פרויקט ברגרסיה ליניארית

פרטי המגישות:

תום דמארי

מיה יערי

תוכן עניינים

תוכן

תקציר מנהלים:

בפרויקט זה ניתחנו נתונים בעזרת תוכנת ה-RStudio במטרה להעריך את תוחלת החיים במדינות שונות ואת הגורמים השונים המשפיעים עליה. בעזרת התוכנה בנינו מודל רגרסיה הכולל את המשתנה המוסבר, תוחלת החיים, ותשעה משתנים מסבירים שונים.
שבעה מתוך המשתנים המסבירים הינם משתנים רציפים: מדד זיהום האוויר, מספר החולים ב-HIV, מספר החולים במלריה, הכנסה שנתית ממוצעת לאדם בדולרים, צריכת אלכוהול שנתית לאדם בליטר, צפיפות אוכלוסין ליחידת שטח ואחוז צרכני הסיגריות, בעוד שהנותרים הם משתנים קטגוריאליים: מספר היבשת ושייכות לארגון ה-OECD.

בחלק הראשון של העבודה ניתחנו את הנתונים ומצאנו: ממוצע, סטיית תקן, חציון ותחום בין רבעוני. השתמשנו ב-BOX PLOT , בגרפים, בהיסטוגרמות ובטבלאות שכיחות. עשינו זאת על מנת להבין קשרים בין משתנים מסבירים שונים ובין כל משתנה מסביר למוסבר, כדי לזהות נתונים חריגים ולמפות אותם במידת הצורך, וכדי לזהות מגמות עיקריות בנתונים.

בחלק השני של העבודה חישבנו את מקדם פירסון ואת רמת המובהקות של כל אחד מהמשתנים המסבירים הרציפים במודל. כך בחנו את השפעת כל אחד מהמשתנים המסבירים על המשתנה המוסבר ובחרנו אילו משתנים נסיר מהמודל. אנו בחרנו להסיר מהמודל את המשתנים של צריכת אלכוהול לאדם בליטר וצפיפות אוכלוסין ליחידת שטח, אשר מקדם המתאם שלהם היה יחסית נמוך ורמת המובהקות שלהם יחסית גבוהה.
לאחר מכן, ביצענו התאמת משתנים ואיחוד קטגוריות לטובת שיפור המודל. בחרנו לאחד את יבשת אירופה עם יבשת אמריקה הצפונית לכדי קטגוריה חדשה של יבשות מפותחות.
בשלב הבא הגדרנו משתני דמה עבור כל אחד מהמשתנים הקטגוריאליים ובחנו הוספה של משתני אינטראקציה עבור משתנים שהפילוג שלהם לקטגוריות שונות משפיע על הקשר בין המסביר למוסבר. לאחר מכן, בנינו את המודל ובצענו רגרסיה בצעדים על ידי שלוש שיטות שונות: Forward Selection, Backward Elimination, Stepwise Regression.
בחנו את התוצאות בעזרת המדדים של AIC ו-BIC וקיבלנו כי ערך המדדים הנמוך ביותר שהתקבל מצביע על אותו מודל כמו המדד שעל פיו ביצענו את ההשוואה בין המודלים, מדד ה-R square Adjusted. קיבלנו כי כלל המדדים והאלגוריתמים השונים הצביעו על אותו המודל. על אותו מודל בדקנו את קיומן של שלושת ההנחות: הנחת הנורמאליות של השגיאות, הנחת שוויון שונויות, והנחת הלינאריות. מצאנו כי כל ההנחות מתקיימות.
בשלב האחרון למרות שכל ההנחות התקיימו עבור המודל שבחרנו, רצינו לבחון האם ישנה אפשרות לשפר עוד יותר את המודל ולכן בחרנו לנסות את השימוש בטרנספורמציה של BoxCox על משתנה המוסבר. השווינו על פי מדד ה-R square Adjusted ועל פיו התקבל כי המודל הטוב יותר הינו המודל שהתקבל לאחר ביצוע הטרנספורמציה על Y.

טבלת המשתנים מחלק א'

סוג המשתנה - מוסבר/מסביר	סימון במודל	יחידת מידה	סוג המשתנה – רציף / קטגוריאלי	הסבר קצר על המשתנה
משתנה מוסבר	Y	שנים	קטגוריאלי	תוחלת החיים במדינה

משתנה מסביר	X_1	%	רציף	מדד זיהום האוויר במדינה
	X_2	מספר	רציף	מספר האנשים החולים בHIV
	X_3	מספר	רציף	מספר האנשים החולים במלריה
	X_4	$	רציף	הכנסה שנתית ממוצעת לאדם בדולרים
	X_5	ליטר לאדם פר שנה	רציף	צריכת אלכוהול שנתית לאדם בליטר
	X_6	km^2	רציף	צפיפות אוכלוסין ליחידת שטח
	X_7	%	רציף	אחוז צרכני הסיגריות
	X_8	-	קטגוריאלי	יבשת
	X_9	-	קטגוריאלי	שייכות לארגון הOECD

עיבוד מקדים:

הסרה של משתנים:

במטרה לקבוע איזה מהמשתנים המסבירים הרציפים כדאי להסיר מהמודל נרצה לבחון את המתאם שלהם מול המשתנה המוסבר. לשם כך, נשתמש במקדם מתאם של פירסון (נספח 1) שהינו מדד לקשר ליניארי בין שני משתנים כמותיים שערכיהם מתקבלים במדגם. עבור כל קשר בין משתנה מסביר למוסבר נסתכל על מקדם המתאם ורמת מובהקות. נבחר להסיר משתנים בעלי רמת מובהקות שגבוהה מ0.05 ומקדם מתאם נמוך מאוד (ערכו של מקדם המתאם קרוב לאפס מעיד על קשר ליניארי חלש).

שם משתנה	Outdoor Air Pollution	HIV	Malaria	Average Income	Alcohol Consumption	Density	Cigarette Consumption
מקדם מתאם	-0.2074854	-0.213534	-0.3420889	0.5805813	**0.01830823**	**0.08771903**	0.163533
רמת מובהקות	0.03735	0.03203	0.0004628	1.966e-10	**0.8565**	**0.3831**	0.1022

עפ"י הטבלה ניתן לראות כי עבור מרבית המשתנים התקבל מקדם מתאם גבוהה יחסית ולכן נבחר לשמור אותם במודל שלנו כי מדד זה מעיד על קשר לינארי חזק יחסית בין המשתנה המסביר למשתנה המוסבר. בנוסף, ניתן לראות כי עבור מרבית המשתנים התקבל ערך P value שקטן מ-0.05. כלומר, מדובר בקשר מובהק בו המשתנה המסביר מצליח להסביר את המוסבר בצורה טובה ולכן גם במצב כזה נבחר להשאיר את אותם משתנים במודל שלנו.

בהתאם לטבלה המוצגת לעיל ועל פי הקריטריונים שהוסברו קודם לכן, מצאנו שני משתנים פוטנציאליים להסרה מהמודל: Alcohol Consumption ו-Density. כפי שניתן לראות המשתנה Alcohol Consumption הינו בעל רמת מובהקות גדולה מאוד וגם רמת השפעה שלו היא נמוכה ביותר ביחס למשתנים האחרים ולכן נבחר להסירו. כמו כן, מאותה סיבה בדיוק נבחר להסיר את המשתנה Density מכיוון שלא רק שמידת ההשפעה שלו נמוכה יחסית, אלא גם רמת המובהקות שלו גבוהה באופן יחסי לשאר המשתנים במודל.

ניתוח המשתנים הנותרים:

Outdoor Air Pollution - קיים קשר לינארי שלילי בין משתנה זה למשתנה המוסבר תוחלת חיי האדם וזאת מכיוון שבמדינות בהן אחוז הזיהום עולה ישנה השפעה על רמת התחלואה במדינה ולכן הגיוני כי תוחלת חיי האדם במדינה תרד. רמת ההשפעה היא אומנם נמוכה אולם רמת המובהקות שלו נמוכה מ0.05 ולכן נרצה להשאירו.

HIV - קיים קשר לינארי שלילי בין משתנה זה למשתנה המוסבר תוחלת חיי האדם וזאת מכיוון שבמדינות בהן מספר החולים בHIV גדל ישנה השפעה שלילית על רמת התחלואה במדינה עקב התפשטות המחלה ולכן הגיוני כי תוחלת חיי האדם במדינה תרד כי מקדם ההדבקה גדל. רמת ההשפעה היא אומנם נמוכה אולם רמת המובהקות שלו נמוכה מ0.05 ולכן נרצה להשאירו.

Malaria - קיים קשר לינארי שלילי בין משתנה זה למשתנה המוסבר תוחלת חיי האדם. הסיבה לכך היא שבמדינות בהן מספר החולים במלריה גדל ישנה השפעה שלילית על רמת התחלואה במדינה עקב התפשטות המחלה ולכן הגיוני כי תוחלת חיי האדם במדינה תרד כי מקדם ההדבקה גדל. מספר החולים במלריה מקטין את תוחלת חיי האדם במדינה בצורה גבוהה יחסית לשאר המשתנים במודל על פי מדד מקדם המתאם ורמת המובהקות שלו נמוכה מ0.05 ולכן נרצה להשאירו.

Average Income - ניתן לראות כי קיים קשר ליניארי חיובי בין המשתנים וזאת מכיוון שההכנסה ממוצעת גבוה יותר לאדם תורמת לאיכות החיים שלו וכיוצא בכך מאריכה את תוחלת חיי האדם במדינה. זאת כי מדינה שההכנסה בה גדולה יותר מעידה על מדינה מפותחת יותר ולכן

התשתיות והשירותים בה ברמה גבוהה יותר וכך גם תוחלת חיי האדם בה גדלה .הכנסה ממוצעת שנתית לאדם מאריכה את תוחלת חיי האדם בצורה גבוהה ביחס לשאר המשתנים במודל על פי מדד מקדם המתאם ובעל רמת מובהקות גבוהה מאוד ולכן נרצה להשאירו .

Cigarette Consumption - קיים קשר ליניארי חיובי בין אחוז המעשנים במדינה לתוחלת חיי האדם בה. נסיק מכך כי נתוני התחלואה מושפעים מאחוז צריכת הסיגריות של כל אדם ולא מאחוז המעשנים במדינה כולה. מהנתונים ניתן להניח כי עישון כיום היא תופעה רווחת. רמת המובהקות של משתנה זה אמנם גדולה מ-0.05, אך גם במקרה זה בחרנו להשאיר את המשתנה כיוון שבאופן יחסי רמת המובהקות שלו לא קיצונית.

<u>התאמת משתנים:</u>

<u>איחוד קטגוריות</u>

לאחר שעבדנו עם הנתונים הבחנו כי בנתוני המשתנה הקטגוריאלי Continent יבשת אירופה (יבשת 4) מכילה אך ורק 2 דגימות. 2 דגימות הן לא מדגם מייצג המאפשר לנתח מידע ולהסיק על כלל המדינות באירופה, לכן נרצה לאחד את קטגוריה זו עם קטגוריה אחרת. מניתוח שאר המדינות מצאנו דמיון עם יבשת אמריקה הצפונית, שכן שתיהן יבשות עם מדינות מפותחות ובהתאמה ניתן לראות כי בשתי יבשות אלה תוחלת החיים דומה וגבוהה באופן יחסי. לכן, ראינו לנכון לאחד ביניהן לאור המאפיינים ואורח החיים הדומה שלהן. בחרנו לאחד בין 2 קטגוריות אלו ולהמיר את הנתונים שמשויכים לאמריקה הצפונית (קטגוריה 5) לקטגוריה 4 מאוחדת של יבשות מפותחות (נספח 2). בצורה זו לא השפענו על הנתונים אך כעת ניתן להסיק בצורה טובה יותר על המגמה ביבשות אלה.

<u>דיסקרטיזציה</u>

דיסקרטיזציה הינו תהליך הפיכת משתנה רציף למשתנה קטגוריאלי.
הערכים של משתנים קטגוריאליים נקבעים בדרך כלל על ידי תכונה איכותית, כלומר תכונה שלא ניתן למדוד באופן כמותי. אולם, במודל שלנו ניכר כי כל המשתנים הרציפים הם משתנים שאכן ניתן למדוד אותם באופן כמותי, לדוגמה: אחוזים וכסף.
כמו כן, משתנה קטגוריאלי משמש גם למצב בו יש צורך למיון או הבחנה בין קבוצות ובמודל שלנו לא מצאנו משתנים רציפים אשר מקיימים שוני מהותי בין טווחי הערכים. לכן, הגענו למסקנה כי אין הצדקה להמרה כזו במודל שלנו.

<u>הגדרת משתנה דמה:</u>

עבור המשתנים הקטגוריאליים Continent ו-Member of OECD נגדיר משתני דמה בעזרת הפונקציה פקטור המאפשרת התייחסות למשתנה כקטגוריאלי כדי שנוכל לבצע ניתוחים סטטיסטיים מתקדמים (נספח 3).

<u>יבשות</u> - יבשת הינה משתנה קטגוריאלי בעלת 4 קבוצות (לאחר שאיחדנו את יבשות 4 ו-5 לקטגוריה מאוחדת 4 של יבשות מפותחות).
קבוצת הבסיס תהיה אסיה.
קבוצה 1 תקבל את הערך 1 אם המדינה היא מיבשת אפריקה.
קבוצה 2 תקבל את הערך 1 אם המדינה היא מיבשת אמריקה הדרומית.
קבוצה 3 תקבל את הערך 1 אם המדינה היא מיבשות מפותחות (אירופה ואמריקה הצפונית).

$$C_1 = \begin{cases} 1\,, if\ country\ is\ from\ Africa \\ 0\,, else \end{cases}$$

$$C_2 = \begin{cases} 1\,, if\ country\ is\ from\ South\ America \\ 0\,, else \end{cases}$$

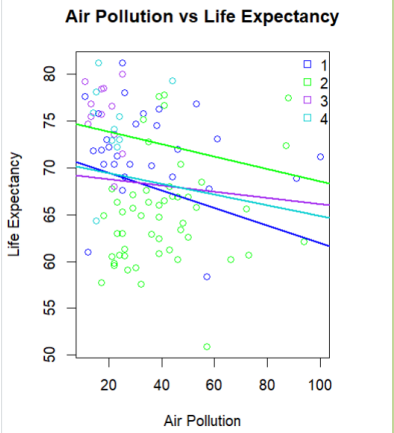
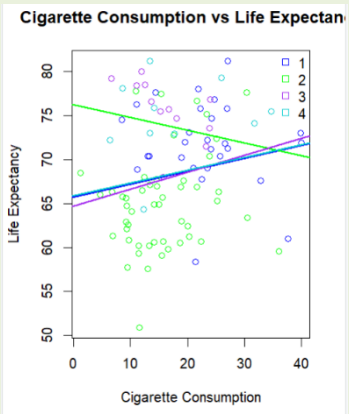
$$C_3 = \begin{cases} 1\,, if\ country\ is\ from\ Evolving \\ 0\,, else \end{cases}$$

<u>השתייכות לארגון הOECD</u> – השתייכות לארגון הOECD הינו משתנה קטגוריאלי בינארי בעל שתי קבוצות, מדינה המשתייכת לארגון תקבל ערך 1, אחרת 0.
קבוצת הבסיס תהיה אי השתייכות לארגון הOECD.
קבוצה 1 תקבל את הערך 1 אם המדינה משתייכת לארגון הOECD, אחרת 0.

$$M = \begin{cases} 1\,, if\ country\ is\ associated\ with\ the\ OECD\ organization \\ 0\,, else \end{cases}$$

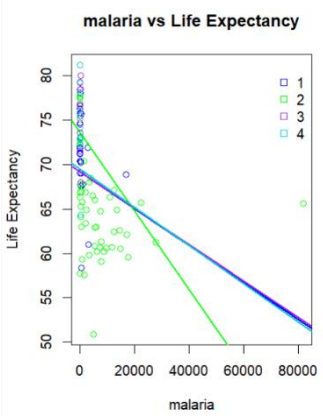
<u>הגדרה והוספה של משתני אינטראקציה</u>

משתני אינטראקציה יעזרו להבין את השפעת הגורמים השונים בכל קטגוריה על קו הרגרסיה. נחפש לזהות את הקשר בין המשתנים הרציפים השונים בכל קטגוריה למשתנה המוסבר. נראה כיצד החלוקה לקטגוריות משפיעה על המשתנים המסבירים בעזרת תרשימי פיזור עם קווי רגרסיה **(נספח 4)**. נבחר להתמקד באינטראקציות המשמעותיות ביותר במודל:

יבשת- אחוז זיהום האוויר נבחן את האינטראקציה בין ממד זיהום האוויר ותוחלת החיים כאשר הנתונים מפולגים ליבשות. קיבלנו קווי רגרסיה שונים בעלי שיפועים וחותכים שונים (אין קווים מקבילים). לכן ניתן לומר כי תוחלת החיים ביחס לזיהום האוויר אכן מושפעת מסוג היבשת. **הקשר מעיד על אינטראקציה אפשרית.**	
יבשת- אחוז צרכני הסיגריות נבחן את האינטראקציה בין אחוז צרכני הסיגריות ותוחלת החיים כאשר הנתונים מפולגים לפי יבשות. קיבלנו קווי רגרסיה בעלי שיפועים וחותכים שונים (לא מקבילים). לכן ניתן להניח כי הקשר בין תוחלת החיים לאחוז צריכת הסיגריות מושפע מסוג היבשת. **הקשר מעיד על אינטראקציה אפשרית.**	

יבשת- מספר החולים במלריה

נבחן את האינטראקציה בין מספר החולים במלריה ותוחלת החיים כאשר הנתונים מפולגים לפי יבשות.
גם כאן קיבלנו קווי רגרסיה עם שיפועים וחותכים שונים. לכן ניתן להניח כי הקשר בין תוחלת החיים לאחוז צריכת הסיגריות מושפע מסוג היבשת.
הקשר מעיד על אינטראקציה אפשרית.



התאמת המודל ובדיקת הנחות המודל:

בחירת משתני המודל:

לאחר הסרת המשתנים שביצענו, איחוד קטגוריות, הוספת משתני הדמה ובחינת הוספת משתני האינטראקציה, נגדיר כעת את משתני מודל הרגרסיה לאחר השינויים:

$$y = Life\ Expectancy$$

$$x_1 = Outdoor\ Air\ Pollution$$

$$x_2 = HIV$$

$$x_3 = Malaria$$

$$x_4 = Average\ Income$$

$$x_5 = Cigarette\ Consumption$$

$$C_1 = \begin{cases} 1\ , if\ country\ is\ from\ Africa \\ 0\ , else \end{cases}$$

$$C_2 = \begin{cases} 1\ , if\ country\ is\ from\ South\ America \\ 0\ , else \end{cases}$$

$$C_3 = \begin{cases} 1\ , if\ country\ is\ from\ Evolving \\ 0\ , else \end{cases}$$

$$M = \begin{cases} 1\ , if\ country\ is\ associated\ with\ the\ OECD\ organization \\ 0\ , else \end{cases}$$

בנוסף, המודל הנ"ל מכיל את האינטראקציות שהוצגו קודם לכן.

בחרנו להשתמש באלגוריתמים הבאים לבחירת משתני המודל:

Forward Selection (נספח 5) – בהתחלה נבחן מודל ללא משתנים כלל ואליו נכניס משתנים בזה אחר זה. המשתנה המובהק ביותר (בעל ה-Fst החלקי הגדול ביותר מבין הסטטיסטים המובהקים) ייבחן להכנסה למודל. לאחר מכן, נבדוק באמצעות מבחן F חלקי הוספה של משתנה נוסף, המובהק ביותר. נעצור כאשר נגיע למועמד הטוב ביותר להיכנס מהמודל כלומר כאשר לא נדחה את השערת האפס.

Backward Elimination (נספח 6) – בהתחלה נבחן את המודל המלא עם כל המשתנים וממנו נבחן הוצאה של משתנים בזה אחר זה. בכל שלב נבחן הוצאה של המשתנה הכי פחות מובהק (בעל ה-Fst החלקי הכי קטן) מהמודל. לאחר מכן, נבדוק באמצעות מבחן F חלקי הורדה של משתנה נוסף, הכי פחות מובהק. נעצור כאשר נגיע למועמד הטוב ביותר ליציאה מהמודל כלומר כאשר נדחה את השערת האפס.

Stepwise Regression (נספח 7) - גישה אשר משלבת את שתי הגישות שהצגנו.
בכל שלב נבחן האם להוציא או להכניס משתנים מתוך אלו שהוספנו למודל בצעדים הקודמים.
נדחה את השערת האפס עבור המועמד הטוב ביותר ליציאה ונדחה את השערת האפס עבור המועמד הטוב ביותר לכניסה.

שלושת האלגוריתמים שהוצגו לעיל העדיפו את אותו המודל: Continent + Average Income.

בחרנו במדד $R^2_{A\,dj}$ לבחירה בין החלופות, המודל המלא והמודל החלקי שהתקבל לאחר ביצוע האלגוריתמים, מכיוון שמדד זה מתאים את עצמו בצורה טובה להשוואה בין מודלים בגדלים שונים. מדד $R^2_{A\,dj}$ (נספח 8): מודל מלא – 0.6223, מודל חלקי – 0.6344.

כלומר, על פי מדד $R^2_{A\,dj}$ נקבל כי המודל הטוב יותר הינו המודל החלקי שהתקבל מביצוע שלושת האלגוריתמים לעיל.

מדד BIC (נספח 9): מודל מלא – 610.5423 , מודל חלקי – 589.5998.
לפיכך, גם על פי מדד BIC התקבלה אותה המסקנה, נבחר במודל החלקי כאשר המסבירים של תוחלת החיים במדינה הם היבשת וההכנסה הממוצעת השנתית לאדם.

מכאן שעל פי כל האלגוריתמים והמדדים השונים הגענו לאותה המסקנה לגבי המודל הטוב ביותר:

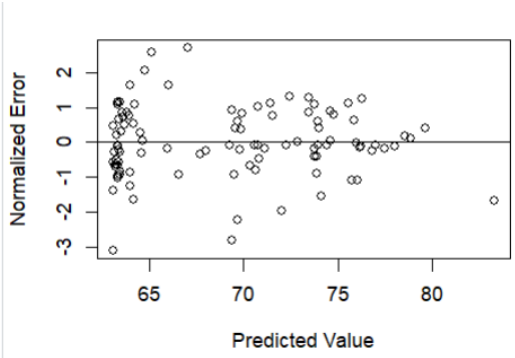
$$y = \beta_0 + \beta_1 C_1 + \beta_2 C_2 + \beta_3 C_3 + \beta_4 X$$

לאחר ההצבה (נספח 10):

$$y = 68.8460274725 - 5.9726874321 C_1 + 2.9881679395 C_2 + 2.6604494895 C_3 + 0.0003014502\, X$$

בדיקת הנחות המודל (נספח 11):

בדיקת הנחת שוויון שונויות



על מנת לבדוק את הנחת שוויון שוניות נשתמש בתרשים השאריות כפונקציה של התחזיות. ניתן לראות בתרשים כי ישנו פיזור מסוים של נתונים סביב קו ה-0 אך הוא אינו אחיד. לא ניתן לקבוע באופן חד משמעי כי הנחת שוויון השונויות מתקיימת.

כדי לקבוע באופן חד משמעי נשתמש במבחן GQ להשוואת שונויות:

```
        Goldfeld-Quandt test

data:  bestModel
GQ = 0.72147, df1 = 46, df2 = 45, p-value = 0.8631
alternative hypothesis: variance increases from segment 1 to 2
```

מתוצאות המבחן נקבל Pval גדול מ-0.05 ולכן לא נדחה את השערת ה-0 ונאמר כי **הנחת שוויון השונויות מתקיימת**.

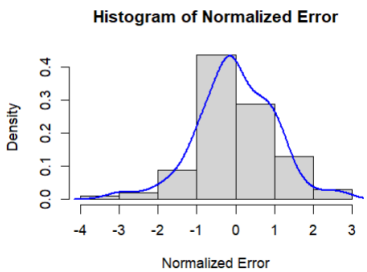
נשתמש במבחן F ליחס שונויות כדי לקבוע האם השונות היא קבועה. ניתן לראות כי Pval גדול מ-0.05 ולכן נדחה את השערת ה-0 ונאמר כי **השונות היא קבועה**.

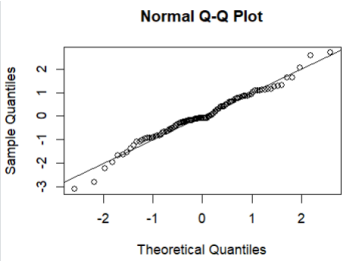
```
        F test to compare two variances

data:  third_vec_values and third23_vec_values
F = 1.716, num df = 33, denom df = 34, p-value = 0.1222
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8631144 3.4239377
sample estimates:
ratio of variances
          1.716016
```

בדיקת הנחת הנורמאליות של השגיאות

כדי לבדוק את ההנחה נייצא תרשים qq plot ההיסטוגרמה:





בתרשים ה-qq plot ניתן לראות כי לא כל הנקודות מסודרות בדיוק על הקו. ישנן נקודות רחוקות מהקו בקצוות התרשים.

מתרשים ההיסטוגרמה ניתן לראות שפונקציית הצפיפות המתקבלת דומה לפונקציית הצפיפות של ההתפלגות הנורמאלית עם תוחלת 0.

כדי לקבוע אם אכן מתפלג נורמאלי נשתמש במבנים סטטיסטים:

```
        Shapiro-Wilk normality test

data:  newDataset$standardResiduals
W = 0.98207, p-value = 0.1871
```

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  newDataset$standardResiduals
D = 0.071035, p-value = 0.688
alternative hypothesis: two-sided
```

ממבחני ה-SW,KS קיבלנו רמת מובהקות אשר גדולה מ-0.05 ולכן לא נדחה את השערת ה-0 ונאמר כי **השגיאות המתוקננות מתפלגות נורמלית**.

בדיקת הנחת הליניאריות

בתרשים השאריות אל מול החיזוי שבדקנו בהנחת שוויון השונויות ניתן לזהות מגמה ליניארית בנתונים. עם זאת כדי להיות בטוחים נשתמש במבחן CHOW:

```
        M-fluctuation test

data:  bestModel
f(efp) = 0.89596, p-value = 0.9212
```

התקבל Pval גדול מ-0.05 ולכן לא נדחה את השערת ה-0 ונאמר כי **הנחת הליניאריות מתקיימת.**

שיפור המודל

כפי שניתן לראות המודל הסופי שהתקבל בתום סעיף 3 עומד בכלל ההנחות של מודל רגרסיה לינארית: הנחת שוויון השונויות, נורמליות השגיאות והנחת הלינאריות. בחרנו לנסות לבחון טרנספורמציות על המשתנה המוסבר במודל שלנו, תוחלת חיי האדם במדינה, במטרה לשפר את המודל עד כמה שניתן למרות שהנחות המודל כבר מתקיימות כפי שהראינו קודם לכן.
יתר על כן, כדי להישאר עקביות נבחר להשוות בין המודל שהתקבל בתום הסעיף הקודם לבין המודל הנוכחי שהתקבל לאחר ביצוע הטרנספורמציה בעזרת אותו מדד $R^2{}_{adj}$.

בחרנו להשתמש בטרנספורמציה של Box-Cox בשאיפה למצוא את המודל הטוב ביותר.
מדובר בטכניקה סטטיסטית המשמשת להפיכת נתונים לא נורמליים להתפלגות נורמלית.
טרנספורמציה זו יכולה לשפר את הדיוק של תחזיות שנעשות באמצעות רגרסיה ליניארית.
ניתן להשתמש בטרנספורמציה זו על גבי נתונים שאינם מתפלגים נורמלית, נתונים מוטים (בעלי זנב ימני או שמאלי) או חריגים.

הטרנספורמציה של Y היא מהצורה:

$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log y, & \text{if } \lambda = 0. \end{cases}$$

הלמדא שהתקבלה מביצוע הטרנספורמציה של Box Cox היא $\lambda = 2$ (**נספח 12**).
מאחר והתקבלה למדא שונה מאפס ניישם את הטרנספורמציה שתוארה לעיל על המשתנה המוסבר במודל שלנו.

לעיתים עשויים להיות שינויים בבחירת המשתנים המסבירים בהתאם לשינוי במשתנה המוסבר ולכן נבדוק את האלגוריתמים השונים לאחר ביצוע הטרנספורמציה על Y. על פי האלגוריתמים Forward Selection ו-Stepwise Regression התקבל כי בחירת המשתנים המסבירים נותרה כפי שהייתה לפני הטרנספורמציה ללא שינוי (**נספח 13**).
לכן נשאיר את המשתנים המסבירים כפי שהם.

נמשיך ונבדוק גם את המדד $R^2_{A\,dj}$ לטובת בחירה בין החלופות. נשווה את המדד כפי שהתקבל לפני ואחרי הטרנספורמציה.
מדד $R^2_{A\,dj}$ (**נספח 14**): מודל קודם – 0.6344, מודל לאחר טרנספורמציה – 0.6477.

כלומר, על פי מדד $R^2_{A\,dj}$ נקבל כי המודל הטוב ביותר הינו המודל שהתקבל לאחר יישום הטרנספורמציה ולכן נבחר בו כמודל האופטימלי עבור המודל שלנו.

המודל הסופי שהתקבל הינו מהצורה:

$$\frac{y^2-1}{2} = \beta_0 + \beta_1 C_1 + \beta_2 C_2 + \beta_3 C_3 + \beta_4 X$$

לאחר ההצבה (**נספח 15**):

$$\frac{y^2-1}{2} = 2381.16372222 - 403.00889020\, C_1 + 218.13782895 C_2 + 194.24356281 C_3 + 0.02112584 X$$

נספחים

נספח 1 – מקדם המתאם של פירסון

```
> cor.test(dataset$Outdoor.air.pollution...., dataset$Life_expectancy, method = c("pearson"))

        Pearson's product-moment correlation

data:  dataset$Outdoor.air.pollution.... and dataset$Life_expectancy
t = -2.1104, df = 99, p-value = 0.03735
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.38722251 -0.01255529
sample estimates:
       cor
-0.2074854

> cor.test(dataset$HIV...Estimated.number.of.people.that.have.been.infected, dataset$Life_expectancy, method = c("pearson"))

        Pearson's product-moment correlation

data:  dataset$HIV...Estimated.number.of.people.that.have.been.infected and dataset$Life_expectancy
t = -2.1748, df = 99, p-value = 0.03203
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.39258933 -0.01888273
sample estimates:
      cor
-0.213534

> cor.test(dataset$malaria...Estimated.number.of.people.that.have.been.infected, dataset$Life_expectancy, method = c("pearson"))

        Pearson's product-moment correlation

data:  dataset$malaria...Estimated.number.of.people.that.have.been.infected and dataset$Life_expectancy
t = -3.6223, df = 99, p-value = 0.0004628
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.5038425 -0.1571568
sample estimates:
       cor
-0.3420889

> cor.test(dataset$Average.income.per.person...., dataset$Life_expectancy, method = c("pearson"))

        Pearson's product-moment correlation

data:  dataset$Average.income.per.person.... and dataset$Life_expectancy
t = 7.0949, df = 99, p-value = 1.966e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4344369 0.6969399
sample estimates:
      cor
0.5805813
```

```
> cor.test(dataset$Alcohol.consumption.per.person..liters..year., dataset$Life_expectancy, method = c("pearson"))

        Pearson's product-moment correlation

data:  dataset$Alcohol.consumption.per.person..liters..year. and dataset$Life_expectancy
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$density.per.square..km., dataset$Life_expectancy, method = c("pearson"))

        Pearson's product-moment correlation

data:  dataset$density.per.square..km. and dataset$Life_expectancy
t = 0.87617, df = 99, p-value = 0.3831
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1095992  0.2783857
sample estimates:
       cor
0.08771903

> cor.test(dataset$Cigarette.consumption...., dataset$Life_expectancy, method = c("pearson"))

        Pearson's product-moment correlation

data:  dataset$Cigarette.consumption.... and dataset$Life_expectancy
t = 1.6493, df = 99, p-value = 0.1022
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.03295972  0.34785453
sample estimates:
     cor
0.163533
```

נספח 2 – איחוד קטגוריות

```
# unite between Europe with North America to the category Developed Continents
newDataset <- dataset[,c(2,3,4,5,6,9,10,11)] #sub table without the relevant variables
for(i in 1:101){
  if(newDataset[i, 7] == "5")
    newDataset[i, 7] = "4"
}
```

נספח 3 – הגדרת משתני הדמה (קטגוריאליים)

```
# define Continent and Member of OECD as categorical variables and therefore are adequate to become the dummy variables
Continent_dummy <- factor(newDataset$Continent)
Continent_dummy <- relevel(Continent_dummy, ref=c(1))
OECD_dummy <- factor(newDataset$Member.of.OECD)
OECD_dummy <- relevel(OECD_dummy, ref=c(1))
```

נספח 4 – בדיקות להוספת משתני אינטראקציה

```
> max<-max(newDataset$Outdoor.air.pollution....)%>%print() #max
[1] 100
> min<-min(newDataset$Outdoor.air.pollution....)%>%print() #min
[1] 11
>
> check1 <- lm(formula = Life_expectancy ~ Outdoor.air.pollution.... * factor(newDataset$Continent=='1'), data = newDataset)
> check2 <- lm(formula = Life_expectancy ~ Outdoor.air.pollution.... * factor(newDataset$Continent=='2'), data = newDataset)
> check3 <- lm(formula = Life_expectancy ~ Outdoor.air.pollution.... * factor(newDataset$Continent=='3'), data = newDataset)
> check4 <- lm(formula = Life_expectancy ~ Outdoor.air.pollution.... * factor(newDataset$Continent=='4'), data = newDataset)
>
> plot(newDataset$Outdoor.air.pollution....[newDataset$Continent=='1'],newDataset$Life_expectancy[newDataset$Continent=='1'],
+      col="blue",xlab="Air Pollution",ylab="Life Expectancy",main="Air Pollution vs Life Expectancy",xlim=c(11.00,100.00),ylim=c(50.90,81.20))
>
> points(newDataset$Outdoor.air.pollution[newDataset$Continent=='2'],newDataset$Life_expectancy[newDataset$Continent=='2'],
+        col="green")
> points(newDataset$Outdoor.air.pollution[newDataset$Continent=='3'],newDataset$Life_expectancy[newDataset$Continent=='3'],
+        col="PURPLE")
> points(newDataset$Outdoor.air.pollution[newDataset$Continent=='4'],newDataset$Life_expectancy[newDataset$Continent=='4'],
+        col="darkturquoise")
Warning message:
In abline(check1, col = "blue", lwd = 2) :
  only using the first two of 4 regression coefficients
> abline(check2,col="green", lwd=2)
Warning message:
In abline(check2, col = "green", lwd = 2) :
  only using the first two of 4 regression coefficients
> abline(check3,col="PURPLE", lwd=2)
Warning message:
In abline(check3, col = "PURPLE", lwd = 2) :
  only using the first two of 4 regression coefficients
> abline(check4,col="darkturquoise", lwd=2)
Warning message:
In abline(check4, col = "darkturquoise", lwd = 2) :
  only using the first two of 4 regression coefficients
```

```
> max<-max(newDataset$malaria...Estimated.number.of.people.that.have.been.infected)%>%print() #max
> [1] 81640
> min<-min(newDataset$malaria...Estimated.number.of.people.that.have.been.infected)%>%print() #min
> [1] 0
>
> check1 <- lm(formula = Life_expectancy ~ malaria...Estimated.number.of.people.that.have.been.infected * factor(newDataset$Continent=='1'), data = newDataset)
> check2 <- lm(formula = Life_expectancy ~ malaria...Estimated.number.of.people.that.have.been.infected * factor(newDataset$Continent=='2'), data = newDataset)
> check3 <- lm(formula = Life_expectancy ~ malaria...Estimated.number.of.people.that.have.been.infected * factor(newDataset$Continent=='3'), data = newDataset)
> check4 <- lm(formula = Life_expectancy ~ malaria...Estimated.number.of.people.that.have.been.infected * factor(newDataset$Continent=='4'), data = newDataset)
>
> plot(newDataset$malaria...Estimated.number.of.people.that.have.been.infected[newDataset$Continent=='1'],newDataset$Life_expectancy[newDataset$Continent=='1'],
+      col="blue",xlab="malaria",ylab="Life Expectancy",main="malaria vs Life Expectancy", xlim=c(0,81640),ylim=c(50.90,81.20))
>
> points(newDataset$malaria...Estimated.number.of.people.that.have.been.infected[newDataset$Continent=='2'],newDataset$Life_expectancy[newDataset$Continent=='2'],
+        col="green")
> points(newDataset$malaria...Estimated.number.of.people.that.have.been.infected[newDataset$Continent=='3'],newDataset$Life_expectancy[newDataset$Continent=='3'],
+        col="PURPLE")
> points(newDataset$malaria...Estimated.number.of.people.that.have.been.infected[newDataset$Continent=='4'],newDataset$Life_expectancy[newDataset$Continent=='4'],
+        col="darkturquoise")
> legend(70000,81.20,legend=c("1","2","3","4"),col=c("blue","green","PURPLE","darkturquoise"),pch=c(0.5,0.5,0.5,0.5),bty="n")
>
> abline(check1,col="blue", lwd=2)
Warning message:
In abline(check1, col = "blue", lwd = 2) :
  only using the first two of 4 regression coefficients
> abline(check2,col="green", lwd=2)
Warning message:
In abline(check2, col = "green", lwd = 2) :
  only using the first two of 4 regression coefficients
> abline(check3,col="PURPLE", lwd=2)
Warning message:
In abline(check3, col = "PURPLE", lwd = 2) :
  only using the first two of 4 regression coefficients
> abline(check4,col="darkturquoise", lwd=2)
Warning message:
In abline(check4, col = "darkturquoise", lwd = 2) :
  only using the first two of 4 regression coefficients
> abline(check4,col="darkturquoise", lwd=2)
Warning message:
In abline(check4, col = "darkturquoise", lwd = 2) :
  only using the first two of 4 regression coefficients
```

```
> max<-max(newDataset$HIV...Estimated.number.of.people.that.have.been.infected)%>%print() #max
[1] 7700000
> min<-min(newDataset$HIV...Estimated.number.of.people.that.have.been.infected)%>%print() #min
[1] 200
>
> check1 <- lm(formula = Life_expectancy ~ HIV...Estimated.number.of.people.that.have.been.infected * factor(newDataset$Continent=='1'), data = newDataset)
> check2 <- lm(formula = Life_expectancy ~ HIV...Estimated.number.of.people.that.have.been.infected * factor(newDataset$Continent=='2'), data = newDataset)
> check3 <- lm(formula = Life_expectancy ~ HIV...Estimated.number.of.people.that.have.been.infected * factor(newDataset$Continent=='3'), data = newDataset)
> check4 <- lm(formula = Life_expectancy ~ HIV...Estimated.number.of.people.that.have.been.infected * factor(newDataset$Continent=='4'), data = newDataset)
>
> plot(newDataset$HIV...Estimated.number.of.people.that.have.been.infected[newDataset$Continent=='1'],newDataset$Life_expectancy[newDataset$Continent=='1'],
+      col="blue",xlab="HIV",ylab="Life Expectancy",main="HIV vs Life Expectancy", xlim=c(200,7700000),ylim=c(50.90,81.20))
>
> points(newDataset$HIV...Estimated.number.of.people.that.have.been.infected[newDataset$Continent=='2'],newDataset$Life_expectancy[newDataset$Continent=='2'],
+        col="green")
> points(newDataset$HIV...Estimated.number.of.people.that.have.been.infected[newDataset$Continent=='3'],newDataset$Life_expectancy[newDataset$Continent=='3'],
+        col="PURPLE")
> points(newDataset$HIV...Estimated.number.of.people.that.have.been.infected[newDataset$Continent=='4'],newDataset$Life_expectancy[newDataset$Continent=='4'],
+        col="darkturquoise")
> legend(7700000,81.20,legend=c("1","2","3","4"),col=c("blue","green","PURPLE","darkturquoise"),pch=c(0.5,0.5,0.5,0.5),bty="n")
>
> abline(check1,col="blue", lwd=2)
Warning message:
In abline(check1, col = "blue", lwd = 2) :
  only using the first two of 4 regression coefficients
> abline(check2,col="green", lwd=2)
Warning message:
In abline(check2, col = "green", lwd = 2) :
  only using the first two of 4 regression coefficients
> abline(check3,col="PURPLE", lwd=2)
Warning message:
In abline(check3, col = "PURPLE", lwd = 2) :
  only using the first two of 4 regression coefficients
> abline(check4,col="darkturquoise", lwd=2)
Warning message:
In abline(check4, col = "darkturquoise", lwd = 2) :
  only using the first two of 4 regression coefficients
```

```
> max<-max(newDataset$Cigarette.consumption....)%>%print() #max
[1] 39.9
> min<-min(newDataset$Cigarette.consumption....)%>%print() #min
[1] 1.3
>
> check1 <- lm(formula = Life_expectancy ~ Cigarette.consumption.... * factor(newDataset$Continent=='1'), data = newDataset)
> check2 <- lm(formula = Life_expectancy ~ Cigarette.consumption.... * factor(newDataset$Continent=='2'), data = newDataset)
> check3 <- lm(formula = Life_expectancy ~ Cigarette.consumption.... * factor(newDataset$Continent=='3'), data = newDataset)
> check4 <- lm(formula = Life_expectancy ~ Cigarette.consumption.... * factor(newDataset$Continent=='4'), data = newDataset)
>
> plot(newDataset$Cigarette.consumption....[newDataset$Continent=='1'],newDataset$Life_expectancy[newDataset$Continent=='1'],
+      col="blue",xlab="Cigarette Consumption",ylab="Life Expectancy",main="Cigarette Consumption vs Life Expectancy", xlim=c(1.3,39.9),ylim=c(50.90,81.20))
>
> points(newDataset$Cigarette.consumption....[newDataset$Continent=='2'],newDataset$Life_expectancy[newDataset$Continent=='2'],
+        col="green")
> points(newDataset$Cigarette.consumption....[newDataset$Continent=='3'],newDataset$Life_expectancy[newDataset$Continent=='3'],
+        col="PURPLE")
> points(newDataset$Cigarette.consumption....[newDataset$Continent=='4'],newDataset$Life_expectancy[newDataset$Continent=='4'],
+        col="darkturquoise")
> legend(35,83,legend=c("1","2","3","4"),col=c("blue","green","PURPLE","darkturquoise"),pch=c(0.5,0.5,0.5,0.5),bty="n")
>
> abline(check1,col="blue", lwd=2)
Warning message:
In abline(check1, col = "blue", lwd = 2) :
  only using the first two of 4 regression coefficients
> abline(check2,col="green", lwd=2)
Warning message:
In abline(check2, col = "green", lwd = 2) :
  only using the first two of 4 regression coefficients
> abline(check3,col="PURPLE", lwd=2)
Warning message:
In abline(check3, col = "PURPLE", lwd = 2) :
  only using the first two of 4 regression coefficients
> abline(check4,col="darkturquoise", lwd=2)
Warning message:
In abline(check4, col = "darkturquoise", lwd = 2) :
  only using the first two of 4 regression coefficients
```

Step Forward – נספח 5

```
> fwd.model <- step(Emp, direction = 'forward', scope = formula(Full))
Start:  AIC=383.02
newDataset$Life_expectancy ~ 1

                                                                  Df Sum of Sq    RSS    AIC
+ Continent_dummy                                                  3   1885.30 2507.0 332.38
+ newDataset$Average.income.per.person....                         1   1480.53 2911.8 343.50
+ newDataset$malaria...Estimated.number.of.people.that.have.been.infected  1    514.01 3878.3 372.45
+ newDataset$HIV...Estimated.number.of.people.that.have.been.infected      1    200.27 4192.0 380.31
+ newDataset$Outdoor.air.pollution....                             1    189.09 4203.2 380.58
+ OECD_dummy                                                       1    159.05 4233.2 381.30
+ newDataset$Cigarette.consumption....                             1    117.46 4274.8 382.28
<none>                                                                         4392.3 383.02

Step:  AIC=332.38
newDataset$Life_expectancy ~ Continent_dummy

                                                                  Df Sum of Sq    RSS    AIC
+ newDataset$Average.income.per.person....                         1    965.19 1541.8 285.28
+ newDataset$malaria...Estimated.number.of.people.that.have.been.infected  1     68.89 2438.1 331.57
<none>                                                                         2507.0 332.38
+ newDataset$HIV...Estimated.number.of.people.that.have.been.infected      1     33.28 2473.7 333.03
+ OECD_dummy                                                       1     16.90 2490.1 333.70
+ newDataset$Outdoor.air.pollution....                             1     10.74 2496.2 333.95
+ newDataset$Cigarette.consumption....                             1      6.38 2500.6 334.13

Step:  AIC=285.28
newDataset$Life_expectancy ~ Continent_dummy + newDataset$Average.income.per.person....

                                                                  Df Sum of Sq    RSS    AIC
<none>                                                                         1541.8 285.28
+ newDataset$HIV...Estimated.number.of.people.that.have.been.infected      1   23.3878 1518.4 285.74
+ newDataset$malaria...Estimated.number.of.people.that.have.been.infected  1    9.1240 1532.7 286.68
+ newDataset$Outdoor.air.pollution....                             1    3.3512 1538.4 287.06
+ newDataset$Cigarette.consumption....                             1    0.4051 1541.4 287.26
+ OECD_dummy                                                       1    0.0067 1541.8 287.28
>
```

Step Backward – נספח 6

```
> bw.model <- step(Full, direction = 'backward', scope = ~ 1)
Start:  AIC=300.74
newDataset$Life_expectancy ~ newDataset$Outdoor.air.pollution.... +
    newDataset$HIV...Estimated.number.of.people.that.have.been.infected +
    newDataset$malaria...Estimated.number.of.people.that.have.been.infected +
    newDataset$Average.income.per.person.... + newDataset$Cigarette.consumption.... +
    Continent_dummy + OECD_dummy + newDataset$Outdoor.air.pollution.... *
    Continent_dummy + newDataset$malaria...Estimated.number.of.people.that.have.been.infected *
    Continent_dummy + newDataset$Cigarette.consumption.... *
    Continent_dummy

                                                                                  Df Sum of Sq    RSS    AIC
- newDataset$Outdoor.air.pollution....:Continent_dummy                             3      7.38 1369.0 295.28
- OECD_dummy                                                                       1      0.38 1362.0 298.76
- newDataset$Cigarette.consumption....:Continent_dummy                             3     58.99 1420.7 299.02
- newDataset$HIV...Estimated.number.of.people.that.have.been.infected              1      8.16 1369.8 299.34
- newDataset$malaria...Estimated.number.of.people.that.have.been.infected:Continent_dummy  3     82.81 1444.5 300.70
<none>                                                                                         1361.7 300.74
- newDataset$Average.income.per.person....                                         1    670.17 2031.8 339.16

Step:  AIC=295.28
newDataset$Life_expectancy ~ newDataset$Outdoor.air.pollution.... +
    newDataset$HIV...Estimated.number.of.people.that.have.been.infected +
    newDataset$malaria...Estimated.number.of.people.that.have.been.infected +
    newDataset$Average.income.per.person.... + newDataset$Cigarette.consumption.... +
    Continent_dummy + OECD_dummy + newDataset$malaria...Estimated.number.of.people.that.have.been.infected:Continent_dummy +
    newDataset$Cigarette.consumption....:Continent_dummy

                                                                                  Df Sum of Sq    RSS    AIC
- OECD_dummy                                                                       1      1.09 1370.1 293.36
- newDataset$Outdoor.air.pollution....                                             1      6.84 1375.9 293.79
- newDataset$Cigarette.consumption....:Continent_dummy                             3     62.76 1431.8 293.81
- newDataset$HIV...Estimated.number.of.people.that.have.been.infected              1     10.08 1379.1 294.02
- newDataset$malaria...Estimated.number.of.people.that.have.been.infected:Continent_dummy  3     82.14 1451.2 295.17
<none>                                                                                         1369.0 295.28
- newDataset$Average.income.per.person....                                         1    770.66 2139.7 338.38
```

```
Step:  AIC=293.36
newDataset$Life_expectancy ~ newDataset$Outdoor.air.pollution.... +
    newDataset$HIV...Estimated.number.of.people.that.have.been.infected +
    newDataset$malaria...Estimated.number.of.people.that.have.been.infected +
    newDataset$Average.income.per.person.... + newDataset$Cigarette.consumption.... +
    Continent_dummy + newDataset$malaria...Estimated.number.of.people.that.have.been.infected:Continent_dummy +
    newDataset$Cigarette.consumption....:Continent_dummy

                                                                                  Df Sum of Sq    RSS    AIC
- newDataset$Outdoor.air.pollution....                                             1      6.39 1376.5 291.83
- newDataset$Cigarette.consumption....:Continent_dummy                             3     62.23 1432.4 291.85
- newDataset$HIV...Estimated.number.of.people.that.have.been.infected              1     10.23 1380.4 292.11
- newDataset$malaria...Estimated.number.of.people.that.have.been.infected:Continent_dummy  3     81.49 1451.6 293.20
<none>                                                                                         1370.1 293.36
- newDataset$Average.income.per.person....                                         1    773.99 2144.1 336.59

Step:  AIC=291.83
newDataset$Life_expectancy ~ newDataset$HIV...Estimated.number.of.people.that.have.been.infected +
    newDataset$malaria...Estimated.number.of.people.that.have.been.infected +
    newDataset$Average.income.per.person.... + newDataset$Cigarette.consumption.... +
    Continent_dummy + newDataset$malaria...Estimated.number.of.people.that.have.been.infected:Continent_dummy +
    newDataset$Cigarette.consumption....:Continent_dummy

                                                                                  Df Sum of Sq    RSS    AIC
- newDataset$Cigarette.consumption....:Continent_dummy                             3     63.05 1439.6 290.36
- newDataset$HIV...Estimated.number.of.people.that.have.been.infected              1     13.57 1390.1 290.82
- newDataset$malaria...Estimated.number.of.people.that.have.been.infected:Continent_dummy  3     78.09 1454.6 291.40
<none>                                                                                         1376.5 291.83
- newDataset$Average.income.per.person....                                         1    794.25 2170.8 335.84

Step:  AIC=290.36
newDataset$Life_expectancy ~ newDataset$HIV...Estimated.number.of.people.that.have.been.infected +
    newDataset$malaria...Estimated.number.of.people.that.have.been.infected +
    newDataset$Average.income.per.person.... + newDataset$Cigarette.consumption.... +
    Continent_dummy + newDataset$malaria...Estimated.number.of.people.that.have.been.infected:Continent_dummy

                                                                                  Df Sum of Sq    RSS    AIC
- newDataset$Cigarette.consumption....                                             1      0.23 1439.8 288.37
- newDataset$HIV...Estimated.number.of.people.that.have.been.infected              1     12.46 1452.0 289.23
- newDataset$malaria...Estimated.number.of.people.that.have.been.infected:Continent_dummy  3     74.79 1514.4 289.47
<none>                                                                                         1439.6 290.36
- newDataset$Average.income.per.person....                                         1    852.52 2292.1 335.33
```

```
Step:  AIC=288.37
newDataset$Life_expectancy ~ newDataset$HIV...Estimated.number.of.people.that.have.been.infected +
    newDataset$malaria...Estimated.number.of.people.that.have.been.infected +
    newDataset$Average.income.per.person.... + Continent_dummy +
    newDataset$malaria...Estimated.number.of.people.that.have.been.infected:Continent_dummy

                                                                                  Df Sum of Sq    RSS    AIC
- newDataset$HIV...Estimated.number.of.people.that.have.been.infected              1     12.52 1452.3 287.25
- newDataset$malaria...Estimated.number.of.people.that.have.been.infected:Continent_dummy  3     74.69 1514.5 287.48
<none>                                                                                         1439.8 288.37
- newDataset$Average.income.per.person....                                         1    852.60 2292.4 333.35

Step:  AIC=287.25
newDataset$Life_expectancy ~ newDataset$malaria...Estimated.number.of.people.that.have.been.infected +
    newDataset$Average.income.per.person.... + Continent_dummy +
    newDataset$malaria...Estimated.number.of.people.that.have.been.infected:Continent_dummy

                                                                                  Df Sum of Sq    RSS    AIC
- newDataset$malaria...Estimated.number.of.people.that.have.been.infected:Continent_dummy  3     80.34 1532.7 286.68
<none>                                                                                         1452.3 287.25
- newDataset$Average.income.per.person....                                         1    849.79 2302.1 331.77

Step:  AIC=286.68
newDataset$Life_expectancy ~ newDataset$malaria...Estimated.number.of.people.that.have.been.infected +
    newDataset$Average.income.per.person.... + Continent_dummy

                                                                              Df Sum of Sq    RSS    AIC
- newDataset$malaria...Estimated.number.of.people.that.have.been.infected  1      9.12 1541.8 285.28
<none>                                                                                   1532.7 286.68
- newDataset$Average.income.per.person....                                 1    905.42 2438.1 331.57
- Continent_dummy                                                          3   1165.57 2698.2 337.81

Step:  AIC=285.28
newDataset$Life_expectancy ~ newDataset$Average.income.per.person.... +
    Continent_dummy

                                        Df Sum of Sq    RSS    AIC
<none>                                                 1541.8 285.28
- newDataset$Average.income.per.person....  1    965.19 2507.0 332.38
- Continent_dummy                           3   1369.97 2911.8 343.50
```

נספח 7 – Stepwise

```
> sw.model <- step(Emp, direction = 'both', scope = formula(Full))
Start:  AIC=383.02
newDataset$Life_expectancy ~ 1

                                                                  Df Sum of Sq    RSS    AIC
+ Continent_dummy                                                  3   1885.30 2507.0 332.38
+ newDataset$Average.income.per.person....                         1   1480.53 2911.8 343.50
+ newDataset$malaria...Estimated.number.of.people.that.have.been.infected 1    514.01 3878.3 372.45
+ newDataset$HIV...Estimated.number.of.people.that.have.been.infected 1    200.27 4192.0 380.31
+ newDataset$Outdoor.air.pollution....                             1    189.09 4203.2 380.58
+ OECD_dummy                                                       1    159.05 4233.2 381.30
+ newDataset$Cigarette.consumption....                             1    117.46 4274.8 382.28
<none>                                                                         4392.3 383.02

Step:  AIC=332.38
newDataset$Life_expectancy ~ Continent_dummy

                                                                  Df Sum of Sq    RSS    AIC
+ newDataset$Average.income.per.person....                         1    965.19 1541.8 285.28
+ newDataset$malaria...Estimated.number.of.people.that.have.been.infected 1     68.89 2438.1 331.57
<none>                                                                         2507.0 332.38
+ newDataset$HIV...Estimated.number.of.people.that.have.been.infected 1     33.28 2473.7 333.03
+ OECD_dummy                                                       1     16.90 2490.1 333.70
+ newDataset$Outdoor.air.pollution....                             1     10.74 2496.2 333.95
+ newDataset$Cigarette.consumption....                             1      6.38 2500.6 334.13
- Continent_dummy                                                  3   1885.30 4392.3 383.02

Step:  AIC=285.28
newDataset$Life_expectancy ~ Continent_dummy + newDataset$Average.income.per.person....

                                                                  Df Sum of Sq    RSS    AIC
<none>                                                                         1541.8 285.28
+ newDataset$HIV...Estimated.number.of.people.that.have.been.infected 1     23.39 1518.4 285.74
+ newDataset$malaria...Estimated.number.of.people.that.have.been.infected 1      9.12 1532.7 286.68
+ newDataset$Outdoor.air.pollution....                             1      3.35 1538.4 287.06
+ newDataset$Cigarette.consumption....                             1      0.41 1541.4 287.26
+ OECD_dummy                                                       1      0.01 1541.8 287.28
- newDataset$Average.income.per.person....                         1    965.19 2507.0 332.38
- Continent_dummy                                                  3   1369.97 2911.8 343.50
```

נספח 8 – השוואה על פי מדד Rsquare adjusted

```
> AIC.model <- lm(formula = newDataset$Life_expectancy ~ Continent_dummy + newDataset$Average.income.per.person.... , data = newDataset)
> summary.lm(AIC.model)

Call:
lm(formula = newDataset$Life_expectancy ~ Continent_dummy + newDataset$Average.income.per.person....,
    data = newDataset)

Residuals:
     Min       1Q   Median       3Q      Max
-12.1747  -2.2859  -0.3223   2.8821  10.5666

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                          6.885e+01  8.476e-01  81.225  < 2e-16 ***
Continent_dummy2                    -5.973e+00  9.483e-01  -6.298 9.05e-09 ***
Continent_dummy3                     2.988e+00  1.388e+00   2.153   0.0338 *
Continent_dummy4                     2.660e+00  1.477e+00   1.802   0.0747 .
newDataset$Average.income.per.person....  3.014e-04  3.889e-05   7.752 9.46e-12 ***
---
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:  0.6344
F-statistic: 44.37 on 4 and 96 DF,  p-value: < 2.2e-16
```

```
> summary.lm(Full)

Call:
lm(formula = newDataset$Life_expectancy ~ newDataset$Outdoor.air.pollution.... +
    newDataset$HIV...Estimated.number.of.people.that.have.been.infected +
    newDataset$malaria...Estimated.number.of.people.that.have.been.infected +
    newDataset$Average.income.per.person.... + newDataset$Cigarette.consumption.... +
    Continent_dummy + OECD_dummy + newDataset$Outdoor.air.pollution.... *
    Continent_dummy + newDataset$malaria...Estimated.number.of.people.that.have.been.infected *
    Continent_dummy + newDataset$Cigarette.consumption.... *
    Continent_dummy, data = newDataset)

Residuals:
     Min       1Q   Median       3Q      Max
-12.5641  -1.9978  -0.0558   2.4290   9.1376

Coefficients:
                                                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                                                      7.127e+01  3.326e+00  21.429  < 2e-16 ***
newDataset$Outdoor.air.pollution....                                             6.127e-03  4.178e-02   0.147  0.88377
newDataset$HIV...Estimated.number.of.people.that.have.been.infected             -3.589e-07  5.120e-07  -0.701  0.48536
newDataset$malaria...Estimated.number.of.people.that.have.been.infected         -2.535e-04  2.784e-04  -0.911  0.36511
newDataset$Average.income.per.person....                                         2.737e-04  4.309e-05   6.353 1.12e-08 ***
newDataset$Cigarette.consumption....                                            -9.058e-02  1.040e-01  -0.871  0.38629
Continent_dummy2                                                                -1.061e+01  3.975e+00  -2.669  0.00916 **
Continent_dummy3                                                                 6.281e+00  6.226e+00   1.009  0.31604
Continent_dummy4                                                                 3.248e+00  5.768e+00   0.563  0.57484
OECD_dummy1                                                                      6.218e-01  4.090e+00   0.152  0.87955
newDataset$Outdoor.air.pollution....:Continent_dummy2                            2.141e-02  5.586e-02   0.383  0.70250
newDataset$Outdoor.air.pollution....:Continent_dummy3                           -6.331e-02  2.613e-01  -0.242  0.80920
newDataset$Outdoor.air.pollution....:Continent_dummy4                           -1.029e-01  2.339e-01  -0.440  0.66102
newDataset$malaria...Estimated.number.of.people.that.have.been.infected:Continent_dummy2  2.278e-04  2.805e-04   0.812  0.41907
newDataset$malaria...Estimated.number.of.people.that.have.been.infected:Continent_dummy3 -1.266e-03  9.519e-03  -0.133  0.89449
newDataset$malaria...Estimated.number.of.people.that.have.been.infected:Continent_dummy4 -1.162e-01  5.622e-02  -2.067  0.04187 *
newDataset$Cigarette.consumption....:Continent_dummy2                            1.956e-01  1.348e-01   1.451  0.15062
newDataset$Cigarette.consumption....:Continent_dummy3                           -1.746e-01  2.560e-01  -0.682  0.49721
newDataset$Cigarette.consumption....:Continent_dummy4                            1.085e-01  1.911e-01   0.568  0.57167
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.075 on 82 degrees of freedom
Multiple R-squared:   0.69,	Adjusted R-squared:  0.6219
F-statistic: 10.14 on 18 and 82 DF,  p-value: 3.547e-14
```

נספח 9 – השוואה על פי מדד BIC

```
> #comparing the chosen model from AIC to the full
> BIC(Full)
[1] 641.6639
> BIC(AIC.model)
[1] 589.5998
```

נספח 10 – מקדמי המודל הנבחר

```
> bestModel <- AIC.model
> coefficients(bestModel)
                        (Intercept)                Continent_dummy2                Continent_dummy3                Continent_dummy4
                      68.8460274725                   -5.9726874321                    2.9881679395                    2.6604494895
newDataset$Average.income.per.person....
                       0.0003014502
> bw.model <- step(Full, direction = 'backward', scope = ~ 1)
```

נספח 11 – הנחות

```
> # Errors & Linear
> summary(bestModel)

Call:
lm(formula = newDataset$Life_expectancy ~ Continent_dummy + newDataset$Average.income.per.person....,
    data = newDataset)

Residuals:
     Min       1Q   Median       3Q      Max
-12.1747  -2.2859  -0.3223   2.8821  10.5666

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                              6.885e+01  8.476e-01  81.225  < 2e-16 ***
Continent_dummy2                        -5.973e+00  9.483e-01  -6.298 9.05e-09 ***
Continent_dummy3                         2.988e+00  1.388e+00   2.153   0.0338 *
Continent_dummy4                         2.660e+00  1.477e+00   1.802   0.0747 .
newDataset$Average.income.per.person....  3.014e-04  3.889e-05   7.752 9.46e-12 ***
---
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:  0.6344
F-statistic: 44.37 on 4 and 96 DF,  p-value: < 2.2e-16

> newDataset$fitted <- fitted(bestModel) #predicted values
> newDataset$residuals <- residuals(bestModel) #residuals
> s.r_res <- sqrt(var(newDataset$residuals)) #calculating the standard deviation of the errors
> newDataset$standardResiduals <- (residuals(bestModel)/s.r_res) #saving the data set after the calculations
> plot(newDataset$fitted, newDataset$standardResiduals, xlab = "Predicted Value", ylab = "Normalized Error")
> abline(0, 0)
```

```
# F test for equal vars - not relavent
Life_expectancy_vec <- newDataset$Life_expectancy
Life_expectancy_vec_sort <- sort(Life_expectancy_vec)
third_vec_length <- round (length(Life_expectancy_vec_sort)/3)
third23_vec_length <- round (length(Life_expectancy_vec_sort)*2/3)
third_vec_values <- Life_expectancy_vec_sort[1:third_vec_length] #third first vaues
third23_vec_values <-  Life_expectancy_vec_sort[third23_vec_length:length(Life_expectancy_vec_sort)]#third last values
var.test(x = third_vec_values, y = third23_vec_values, ratio = 1,
         alternative = c("two.sided"), conf.level = 0.95)
```

```
> #SW Test
> shapiro.test(newDataset$standardResiduals)

        Shapiro-Wilk normality test

data:  newDataset$standardResiduals
W = 0.98207, p-value = 0.1871

>
> #GQ Test- full or fwd.model, we need to put hear the finel model
> gqtest(bestModel)

        Goldfeld-Quandt test

data:  bestModel
GQ = 0.72147, df1 = 46, df2 = 45, p-value = 0.8631
alternative hypothesis: variance increases from segment 1 to 2
```

```
> #Chow Test- full or fwd.model, we need to put hear the finel model
> sctest(bestModel)

        M-fluctuation test

data:  bestModel
f(efp) = 0.89596, p-value = 0.9212
```

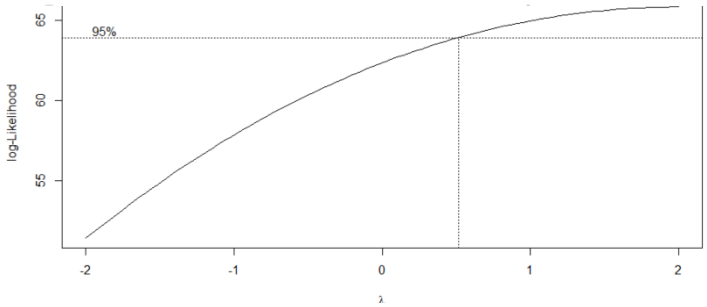
```
> # Normal
> qqnorm(newDataset$standardResiduals)
> abline(a = 0, b = 1)
> hist(newDataset$standardResiduals, prob = TRUE, xlab = "Normalized Error", main = "Histogram of Normalized Error")
> lines(density(newDataset$standardResiduals), col = "blue", lwd = 2)
>
> #KS Test
> ks.test(x = newDataset$standardResiduals, y ="pnorm", alternative = "two.sided", exact = NULL)

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  newDataset$standardResiduals
D = 0.071035, p-value = 0.688
alternative hypothesis: two-sided
```

מודל 12- מציאת למדא מתאימה לפי פונקציית Box Cox

```
> library(MASS)
> bc <- boxcox(newDataset$Life_expectancy ~ Continent_dummy + new_x_exact , data = newDataset)
> lambda <- bc$x[which.max(bc$y)] # Exact lambda
> lambda
[1] 2
```



מודל 13- אלגוריתמים

```
> # FORWARD
> fwd.model <- step(Emp, direction = 'forward', scope = formula(Full))
Start:  AIC=1237.33
newForm ~ 1

                                                                  Df Sum of Sq      RSS    AIC
+ Continent_dummy                                                  3   8964436 11744593 1186.0
+ newDataset$Average.income.per.person....                         1   7227711 13481318 1196.0
+ newDataset$malaria...Estimated.number.of.people.that.have.been.infected  1   2469066 18239963 1226.5
+ newDataset$HIV...Estimated.number.of.people.that.have.been.infected      1    966857 19742172 1234.5
+ newDataset$Outdoor.air.pollution....                             1    925157 19783872 1234.7
+ OECD_dummy                                                       1    805744 19903286 1235.3
+ newDataset$Cigarette.consumption....                             1    518541 20190488 1236.8
<none>                                                                         20709029 1237.3

Step:  AIC=1186.04
newForm ~ Continent_dummy

                                                                  Df Sum of Sq      RSS    AIC
+ newDataset$Average.income.per.person....                         1   4740347  7004246 1135.8
+ newDataset$malaria...Estimated.number.of.people.that.have.been.infected  1    341407 11403186 1185.1
<none>                                                                         11744593 1186.0
+ newDataset$HIV...Estimated.number.of.people.that.have.been.infected      1    165218 11579375 1186.6
+ OECD_dummy                                                       1     92764 11651829 1187.2
+ newDataset$Outdoor.air.pollution....                             1     48376 11696218 1187.6
+ newDataset$Cigarette.consumption....                             1     26002 11718591 1187.8

Step:  AIC=1135.84
newForm ~ Continent_dummy + newDataset$Average.income.per.person....

                                                                  Df Sum of Sq     RSS    AIC
<none>                                                                          7004246 1135.8
+ newDataset$HIV...Estimated.number.of.people.that.have.been.infected      1    116339 6887907 1136.2
+ newDataset$malaria...Estimated.number.of.people.that.have.been.infected  1     45944 6958302 1137.2
+ newDataset$Outdoor.air.pollution....                             1     14054 6990192 1137.6
+ newDataset$Cigarette.consumption....                             1       833 7003413 1137.8
+ OECD_dummy                                                       1       498 7003748 1137.8
```

```
> # BACKWARD
> bw.model <- step(Full, direction = 'backward', scope = ~ 1)
Start:  AIC=1149.83
newForm ~ newDataset$Outdoor.air.pollution.... + newDataset$HIV...Estimated.number.of.people.that.have.been.infected +
    newDataset$malaria...Estimated.number.of.people.that.have.been.infected +
    newDataset$Average.income.per.person.... + newDataset$Cigarette.consumption.... +
    Continent_dummy + OECD_dummy + newDataset$Outdoor.air.pollution.... "
    Continent_dummy + newDataset$malaria...Estimated.number.of.people.that.have.been.infected "
    Continent_dummy + newDataset$Cigarette.consumption.... "
    Continent_dummy

                                                                                  Df Sum of Sq     RSS    AIC
- newDataset$Outdoor.air.pollution....:Continent_dummy                             3     36205 6133427 1144.4
- OECD_dummy                                                                       1      2981 6100202 1147.9
- newDataset$HIV...Estimated.number.of.people.that.have.been.infected              1     40746 6137968 1148.5
- newDataset$Cigarette.consumption....:Continent_dummy                             3    301217 6398439 1148.7
<none>                                                                                          6097222 1149.8
- newDataset$malaria...Estimated.number.of.people.that.have.been.infected:Continent_dummy  3    414846 6512068 1150.5
- newDataset$Average.income.per.person....                                         1   3286646 9383868 1191.4

Step:  AIC=1144.43
newForm ~ newDataset$Outdoor.air.pollution.... + newDataset$HIV...Estimated.number.of.people.that.have.been.infected +
    newDataset$malaria...Estimated.number.of.people.that.have.been.infected +
    newDataset$Average.income.per.person.... + newDataset$Cigarette.consumption.... +
    Continent_dummy + OECD_dummy + newDataset$malaria...Estimated.number.of.people.that.have.been.infected:Continent_dummy +
    newDataset$Cigarette.consumption....:Continent_dummy

                                                                                  Df Sum of Sq     RSS    AIC
- OECD_dummy                                                                       1      3507 6136934 1142.5
- newDataset$Outdoor.air.pollution....                                             1     30743 6164170 1142.9
- newDataset$HIV...Estimated.number.of.people.that.have.been.infected              1     50090 6183517 1143.2
- newDataset$Cigarette.consumption....:Continent_dummy                             3    323611 6457038 1143.6
<none>                                                                                          6133427 1144.4
- newDataset$malaria...Estimated.number.of.people.that.have.been.infected:Continent_dummy  3    411547 6544974 1145.0
- newDataset$Average.income.per.person....                                         1   3778039 9911466 1190.9

Step:  AIC=1142.49
newForm ~ newDataset$Outdoor.air.pollution.... + newDataset$HIV...Estimated.number.of.people.that.have.been.infected +
    newDataset$malaria...Estimated.number.of.people.that.have.been.infected +
    newDataset$Average.income.per.person.... + newDataset$Cigarette.consumption.... +
    Continent_dummy + newDataset$malaria...Estimated.number.of.people.that.have.been.infected:Continent_dummy +
    newDataset$Cigarette.consumption....:Continent_dummy

                                                                                  Df Sum of Sq     RSS    AIC
- newDataset$Outdoor.air.pollution....                                             1     29040 6165974 1141.0
- newDataset$HIV...Estimated.number.of.people.that.have.been.infected              1     50692 6187626 1141.3
- newDataset$Cigarette.consumption....:Continent_dummy                             3    321285 6458219 1141.6
<none>                                                                                          6136934 1142.5
- newDataset$malaria...Estimated.number.of.people.that.have.been.infected:Continent_dummy  3    409065 6545999 1143.0
- newDataset$Average.income.per.person....                                         1   3800493 9937427 1189.2

Step:  AIC=1140.96
newForm ~ newDataset$HIV...Estimated.number.of.people.that.have.been.infected +
    newDataset$malaria...Estimated.number.of.people.that.have.been.infected +
    newDataset$Average.income.per.person.... + newDataset$Cigarette.consumption.... +
    Continent_dummy + newDataset$malaria...Estimated.number.of.people.that.have.been.infected:Continent_dummy +
    newDataset$Cigarette.consumption....:Continent_dummy

                                                                                  Df Sum of Sq      RSS    AIC
- newDataset$HIV...Estimated.number.of.people.that.have.been.infected              1     66608  6232582 1140.0
- newDataset$Cigarette.consumption....:Continent_dummy                             3    323926  6489900 1140.1
<none>                                                                                           6165974 1141.0
- newDataset$malaria...Estimated.number.of.people.that.have.been.infected:Continent_dummy  3    392774  6558748 1141.2
- newDataset$Average.income.per.person....                                         1   3897476 10063450 1188.4

Step:  AIC=1140.05
newForm ~ newDataset$malaria...Estimated.number.of.people.that.have.been.infected +
    newDataset$Average.income.per.person.... + newDataset$Cigarette.consumption.... +
    Continent_dummy + newDataset$malaria...Estimated.number.of.people.that.have.been.infected:Continent_dummy +
    newDataset$Cigarette.consumption....:Continent_dummy

                                                                                  Df Sum of Sq      RSS    AIC
- newDataset$Cigarette.consumption....:Continent_dummy                             3    318337  6550919 1139.1
<none>                                                                                           6232582 1140.0
- newDataset$malaria...Estimated.number.of.people.that.have.been.infected:Continent_dummy  3    427415  6659997 1140.8
- newDataset$Average.income.per.person....                                         1   3886009 10118591 1187.0

Step:  AIC=1139.08
newForm ~ newDataset$malaria...Estimated.number.of.people.that.have.been.infected +
    newDataset$Average.income.per.person.... + newDataset$Cigarette.consumption.... +
    Continent_dummy + newDataset$malaria...Estimated.number.of.people.that.have.been.infected:Continent_dummy

                                                                                  Df Sum of Sq      RSS    AIC
- newDataset$Cigarette.consumption....                                             1      3010  6553929 1137.1
<none>                                                                                           6550919 1139.1
- newDataset$malaria...Estimated.number.of.people.that.have.been.infected:Continent_dummy  3    407316  6958235 1139.2
- newDataset$Average.income.per.person....                                         1   4172587 10723506 1186.9

Step:  AIC=1137.13
newForm ~ newDataset$malaria...Estimated.number.of.people.that.have.been.infected +
    newDataset$Average.income.per.person.... + Continent_dummy +
    newDataset$malaria...Estimated.number.of.people.that.have.been.infected:Continent_dummy

                                                                                  Df Sum of Sq      RSS    AIC
<none>                                                                                           6553929 1137.1
- newDataset$malaria...Estimated.number.of.people.that.have.been.infected:Continent_dummy  3    404372  6958302 1137.2
- newDataset$Average.income.per.person....                                         1   4169866 10723795 1184.9
```

```
> # STEPWISE
> sw.model <- step(Emp, direction = 'both', scope = formula(Full))
Start:  AIC=1237.33
newForm ~ 1

                                                                    Df Sum of Sq      RSS    AIC
+ Continent_dummy                                                    3   8964436 11744593 1186.0
+ newDataset$Average.income.per.person....                           1   7227711 13481318 1196.0
+ newDataset$malaria...Estimated.number.of.people.that.have.been.infected  1   2469066 18239963 1226.5
+ newDataset$HIV...Estimated.number.of.people.that.have.been.infected      1    966857 19742172 1234.5
+ newDataset$Outdoor.air.pollution....                               1    925157 19783872 1234.7
+ OECD_dummy                                                         1    805744 19903286 1235.3
+ newDataset$Cigarette.consumption....                               1    518541 20190488 1236.8
<none>                                                                           20709029 1237.3

Step:  AIC=1186.04
newForm ~ Continent_dummy

                                                                    Df Sum of Sq      RSS    AIC
+ newDataset$Average.income.per.person....                           1   4740347  7004246 1135.8
+ newDataset$malaria...Estimated.number.of.people.that.have.been.infected  1    341407 11403186 1185.1
<none>                                                                           11744593 1186.0
+ newDataset$HIV...Estimated.number.of.people.that.have.been.infected      1    165218 11579375 1186.6
+ OECD_dummy                                                         1     92764 11651829 1187.2
+ newDataset$Outdoor.air.pollution....                               1     48376 11696218 1187.6
+ newDataset$Cigarette.consumption....                               1     26002 11718591 1187.8
- Continent_dummy                                                    3   8964436 20709029 1237.3

Step:  AIC=1135.84
newForm ~ Continent_dummy + newDataset$Average.income.per.person....

                                                                    Df Sum of Sq      RSS    AIC
<none>                                                                            7004246 1135.8
+ newDataset$HIV...Estimated.number.of.people.that.have.been.infected      1    116339  6887907 1136.2
+ newDataset$malaria...Estimated.number.of.people.that.have.been.infected  1     45944  6958302 1137.2
+ newDataset$Outdoor.air.pollution....                               1     14054  6990192 1137.6
+ newDataset$Cigarette.consumption....                               1       833  7003413 1137.8
+ OECD_dummy                                                         1       498  7003748 1137.8
- newDataset$Average.income.per.person....                           1   4740347 11744593 1186.0
- Continent_dummy                                                    3   6477072 13481318 1196.0
```

נספח 14- R square adjusted

```
> finalModel <- lm(formula = newForm ~ newDataset$Average.income.per.person.... + Continent_dummy, data = newDataset)
> summary.lm(finalModel)

Call:
lm(formula = newForm ~ newDataset$Average.income.per.person.... +
    Continent_dummy, data = newDataset)

Residuals:
    Min      1Q  Median      3Q     Max
-714.41 -145.23  -33.47  187.32  740.69

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                            2.381e+03  5.713e+01  41.680  < 2e-16 ***
newDataset$Average.income.per.person....  2.113e-02  2.621e-03   8.060 2.12e-12 ***
Continent_dummy2                      -4.030e+02  6.392e+01  -6.305 8.77e-09 ***
Continent_dummy3                       2.181e+02  9.354e+01   2.332   0.0218 *
Continent_dummy4                       1.942e+02  9.952e+01   1.952   0.0539 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 270.1 on 96 degrees of freedom
Multiple R-squared:  0.6618,    Adjusted R-squared:  0.6477
F-statistic: 46.96 on 4 and 96 DF,  p-value: < 2.2e-16
```

נספח 15- מקדמי המודל הסופי לאחר הטרנספורמציה

```
> coefficients(finalModel)
                                (Intercept)
                              2381.16372222
newDataset$Average.income.per.person....
                               0.02112584
                          Continent_dummy2
                             -403.00889020
                          Continent_dummy3
                              218.13782895
                          Continent_dummy4
                              194.24356281
```