דוח פרויקט ברגרסיה

פרויקט בניתוח סטטיסטי של מאגרי

נתונים טבלאיים - חלק ב'

<u>מגישות:</u>

קבוצה 11

<u>תאריך הגשה:</u>

29.06.17

תוכן עניינים

תקציר מנהלים 4

עיבוד מקדים 5

1.הגדרת משתנים 5

הסבר על משתנים 6

2.הסרה של משתנים 6

מדד פירסון (מקדם המתאם של פירסון) 6

בחינת התאמה בין משתנים מסבירים למשתנה המוסבר 7

בחינת התאמה בין משתנים מסבירים למשתנים מסבירים 7

3.התאמת המשתנים 8

משתנים קטגוריאליים 8

משתנים רציפים 8

4.הגדרת משתני דמה 8

5.הגדרת משתני אינטראקציה 9

התאמת המודל ובדיקת הנחות המודל 12

1.בחירת משתני המודל (נספח 7) 12

1. Forward Selection 12

2. Backward Elimination 13

3. Stepwise Regression 13

ניתוח תוצאות האלגוריתמים 13

2.בדיקת הנחות המודל 14

הנחת הליניאריות 14

הנחת שוויון שונויות 14

הנחת הנורמליות של השגיאות 14

Kolmogorov-Smirnov (KS) 15

Shapiro-Wilk (SW) 15

3.דוגמא לשימוש במודל הנבחר 15

4.בדיקת השערות המודל 16

ניסוח השערות המודל 16

שיפור המודל (נספח 9) 17

מסקנות והמלצות 18

נספחים : 19

נספח 1 : תרשימי פיזור – משתנים מסבירים ומשתנה מוסבר 19

איור 1 - תרשים פיזור מספר רצחות ותוחלת חיים 19
איור 2 - תרשים פיזור מספר רצחות ומדד ג'יני 19
איור 3 - תרשים פיזור מספר רצחות ומדד GDP 20
איור 4 - תרשים פיזור מספר רצחות ודת עיקרית 20
איור 5 - תרשים פיזור מספר רצחות וצריכת אלכוהול 20
איור 6 - תרשים פיזור מספר רצחות וצפיפות האוכלוסייה 21
איור 7 - תרשים פיזור מספר רצחות ושיעור האוכלוסייה מתחת לקו העוני 21
איור 8- תרשים פיזור מספר רצחות ושיעור המעשנים 21
איור 9 - תרשים פיזור מספר רצחות והיתר עונש מוות 22
איור 10 - תרשים פיזור מספר רצחות ושיעור גירושים 22
איור 11 - תרשים פיזור מספר רצחות וטמפרטורה ממוצעת 22
איור 12 - תרשים פיזור מספר רצחות ושיעור שטחים ירוקים 23
נספח 2 : מדד פירסון 23
איור 13 טבלת קשרי פירסון 24
נספח 3 : תרשימי פיזור של המשתנים המסבירים בינם לבין עצמם 24
איור 14 תרשימי פיזור משתנים מסבירים 24
נספח 4 : תרשים Box Plot משתנה מסביר 'דת עיקרית' 24
איור 15 – דת עיקרית 25
נספח 5 : משתני אינטראקציה – היתר עונש מוות 25
איור 16 תוחלת חיים והיתר עונש מוות – אינטרקציה 25
איור 17 GDP והיתר עונש מוות – אינטרקציה 26
איור 18 מדד ג'יני והיתר עונש מוות – אינטרקציה 26
איור 19 שיעור העוני והיתר עונש מוות – אינטרקציה 27
איור 20 שיעור הגירושים והיתר עונש מוות – אינטרקציה 27
נספח 6 : משתני אינטראקציה - דתות – הצגת השימוש ב-3 קטגוריות לעומת 2 קטגוריות 27
איור 21 תוחלת חיים ודתות - אינטראקציה 28
איור 22 מדד GDP ודתות - אינטראקציה 28
איור 23 מדד ג'יני ודתות – אינטראקציה 28
איור 24 שיעור העוני ודתות – אינטראקציה 29
איור 25 שיעור הגירושים ודתות – אינטראקציה 29
נספח 7 : בחירת משתני המודל 29
איור 26 Forward Selection - מבחן F חלקי 29
איור 27 Forward Selection - מבחן AIC 29
איור 28 Backward Selection - מבחן F חלקי 30

איור 29 Backward Selection - מבחן AIC 30
איור 30 Stepwise Regression מבחן AIC 30
נספח 8 : בדיקת הנחות המודל 31
איור 31 תרשים פיזור של השגיאות המתוקננות מול הערך הצפוי 31
איור 32 תרשים היסטוגרמה..... 31
איור 33 תרשים QQPlot 32
נספח 9 : שיפור המודל 33
איור 34 תרשים Box-Cox..... 33
איור 35 תרשים פיזור של השגיאות המתוקננות מול הערך הצפוי 33
איור 36- תרשים היסטוגרמה 34
איור 37 תרשים QQPlot 34
איור 38 טרנספורמציה על תוחלת חיים 35
איור 39 טרנספורמציה על מדד ג'יני 35
איור 40 טרנספורמציה על שיעור העוני 36
איור 41 טרנספורמציה על משתנה דת..... 36
איור42 טרנספורמציה על היתר עונש מוות..... 37
איור 43 BackWard - AIC 37
נספח 10 : גיבוי קוד 38

תקציר מנהלים

בפרויקט זה בחרנו לבחון את הקשר בין מספר משתנים לבין מדד הפשיעה במדינות שונות בעולם, הנמדד על ידי מספר הרציחות לכל 100,000 איש. מטרת הפרויקט הייתה ליצור מודל החוזה בצורה אופטימאלית את מדד הפשיעה בכל מדינה, להלן 'המשתנה המוסבר'. לשם כך, בחרנו ובחנו מספר משתנים מסבירים במטרה לבחון את השפעתם האפשרית על מדד הפשיעה:

- משתנים רציפים: תוחלת חיים, מדד ג'יני, תמ"ג (GDP), צפיפות האוכלוסייה, שיעור האוכלוסייה מתחת לקו העוני, אחוז מעשנים יומי, שיעור מקרי הגירושים, טמפרטורה ממוצעת ושיעור השטחים הירוקים במדינה.
- משתנים קטגוריאליים: דת עיקרית, צריכת אלכוהול והיתר עונש מוות במדינה.

תחילה הסרנו חמישה משתנים מסבירים אשר מקדם המתאם שלהם למול המשתנה המוסבר היה הנמוך ביותר.

לאחר מכן, מידלנו את המשתנים הקטגוריאליים. הגדרנו מחדש את המשתנה הקטגוריאלי 'דת עיקרית' תוך ביצוע איחוד קטגוריות, והמרנו אותו ואת המשתנה הקטגוריאלי הנוסף 'היתר עונש מוות במדינה' למשתני דמה ולאחר מכן, בחירת משתני אינטראקציה רלוונטיים.

על פי שלושת השיטות שנלמדו לבחירת משתני המודל, נבחנו מספר מודלים ונערכו השוואות ביניהם כאשר המדד המרכזי שנבחן הינו R^2_{adj}. לבסוף נבחר המודל הטוב ביותר על פי מדד זה שהתקבל בשיטת הרגרסיה לאחור.

על המודל הנבחר, נבדקו שלוש הנחות הרגרסיה הליניארית: הנחת הלינאריות, שוויון שונויות והנחת הנורמליות על השגיאות. התקבל כי אף אחת מן ההנחות לא מתקיימת במודל.

ביצענו בדיקת השערה על מובהקות הרגרסיה באמצעות מבחן F. כלומר, האם קיים קשר בין המשתנה התלוי (המוסבר) לבין לפחות אחד מהמשתנים המסבירים והגענו למסקנה שלפחות אחד מהמשתנים המסבירים משפיעים על המשתנה המוסבר באופן מובהק. בנוסף, הראנו דוגמא רלוונטית לשימוש במודל.

על מנת לשפר את המודל, ביצענו שימוש בטרנספורמציות שונות הן על המשתנה המוסבר והן על כל אחד מהמשתנים המסבירים שנותרו במודל, ביניהם נבחר משתנה הטרנספורמציה שמקסם את מדד R^2_{adj}. על המודל המשופר בוצעה בחינת מדד R^2_{adj} לאחר ביצוע אלגוריתם הרגרסיה לאחור לפי קריטריון AIC.

המודל הסופי שהתקבל לפני שלב שיפור המודל הוא:

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1}X_1 + \widehat{\beta_2}X_2 + \widehat{\beta_4}X_{4.2} + \widehat{\beta_7}X_7 + \widehat{\beta_9}X_{9.2} + \widehat{\beta_{13}}X_{4.2}X_1$$

עיבוד מקדים

1. הגדרת משתנים

להלן משתני המודל, כפי שהוצגו בחלק א' של הפרויקט. קיים משתנה מוסבר אחד (Y – מדד הפשיעה במדינה) ושנים-עשר משתנים מסבירים (X).

סוג המשתנה	סימון	תיאור מילולי	סוג משתנה	יחידות מידה
מוסבר	Y	מדד הפשע	רציף	מספר מקרי הרצח על כל 100,000 איש
מסביר	X1	תוחלת חיים	רציף	שנים
מסביר	X2	מדד ג'יני	רציף	סקלה בין 0 ל-100 (0 מעיד על שוויון מוחלט ו-100 מעיד על היעדר שוויון מוחלט)
מסביר	X3	תמ"ג (GPD)	רציף	מונחי כסף
מסביר	X4	דת עיקרית במדינה	קטגוריאלי	• נצרות • אסלאם • בודהיזם • יהדות • אתיאיזם
מסביר	X5	צריכת אלכוהול	קטגוריאלי	ליטר לבן-אדם לשנה 1- נמוך (קטן מ5) 2- בינוני (בין 5-10) 3- גבוה (גדול מ10)
מסביר	X6	צפיפות אוכלוסייה	רציף	מספר בני אדם/2^(ק"מ)
מסביר	X7	שיעור האוכלוסייה מתחת לקו העוני	רציף	אחוזים
מסביר	X8	אחוז מעשנים יומי	רציף	אחוזים
מסביר	X9	היתר הוצאה לפועל של עונש מוות במדינה	קטגוריאלי	0 – חל איסור 1 – מאושר על פי חוק
מסביר	X10	שיעור מקרי הגירושים	רציף	אחוזים
מסביר	X11	טמפרטורה ממוצעת	רציף	מעלות צלסיוס
מסביר	X12	שיעור השטחים הירוקים במדינה	רציף	אחוזים

להלן המודל:

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1}X_1 + \widehat{\beta_2}X_2 + \widehat{\beta_3}X_3 + \widehat{\beta_4}X_4 + \widehat{\beta_5}X_5 + \widehat{\beta_6}X_6 + \widehat{\beta_7}X_7 + \widehat{\beta_8}X_8 + \widehat{\beta_9}X_9 + \widehat{\beta_{10}}X_{10} + \widehat{\beta_{11}}X_{11} + \widehat{\beta_{12}}X_{12}$$

הסבר על משתנים

1. **(Y) שיעור הפשיעה במדינה** - המדד נקבע לפי מספר הרציחות המתרחשות במדינה לכל 100,000 איש.
2. **(X_1) תוחלת החיים-** מדד סטטיסטי לתוחלת של הזמן הנותר לפרטים חיים מקבוצה נתונה להישאר בחיים.
3. **(X_2) מדד ג'יני -** מדד לאי שוויון בחלוקת ההכנסות.
4. **GDP per capita (X_3)** - תוצר מקומי גולמי לנפש. מדד המבטא את הערך הכולל של הסחורות והשירותים שיוצרו במדינה במהלך השנה הלוקח בחשבון את גודל האוכלוסייה.
5. **(X_4) הדת העיקרית** - מייצג את הדת אליה משויכים פלח האוכלוסייה הגדול ביותר באותה המדינה. הקטגוריות השונות הינן: נצרות, אסלאם וקטגוריה של דתות מיעוטים (במובן של מיעוט מדינות) הכוללת את: הדת היהודית (מדינה אחת), אתאיזם (מדינה אחת) ובודהיזם (שמונה מדינות). שתי האחרונות הינן בעלות אופי ומאפייני דת דומים ועל כן בוצע האיחוד.
6. **(X_5) צריכת אלכוהול** – כמות האלכוהול הממוצעת בליטרים שבן אדם שותה בשנה.
7. **(X_6) צפיפות האוכלוסיה** - מונח בתחום הדמוגרפיה, המציין את יחס האוכלוסייה לשטח עבור מרחב גיאוגרפי (נמדד במספר בני אדם לק"מ רבוע).
8. **(X_7) שיעור האוכלוסיה מתחת לקו העוני** - מדד חברתי כלכלי, המתייחס לרמות ההכנסה המינימאליות הנדרשות לאדם או משפחה.
9. **(X_8) אחוז מעשנים יומי** - יחס בין מספר האנשים במדינה אשר מעשנים סיגריות לבין כלל האוכלוסייה.
10. **(X_9) היתר הוצאה לפועל של עונש מוות** – האם במדינה מסוימת מותר לדון נאשמים במשפט לעונש מוות על פי חוק. 0 – מדינות בהן עונש מוות אסור, 1 – מדינות בהן עונש מוות מותר.
11. **(X_{10}) שיעור מקרי הגירושים** – אחוז הזוגות שעברו גירושים פורמליים מתוך כלל הזוגות הנשואים במדינה.
12. **(X_{11}) טמפרטורה ממוצעת** - טמפרטורה שנתית ממוצעת במדינה, נמדד במעלות צלזיוס.
13. **(X_{12}) שיעור השטחים הירוקים במדינה** – שיעור השטחים במדינה המוגדרים כשטחים ירוקים, כלומר שטחים מסוג "ריאה ירוקה" או שטחים פתוחים בתחומים עירוניים.

2. הסרה של משתנים

כדי לבדוק את המתאם בין כל אחד מהמשתנים המסבירים למשתנה המוסבר, השתמשנו בשני כלים עיקריים - מדד פירסון ותרשימי פיזור.

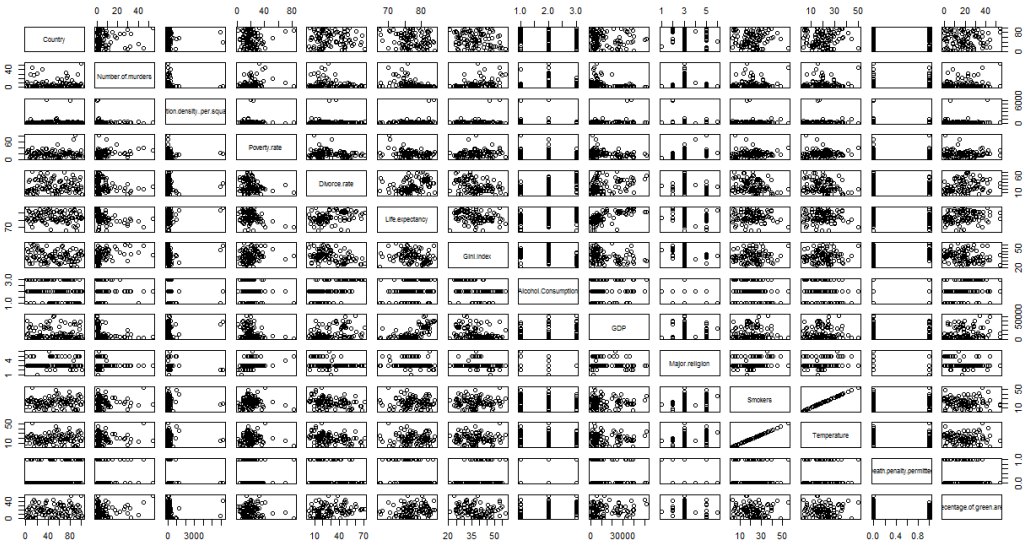
מדד פירסון (מקדם המתאם של פירסון)

מדד למתאם ליניארי בין שתי קבוצות של מספרים. במסגרת הפרויקט נרצה לבחון קורלציה בין שני משתנים, בעזרת המתאם הלינארי ביניהם. ערך המדד נע בין הערכים 1 ל- (1-). ככל שערכו המוחלט של המדד קרוב יותר ל-1, כך המדד יצביע על קשר לינארי חזק יותר בין המשתנים, ובמידה ומדובר בשני משתנים מסבירים, ייתכן כי אחד מהם מיותר. כמו כן, מקדם מתאם פירסון נמוך בערכו המוחלט (קרוב לאפס) של משתנה מסביר ביחס למשתנה מוסבר, מעיד על כך שמשתנה מסביר זה איננו מסביר טוב (יחסית) את המשתנה המוסבר ולכן גם אותו נבחר להסיר.

בחינת התאמה בין משתנים מסבירים למשתנה המוסבר

על סמך תרשימי הפיזור (נספח 1) ומטריצת הקשרים (correlation) לבחינת מדד פירסון (נספח 2) ניתן לראות כי התקבל מתאם נמוך בין המשתנה המוסבר Y (מספר רציחות לכל 100,000 איש במדינה) לבין המשתנים המסבירים שלהלן:

- X_5: צריכת אלכוהול – 0.067 בערך מוחלט.
- X_6: צפיפות האוכלוסייה - 0.103 בערך מוחלט.
- X_8: שיעור המעשנים- 0.048 בערך מוחלט.
- X_{11}: טמפרטורה ממוצעת – 0.048 בערך מוחלט.
- X_{12}: שיעור השטחים הירוקים במדינה– 0.117 בערך מוחלט.



המשתנה המסביר "צריכת אלכוהול" הינו קטגוריאלי ועל כן לא ניתן ללמוד מתרשים הפיזור של משתנה זה. בתרשימי הפיזור של ארבעת המשתנים המסבירים האחרים ניתן לראות כי לא קיים קשר ליניארי בינם לבין המשתנה המוסבר. כמו כן, ניתן לראות כי קו הרגרסיה הינו בעל שיפוע מתון מאוד וכי קיים פיזור גדול של התצפיות. על כן נוכל להסיק כי משתנים אלו אינם תורמים להבנת המשתנה המוסבר ולאור זאת, נבחר להסירם מהמודל.

בחינת התאמה בין משתנים מסבירים למשתנים מסבירים

על סמך תרשימי הפיזור (נספח 3) ומטריצת הקשרים (correlation) לבחינת מדד פירסון (נספח 2) ניתן לראות כי המתאם הגבוה ביותר הוא בין תוחלת חיים ומדד GDP וערכו 0.6696. ערך זה הינו גבוה יחסית ליתר מקדמי המתאם, אך אינו גבוה באופן קיצוני. בנוסף, לאור מיעוט משתנים מסבירים במודל לאחר ההסרה לעיל נבחר לא להסיר את המשתנים הנ"ל מהמודל.

להלן המודל המעודכן:

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1}X_1 + \widehat{\beta_2}X_2 + \widehat{\beta_3}X_3 + \widehat{\beta_4}X_4 + \widehat{\beta_7}X_7 + \widehat{\beta_9}X_9 + \widehat{\beta_{10}}X_{10}$$

3. התאמת המשתנים

משתנים קטגוריאליים

בעקבות חלק א' של הפרויקט, בחנו את משתנה "הדת עיקרית במדינה" והחלטנו כי נחוץ איחוד של המיעוטים לקבוצה קטגוריאלית אחת. הסקנו כי איחוד נכון יבוצע לפי קטגוריות בעלות השפעה דומה על המשתנה המוסבר.

ע"פ תרשים הBoxplot בנספח 4 ניתן לראות כי לדתות היהדות והאתאיזם קיימת השפעה דומה על כמות מקרי הרצח לכל 100,000 איש.

כלומר יצרנו שלוש קטגוריות עבור משתנה זה:

* יהדות ואתאיזם
* בודהיזם ואיסלאם
* נצרות

הערה: לאחר השלמת סעיפים 4 ו-5, התגלה צורך לאחד קטגוריות נוספות במשתנה קטגוריאלי זה. הסבר מפורט אודות האיחודים שביצענו ניתן לראות בסעיף 5.

בחנו את המשתנה הקטגוריאלי "היתר עונש מוות" ונוכחנו לראות כי הערכים שלו הם 0 או 1 (2 קטגוריות). בשתיהן השכיחויות גבוהות ולכן אין צורך בהתאמת משתנה זה.

בנוסף, את המשתנה צריכת אלכוהול החלטנו להסיר בסעיף הקודם ועל כן אין צורך בהתאמתו.

משתנים רציפים

לאחר בחינה של המשתנים הרציפים החלטנו כי לא נדרשת התאמה כלשהי. ניתן לראות כי המשתנים הרציפים הינם בעלי טווח ערכים מצומצם ועל כן לא קיימת הפרדה ברורה בין הערכים של המשתנים. כתוצאה מכך, בחרנו לא להפוך אף אחד מהמשתנים הרציפים לקטגוריאלי.

4. הגדרת משתני דמה

מטרת משתנה דמה הינה לציין את התרומה השולית של משתנים שלא נמצאים בקבוצת בסיס על החותך β_0. עבור שני המשתנים הקטגוריאליים, נגדיר משתני דמה על ידי הוספת משתנים כמספר הקטגוריות בכל משתנה פחות אחד. משתנים קטגוריאליים אינם מראים את הקשר הלינארי בצורה רצויה ולכן נגדירם באופן הבא:

1. **היתר עונש מוות במדינה** – משתנה זה מציין האם עונש מוות מאושר על פי החוק במדינה. כאשר קבוצת הבסיס הינה איסור על עונש מוות.

משתנה דמה	לא חוקי	חוקי
X9.2	0	1

$$\boldsymbol{approved} = \begin{cases} 1, if\ death\ penalty\ is\ approved \\ 0, else \end{cases}$$

2. **דת עיקרית** – את המשתנה הקטגוריאלי הזה חילקנו לשתי קטגוריות בלבד כאשר קבוצת הבסיס הינה נצרות . (הסיבה לחלוקה לשתי קטגוריות במקום ארבע מפורטת בהרחבה בסעיף 5).

משתנה דמה	1 – נצרות	2 – שאר הדתות
X4.2	0	1

$$\boldsymbol{Religion} = \begin{cases} 0, if \text{ Christianity is the main religion in the country} \\ 1\ , else \end{cases}$$

להלן המודל המעודכן:

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1}X_1 + \widehat{\beta_2}X_2 + \widehat{\beta_3}X_3 + \widehat{\beta_4}X_{4.2} + \widehat{\beta_7}X_7 + \widehat{\beta_9}X_{9.2} + \widehat{\beta_{10}}X_{10}$$

5. הגדרת משתני אינטראקציה

משתנה אינטראקציה הינו משתנה המציין את התרומה השולית על השיפוע, עבור שילוב של שני משתנים מסבירים יחדיו, על המשתנה המוסבר . בסעיף זה נרצה למצוא את ההשפעה של המשתנים הקטגוריאליים על שיפוע קו הרגרסיה עבור שילוב עם משתנים מסבירים שונים.

ביצענו בדיקה עבור כל משתני הדמה מול כל המשתנים המסבירים (לאחר שלב ההסרה) על ידי תרשימי פיזור (נספחים 5 + 6). בחרנו את אלו שבשילוב עם אחד מהמשתנים המסבירים הרציפים הייתה להם השפעה על המשתנה המוסבר, וכך ניתן לראות את התרומה השולית של משתנה הדמה לשיפוע.

1. משתנה קטגוריאלי: אישור עונש מוות (נספח 5)

חום – 0 (עונש מוות אסור במדינה), כחול - 1 (עונש מוות מאושר במדינה)

1. **תוחלת חיים -** תוחלת החיים היא משתנה מסביר למספר מקרי הרצח. (כיוון שתוחלת חיים גבוהה מעידה על מדינה מפותחת). בגרף זה החיתוך עם ציר ה-y הינו x=65, כלומר אין משמעות לתוחלת חיים אפס. ניתן לראות מגמה זהה בשני קווי הרגרסיה (כשמשתנה הדמה 0 וגם כשהוא 1)– יחס הפוך בין מס' הרציחות לבין תוחלת החיים, בין אם עונש מוות אסור או מאושר במדינה. כמו כן, במדינות בהן יש עונש מוות מס' הרציחות גדול יותר כאשר תוחלת החיים נמוכה, וקטן יותר כאשר תוחלת החיים גבוהה יחסית (מעל 80).

   משתנה האינטראקציה ייקרא:

$$deathPenalty * X_1 = \begin{cases} deathPenalty * X_1, & if\ death\ penalty\ is\ approved \\ 0 & , else \end{cases}$$

   המשמעות של משתנה אינטראקציה זה הינו תוחלת החיים במדינה במידה ובמדינה הדת העיקרית הינה נצרות, אחרת ערכו יהיה אפס (במדינה בה הדת העיקרית איננה נצרות).

2. **שיעור העוני** – שיעור העוני הוא משתנה מסביר למספר מקרי הרצח.

לפי השיפועים ניתן לראות כי במדינות בהן עונש מוות אסור ישנו יחס ישר בין מס' הרציחות לשיעור העוני, כלומר ככל ששיעור העוני עולה כך גם מס' הרציחות גדל, ואילו במדינות בהן עונש מוות מותר, אין תלות משמעותית בין המשתנים הללו – ניתן לראות ששיפוע הגרף הכחול שואף לאפס, ולכן נסיק כי אין השפעה בין שיעור העוני לבין מס' הרציחות במדינות בהן מאושר עונש מוות.

   משתנה האינטראקציה ייקרא:

$$deathPenalty * X_7 = \begin{cases} deathPenalty * X_7, & if\ death\ penalty\ is\ approved \\ 0 & , else \end{cases}$$

   המשמעות של משתנה אינטראקציה זה הינו שיעור העוני במדינה במידה ובמדינה קיים עונש מוות, אחרת ערכו יהיה אפס (במדינה חל איסור על עונש מוות).

3. **מדד ג'יני:** ניתן לראות בגרף זה כי החותכים קרובים (ושניהם שליליים) והשיפועים דומים מאוד, ועל כן אין הבדל כמעט בהשפעה של מדד ג'יני על מספר מקרי הרצח בין מדינות שעונש המוות בהן מותר או אסור. לכן לא נוסיף משתנה אינטראקציה.

4. **שיעור גירושין:** ניתן לראות בגרף זה כי ההפרש בין החותכים קטן וכי השיפועים יחסית דומים. על כן, אין הבדל משמעותי בהשפעה של שיעור הגירושין על מספר הרציחות בין מדינות שעונש המוות בהן מותר או אסור. לא נוסיף משתנה אינטראקציה.

5. **GDP:** ניתן לראות בגרף זה כי החותכים קרובים והשיפועים זהים, ועל כן אין הבדל בהשפעה של GDP על מספר הרציחות בין מדינות שעונש המוות בהן מותר או אסור. לא נוסיף משתנה אינטראקציה.

2. **<u>משתנה קטגוריאלי: דתות (נספח 6)</u>**

אדום-1(יהדות ואתאיזם), ירוק–2 (בודהיזם ואיסלם), סגול – 3(נצרות)

להלן מוצגת החלוקה הראשונית שביצענו עבור המשתנה הקטגוריאלי 'דתות'. על פי הגרפים המוצגים בנספחים ניתן לראות כי בכל חמשת הגרפים התנהגותם של קווי הרגרסיה האדום והירוק דומה ובחלק מהגרפים אף זהה לחלוטין. כלומר, בגרפים ישנם שני קווי רגרסיה בעלי שיפוע כמעט זהה וקו רגרסיה נוסף בעל שיפוע חד משמעותית ושונה משני הקווים הראשונים. לכן הבנו כי נדרש לאחד את קטגוריות היהדות והאיסלם לקטגוריה משותפת. בעקבות זאת ביצענו עדכון של סעיף 4 בתרגיל. להלן מוצגת החלוקה החדשה:

ירוק -1(else) , סגול–0 (נצרות)

1. **תוחלת חיים:** ניתן לראות כי עבור מדינות בהן הנצרות היא אינה הדת העיקרית מספר הרציחות הינו אפסי (מודל רגרסיה ללא שיפוע – השיפוע שואף לאפס).
לפי השיפועים ניתן לראות כי במדינות בהן הדת העיקרית הינה נצרות ישנו יחס הפוך (בעל מתאם שלילי) בין מס' הרציחות לתוחלת החיים, כלומר ככל שתוחלת החיים עולה כך מס' הרציחות קטן, ואילו במדינות בהן הנצרות אינה הדת העיקרית במדינה, אין תלות בין המשתנים הללו – ניתן לראות ששיפוע הגרף הירוק שואף לאפס, ולכן נסיק כי אין השפעה בין תוחלת החיים לבין מס' הרציחות במדינות מסוג זה.

משתנה האינטראקציה ייקרא:

$$religion * X_1 = \begin{cases} religion * X_1, & if\ main\ religion\ in\ country\ is\ Christianity \\ 0, else \end{cases}$$

משמעות של משתנה אינטראקציה זה הינו תוחלת החיים במדינה במידה ובמדינה הדת העיקרית הינה נצרות, אחרת ערכו יהיה אפס (במדינה בה הדת העיקרית איננה נצרות).

2. **מדד ג'יני:** מדד ג'יני הוא משתנה מסביר למידת הפשיעה. לפי החותכים ניתן לראות כי במדינות בהן הנצרות הינה הדת העיקרית, מדד ג'יני הינו בעל ערך שלילי (לא קיימת משמעות למספר רציחות שלילי). בנוסף, עבור מדינות בהן הנצרות היא אינה הדת העיקרית מספר הרציחות הינו אפסי (מודל רגרסיה ללא שיפוע – השיפוע שואף לאפס).
לפי השיפועים ניתן לראות כי במדינות בהן הדת העיקרית היא נצרות ישנו יחס ישר (בעל מתאם חיובי) בין מס' הרציחות למדד ג'יני, כלומר ככל שמדד ג'יני עולה כך גם מס' הרציחות גדל, ואילו במדינות בהן הנצרות אינה הדת העיקרית במדינה, אין תלות בין המשתנים הללו – ניתן לראות ששיפוע הגרף הירוק שואף לאפס, ולכן נסיק כי אין השפעה בין מדד ג'יני לבין מס' הרציחות במדינות מסוג זה.

משתנה האינטראקציה ייקרא:

$$religion * X_2 = \begin{cases} religion * X_2, & if\ main\ religion\ in\ country\ is\ Christianity \\ 0, else \end{cases}$$

משמעות של משתנה אינטראקציה זה הינו מדד ג'יני במדינה במידה ובמדינה הדת העיקרית הינה נצרות, אחרת ערכו יהיה אפס (במדינה בה הדת העיקרית איננה נצרות).

3. **GDP:** ניתן לראות בגרף זה כי החותכים והשיפועים יחסית קרובים, ועל כן אין הבדל בהשפעה של GDP על מספר הרציחות בין מדינות בהן הדת העיקרית הינה נצרות או אחרת. לא נוסיף משתנה אינטראקציה.

4. **<u>שיעור האוכלוסיה מתחת לקו העוני:</u>** שיעור העוני הוא משתנה מסביר למידת הפשיעה. לפי החותכים ניתן לראות כי במדינות בהן הנצרות הינה הדת העיקרית, שיעור העוני הוא בעל ערך שלילי (לא קיימת משמעות למספר רציחות שלילי). עבור מדינות בהן הנצרות היא אינה הדת העיקרית ההשפעה על מספר הרציחות הינה אפסית (מודל רגרסיה ללא שיפוע – השיפוע שואף לאפס). לפי השיפועים ניתן לראות כי במדינות בהן הדת העיקרית הינה נצרות ישנו יחס ישר (בעל מתאם חיובי) בין מס' הרציחות לשיעור העוני, ואילו במדינות בהן הנצרות אינה הדת העיקרית במדינה, אין תלות בין המשתנים הללו – ניתן לראות ששיפוע הגרף הירוק שואף לאפס, ולכן נסיק כי אין השפעה בין שיעור העוני במדינה לבין מס' הרציחות במדינות מסוג זה.

משתנה האינטראקציה ייקרא:

$$religion * X_7 = \begin{cases} religion * X_7, & if\ main\ religion\ in\ country\ is\ Christianity \\ 0, else \end{cases}$$

משמעות משתנה אינטראקציה זה הינו שיעור האוכלוסיה מתחת לקו העוני במדינה במידה והדת העיקרית בה היא נצרות, אחרת ערכו יהיה אפס (במדינה בה הדת העיקרית איננה נצרות).

5. **שיעור הגירושים :** ניתן לראות בגרף זה כי החותכים והשיפועים יחסית קרובים, ועל כן אין הבדל בהשפעה של שיעור הגירושים על מספר הרציחות בין מדינות בהן הדת העיקרית הינה נצרות או אחרת. לא נוסיף משתנה אינטראקציה.

הבהרה: מאחר ועל פי הנחיות עבודה זו אין צורך בהוספת כל משתני האינטראקציה האפשריים, מתוך החמישה נבחר לא להוסיף את משתנה ה $deathPenalty * X_1$ ואת משתנה ה $religion * X_2$ מאחר ושיפועם קטן יותר באופן יחסי לאחרים.

המודל החדש הוא:

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1}X_1 + \widehat{\beta_2}X_2 + \widehat{\beta_3}X_3 + \widehat{\beta_4}X_{4.2} + \widehat{\beta_7}X_7 + \widehat{\beta_9}X_{9.2} + \widehat{\beta_{10}}X_{10} + \widehat{\beta_{13}}X_{4.2}X_1 + \widehat{\beta_{14}}\,X_{4.2}X_7 + \widehat{\beta_{15}}X_{9.2}X_7$$

התאמת המודל ובדיקת הנחות המודל

1. בחירת משתני המודל (נספח 7)

לשם בחירת משתני המודל נשתמש בשלושה אלגוריתמים:

1. Forward Selection
2. Backward Elimination
3. Stepwise Regression

את שני האלגוריתמים הראשונים נבחן באמצעות מבחן F חלקי ו-AIC , ואת השלישי באמצעות מבחן AIC בלבד.

- **מבחן AIC** - בכל איטרציה נחשב את ערכי AIC עבור כל משתנה בנפרד, ונבחר במודל בעל ערך ה-AIC הקטן ביותר (כלומר, המובהק ביותר) והקטן מה-AIC של המודל מהשלב הקודם .נעצור כאשר ה-AIC הקטן ביותר המתקבל גדול או שווה ל-AIC של המודל מהשלב הקודם.
- **מבחן F חלקי** – הסבר למבחן זה מפורט בכל אחד מן האלגוריתמים.

מדד R^2 גדל ככל שמעלים את מספר המסבירים, ולכן את המודלים שהתקבלו נשווה על ידי מדד R^2_{adj} - מדד זה קונס על הוספת משתנים מיותרים. המדד מייצג את אחוז השונות המוסברת במודל תוך התמודדות עם ההשפעה המלאכותית של מספר המשתנים המסבירים על המדד, ונבחר במודל שערך המדד שלו הוא מקסימלי.

- **מדד R^2** - $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
- **מדד R^2_{adj}** - $R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$
- **מדד AIC** - $AIC = n\log(SSE/n) + 2(k+1)$
- **מדד BIC** - $BIC = n\log(SSE/n) + \log(n)(k+1)$
- **מדד C_p** - $C_p = \frac{SSE_{RM}}{MSE_{FM}} - (n-2p)$

1. Forward Selection

נתחיל במודל ללא משתנים. בכל איטרציה נוסיף למודל משתנה נוסף בעל ערך ה- F_{st} הגדול ביותר, כלומר המשתנה המובהק ביותר. לאחר מכן, נבדוק במבחן F חלקי איזה משתנים מובהקים ונכניס את המשתנה המובהק ביותר במבחן F.

מבחן F חלקי

בכל איטרציה נחשב סטטיסטיים עבור ערכי F חלקיים לכל משתנה שמועמד להיכנס למודל ונבחר את המשתנה בעל ערך F חלקי המקסימלי, F_m. אם $F_m > F^{1-\alpha}_{1,n-2}$ נוסיף את המשתנה הנ"ל ונמשיך לבצע איטרציות בדרך הנ"ל. אחרת, נעצור ונבחר במודל הקודם כמודל הסופי. (איור 26)

מבחן AIC – בכל איטרציה, נחשב את ערכי AIC עבור המודל המתקבל תוך כדי התחשבות בהשפעה של כל אחד מהמשתנים. המודל שנבחר יהיה בעל הערך ה- AIC הקטן ביותר והקטן מה-AIC של המודל מהשלב הקודם.

נפסיק כאשר ה-AIC הקטן ביותר המתקבל גדול או שווה ל-AIC של המודל מהשלב הקודם(איור 27).

2. Backward Elimination

נתחיל במודל מלא המכיל את כל המשתנים. נבדוק את המובהקות של כל אחד מהמשתנים, כאשר בכל איטרציה נוריד את המשתנה בעל ה- F_{st} החלקי הכי קטן. באלגוריתם זה נשאף שהSSE יהיה הגבוה ביותר.

מבחן F חלקי

בכל איטרציה, נבצע חישוב הסטטיסטיים עבור מבחני F החלקיים לכל משתנה מסביר בהנחה שהוא האחרון שנכנס למודל. נבחר את המשתנה בעל ערך F חלקי מינימלי FL.

אם $F_L < F_{1,n-k-1}^{1-\propto}$:

1) נוציא מהמודל את המשתנה המסביר שעבורו התקבל הערך
2) נחשב מודל המכיל את שאר המשתנים
3) נמשיך לבצע איטרציות בדרך הנ"ל

אחרת, נעצור ונבחר במודל הקודם כמודל הסופי (איור 28).

מבחן AIC – נבצע איטרציות בהן נסיר כל אחד מהמשתנים ונחשב את ערכי AIC עבור המודל המתקבל. נבחר את המודל עם ערך AIC הקטן ביותר וכל עוד הוא קטן מה-AIC של המודל מהשלב הקודם. כלומר נעצור כאשר ה-AIC הקטן ביותר המתקבל גדול או שווה ל-AIC של המודל מהשלב הקודם (איור 29).

3. Stepwise Regression

שיטה זו משלבת את העקרונות של שתי השיטות הקודמות. נתחיל ממודל ללא משתנים כלל, כאשר בכל איטרציה נבצע שתי פעולות. הראשונה, הוספה של משתנה נוסף ובדיקת המודל שהתגבש. השנייה, הוצאת משתנים אשר מתגלים בבדיקה כלא מתאימים. לאחר ביצוע האיטרציות נקבל את המודל בעל המשתנים המתאימים ביותר.

מבחן AIC – (איור 30)

ניתוח תוצאות האלגוריתמים

להלן מוצגת טבלה המרכזת את תוצאות הR_{adj}^2 של המבחנים השונים אותם הרצנו:

מבחן	Forward	Backward	Stepwise Regression
F חלקי	0.4102	0.4405	X
AIC	0.4405	0.4405	0.4405

לפי מבחן ה-AIC

ניתן לראות כי מדד ה-R_{adj}^2 בשלוש השיטות (רגרסיה לאחור, לפנים והלוך ושוב) זהה ושווה ל-0.4405, זאת מכיוון שקיבלנו בשלוש הדרכים את אותו המודל בעל אותם המשתנים המסבירים.

לפי מבחן F חלקי

ניתן לראות כי מדד ה- R_{adj}^2 המתקבל בשיטת הרגרסיה לאחור (0.4405) גדול יותר מאשר רגרסיה לאחור (0.4102) ושווה לערכי ה- R_{adj}^2 של הרצות הAIC (כיוון שגם שם מתקבל אותו המודל) ולכן נעדיף מודל זה.

המודל הסופי

המודל הסופי בו נבחר הוא המודל עם מדד ה-R_{adj}^2 הגבוה ביותר שהתקבל והוא בכל שיטות הרגרסיה לפי AIC וגם בשיטת רגרסיה לאחור במבחן F החלקי. הערך הינו 0.4405.

המודל המתקבל

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1}X_1 + \widehat{\beta_2}X_2 + \widehat{\beta_4}X_{4.2} + \widehat{\beta_7}X_7 + \widehat{\beta_9}X_{9.2} + \widehat{\beta_{13}}X_1X_{4.2}$$

לאחר הצבת ערכי האמדים

$$\hat{Y} = 64.3096 - 0.9566X_1 + 0.38623X_2 - 86.33197X_{4.2} + 0.15333X_7 + 4.79898X_{9.2} + 0.9997X_1X_{4.2}$$

2. <u>בדיקת הנחות המודל</u>

הנחת הליניאריות

ננתח את הנחת הלינאריות לפי **תרשים פיזור שגיאות (standardized residuals) ביחס ל- $\hat{Y}$ (Fitted Values)**, (נספח 8 – איור 31).

בבחינת תרשים פיזור זה נשים לב לשני קריטריונים מרכזיים:

1. <u>פיזור התצפיות בתרשים</u> – נצפה לראות פיזור אקראי של התצפיות בכל טווח התרשים. בתרשים שקיבלנו, ניתן לראות כי התצפיות אינן מפוזרות בצורה אקראית ובטווח ערכים רחב. מרבית התצפיות קרובות לקו ה-0 ולא קיים פיזור אקראי שלהן.
2. <u>תבנית הנתונים</u> – נצפה לראות פיזור של הנתונים כך שהריכוז הינו אחיד . בתרשים שקיבלנו, קיים ריכוז נתונים גדול עבור ערכים נמוכים ופיזור נמוך יותר עבור ערכים גבוהים יותר (מעל 10). לא ניתן להסיק אודות תבנית ספציפית של הנתונים אך ניתן לראות כי הן אינן מפוזרות כפי שאנו מצפים לראות.
לכן, ניתן להסיק **כי הנחת הלינאריות איננה מתקיימת במודל.**

הנחת שוויון שונויות

נבחן הנחה זו באמצעות תרשים פיזור שגיאות (standardized residuals) ביחס ל- $\hat{Y}$ (Fitted Values), על פי הגרף שהתקבל. במצב בו מתקיימת הנחת שוויון השונויות, נצפה לקבל פיזור אחיד של השגיאות סביב קו האפס.

בתרשים שקיבלנו, ניתן לראות כי השאריות אינן מפוזרות באופן אחיד סביב קו האפס, ויש מגמה כלשהי, על כן המודל הנבחר לא מתאים לנתונים ולכן לא ניתן להסיק כי השונות קבועה, כלומר, **הנחת שוויון השונויות אינה מתקיימת.**

הנחת הנורמליות של השגיאות

- **<u>תרשים היסטוגרמה</u>** – נבחן את התרשים ונבדוק האם מזכיר בצורתו את צורת ההתפלגות הנורמלית.
בתרשים שקיבלנו, הגרף מתנהג בצורה שמזכירה במידה מסוימת את ההתפלגות הנורמלית מכיוון שישנה עלייה ולאחריה ירידה, אך הערכים הגבוהים לא נמצאים בדיוק במרכז ויש סוג של זנב ימני (גם בזנב הירידה איננה קבועה לחלוטין). על מנת לוודא אם הנחת הנורמליות מתקיימת, נבדוק את תרשים הכמותונים הנורמלי ונבצע מבחנים סטטיסטים מתאימים. (נספח 8 – איור 32)
- **<u>תרשים כמותונים נורמאלי (QQplot) –</u>** בתרשים זה ציר ה-X מייצג את האחוזים של ההתפלגות הנורמלית, וציר ה-Y את השגיאות המתוקננות. הקו הלינארי מייצג את פיזור האחוזונים בהתפלגות הנורמלית. נשווה את השגיאות למול האחוזונים של ההתפלגות הנורמלית, כך שככל שהנקודות יהיו יותר קרובות לקו הלינארי, נסיק שהתפלגות השגיאות דומה יותר להתפלגות הנורמלית.
לפי התרשים ניתן לראות כי ישנו פיזור יחסי סביב קו ה- 45° אך לא בצורה מובהקת, ובעיקר בין אחוזון (1-) ל-1, כך שבקצוות ישנה התרחקות מהקו, (מצב בו הקצוות של הגרף מתבדרים הוא מצב לא תקין). כלומר, נסיק כי השגיאות לא מתפלגות נורמלית. (נספח 8 – איור 33).

משני התרשימים הנ"ל ניתן לקבוע כי **השגיאות אינן מתפלגות באופן נורמלי**, ועל כן לא נוכל לבצע מבחנים סטטיסטיים רגילים. לכן נבצע מבחנים סטטיסטיים א-פרמטרים (כאשר השערת האפס תהיה שהנתונים מתפלגים נורמלית):

Kolmogorov-Smirnov (KS)

```
> ks.test(x=dataset$stan_residuals, y="pnorm", alternative = "two.si
ded", exact=NULL)

        One-sample Kolmogorov-Smirnov test

data:  dataset$stan_residuals
D = 0.16948, p-value = 0.006041
alternative hypothesis: two-sided
```

Shapiro-Wilk (SW)

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

        Shapiro-Wilk normality test

data:  dataset$stan_residuals
W = 0.82387, p-value = 1.258e-09
```

קיבלנו בשני המבחנים כי P_value<0.05 ← נדחה את השערת האפס ונקבע כי **השגיאות המתוקננות אינן מתפלגות נורמלית.**

3. דוגמא לשימוש במודל הנבחר

לישראל והלל הציעו רילוקיישן לאיטליה. הם החליטו כי תחושת הביטחון האישי חשובה להם והם מעוניינים לעבור למדינה בעלת שיעור רציחות נמוך בלבד. מכיוון שאין נתונים על מדינה זו במאגר הנתונים בו השתמשנו, הן נעזרו במודל הרגרסיה המרובה שיצרנו.

נתוני אמת שידועים על איטליה:

- תוחלת החיים במדינה היא 81.7
- מדד הג'יני הוא 32.7
- הדת העיקרית במדינה היא נצרות, ולכן תקבל את הערך 0
- שיעור האוכלוסייה מתחת לקו העוני הוא 12.5%
- עונש המוות אינו מותר במדינה לכן יקבל את הערך 0
- משתנה האינטראקציה Major.ReligionLife.Expectancy וערכו 0

התוצאה שהתקבלה מהצבת נתוני האמת במודל הרגרסיה:

$$\hat{Y} = 64.3096 - 0.9566X_1 + 0.38623X_2 - 86.33197X_{4.2} + 0.15333X_7 + 4.79898X_{9.2} + 0.9997X_1X_{4.2}$$

$$\hat{Y} = 64.3096 - 0.9566 \cdot 81.7 + 0.38623 \cdot 32.7 - 86.33197 \cdot 0 + 0.15333 \cdot 12.5 + 4.79898 \cdot 0 - 0.9997 \cdot 0 = 0.7013$$

לפי מודל הרגרסיה, קיבלנו שמספר הרציחות במדינה הינו **0.7013** לכל 100,000 איש.

לפי האתר http://www.nationmaster.com/country-info/stats/Crime/Violent-crime/Murder-rate-per-million-people המתייחס לנתוני הרצח במדינות, במציאות מספר הרציחות באיטליה לכל 100,000 איש הוא **0.875**. כלומר, המודל נתן לנו תוצאה יחסית קרובה למציאות (הפרש של 0.1737 בערך מוחלט).

4. בדיקת השערות המודל

נבצע בדיקת השערה באמצעות מבחן F למקדמי הרגרסיה, על הקשר בין המשתנה המוסבר למשתנים המסבירים. מבחן זה בודק את מובהקות הרגרסיה, כלומר האם קיים קשר ליניארי בין המשתנה התלוי, מספר הרציחות על כל 100,000 אנשים, לבין לפחות אחד מהמשתנים המסבירים. השערת האפס גורסת כי מקדמי הרגרסיה שווים לאפס, דהיינו, לא קיים קשר בין המשתנה המוסבר לבין אף אחד מהמשתנים המסבירים.

ניסוח השערות המודל

$$H_0: \beta_1 = \beta_2 = \beta_4 = \beta_7 = \beta_9 = \beta_{14} = 0$$

$$H_1: else, at\ least\ one\ of\ the\ \beta\ is\ not\ \ 0$$

סטטיסטי המבחן: $F_{st} = \dfrac{MSR}{MSE}$

נדחה את H_0 אם $F_{st} \geq F_{k,n-k-1}^{1-\alpha}$.

במודל שלנו ישנם 6 משתנים מסבירים ועל כן k=6 והמדגם בגודל n=101.

מטבלת F הוצאנו את הערך הקריטי ברמת מובהקות של 5%.

$$F_{cr} = F_{k,n-k-1}^{0.95} = F_{6,94}^{0.95} \cong 2.3291$$

```
Call:
lm(formula = Number.of.murders ~ Life.expectancy + Gini.Index +
    Poverty.rate + religion.category + Death.penalty.permitted.category +
    Life.expectancy:religion.category, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-11.843  -3.903  -0.717   2.703  42.216

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                         64.30960   22.71248   2.831 0.005668 **
Life.expectancy                     -0.95660    0.27228  -3.513 0.000682 ***
Gini.Index                           0.38623    0.10518   3.672 0.000399 ***
Poverty.rate                         0.15333    0.07101   2.159 0.033365 *
religion.category1                 -86.33197   38.42345  -2.247 0.026989 *
Death.penalty.permitted.category1    4.79898    1.98301   2.420 0.017441 *
Life.expectancy:religion.category1   0.99970    0.49874   2.004 0.047897 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.698 on 94 degrees of freedom
Multiple R-squared:  0.4741,	Adjusted R-squared:  0.4405
F-statistic: 14.12 on 6 and 94 DF,  p-value: 2.114e-11
```

על פי האלגוריתם שהרצנו בסעיפים הקודמים ניתן לראות כי הערך הסטטיסטי ברמת מובהקות של 5% הינו $F_{st} = 14.12$.

$$F_{st} > F_{cr}$$

$$14.12 > 2.3291$$

קיבלנו שערך הסטטיסטי גדול מהערך הקריטי ולכן נגיד כי קיימים מספיק נתונים על מנת לדחות את השערת האפס ברמת מובהקות 5%.

כלומר, לפחות אחד מהמשתנים המסבירים במודל הוא מובהק בהשפעתו על המשתנה המוסבר.

```
> fit=lm(dataset$Number.of.murders ~ dataset$Gini.Index + dataset$religion.new + dat
expectancy)
> anova(fit)
Analysis of Variance Table

Response: dataset$Number.of.murders
                                          Df Sum Sq Mean Sq F value    Pr(>F)
dataset$Gini.Index                         1 2122.2 2122.15 35.8126 3.925e-08 ***
dataset$religion.new                       1 1166.7 1166.75 19.6896 2.475e-05 ***
dataset$Life.expectancy                    1 1006.8 1006.77 16.9899 8.095e-05 ***
dataset$Death.penalty.permitted            1  298.6  298.62  5.0394   0.02712 *
dataset$Poverty.rate                       1  188.9  188.89  3.1877   0.07742 .
dataset$religion.new:dataset$Life.expectancy 1  238.1  238.09  4.0179   0.04790 *
Residuals                                 94 5570.2   59.26
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

פירוט נוסף של חישוב סטטיסטי המבחן על פי טבלת ANOVA:

$$SSR = 2122.2 + 1166.7 + 1006.8 + 298.6 + 188.89 + 238.09 = 5021.28$$
$$df = 6$$
$$MSR = \frac{SSR}{df} = \frac{5021.28}{6} = 836.88$$

$$SSE = 5570.2$$
$$MSE = \frac{SSE}{df} = \frac{5570.2}{94} = 59.257$$
$$F_{st} = \frac{MSR}{MSE} = \frac{836.88}{59.257} = 14.12$$

שיפור המודל (נספח 9)

ראשית על מנת לטפל בחוסר שיוויון השונויות השתמשנו בטרנספורמציית החזקה של boxcox

$$y(\lambda) = \begin{cases} (y^{\lambda} - 1)/\lambda, & \lambda \neq 0 \\ \ln(y), & \lambda = 0 \end{cases}$$

שלאחריה מודל הרגרסיה יראה מהצורה:

$$y_i(\lambda) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i$$

לאחר הרצת הפקודה הרלוונטית בR (נספח 9 איור 34) ניתן לראות שברמת בטחון של 95% הלמבדה נמצאת בין 0.22 ל0.35, לכן בחרנו למדה 0.3 אשר ממקסמת את הנראות.
לאחר עדכון המודל בדקנו שוב את הנחות המודל במטרה לראות שיפור.

על פי תרשים השאריות המתוקננות ניתן לראות שהנחת שיוויון השונויות השתפרה ויחד איתה גם הנחת הליניאריות באופן משמעותי (נספח 9 איורים 35-37).
לאחר מכן בדקנו האם לאחד המסבירים מתאם גבוה כאשר הוא מועלה בריבוע (נספח 9 איורים 38-42).
את המתאם בחרנו לבדוק באמצעות מדד ה- R^2_{adj} המקסימלי שהתקבל:

עבור משתנה מסביר "תוחלת חיים" התקבל **0.2838.**
עבור משתנה מסביר "מדד ג'יני" התקבל 0.2392.
עבור משתנה מסביר "שיעור העוני" התקבל 0.0583.
עבור משתנה מסביר "דת עיקרית" התקבל 0.04322.
עבור משתנה מסביר "היתר עונש מוות" התקבל 0.05488.

ערך R^2_{adj} המקסימלי התקבל עבור המשתנה המסביר "תוחלת החיים" בטרנספורמציה של העלאה בריבוע, ולכן נבחר להוסיף משתנה זה למודל.
לאחר ההוספה ביצענו רגרסיה לאחור ע"פ קריטריון AIC , על מנת לבחון אם הוספת המשתנה המסביר הנ"ל וביצוע הטרנספורמציה על המשתנה המוסבר אכן שיפרה את המודל. להלן התוצאות בהשוואה על פי מדד R^2_{adj} :

- מודל לפני טרנספורמציה – שהתקבל לפי שיטת רגרסיה לאחור– 0.4055.
- מודל משופר שהתקבל לפי שיטת רגרסיה לאחור - 0.541.

ניתן לראות כי ביצוע הטרנספורמציה אכן שיפרה את המודל .

מסקנות והמלצות

מטרת חלק ב' של הפרויקט הייתה למצוא כיצד משתנים שונים משפיעים על מספר הרציחות על כל 100,000 איש ב-101 מדינות . המודל הראשוני כלל את כל המשתנים שנבחרו והוצגו בחלק א' של הפרויקט, מתוך מחשבה כי כל אחד מהם עשוי להשפיע על מספר הרציחות. תהליך ההתאמה ושיפור המודל התבצעו במספר שלבים.

- המודל הראשוני שהתקבל הינו

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1}X_1 + \widehat{\beta_2}X_2 + \widehat{\beta_3}X_3 + \widehat{\beta_4}X_4 + \widehat{\beta_5}X_5 + \widehat{\beta_6}X_6 + \widehat{\beta_7}X_7 + \widehat{\beta_8}X_8 + \widehat{\beta_9}X_9 + \widehat{\beta_{10}}X_{10} + \widehat{\beta_{11}}X_{11} + \widehat{\beta_{12}}X_{12}$$

- ביצענו עיבוד ראשוני על הנתונים אשר כלל:
  - בדיקת התאמה בין המשתנים המסבירים לבין המשתנה המוסבר וכן בין המשתנים המסבירים לבין עצמם. הוחלט להסיר חמישה משתנים מסבירים.
  - איחוד קטגוריות של המשתנה הקטגוריאלי 'דתות'.
  - הגדרת משתני דמה ומשתני אינטראקציה עבור המשתנים הקטגוריאליים.
- על מנת למצוא את מודל הרגרסיה המתאים ביותר השתמשנו בשלושת האלגוריתמים הבאים:
  - Forward selection
  - Backward elimination
  - Stepwise Regression

  עבור אלגוריתמים אלו ביצענו מבחן F חלקי ו-AIC. המודל הנבחר היה המודל בעל מדד הR^2_{adj} הגבוה ביותר. בסופו של דבר נבחר המודל בשיטת הרגרסיה לאחור. ניתן לראות כי מספר משתנים מסבירים הושמטו מן המודל משום שתרומתם לא הייתה מספיק משמעותית להסברת המשתנה המוסבר.
- ביצענו בחינה של הנחות המודל אשר כלל בדיקת:
  - שוויון שונויות
  - הנחת הלינאריות
  - התפלגות נורמאלית של השגיאות

התקבל כי אף אחת מההנחות לא מתקיימת.

- ביצענו תיקון למשתנה המוסבר באמצעות טרנספורמציית החזקה של boxcox, אשר לאחריו ניתן היה לראות שיפור משמעותי בקיום הנחות המודל.
- בחנו האם יש הצדקה בהוספת משתנה טרנספורמציה, על מנת לחזק את הקשר בין המשתנים המסבירים למשתנה המוסבר וזאת במטרה לשפר את המודל. בדקנו את מקדם המתאם של משתנה הטרנספורמציה עם המשתנה המסביר ובחרנו להכניס למודל את משתנה "תוחלת החיים" לאחר העלאתו בריבוע, לאור זאת שמקדם המתאם שלו הינו הגבוה ביותר מבין כל משתני הטרנספורמציה שנבחנו.

על מנת לחזות בצורה הטובה ביותר את כמות הרציחות במדינה מסוימת על כל 100,000 איש המלצותינו למחקר עתידי הינם:

א) הגדלת מאגר התצפיות עבור כל משתנה - בחירה במספר גדול של מדינות מאזורים שונים בעולם יוכל לספק מספר תצפיות רב שישקף בצורה טובה יותר את המצב הקיים ולספק מודל מבוסס ומדויק יותר.

ב) הוספת משתנים מסבירים נוספים למודל שיוכלו לתרום להשפעה על כמות הרציחות לכל 100,000 איש. משתנים לדוגמא הינם:

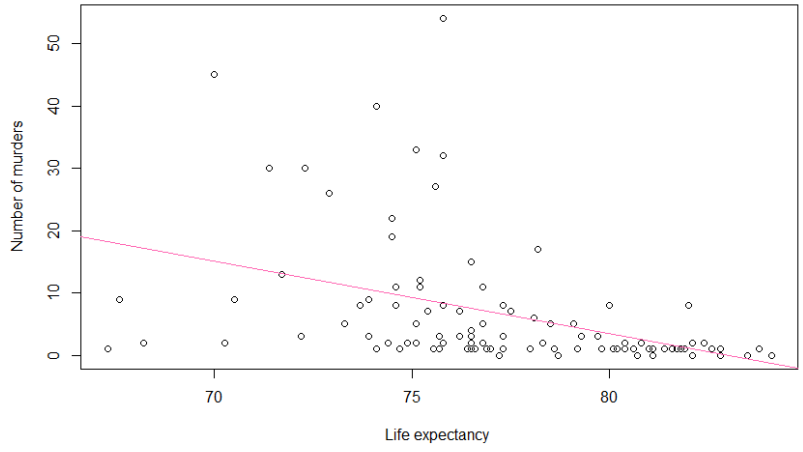
   a. משתנה רציף – אחוז בוגרי 12 שנות לימוד, גיל נישואים ממוצע, אחוז המועסקים.

   b. משתנה קטגוריאלי – מספר ימי חופשה במדינה (בחלוקה לתחומים).

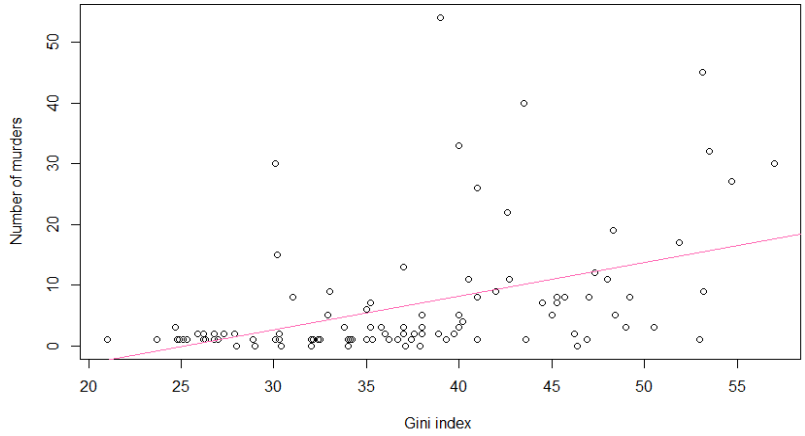
ג) חיפוש מקורות נוספים שונים לנתונים שהתקבלו.

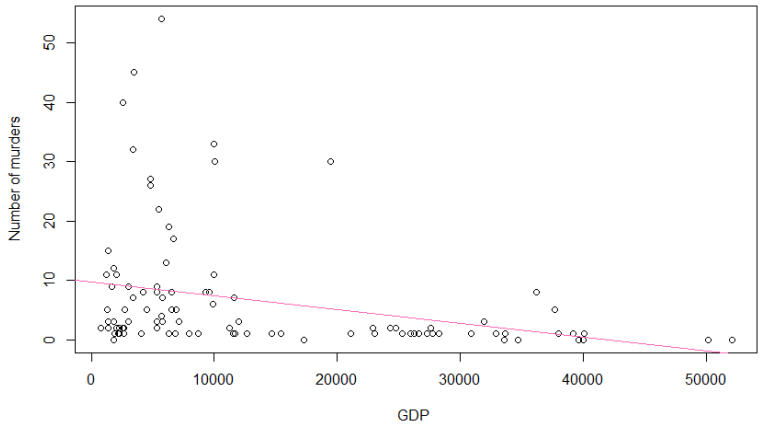
נספחים :

נספח 1: תרשימי פיזור – משתנים מסבירים ומשתנה מוסבר



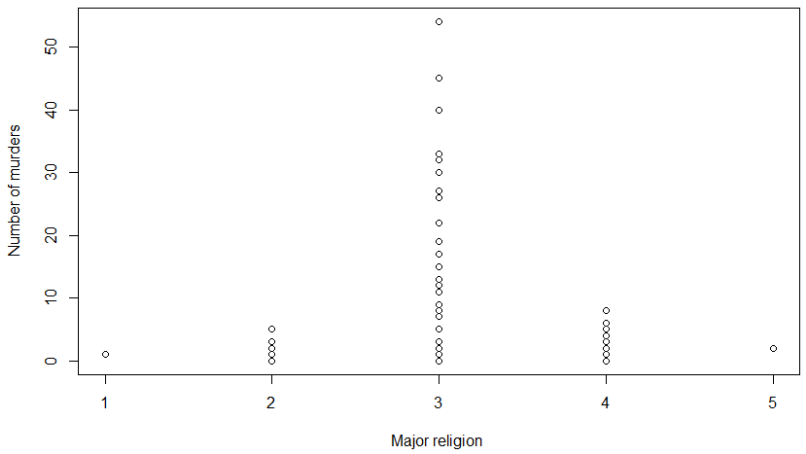
איור 1 - תרשים פיזור מספר רציחות ותוחלת חיים



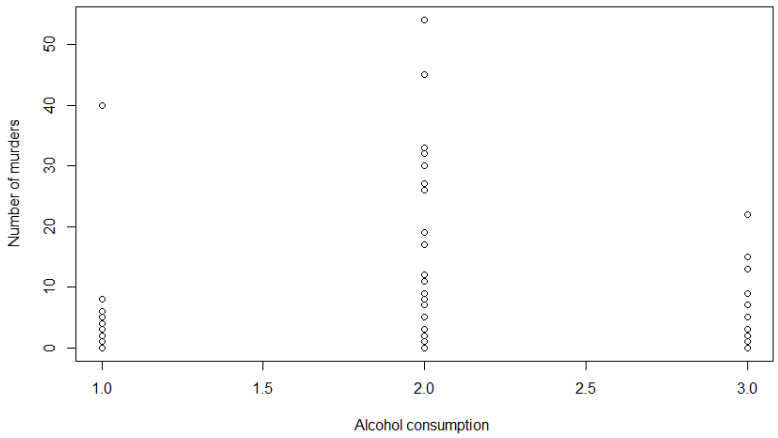


איור 2 - תרשים פיזור מספר רציחות ומדד ג'יני

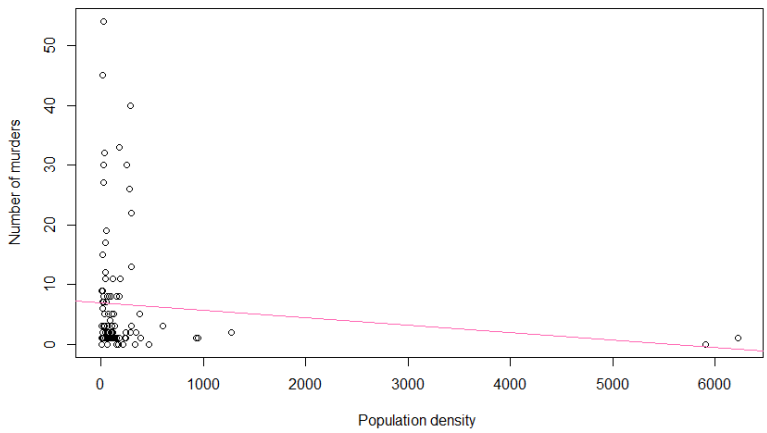
איור 3 - תרשים פיזור מספר רציחות ומדד GDP



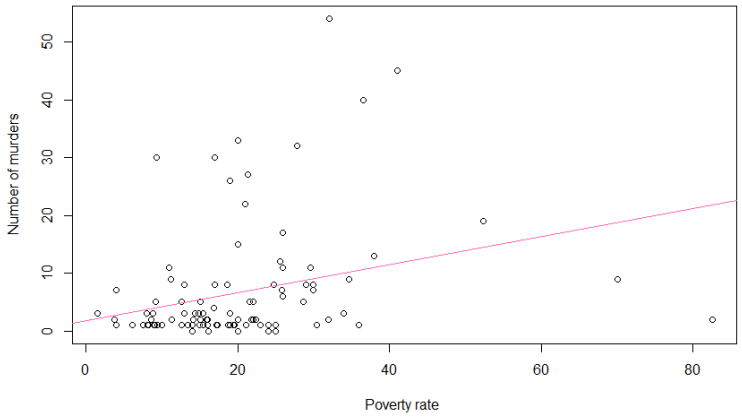
איור 4 - תרשים פיזור מספר רציחות ודת עיקרית



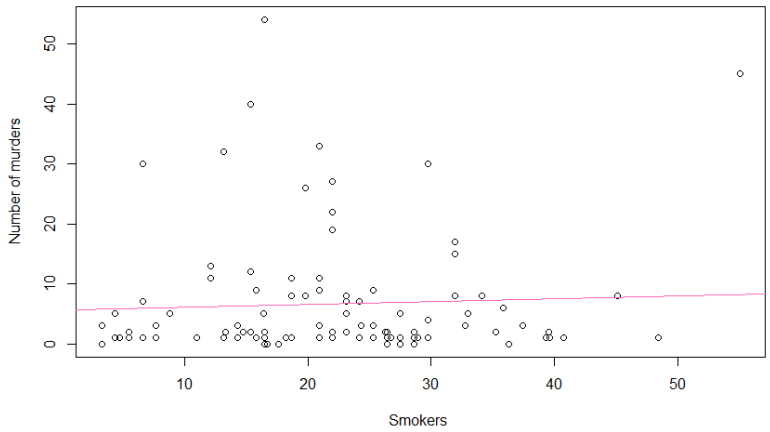
איור 5 - תרשים פיזור מספר רציחות וצריכת אלכוהול



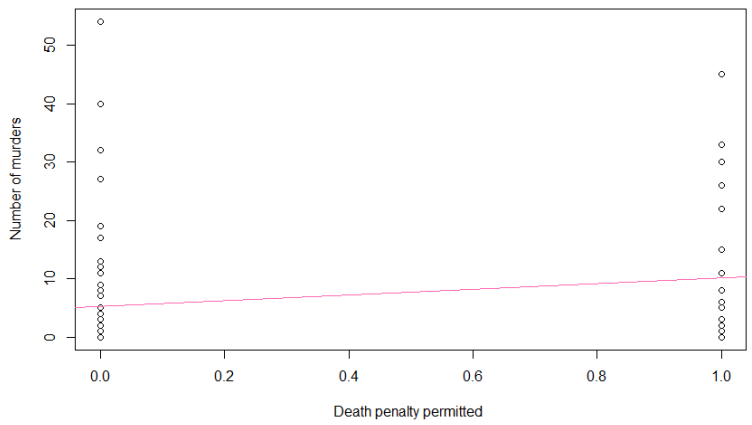
איור 6 - תרשים פיזור מספר רציחות וצפיפות האוכלוסייה



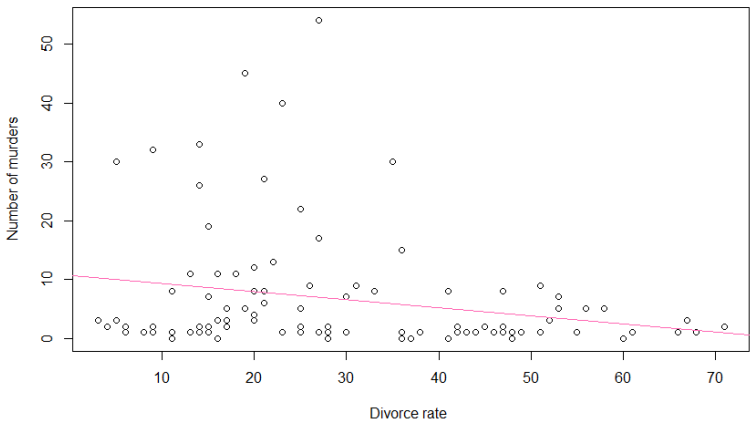
איור 7 - תרשים פיזור מספר רציחות ושיעור האוכלוסייה מתחת לקו העוני



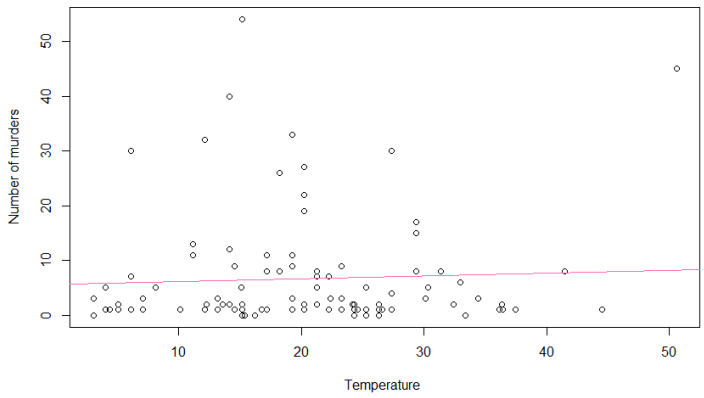
איור 8- תרשים פיזור מספר רציחות ושיעור המעשנים



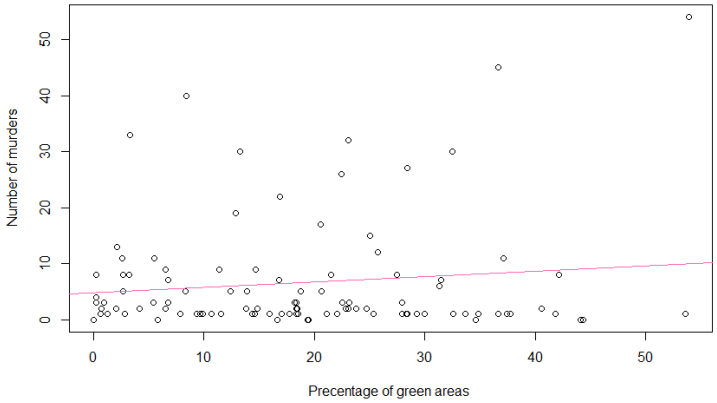
איור 9 - תרשים פיזור מספר רציחות והיתר עונש מוות



איור 10 - תרשים פיזור מספר רציחות ושיעור גירושים



איור 11 - תרשים פיזור מספר רציחות וטמפרטורה ממוצעת



איור 12 - תרשים פיזור מספר רציחות ושיעור שטחים ירוקים

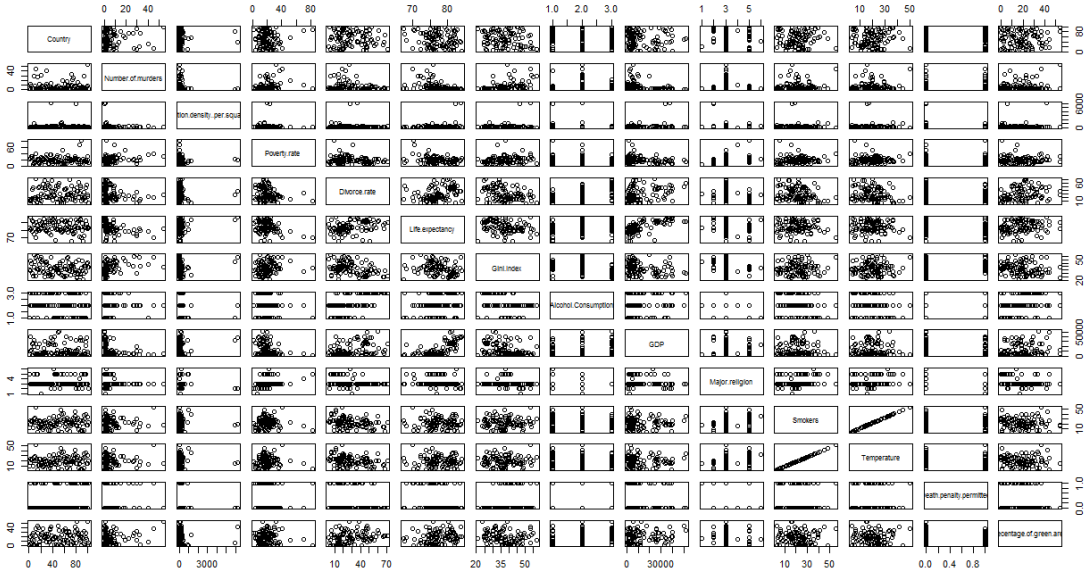
נספח 2: מדד פירסון

	Number.of.murders	Population.density..per.square.km.	Poverty.rate	Divorce.rate	Life.expectancy	Gini.Index
Number.of.murders	1.00000000	-0.10302219	0.281637785	-0.22888151	-0.41654916	0.44762107
Population.density..per.square.km.	-0.10302219	1.00000000	-0.013028985	-0.01174654	0.23914810	0.17770895
Poverty.rate	0.28163779	-0.01302899	1.000000000	-0.22093172	-0.33311345	0.19444165
Divorce.rate	-0.22888151	-0.01174654	-0.220931718	1.00000000	0.35383198	-0.39550136
Life.expectancy	-0.41654916	0.23914810	-0.333113454	0.35383198	1.00000000	-0.25211113
Gini.Index	0.44762107	0.17770895	0.194441653	-0.39550136	-0.25211113	1.00000000
Alcohol.Consumption	-0.06701583	-0.13891488	-0.153009997	0.59295987	0.26469999	-0.34669452
GDP	-0.29538679	0.25253376	-0.249727071	0.35615171	0.66962292	-0.26803034
Smokers	0.04809660	-0.01998788	0.005587341	-0.25213513	-0.04423896	0.02069348
Temperature	0.04809619	-0.01999068	0.005587489	-0.25213375	-0.04423933	0.02069139
Death.penalty.permitted	0.20791071	-0.07601948	0.006725505	-0.23694021	-0.22490082	0.29520633
Precentage.of.green.areas	0.11737958	0.02048545	-0.037588296	0.27147607	0.18931764	-0.18217329

	Alcohol.Consumption	GDP	Smokers	Temperature	Death.penalty.permitted	Precentage.of.green.areas
Number.of.murders	-0.06701583	-0.29538679	0.048096596	0.048096190	0.207910712	0.11737958
Population.density..per.square.km.	-0.13891488	0.25253376	-0.019987884	-0.019990684	-0.076019478	0.02048545
Poverty.rate	-0.15301000	-0.24972707	0.005587341	0.005587489	0.006725505	-0.03758830
Divorce.rate	0.59295987	0.35615171	-0.252135129	-0.252133754	-0.236940215	0.27147607
Life.expectancy	0.26469999	0.66962292	-0.044238958	-0.044239331	-0.224900818	0.18931764
Gini.Index	-0.34669452	-0.26803034	0.020693480	0.020691389	0.295206335	-0.18217329
Alcohol.Consumption	1.00000000	0.16885242	-0.123672684	-0.123670359	-0.340855715	0.33894082
GDP	0.16885242	1.00000000	-0.028752364	-0.028753597	-0.027236662	0.17485385
Smokers	-0.12367268	-0.02875236	1.000000000	1.000000000	-0.009885049	-0.02534466
Temperature	-0.12367036	-0.02875360	1.000000000	1.000000000	-0.009887234	-0.02534395
Death.penalty.permitted	-0.34085572	-0.02723666	-0.009885049	-0.009887234	1.000000000	-0.21744211
Precentage.of.green.areas	0.33894082	0.17485385	-0.025344661	-0.025343952	-0.217442113	1.00000000

איור 2 טבלת קשרי פירסון

נספח 3: תרשימי פיזור של המשתנים המסבירים בינם לבין עצמם



איור 3 תרשימי פיזור משתנים מסבירים

מקרא:

1-אתאיזם

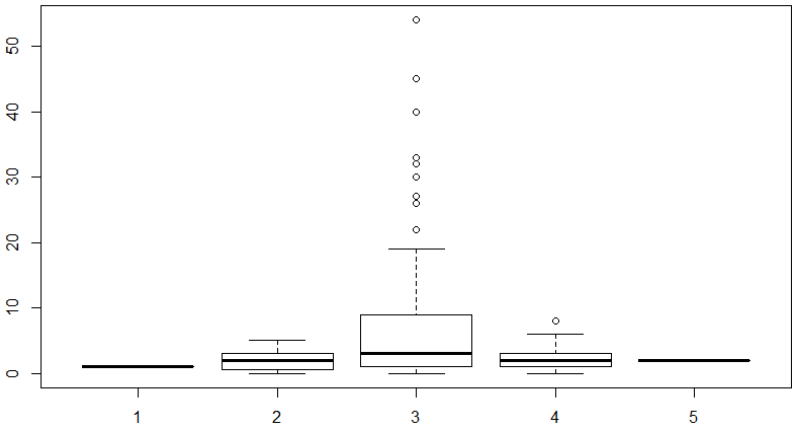
2- בודהיזם

3- נצרות

4- איסלאם

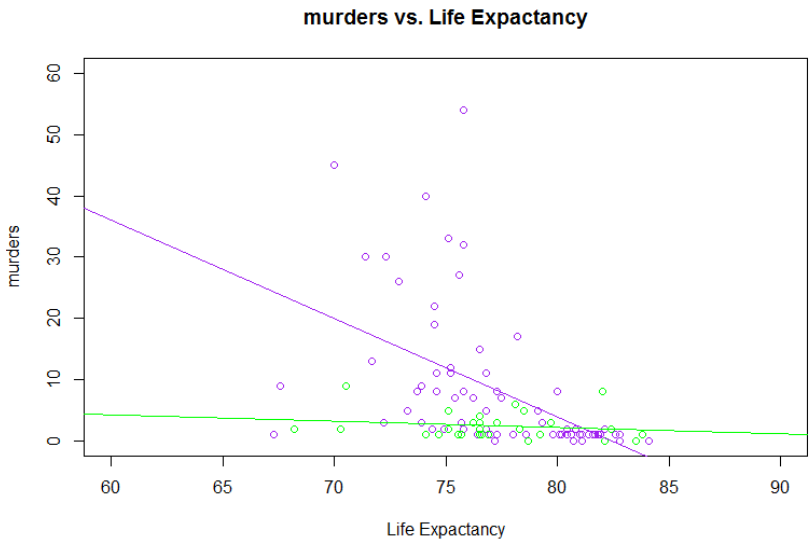
5- יהדות

נספח 4: תרשים Box Plot משתנה מסביר 'דת עיקרית'

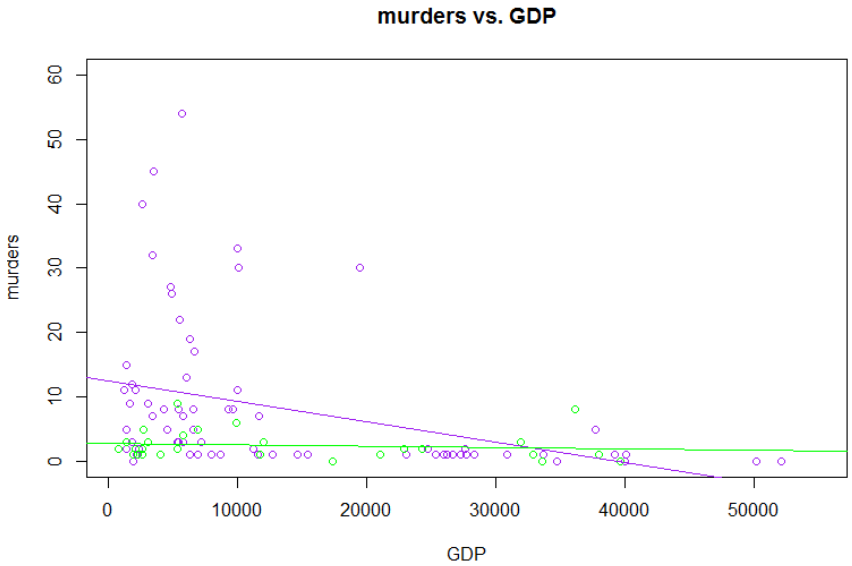


איור 15 – דת עיקרית

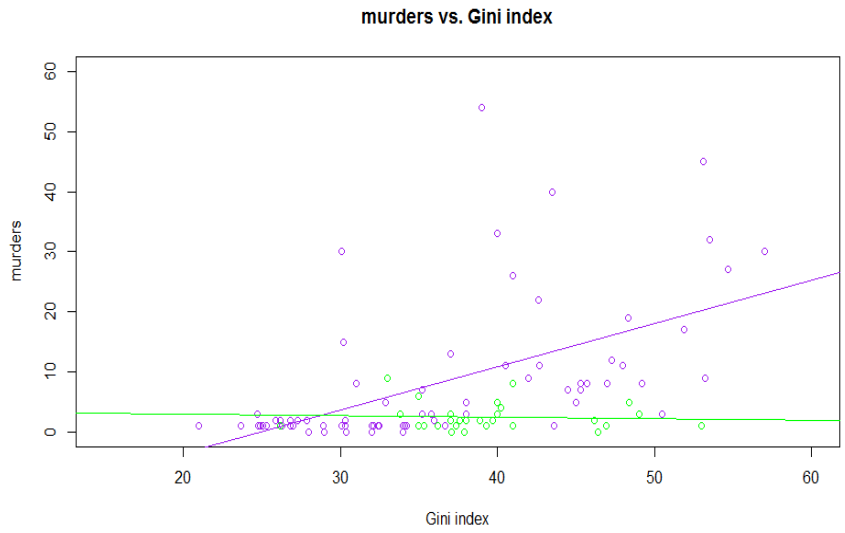
נספח 5: משתני אינטראקציה – היתר עונש מוות



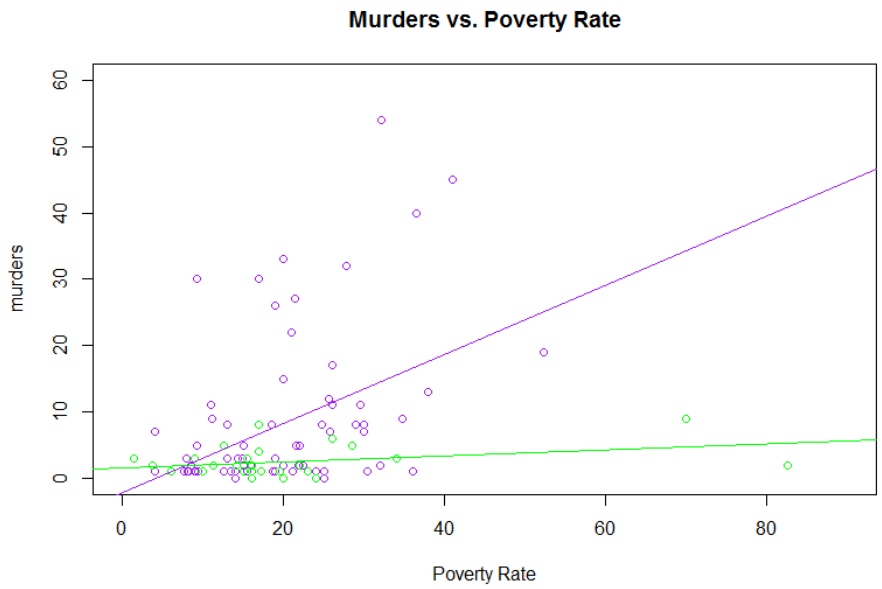
איור 4 תוחלת חיים והיתר עונש מוות – אינטרקציה



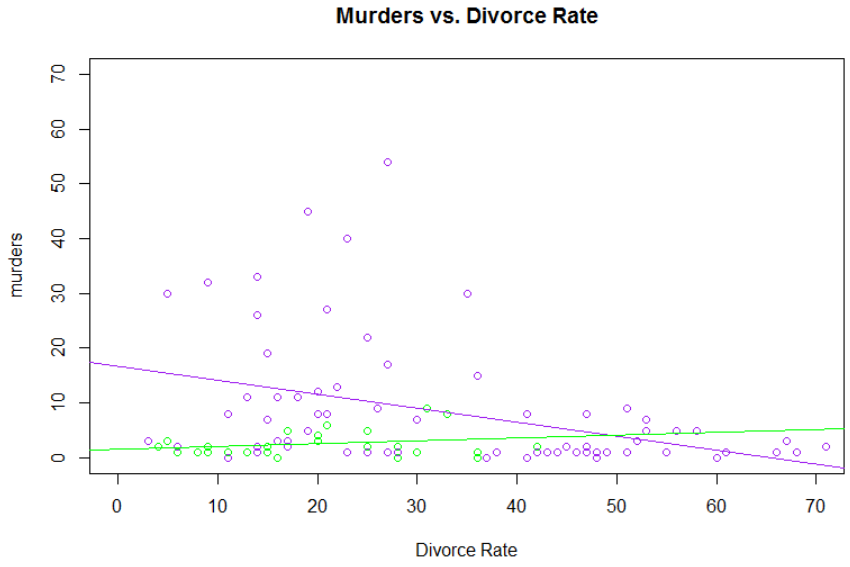
איור 5 GDP והיתר עונש מוות – אינטרקציה



איור 6 מדד ג'יני והיתר עונש מוות – אינטרקציה



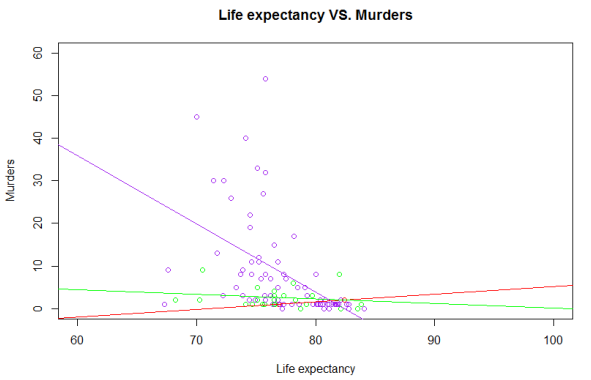
איור 7 שיעור העוני והיתר עונש מוות – אינטרקציה

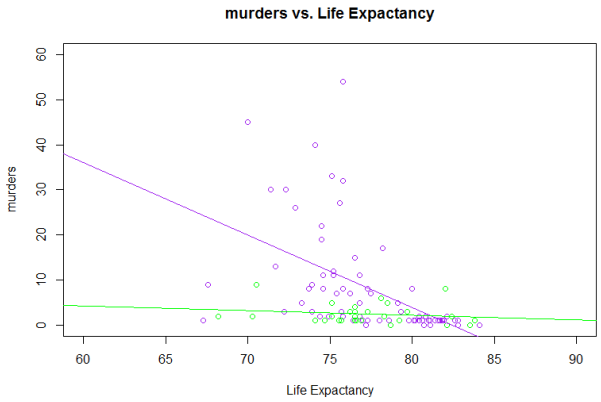


איור 20 שיעור הגירושים והיתר עונש מוות – אינטרקציה

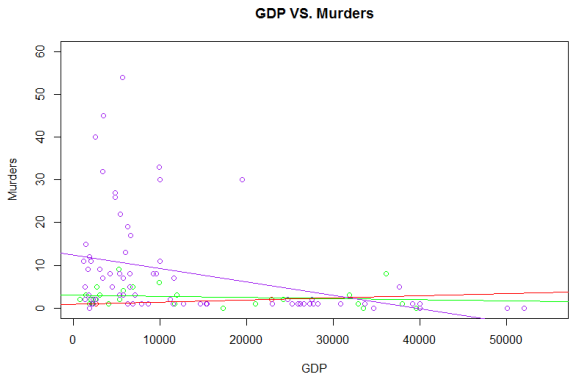
נספח 6: משתני אינטראקציה - דתות – הצגת השימוש ב-3 קטגוריות לעומת 2 קטגוריות

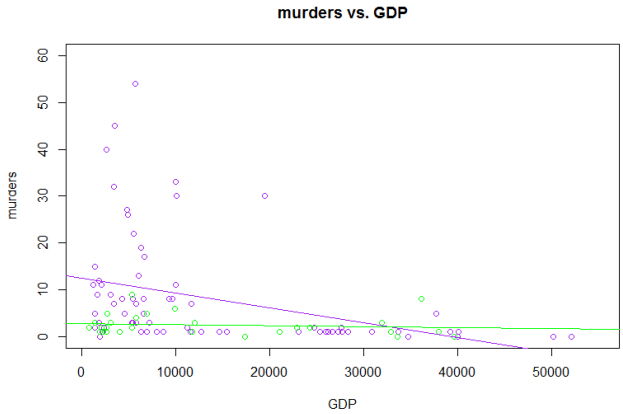
תוחלת חיים



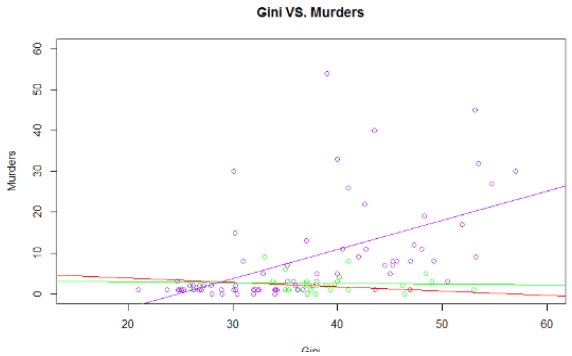


איור 8 תוחלת חיים ודתות - אינטראקציה

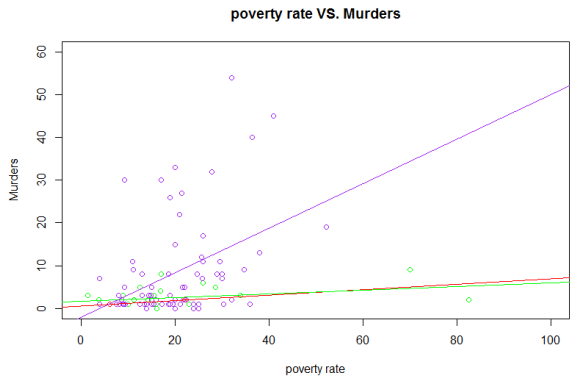


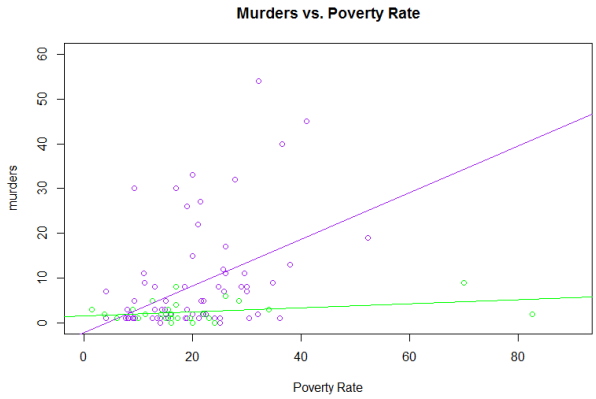


איור 9 מדד GDP ודתות - אינטראקציה

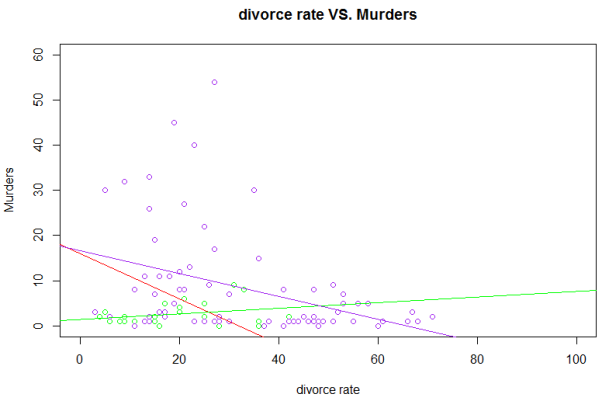


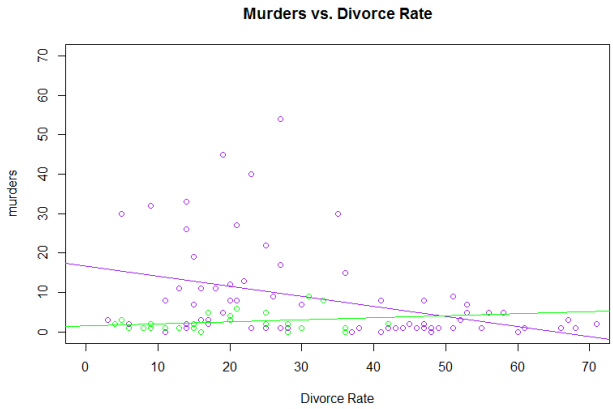
איור 10 מדד ג'יני ודתות – אינטראקציה





איור 11 שיעור העוני ודתות – אינטראקציה





איור 25 שיעור הגירושים ודתות – אינטראקציה

נספח 7: בחירת משתני המודל

```
Call:
lm(formula = Number.of.murders ~ Gini.Index + religion.category +
    Life.expectancy + Death.penalty.permitted.category, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-10.561  -4.493  -0.952   1.961  44.020

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                        51.9633    18.6093   2.792 0.006316 **
Gini.Index                          0.4780     0.1014   4.713 8.25e-06 ***
religion.category1                 -9.2218     1.9340  -4.768 6.62e-06 ***
Life.expectancy                    -0.7998     0.2255  -3.547 0.000604 ***
Death.penalty.permitted.category1   4.4181     2.0208   2.186 0.031219 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.904 on 96 degrees of freedom
Multiple R-squared:  0.4338,    Adjusted R-squared:  0.4102
F-statistic: 18.39 on 4 and 96 DF,  p-value: 3.037e-11
```

איור 12 Forward Selection - מבחן F חלקי

```
Call:
lm(formula = dataset$Number.of.murders ~ Gini.Index + religion.category +
    Life.expectancy + Death.penalty.permitted.category + Poverty.rate +
    religion.category:Life.expectancy, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-11.843  -3.903  -0.717   2.703  42.216

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                        64.30960   22.71248   2.831 0.005668 **
Gini.Index                          0.38623    0.10518   3.672 0.000399 ***
religion.category1                -86.33197   38.42345  -2.247 0.026989 *
Life.expectancy                    -0.95660    0.27228  -3.513 0.000682 ***
Death.penalty.permitted.category1   4.79898    1.98301   2.420 0.017441 *
Poverty.rate                        0.15333    0.07101   2.159 0.033365 *
religion.category1:Life.expectancy  0.99970    0.49874   2.004 0.047897 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.698 on 94 degrees of freedom
Multiple R-squared:  0.4741,    Adjusted R-squared:  0.4405
F-statistic: 14.12 on 6 and 94 DF,  p-value: 2.114e-11
```

איור 13 Forward Selection - מבחן AIC

```
Call:
lm(formula = Number.of.murders ~ Life.expectancy + Gini.Index +
    Poverty.rate + religion.category + Death.penalty.permitted.category +
    Life.expectancy:religion.category, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-11.843  -3.903  -0.717   2.703  42.216

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                        64.30960   22.71248   2.831 0.005668 **
Life.expectancy                    -0.95660    0.27228  -3.513 0.000682 ***
Gini.Index                          0.38623    0.10518   3.672 0.000399 ***
Poverty.rate                        0.15333    0.07101   2.159 0.033365 *
religion.category1                -86.33197   38.42345  -2.247 0.026989 *
Death.penalty.permitted.category1   4.79898    1.98301   2.420 0.017441 *
Life.expectancy:religion.category1  0.99970    0.49874   2.004 0.047897 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.698 on 94 degrees of freedom
Multiple R-squared:  0.4741,	Adjusted R-squared:  0.4405
F-statistic: 14.12 on 6 and 94 DF,  p-value: 2.114e-11
```

איור 14 Backward Selection - מבחן F חלקי

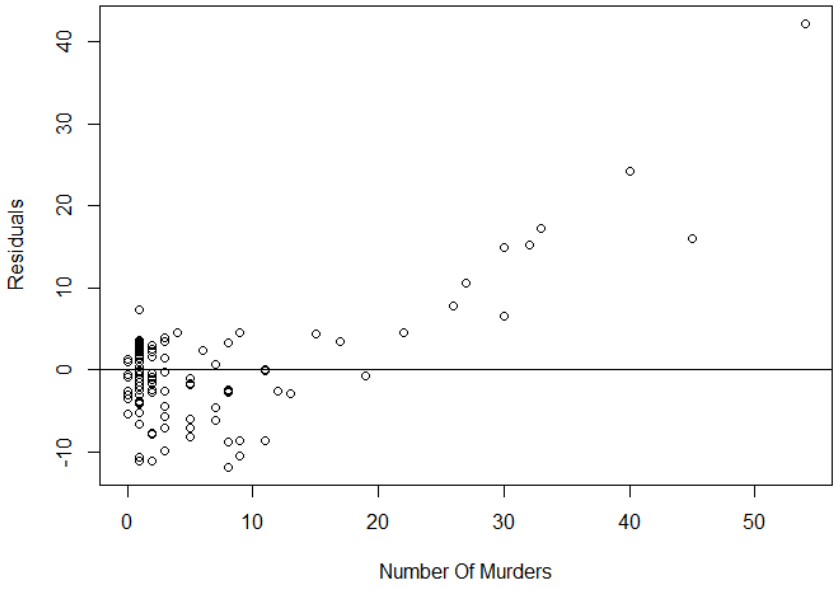

```
Call:
lm(formula = Number.of.murders ~ Life.expectancy + Gini.Index +
    Poverty.rate + religion.category + Death.penalty.permitted.category +
    Life.expectancy:religion.category, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-11.843  -3.903  -0.717   2.703  42.216

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                        64.30960   22.71248   2.831 0.005668 **
Life.expectancy                    -0.95660    0.27228  -3.513 0.000682 ***
Gini.Index                          0.38623    0.10518   3.672 0.000399 ***
Poverty.rate                        0.15333    0.07101   2.159 0.033365 *
religion.category1                -86.33197   38.42345  -2.247 0.026989 *
Death.penalty.permitted.category1   4.79898    1.98301   2.420 0.017441 *
Life.expectancy:religion.category1  0.99970    0.49874   2.004 0.047897 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.698 on 94 degrees of freedom
Multiple R-squared:  0.4741,	Adjusted R-squared:  0.4405
F-statistic: 14.12 on 6 and 94 DF,  p-value: 2.114e-11
```


איור 15 Backward Selection - מבחן AIC

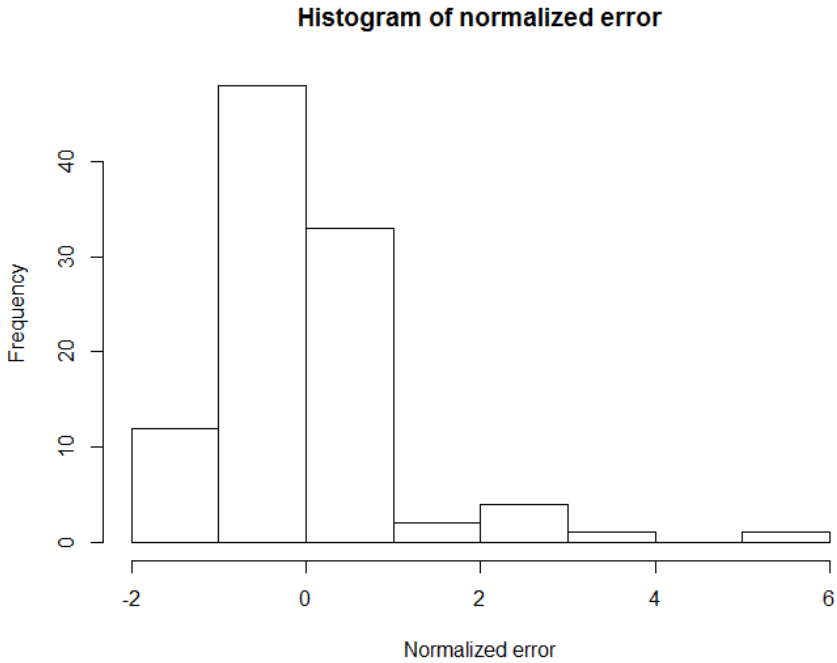
```
Call:
lm(formula = dataset$Number.of.murders ~ Gini.Index + religion.category +
    Life.expectancy + Death.penalty.permitted.category + Poverty.rate +
    religion.category:Life.expectancy, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-11.843  -3.903  -0.717   2.703  42.216

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                        64.30960   22.71248   2.831 0.005668 **
Gini.Index                          0.38623    0.10518   3.672 0.000399 ***
religion.category1                -86.33197   38.42345  -2.247 0.026989 *
Life.expectancy                    -0.95660    0.27228  -3.513 0.000682 ***
Death.penalty.permitted.category1   4.79898    1.98301   2.420 0.017441 *
Poverty.rate                        0.15333    0.07101   2.159 0.033365 *
religion.category1:Life.expectancy  0.99970    0.49874   2.004 0.047897 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.698 on 94 degrees of freedom
Multiple R-squared:  0.4741,	Adjusted R-squared:  0.4405
F-statistic: 14.12 on 6 and 94 DF,  p-value: 2.114e-11
```

איור 16 Stepwise Regression מבחן AIC

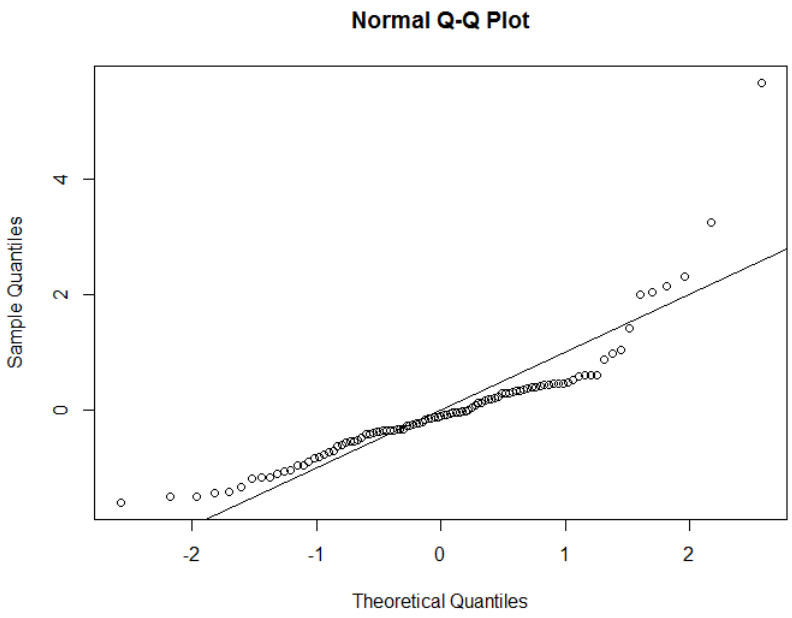
נספח 8: בדיקת הנחות המודל



איור 17 תרשים פיזור של השגיאות המתוקננות מול הערך הצפוי

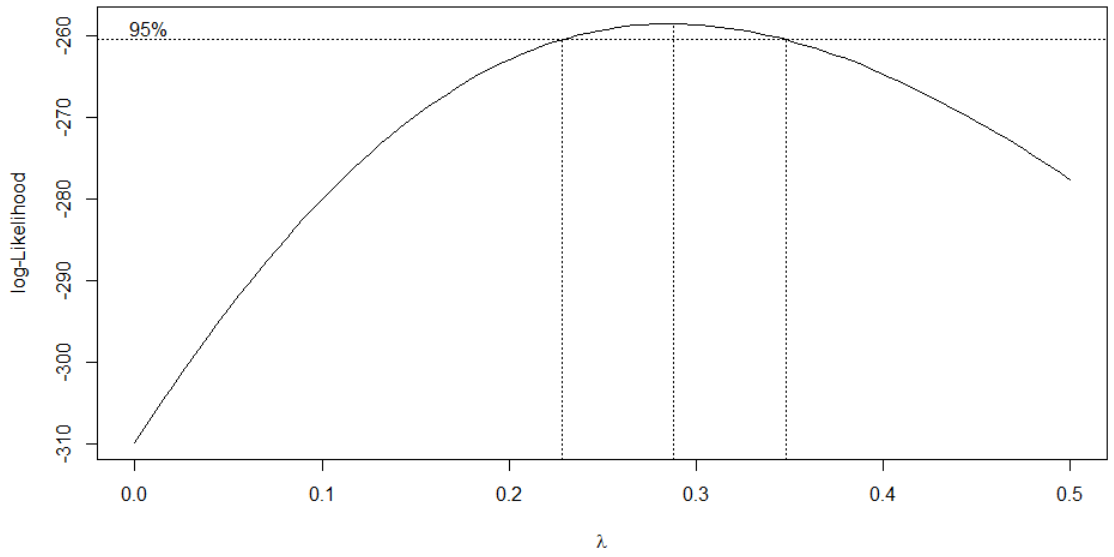


איור 18 תרשים היסטוגרמה

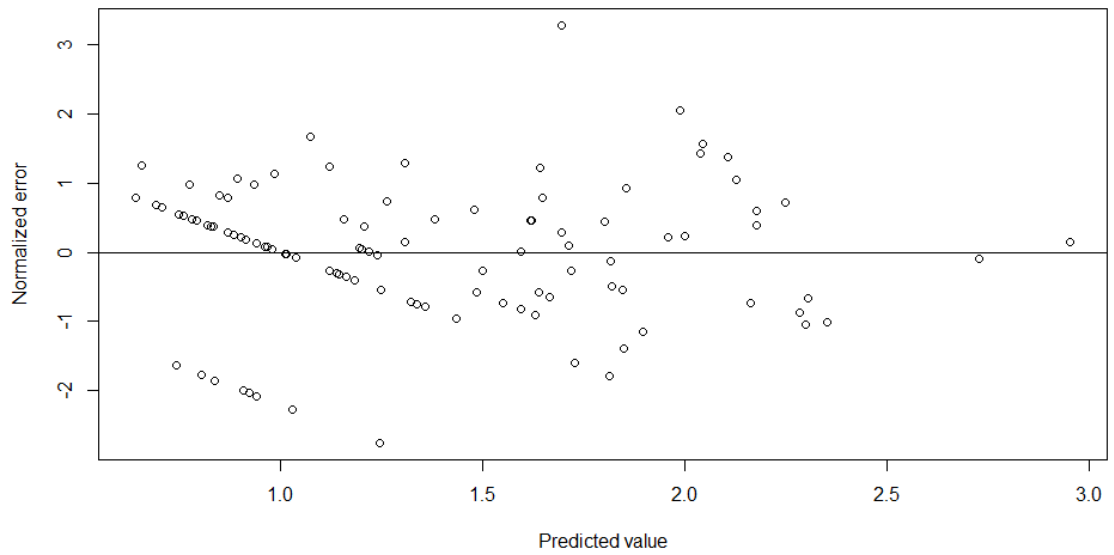


איור 19 תרשים QQPlot

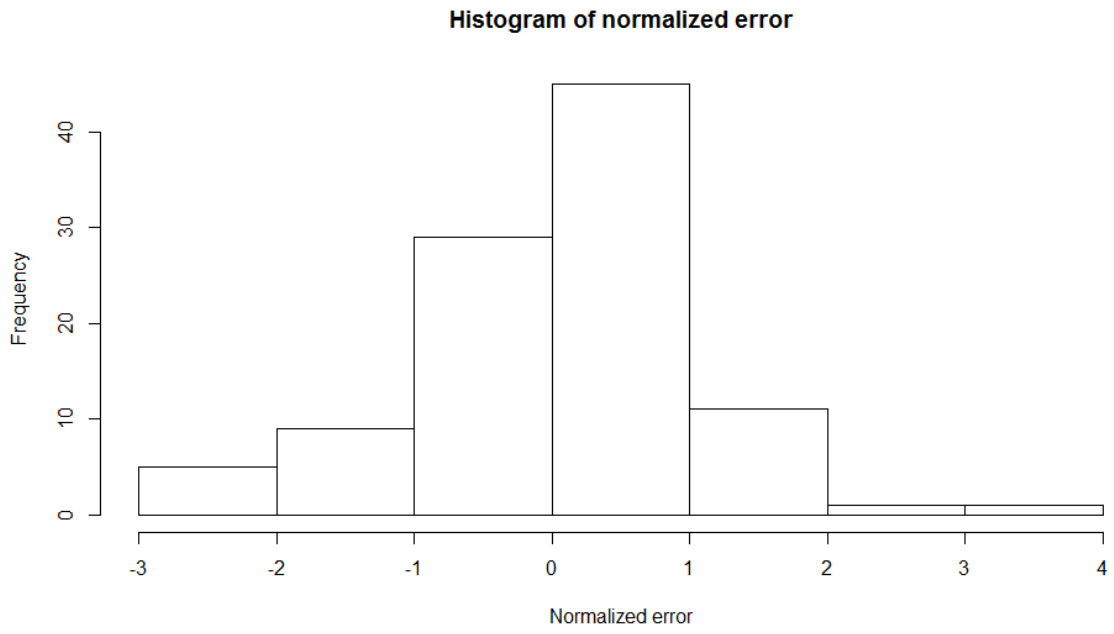
נספח 9: שיפור המודל



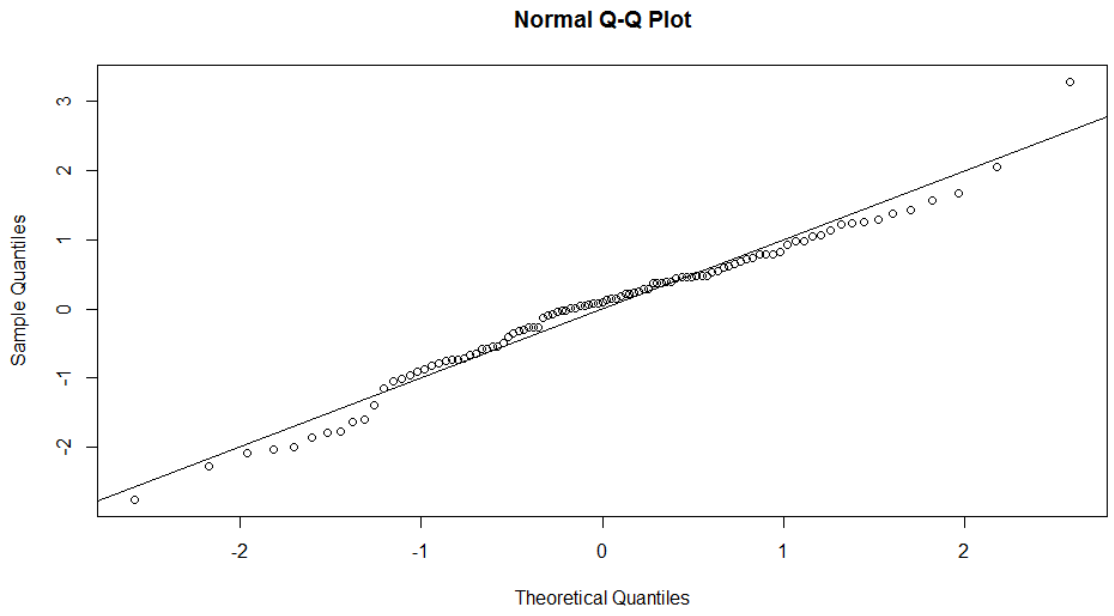
איור 34 תרשים Box-Cox



איור 35 תרשים פיזור של השגיאות המתוקננות מול הערך הצפוי



איור 36- תרשים היסטוגרמה



איור 37 תרשים QQPlot

```
> fit<-lm(dataset$Number.of.murders^(0.3)~(dataset$Life.expectancy)^2)
> summary(fit)

Call:
lm(formula = dataset$Number.of.murders^(0.3) ~ (dataset$Life.expectancy)^2)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.45530 -0.40829 -0.01184  0.33686  1.74295 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)              9.49475    1.26924   7.481 3.05e-11 ***
dataset$Life.expectancy -0.10460    0.01641  -6.373 5.89e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6048 on 99 degrees of freedom
Multiple R-squared:  0.2909,	Adjusted R-squared:  0.2838 
F-statistic: 40.62 on 1 and 99 DF,  p-value: 5.889e-09
```

איור 20 טרנספורמציה על תוחלת חיים

```
> fit<-lm(dataset$Number.of.murders^(0.3)~(dataset$Gini.Index)^2)
> summary(fit)

Call:
lm(formula = dataset$Number.of.murders^(0.3) ~ (dataset$Gini.Index)^2)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.80510 -0.27771  0.03544  0.28518  1.81879 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)        -0.168356   0.284718  -0.591    0.556    
dataset$Gini.Index  0.042531   0.007467   5.696 1.26e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6233 on 99 degrees of freedom
Multiple R-squared:  0.2468,	Adjusted R-squared:  0.2392 
F-statistic: 32.44 on 1 and 99 DF,  p-value: 1.263e-07
```

איור 21 טרנספורמציה על מדד ג'יני

```
> fit<-lm(dataset$Number.of.murders^(0.3)~(dataset$Poverty.rate)^2)
> summary(fit)

Call:
lm(formula = dataset$Number.of.murders^(0.3) ~ (dataset$Poverty.rate)^2)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.4902 -0.3419 -0.1170  0.3240  1.7082 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)          1.100106   0.135947   8.092 1.52e-12 ***
dataset$Poverty.rate 0.015603   0.005815   2.683  0.00855 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6935 on 99 degrees of freedom
Multiple R-squared:  0.06779,	Adjusted R-squared:  0.05837 
F-statistic: 7.199 on 1 and 99 DF,  p-value: 0.008551
```

איור 22 טרנספורמציה על שיעור העוני

```
> fit<-lm(dataset$Number.of.murders^(0.3)~(dataset$religion.new)^2)
> summary(fit)

Call:
lm(formula = dataset$Number.of.murders^(0.3) ~ (dataset$religion.new)^2)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.5156 -0.5156 -0.1252  0.3651  1.7936 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)          1.51556    0.08181  18.525   <2e-16 ***
dataset$religion.new -0.36497    0.15538  -2.349   0.0208 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.699 on 99 degrees of freedom
Multiple R-squared:  0.05279,	Adjusted R-squared:  0.04322 
F-statistic: 5.517 on 1 and 99 DF,  p-value: 0.02082
```

איור 23 טרנספורמציה על משתנה דת

```
> fit<-lm(dataset$Number.of.murders^(0.3)~(dataset$Death.penalty.permitted)^2)
> summary(fit)

Call:
lm(formula = dataset$Number.of.murders^(0.3) ~ (dataset$Death.penalty.permitted)^2)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.70561 -0.30268 -0.08495  0.49011  2.00648 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                    1.30268    0.08131  16.021   <2e-16 ***
dataset$Death.penalty.permitted  0.40293    0.15443   2.609   0.0105 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6947 on 99 degrees of freedom
Multiple R-squared:  0.06434,	Adjusted R-squared:  0.05488 
F-statistic: 6.807 on 1 and 99 DF,  p-value: 0.01049
```

איור24 טרנספורמציה על היתר עונש מוות

```
> lm.full<-lm(dataset$Number.of.murders^(0.3)~dataset$Life.expectancy + (dataset$Life.expe
.Index + dataset$Poverty.rate + dataset$religion.new + dataset$Death.penalty.permitted + d
dataset$religion.new)
> model.aic.backward <- step(lm.full, direction = "backward", trace = 1)
Start:  AIC=-140.2
dataset$Number.of.murders^(0.3) ~ dataset$Life.expectancy + (dataset$Life.expectancy)^2 + 
    dataset$Gini.Index + dataset$Poverty.rate + dataset$religion.new + 
    dataset$Death.penalty.permitted + dataset$Life.expectancy:dataset$religion.new

                                            Df Sum of Sq    RSS     AIC
- dataset$Poverty.rate                       1    0.3335 22.276 -140.67
- dataset$Life.expectancy:dataset$religion.new  1    0.3431 22.286 -140.63
<none>                                                  21.943 -140.20
- dataset$Death.penalty.permitted            1    1.7135 23.656 -134.60
- dataset$Gini.Index                         1    5.1744 27.117 -120.81

Step:  AIC=-140.67
dataset$Number.of.murders^(0.3) ~ dataset$Life.expectancy + dataset$Gini.Index + 
    dataset$religion.new + dataset$Death.penalty.permitted + 
    dataset$Life.expectancy:dataset$religion.new

                                            Df Sum of Sq    RSS     AIC
- dataset$Life.expectancy:dataset$religion.new  1    0.2277 22.504 -141.65
<none>                                                  22.276 -140.67
- dataset$Death.penalty.permitted            1    1.5631 23.839 -135.82
- dataset$Gini.Index                         1    5.9284 28.204 -118.84

Step:  AIC=-141.65
dataset$Number.of.murders^(0.3) ~ dataset$Life.expectancy + dataset$Gini.Index + 
    dataset$religion.new + dataset$Death.penalty.permitted

                                Df Sum of Sq    RSS     AIC
<none>                                      22.504 -141.65
- dataset$Death.penalty.permitted  1    1.6106 24.114 -136.66
- dataset$religion.new             1    6.5134 29.017 -117.97
- dataset$Gini.Index               1    7.2132 29.717 -115.56
- dataset$Life.expectancy          1    7.4947 29.998 -114.61
> summary(model.aic.backward)

Call:
lm(formula = dataset$Number.of.murders^(0.3) ~ dataset$Life.expectancy + 
    dataset$Gini.Index + dataset$religion.new + dataset$Death.penalty.permitted)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.21852 -0.30086  0.03593  0.28982  1.63667 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                   6.247891   1.139946   5.481 3.41e-07 ***
dataset$Life.expectancy      -0.078094   0.013811  -5.654 1.61e-07 ***
dataset$Gini.Index            0.034464   0.006213   5.547 2.56e-07 ***
dataset$religion.new         -0.624498   0.118473  -5.271 8.33e-07 ***
dataset$Death.penalty.permitted  0.324466   0.123786   2.621   0.0102 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4842 on 96 degrees of freedom
Multiple R-squared:  0.5593,	Adjusted R-squared:  0.541 
F-statistic: 30.46 on 4 and 96 DF,  p-value: 2.299e-16
```

איור 25 BackWard - AIC

נספח 10: גיבוי קוד

תרשימי פיזור משתנים מסבירים

```
> install.packages("moments)"
> library(moments(
> dataset<-read.csv(file.choose(),header = T(
> plot(dataset)
> plot(x=dataset$Life.expectancy,y=dataset$Number.of.murders,xlab="Life expectancy ",ylab="Number
of murders")
> abline(lm(dataset$Number.of.murders~dataset$Life.expectancy), col="hotpink")
> plot(x=dataset$Gini.Index,y=dataset$Number.of.murders,xlab="Gini index ",ylab="Number of murders
")
> abline(lm(dataset$Number.of.murders~dataset$Gini.Index), col="hotpink")
> plot(x=dataset$GDP,y=dataset$Number.of.murders,xlab="GDP ",ylab="Number of murders")
> abline(lm(dataset$Number.of.murders~dataset$GDP), col="hotpink")
> plot(x=dataset$Alcohol.Consumption,y=dataset$Number.of.murders,xlab="Alcohol consumption ",ylab=
"Number of murders")
> abline(lm(dataset$Number.of.murders~dataset$Alcohol.Consumption), col="hotpink")
> plot(x=dataset$Population.density..per.square.km.,y=dataset$Number.of.murders,xlab="Population d
ensity ",ylab="Number of murders")
> abline(lm(dataset$Number.of.murders~dataset$Population.density..per.square.km.), col="hotpink")
> plot(x=dataset$Poverty.rate,y=dataset$Number.of.murders,xlab="Poverty rate ",ylab="Number of mur
ders")
> abline(lm(dataset$Number.of.murders~dataset$Poverty.rate), col="hotpink")
> plot(x=dataset$Smokers,y=dataset$Number.of.murders,xlab="Smokers ",ylab="Number of murders")
> abline(lm(dataset$Number.of.murders~dataset$Smokers), col="hotpink")
> plot(x=dataset$Death.penalty.permitted,y=dataset$Number.of.murders,xlab="Death penalty permitted
",ylab="Number of murders")
> abline(lm(dataset$Number.of.murders~dataset$Death.penalty.permitted), col="hotpink")
> plot(x=dataset$Divorce.rate,y=dataset$Number.of.murders,xlab="Divorce rate ",ylab="Number of mur
ders")
> abline(lm(dataset$Number.of.murders~dataset$Divorce.rate), col="hotpink")
> plot(x=dataset$Temperature,y=dataset$Number.of.murders,xlab="Temperature ",ylab="Number of murde
rs")
> abline(lm(dataset$Number.of.murders~dataset$Temperature), col="hotpink")
> plot(x=dataset$Precentage.of.green.areas,y=dataset$Number.of.murders,xlab="Precentage of green a
reas ",ylab="Number of murders")
> abline(lm(dataset$Number.of.murders~dataset$Precentage.of.green.areas), col="hotpink")

> boxplot(dataset$Number.of.murders~dataset$Major.religion)
> cor(dataset[sapply(dataset, function(x) !is.factor(x))],method = "pearson")
```

משתני אינטרקציה

```
> mod<-lm(formula=dataset$Number.of.murders~dataset$Gini.Index*factor(dataset$Death.penalty.permit
ted))
> summary(mod)
> plot(dataset$Gini.Index[dataset$Death.penalty.permitted=='0'] ,dataset$Number.of.murders[dataset
$Death.penalty.permitted=='0'], col="brown",xlim = c(15,60),ylim=c(0,60) ,xlab="Gini index",ylab="
murders", main = "murders vs. Gini index")
> points(dataset$Gini.Index[dataset$Death.penalty.permitted=='1'] ,dataset$Number.of.murders[datas
et$Death.penalty.permitted=='1'], col="blue")
> abline(a=-12.98329,b=0.51263, col="brown")
> abline(a=-12.98329-0.25138, b=0.51263+0.05343, col='blue')

>  mod<-lm(formula=dataset$Number.of.murders~dataset$Life.expectancy*factor(dataset$Death.penalty.p
ermitted))
> summary(mod)
> plot(dataset$Life.expectancy[dataset$Death.penalty.permitted=='0'] ,dataset$Number.of.murders[da
taset$Death.penalty.permitted=='0'], col="brown",xlim = c(60,100),ylim=c(0,60) ,xlab="life expecta
ncy",ylab="murders", main = "murders vs. life expectancy")
> points(dataset$Life.expectancy[dataset$Death.penalty.permitted=='1'] ,dataset$Number.of.murders[
dataset$Death.penalty.permitted=='1'], col="blue")
> abline(a=73.3275,b=-0.8746, col="brown")
> abline(a=73.3275+66.1327, b=-0.8746-0.8297, col='blue')

> mod<-lm(formula=dataset$Number.of.murders~dataset$GDP*factor(dataset$Death.penalty.permitted))
> summary(mod)
> plot(dataset$GDP[dataset$Death.penalty.permitted=='0'] ,dataset$Number.of.murders[dataset$Death.
penalty.permitted=='0'], col="brown",xlim = c(400,55000),ylim=c(0,60) ,xlab="GDP",ylab="murders",
main = "murders vs. GDP")
> points(dataset$GDP[dataset$Death.penalty.permitted=='1'] ,dataset$Number.of.murders[dataset$Deat
h.penalty.permitted=='1'], col="blue")
> abline(a=8.462e+00,b=-2.306e-04, col="brown")
> abline(a=8.462e+00+4.474e+00, b=-2.306e-04+7.782e-06, col='blue')

> mod<-lm(formula=dataset$Number.of.murders~dataset$Poverty.rate*factor(dataset$Death.penalty.perm
itted))
> summary(mod)
> plot(dataset$Poverty.rate[dataset$Death.penalty.permitted=='0'] ,dataset$Number.of.murders[datas
et$Death.penalty.permitted=='0'], col="brown",xlim = c(0,90),ylim=c(0,60) ,xlab="poverty rate",yla
b="murders", main = "murders vs. poverty rate")
> points(dataset$Poverty.rate[dataset$Death.penalty.permitted=='1'] ,dataset$Number.of.murders[dat
aset$Death.penalty.permitted=='1'], col="blue")
> abline(a= -1.8186,b=0.3550, col="brown")
> abline(a=-1.8186+10.0502, b=0.3550-0.2643, col='blue')

> mod<-lm(formula=dataset$Number.of.murders~dataset$Divorce.rate*factor(dataset$Death.penalty.perm
itted))
> summary(mod)> plot(dataset$Divorce.rate[dataset$Death.penalty.permitted=='0'] ,dataset$Number.of
.murders[dataset$Death.penalty.permitted=='0'], col="brown",xlim = c(0,80),ylim=c(0,60) ,xlab="div
orce rate",ylab="murders", main = "murders vs. divorce rate")
> points(dataset$Divorce.rate[dataset$Death.penalty.permitted=='1'] ,dataset$Number.of.murders[dat
aset$Death.penalty.permitted=='1'], col="blue")
> abline(a=9.21654,b= -0.12119, col="brown")
```

```
> abline(a= 9.21654+2.90973, b=-0.12119+0.03268, col='blue')
```

דתות(3 דתות)

```
> dataset$religion.fac <- factor(dataset$religion.cat)
> summary(mod)
> mod<-lm(formula=dataset$Number.of.murders ~ dataset$Gini.Index*dataset$religion.fac)
> plot(dataset$Gini.Index[dataset$religion.cat=='1'],dataset$Number.of.murders[dataset$religion.ca
t=='1'], col="red",xlim = c(15,60), ylim=c(0,60) ,xlab="Gini",ylab="Murders", main = "Gini VS. Mur
ders")
> points(dataset$Gini.Index[dataset$religion.cat=='2'],dataset$Number.of.murders[dataset$religion.
cat=='2'], col="Green" )
> points(dataset$Gini.Index[dataset$religion.cat=='3'],dataset$Number.of.murders[dataset$religion.
cat=='3'], col="purple" )
> abline(a= 6.04301, b=-0.10753, col="red")
> abline(a= 6.04301 + -2.80132, b=-0.10753 + 0.09062, col="Green")
> abline(a= 6.04301 -23.89823, b=-0.10753 + 0.82526, col="purple")

>mod<-lm(formula=dataset$Number.of.murders ~ dataset$Life.expectancy*dataset$religion.fac)
> summary(mod)
> plot(dataset$Life.expectancy[dataset$religion.cat=='1'],
dataset$Number.of.murders[dataset$religion.cat=='1'],
 col="red",xlim = c(60,100), ylim=c(0,60) ,xlab="Life expectancy",ylab="Murders",
 main = "Life expectancy VS. Murders")
> points(dataset$Life.expectancy[dataset$religion.cat=='2'],
dataset$Number.of.murders[dataset$religion.cat=='2'], col="Green" )
> points(dataset$Life.expectancy[dataset$religion.cat=='3'],
dataset$Number.of.murders[dataset$religion.cat=='3'], col="purple" )
> abline(a= -12.9818, b=0.1818, col="red")
> abline(a= -12.9818+23.6053, b=0.1818-0.2865, col="green")
> abline(a= -12.9818+145.1816, b=0.1818-1.7854, col="purple")

> mod<-lm(formula=dataset$Number.of.murders ~ dataset$Poverty.rate*dataset$religion.fac)
> summary(mod)
> plot(dataset$Poverty.rate[dataset$religion.cat=='1'],dataset$Number.of.murders[dataset$religion.
cat=='1'], col="red",xlim = c(0,100), ylim=c(0,60) ,xlab="poverty rate",ylab="Murders", main = "po
verty rate VS. Murders")
> points(dataset$Poverty.rate[dataset$religion.cat=='2'],dataset$Number.of.murders[dataset$religio
n.cat=='2'], col="Green" )
> points(dataset$Poverty.rate[dataset$religion.cat=='3'],dataset$Number.of.murders[dataset$religio
n.cat=='3'], col="purple" )
> abline(a= 0.61635, b=0.06289, col="red")
> abline(a= 0.61635+1.04135, b=0.06289-0.01985, col="green")
> abline(a= 0.61635-2.74946, b=0.06289+0.45790, col="purple")

> mod<-lm(formula=dataset$Number.of.murders ~ dataset$Divorce.rate*dataset$religion.fac)
> summary(mod)
> plot(dataset$Divorce.rate[dataset$religion.cat=='1'],dataset$Number.of.murders[dataset$religion.
cat=='1'], col="red",xlim = c(0,100), ylim=c(0,60) ,xlab="divorce rate",ylab="Murders", main = "di
vorce rate VS. Murders")
> points(dataset$Divorce.rate[dataset$religion.cat=='2'],dataset$Number.of.murders[dataset$religio
n.cat=='2'], col="Green" )
> points(dataset$Divorce.rate[dataset$religion.cat=='3'],dataset$Number.of.murders[dataset$religio
n.cat=='3'], col="purple" )
> abline(a= 16.0000, b=-0.5000, col="red")
> abline(a= 16.0000-14.6331, b=-0.5000+0.5629, col="green")
> abline(a= 16.0000+0.7379, b=-0.5000+0.2456 , col="purple")

> mod<-lm(formula=dataset$Number.of.murders ~ dataset$GDP*dataset$religion.fac)
> summary(mod)
> plot(dataset$GDP[dataset$religion.cat=='1'],dataset$Number.of.murders[dataset$religion.cat=='1']
, col="red",xlim = c(750,55000), ylim=c(0,60) ,xlab="GDP",ylab="Murders", main = "GDP VS. Murders"
)
> points(dataset$GDP[dataset$religion.cat=='2'],dataset$Number.of.murders[dataset$religion.cat=='2
'], col="Green" )
> points(dataset$GDP[dataset$religion.cat=='3'],dataset$Number.of.murders[dataset$religion.cat=='3
'], col="purple" )
> abline(a=  8.695e-01, b= 4.946e-05, col="red")
> abline(a=  8.695e-01+ 2.054e+00, b= 4.946e-05-7.485e-05 , col="green")
> abline(a=  8.695e-01+ 1.156e+01, b= 4.946e-05-3.640e-04, col="purple")

> dataset<-read.csv(file.choose(),header = T)
> mod<-lm(formula=dataset$Number.of.murders~dataset$Gini.Index*factor(dataset$religion.new))
> summary(mod)

> plot(dataset$Divorce.rate[dataset$religion.new=='1'],dataset$Number.of.murders[dataset$religion.
new=='1'], col="red",xlim = c(0,100), ylim=c(0,60) ,xlab="divorce rate",ylab="Murders", main = "di
vorce rate VS. Murders")
> points(dataset$Divorce.rate[dataset$religion.new=='2'],dataset$Number.of.murders[dataset$religio
n.new=='2'], col="Green" )
> points(dataset$Divorce.rate[dataset$religion.new=='3'],dataset$Number.of.murders[dataset$religio
n.new=='3'], col="purple" )
> abline(a=  2.000e+00,  1.075e-13, col="red")
> abline(a=  2.000e+00-1.181e+00,  1.075e-13+ 3.794e-02, col="green")

> dataset<-read.csv(file.choose(),header = T)
> mod<-lm(formula=dataset$Number.of.murders~dataset$Gini.Index*factor(dataset$religion.new))
> summary(mod)
dataset$religion.fac<-factor(dataset$religion.new)
plot(dataset$Gini.Index[dataset$religion.fac=='0'] ,dataset$Number.of.murders[dataset$religion.fac
=='0'], col="purple",xlim = c(15,60),ylim=c(0,60) ,xlab="Gini index",ylab="murders", main = "murde
rs vs. Gini index")
points(dataset$Gini.Index[dataset$religion.fac=='1'] ,dataset$Number.of.murders[dataset$religion.f
ac=='1'], col="green")
> abline(a=-17.8552,b=0.7177, col="purple")
> abline(a=-17.8552+21.4633, b=0.7177-0.7458, col='green')

> plot(dataset$Life.expectancy[dataset$religion.fac=='0'],dataset$Number.of.murders[dataset$religi
on.fac=='0'], col="purple",xlim = c(60,90),ylim=c(0,60),xlab="Life Expactancy",ylab="murders", mai
n = "murders vs. Life Expactancy")
```

```
> points(dataset$Life.expectancy[dataset$religion.fac=='1'] ,dataset$Number.of.murders[dataset$rel
igion.fac=='1'], col="green")
>
> abline(a=132.1998,b=-1.6036, col="purple")
>
> abline(a=132.1998-121.6926, b=-1.6036+1.4997, col='green')
```

GDP

```
plot(dataset$GDP[dataset$religion.fac=='0'],dataset$Number.of.murders[dataset$religion.fac=='0'],
col="purple",xlim = c(500,55000),ylim=c(0,60),xlab="GDP",ylab="murders", main = "murders vs. GDP")
> points(dataset$GDP[dataset$religion.fac=='1'] ,dataset$Number.of.murders[dataset$religion.fac=='
1'], col="green")
> abline(a=1.243e+01,b= -3.145e-04, col="purple")
> abline(a= 1.243e+01-9.630e+00, b=-3.145e-04+2.927e-04, col='green')
```

עוני

```
> plot(dataset$Poverty.rate[dataset$religion.fac=='0'],dataset$Number.of.murders[dataset$religion.
fac=='0'], col="purple",xlim = c(0,90),ylim=c(0,60),xlab="Poverty Rate",ylab="murders", main = "Mu
rders vs. Poverty Rate")
> points(dataset$Poverty.rate[dataset$religion.fac=='1'] ,dataset$Number.of.murders[dataset$religi
on.fac=='1'], col="green")
> abline(a=-2.1331,b= 0.5208, col="purple")
> abline(a= -2.1331+ 3.7031, b=0.5208-0.4761, col='green')
```

גירושין

```
> plot(dataset$Divorce.rate[dataset$religion.fac=='0'],dataset$Number.of.murders[dataset$religion.
fac=='0'], col="purple",xlim = c(0,70),ylim=c(0,70),xlab="Divorce Rate",ylab="murders", main = "Mu
rders vs. Divorce Rate")
> points(dataset$Divorce.rate[dataset$religion.fac=='1'] ,dataset$Number.of.murders[dataset$religi
on.fac=='1'], col="green")
> abline(a=16.73791,b=  -0.25466, col="purple")
> abline(a=16.73791-15.26996,b=  -0.25466+0.30644, col="green")
```

<u>הכרזת קטגוריאליים ומודל מלא עם האינטרקציות</u>

```
> install.packages("moments")
> library(moments)
> dataset<-read.csv(file.choose(),header = T)

> Death.penalty.permitted.category<-as.factor(dataset$Death.penalty.permitted)
> religion.category<-as.factor(dataset$religion.new)
> lm.full <- lm(Number.of.murders~Life.expectancy+Gini.Index+GDP+Poverty.rate+religion.category+Di
vorce.rate+Death.penalty.permitted.category+religion.category:Life.expectancy+religion.category:Po
verty.rate+Death.penalty.permitted.category:Poverty.rate, data=dataset)
```

שיטת Forward elimination:

<u>AIC</u>

```
> lm.null <- lm(dataset$Number.of.murders ~ 1, data = dataset)
> model.aic.forward <- step(lm.null, direction = "forward", trace = 1, scope =  ~Life.expectancy+G
ini.Index+GDP+Poverty.rate+religion.category+Divorce.rate+Death.penalty.permitted.category+religio
n.category:Life.expectancy+religion.category:Poverty.rate+Death.penalty.permitted.category:Poverty
.rate)
> summary(model.aic.forward)
```

<u>F חלקי</u>

```
> lm.full <- lm(Number.of.murders~Life.expectancy+Gini.Index+GDP+Poverty.rate+religion.category+Di
vorce.rate+Death.penalty.permitted.category+religion.category:Life.expectancy+religion.category:Po
verty.rate+Death.penalty.permitted.category:Poverty.rate, data = dataset)
> lm.null <- lm(Number.of.murders ~ 1, data = dataset)
> add1(lm.null, scope =  ~Life.expectancy+Gini.Index+GDP+Poverty.rate+religion.category+Divorce.ra
te+Death.penalty.permitted.category+religion.category:Life.expectancy+religion.category:Poverty.ra
te+Death.penalty.permitted.category:Poverty.rate, data = dataset, test = "F")
> add1(update(lm.null, ~ . +Gini.Index),  scope =~Life.expectancy+Gini.Index+GDP+Poverty.rate+reli
gion.category+Divorce.rate+Death.penalty.permitted.category+religion.category:Life.expectancy+reli
gion.category:Poverty.rate+Death.penalty.permitted.category:Poverty.rate, data = dataset, test = "
F")
> add1(update(lm.null, ~ . +Gini.Index+religion.category),  scope =~Life.expectancy+Gini.Index+GDP
+Poverty.rate+religion.category+Divorce.rate+Death.penalty.permitted.category+religion.category:Li
fe.expectancy+religion.category:Poverty.rate+Death.penalty.permitted.category:Poverty.rate, data =
dataset, test = "F")
> add1(update(lm.null, ~ . +Gini.Index+religion.category+Life.expectancy),  scope =~Life.expectanc
y+Gini.Index+GDP+Poverty.rate+religion.category+Divorce.rate+Death.penalty.permitted.category+reli
gion.category:Life.expectancy+religion.category:Poverty.rate+Death.penalty.permitted.category:Pove
rty.rate, data = dataset, test = "F")
> add1(update(lm.null, ~ . +Gini.Index+religion.category+Life.expectancy+Death.penalty.permitted.c
ategory),  scope =~Life.expectancy+Gini.Index+GDP+Poverty.rate+religion.category+Divorce.rate+Deat
h.penalty.permitted.category+religion.category:Life.expectancy+religion.category:Poverty.rate+Deat
h.penalty.permitted.category:Poverty.rate, data = dataset, test = "F")
```

שיטת Backward elimination:

AIC

```
> model.aic.backward <- step(lm.full, direction = "backward", trace = 1)
> summary(model.aic.backward)
```

F חלקי:

```
> lm.full <- lm(Number.of.murders~Life.expectancy+Gini.Index+GDP+Poverty.rate+religion.category+Di
vorce.rate+Death.penalty.permitted.category+religion.category:Life.expectancy+religion.category:Po
verty.rate+Death.penalty.permitted.category:Poverty.rate, data = dataset)
> lm.null <- lm(Number.of.murders ~ 1, data = dataset)
> drop1(lm.full, test = "F")
> drop1(update(lm.full, ~ . -GDP), test = "F")
> drop1(update(lm.full, ~ . -GDP-Poverty.rate:Death.penalty.permitted.category), test = "F")
> drop1(update(lm.full, ~ . -GDP-Poverty.rate:Death.penalty.permitted.category-Divorce.rate), test
= "F")
> drop1(update(lm.full, ~ . -GDP-Poverty.rate:Death.penalty.permitted.category-Divorce.rate-Povert
y.rate:religion.category), test = "F")
```

שיטת stepwise regression

```
> model.aic.both <- step(lm.null, direction = "both", trace = 1, scope = ~Life.expectancy+Gini.Ind
ex+GDP+Poverty.rate+religion.category+Divorce.rate+Death.penalty.permitted.category+religion.categ
ory:Life.expectancy+religion.category:Poverty.rate+Death.penalty.permitted.category:Poverty.rate)
> summary(model.aic.both)
```

בדיקת הנחות המודל:

מודל סופי

```
lm(formula = dataset$Number.of.murders ~ Gini.Index + religion.new +
    Life.expectancy + Death.penalty.permitted + Poverty.rate +
    religion.new:Life.expectancy, data = dataset)
```

הנחת שוויון שונויות

תרשים פיזור של השגיאות המתוקננות מול הערך הצפוי:

```
> data.lm = lm(Number.of.murders ~ Life.expectancy + Gini.Index + Poverty.rate + religion.new + De
ath.penalty.permitted + Life.expectancy:religion.new
+ , data=dataset)
> data.res = resid(data.lm)
> plot(dataset$Number.of.murders, data.res,ylab ="Residuals",xlab="Number Of Murders")
> abline(0,0)
```

הנחת הנורמליות של השגיאות:

```
> dataset$fitted<-fitted(data.lm)
> dataset$residuals<-residuals(data.lm)
> s.e_res<-sqrt(var(dataset$residuals))
> dataset$stan_residuals<-(residuals(data.lm)/s.e_res)
> qqnorm(dataset$stan_residuals)
> abline(a=0,b=1)

hist(dataset$stan_residuals,xlab="Normalized error", main= "Histogram of normalized error")
```

מבחן KS ו-SW:

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

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

שיפור המודל:

גרף BOXCOX

```
library(MASS)

modelBC<- boxcox(I(dataset$Number.of.murders+ 0.0005)~Life.expectancy + Gini.Index + Poverty.rate
+ religion.new + Death.penalty.permitted + Life.expectancy:religion.new, data=dataset,lambda = seq
(0, 0.5, 1/20))
> cbind(modelBC$x,modelBC$y)[order(modelBC$y),]
```

כתיבת מודל חדש עם העלאת Y בחזקת 0.29:

```
 NewMod2<-lm(dataset$Number.of.murders^(0.29)~dataset$Life.expectancy + dataset$Gini.Index + datas
et$Poverty.rate + dataset$religion.new + dataset$Death.penalty.permitted + dataset$Life.expectancy
:dataset$religion.new)
> NewMod2F<-fitted(NewMod2)
> res=resid(NewMod2)
> s.e_res<-sqrt(var(res))
> stan_res<-(residuals(NewMod2)/s.e_res)
> plot(NewMod2F, stan_res, ylab="Normalized error", xlab="Predicted value")
> abline(0,0)

> qqnorm(stan_res)
> abline(0,1)
 hist(stan_res, xlab= "Normalized error", main="Histogram of normalized error")
```