3 -1.257e-01  8.312e-02  -1.512  0.13425    
DatasetNew$X2:factor(DatasetNew$X8)5 -2.227e-01  1.012e+00  -0.220  0.82639    
factor(DatasetNew$X8)2:DatasetNew$X3  1.824e+01  7.111e+00   2.565  0.01203 *  
factor(DatasetNew$X8)3:DatasetNew$X3 -1.499e+01  1.240e+02  -0.121  0.90406    
factor(DatasetNew$X8)5:DatasetNew$X3 -1.497e+03  1.758e+03  -0.851  0.39689    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 53790 on 87 degrees of freedom
Multiple R-squared:  0.714,	Adjusted R-squared:  0.6745 
F-statistic:  18.1 on 12 and 87 DF,  p-value: < 2.2e-16
```

קיום הנחות המודל:

```
        Shapiro-Wilk normality test

data:  Dataset$stan_residuals2
W = 0.97991, p-value = 0.1307
```

```
        Goldfeld-Quandt test

data:  FinalModelTrans
GQ = 0.75375, df1 = 22, df2 = 22, p-value = 0.5128
alternative hypothesis: variance changes from segment 1 to 2
```

```
        M-fluctuation test

data:  FinalModelTrans
f(efp) = 1.7236, p-value = 0.06622
```