דו"ח פרויקט - חלק ב'

רגרסיה לינארית

313520389 | 205918477

תאריך הגשה :30.06.2022

תוכן עניינים

תקציר מנהלים

מטרת על – ביצוע תהליך ניתוח נתונים סטטיסטי מתחילתו ועד סופו להתאמת מודל רגרסיה מיטבי אשר יסביר בצורה הטובה ביותר את רווח מוסד אקדמאי.

מטרות משנה: 1) התאמת בסיס הנתונים לעבודה עם מספרים תוך התאמת משתני דמה ואינטרקציה.

2) בחירת מודל ראשוני ע"פ שלבי ביצוע מבוססים מדדי איכות.

3) ביצוע מבחנים לבדיקת קיום הנחות המודל וביצוע טרנספורמציות לשיפור המודל.

צעדים שננקטו:

על המודל שקיבלנו לאמוד את רווח מוסד אקדמי תוך שימוש במספר משתנים מסבירים. השתמשנו בטבלת הנתונים שקיבלנו בחלק א' שמכילה 8 משתנים מסבירים, 5 מהם רציפים ו-3 קטגוריאליים. בתחילת חלק ב' בדקנו כל אחד מהמשתנים וביצענו מבחן על סמך P_Value במטרה לראות עד כמה הוא נחוץ ומתאים להסברת רווח המוסד האקדמי. בשלב ראשוני זה, על פי רמת המובהקות שביצענו במבחן, הסרנו משתנים לפי קטלוג של נכנס או יוצא למודל. המשתנים **הוסרו רק עקב המבחן שביצענו**, ללא ניסיון לתמוך את הסרתם מאופן הקשרם לרווח המוסד.

לאחר מכן, ניתחנו את המשתנים המסבירים שנותרו **וביצענו התאמות** לקבוצות פנימיות **עבור כל משתנה מסביר בנפרד**. לאלו אשר היה צורך התאמנו משתני דמה ואינטרקציה. מכיוון שנותרנו לאחר ההסרה רק עם משתנה רציף בודד, **בדקנו** עליו את כל **האינטרקציות האפשריות**. לאחר הגדרת משתני המודל מתוך האינטרקציות שנבחרו, ניסינו אלגוריתמי S.R , B.E , F.S ובחרנו **שהמדד** המתאים ביותר לנו **להמשך הבדיקות יהיה** R_{adj}. בנוסף, השתמשנו במהלך חלק זה במדדים AIC, BIC. בדקנו על ידי מספר מבחנים את הנחות המודל (נורמליות, ליניאריות ושוויון שונויות)**, בין המבחנים:** Shapiro-Wilk, Chow, KS, Goldfeld-Quandt. לבסוף, ביצענו טרנספורמציות למשתנים המסבירים כדי לקיים את הנחות המודל.

לשיפור המודל, ביצענו **מבחני פירסון מחודשים** עבור טרנספורמציות שונות **על המשתנים שהסרנו** בתחילת התהליך ע"מ לבדוק האם בעזרתם נוכל למצוא מודל שערך המדד שלו יהיה טוב יותר. בעזרת שימוש נוסף של האלגוריתמים לבחירת משתני המודל, **מצאנו את המודל הטוב ביותר** שהצלחנו לקבל משלל **האפשרויות שניסינו.**

מסקנות- על אף שבחרנו להסיר משתנים מסוימים בתחילת העבודה על חלק זה, גילינו שתרומתם למודל ביחד עם משתני אינטרקציה נוספים יכולה להיות גדולה יותר ע"פ המדד שבחרנו. מסקנה נוספת היא שביצוע טרנספורמציות למשתנים מסבירים הוא השלב שהיווה את השיפור הטוב ביותר עבור המדד שבחרנו.

תוצאות- עבור המודל הסופי שלנו, ערכו של R_{adj} = 0.6443, שיפור של 17% לעומת המודל הראשוני לפני ניסיון שיפור המודל. המודל הסופי מכיל 6 משתנים מסבירים, שחלקם הוגדרו כמשתני דמה, בנוסף ישנם עוד 12 משתני אינטרקציה שנוצרו כתוצאה מהתאמת המודל.

אנו צופים כי צלחנו את המטרה והתאמנו את מודל הרגרסיה הטוב ביותר לנתונים שקיבלנו.

טבלת המשתנים מחלק א' של הפרויקט - **Student's Earnings.**

שם משתנה	מוסבר/ מסביר	סימון במודל	יחידת מידה	רציף/ קטגוריאלי	הסבר קצר על המשתנה
אינדקס למל"ג בארה"ב	אינדקס	row_id	-	-	המספר המזהה של כל מוסד במל"ג
התואר הגבוה ביותר המוענק במוסד	מסביר	highest_degrees_awarded	-	קטגוריאלי	סוג התואר הגבוה ביותר שניתן לקבל במוסד
שכר סגל ממוצע	מסביר	faculty_salary	$	רציף	כמות כספית שמקבל חבר סגל במוסד
שיעור הסגל במשרה מלאה	מסביר	ft_faculty_rate	%	רציף	אחוז הסגל שעובדים במשרה מלאה במוסד
שליטה במוסד	מסביר	ownership	-	קטגוריאלי	האם המוסד ציבורי או פרטי עם/בלי רווחים
גיל כניסה ממוצע	מסביר	demographics_age_entry	מספר שנים	רציף	מספר השנים הממוצע של סטודנט בעת הכניסה למוסד לימודים
נתח סטודנטיות	מסביר	demographics_female_share	%	רציף	אחוז הסטודנטיות מכלל הלומדים במוסד
סוג לימודים	מסביר	online_only	בינארי	קטגוריאלי	האם הלימודים פרונטליים או לא. 1- מייצג רק פרונטלי.
אחוז סטודנטים שרמת ההשכלה הגבוהה של הוריהם היא תיכון	מסביר	parents_highschool	%	רציף	ערך מספרי שמגדיר את רמת ההשכלה של הורי הסטודנטים הלומדים במוסד
רווח המוסד באלפי דולרים	מוסבר	income	אלפי דולרים (K$)	רציף	כמות הכסף שהמוסד מרוויח באלפי דולרים

עיבוד מקדים

הסרת משתנים

עבור בדיקת איזה משתנים כדאי לנו להוציא, בחרנו לבצע בעזרת **מבחן פירסון** אשר מתאר קשר לינארי בין שני משתנים. המבחן נותן מידע על מקדם המתאם של פירסון ועל רמת המובהקות (P_Value) בין אחד מהמשתנים המסבירים למשתנה המוסבר. את המשתנים בעלי רמת מובהקות גבוהה מ-0.05 וכן משתנים בעלי מקדם מתאם שקרוב ל-0 נבחר להסיר מהמודל בשלב זה. נספח 1

שם משתנה\ סוג קשר	שכר סגל ממוצע	שיעור הסגל במשרה מלאה	גיל כניסה ממוצע	נתח סטודנטיות	אחוז סטודנטים שרמת ההשכלה הגבוהה של הוריהם היא תיכון	התואר הגבוה ביותר המוענק במוסד	סוג לימודים	שליטה במוסד
מקדם מתאם ע"פ פירסון	0.521	0.161	-0.075	-0.351	-0.118	0.354	0.007	-0.201
P_Value	5.265e-11 ~ 0	0.059	0.377	2.337e-5	0.167	2.031e-5	0.933	0.017

ניתן לראות ע"פ הטבלה שישנם משתנים אשר בהתאמה מקדם המתאם שלהם קטן יחסית (כלומר קרוב יותר ל-0 מאחד משני הצדדים) ורמת המובהקות שלהם גבוהה מ- α שנבחרה (0.05). את משתנים אלו נבחר להסיר מהמודל שלנו- שיעור הסגל במשרה מלאה, גיל כניסה ממוצע, אחוז סטודנטים שרמת ההשכלה של הוריהם היא תיכון וסוג הלימודים.
לאחר הסרת ארבעת המשתנים המסבירים הללו, נישאר עם 4 משתנים מסבירים שנותרו לנו.

לאחר הסרת המשתנים הרלוונטיים, ננתח את המשתנים שנותרו:

- **שכר סגל ממוצע** – לפי הטבלה נראה כי בין משתנה זה לרווח המוסד קיים קשר לינארי חיובי גבוהה (יחסית לשאר המשתנים המסבירים) וכן רמת המובהקות של הקשר שואפת ל-0 (10 אפסים אחרי הנקודה). נרצה להשאירו כי הוא אכן משפיע, מה שמתאים הגיונית היות ששכר סגל גבוה מצביע על איכות למידה גבוהה שיכולה להתאים לתארים מתקדמים שעולים בספית יקר יותר, לכן ההערכה היא שרווח מוסד הלימודים יושפע חיובית.

- **נתח סטודנטיות** – לפי הטבלה נראה כי בין משתנה זה לרווח המוסד קיים קשר לינארי שלילי וכן רמת המובהקות נמוכה בהרבה מ-0.05. נרצה להשאיר משתנה משפיע זה בגלל הנתונים הגבוהים אשר מראים על קשר מסוים, נתונים שבהמשך התהליך הניתוחי נוכל להסבירם בצורה טובה יותר מעין הם נובעים.
- **התואר הגבוה ביותר המוענק במוסד** – לפי הטבלה נראה כין משתנה זה לרווח המוסד קיים קשר לינארי חיובי ורמת המובהקות נמוכה בהרבה מ-0.05. תארים גבוהים יותר עולים כספית יקר יותר לסטודנטים לכן הרווחים של מוסד הלימודים עליהם גבוהים יותר ומכאן נובע הקשר הלינארי החיובי. נרצה להשאיר משתנה משפיע זה.
- **שליטה במוסד** – ניתן לראות עבור משתנה זה כי אכן רמת המובהקות שלו קטנה מ-0.05 ואכן יש לו סוג של קשר לינארי. עם זאת, חלוקת הדרגות שעשינו [1:3] לפי כל סוג שליטה במוסד הייתה לפי הערכה שלנו (כאשר 1 מייצג מוסד פרטי ללא מטרות רווח- לכן נצפה לרווח נמוך יותר, ו-3 הינו מוסד פרטי למטרות רווח- לכן נצפה לרווח גדול יותר). בהמשך התהליך נבחן האם יש צורך להתאים את החלוקה ולשנות אותו כך שתיתן הסבר טוב אף יותר מאשר ההערכה ההתחלתית שלנו.

התאמת משתנים

לאחר בחינת המשתנים המסבירים שנותרו לנו, מצאנו לנכון לבצע 2 פעולות על הנתונים- דיסקרטיזציה עבור נתח הסטודנטיות ואיחוד קטגוריות בין המוסדות שבהם התואר הגבוה ביותר שמוענק שונה.

עבור הדיסקרטיזציה, ניסינו תחילה לחלק את הנתונים בהקשר רוב נשים (<50%) מכלל הסטודנטים. ראינו כי בעת חלוקה, אמנם ישנם הבדלים בין הקבוצות השונות אך ישנה קבוצה שלמה שמתפספסת ומבליטה את הלינאריות בקשר. לכן בחרנו לחלק את המשתנה לשלוש קטגוריות שונות: רוב לאחוז הנשים, רוב לאחוז הגברים וקבוצת ביניים שאין לה רוב מובהק עבור אחד מהמינים ($50\% \pm 10\%$). כאשר השתמשנו בפקודת Summary זיהינו כי יש בסיס להמרת הנתונים מבחינת הפרש הממוצעים הלינארי בין 3 הקבוצות, כאשר בין קבוצת הביניים לקבוצה המובהקת של הנשים ישנו פער של יותר מ6 אלף דולר בממוצע, בעוד הפער בין קבוצות הקצה אף גבוה יותר – מעל 10 אלף דולר. נספח 2

עבור איחוד הקטגוריות, מכיוון שישנם רק 2 מוסדות בהם התואר הגבוה ביותר שמוענק הינו Certified degree, יש צורך באיחוד נתונים עבור קטגוריה זו שכן לא ניתן לנתח את הנתונים מ-2

דגימות ולהסיק מכך על הכלל – אינה מהווה מדגם מייצג. בנוסף, על ידי פונקציית Summary ניתן לראות כי קטגוריות Non-Degree (1) ו- Graduate-Degree (5) קרובות מבחינת נתוני הממוצע והחציון, כך שהחלטנו שאיחוד קטגוריות אלו יוכל להועיל לנו. נספח 3

הגדרת משתני דמה

יש צורך בהגדרת משתני דמה עבור כל אחד מהמשתנים שמוצג בצורה קטגוריאלית. לאחר הדיסקרטיזציה שביצענו לנתח הסטודנטיות בסעיף הקודם, ישנם 3 משתנים שנצטרך להגדיר להם משתני דמה ובעקבות זאת- משתני אינטרקציה תואמים.

נתח הסטודנטיות: קבוצת הבסיס תהיה קטגוריה 1 בה יש רוב מובהק לאחוז הגברים במוסד הלימודים (פחות מ-40% נשים) ותקבל את הערך 1. קבוצה 2 שהינה קבוצת הביניים תקבל את הערך 1 אם אין רוב מובהק לאחד המינים. קבוצה 3 תקבל את הערך 1 אם יש רוב מובהק לאחוז הנשים (יותר מ-60%).

$$A_2 = \begin{cases} 1, if \;\; 40\% \leq female\_share \leq 60\% \\ 0, else \end{cases} \qquad A_3 = \begin{cases} 1, if \;\; female\_share \geq 60\% \\ 0, else \end{cases}$$

שליטה במוסד: קבוצת הבסיס הינה קבוצה 1 -מוסדות פרטיים ללא מטרות רווח- ותקבל את הערך 1. קבוצה 2 תקבל את הערך 1 כאשר המוסד הינו ציבורי. קבוצה 3 תקבל את הערך 1 כאשר המוסד פרטי למטרות רווח (מתואר בטבלה כאשר ערך הקבוצה הינו 3).

$$B_2 = \begin{cases} 1, if \;\; Ownership \;=\; Public \; (2) \\ 0, else \end{cases} \qquad B_3 = \begin{cases} 1, if \; Ownership = Private \; for \; profit \; (3) \\ 0, else \end{cases}$$

התואר הגבוה ביותר המוענק במוסד: קבוצת הבסיס היא הקבוצה המאוחדת- מוסדות שבם התואר הגבוה ביותר הוא אחד מ- Certified/Associate ותקבל את הערך 1. קבוצה 5 תקבל את הערך 1 כאשר התואר הוא Graduate/Non-Degree. קבוצה 4 תקבל את הערך 1 כאשר התואר הוא Bachelor's.

$$C_4 = \begin{cases} 1, if \;\; degree \;=\; Bachelor's \\ 0, else \end{cases} \qquad C_5 = \begin{cases} 1, if \; degree \;=\; Graduate/Non-Degree \\ 0, else \end{cases}$$

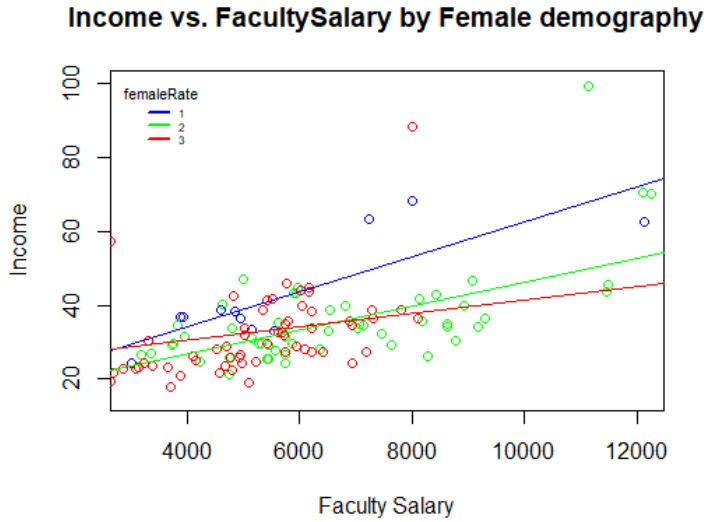
הוספת משתני אינטרקציה

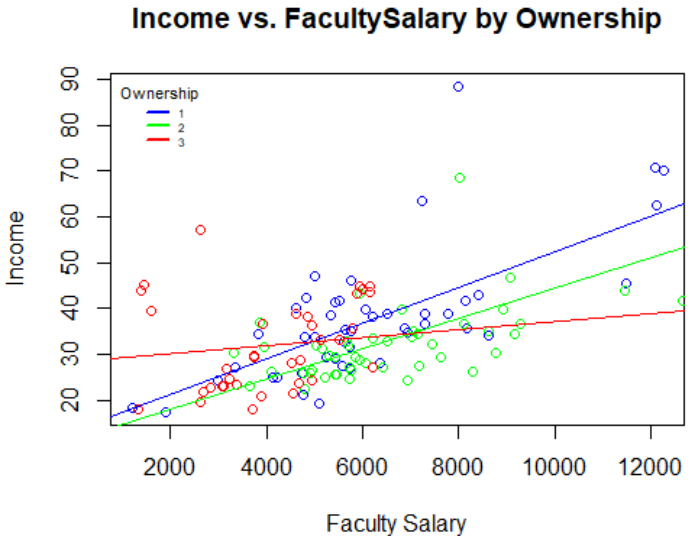
לאחר הגדרת משתני הדמה הרלוונטיים שיעזרו באמידת החותך, נרצה לאמוד את התרומה השולית לשיפוע שנובעת מהוספת משתנה אינטרקציה. נרצה שהוספת משתנים אלו יעזרו לבדוק את השפעת הקטגוריות השונות על קו הרגרסיה, כל אחד עבור המשתנה המסביר המתאים לו. נצפה לקבל השפעות שונות על רווח המוסד כתוצאה מפילוג התוצאות המתקבלות בעזרת שילוב משתנה הדמה עם משתנה האינטרקציה במודל הרגרסיה.

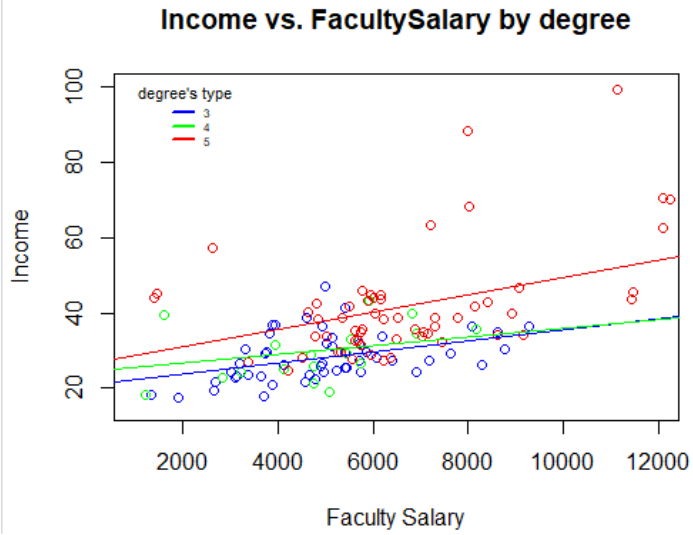
השתמשנו בתרשים פיזור Plot וכן בפונקציית abline שמציירת על גבי התרשים קו רגרסיה לינארי. באמצעות התרשים נוכל לבחון האם ישנו קשר בין המשתנה המוסבר לאחת מהקבוצות השונות שחילקנו בכל קטגוריה (משתנה מסביר). כמו כן נבדוק מקרים בהם ישנה אינטרקציה בין 2 קטגוריות שונות (משתנים מסבירים).

בעקבות התאמת המשתנים, נשארנו עם משתנה רציף יחיד – שכר סגל ממוצע. בהתאם לכך, ביצענו מספר מבחנים על מנת לבדוק האם קיימת אינטרקציה בינו לבין אחד מהמשתנים הקטגוריאליים שנותרו לנו – נתח הסטודנטיות, שליטה במוסד, התואר הגבוה ביותר המוענק במוסד.

לאור שלושת תרשימי הפיזור, החלטנו להכניס כמשתני אינטרקציה את שלושת המשתנים הקטגוריאליים שבחנו. בכל אחד מתרשימי הפיזור ניתן לראות שיפועים שונים עבור קבוצות שונות. בתרשימים על פי סוג התואר הגבוה ביותר ועל פינתח הסטודנטיות, ניתן לראות מגמות עלייה בכל אחת מתתי הקבוצות, חלקם בשיפוע גדול יותר וחלקם בקטן יותר. הקבוצות היחידות שהתלבטנו האם להכניסם היו תחת משתנה Ownership, כאשר לקבוצות 1 ו-2 ישנם שיפועים דומים, אך מפאת שלא הצלחנו לקשר מבחינה הגיונית בין הנתונים בחרנו לא לאחד או לפסול משתנה אינטרקציה זה. נספח 4







התאמת המודל ובדיקת הנחות

בחירת משתני המודל

לאחר בחינת משתני האינטראקציה אשר משפיעים על המשתנה המוסבר, קיבלנו את המודל הבא:

X – שכר סגל ממוצע.

A_2 – קבוצת בה אחוז הנשים בין 40-60%.

A_3 – קבוצה בה ישנו רוב נשי מובהק.

B_2 – בעלות ציבורית על המוסד.

B_3 – בעלות פרטית למטרות רווח במוסד.

C_4 – מוסדות בהם התואר הגבוה ביותר הינו Bachelor's.

C_5 – מוסדות בהם התואר הגבוה ביותר הינו Graduate/Non-Degree.

β_0 – תוחלת רווח מוסד עבור שכר סגל ממוצע במוסדות בהן יש רוב גברי מובהק (A_1), הבעלות הינה פרטית ללא מטרות רווח (B_1) וכן התואר הגבוה ביותר שמוענק הוא אחד מ Certified/Associate(C_3).

***A_1, B_1, C_3 – קבוצת הבסיס.**

$\beta_0 + \beta_{0j}$ – תוחלת רווח המוסד עבור שיעור נשים j (j=2,3).

$\beta_0 + \beta_{0k}$ – תוחלת רווח המוסד עבור סוג השליטה במוסד k (k=2,3).

$\beta_0 + \beta_{0i}$ – תוחלת רווח המוסד עבור סוג התואר הגבוה ביותר שניתן i (i=4,5).

β_{0j} – תוספת שולית לתוחלת רווח המוסד עבור שיעור נשים j (j=2,3), מעבר לתוחלת רווח המוסד ברוב גברי מובהק (A_1)

β_{0k} - תוספת שולית לתוחלת רווח המוסד עבור שליטה במוסד k (k=2,3), מעבר לתוחלת רווח המוסד שנמצאת בבעלות פרטית ללא מטרות רווח (B_1).

β_{0i} - תוספת שולית לתוחלת רווח המוסד עבור סוג התואר הגבוה ביותר שניתן i (i=4,5).

$$Y = \beta_0 + \beta_1 X + \beta_{02} A_2 + X\beta_{12} A_2 + \beta_{03} A_3 + X\beta_{13} A_3 + \beta_{04} B_2 + X\beta_{14} B_2 + \beta_{05} B_3 + X\beta_{15} B_3 + \beta_{06} C_4 + X\beta_{16} C_4 + \beta_{07} C_5 + X\beta_{17} C_5$$

הקריטריון שבחרנו להשתמש עבור בחירת החלופה המועדפת מבין כלל הקריטריונים שלמדנו במהלך הקורס הינו R^2_{adj}, כיוון שזה הקריטריון המתאים ביותר אם ברצוננו לאמוד את מובהקות התוצאה ברגרסיה מרובה. את מדד זה נרצה למקסם. בנוסף, ניעזר בשני מדדים נוספים AIC, BIC אשר מתחשבים במספר התצפיות וכן מספר הפרמטרים במובהקות התוצאה, את מדדים אלו נרצה למזער.

ערך המדד ע"פ המודל המלא: $R^2_{adj} = 0.4615$, AIC = 623.87, BIC = 628.1 (נספח 5)

למציאת המודל המתאים, השתמשנו ב-3 סוגי האלגוריתמים שנלמדו המשתמשים בגישות שונות:

1. Forward Selection – המודל מתחיל ריק, כלומר ללא משתנים. בכל איטרציה נכניס אליו את המשתנה אשר מקבל את ה- F_{st} הגדול ביותר (כלומר המובהק ביותר). נעצור את האלגוריתם כאשר מתוך המשתנים אשר ניתן להכניס, נדחה את השערת H_0 עבור המשתנה בעל ה- F_{st} המירבי.
2. Backward Elimination – הגישה מתחילה מהמודל המלא, כאשר בכל איטרציה מתבצע מבחן F חלקי והמשתנה הכי פחות מובהק יצא מהמודל. נעצור את האלגוריתם כאשר נדחה את השערת H_0 למועמד הטוב ביותר שנשאר לנו להוציא.
3. Stepwise Regression – גישה שמשלבת צעידה לאחור וצעידה לפנים, בכל שלב ניתן להחליט האם להוציא משתנה או להכניסו. נעצור כאשר יש דחייה להשערת H_0 עבור המועמד הטוב ביותר לצאת וכן נדחה גם עבור המועמד הטוב ביותר להיכנס.

ניתן לראות כי בסיכום שלב זה, לאחר הרצת שלושת השיטות שנכתבו לעיל, עבור שיטת Backward Elimination קיבלנו את ערך המדד הטוב ביותר גם מבחינת $R^2_{adj} = 0.4768$ גם AIC = 616.26 וגם BIC = 619.27. נספח 6

כלומר הגישה הטובה ביותר הגיעה למודל הבא:

$$Y = \beta_0 + \beta_1 X + \beta_{02} A_2 + \beta_{03} A_3 + \beta_{04} B_2 + X\beta_{14} B_2 + \beta_{05} B_3 + X\beta_{15} B_3 + \beta_{06} C_4 + \beta_{07} C_5$$

ולאחר הצבת ה-β :

$$Y = 20.79 + 3.062 * 10^{-3} * X - 11.79 * A_2 - 10.14 * A_3 + 3.888 * B_2 - 6.731 * 10^{-4} * X * B_2 + 17.73 * B_3 - 3.638 * 10^{-3} * X * B_3 + 3.101 * C_4 + 10.92 * C_5$$

בדיקת הנחות המודל

הנחת הלינאריות:

```
> ##בדיקת הנחות המודל###
>
> #perform Chow test
>
> dataset_high <- subset(dataset, dataset$faculty_salary > 5500)
> dataset_low <- subset(dataset, dataset$faculty_salary < 5500)
>
> xxx <- lm(dataset_high$income~dataset_high$faculty_salary, data=dataset_high)
> anova(xxx) # SSE_High = 9681.8
Analysis of Variance Table

Response: dataset_high$income
                          Df Sum Sq Mean Sq F value    Pr(>F)
dataset_high$faculty_salary  1 3547.5  3547.5  24.549 5.192e-06 ***
Residuals                   67 9681.8   144.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> xxy <- lm(dataset_low$income~dataset_low$faculty_salary, data=dataset_low)
> anova(xxy) # SSE_Low = 4780.4
Analysis of Variance Table

Response: dataset_low$income
                          Df Sum Sq Mean Sq F value Pr(>F)
dataset_low$faculty_salary  1    0.2   0.217   0.003 0.9562
Residuals                  67 4780.4  71.349
>
> xxz <- lm(dataset$income~dataset$faculty_salary, data=dataset)
> anova(xxz) #Residuals --> SSE_regular = 15457.3
Analysis of Variance Table

Response: dataset$income
                      Df  Sum Sq Mean Sq F value    Pr(>F)
dataset$faculty_salary   1  5784.1  5784.1  50.891 5.265e-11 ***
Residuals              136 15457.3   113.7
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> # SSE_AVG= SSE_regular - SSE_Low - SSE_High = 15457.3 - 9681.8 - 4780.4 = 995.1
>
> F_st <- (995.1/2)/((9681.8 + 4780.4)/134)  #(= 4.610066)
> # F_cr = 3.072 (From F_table)
> # F_cr < F_st ---> Rejected! No linear
```

על מנת לבדוק את הנחת הלינאריות ביצענו מבחן Chow, חישבנו את הערכים השונים על פי המבחן וקיבלנו ש- F_cr < F_st ולכן נדחה את השערת H_0 כלומר אנו לא מקבלים את הנחת הלינאריות במודל.

הנחת שוויון השונויות:

```
> #perform the Goldfeld Quandt test
> gqtest(xxz, order.by = ~dataset$faculty_salary, data = dataset, fraction = 27)

        Goldfeld-Quandt test

data:  xxz
GQ = 2.1021, df1 = 54, df2 = 53, p-value = 0.003789
alternative hypothesis: variance increases from segment 1 to 2

> # GQ = F_st = 2.1021, P_value = 0.003789
> # F_cr by df1=54, df2=53 --> 1.5343
> #F_cr < F_st --> Rejected! There is Heteroskedasticity
```

על מנת לבדוק את ההנחה ביצענו מבחן Goldfeld-Quandt בו קיבלנו כי P_Value נמוך מאוד ושווה ל – 0.003789, מה שמצביע על כך שפיזור השונויות אינו אקראי ובעל סימטריה סביב ציר X. לכן נדחה את השערת H_0 ונאמר כי הנחת שוויון השונויות אינה מתקיימת.

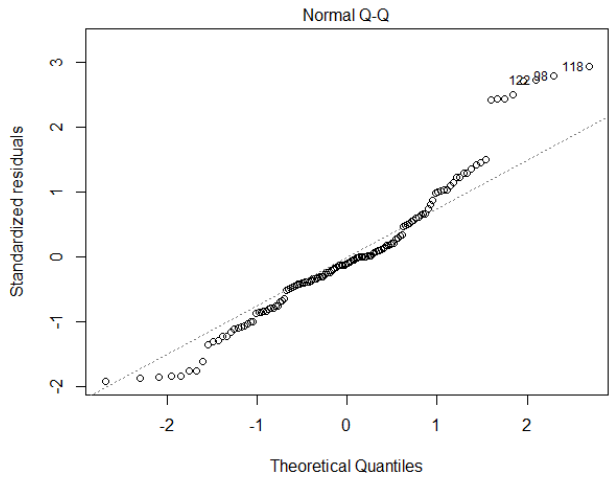
הנחת הנורמליות של השגיאות:

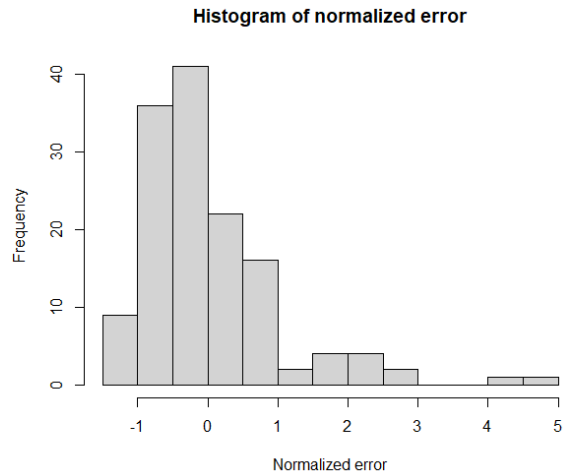
```
> # K.s and Shapiro tests for normalized
> mod1 <- lm(dataset$income~dataset$faculty_salary, data=dataset)
> dataset$fitted<-fitted(mod1) # predicted values
> dataset$residuals<-residuals(mod1) # residuals
> s.e_res <- sqrt(var(dataset$residuals))
> dataset$stan_residuals<-(residuals(mod1)/s.e_res)
>
> shapiro.test(dataset$stan_residuals) # בדיקת נורמליות לא מתקיימת - p-value = 1.472e-11

        Shapiro-Wilk normality test

data:  dataset$stan_residuals
W = 0.82949, p-value = 2.309e-11

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

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  dataset$stan_residuals
D = 0.15493, p-value = 0.002655
alternative hypothesis: two-sided
```

כדי לבדוק האם השגיאות מתפלגות נורמלית, נבצע מספר מבחנים ונבחן אותם לפי התוצאות המתקבלות ממבחנים אלו. לפי התוצאות שקיבלנו מההיסטוגרמה שעשינו ניתן לראות שההתפלגות אינה מזכירה את ההתפלגות הנורמלית, כלומר-

התפלגות א-סימטרית בעלת זנב ימני. במבחנים Kolmogorov-Smirnov ו-Shapiro קיבלנו כי

0.05 > P_value ולכן **נדחה את השערת H_0** ונאמר כי השגיאות המתוקננות אינן מתפלגות נורמלית.





שיפור המודל

לאחר בדיקת ההנחות והבנתנו שהן אינן מתקיימות, החלטנו להתחיל מביצוע טרנספורמציות על המודל על מנת לקיים אותן. בחרנו לעבוד כך שבכל שינוי שנעשה, או ביצוע טרנספורמציה, נבדוק שאכן מתבצע שיפור ואנו מאששים את הנחות המודל.

תחילה בחרנו ליצור שוויון שוניות על ידי ביצוע טרנספורמציות: $\sqrt{Y}$, ln(Y), אותן נבצע כמובן על המשתנה המוסבר. ביצענו את הטרנספורמציות הבאות על המודל הנוכחי ובדקנו שוב בעזרת מבחן Goldfeld האם כעת באחת מהאופציות הנחת שוויון השוניות מתקיימת. מצאנו כי אכן עבור שתי הטרנספורמציות מתקיימת ההנחה, עבור ln(Y) רמת

```
Step:  AIC=-415.13
dataset$incomeLan ~ highest_degrees_Dum + dataset$faculty_salary +
    female_Dum

                                          Df Sum of Sq    RSS     AIC
<none>                                                  6.2470 -415.13
+ dataset$faculty_salary:highest_degrees_Dum  2   0.07150 6.1755 -412.72
+ dataset$faculty_salary:female_Dum           2   0.05889 6.1881 -412.44
+ ownership_Dum                               2   0.02942 6.2176 -411.78
- female_Dum                                  2   0.82231 7.0693 -402.07
- dataset$faculty_salary                      1   1.44421 7.6912 -388.43
- highest_degrees_Dum                         2   2.59997 8.8470 -371.11
>
> summary(bwd.model)

Call:
lm(formula = dataset$incomeLan ~ dataset$faculty_salary + highest_degrees_Dum +
    female_Dum + ownership_Dum + dataset$faculty_salary:highest_degrees_Dum +
    dataset$faculty_salary:ownership_Dum)

Residuals:
     Min       1Q   Median       3Q      Max
-0.37460 -0.12919 -0.03307  0.12719  0.79602

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                              2.989e+00  1.378e-01  21.695  < 2e-16 ***
dataset$faculty_salary                   1.074e-04  2.223e-05   4.832 3.85e-06 ***
highest_degrees_Dum4                     2.138e-01  1.588e-01   1.346 0.180704
highest_degrees_Dum5                     5.406e-01  1.205e-01   4.487 1.61e-05 ***
female_Dum2                             -3.111e-01  7.352e-02  -4.232 4.43e-05 ***
female_Dum3                             -2.948e-01  6.977e-02  -4.225 4.54e-05 ***
ownership_Dum2                           2.138e-01  1.424e-01   1.501 0.135798
ownership_Dum3                           4.506e-01  1.330e-01   3.389 0.000937 ***
dataset$faculty_salary:highest_degrees_Dum4 -2.459e-05  3.214e-05  -0.765 0.445634
dataset$faculty_salary:highest_degrees_Dum5 -4.036e-05  2.074e-05  -1.947 0.053798 .
dataset$faculty_salary:ownership_Dum2   -3.389e-05  2.053e-05  -1.650 0.101369
dataset$faculty_salary:ownership_Dum3   -8.800e-05  2.693e-05  -3.267 0.001399 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2116 on 126 degrees of freedom
Multiple R-squared:  0.576,     Adjusted R-squared:  0.539
F-statistic: 15.56 on 11 and 126 DF,  p-value: < 2.2e-16

> extractAIC(bwd.model,k=log(10)) #BIC value of bwd.model
[1]   12.0000 -413.5854
> #bw.model is still the best model --> AIC= -417.22 , BIC = -413.58 , R-adj = 0.539
```

המובהקות הייתה גדולה בהרבה ולכן בחרנו להמשיך עם טרנספורמציה זו. לאחר מצב זה רצינו לוודא שערכי המדד שלנו אינם משתנים ולכן הרצנו שוב את אלגוריתמי הרגרסיה לאחור, לפנים ובצעדים. ניתן לראות כי במצב זה הערכים השתפרו מאוד וכעת מדד R_{adj} = 0.539.

```
> ####שיפור המודל ######
>
> #יצירת שיוויון שונויות#
>
> dataset$incomeSquareRoot <- sqrt(dataset$income) #הוספת עמודת שורש הרווח
> dataset$incomeLan <- ln(dataset$income) #הוספת עמודת לאן הרווח
>
> rootModel <- lm(dataset$incomeSquareRoot~dataset$faculty_salary, data=dataset)
> lnModel<- lm(dataset$incomeLan~dataset$faculty_salary, data=dataset)
>
> gqtest(rootModel, order.by = ~dataset$faculty_salary, data = dataset, fraction = 27)

        Goldfeld-Quandt test

data:  rootModel
GQ = 1.2802, df1 = 54, df2 = 53, p-value = 0.1849
alternative hypothesis: variance increases from segment 1 to 2

> #GQ = 1.2802, p-value = 0.1849 --> מתקיים שיוויון שונויות
> gqtest(lnModel, order.by = ~dataset$faculty_salary, data = dataset, fraction = 27)

        Goldfeld-Quandt test

data:  lnModel
GQ = 0.81186, df1 = 54, df2 = 53, p-value = 0.7761
alternative hypothesis: variance increases from segment 1 to 2

> #GQ = 0.81186, p-value = 0.7761 --> מתקיים שיוויון שונויות - מובחקות טובה יותר
>
```

לאחר מכן, בדקנו עבור המודל החדש האם הנחת הלינאריות מתקיימת. ביצענו את מבחן Chow עם הערכים החדשים, מתוך טבלאות Anova החדשות לקחנו את ערכי SSE וחישבנו את SSE_Avg. קיבלנו F_st = 2.39 , אשר קטן מ-F_cr מתוך הטבלה ולכן ניתן לומר כי לא נדחה את השערת H_0 ונאמר כי המודל מקיים את הנחת הלינאריות.

```
> # SSE_AVG= SSE_regular - SSE_Low - SSE_High = 9.4308 - 5.0156 - 4.0899 = 0.3253
> (0.3253/2)/((4.0899 + 5.0156)/134)
[1] 2.393619
> F_st <- (0.3253/2)/((4.0899 + 5.0156)/134)  #(= 2.393619)
> # F_cr = 3.072 (From F_table)
> #F_st= 2.3936 ,F_cr = 3.072 --> F_st < F_cr --> לא נדחה את השערת האפס ונאמר כי המודל לינארי
```

בעת עברנו לשלב שיפור המודל, תוך ניסיון לאשש את הנחת הנורמליות על ידי ביצוע טרנספורמציות למשתנים המסבירים. עבור המשתנה הרציף היחיד שנשאר לנו במודל הנוכחי – שכר סגל ממוצע- בחרנו לנסות ולבדוק 3 סוגי טרנספורמציות שיוכלו לשפר אותו - $\sqrt{X}$, log(X), X^2. מצאנו כי עבור X^2 הקורלציה שלו עם המשתנה המוסבר החדש היא הטובה ביותר. עם זאת, המדד איתו בחרנו לעבוד ירד (0.526) ולכן בחרנו לבדוק דרכים נוספות לשיפור המודל.

בחרנו לנסות ולהכניס חלק מהמשתנים שהסרנו בסעיף הסרת משתנים. מכיוון שבחלק מהמשתנים אותם הסרנו, לא בחנו עד כמה רמת ה-*P_Value* שלהם קרובה ל0.05 אלא רק האם הם מובהקים על פי הגדרה או לא, במידה ונכניס אותם כעת למודל קיים סיכוי לא רע שהמשתנים בעלי רמת מובהקות קרובה ל5% יתרמו להסברת המשתנה המסביר ויתאימו למודל החדש בעל הטרנספורמציה. נספח 7

שם משתנה	טרנספורמציה	מקדם מתאם ע"פ פירסון	P_Value
שיעור הסגל במשרה מלאה	X	0.1611396	0.05901
	Log(X)	0.1409539	0.09914
	X^2	0.1570465	0.06584
	$\sqrt{X}$	0.1559055	0.06785
גיל כניסה ממוצע למוסד	X	-0.07572984	0.3773
	Log(X)	-0.1078483	0.208
	X^2	-0.04266536	0.6193
	$\sqrt{X}$	-0.09196915	0.2833
אחוז סטודנטים שרמת ההשכלה הגבוהה של הוריהם היא תיכון	X	-0.1182187	0.1673
	Log(X)	-0.1424276	0.09563
	X^2	-0.07574908	0.3772
	$\sqrt{X}$	-0.1330885	0.1197

ניתן לראות מהטבלה כי עבור כל משתנה שהסרנו צמאנו מה הטרנספורמציה המתאימה ביותר עבור – ע"פ מבחן פירסון על רווח המוסד האקדמי. עבור המשתנה המתאר את שיעור הסגל במשרה מלאה, במידה ולא נבצע עליו טרנספורמציה נקבל את רמת המובהקות הקרובה ביותר ל5% לכן נבחר להוסיף משתנה זה ללא טרנספורמציה. עבור 2 המשתנים האחרים, ניתן לראות כי טרנספורמציית Log(X) תיתן עבורם את רמת המובהקות הטובה ביותר.

בנוסף, בשלב זה בחנו האם כדאי להכניס רק את חלק מהמשתנים, או את כולם. התחלנו בהוספת המשתנה בעל רמת המובהקות הטובה ביותר עד זה שבעל מובהקות נמוכה ביחס לשאר. ראינו שלאחר כל הוספה של משתנה למודל, אנו ממשיכים לשפר את המדד שבחרנו לעבוד על פיו. כמו כן, גם המדדים הנוספים (AIC,BIC) איתם ביצענו בדיקה נוספת הלכו ונהיו יותר ויותר קטנים.

```
Residual standard error: 0.212 on 117 degrees of freedom
Multiple R-squared:  0.6047,    Adjusted R-squared:  0.5371
F-statistic:  8.95 on 20 and 117 DF,  p-value: 1.444e-15
```

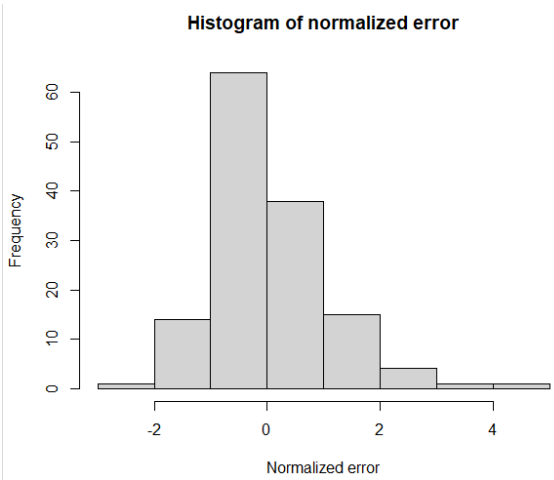
```
Residual standard error: 0.1932 on 103 degrees of freedom
Multiple R-squared:  0.7111,    Adjusted R-squared:  0.6157
F-statistic: 7.457 on 34 and 103 DF,  p-value: 1.06e-15
```

ניתן לומר כי ע"פ המדדים שהשווינו אכן המודל החדש מתאים יותר. על מנת לבצע סידור אחרון למודל ולהכין מה הם משתני האינטרקציה המינימליים הדרושים לנו להסברת רווח המוסד האקדמי, הרצנו שוב את אלגוריתמי הרגרסיה.

לבסוף הגענו לערך המדד הטוב ביותר שהצלחנו להוציא: $R_{adj} = 0.6443$, AIC = -444.44, BIC =-439.32.

כעת על מנת לוודא שאכן המודל שלנו עומד בהנחת הנורמליות על השגיאות, ביצענו מבחן Kolmogrov-Smirnov למודל החדש. בנוסף למבחן, יצרנו תרשים היסטוגרמה על מנת לראות מוחשית את התפלגות השגיאות. במבחן אכן ניתן לראות כי ערכו של P_Value גבוה מ-0.05 (0.1541) ולכן ניתן לומר כי עבור מודל זה לא נדחה את השערת H_0 ונאמר כי השגיאות מתפלגות בצורה נורמלית.

```
> ks.test(x=dataset$stan_residuals, y='pnorm', alternative = 'two.sided',
+         exact=NULL) #הנחת נורמליות כן מתקיימת -  p-value = 0.1541

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  dataset$stan_residuals
D = 0.09636, p-value = 0.1541
alternative hypothesis: two-sided
```



לאחר הוספת המשתנים הרציפים, נוספו לנו עוד 2 משתנים רציפים למודל:

X_2 – [log()] של אחוז ההורים בעלי השכלה תיכונית בלבד.

X_3 –[log()] של גיל הכניסה הממוצע במוסד.

לסיכום, המודל הנבחר הינו:

$$
\begin{aligned}
ln(Y) = \ & \beta_0 + \beta_1 X_1 + \beta_2 C_4 + \beta_3 C_5 + \beta_4 A_2 + \beta_5 A_3 + \beta_6 B_2 + \beta_7 B_3 \\
& + \beta_8 X_2 + \beta_9 X_3 + X_1 \beta_{10} B_2 + X_1 \beta_{11} B_3 + X_2 \beta_{12} A_2 + X_2 \beta_{13} A_3 \\
& + X_2 \beta_{14} B_2 + X_2 \beta_{15} B_3 + X_3 \beta_{16} C_4 + X_3 \beta_{17} C_5 + X_3 \beta_{18} A_2 \\
& + X_3 \beta_{19} A_3 + X_3 \beta_{20} B_2 + X_3 \beta_{21} B_3
\end{aligned}
$$

נספחים

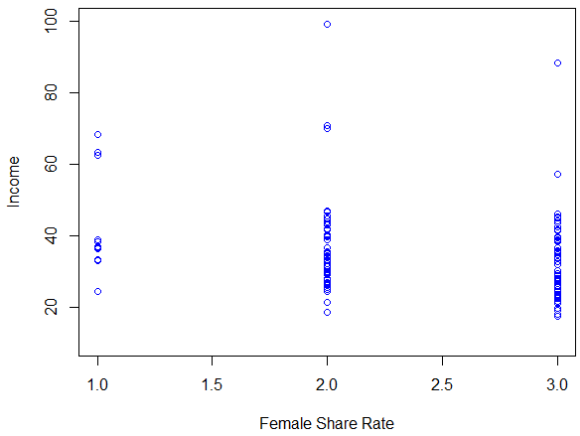
נספח 2.1 – ביצוע מבחן פירסון לבדיקת התאמה

```
> #סעיף א#
> cor.test(dataset$faculty_salary,dataset$income,method = "pearson")

        Pearson's product-moment correlation

data:  dataset$faculty_salary and dataset$income
t = 7.1338, df = 136, p-value = 5.265e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3886102 0.6336764
sample estimates:
      cor
0.5218282

> cor.test(dataset$ft_faculty_rate,dataset$income,method = "pearson")

        Pearson's product-moment correlation

data:  dataset$ft_faculty_rate and dataset$income
t = 1.9041, df = 136, p-value = 0.05901
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.006130298  0.319637502
sample estimates:
      cor
0.1611396

> cor.test(dataset$demographics_age_entry,dataset$income,method = "pearson")

        Pearson's product-moment correlation

data:  dataset$demographics_age_entry and dataset$income
t = -0.8857, df = 136, p-value = 0.3773
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.23980005  0.09254615
sample estimates:
         cor
-0.07572984

> cor.test(dataset$demographics_female_share,dataset$income,method = "pearson")

        Pearson's product-moment correlation

data:  dataset$demographics_female_share and dataset$income
t = -4.3815, df = 136, p-value = 2.337e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.4900108 -0.1961265
sample estimates:
      cor
-0.3517047
```

```
> cor.test(dataset$parents_highschool,dataset$income,method = "pearson")

        Pearson's product-moment correlation

data:  dataset$parents_highschool and dataset$income
t = -1.3884, df = 136, p-value = 0.1673
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.27979616  0.04987137
sample estimates:
        cor
-0.1182187

> cor.test(dataset$highest_degrees_Number,dataset$income,method = "pearson")

        Pearson's product-moment correlation

data:  dataset$highest_degrees_Number and dataset$income
t = 4.4163, df = 136, p-value = 2.031e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1988155 0.4921341
sample estimates:
      cor
0.3541542

> cor.test(dataset$online_binarey,dataset$income,method = "pearson")

        Pearson's product-moment correlation

data:  dataset$online_binarey and dataset$income
t = 0.08319, df = 136, p-value = 0.9338
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1601624  0.1740307
sample estimates:
         cor
0.007133306

> cor.test(dataset$ownership_Number,dataset$income,method = "pearson")

        Pearson's product-moment correlation

data:  dataset$ownership_Number and dataset$income
t = -2.4044, df = 136, p-value = 0.01754
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.35699053 -0.03604331
sample estimates:
      cor
-0.2019319
```

נספח 2.2 – הפיכת משתנה רציף לקטגוריאלי

```
> #####סעיף ב#####
> ####הפיכת משתנה רציף למשתנה קטגוריאלי#######
> dataset$demographics_female_share <- ifelse(dataset$demographics_female_share<0.4,1,
+                                             ifelse(dataset$demographics_female_share>0.6,3,2))
>
> tableWomenRateLow <-  subset(dataset,dataset$demographics_female_share==1)
> summary(tableWomenRateLow$income)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  24.40   33.30   36.80   42.11   44.70   68.40
>
> tableWomenRateMedium <-  subset(dataset,dataset$demographics_female_share==2)
> summary(tableWomenRateMedium$income)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.40   29.10   33.90   35.94   40.20   99.20
>
> tableWomenRateHigh <-  subset(dataset,dataset$demographics_female_share==3)
> summary(tableWomenRateHigh$income)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   17.5    24.3    29.5    32.0    38.8    88.3
>
> plot(dataset$demographics_female_share,dataset$income,ylim = c(10,100)
+      ,xlab = "Female Share Rate",ylab = "Income",col="blue")
```



נספח 2.3 – איחוד קטגוריות

```
> NonDegree <-  subset(dataset,dataset$highest_degrees_Number==1)
> summary(NonDegree $income)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  28.10   38.30   43.80   42.44   44.90   57.10
>
> CertificateDegree <-  subset(dataset,dataset$highest_degrees_Number==2)
> summary(CertificateDegree $income)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   26.1    31.3    36.5    36.5    41.7    46.9
>
> AssociateDegree <-  subset(dataset,dataset$highest_degrees_Number==3)
> summary(AssociateDegree $income)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  17.50   23.95   27.20   27.83   31.50   41.40
>
> BachelorDegree <-  subset(dataset,dataset$highest_degrees_Number==4)
> summary(BachelorDegree $income)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.40   23.60   28.70   29.52   35.10   43.20
>
> GraduateDegree<-  subset(dataset,dataset$highest_degrees_Number==5)
> summary(GraduateDegree $income)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  24.80   33.90   38.90   41.85   44.20   99.20
```

נספח 2.4 – משתני דמה ואינטרקציה

```
Call:
lm(formula = dataset$income ~ dataset$faculty_salary * female_Dum)

Residuals:
    Min      1Q  Median      3Q     Max
-14.381  -6.279  -2.390   3.478  50.561

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    15.004446   7.840140   1.914 0.057811 .
dataset$faculty_salary          0.004759   0.001273   3.737 0.000276 ***
female_Dum2                    -1.145286   8.777403  -0.130 0.896384
female_Dum3                     8.355827   8.668792   0.964 0.336860
dataset$faculty_salary:female_Dum2 -0.001534   0.001383  -1.109 0.269624
dataset$faculty_salary:female_Dum3 -0.002961   0.001465  -2.021 0.045301 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.32 on 132 degrees of freedom
Multiple R-squared:  0.338,     Adjusted R-squared:  0.3129
F-statistic: 13.48 on 5 and 132 DF,  p-value: 1.283e-10
```

```
lm(formula = dataset$income ~ dataset$faculty_salary * highest_degrees_Dum)

Residuals:
    Min      1Q  Median      3Q     Max 
-13.869  -5.837  -2.467   3.477  47.047 

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)                             20.6786060  3.9397645   5.249 5.95e-07 ***
dataset$faculty_salary                   0.0014760  0.0007376   2.001   0.0475 *  
highest_degrees_Dum4                     3.5901039  7.0089283   0.512   0.6094    
highest_degrees_Dum5                     5.5647133  5.2423170   1.061   0.2904    
dataset$faculty_salary:highest_degrees_Dum4 -0.0003063  0.0014029  -0.218   0.8275    
dataset$faculty_salary:highest_degrees_Dum5  0.0008507  0.0008803   0.966   0.3356    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.691 on 132 degrees of freedom
Multiple R-squared:  0.4163,	Adjusted R-squared:  0.3942 
F-statistic: 18.83 on 5 and 132 DF,  p-value: 4.241e-14
```

```
Call:
lm(formula = dataset$income ~ dataset$faculty_salary * ownership_Dum)

Residuals:
    Min      1Q  Median      3Q     Max 
-14.150  -5.910  -1.486   4.175  50.930 

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)                           13.5161933  4.0440004   3.342  0.00108 ** 
dataset$faculty_salary                 0.0038935  0.0005999   6.490 1.59e-09 ***
ownership_Dum2                        -2.1277927  6.4084675  -0.332  0.74039    
ownership_Dum3                        14.8163438  6.0850739   2.435  0.01623 *  
dataset$faculty_salary:ownership_Dum2 -0.0005816  0.0009351  -0.622  0.53506    
dataset$faculty_salary:ownership_Dum3 -0.0030150  0.0012387  -2.434  0.01627 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.17 on 132 degrees of freedom
Multiple R-squared:  0.3566,	Adjusted R-squared:  0.3323 
F-statistic: 14.63 on 5 and 132 DF,  p-value: 2.098e-11
```

נספח 3.1 – בחירת משתני המודל והתאמות

```
> #בחירת משתני המודל#
>
> originalModel <- lm(dataset$income~dataset$faculty_salary*highest_degrees_Dum+dataset$faculty_salary*
+                     female_Dum+dataset$faculty_salary*ownership_Dum) #Original Model
>
> Emp <- lm(dataset$income~1,data=dataset)
> Full <- lm(originalModel)
> summary(originalModel) #Adjusted R-squared = 0.4615

Call:
lm(formula = dataset$income ~ dataset$faculty_salary * highest_degrees_Dum + 
    dataset$faculty_salary * female_Dum + dataset$faculty_salary * 
    ownership_Dum)

Residuals:
    Min      1Q  Median      3Q     Max 
-11.897  -5.801  -1.225   3.411  49.309 

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)                               1.837e+01  8.332e+00   2.205  0.02932 *  
dataset$faculty_salary                    3.491e-03  1.419e-03   2.460  0.01527 *  
highest_degrees_Dum4                      4.887e+00  6.928e+00   0.705  0.48187    
highest_degrees_Dum5                      1.131e+01  5.359e+00   2.111  0.03680 *  
female_Dum2                              -9.069e+00  8.194e+00  -1.107  0.27053    
female_Dum3                              -9.971e+00  9.083e+00  -1.098  0.27443    
ownership_Dum2                            3.908e+00  6.225e+00   0.628  0.53125    
ownership_Dum3                            1.920e+01  7.557e+00   2.540  0.01232 *  
dataset$faculty_salary:highest_degrees_Dum4 -4.002e-04  1.400e-03  -0.286  0.77551    
dataset$faculty_salary:highest_degrees_Dum5 -9.250e-05  9.316e-04  -0.099  0.92107    
dataset$faculty_salary:female_Dum2       -4.442e-04  1.268e-03  -0.350  0.72669    
dataset$faculty_salary:female_Dum3       -7.914e-06  1.552e-03  -0.005  0.99594    
dataset$faculty_salary:ownership_Dum2    -6.756e-04  9.011e-04  -0.750  0.45486    
dataset$faculty_salary:ownership_Dum3    -3.892e-03  1.477e-03  -2.635  0.00947 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.138 on 124 degrees of freedom
Multiple R-squared:  0.5126,	Adjusted R-squared:  0.4615 
F-statistic: 10.03 on 13 and 124 DF,  p-value: 4.092e-14

> extractAIC(originalModel,k=2) #AIC value for the original model = 623.86
[1]  14.0000 623.8653
> extractAIC(originalModel,k=log(10)) #BIC value the original model = 628.1
[1]  14.0000 628.1015
```

נספח 3.2 – הרצת אלגוריתמי רגרסיה לפנים, רגרסיה לאחור ורגרסיה בצעדים

רגרסיה לפנים:

```
> fwd.model <- step(Emp,direction = 'forward',scope = ~dataset$faculty_salary*highest_degrees_Dum+
+                   dataset$faculty_salary*female_Dum+dataset$faculty_salary*ownership_Dum)
Start:  AIC=697.03
dataset$income ~ 1

                        Df Sum of Sq   RSS    AIC
+ highest_degrees_Dum    2    6174.4 15067 653.64
+ dataset$faculty_salary 1    5784.1 15457 655.16
+ female_Dum             2    1245.4 19996 692.69
+ ownership_Dum          2     939.2 20302 694.79
<none>                               21241 697.03

Step:  AIC=653.64
dataset$income ~ highest_degrees_Dum

                        Df Sum of Sq   RSS    AIC
+ dataset$faculty_salary 1   2533.05 12534 630.23
+ female_Dum             2   1236.18 13831 645.82
<none>                               15067 653.64
+ ownership_Dum          2    103.48 14964 656.68

Step:  AIC=630.23
dataset$income ~ highest_degrees_Dum + dataset$faculty_salary

                                         Df Sum of Sq   RSS    AIC
+ female_Dum                              2   1207.51 11326 620.25
<none>                                                12534 630.23
+ ownership_Dum                           2    336.14 12198 630.48
+ dataset$faculty_salary:highest_degrees_Dum  2    136.07 12398 632.73

Step:  AIC=620.25
dataset$income ~ highest_degrees_Dum + dataset$faculty_salary +
    female_Dum

                                         Df Sum of Sq   RSS    AIC
+ dataset$faculty_salary:female_Dum       2    348.19 10978 619.95
<none>                                                11326 620.25
+ dataset$faculty_salary:highest_degrees_Dum  2     81.91 11245 623.25
+ ownership_Dum                           2     75.92 11250 623.33

Step:  AIC=619.95
dataset$income ~ highest_degrees_Dum + dataset$faculty_salary +
    female_Dum + dataset$faculty_salary:female_Dum

                                         Df Sum of Sq   RSS    AIC
<none>                                                10978 619.95
+ dataset$faculty_salary:highest_degrees_Dum  2    19.932 10958 623.70
+ ownership_Dum                           2    19.243 10959 623.70
```

```
> summary(fwd.model) #Adjusted R-squared = 0.4553

Call:
lm(formula = dataset$income ~ highest_degrees_Dum + dataset$faculty_salary +
    female_Dum + dataset$faculty_salary:female_Dum, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-13.382  -5.517  -1.886   4.164  47.867

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                        19.9488844  7.0293833   2.838  0.00527 **
highest_degrees_Dum4                3.2135759  2.4946078   1.288  0.19996
highest_degrees_Dum5               10.9028709  1.8167454   6.001 1.83e-08 ***
dataset$faculty_salary              0.0032056  0.0011627   2.757  0.00667 **
female_Dum2                        -7.8837152  7.9495154  -0.992  0.32318
female_Dum3                         2.0426532  7.8273890   0.261  0.79453
dataset$faculty_salary:female_Dum2 -0.0006584  0.0012448  -0.529  0.59775
dataset$faculty_salary:female_Dum3 -0.0020983  0.0013189  -1.591  0.11404
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.19 on 130 degrees of freedom
Multiple R-squared:  0.4832,    Adjusted R-squared:  0.4553
F-statistic: 17.36 on 7 and 130 DF,  p-value: 4.317e-16

> extractAIC(fwd.model,k=2) #AIC value for the fwd model = 619.9459
[1]   8.0000 619.9459
> extractAIC(fwd.model,k=log(10)) #BIC value the fwd model = 622.3666
[1]   8.0000 622.3666
```

רגרסיה לאחור:

```
> bwd.model <- step(Full,direction = 'backward',k=2,scope=~1)
Start:  AIC=623.87
dataset$income ~ dataset$faculty_salary * highest_degrees_Dum +
    dataset$faculty_salary * female_Dum + dataset$faculty_salary *
    ownership_Dum

                                         Df Sum of Sq   RSS    AIC
- dataset$faculty_salary:highest_degrees_Dum  2      6.96 10361 619.96
- dataset$faculty_salary:female_Dum           2     18.60 10372 620.11
<none>                                                   10354 623.87
- dataset$faculty_salary:ownership_Dum        2    581.56 10936 627.41

Step:  AIC=619.96
dataset$income ~ dataset$faculty_salary + highest_degrees_Dum +
    female_Dum + ownership_Dum + dataset$faculty_salary:female_Dum +
    dataset$faculty_salary:ownership_Dum

                                       Df Sum of Sq   RSS    AIC
- dataset$faculty_salary:female_Dum     2     22.18 10383 616.25
<none>                                              10361 619.96
- dataset$faculty_salary:ownership_Dum  2    598.17 10959 623.70
- highest_degrees_Dum                   2   2408.59 12769 644.80

Step:  AIC=616.25
dataset$income ~ dataset$faculty_salary + highest_degrees_Dum +
    female_Dum + ownership_Dum + dataset$faculty_salary:ownership_Dum

                                       Df Sum of Sq   RSS    AIC
<none>                                              10383 616.25
- dataset$faculty_salary:ownership_Dum  2    867.51 11250 623.33
- female_Dum                            2   1186.69 11570 627.19
- highest_degrees_Dum                   2   2452.63 12836 641.52
```

```
> summary(bwd.model) #Adjusted R-squared = 0.4768 !MAX!

Call:
lm(formula = dataset$income ~ dataset$faculty_salary + highest_degrees_Dum +
    female_Dum + ownership_Dum + dataset$faculty_salary:ownership_Dum)

Residuals:
    Min      1Q  Median      3Q     Max
-12.220  -5.768  -1.356   3.106  48.789

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                        2.079e+01  4.709e+00   4.415 2.13e-05 ***
dataset$faculty_salary             3.062e-03  5.557e-04   5.509 1.90e-07 ***
highest_degrees_Dum4               3.101e+00  2.571e+00   1.206 0.229998
highest_degrees_Dum5               1.092e+01  2.033e+00   5.374 3.52e-07 ***
female_Dum2                       -1.179e+01  3.122e+00  -3.776 0.000243 ***
female_Dum3                       -1.014e+01  2.956e+00  -3.432 0.000807 ***
ownership_Dum2                     3.888e+00  6.020e+00   0.646 0.519451
ownership_Dum3                     1.773e+01  5.472e+00   3.241 0.001519 **
dataset$faculty_salary:ownership_Dum2 -6.731e-04  8.522e-04  -0.790 0.431096
dataset$faculty_salary:ownership_Dum3 -3.638e-03  1.114e-03  -3.265 0.001406 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.007 on 128 degrees of freedom
Multiple R-squared:  0.5112,    Adjusted R-squared:  0.4768
F-statistic: 14.87 on 9 and 128 DF,  p-value: 2.606e-16

> extractAIC(bwd.model,k=2) #AIC value for bwd model = 616.2531 !LOWEST!
[1]  10.0000 616.2531
> extractAIC(bwd.model,k=log(10)) #BIC value for bwd model =  619.279 !LOWEST!
[1]  10.000 619.279
```

רגרסיה בצעדים:

```
> sw.model <- step(Emp,direction = 'both',scope = ~dataset$faculty_salary*highest_degrees_Dum+
+                  dataset$faculty_salary*female_Dum+dataset$faculty_salary*ownership_Dum)
Start:  AIC=697.03
dataset$income ~ 1

                        Df Sum of Sq   RSS    AIC
+ highest_degrees_Dum    2    6174.4 15067 653.64
+ dataset$faculty_salary 1    5784.1 15457 655.16
+ female_Dum             2    1245.4 19996 692.69
+ ownership_Dum          2     939.2 20302 694.79
<none>                               21241 697.03

Step:  AIC=653.64
dataset$income ~ highest_degrees_Dum

                        Df Sum of Sq   RSS    AIC
+ dataset$faculty_salary 1    2533.1 12534 630.23
+ female_Dum             2    1236.2 13831 645.82
<none>                               15067 653.64
+ ownership_Dum          2     103.5 14964 656.68
- highest_degrees_Dum    2    6174.4 21241 697.03

Step:  AIC=630.23
dataset$income ~ highest_degrees_Dum + dataset$faculty_salary

                                         Df Sum of Sq   RSS    AIC
+ female_Dum                              2   1207.51 11326 620.25
<none>                                                12534 630.23
+ ownership_Dum                           2    336.14 12198 630.48
+ dataset$faculty_salary:highest_degrees_Dum  2    136.07 12398 632.73
- dataset$faculty_salary                  1   2533.05 15067 653.64
- highest_degrees_Dum                     2   2923.33 15457 655.16

Step:  AIC=620.25
dataset$income ~ highest_degrees_Dum + dataset$faculty_salary +
    female_Dum

                                         Df Sum of Sq   RSS    AIC
+ dataset$faculty_salary:female_Dum       2     348.2 10978 619.95
<none>                                                11326 620.25
+ dataset$faculty_salary:highest_degrees_Dum  2      81.9 11245 623.25
+ ownership_Dum                           2      75.9 11250 623.33
- female_Dum                              2    1207.5 12534 630.23
- dataset$faculty_salary                  1    2504.4 13831 645.82
- highest_degrees_Dum                     2    3247.5 14574 651.04

Step:  AIC=619.95
dataset$income ~ highest_degrees_Dum + dataset$faculty_salary +
    female_Dum + dataset$faculty_salary:female_Dum

                                         Df Sum of Sq   RSS    AIC
<none>                                                10978 619.95
- dataset$faculty_salary:female_Dum       2    348.19 11326 620.25
+ dataset$faculty_salary:highest_degrees_Dum  2     19.93 10958 623.70
+ ownership_Dum                           2     19.24 10959 623.70
- highest_degrees_Dum                     2   3084.28 14063 650.11
```

```
> summary(sw.model) #Adjusted R-squared = 0.4553

Call:
lm(formula = dataset$income ~ highest_degrees_Dum + dataset$faculty_salary +
    female_Dum + dataset$faculty_salary:female_Dum, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-13.382  -5.517  -1.886   4.164  47.867

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     19.9488844  7.0293833   2.838  0.00527 **
highest_degrees_Dum4             3.2135759  2.4946078   1.288  0.19996
highest_degrees_Dum5            10.9028709  1.8167454   6.001 1.83e-08 ***
dataset$faculty_salary           0.0032056  0.0011627   2.757  0.00667 **
female_Dum2                     -7.8837152  7.9495154  -0.992  0.32318
female_Dum3                      2.0426532  7.8273890   0.261  0.79453
dataset$faculty_salary:female_Dum2 -0.0006584  0.0012448  -0.529  0.59775
dataset$faculty_salary:female_Dum3 -0.0020983  0.0013189  -1.591  0.11404
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.19 on 130 degrees of freedom
Multiple R-squared:  0.4832,    Adjusted R-squared:  0.4553
F-statistic: 17.36 on 7 and 130 DF,  p-value: 4.317e-16

> extractAIC(sw.model,k=2) #AIC value for the fwd model = 619.9459
[1]   8.0000 619.9459
> extractAIC(sw.model,k=log(10)) #BIC value the fwd model = 622.3666
[1]   8.0000 622.3666
```

נספח 4.1 – מבחן פירסון מחודשים

```
> cor.test(dataset$faculty_salary,dataset$incomeLan,method = "pearson")  # X

        Pearson's product-moment correlation

data:  dataset$faculty_salary and dataset$incomeLan
t = 7.4727, df = 136, p-value = 8.622e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4093162 0.6481856
sample estimates:
      cor
0.5395187
```

```
> cor.test(log(dataset$faculty_salary),dataset$incomeLan,method = "pearson") #log(X)

        Pearson's product-moment correlation

data:  log(dataset$faculty_salary) and dataset$incomeLan
t = 5.8142, df = 136, p-value = 4.14e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3015637 0.5707352
sample estimates:
      cor
0.4461841
```

```
> cor.test(dataset$faculty_salary_X2,dataset$incomeLan,method = "pearson") #X^2 --> הטוב ביותר

        Pearson's product-moment correlation

data:  dataset$faculty_salary_X2 and dataset$incomeLan
t = 7.8355, df = 136, p-value = 1.199e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4307363 0.6630147
sample estimates:
     cor
0.557699
```

```
> cor.test(sqrt(dataset$faculty_salary),dataset$incomeLan,method = "pearson") #X^0.5

        Pearson's product-moment correlation

data:  sqrt(dataset$faculty_salary) and dataset$incomeLan
t = 6.7999, df = 136, p-value = 3.023e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3675438 0.6187357
sample estimates:
      cor
0.5037115
```

```
> cor.test(dataset$ft_faculty_rate,dataset$income,method = "pearson") # X

        Pearson's product-moment correlation

data:  dataset$ft_faculty_rate and dataset$income
t = 1.9041, df = 136, p-value = 0.05901
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.006130298  0.319637502
sample estimates:
      cor
0.1611396

> cor.test(log(dataset$ft_faculty_rate),dataset$income,method = "pearson") #log(X)

        Pearson's product-moment correlation

data:  log(dataset$ft_faculty_rate) and dataset$income
t = 1.6604, df = 136, p-value = 0.09914
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.02678171  0.30096971
sample estimates:
      cor
0.1409539

> cor.test(dataset$ft_faculty_rate^2,dataset$income,method = "pearson") #X^2 --> הטוב ביותר מביניהם

        Pearson's product-moment correlation

data:  dataset$ft_faculty_rate^2 and dataset$income
t = 1.8545, df = 136, p-value = 0.06584
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.01032944  0.31586207
sample estimates:
      cor
0.1570465

> cor.test(sqrt(dataset$ft_faculty_rate),dataset$income,method = "pearson") #X^0.5

        Pearson's product-moment correlation

data:  sqrt(dataset$ft_faculty_rate) and dataset$income
t = 1.8407, df = 136, p-value = 0.06785
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.01149895  0.31480872
sample estimates:
      cor
0.1559055
```

```
> cor.test(dataset$demographics_age_entry,dataset$income,method = "pearson") # X

        Pearson's product-moment correlation

data:  dataset$demographics_age_entry and dataset$income
t = -0.8857, df = 136, p-value = 0.3773
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.23980005  0.09254615
sample estimates:
        cor
-0.07572984

> cor.test(log(dataset$demographics_age_entry),dataset$income,method = "pearson") #log(X) --> הטוב ביותר מביניהם

        Pearson's product-moment correlation

data:  log(dataset$demographics_age_entry) and dataset$income
t = -1.2651, df = 136, p-value = 0.208
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.27008565  0.06034407
sample estimates:
       cor
-0.1078483

> cor.test(dataset$demographics_age_entry^2,dataset$income,method = "pearson") #X^2

        Pearson's product-moment correlation

data:  dataset$demographics_age_entry^2 and dataset$income
t = -0.49801, df = 136, p-value = 0.6193
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2082852  0.1253330
sample estimates:
        cor
-0.04266536

> cor.test(sqrt(dataset$demographics_age_entry),dataset$income,method = "pearson") #X^0.5

        Pearson's product-moment correlation

data:  sqrt(dataset$demographics_age_entry) and dataset$income
t = -1.0771, df = 136, p-value = 0.2833
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.25515268  0.07630843
sample estimates:
        cor
-0.09196915
```

מבחן קיום נורמליות

```
> cor.test(dataset$parents_highschool,dataset$income,method = "pearson") # X

        Pearson's product-moment correlation

data:  dataset$parents_highschool and dataset$income
t = -1.3884, df = 136, p-value = 0.1673
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.27979616  0.04987137
sample estimates:
       cor
-0.1182187

> cor.test(log(dataset$parents_highschool),dataset$income,method = "pearson") #log(X) --> הטוב ביותר מביניהם

        Pearson's product-moment correlation

data:  log(dataset$parents_highschool) and dataset$income
t = -1.6781, df = 136, p-value = 0.09563
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.3023367  0.0252789
sample estimates:
       cor
-0.1424276

> cor.test(dataset$parents_highschool^2,dataset$income,method = "pearson") #X^2

        Pearson's product-moment correlation

data:  dataset$parents_highschool^2 and dataset$income
t = -0.88592, df = 136, p-value = 0.3772
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.23981829  0.09252697
sample estimates:
        cor
-0.07574908

> cor.test(sqrt(dataset$parents_highschool),dataset$income,method = "pearson") #X^0.5

        Pearson's product-moment correlation

data:  sqrt(dataset$parents_highschool) and dataset$income
t = -1.566, df = 136, p-value = 0.1197
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2936623  0.0347901
sample estimates:
       cor
-0.1330885
```