פרויקט ברגרסיה לינארית

ניתוח פורמאלי של מאגר הנתונים

קבוצה 8

אופיר צרפתי 308244573

נעה מתתיהו 315474726

נוי כשר 314963810

תוכן עניינים

1. תקציר מנהלים

מטרת הפרויקט הינה יצירת מודל רגרסיה לינארית חזק ומדויק ככל הניתן לחיזוי תוחלת החיים. לצורך מטרה זו, השתמשנו בבסיס נתונים של WHO – ארגון הבריאות העולמי, על מנת לחזות את תוחלת החיים. בבסיס הנתונים היו בתחילה כ-20 משתנים ותצפיות על יותר מ-200 מדינות.

בחלק זה של הפרויקט, בחנו לעומק בעזרת כלים סטטיסטיים את מידת ההתאמה של המשתנים המסבירים למשתנה המוסבר. נעזרנו בתרשימי פיזור (המחשה גרפית) ומבחנים סטטיסטים שנלמדו בכיתה.

לאחר ניתוח ראשוני, בחנו את הסרתם של שישה משתנים מסבירים (אך לא הסרנו אותם). בנוסף, הפכנו שני משתנים רציפים לקטגוריאליים והוספנו שלושה משתנים אינטראקציה.

לאחר מכן, בחנו את המודל בעזרת המדדים שנלמדו בכיתה: AIC ו – BIC. ביצענו רגרסיה לפנים, רגרסיה לאחור ורגרסיה בצעדים בעבור כל אחד מהם, ובדקנו באיזו חלופה התקבל מדד R^2_{adj} הגבוה ביותר. בהתאם לזאת, בחרנו במודל שהתקבל לפי AIC ורגרסיה לאחור.

לאחר בחירת המודל, בחנו האם מתקיימות ההנחות של מודל רגרסיה לינארית; שוויון שונויות, הנחת הלינאריות והנחת הנורמליות. לאחר הבדיקה, נוכחנו לגלות שהנחת הנורמליות איננה מתקיימת.

בשלב האחרון, בדקנו האם ניתן לשפר את המודל שלנו בעזרת טרנספורמציות על משתנים. לאחר בדיקת הטרנספורמציות, גילינו כי ישנם כמה משתנים מסבירים שהקורלציה שלהם עם המשתנה המוסבר עלתה בעקבות הטרנספורמציה. לאחר ביצוע הטרנספורמציות הנבחרות, ביצענו שוב רגרסיה על כל המדדים והשווינו בין החלופות שהתקבלו. לאחר בחינת הממצאים, בחרנו הפעם במודל במלא, מכיוון שבעבורו המדד R^2_{adj} היה הגבוה ביותר. המודל שקיבלנו לאחר הטרנספורמציות היה חזק מהמודל שקיבלנו לפני הטרנספורמציות, ולכן בחרנו בו כמודל הסופי.

המודל שקיבלנו הינו מדויק וחזק. ערך המדד שהתקבל הינו 0.8978, ערך המעיד על טיב המודל הסופי שנבחר.

2. עיבוד מקדים

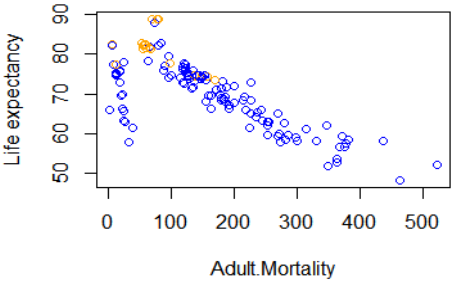
2.1 הסרה של משתנים

נבדוק אם הסרת משתנים תורמת ליצירת מודל טוב יותר. על מנת להבין אילו משתנים מועמדים להסרה, נבצע מבחן פירסון בין המשתנים הרציפים ונבדוק את ערכי הקצה.

תמותת מבוגרים:

מתאם פירסון: -0.7700689

מקדם המתאם גבוה מאוד ומרבית התצפיות בתרשים הפיזור נמצאות סביב אותו קו מגמה.

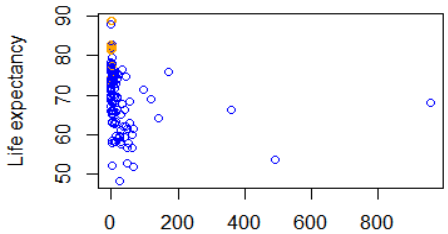


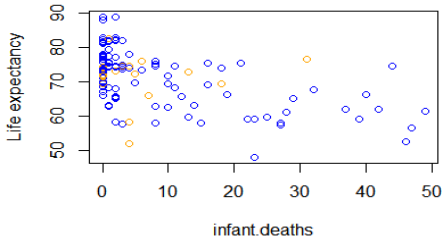
תמותת תינוקות:

מתאם פירסון: -0.2009425

מקדם פירסון לאחר חריגים: -0.528

מקדם המתאם נמוך ובעוד אנו מסתכלים על תרשים הפיזור ניתן לראות כי מרבית התצפיות מרוכזות סביב ה-0. לעומת זאת, כאשר מסירים את התצפיות החריגות, ניתן לראות כי מקדם פירסון גדל וניתן לראות קו מגמה ברור יותר.

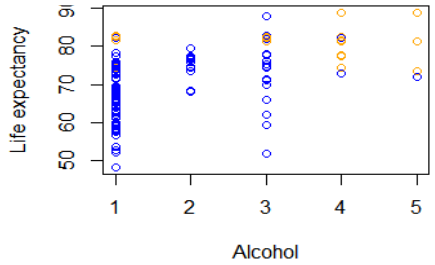




אלכוהול:

מתאם פירסון: 0.50187

זהו משתנה קטגוריאלי. מקדם הקורלציה הינו ממוצע, וניתן לראות בתרשים הפיזור כי ההתפלגות הינה בעלת זנב שמאלי.

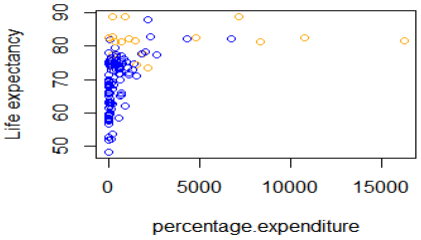


הוצאות בריאות כאחוז מהתוצר המקומי הגולמי לנפש:

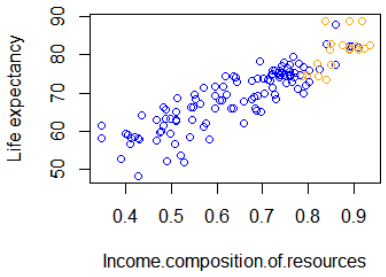
מתאם פירסון: 0.41259

מתאם פירסון לאחר חריגים: 0.3964355

מקדם הקורלציה קרוב לממוצע אך מעט נמוך ממנו. לאחר הסרת התצפיות החריגות ניתן לראות המתאם ירד ל-0.3964 ומרבית הרשומות נמצאות סביב ה-0 ולא סביב קו מגמה ברור.



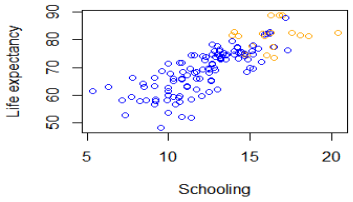
<u>מדד התפתחות האדם מבחינת הרכב ההכנסות של המשאבים:</u>



מתאם פירסון: 0.892017

קיים קשר חזק מאוד בין התפחות האדם מבחינת הרכבת ההכנסות לבין תוחלת החיים, ובתרשים הפיזור ניתן לראות כי קיים קו מגמה ברור.

<u>מספר שנות לימוד:</u>



מתאם פירסון: 0.8017963

מתאם פירסון גבוה מאוד ובתרשים הפיזור ניתן לראות קו מגמה ברור – ככל שמספר שנות הלימוד עולה, כך גם תוחלת החיים עולה.

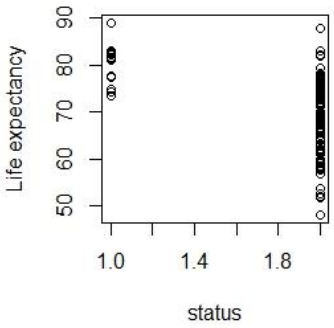
משתנים נוספים יפורטו בנספחים.

<u>לסיכום:</u>

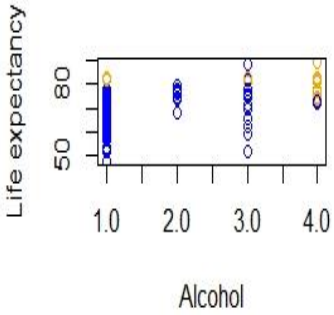
משתנים שנשקול להוריד הינ: הוצאות בריאות כאחוז מהתוצר המקומי הגולמי לנפש, צהבת B, שחפת, פוליו, כיסוי החיסונים נגד טטנוס ושעלת בקרבי ילדים עד גיל שנה ואוכלוסייה.

2.2 התאמת משתנים

משתנים קטגוריים:



- **סטטוס מדינה** X_1 (כאשר 1 – מדינה מפותחת, 2 – מדינה מתפתחת):
  ניתן לראות בגרף הפיזור כי קיים פיזור הגיוני בין המדינות המתפתחות למפותחות. במדינות מתפתחות בהן התברואה נמוכה, פיזור תוחלת החיים גדול יותר ממדינות מפותחות, ומושפע מגורמים רבים.
  לאחר בחינה של הנושא אנו מבינים כי אין באפשרותנו לאחד קטגוריות כיוון שמשתנה המבדיל בין סוגי המדינות הוא משמעותי לתיאור המודל.



- **צריכת אלכוהול לנפש מעל גיל 15 בליטרים** X_4:
  ניתן לראות בגרף הפיזור כי ישנו פיזור גדול כאשר הקטגוריה הינה 1 או 3. כמו כן, ניתן להסיק מהגרף כי במרבית המדינות המפותחות צריכת האלכוהול גבוהה ביחס למדינות מתפתחות (קטגוריה 4), אך עם זאת תוחלת החיים הינה גבוהה ועל כן נאמר כי אינה מוסברת על ידי משתנה זה. לאור מסקנות אלו, איחוד הקטגוריות עלול ליצור קשר לינארי שאינו מדויק או אמיתי ולפגוע בקורלציה בין המשתנים, ואכן הקורלציה ירדה מ-0.509 ל– 0.495.

משתנים רציפים:

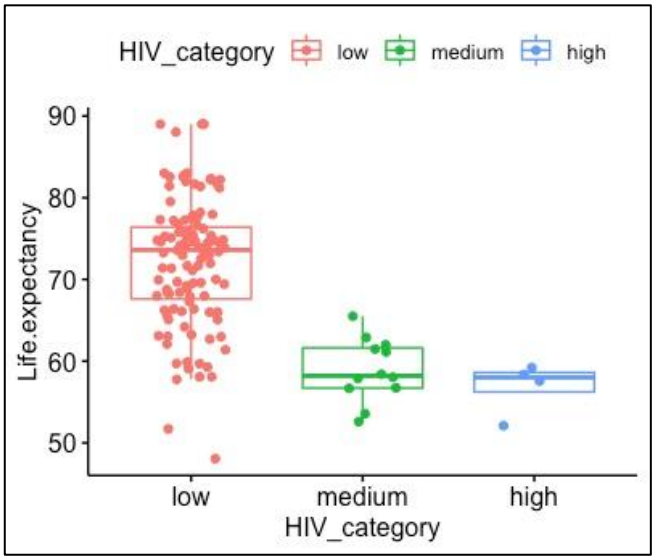
- **מספר מקרי מוות בקרב לידות נשאי איידס בגילאים 0-4 שנים X_{13}:**

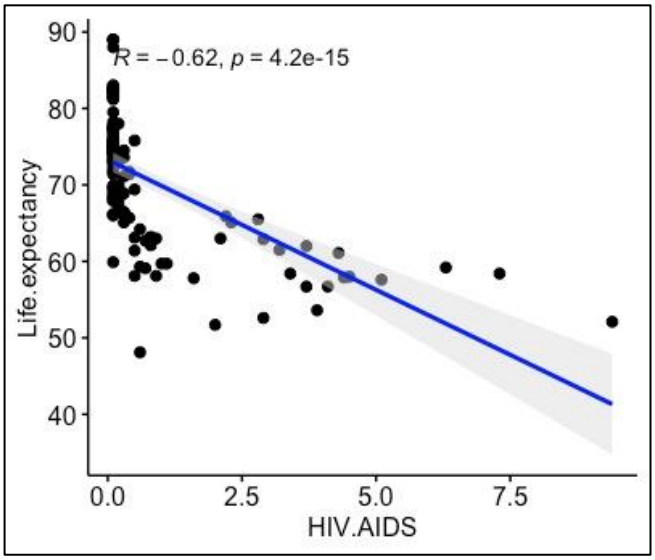
על פי התרשים, אנו רואים כי ישנו ריכוז גבוה של נתונים בערך הקטן מ 2.5, ולכן נבחן את המשתנה הזה כקטגוריאלי. בחרנו לייצג את הערכים בצורה הבאה:

$$low \rightarrow x < 2.5$$

$$medium \rightarrow \ 2.5 \leq x < 5$$

$$high \rightarrow x \geq 5$$





לאחר ביצוע מבחן התאמה עם המשתנה הקטגוריאלי קיבלנו רמת מובהקות הרבה פחות טובה. **בעקבות כך החלטנו להשאיר משתנה זה כמשתנה רציף.**

מבחן התאמה למשתנה רציף:

```
Call:
lm(formula = data$Life.expectancy ~ data$HIV.AIDS, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-23.1335  -4.4133   0.7067   3.4667  16.0667 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    73.2733     0.6695 109.440  < 2e-16 ***
data$HIV.AIDS  -3.3996     0.3817  -8.907 4.19e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.797 on 129 degrees of freedom
Multiple R-squared:  0.3808,	Adjusted R-squared:  0.376 
F-statistic: 79.34 on 1 and 129 DF,  p-value: 4.192e-15
```

מבחן התאמה משתנה קטגוריאלי:

```
Call:
lm(formula = data$Life.expectancy ~ data$HIV_category, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-24.108  -4.258   1.192   3.742  16.792 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)              72.2078     0.6862 105.232  < 2e-16 ***
data$HIV_categorymedium -13.2995     2.2323  -5.958 2.32e-08 ***
data$HIV_categoryhigh   -15.3828     3.7426  -4.110 7.02e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.358 on 128 degrees of freedom
Multiple R-squared:  0.28,	Adjusted R-squared:  0.2688 
F-statistic: 24.89 on 2 and 128 DF,  p-value: 7.375e-10
```

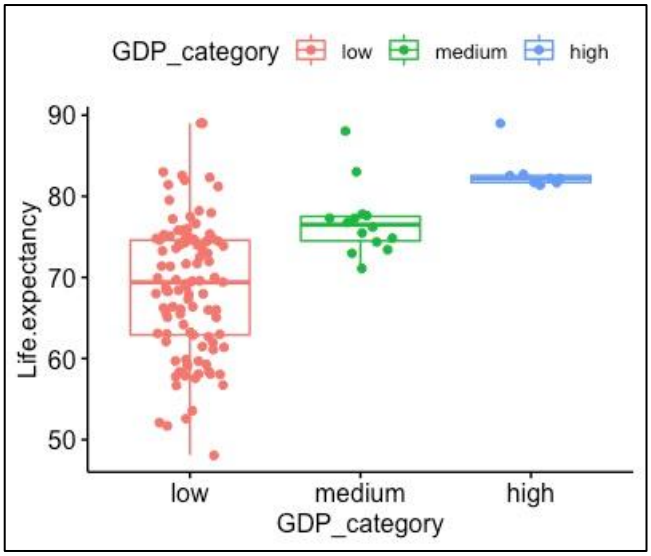
- **תוצר מקומי גולמי לנפש X_{14}:**

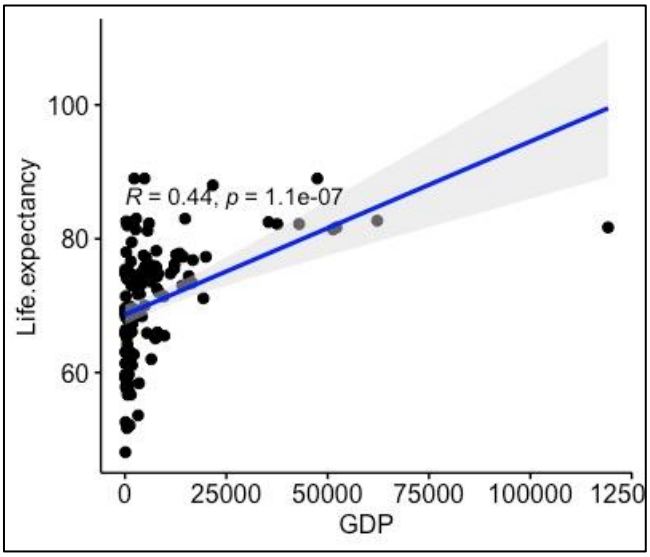
על פי התרשים, ניתן לראות כי ישנו ריכוז גבוה של נתונים בערך הקטן מ 10000, ולכן נבחן את המשתנה הזה כקטגוריאלי. בחרנו לייצג את הערכים בצורה הבאה:

$$low \rightarrow x < 10k$$

$$medium \rightarrow \ 10k \leq x < 25k$$

$$high \rightarrow x \geq 25k$$





לאחר ביצוע מבחן התאמה עם המשתנה הקטיגוריאלי קיבלנו רמת מובהקות יותר טובה ממה שקיבלנו טובה. **בעקבות כך החלטנו להחליף משתנה רציף זה למשתנה קטגוריאלי עם הערכים הנ״ל.**

מבחן התאמה משתנה רציף:

```
Call:
lm(formula = data$Life.expectancy ~ data$GDP, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-20.5603  -5.6061   0.9268   5.0740  19.7700 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.864e+01  7.544e-01  90.981  < 2e-16 ***
data$GDP    2.590e-04  4.606e-05   5.624  1.1e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.741 on 129 degrees of freedom
Multiple R-squared:  0.1969,	Adjusted R-squared:  0.1907 
F-statistic: 31.63 on 1 and 129 DF,  p-value: 1.103e-07
```

מבחן התאמה משתנה קטגוריאלי:

```
Call:
lm(formula = data$Life.expectancy ~ data$GDP_category, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-20.6936  -5.0936  -0.0714   5.7064  20.2064 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)             68.7936     0.7335  93.793  < 2e-16 ***
data$GDP_categorymedium  8.0779     2.1740   3.716 0.000302 ***
data$GDP_categoryhigh   14.1314     2.8050   5.038 1.57e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.658 on 128 degrees of freedom
Multiple R-squared:  0.2203,	Adjusted R-squared:  0.2081 
F-statistic: 18.08 on 2 and 128 DF,  p-value: 1.212e-07
```

2.3 הגדרת משתני דמה

במודל שלנו רוב המשתנים המסבירים הינם כמותיים. בנוסף אליהם, קיימים משתנים מסבירים שהינם משתנים איכותיים, כלומר קטגוריים. כדי לשלבם ברגרסיה כמשתנים מסבירים, נבצע שימוש במשתני דמי ונגדיר את קבוצת הבסיס להיות קטגוריה 1. משתני הדמי יציינו את התרומה השולית של המשתנים שאינם בקבוצת הבסיס עבור החותך שכבר קיים במודל.

המשתנים הקטגוריים:

- **סטטוס:** 1 – מדינה מפותחת, 2 – מדינה מתפתחת.
- **אלכוהול:** צריכת אלכוהול לנפש מעל גיל 15 בליטרים כאשר:

  [0.00518, 3.0460) – 1

  [3.046, 6.082) – 2

  [6.082, 9.118) – 3

  [9.118, 12.154) – 4

- **תוצר מקומי לנפש:**

$$x < 10k : 1$$
$$10k \le x < 25k : 2$$
$$x \ge 25k : 3$$

$$Status = \begin{cases} 0, & else \\ 1, & developing \end{cases} \qquad GDP_2 = \begin{cases} 0, & else \\ 1, & 10k \le x < 25k \end{cases}$$

$$Alcohol_2 = \begin{cases} 0, & else \\ 1, & [3.046, 6.082) \end{cases} \qquad GDP_3 = \begin{cases} 0, & else \\ 1, & x \ge 25k \end{cases}$$

$$Alcohol_3 = \begin{cases} 0, & else \\ 1, & [6.082, 9.118) \end{cases}$$

$$Alcohol_4 = \begin{cases} 0, & else \\ 1, & [9.118, 12.154) \end{cases}$$

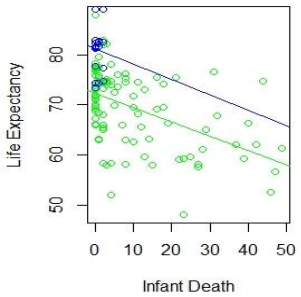
2.4 הוספת משתני אינטראקציה

משתני האינטראקציה יציינו את התרומה השולית של המשתנים שאינם נמצאים בקבוצת הבסיס עבור השיפוע. נבחן משתני אינטראקציה מעניינים:

- **סטטוס המדינה X_1 * תמותת תינוקות X_3:**

אנו מניחים כי קיים ישיר בין מצב המדינה, כלומר האם המדינה הינה מדינה מפותחת או מתפתחת, לבין מספר מקרי המוות בקרב תינוקות.

כאשר: ירוק – מדינות מתפתחות, כחול – מדינות מפותחות.



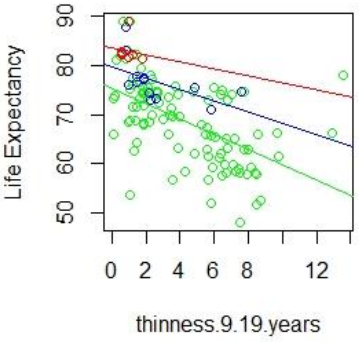

```
Coefficients:
                                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                                             81.22034    1.88732  43.035  < 2e-16 ***
no_harigim$infant.deaths                                -0.30720    1.83953  -0.167    0.868
factor(no_harigim$Status)2                              -8.89505    2.05955  -4.319 3.38e-05 ***
no_harigim$infant.deaths:factor(no_harigim$Status)2  0.02355    1.84022   0.013    0.990
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.483 on 113 degrees of freedom
Multiple R-squared:  0.4204,    Adjusted R-squared:  0.405
F-statistic: 27.32 on 3 and 113 DF,  p-value: 2.32e-13
```

כפי שציפינו, במדינות מפותחות מספר מקרי המוות בקרב תינוקות הוא נמוך ותוחלת החיים גבוהה, ואילו במדינות מתפתחות קיים קשר הפוך בין השניים. קיים הבדל התחלתי מובהק בתוחלת החיים בין מדינות מפותחות למתפתחות, עם זאת, לא קיים הבדל מובהק בקצב השינוי של תוחלת החיים בין שני סוגי המדינות, זאת בשל פיזור קטן מאוד במדינות המפותחות. כמו כן, F סטטיסטי מאוד גדול ו-pvalue מאוד קטן ועל כן המודל מובהק **ונוסיף משתנה זה למודל**.

- **תמ"ג X_{14} * שכיחות הרזון בקרב ילדים ובני נוער בגילאי 10-19 X_{16}**

לאחר שינוי משתנה התמ"ג למשתנה קטגוריאלי רצינו לבדוק כיצד זה משפיע על שכיחות הרזון בקרב בני הנוער.

כאשר: ירוק – תמ"ג נמוך, כחול – תמ"ג בינוני, אדום – תמ"ג גבוה.

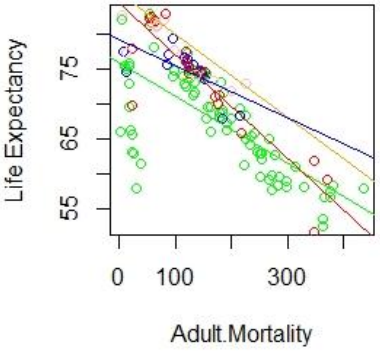



```
Coefficients:
                                                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                                                        75.5084     1.1462  65.876  < 2e-16 ***
no_harigim$thinness.9.19.years                                                     -1.5662     0.2206  -7.099 1.01e-10 ***
factor(no_harigim$GDP_category)medium                                               4.3831     3.1564   1.389    0.168
factor(no_harigim$GDP_category)high                                                 8.1045     6.5145   1.244    0.216
no_harigim$thinness.9.19.years:factor(no_harigim$GDP_category)medium   0.4141     0.9254   0.448    0.655
no_harigim$thinness.9.19.years:factor(no_harigim$GDP_category)high     0.8606     6.1323   0.140    0.889
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.586 on 118 degrees of freedom
Multiple R-squared:  0.4593,    Adjusted R-squared:  0.4364
F-statistic: 20.05 on 5 and 118 DF,  p-value: 1.979e-14
```

כפי שציפינו, קיים הבדל התחלתי מובהק בתוחלת החיים בין מדינות בעלות תמ"ג נמוך, בינוני או גבוה. בנוסף, קיים דפוס שבו במדינות בהן התמ"ג גבוה תוחלת החיים גבוהה ושכיחות הרזון נמוכה, ובמדינות בהן התמ"ג בינוני תוחלת החיים גבוהה יותר מאשר במדינות בהן התמ"ג נמוך. כמו כן, F סטטיסטי גדול ו-pvalue מאוד קטן ועל כן המודל מובהק **ונוסיף משתנה זה למודל**.

- **אלכוהול X_4 * תמותה בקרב מבוגרים X_2**

רצינו לבדוק האם יש הבדל בהשפעת התמותה בקרב מבוגרים על תוחלת החיים במדינה, כתלות בצריכת האלכוהול באותה המדינה.




```
Coefficients:
                                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                                      76.061884   1.011876  75.169  < 2e-16 ***
no_harigim$Adult.Mortality                                       -0.047869   0.004911  -9.748  < 2e-16 ***
factor(no_harigim$Alcohol)2                                       3.184495   3.138652   1.015 0.312318
factor(no_harigim$Alcohol)3                                       8.442530   2.072687   4.073 8.32e-05 ***
factor(no_harigim$Alcohol)4                                       9.952198   2.549039   3.904 0.000156 ***
no_harigim$Adult.Mortality:factor(no_harigim$Alcohol)2  0.010676   0.023613   0.452 0.651985
no_harigim$Adult.Mortality:factor(no_harigim$Alcohol)3 -0.025921   0.011410  -2.272 0.024864 *
no_harigim$Adult.Mortality:factor(no_harigim$Alcohol)4 -0.011757   0.021033  -0.559 0.577188
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.753 on 121 degrees of freedom
Multiple R-squared:  0.6884,    Adjusted R-squared:  0.6704
F-statistic: 38.19 on 7 and 121 DF,  p-value: < 2.2e-16
```

קיים הבדל התחלתי מובהק בתוחלת החיים בין מדינות עם צריכת אלכוהול שונה. בכלל המדינות מדובר בקשר הפוך בין תמותה בקרב מבוגרים לבין תוחלת החיים, כלומר ככל שהתמותה עולה כך תוחלת החיים יורדת. ניתן לראות כי קיים לא קיים הבדל מובהק בקצב השינוי של תוחלת החיים פרט למדינות בהן צריכת האלכוהול מוגברת. כלומר האלכוהול אכן מאיץ את הקטנת מדד תוחלת החיים במקרה הזה. בנוסף, F סטטיסטי מאוד גדול ו-pvalue מאוד קטן ועל כל המודל מובהק **ונוסיף משתנה זה למודל.**

לסיכום חלק 2:

התאמת משתנים: הפיכת המשתנה המסביר תמ"ג X_{14} למשתנה קטגוריאלי.

משתני דמה: סטטוס X_1, אלכוהול X_4 ותמ"ג X_{14}.

משתני אינטראקציה:

סטטוס המדינה X_1 * תמותת תינוקות X_3,

תמ"ג X_{14} * שכיחות הרזון בקרב ילדים ובני נוער בגילאי 10-19 X_{16},

אלכוהול X_4 * תמותה בקרב מבוגרים X_2

3. התאמת המודל ובדיקת הנחות המודל:

3.1 בחירת משתני המודל:

ישנם מספר מדדים המשמשים לבחירת מודלים. נבחר במדדים **AIC** ו-**BIC** אשר בוחנים את טיב ההתאמה של המודל לנתונים מבחינת הנראות שלו (ככל שהנראות יותר גדולה כך ערך המדדים קטן יותר), וגם קונס את המודל על פי מספר הפרמטרים בו (במדד AIC) ועל פי מספר התצפיות (במדד BIC). בהינתן מספר מודלים מועמדים, המודל בעל ה- AIC או ה- BIC **המינימלי** הוא המועדף.

על מנת לבחור את האפשרויות השונות נשתמש במדד $\mathbf{R^2}_{adj}$ אשר מציין את אחוז השונות המוסבר. נשאף שיהיה **מקסימלי** מדד זה. המדד מחושב באופן הבא:

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

(כאשר n - מספר התצפיות הכולל ו - p – מספר המשתנים המסבירים).

סיכום התוצאות בטבלה:

		$R^2{}_{adj}$
מודל מלא		0.8633
רגרסיה לפנים	**AIC**	0.8701
	BIC	0.8701
רגרסיה לאחור	**AIC**	**0.8737**
	BIC	0.8701
רגרסיה בצעדים	**AIC**	0.8701
	BIC	0.8701

ניתן לראות כי ערך המדד הגבוה ביותר הינו של המודל רגרסיה לאחור לפי קריטריון AIC.

המודל שהתקבל הוא:

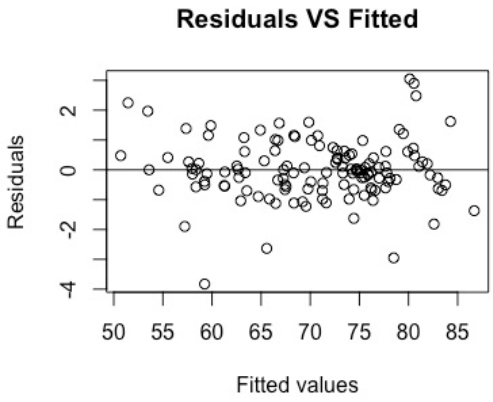
```
> summary(ourModel)

Call:
lm(formula = y ~ x2 + x3 + x9 + x11 + x13 + x18 + I4 * x2, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.1498  -1.6134  -0.2004   1.4086   8.8777 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 48.271954   2.340982  20.620  < 2e-16 ***
x2          -0.014370   0.004172  -3.444 0.000793 ***
x3           0.042073   0.030106   1.398 0.164884    
x9          -0.034397   0.022887  -1.503 0.135534    
x11          0.291814   0.113697   2.567 0.011521 *  
x13         -0.800062   0.234362  -3.414 0.000879 ***
x18         34.674897   2.984435  11.619  < 2e-16 ***
I42          1.661177   2.046457   0.812 0.418579    
I43          2.283957   1.403841   1.627 0.106419    
I44          3.513755   1.751345   2.006 0.047109 *  
x2:I42      -0.001273   0.016036  -0.079 0.936845    
x2:I43      -0.015267   0.007300  -2.091 0.038652 *  
x2:I44      -0.023442   0.013695  -1.712 0.089575 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.058 on 118 degrees of freedom
Multiple R-squared:  0.8854,	Adjusted R-squared:  0.8737 
F-statistic: 75.97 on 12 and 118 DF,  p-value: < 2.2e-16
```

$$\hat{Y} = 48.271954 - 0.01437X_2 + 0.042073X_3 - 0.034397X_9 + 0.291814X_{11} - 0.800062X_{13} + 34.674897X_{18} + 1.661177I_{42} + 2.283957I_{43} + 3.513755I_{44} - 0.001273X_2 * I_{42} - 0.015267X_2 * I_{43} - 0.023442X_2 * I_{44}$$

<u>מקרא:</u>

y - תוחלת חיים ממוצעת (בשנים)
I4 - צריכת אלכוהול
X2 – תמותה בקרב מבוגרים
X3 – תמותת תינוקות
X9 – מקרי מוות מתחת לגיל 5
X11 – אחוז הוצאה ממשלתית כללית לבריאות
X13 – מקרי מוות בקרב לידות נשאי איידס
X18 – מספר שנות לימוד

3.2 בדיקת הנחות המודל:



תחילה, נציג את תרשים פיזור השגיאות הרלוונטי לשתי בדיקות- הנחת שיוויון שונויות והנחת הלינאריות. השגיאות יחושבו באופן מנורמל, על פי הנוסחה: $e^*_{i,j} = \frac{(y_i - \hat{y}_i)}{\sqrt{V(e_i)}} = \frac{e_i}{s.e(e_i)}$.

3.2.1 הנחת שוויון שונויות:

כדי להסיק על הנחת שיוויון השונויות, נתבונן בתרשים פיזור השגיאות. כאשר ההנחה מתקיימת נצפה לראות פיזור וצפיפות אחידים. בתחילה נרצה לבדוק האם נוצרה צורה של מעין משפך. בתרשים פיזור השגיאות שנוצר, ניתן לראות שהנקודות אינו יוצרות מעין צורה של משפך (צורת משפך הכוונה היא רחב בצד שמאל והולך ונהיה צר ככל שזזים ימינה בגרף, כלומר הנתונים הולכים ונהיים צפופים יותר). בעקבות העובדה כי לא נוצרה צורת משפך לא ניתן להסיק ישר כי הנחת שיוויון השונויות מתקיימת.

לכן, נבצע **מבחן F ליחס שונויות** על השליש התחתון והשליש העליון של הרשומות.

הP-value שהתקבל הוא 0.0591. כיוון שהערך שהתקבל גבוה מ - 0.05 לא נדחה את השערת האפס. **כלומר הנחת שיוויון השונויות מתקיימת ברמת מובהקות של 5%.**

3.2.2 הנחת לינאריות:

כדי להסיק על הנחת הלינאריות, נתבונן בתרשים פיזור השגיאות. אם ההנחה אכן מתקיימת נצפה לראות פיזור סימטרי לאורך הגרף והתקזזות במרחקי השגיאות . מהתבוננות ראשונית בגרף, ניתן לראות כי השגיאות מפוזרות בצורה סימטרית סביב הציר האופקי לאורך הגרף ויש קיזוז במרחקי השגיאות, אך ישנם מקטעים קטנים בהם הפיזור סביב הציר האופקי הוא איננו אחיד (ניתן לראות אפילו שיש מקטע שבו כל הנקודות הן רק מעל לציר האופקי).

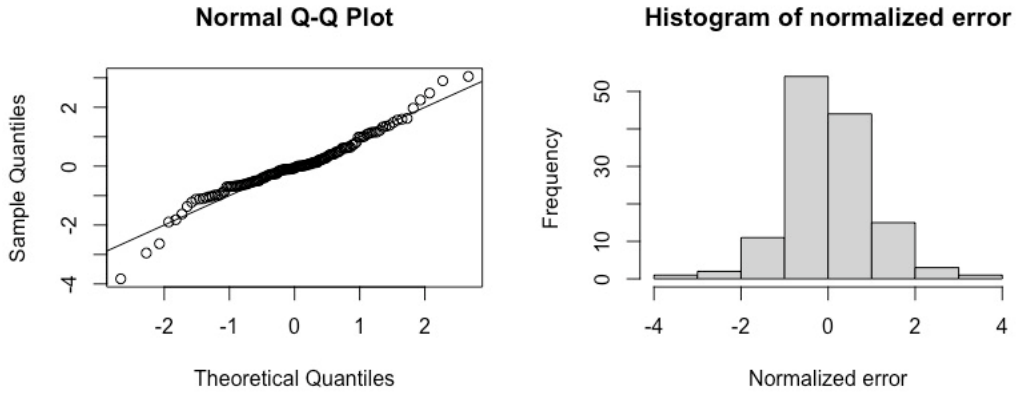
לכן, נבצע מבחן סטטיסטי- **מבחן Chow:**

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

        M-fluctuation test

data:  bcdAIC
f(efp) = 0.95859, p-value = 0.993
```

ערך ה P-value שהתקבל הוא 0.993. כיוון שהערך שהתקבל גבוה מ 0.05 לא נדחה את השערת האפס. **כלומר הנחת הלינאריות מתקיימת ברמת מובהקות של 5%.**

3.2.3 הנחת נורמאליות של השגיאות:

בכדי לבדוק הנחה זו נשתמש בשני הגרפים הבאים: היסטוגרמה של השגיאות ו- QQplot. מהתבוננות ראשונית בהיסטוגרמה ניתן לראות כי הגרף מזכיר גרף של התפלגות נורמלית.



בגרף ה- QQplot ניתן לראות כי רוב הנקודות נמצאות על הקו או בסמוך אליו (הנקודות שבקצוות מרוחקות יותר מהקו, כמצופה בהתאם למה שנלמד בהרצאה ובתרגול), אך ניתן לראות שישנן נקודות חריגות. לפיכך, על מנת לקבוע באופן חד משמעי לגבי הנורמליות של השגיאות, נבצע שני מבחנים סטטיסטיים על מנת לאשש / להפריך הנחה זו- מבחן **KS** ו- **SW**.

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

        One-sample Kolmogorov-Smirnov test

data:  data$stan_residuals
D = 0.096302, p-value = 0.176
alternative hypothesis: two-sided

> shapiro.test(data$stan_residuals)

        Shapiro-Wilk normality test

data:  data$stan_residuals
W = 0.95805, p-value = 0.0004692
```

ניתן לראות כי במבחן KS ה- D קרוב ל-0 ולפי מדד SW ה- W שואף ל-1.

נשים לב כי הP-value לפי מבחן KS גדול מ0.05 וזאת בניגוד למבחן SW. במקרים שבהן יש סתירה נלך לפי SW משום שהעוצמה שלו גדולה יותר ועל כן נדחה את השערת האפס ונאמר כי **השגיאות אינן מתפלגות נורמלית ברמת מובהקות של 5%.** (אך כשמסתכלים על הגרף נראה כי התצפיות לא מאוד רחוקות מהתפלגות זו).

3.3 שימוש במודל הנבחר:

למודל רגרסיה מרובה שתי מטרות עיקריות: האחת הינה חיזוי, ואילו השנייה הינה בדיקת קשר סיבתי בין משתנים מסבירים לבין משתנה מוסבר. המודל בו בחרנו בסעיף "בחירת משתני המודל" התקבל באלגוריתם רגרסיה לאחור לפי קריטריון AIC אשר נלמד בכיתה. ערך המדד R^2_{adj} אשר התקבל הינו 0.8737 וזהו ערך גבוה מאוד המרמז על טיב המודל. לפיכך, שימוש במודל זה **למטרות חיזוי** עשוי להניב תוצאות יחסית מדויקות השואפות לערך האמיתי והלא ידוע. דוגמא לשימוש במודל הנבחר היא חיזוי תוחלת חיים עבור מדינה שאינה נמצאת במדגם שלנו:

I4- צריכת אלכוהול	3
X2 – תמותה בקרב מבוגרים	56
X3 – תמותת תינוקות	72
X9 – מקרי מוות מתחת לגיל 5	89
X11 – אחוז הוצאה ממשלתית כללית לבריאות	4.32
X13 – מקרי מוות בקרב לידות נשאי איידס	1.8

```
> predict (ourModel,data.frame(I4='3',x2= 52, x3=72, x9= 89, x11=4.32 ,x13=1.8,x19=11.8), interval = "predict")
       fit      lwr      upr
1 72.96215 65.18411 80.74019
```

עפ"י המודל שלנו, הערך החזוי של תוחלת החיים הממוצעת במדינה זו הוא 72.96215 שנים. בנוסף, נשתמש ברשומה מבסיס הנתונים שבידינו על מנת להראות את דיוק המודל ולבדוק את נכונותו ואת טיבו. נבחר ברשומה מספר 4 בבסיס הנתונים:

I4- צריכת אלכוהול	1
X2 – תמותה בקרב מבוגרים	11
X3 – תמותת תינוקות	21
X9 – מקרי מוות מתחת לגיל 5	24
X11 – אחוז הוצאה ממשלתית כללית לבריאות	7.21
X13 – מקרי מוות בקרב לידות נשאי איידס	0.1
X19 – מספר שנות לימוד	14.4

```
> predict (ourModel,data.frame(I4='1', x2= 11, x3=21, x9= 24, x11=7.21 ,x13=0.1,  x19=14.4), interval = "predict")
       fit      lwr      upr
1 76.74331 69.33661 84.15001
```

הערך שהתקבל בחיזוי הינו 76.74331 והערך האמיתי של תוחלת החיים הממוצעת במדינה הנבחרת הינו 75.4 שנים, סטייה של 1.78%. בנוסף, הערך האמיתי נמצא ברווח הסמך בר"ב של 95%, דבר המעיד על חוזקו ורמת דיוקו של המודל.

4. שיפור המודל

לאחר בדיקת המודל, נבחן האם ביצוע טרנספורמציה על המשתנים במודל תשפר את תוצאותיו. כלומר, ברצוננו לשפר את מדד R^2_{adj} על מנת לקבל מודל מדויק יותר.

מכיוון שהראנו כי המודל מקיים את הנחות הלינאריות ושוויון השונויות אך לא מקיים את הנחת הנורמליות, נבצע בדיקה כללית ולא ספציפית.

נבצע את טרנספורמציות על X, הן פולינומיאליות והן מייצבות שונות; X^2 , $\ln(X)$, $\sqrt{x}$.

נבחן את יעילות הטרנספורמציות ע"י בדיקה האם חל שיפור במתאם פירסון.

מתאם פירסון עבור המודל הקיים:

	Life.expectancy	Adult.Mortality	infant.deaths	Alcohol	under.five.deaths	Total.expenditure	HIV.AIDS	Schooling
Life.expectancy	1.0000000	-0.7700689	-0.20094246	0.50909446	-0.22857019	0.31931099	-0.61710562	0.8017963
Adult.Mortality	-0.7700689	1.0000000	0.15428737	-0.26010165	0.17906676	-0.13952260	0.65087780	-0.5708780
infant.deaths	-0.2009425	0.1542874	1.00000000	-0.07654490	0.99538566	-0.15263474	0.06925466	-0.2111162
Alcohol	0.5090945	-0.2601017	-0.07654490	1.00000000	-0.08620968	0.29180703	-0.19975870	0.5725941
under.five.deaths	-0.2285702	0.1790668	0.99538566	-0.08620968	1.00000000	-0.15659779	0.09445192	-0.2289669
Total.expenditure	0.3193110	-0.1395226	-0.15263474	0.29180703	-0.15659779	1.00000000	-0.09642622	0.3020399
HIV.AIDS	-0.6171056	0.6508778	0.06925466	-0.19975870	0.09445192	-0.09642622	1.00000000	-0.3857052
Schooling	0.8017963	-0.5708780	-0.21111617	0.57259415	-0.22896687	0.30203992	-0.38570522	1.0000000

כעת נבצע את הטרנספורמציות ונבדוק את המתאם החדש –

עבור X^2:

	Life.expectancy	Adult.Mortality	infant.deaths	Alcohol	under.five.deaths	Total.expenditure	HIV.AIDS	Schooling
Life.expectancy	1.0000000	-0.74413670	-0.081691398	0.527403329	-0.10139594	0.34124594	-0.428270045	0.80853622
Adult.Mortality	-0.7441367	1.00000000	0.036098748	-0.246503358	0.05909179	-0.08027226	0.697130623	-0.50850449
infant.deaths	-0.0816914	0.03609875	1.000000000	-0.007455785	0.99224221	-0.09137273	-0.003006119	-0.08758148
Alcohol	0.5274033	-0.24650336	-0.007455785	1.000000000	-0.01577365	0.29703372	-0.126273526	0.58897130
under.five.deaths	-0.1013959	0.05909179	0.992242208	-0.015773646	1.00000000	-0.09791214	0.010713801	-0.09711582
Total.expenditure	0.3412459	-0.08027226	-0.091372728	0.297033718	-0.09791214	1.00000000	-0.042123336	0.29989140
HIV.AIDS	-0.4282700	0.69713062	-0.003006119	-0.126273526	0.01071380	-0.04212334	1.000000000	-0.23858268
Schooling	0.8085362	-0.50850449	-0.087581475	0.588971299	-0.09711582	0.29989140	-0.238582679	1.00000000

ניתן לראות כי הקורלציה של שני המשתנים הוצאה כוללת לבריאות ומספר שנות הלימוד השתפרה בעקבות הטרנספורמציה.

עבור $\ln(X)$:

	Life.expectancy	Adult.Mortality	infant.deaths	Alcohol	under.five.deaths	Total.expenditure	HIV.AIDS	Schooling
Life.expectancy	1.0000000	-0.51772715	NaN	0.4874188	NaN	0.25644910	-0.6373559	0.7695071
Adult.Mortality	-0.5177272	1.00000000	NaN	-0.1596811	NaN	-0.03063865	0.3813304	-0.4353514
infant.deaths	NaN	NaN	1	NaN	NaN	NaN	NaN	NaN
Alcohol	0.4874188	-0.15968106	NaN	1.0000000	NaN	0.23386447	-0.1997587	0.5382400
under.five.deaths	NaN	NaN	NaN	NaN	1	NaN	NaN	NaN
Total.expenditure	0.2564491	-0.03063865	NaN	0.2338645	NaN	1.00000000	-0.1113510	0.2560032
HIV.AIDS	-0.6373559	0.38133041	NaN	-0.1997587	NaN	-0.11135096	1.0000000	-0.3768640
Schooling	0.7695071	-0.43535143	NaN	0.5382400	NaN	0.25600323	-0.3768640	1.0000000

ניתן לראות כי טרנספורמציה זו אינה מתאימה לרוב המשתנים, מכיוון שבחלק מהמשתנים חלק מהתוצאות שוות ל-0, ו– $\ln(0)$ שואף למינוף אינסוף. עם זאת, הקורלציה של המשתנה תמותה מHIV השתפרה.

עבור $\sqrt{x}$:

	Life.expectancy	Adult.Mortality	infant.deaths	Alcohol	under.five.deaths	Total.expenditure	HIV.AIDS	Schooling
Life.expectancy	1.0000000	-0.6875220	-0.4215959	0.4986591	-0.4528224	0.2932248	-0.6278679	0.7893300
Adult.Mortality	-0.6875220	1.0000000	0.3204213	-0.2298079	0.3427069	-0.1144575	0.5335858	-0.5373411
infant.deaths	-0.4215959	0.3204213	1.0000000	-0.2060978	0.9967206	-0.1873403	0.1858543	-0.4222902
Alcohol	0.4986591	-0.2298079	-0.2060978	1.0000000	-0.2129619	0.2699513	-0.1997587	0.5578468
under.five.deaths	-0.4528224	0.3427069	0.9967206	-0.2129619	1.0000000	-0.1937082	0.2114190	-0.4492133
Total.expenditure	0.2932248	-0.1144575	-0.1873403	0.2699513	-0.1937082	1.0000000	-0.1061283	0.2856178
HIV.AIDS	-0.6278679	0.5335858	0.1858543	-0.1997587	0.2114190	-0.1061283	1.0000000	-0.3835237
Schooling	0.7893300	-0.5373411	-0.4222902	0.5578468	-0.4492133	0.2856178	-0.3835237	1.0000000

ניתן לראות כי טרנספורמציה זו שיפרה את הקורלציה של תמותת תינוקות.

כעת, נבצע רגרסיה למודל החדש:

y ~
I1*sqrt(x3)+I4*(x2)^2+I14*x16+(x2)^2+sqrt(x3)+x5+x6+x7+x8+x9+x10+(x11)^2+x12+log(x13)+x15+x16+x17+x18+(x19)^2

R^2_{adj}		
0.8978	**מודל מלא**	
0.8817	**AIC**	**רגרסיה לפנים**
0.8741	**BIC**	
0.8875	**AIC**	**רגרסיה לאחור**
0.8741	**BIC**	
0.8741	**AIC**	**רגרסיה בצעדים**
0.8741	**BIC**	

ניתן לראות כי התוצאה הטובה ביותר התקבלה לפי המודל המלא, והינה טובה יותר מלפני הוספת הטרנספורמציות. לכן המודל החדש הוא:

y ~
I1*sqrt(x3)+I4*(x2)^2+I14*x16+(x2)^2+sqrt(x3)+x5+x6+x7+x8+x9+x10+(x11)^2+x12+log(x13)+x15+x16+x17+x18+(x19)^2

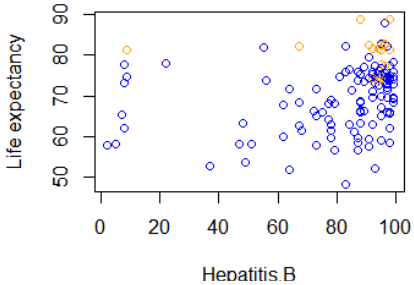
נספחים

2.1 המשך הסרת משתנים:

צהבת B:

מתאם פירסון: 0.252976

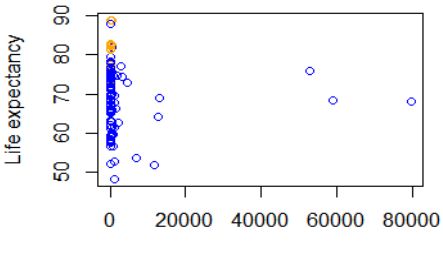
מקדם הקורלציה נמוך ונמצא ברבע התחתון. בנוסף ניתן לראות כי התצפיות מפוזרות ולא נמצאות סביב קו מגמה ברור.



שחפת:

מתאם פירסון:-0.0487

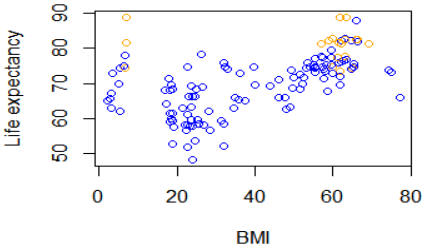
ניתן לראות כי מקדם הקורלציה נמוך מאוד והקשר בין שחפת לבין תוחלת החיים כמעט ולא קיים.



BMI:

מתאם פירסון: 0.5552265

ניתן לראות כי קיים קשר חזק יחסית בים ה-BMI לבין תוחלת החיים, כאשר מקדם הקורלציה יחסית גבוהה.

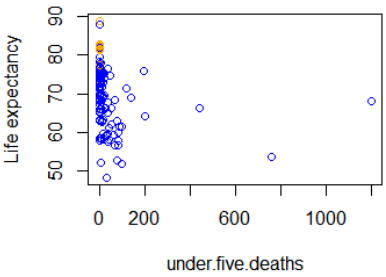


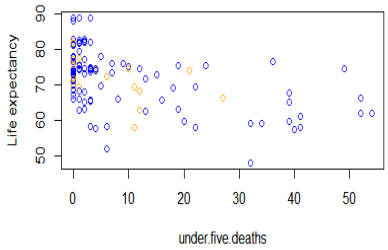
תמותה מתחת לגיל 5:

מתאם פירסון:-0.2285702

מתאם פירסון לאחר הוצאת חריגים: -0.49925

מקדם המתאם נמוך, דבר המעיד על קשר חלש בין משתנה מסביר זה לבין תוחלת החיים. כמו כן, ניתן לראות בתרשים הפיזור כי מרבית התצפיות מרוכזות סביב ה-0 וכי אין מגמה ברורה. עם זאת, לאחר הסרת התצפיות החריגות, הקשר בין השניים התחזק ונצפה קו מגמה ברור יותר.

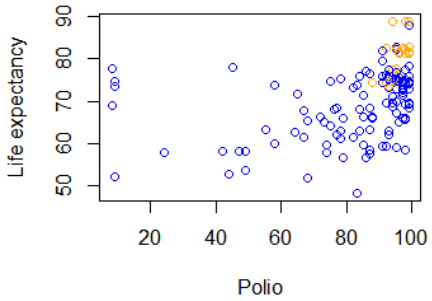




פוליו:

מתאם פירסון: 0.3851558

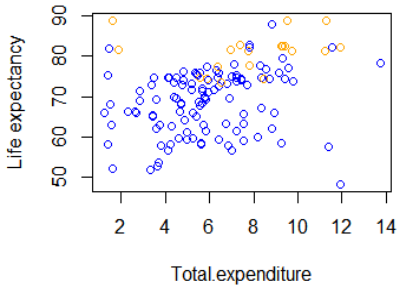
ניתן לראות כי מקדם המתאם ממוצע, עם זאת אין קו מגמה ברור.



הוצאה ממשלתית כללית לבריאות כאחוז מסך ההוצאות הממשלתיות:

מתאם פירסון: 0.319311

המתאם יחסית נמוך, אך ניתן לראות כי ישנו פיזור גבוה וניתן להבין את המגמתיות – ככל שעולה ההשקעה הממשלתית בבריאות, כך עולה גם תוחלת החיים.

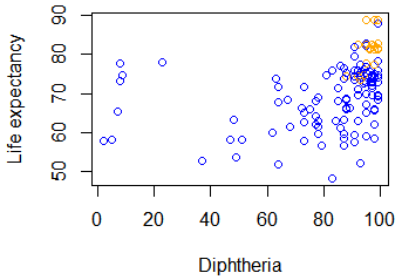


כיסוי החיסונים נגד טטנוס ושעלת בקרבי ילדים עד גיל שנה:

מתאם פירסון: 0.3424636

מקדם פירסון יחסית נמוך, אך ניתן לראות קו מגמה ברור יחסית – ככל שכיסוי החיסונים במידה מסוימת עולה, כך גם עולה תוחלת החיים.

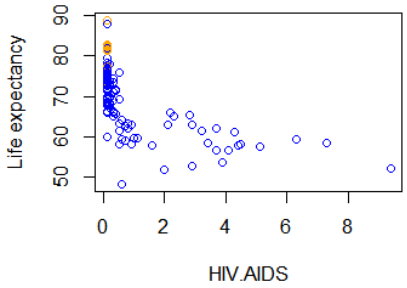
אנו מעריכים כי הקשר הוא מדגמי, והסיבה למגמה הנראית היא שמדינות מפותחת ועשירות יותר מחסנות יותר, ומכאן נובע הקשר.



איידס:

מתאם פירסון:-0.6171

קיים קשר שלילי חזק בין תחלואה באיידס לבין תוחלת החיים.



תוצר לאומי גולמי:

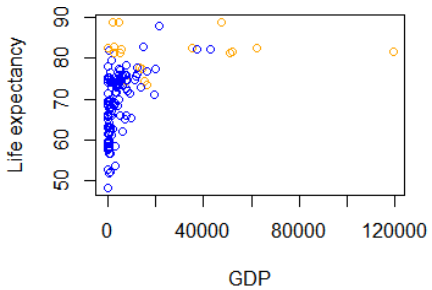
מתאם פירסון: 0.44377

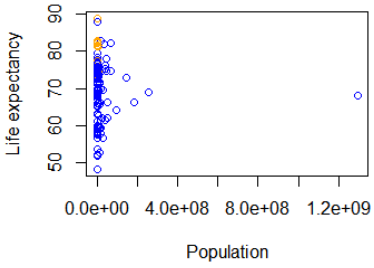
מתאם פירסון לאחר חריגים: 0.437339

קיים קשר יחסית חזק בית התוצר הלאומי הגולמי לבין תוחלת החיים.

בנוסף, מקדם פירסון כמעט ולא השתנה לאחר הסרת החריגים

לכן, נבחר להשאיר משתנה זה

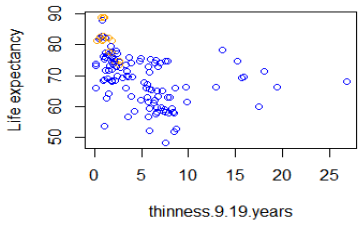




אוכלוסייה:

מתאם פירסון:-0.03594

מקדם המתאם נמוך מאוד ונראה כי הקשר בין גודל האוכלוסייה במדינה לבין תוחלת החיים כמעט ולא קיים.



שכיחות הרזון בקרב ילדים ובני נוער בגילאי 10-19:

מתאם פירסון: -0.4369442

קיים קשר שלילי חזק יחסית בין רזון לבין תוחלת החיים, כאשר המתאם קרוב לממוצע.

3.3 שימוש במודל הנבחר:

מודל מלא:

```
> Full<-lm (y ~ I1*x3+I4*x2+I14*x16+x2+x3+x5+x6+x7+x8+x9+x10+x11+x12+x13+x15+x16+x17+x18+x19,data)
> summary(Full)

Call:
lm(formula = y ~ I1 * x3 + I4 * x2 + I14 * x16 + x2 + x3 + x5 +
    x6 + x7 + x8 + x9 + x10 + x11 + x12 + x13 + x15 + x16 + x17 +
    x18 + x19, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-11.1377  -1.5334  -0.1189   1.5080   7.5117

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.991e+01  3.566e+00  13.997  < 2e-16 ***
I12           -1.299e+00  1.436e+00  -0.904  0.36796
x3            -7.624e-01  1.076e+00  -0.708  0.48026
I42            1.143e+00  2.279e+00   0.502  0.61697
I43            2.227e+00  1.549e+00   1.438  0.15353
I44            2.975e+00  2.429e+00   1.225  0.22354
x2            -1.368e-02  4.462e-03  -3.065  0.00278 **
I14medium      1.259e+00  1.695e+00   0.743  0.45944
I14high        3.066e+00  4.106e+00   0.747  0.45703
x16           -8.362e-02  2.302e-01  -0.363  0.71714
x5            -2.017e-04  3.095e-04  -0.652  0.51612
x6             1.124e-02  3.090e-02   0.364  0.71681
x7            -6.842e-05  5.317e-05  -1.287  0.20107
x8            -1.304e-02  2.089e-02  -0.624  0.53376
x9            -5.011e-02  3.890e-02  -1.288  0.20063
x10            2.555e-03  2.189e-02   0.117  0.90730
x11            2.710e-01  1.281e-01   2.116  0.03681 *
x12            4.836e-03  3.632e-02   0.133  0.89432
x13           -8.016e-01  2.512e-01  -3.191  0.00189 **
x15            3.999e-10  7.092e-09   0.056  0.95514
x17            7.099e-03  2.281e-01   0.031  0.97523
x18            3.861e+01  6.628e+00   5.825 6.68e-08 ***
x19           -2.901e-01  2.919e-01  -0.994  0.32273
I12:x3         8.292e-01  1.073e+00   0.773  0.44157
I42:x2         8.420e-03  1.882e-02   0.447  0.65555
I43:x2        -1.406e-02  7.869e-03  -1.786  0.07704 .
I44:x2        -2.196e-02  1.669e-02  -1.316  0.19111
I14medium:x16 -5.119e-01  4.665e-01  -1.097  0.27506
I14high:x16   -1.167e+00  3.220e+00  -0.363  0.71770
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.182 on 102 degrees of freedom
Multiple R-squared:  0.8927,    Adjusted R-squared:  0.8633
F-statistic: 30.32 on 28 and 102 DF,  p-value: < 2.2e-16
```

רגרסיה לפנים AIC:

```
Step:  AIC=301.47
y ~ x18 + x2 + x13 + x11

       Df Sum of Sq    RSS    AIC
<none>              1212.2 301.47
+ I1    1   14.0983 1198.1 301.94
+ x12   1   10.8688 1201.3 302.29
+ x6    1    9.6476 1202.5 302.43
+ x5    1    7.0891 1205.1 302.70
+ x17   1    5.4352 1206.7 302.88
+ x9    1    3.6605 1208.5 303.08
+ x16   1    2.5798 1209.6 303.19
+ x3    1    2.2083 1210.0 303.23
+ x19   1    1.2706 1210.9 303.33
+ x7    1    0.8147 1211.4 303.38
+ x10   1    0.4995 1211.7 303.42
+ x8    1    0.2329 1211.9 303.45
+ x15   1    0.0156 1212.2 303.47
+ I14   2   13.6323 1198.5 303.99
+ I4    3   21.4144 1190.8 305.14
```

```
> summary(fwdAIC)

Call:
lm(formula = y ~ x18 + x2 + x13 + x11, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.3808  -1.6174  -0.0501   1.6143   9.9760 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 47.614113   2.047692  23.253  < 2e-16 ***
x18         36.285086   2.488491  14.581  < 2e-16 ***
x2          -0.017949   0.003855  -4.656 8.04e-06 ***
x13         -0.844894   0.230049  -3.673 0.000353 ***
x11          0.355162   0.111458   3.187 0.001816 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.102 on 126 degrees of freedom
Multiple R-squared:  0.8741,	Adjusted R-squared:  0.8701 
F-statistic: 218.7 on 4 and 126 DF,  p-value: < 2.2e-16
```

רגרסיה לאחור AIC:

```
Step:  AIC=305.13
y ~ x3 + I4 + x2 + x9 + x11 + x13 + x18 + I4:x2

        Df Sum of Sq    RSS    AIC
<none>               1103.2 305.13
- x3     1     18.26 1121.5 305.28
- x9     1     21.12 1124.3 305.61
- I4:x2  3     63.87 1167.1 306.50
- x11    1     61.59 1164.8 310.25
- x13    1    108.95 1212.2 315.47
- x18    1   1262.05 2365.2 403.04
```

```
> summary(bcdAIC)

Call:
lm(formula = y ~ x3 + I4 + x2 + x9 + x11 + x13 + x18 + I4:x2, 
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.1498  -1.6134  -0.2004   1.4086   8.8777 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 48.271954   2.340982  20.620  < 2e-16 ***
x3           0.042073   0.030106   1.398 0.164884    
I42          1.661177   2.046457   0.812 0.418579    
I43          2.283957   1.403841   1.627 0.106419    
I44          3.513755   1.751345   2.006 0.047109 *  
x2          -0.014370   0.004172  -3.444 0.000793 ***
x9          -0.034397   0.022887  -1.503 0.135534    
x11          0.291814   0.113697   2.567 0.011521 *  
x13         -0.800062   0.234362  -3.414 0.000879 ***
x18         34.674897   2.984435  11.619  < 2e-16 ***
I42:x2      -0.001273   0.016036  -0.079 0.936845    
I43:x2      -0.015267   0.007300  -2.091 0.038652 *  
I44:x2      -0.023442   0.013695  -1.712 0.089575 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.058 on 118 degrees of freedom
Multiple R-squared:  0.8854,	Adjusted R-squared:  0.8737 
F-statistic: 75.97 on 12 and 118 DF,  p-value: < 2.2e-16
```

רגרסיה בצעדים AIC:

```
Step:  AIC=301.47
y ~ x18 + x2 + x13 + x11

       Df Sum of Sq    RSS    AIC
<none>               1212.2 301.47
+ I1    1     14.10 1198.1 301.94
+ x12   1     10.87 1201.3 302.29
+ x6    1      9.65 1202.5 302.43
+ x5    1      7.09 1205.1 302.70
+ x17   1      5.44 1206.7 302.88
+ x9    1      3.66 1208.5 303.08
+ x16   1      2.58 1209.6 303.19
+ x3    1      2.21 1210.0 303.23
+ x19   1      1.27 1210.9 303.33
+ x7    1      0.81 1211.4 303.38
+ x10   1      0.50 1211.7 303.42
+ x8    1      0.23 1211.9 303.45
+ x15   1      0.02 1212.2 303.47
+ I14   2     13.63 1198.5 303.99
+ I4    3     21.41 1190.8 305.14
- x11   1     97.68 1309.9 309.62
- x13   1    129.77 1341.9 312.79
- x2    1    208.58 1420.8 320.27
- x18   1   2045.41 3257.6 428.97
```

```
> summary(swAIC)

Call:
lm(formula = y ~ x18 + x2 + x13 + x11, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.3808  -1.6174  -0.0501   1.6143   9.9760 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 47.614113   2.047692  23.253  < 2e-16 ***
x18         36.285086   2.488491  14.581  < 2e-16 ***
x2          -0.017949   0.003855  -4.656 8.04e-06 ***
x13         -0.844894   0.230049  -3.673 0.000353 ***
x11          0.355162   0.111458   3.187 0.001816 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.102 on 126 degrees of freedom
Multiple R-squared:  0.8741,	Adjusted R-squared:  0.8701 
F-statistic: 218.7 on 4 and 126 DF,  p-value: < 2.2e-16
```

רגרסיה לפנים BIC

```
Step:  AIC=315.85
y ~ x18 + x2 + x13 + x11

       Df Sum of Sq    RSS    AIC
<none>               1212.2 315.85
+ I1    1   14.0983 1198.1 319.19
+ x12   1   10.8688 1201.3 319.54
+ x6    1    9.6476 1202.5 319.68
+ x5    1    7.0891 1205.1 319.95
+ x17   1    5.4352 1206.7 320.13
+ x9    1    3.6605 1208.5 320.33
+ x16   1    2.5798 1209.6 320.44
+ x3    1    2.2083 1210.0 320.48
+ x19   1    1.2706 1210.9 320.59
+ x7    1    0.8147 1211.4 320.63
+ x10   1    0.4995 1211.7 320.67
+ x8    1    0.2329 1211.9 320.70
+ x15   1    0.0156 1212.2 320.72
+ I14   2   13.6323 1198.5 324.12
+ I4    3   21.4144 1190.8 328.14
```

```
> summary(fwdBIC)

Call:
lm(formula = y ~ x18 + x2 + x13 + x11, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.3808  -1.6174  -0.0501   1.6143   9.9760 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 47.614113   2.047692  23.253  < 2e-16 ***
x18         36.285086   2.488491  14.581  < 2e-16 ***
x2          -0.017949   0.003855  -4.656 8.04e-06 ***
x13         -0.844894   0.230049  -3.673 0.000353 ***
x11          0.355162   0.111458   3.187 0.001816 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.102 on 126 degrees of freedom
Multiple R-squared:  0.8741,	Adjusted R-squared:  0.8701 
F-statistic: 218.7 on 4 and 126 DF,  p-value: < 2.2e-16
```

רגרסיה לאחור BIC

```
Step:  AIC=315.85
y ~ x2 + x11 + x13 + x18

       Df Sum of Sq    RSS    AIC
<none>              1212.2 315.85
- x11   1     97.68 1309.9 321.13
- x13   1    129.77 1341.9 324.30
- x2    1    208.58 1420.8 331.77
- x18   1   2045.41 3257.6 440.48
>
```

```
> summary(bcdBIC)

Call:
lm(formula = y ~ x2 + x11 + x13 + x18, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.3808  -1.6174  -0.0501   1.6143   9.9760 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 47.614113   2.047692  23.253  < 2e-16 ***
x2          -0.017949   0.003855  -4.656 8.04e-06 ***
x11          0.355162   0.111458   3.187 0.001816 ** 
x13         -0.844894   0.230049  -3.673 0.000353 ***
x18         36.285086   2.488491  14.581  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.102 on 126 degrees of freedom
Multiple R-squared:  0.8741,	Adjusted R-squared:  0.8701 
F-statistic: 218.7 on 4 and 126 DF,  p-value: < 2.2e-16
```

רגרסיה בצעדים BIC

```
Step:  AIC=315.85
y ~ x18 + x2 + x13 + x11

       Df Sum of Sq    RSS    AIC
<none>              1212.2 315.85
+ I1    1     14.10 1198.1 319.19
+ x12   1     10.87 1201.3 319.54
+ x6    1      9.65 1202.5 319.68
+ x5    1      7.09 1205.1 319.95
+ x17   1      5.44 1206.7 320.13
+ x9    1      3.66 1208.5 320.33
+ x16   1      2.58 1209.6 320.44
+ x3    1      2.21 1210.0 320.48
+ x19   1      1.27 1210.9 320.59
+ x7    1      0.81 1211.4 320.63
+ x10   1      0.50 1211.7 320.67
+ x8    1      0.23 1211.9 320.70
+ x15   1      0.02 1212.2 320.72
- x11   1     97.68 1309.9 321.13
+ I14   2     13.63 1198.5 324.12
- x13   1    129.77 1341.9 324.30
+ I4    3     21.41 1190.8 328.14
- x2    1    208.58 1420.8 331.77
- x18   1   2045.41 3257.6 440.48
```

```
> summary(swBIC)

Call:
lm(formula = y ~ x18 + x2 + x13 + x11, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.3808  -1.6174  -0.0501   1.6143   9.9760 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 47.614113   2.047692  23.253  < 2e-16 ***
x18         36.285086   2.488491  14.581  < 2e-16 ***
x2          -0.017949   0.003855  -4.656 8.04e-06 ***
x13         -0.844894   0.230049  -3.673 0.000353 ***
x11          0.355162   0.111458   3.187 0.001816 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.102 on 126 degrees of freedom
Multiple R-squared:  0.8741,	Adjusted R-squared:  0.8701 
F-statistic: 218.7 on 4 and 126 DF,  p-value: < 2.2e-16
```

סעיף 4

הרצת המודל המלא:

```
Call:
lm(formula = y ~ I1 * sqrt(x3) + I4 * (x2)^2 + I14 * x16 + (x2)^2 +
    sqrt(x3) + x5 + x6 + x7 + x8 + x9 + x10 + (x11)^2 + x12 +
    log(x13) + x15 + x16 + x17 + x18 + (x19)^2, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-10.8203  -1.3719  -0.1022   1.5060   7.6498

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.069e+01  3.600e+00  14.078  < 2e-16 ***
I12           -1.237e+00  1.467e+00  -0.843  0.40096
sqrt(x3)      -9.714e-01  1.421e+00  -0.684  0.49579
I42            9.815e-01  2.211e+00   0.444  0.65800
I43            2.062e+00  1.511e+00   1.365  0.17528
I44            3.278e+00  2.365e+00   1.386  0.16875
x2            -1.293e-02  4.281e-03  -3.020  0.00319 **
I14medium      1.223e+00  1.658e+00   0.737  0.46256
I14high        3.435e+00  4.086e+00   0.841  0.40250
x16            7.718e-02  2.206e-01   0.350  0.72717
x5            -2.300e-04  3.048e-04  -0.755  0.45224
x6             1.925e-03  3.080e-02   0.062  0.95029
x7            -5.113e-05  4.980e-05  -1.027  0.30698
x8            -2.485e-02  2.101e-02  -1.183  0.23970
x9            -2.246e-03  7.295e-03  -0.308  0.75883
x10            8.926e-03  2.164e-02   0.412  0.68085
x11            2.450e-01  1.250e-01   1.960  0.05274 .
x12            6.410e-03  3.626e-02   0.177  0.86001
log(x13)      -1.488e+00  3.564e-01  -4.175  6.3e-05 ***
x15            5.340e-09  5.481e-09   0.974  0.33230
x17           -1.612e-01  2.273e-01  -0.709  0.47971
x18            3.488e+01  6.606e+00   5.281  7.3e-07 ***
x19           -3.096e-01  2.837e-01  -1.091  0.27764

I12:sqrt(x3)   9.203e-01  1.435e+00   0.641  0.52269
I42:x2         1.057e-02  1.824e-02   0.579  0.56360
I43:x2        -1.233e-02  7.726e-03  -1.596  0.11359
I44:x2        -2.488e-02  1.611e-02  -1.544  0.12561
I14medium:x16 -4.886e-01  4.552e-01  -1.074  0.28557
I14high:x16   -1.121e+00  3.174e+00  -0.353  0.72481
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.106 on 102 degrees of freedom
Multiple R-squared:  0.8978,    Adjusted R-squared:  0.8697
F-statistic: 31.99 on 28 and 102 DF,  p-value: < 2.2e-16
```

לפי AIC:

לפנים:

```
Call:
lm(formula = y ~ x18 + x2 + log(x13) + x11 + I1, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.2319  -1.4534  -0.0191   1.5701   9.3050 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 50.26774    2.62599  19.142  < 2e-16 ***
x18         30.26170    2.93953  10.295  < 2e-16 ***
x2          -0.01761    0.00365  -4.825 4.00e-06 ***
log(x13)    -1.43606    0.31135  -4.612 9.71e-06 ***
x11          0.33083    0.10985   3.012  0.00315 ** 
I12         -1.34440    0.90830  -1.480  0.14136    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.018 on 125 degrees of freedom
Multiple R-squared:  0.8817,	Adjusted R-squared:  0.877 
F-statistic: 186.3 on 5 and 125 DF,  p-value: < 2.2e-16
```

לאחור:

```
Call:
lm(formula = y ~ I4 + x2 + x11 + log(x13) + x18 + I4:x2, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.6624  -1.2351  -0.2573   1.4720   8.9995 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 48.240713   2.265043  21.298  < 2e-16 ***
I42          1.471146   1.986924   0.740  0.46050    
I43          2.238748   1.375905   1.627  0.10634    
I44          3.727629   1.725376   2.160  0.03272 *  
x2          -0.014962   0.003940  -3.798  0.00023 ***
x11          0.305582   0.110595   2.763  0.00663 ** 
log(x13)    -1.388903   0.320673  -4.331  3.1e-05 ***
x18         30.758529   3.120885   9.856  < 2e-16 ***
I42:x2      -0.001291   0.014960  -0.086  0.93138    
I43:x2      -0.013222   0.007173  -1.843  0.06777 .  
I44:x2      -0.024865   0.013488  -1.844  0.06772 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.005 on 120 degrees of freedom
Multiple R-squared:  0.8875,	Adjusted R-squared:  0.8781 
F-statistic: 94.62 on 10 and 120 DF,  p-value: < 2.2e-16
```

בצעדים:

```
Call:
lm(formula = y ~ x18 + x2 + x13 + x11, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.3808  -1.6174  -0.0501   1.6143   9.9760 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 47.614113   2.047692  23.253  < 2e-16 ***
x18         36.285086   2.488491  14.581  < 2e-16 ***
x2          -0.017949   0.003855  -4.656 8.04e-06 ***
x13         -0.844894   0.230049  -3.673 0.000353 ***
x11          0.355162   0.111458   3.187 0.001816 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.102 on 126 degrees of freedom
Multiple R-squared:  0.8741,	Adjusted R-squared:  0.8701 
F-statistic: 218.7 on 4 and 126 DF,  p-value: < 2.2e-16
```

<u>לפי BIC</u>

לפנים:

```
Call:
lm(formula = y ~ x18 + x2 + x13 + x11, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.3808  -1.6174  -0.0501   1.6143   9.9760 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 47.614113   2.047692  23.253  < 2e-16 ***
x18         36.285086   2.488491  14.581  < 2e-16 ***
x2          -0.017949   0.003855  -4.656 8.04e-06 ***
x13         -0.844894   0.230049  -3.673 0.000353 ***
x11          0.355162   0.111458   3.187 0.001816 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.102 on 126 degrees of freedom
Multiple R-squared:  0.8741,	Adjusted R-squared:  0.8701 
F-statistic: 218.7 on 4 and 126 DF,  p-value: < 2.2e-16
```

לאחור:

```
Call:
lm(formula = y ~ x2 + x11 + x13 + x18, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.3808  -1.6174  -0.0501   1.6143   9.9760 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 47.614113   2.047692  23.253  < 2e-16 ***
x2          -0.017949   0.003855  -4.656 8.04e-06 ***
x11          0.355162   0.111458   3.187 0.001816 ** 
x13         -0.844894   0.230049  -3.673 0.000353 ***
x18         36.285086   2.488491  14.581  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.102 on 126 degrees of freedom
Multiple R-squared:  0.8741,	Adjusted R-squared:  0.8701 
F-statistic: 218.7 on 4 and 126 DF,  p-value: < 2.2e-16
```

בצעדים:

```
Call:
lm(formula = y ~ x18 + x2 + x13 + x11, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.3808  -1.6174  -0.0501   1.6143   9.9760 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 47.614113   2.047692  23.253  < 2e-16 ***
x18         36.285086   2.488491  14.581  < 2e-16 ***
x2          -0.017949   0.003855  -4.656 8.04e-06 ***
x13         -0.844894   0.230049  -3.673 0.000353 ***
x11          0.355162   0.111458   3.187 0.001816 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.102 on 126 degrees of freedom
Multiple R-squared:  0.8741,	Adjusted R-squared:  0.8701 
F-statistic: 218.7 on 4 and 126 DF,  p-value: < 2.2e-16
```