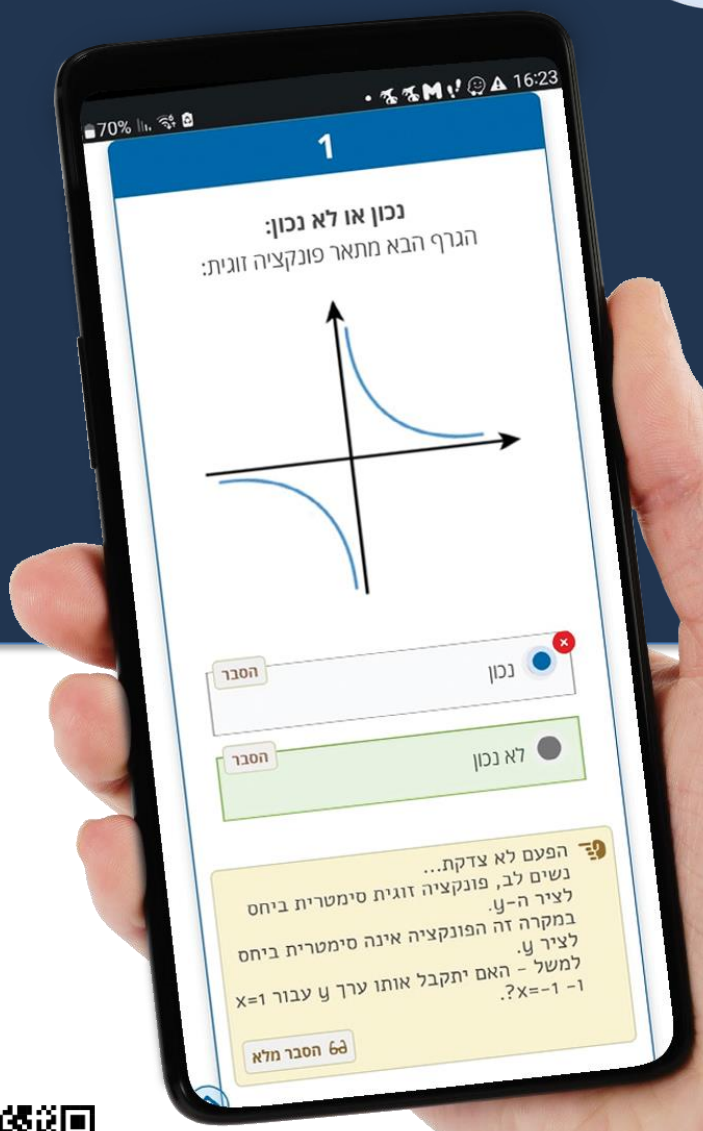
שאלון 471

רגרסיה לינארית

Classit



פתיחת חשבון מורה

וואטסאפ תמיכה

info@classit.co - 073-3744093

מצגת

יחידה ראשונה

קשר לינארי בדיאגרמות פיזור

מבוא: קשרים סטטיסטיים בין משתנים

עד עתה למדנו לחקור משתנים בנפרד, אחד אחד.

בכיתה י' למדנו לחשב למשתנה מסוים את הממוצע, השכיח, החציון, ואת סטית התקן. בכיתה יא' למדנו לחשב את ההסתברויות והשטחים מתחת לעקומת ההתפלגות הנורמלית – שתיארה את האופן שבו משתנה מסוים מתפלג באוכלוסיה (למשל גובה, משכורות, טמפרטורה, וכו')

אבל בתוך העולם הסטטיסטי (וגם בחיים עצמם) יש מקום חשוב לקשרים **בין משתנים**.

למה זה כל כך חשוב?

קשרים סטטיסטיים עוזרים לנו לזהות יחסי גומלין בין משתנים, ומכיוון שקשרים בין משתנים עוזרים לנו לנבא מה צפוי, הם מגדילים את היכולת שלנו לשרוד בעולם. למעשה, המוח שלנו, בנוי כך שהוא ממש "מומחה" בזיהוי קשרים. גם המוח של חיות הוא כזה, אך בדרגת מומחיות נמוכה יותר... והאמת היא שגם כלי AI בנויים על תשתית שמזהה קשרים סטטיסטיים בין משתנים שונים. אבל לא צריך בהכרח קשרים מסובכים כדי לקדם את האנושות, לפעמים מהפכות עולמיות קורות תוך כדי זיהוי של קשרים יומיומיים וטריוויאליים.

במאה ה-18 רופא אנגלי שם לב כי חקלאים שטיפלו בעדרי בקר לא לקו במחלת האבעבועות השחורות – מחלה מדבקת מאוד שהרגה מיליונים רבים של אנשים לאורך מאות שנים - יותר מכל מחלה אחרת.

הרופא חקר את הקשר שגילה, ושם לב שחקלאים שנדבקו מהפרות במחלה בעלת מאפיינים דומים למחלת האבעבועות השחורות - והחלימו ממנה - לא נדבקו בהמשך במחלת האבעבועות הקטלנית.

הקשר שזיהה הרופא היה הבסיס לפיתוח החיסון למחלה שהתבסס על נגיף מוחלש של אבעבועות. בזכות החיסון שגילה הרופא פותח החיסון ההמוני הראשון בהיסטוריית הרפואה, שמיגר כליל את המחלה. מאז ועד היום חיסונים שמבוססים על נגיף מוחלש נפוצים מאוד בעולם הרפואה, והצליחו להדביר מחלות רבות שבעבר היו חשוכות מרפא.

אם כן הקשר הפשוט שגילה הרופא, היה דרמטי בהתפתחות הרפואה והמין האנושי, והציל מליוני אנשים ממוות – כנראה יותר מכל המצאה אנושית אחרת.

לא חייבים להיות דרמטיים כדי להדגים את חשיבות הקשרים בחיינו: למשל, הקשר הפשוט שבין מזג האויר וצבע השמיים – די במבט חטוף מהחלון כדי לזהות שהשמיים אפורים, ולהסיק שקריר בחוץ. או הקשר הפשוט בין מראה פניו של אדם למצב רוחו - די להעיף מבט באדם ולראות שהוא מחייך כדי להסיק שהוא חשב על משהו שעשה לו שמח על הלב...

המשותף לכל הדוגמאות: מידע על משתנה כלשהו (צבע השמיים, חיוך...) עוזר לנו לנבא משתנה אחר (מזג האוויר, מצב רוחו של אדם, וכו').

קשרים לינאריים

בתוך עולם הקשרים הרחב, נתמקד בקשר לינארי.

קשר לינארי הוא קשר בין שני משתנים שאפשר לתאר אותו בעזרת קו ישר. קשרים לינאריים לרוב אפשר לתאר כעליה או ירידה משותפת והדדית של שני המשתנים, למשל: ככל שנהיה חם יותר אנשים מזיעים יותר. או: ככל שיוקר המחיה עולה, כך פוחתת שביעות הרצון של האזרחים.

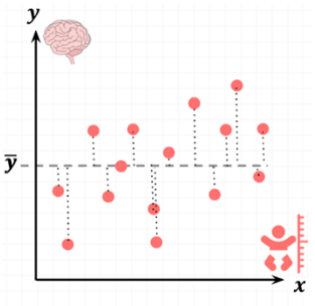
כדי לבדוק אם בין שני משתנים קיים קשר, נאסוף מקבץ של תצפיות (בשפה הסטטיסטית: מדגם) ועבור כל תצפית נמדוד את הערך של משתנה x ושל משתנה y. את הערכים שקיבלנו נציג במערכת צירים, וכך נקבל דיאגרמת פיזור של תצפיות. כדי לבדוק האם קיים קשר לינארי נבדוק האם ניתן לזהות קשר לינארי בפיזור התצפיות בגרף. אך כדי שנוכל לעשות זאת, ראשית נסביר את המושג "דיאגרמת פיזור".

דיאגרמת פיזור

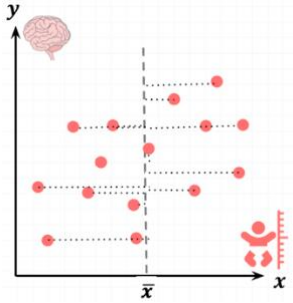
דיאגרמות פיזור הן ייצוג גרפי של קבוצת תצפיות, שלכל אחת מהן ערך X וערך Y. את הערכים שקיבלנו נציג במערכת צירים, כך שערכי ה-x וה-y מתבטאים בשיעורי הנקודה, וכל נקודה היא "תצפית".

למשל, לפנינו דיאגרמת פיזור, שמציגה את המדידות של אחות טיפת חלב ביום מסוים, שבדקה את האורך של התינוקות שהגיעו אליה למרפאה ואת היקף הראש שלהם. כל תינוק הוא נקודה בגרף. ערכי ה-X של הנקודה הם האורך של התינוק, וערכי ה-y של הנקודה הם היקף הראש שלו.

שימו לב, שבדומה למשתנים שלמדנו עליהם בכיתה י', גם כאן אנחנו יכולים לחשב מדדים מוכרים, למשל את הממוצע, את סטית התקן. בדרך כלל לא נידרש לחשב מדדים אלו מתוך הדיאגרמה (אלא רק מתוך נתונים מספריים) אבל בהמשך לימודי הרגרסיה זה שימושי "לקרוא" ולהבין את משמעותם מתוך דיאגרמת הפיזור אז נקדיש רגע כדי להסביר את זה גרפית.



אם נסמן על גבי הגרף את הערך הממוצע של היקף הראש, ונעביר את "קו הממוצע של y" (כלומר את הקו העובר לאורך הדיאגרמה וערכו הוא הממוצע של משתנה y), נוכל לראות את הפיזור סביב הקו.



בדומה לכך, אם נסמן על גבי הגרף את הערך הממוצע של אורכו של התינוק, ונעביר את "קו הממוצע של x" (כלומר את הקו העובר לאורך הדיאגרמה וערכו הוא הממוצע של משתנה x), נוכל לראות את הפיזור סביב הקו.

קשר לינארי בדיאגרמות פיזור

כאמור, קשר לינארי הוא קשר שניתן לבטא כקו ישר.

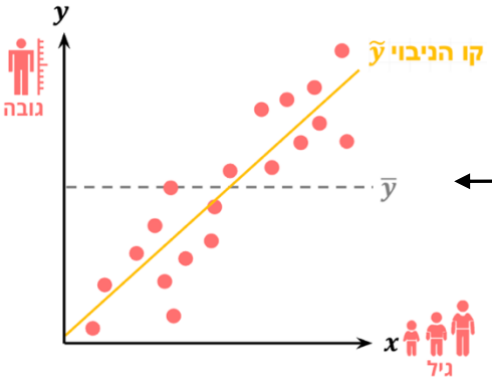
בעולם האמיתי, נדירים המקרים שבהם הקשר הוא מושלם וכל הנקודות מסודרות על קו ישר.

לכן, כדי לזהות אם קיים קשר לינארי או לא, עלינו לבדוק אם לענת התצפיות יש **מגמה של עליה או ירידה**.

כלומר האם קיימת מגמה של שיפוע חיובי או שלילי שאפשר לבטא בעזרת קו ישר.

למשל, בגרף הבא מוצג הקשר בין גילם לגובהם של ילדים, במדגם של ילדים מכיתה א' ועד לכיתה יב'.

ניתן לראות כי ענת התצפיות היא בעלת מגמה של עליה - ככל שגילו של הילד גדול יותר, כך גם הגובה שלו גדול יותר:



ניתן להעביר קו ישר (מסומן בצהוב) שמבטא את המגמה של התצפיות. בהמשך קו זה ישמש אותנו לניבוי ערכי y, ונקרא לו קו הניבוי או **קו הרגרסיה**.

נסמן את הקו ב- $\tilde{y}$.

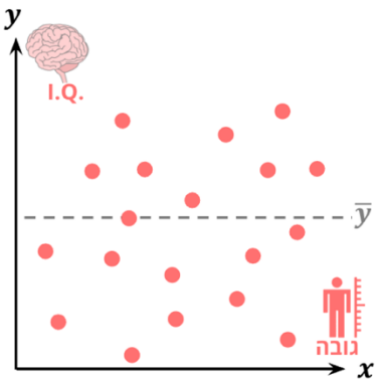
שימו לב: בדיאגרמות פיזור נראה לא פעם את הממוצע מסומן בגרף.

אין צורך לדעת לחשב אותו בעצמנו מתוך התצפיות, אך חשוב להבין את משמעותו בדיאגרמה, זה יהיה לנו שימושי בהמשך הדרך.

כאשר אין קשר לינארי, לתצפיות לא תהיה מגמה כלל - לא של עליה ולא של ירידה.

למשל, נרצה לבחון את הקשר בין גובהו של ילד לגובה האיי.קיו שלו. ברור שאין קשר בין שני הדברים הללו...

בגרף ניתן לראות שאין לענת התצפיות מגמה כלל - לא של עליה ולא של ירידה. התצפיות מפוזרות בצורה אקראית במערכת הצירים:



חשוב: כאשר אין קשר לינארי, לא נוכל להיעזר בערכי משתנה x כדי לנבא את ערכי משתנה y.

במקרה כזה, שבו אין קשר בין שני המשתנים, נחזור לנבא בעזרת מדדי המרכז – כמדדים שמשקפים את "מרכז הכובד" של התצפיות.

הממוצע הוא מדד המרכז הכי שימושי במצב כזה, וישמש מעין baseline, או "ברירת מחדל". בעזרתו ננבא את ערכי y.

נסכם: הניבוי הכי טוב שלנו כשאין לנו ביד שום מידע נוסף, הוא לנבא בעזרת הממוצע של משתנה y.

כיצד נבטא את המגמה של הקשר בעזרת קו?

אם קיימת מגמה לינארית בדיאגרמת הפיזור נרצה לבטא מגמה זו על ידי קו המגמה. קו זה ייקרא בהמשך קו הרגרסיה, או קו הניבוי, והוא ישמש אותנו... ובכן, לניבוי.

- כאשר הקשר הוא חלקי (כלומר קשר לא מושלם), הקו יעבור במרכז ענת התצפיות ויבטא את כיוון הקשר.
- כאשר אין כלל קשר בין המשתנים, ננבא את הממוצע, ולכן במקרה כזה קו הרגרסיה יתלכד עם קו הממוצע, ויהיה זהה לו.
- וכאשר הקשר הלינארי מושלם – הקו יעבור בדיוק על הנקודות של דיאגרמת הפיזור.

בהמשך נלמד לחשב את קו הרגרסיה, ובינתיים נסמן אותו ביד חופשית.

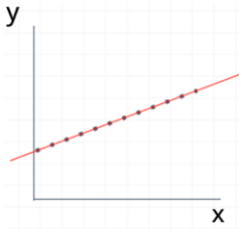
המדד לקשר לינארי – r

המדד המוכר לקשר לינארי הוא מדד בשם מקדם המתאם הלינארי, או מקדם המתאם של פירסון (על שם החוקר פירסון שחישב אותו לראשונה). המדד מסומן באות r. מקדם המתאם יקבל ערכים חיוביים כאשר הקשרים חיוביים, וערכים שליליים כאשר הקשרים שליליים. ערכו של מקדם המתאם נע בין 1 לבין 1-.

עוצמת הקשר הלינארי

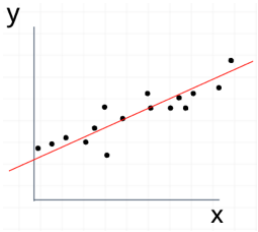
בקשר לינארי לרוב נהוג להתייחס לא רק למקדם המתאם הלינארי עצמו, אלא גם לעוצמת הקשר הלינארי. עוצמת הקשר תסומן $|r|$ והיא מייצגת את חוזק הקשר, ללא תלות בכיוונו של הקשר, שיכול להיות חיובי או שלילי. עוצמת הקשר הלינארי נעה בטווח: $0 \leq |r| \leq 1$.

$r = 1$



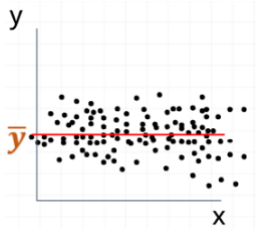
כאשר ערכו של מקדם המתאם הוא 1 קיים **קשר מושלם חיובי**.
כל הנקודות נמצאות על קו ישר עולה, שהוא גם קו הרגרסיה:
במצב זה גם עוצמת המתאם גם היא $|r| = 1$.

$0 \leq r \leq 1$



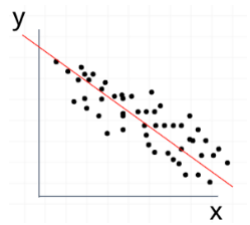
כאשר ערכו של מקדם המתאם נמצא בין 0 ל-1 קיים **קשר חיובי לא מושלם**.
הנקודות יוצרות מגמה של שיפוע עולה (חיובי) וקו הרגרסיה מנבא את המגמה של הנקודות:
עוצמת הקשר הלינארי תהיה: $0 \leq |r| \leq 1$.

$r = 0$



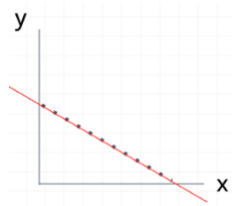
כאשר ערכו של מקדם 0 **אין קשר לינארי בין התצפיות**.
לנקודות אין מגמה של עליה או של ירידה, וקו הניבוי יהיה קו הממוצע של y:
במצב זה גם עוצמת הקשר הלינארי תהיה $|r| = 0$.

$-1 \leq r \leq 0$



כאשר ערכו של מקדם המתאם נמצא בין 0 ל-1- קיים **קשר שלילי לא מושלם**.
הנקודות יוצרות מגמה של שיפוע יורד (שלילי) וקו הרגרסיה מנבא את המגמה של הנקודות:
עוצמת הקשר הלינארי תהיה: $0 \leq |r| \leq 1$.

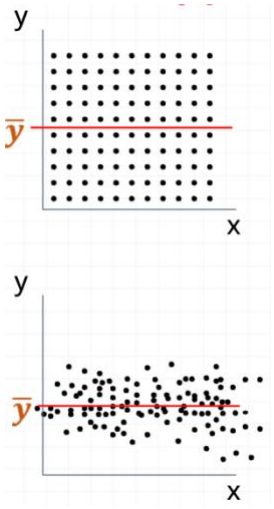
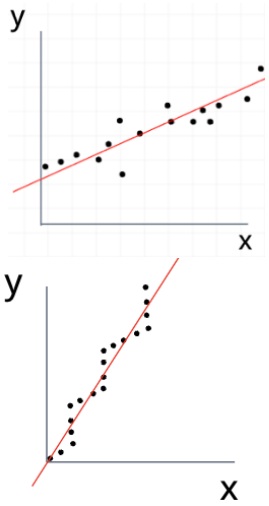
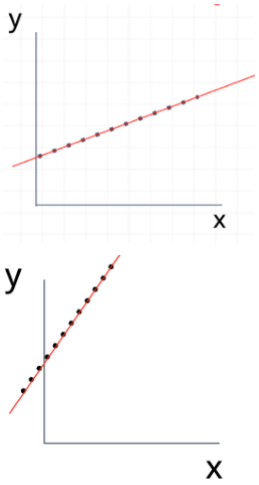
$r = -1$



כאשר ערכו של מקדם המתאם הוא 1- קיים **קשר מושלם שלילי**.
כל הנקודות נמצאות על קו ישר יורד, קו הרגרסיה עובר בדיוק על הנקודות:
במצב זה גם עוצמת המתאם היא $|r| = 1$

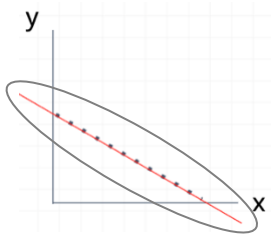
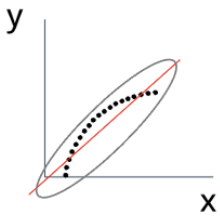
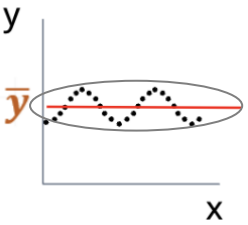
סיכום

ניתן לסכם ולומר שחוזקו של הקשר הלינארי בדיאגרמת פיזור מתבטא במידה שבה הנקודות קרובות לקו ישר (או רחוקות ממנו). ככל שפיזור הנקודות הולך ומתקרב לצורתו של ישר עולה או קו ישר יורד- גדלה עוצמת הקשר. מאידך, ככל שהנקודות הולכות ומתרחקות מקו עולה או קו יורד, עוצמת הקשר נחלשת עוד ועוד, עד למקרה שבו אין כלל קשר לינארי, ומקדם המתאם הוא 0. נראה דוגמאות נוספות לשלושת המצבים הכלליים של קשר לינארי:

אין קשר לינארי	קשר לינארי לא מושלם	קשר לינארי מושלם
		

ניתן לראות שקיימות דיאגרמות פיזור שונות, וחשוב להתרגל למגוון גרפים, ולזהות בהם את אופי הקשר.

בתחילת הדרך עוזר מאד לשרטט "ביצה" מסביב לתצפיות (כפי שסומן ב-3 הדיאגרמות התחתונות), ולראות אם מתקבלת "ביצה עולה" או "ביצה יורדת", וכך לקבוע את כיוון הקשר. אם הביצה "שוכבת" אין קשר....

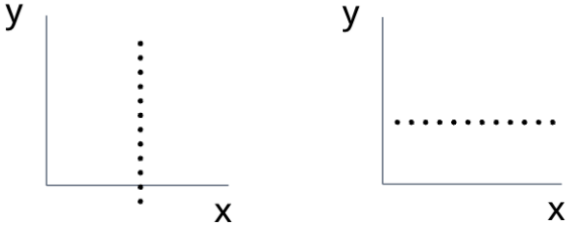
מתי r לא מוגדר

אם בגרף מופיע משתנה אחד ולא שניים - r לא מוגדר.

בגרפים שלפניכם יש משתנה אחד, והמשתנה השני למעשה איננו כלל משתנה, אלא קבוע.

בגרף הימני x משתנה ו-y קבוע. בגרף השמאלי x קבוע, ו-y משתנה.

במקרים אלו r אינו מוגדר, וגם קו הרגרסיה אינו מוגדר:



מספר הערות חשובות:

1. קשר לינארי ניתן לחישוב על משתנים כמותיים בלבד, אבל לא על משתנים איכותיים.
   המשתנים הכמותיים יכולים להיות גם בדידים וגם רציפים.
2. אם משתנה מסוים משפיע על משתנה שני אז בהכרח יהיה ביניהם קשר סטטיסטי. אבל אם קיים קשר סטטיסטי בין שני משתנים זה לא אומר שאחד בהכרח משפיע על השני.

   במילים אחרות: קשר סטטיסטי ≠ סיבתיות.

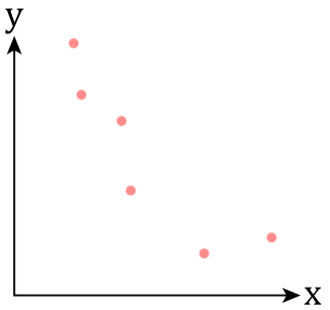
   במקרה של הרופא והחיסון הקשר שנמצא היה סיבתי. העובדה שהחקלאים נדבקו מהפרות במחלה דומה לאבעבועות שחורות, היא שחיסנה אותן בהמשך בפני מחלת האבעבועות השחורות. אבל זה שיש קשר בין צבע העננים למצג האוויר, לא אומר שצבע העננים הוא זה שגורם למזג האוויר. וגם לא שמזג האוויר הוא זה שיוצר את צבע העננים. במילים אחרות קיומו של קשר סטטיסטי לא מעיד בהכרח על סיבתיות.
3. עד עכשיו, גם אם עסקנו בתצוגות גרפיות, עסקנו במשתנה אחד ובשכיחות שלו. אבל כדי לחשב קשר, הכרחי שיהיו בידינו 2 משתנים לפחות. לכן כאשר יש משתנה אחד בלבד, הקשר איננו מוגדר.

קשר לינארי בדיאגרמות פיזור

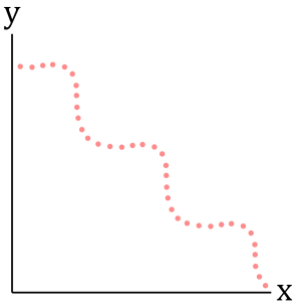
1. עבור כל אחת מדיאגרמות הפיזור הבאות, קבעו מה יתאר אותה באופן הטוב ביותר:

r לא מוגדר	$r=-1$	$-1<r<0$	$r=0$	$0<r<1$	$r=1$

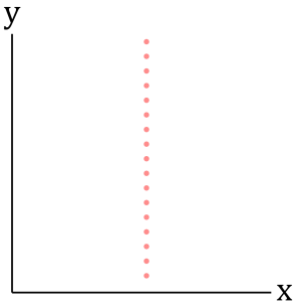
א



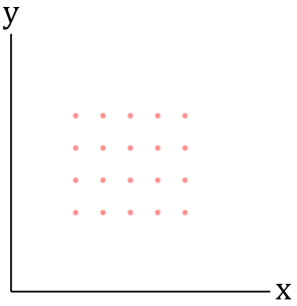
ב



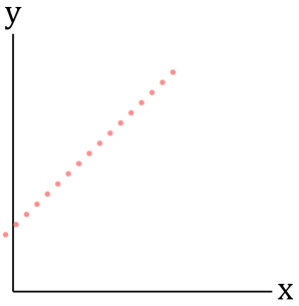
ג



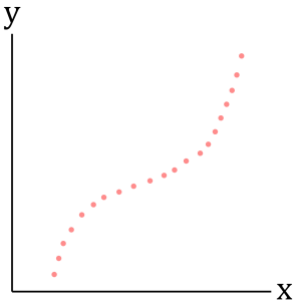
ד



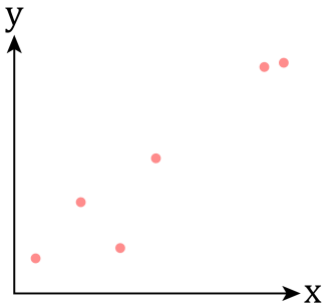
ה



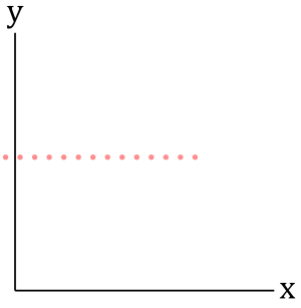
ו



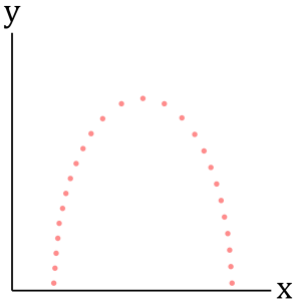
ז



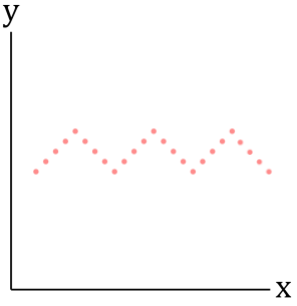
ח



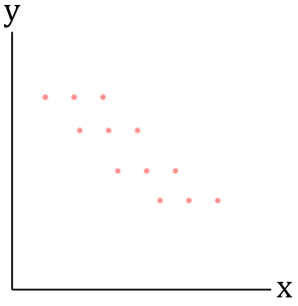
ט



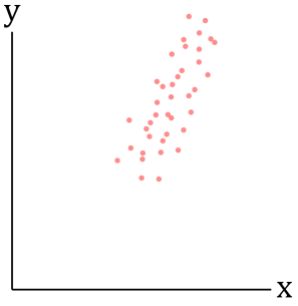
י



יא



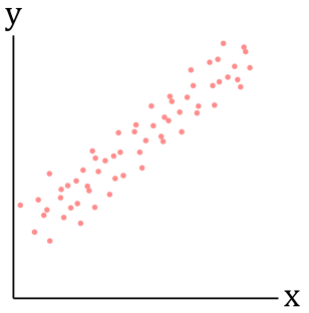
יב



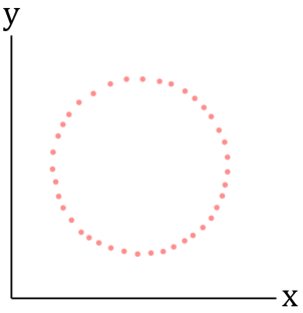
Classit
info@classit.co - 073-3744093

2. לפניך דיאגרמות פיזור. יש לסמן ☑ על כל דיאגרמה בה $r \neq 0$.

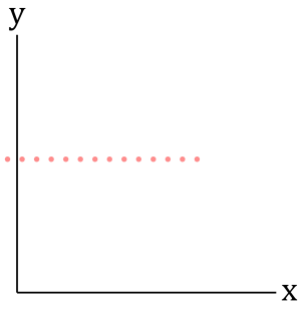
א



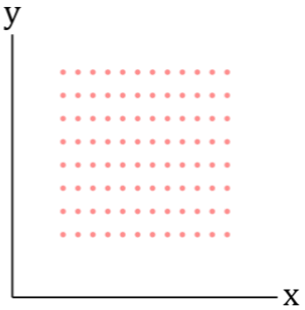
ב



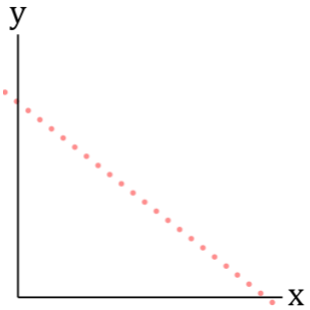
ג



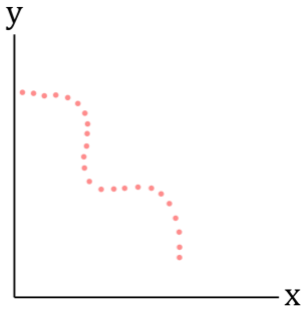
ד



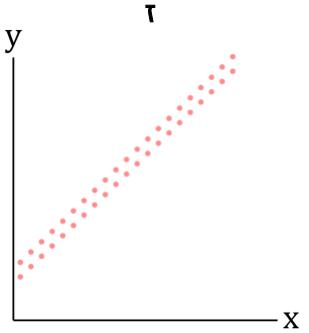
ה



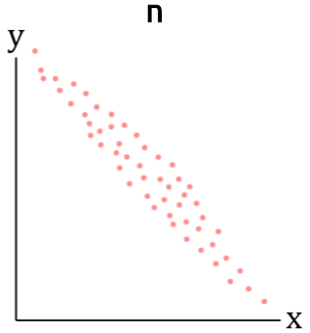
ו



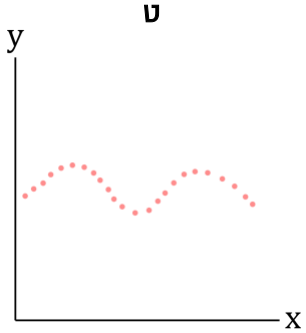
ז



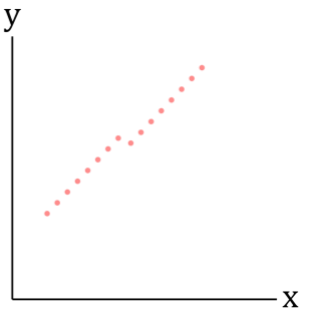
ח



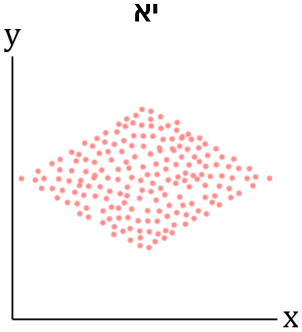
ט



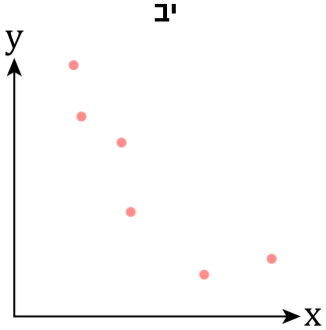
י



יא



יב

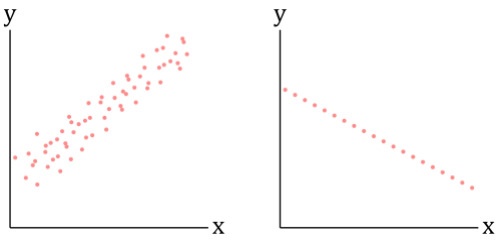


3. לפניכם זוגות של דיאגרמות שונות זו מזו. בכל זוג עליכם להשוות בין הדיאגרמות ולקבוע:

א. סמנו בעזרת $<$ $>$ $=$, באיזו משתי הדיאגרמות מקדם המתאם הלינארי r גדול יותר?

ב. סמנו בעזרת $<$ $>$ $=$, באיזו משתי הדיאגרמות עוצמת הקשר $|r|$ גדולה יותר?

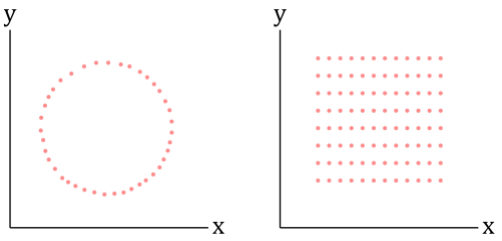
①



r_1 ☐ r_2

$|r_1|$ ☐ $|r_2|$

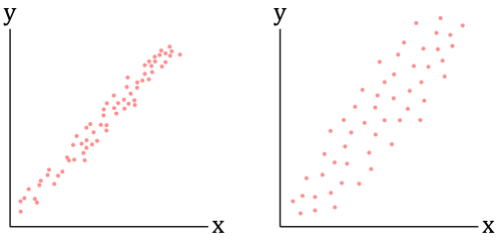
②



r_1 ☐ r_2

$|r_1|$ ☐ $|r_2|$

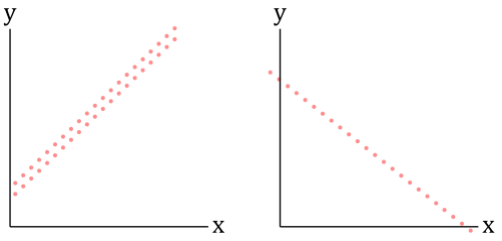
③



r_1 ☐ r_2

$|r_1|$ ☐ $|r_2|$

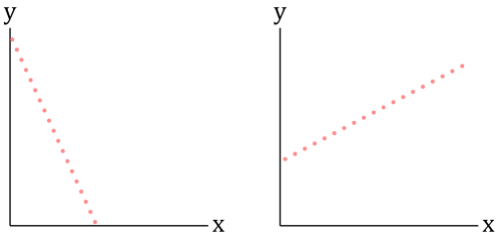
④



r_1 ☐ r_2

$|r_1|$ ☐ $|r_2|$

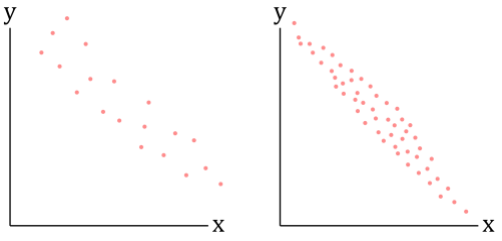
⑤



r_1 ☐ r_2

$|r_1|$ ☐ $|r_2|$

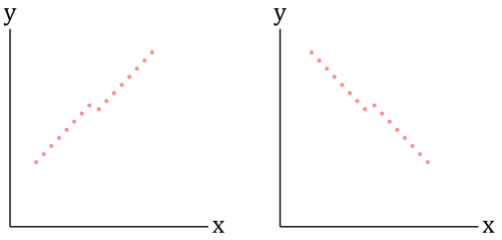
⑥



r_1 ☐ r_2

$|r_1|$ ☐ $|r_2|$

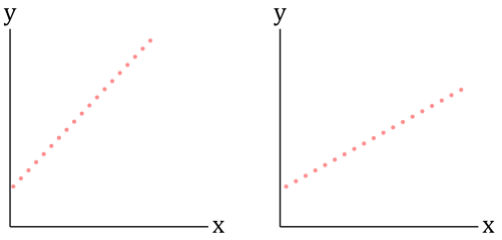
⑦



r_1 ☐ r_2

$|r_1|$ ☐ $|r_2|$

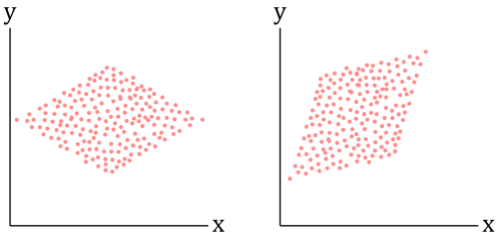
⑧



r_1 ☐ r_2

$|r_1|$ ☐ $|r_2|$

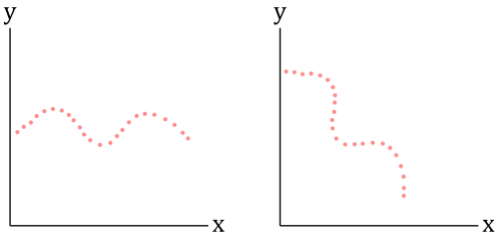
⑨



r_1 ☐ r_2

$|r_1|$ ☐ $|r_2|$

⑩



r_1 ☐ r_2

$|r_1|$ ☐ $|r_2|$

יחידה שניה

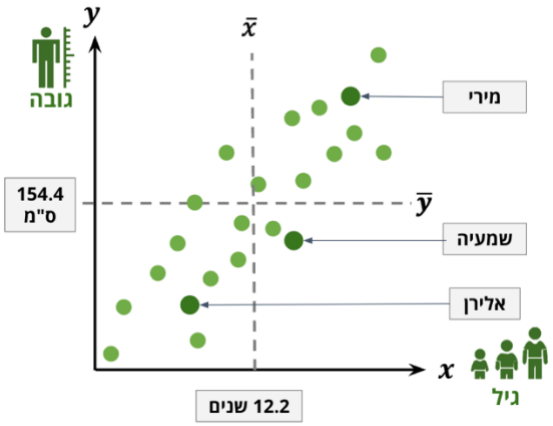
חישוב מקדם המתאם - r

בפרק הקודם למדנו לזהות קשר לינארי בדיאגרמות פיזור. כעת נרצה לחשב את ערכו של r: מקדם המתאם הלינארי. מי שיקפוץ ישר לסוף הפרק יראה נוסחה די ארוכה וקצת מפחידה לחישוב מקדם המתאם.
אבל... האמת היא שהעיקרון של חישוב מקדם המתאם הוא לא כל-כך מסובך. כדי להבין את נוסחת מקדם המתאם ננסה להגדיר מחדש מהו קשר לינארי, הפעם נסתכל על התצפיות אחת אחת.

קשר לינארי: ערכי התצפיות ביחס לממוצעי המשתנים

בפרק של דיאגרמת פיזור למדנו לזהות קשר לינארי בעננת התצפיות כולה, והסקנו שעוצמת הקשר גדלה ככל שפיזור התצפיות הולך ומתקרב לקו ישר. אבל עשינו את זה בעזרת התבוננות גרפית. זיהינו שמבחינה ויזואלית עננת התצפיות "עולה" או "יורדת" כלומר לעננה יש שיפוע חיובי או שלילי. כעת אנחנו מחפשים חישוב מדויק יותר, שבסופו יביא אותנו לנוסחת r.

כדי להגדיר מחדש את הקשר הלינארי נחזור לדוגמה שהתחלנו איתה – הקשר בין גיל הילד לגובהו, בקרב ילדים מכיתה א' עד יב'. הפעם הוספנו סימון על גבי הגרף של ממוצע גיל הילדים, 12.2 שנים, וכן סומן הממוצע של משתנה הגובה, 154.4 ס"מ. ניתן לראות שבגרף קיימת מגמה של שיפוע חיובי, שמשקף קשר חיובי - כלומר ככל שערכי הגיל גדלים, גם ערכי הגובה גדלים. כעת נסתכל על המגמה שיוצרות הנקודות לא מהזווית הגרפית, אלא מהזווית האלגברית, או המספרית שלה.



למשל ניקח את הילדה **מירי**, שגילה 16.8 שנים וגובהה 164 ס"מ. היא יותר גדולה מהגיל הממוצע, והיא גם יותר גבוהה מהגובה הממוצע.
וכן ניקח את **אלירן**, שהוא ילד בן 8 וגובהו 148 ס"מ. כלומר הוא נמוך יותר מהממוצע, בגיל וגם בגובה.
גם מירי וגם אלירן משקפים את הקשר החיובי בין המשתנים - ככל שהילד גדול יותר בגיל, כך הוא גם גבוה יותר בגובה, וככל שהוא קטן יותר בגיל, כך הוא גם נמוך יותר בגובה.

אבל... מה לגבי **שמעיה**?
שמעיה בן 13, וגובהו 150 ס"מ. כלומר הוא גדול יותר מהממוצע בגיל, אך נמוך יותר מהממוצע בגובה.
שמעיה משקף את העובדה שלמרות שברוב המקרים ילדים שגדולים יותר בגיל הם גם גבוהים יותר בגובה, זה לא קורה בכל המקרים. יש ילדים גדולים בגיל שבכל זאת נמוכים מהממוצע בגובה, ולהיפך – יש ילדים קטנים בגיל, שבכל זאת גבוהים מהממוצע בגובה.

כעת נרצה לייצר ביטוי אלגברי עבור הטענה הבאה:

בקשר החיובי בין הגיל לגובה – רוב הזמן ילדים שגדולים מהממוצע בגיל, גם גבוהים מהממוצע בגובה.

כלומר, ננסה לנסח ביטוי אלגברי שיבטא את ההתרחקות המתואמת של הילדים מהממוצע בשני המשתנים, וישקף את העובדה שמירי ואלירן מתואמים בשני המשתנים, ואילו שמעיה לא. לשם כך נבטא אלגברית את המרחק של כל אחד מהילדים מממוצע הגיל, וממוצע הגובה:

א. נביע את המרחק של כל אחד מהילדים מממוצע הגיל $(x_i - \bar{x})$:

עבור מירי (4.6 שנים), עבור אלירן (4.2- שנים), עבור שמעיה (0.8 שנים).

ב. נביע את ההתרחקות של כל אחד מהילדים מממוצע הגובה $(y_i - \bar{y})$:

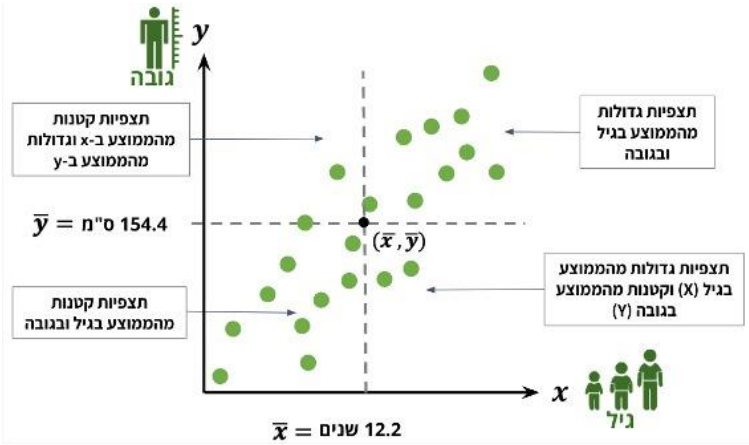
עבור מירי (9.6 ס"מ), עבור אלירן (6.4- ס"מ), עבור שמעיה (4.4- ס"מ).

ג. כעת כדי לבדוק אם ההתרחקות מהממוצעים מתואמת בשני המשתנים, נכפול עבור כל ילד את המרחק שלו מממוצע הגיל במרחק שלו ממוצע הגובה, $(x_i - \bar{x})(y_i - \bar{y})$ ונראה מה יקרה:

עבור מירי 44.16 =(9.6)*(4.6) **עבור אלירן** 26.88 =(-6.4)*(-4.2) **עבור שמעיה** -3.52 =(-4.4)*(0.8)

בעצם קיבלנו ביטוי שנותן למירי ולאלירן ערך חיובי כי הם "הולכים" עם הכיוון החיובי של הקשר, ולשמעיה ערך שלילי, כי הוא הולך "נגד" כיוון הקשר.

בגרף שלפנינו נוכל לראות את העיקרון המנחה, לפיו קוי הממוצעים מחלקים את הגרף ל-4 רביעים, על פי כיוון ההתרחקות מהממוצע. מכפלת המרחקים תתאים לכל נקודה בגרף ערך חיובי או שלילי באופן הבא:



1. אם הילד גבוה בגיל וגם בגובה, הביטוי יקבל ערך חיובי.
2. אם הילד נמוך בגיל וגם בגובה, הביטוי יקבל ערך חיובי.
3. אם הילד גבוה בגיל, אבל נמוך בגובה, הביטוי יקבל ערך שלילי.
4. אם הילד נמוך בגיל, אבל גבוה בגובה, הביטוי יקבל ערך שלילי.

יצרנו ביטוי שמסייע לנו לקבוע את כיוונו של הקשר. אומנם עוד לא הגענו לנוסחת מקדם המתאם, אבל בנינו את החלק המרכזי שלה, שמתיחס למידת התיאום בין המשתנים.

בהמשך הדרך הביטוי שבנינו יהיה חלק **מהמונה של נוסחת r.**

במונה הנוסחה נחשב לכל תצפית את הערך המתאים לה, ונסכום את הערכים האלו עבור כל התצפיות במדגם:

$$(x_{\text{מירי}} - \bar{x})(y_{\text{מירי}} - \bar{y}) + (x_{\text{אלירן}} - \bar{x})(y_{\text{אלירן}} - \bar{y}) + (x_{\text{שמעיה}} - \bar{x})(y_{\text{שמעיה}} - \bar{y}) + \cdots + (x_{\text{ילד אחרון}} - \bar{x})(y_{\text{ילד אחרון}} - \bar{y})$$

הסימון המקובל כמובן לא כולל את שמות הילדים, ויראה כך:

$$(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + (x_3 - \bar{x})(y_3 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})$$

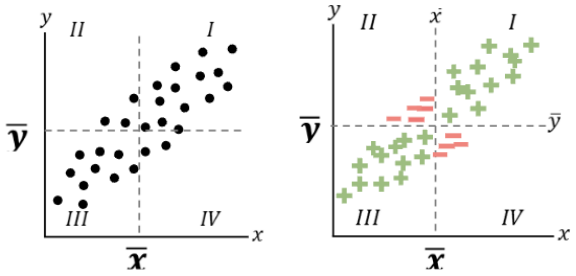
כעת, נכליל את הביטוי שיצרנו לקשרים אחרים, ונראה שבכולם נוצרת התאמה בין המגמה של הקשר בדיאגרמת הפיזור, לבין סכום הערכים שמתקבל בביטוי עבור התצפיות במדגם:

- בקשרים חיוביים, רוב הערכים שיתקבלו בנוסחה יהיו חיוביים, ואם נסכום אותם יתקבל ערך חיובי.
- בקשרים שליליים, רוב הערכים שיתקבלו בנוסחה יהיו שליליים, ואם נסכום אותם יתקבל ערך שלילי.
- כאשר אין קשר, יתקבלו ערכים אקראיים בביטוי, חיוביים ושליליים, שיקזזו זה את זה.

קשר חיובי

אם נסמן כל אחת מהתצפיות בערך חיובי או שלילי בהתאם למיקומה ביחס לקוי הממוצעים, נראה שצפויות להתקבל הרבה תצפיות בעלות ערך חיובי, ומעט תצפיות בעלות ערך שלילי.

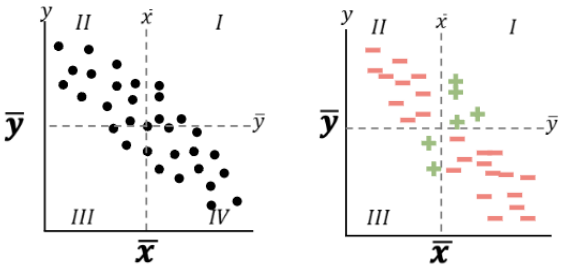
אם נחבר את כל הערכים של התצפיות נקבל בנוסחה שיצרנו **ערך חיובי.**



קשר שלילי

אם נסמן כל אחת מהתצפיות בערך חיובי או שלילי בהתאם למיקומה ביחס לקוי הממוצעים, נראה שצפויות להתקבל הרבה תצפיות בעלות ערך שלילי, ומעט תצפיות בעלות ערך חיובי.

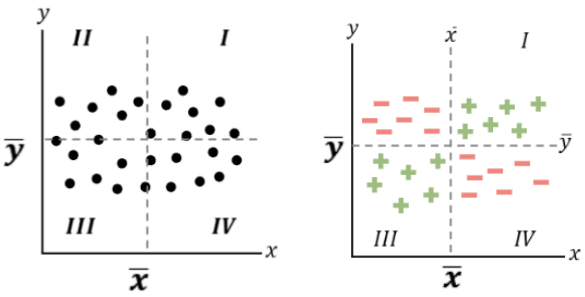
אם נחבר את כל הערכים של התצפיות נקבל בנוסחה שיצרנו **ערך שלילי.**



כשאין קשר

אם נסמן כל אחת מהתצפיות בהתאם למיקומה ביחס לקוי הממוצעים, נראה שצפויות להתקבל גם תצפיות בעלות ערך חיובי, וגם תצפיות בעלות ערך שלילי.

אם נחבר את כל הערכים של התצפיות נקבל בנוסחה שיצרנו ערך **ששואף ל-0**, בדיוק כמו הקשר.



אם כן, הביטוי שיצרנו נותן לנקודות ערך חיובי או שלילי בהתאם למיקומן ולמרחקן מהממוצע בשני המשתנים. (ערכה של הנקודה בביטוי יכול להיות גם אפס אם היא שווה לממוצע באחד המשתנים).

אם נסכום את כל הערכים שהתקבלו בכל אחת מהתצפיות - נקבל ערך שהסימן שלו זהה לסימן של מקדם המתאם r.

האם כבר מצאנו את הנוסחה לחישוב r?

כמעט.... בעמוד הבא נגיע לנוסחה עצמה.

נוסחת מקדם המתאם – r

<u>הערה:</u> כדי להגיע לנוסחת r עצמה עלינו לבצע כמה התאמות על הביטוי שיצרנו. ההסבר על המשך בנית הנוסחה עושה חזרה טובה על מושגים קודמים בסטטיסטיקה, אך אין צורך לשלוט בהבנה שלו לצורך הבגרות ב-4 יח"ל. ההסבר המובא בעמוד זה מיועד בעיקר למורים, תלמידים מוזמנים לקפוץ מיד לנוסחה הסופית שמובאת בסוף העמוד (-:

חלוקה ב-n

כמו כל דבר בסטטיסטיקה ובעולם, עלינו לחלק את הביטוי ב-n, כלומר במספר התצפיות במדגם.

אנחנו עושים את זה כדי לחשב ערך ממוצע, כלומר ערך ממוצע שמבוסס על חישוב כלל התצפיות – שישקף את ההתרחקות הצפויה מהממוצע במשתנה x ובמשתנה y, עבור תצפית בודדת.

$$\frac{(x_1-\bar{x})(y_1-\bar{y})+(x_2-\bar{x})(y_2-\bar{y})\ldots+(x_n-\bar{x})(y_n-\bar{y})}{n}$$

חלוקה בסטיות התקן

לאחר שחילקנו במספר התצפיות נשאר לנו לטפל בבעיה נוספת. הביטוי שיצרנו מאוד רגיש ליחידות המידה של המשתנים.

למשל בדוגמה של הקשר בין הגיל לגובה, האם עוצמת הקשר בין גיל לגובה היתה משתנה אם במקום למדוד את הגיל בשנים היינו מודדים אותו בימים?

ברור שלא. אלו אותן תצפיות, אותם ילדים, אותו גיל, ואותו גובה.

אבל מצד שני מה היה קורה לביטוי שחישבנו קודם?

נבטא את הגיל של הילדים בימים במקום בשנים, ונראה מה קורה לביטוי שחישבנו:

הממוצע של הגיל של הילדים יהיה (12.2 שנים) * (356 יום בשנה) = 4343.2 ימים.

הגיל של מירי יהיה (16.8 שנים) * (356 ימים בשנה) = 5980.8 ימים.

הערך שהיה מתקבל עבור מירי אם היינו כופלים את המרחק שלה מהגיל הממוצע במרחק שלה מהגובה הממוצע: (9.6 ס"מ) * (1637.6 ימים) = **15,720.96** !! הערך המקורי שהתקבל עבור מירי היה 44.16!

הפער בין שני החישובים הוא אדיר.

חשוב להדגיש שוב כי עוצמת הקשר בין הגיל והגובה לא השתנתה, למרות השינוי שחל ביחידות המידה של הגיל. נותר רק לדמיין לעצמנו מה היה קורה אם במקום למדוד את הגיל בשנים, או בימים, היינו מודדים אותו בדקות...

אם כך, עלינו "לכווץ" את יחידות המידה לסקאלה אחידה שלא תושפע מכפל או חילוק של יחידות המידה.

זה מזכיר בעיה שכבר התמודדנו איתה בעבר בלימודי הסטטיסטיקה...

זוכרים מה עשינו כשרצינו כדי לדעת את תצפית רחוקה הרבה או מעט מהממוצע? חריגה או לא? קיצונית או לא?

חישבנו ציוני תקן וקיבלנו את המרחק מהממוצע מבוטא ביחידות של סטיות תקן, ללא תלות ביחידות המידה של המשתנה. זה בדיוק מה שנעשה גם כאן: נחלק את המרחקים שחישבנו מהממוצעים בסטיות התקן של שני המשתנים, וכך "ניפטר" מהרגישות של הנוסחה ליחידות המידה, ונקבל את הנוסחה המלאה של מקדם המתאם:

$$\boldsymbol{r=\frac{(x_1-\bar{x})(y_1-\bar{y})+(x_2-\bar{x})(y_2-\bar{y})\ldots+(x_n-\bar{x})(y_n-\bar{y})}{n\cdot S_x\cdot S_y}}$$

נוסחת מקדם המתאם מקיימת את הכלל שהצגנו בפרק הקודם: ערכו של r יקיים $-1\le r\le 1$.

בתרגול לקראת הבגרות חשוב להציג את הנוסחה כפי שמופיעה בדף הנוסחאות, כדי להתרגל אליה ולעבוד איתה לקראת הבגרות:

$$r = \frac{1}{n \cdot S_x \cdot S_y}[(x_1 - \bar{x})(y_1 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})]$$

לבסוף, מה נחשב קשר חזק?

במדעי החברה נהוג לחלק את עוצמת הקשר ל-3 רמות, עוצמת קשר גבוהה (בין 0.7 ל-1) עוצמה בינונית (בין 0.4 ל-0.7) ועוצמה נמוכה (בין 0 ל-0.4). חשוב לשים לב שמדובר במוסכמה בלבד, ובתחומי דעת שונים נהוגות עוצמות קשר שונות שיוגדרו "חזקות" או "חלשות".

הערות למורה על נוסחת r

- על מנת להשתמש בנוסחת r צריך לחשב את הממוצע של כל אחד מהמשתנים, את סטית התקן, ואז את כל המרחקים של התצפיות מהממוצעים שלהן. כיון שהחישוב ארוך, לרוב נתבקש לחשב את הנוסחה על מספר בודד של תצפיות.
- ברוב חומרי הלימוד בסטטיסטיקה הנוסחאות מופיעות בצורה המקוצרת שלהן עם הסימן סיגמה : Σ סיגמה הוא סימן מאוד נפוץ בסטטיסטיקה שחוסך לנו את הפירוט של: תצפית 1, תצפית 2, תצפית 3, וכו'.

  הסימן סיגמה אומר: "כל מה שבא אחריי יש לחשב לכל אחת מהתצפיות, ואז לסכום"

  ואכן, ניתן לראות שהנוסחה שהתקבלה קצרה ואלגנטית יותר:

$$\frac{\sum(x_1 - \bar{x})(y_1 - \bar{y})}{n \cdot S_x \cdot S_y}$$

- הנוסחה שקיבלנו רגע לפני נוסחת r היא נוסחת השונות המשותפת – Covariance:

$$\frac{\sum(x_1 - \bar{x})(y_1 - \bar{y})}{n}$$

  תלמידי 4 יח"ל לא צריכים להכיר אותה אבל היא מאוד מפורסמת ויהיה מספיק זמן להתיידד איתה באוניברסיטה.
- בהמשך להסבר למעלה על השימוש בחישוב של ציוני תקן, ניתן להציג את נוסחת r גם בעזרת ציוני תקן. (חשוב לשים לב שהנוסחה לחישוב r בציוני תקן מופיעה בדף הנוסחאות, אך אין בה שימוש משמעותי במסגרת התוכנית):

$$r = \frac{\sum(x_1 - \bar{x})(y_1 - \bar{y})}{n \cdot S_x \cdot S_y} = \frac{\sum Z_x \cdot Z_y}{n}$$

- שימו לב, כי בדף הנוסחאות נוסחת מקדם המתאם מוצגת באופן מעט שונה. מומלץ ללמד את נוסחת r כפי שהוסבר בפרק זה כדי לעזור לתלמידים להבין את הנוסחה מבחינה דידקטית. זה יעזור גם לשמור על תאימות לספרי סטטיסטיקה ולמקורות מתמטיים. אולם לקראת הבגרות יש להכיר לתלמידים את הנוסחה כפי שמופיעה בדף הנוסחאות על מנת שיכירו אותה בצורתה זו.

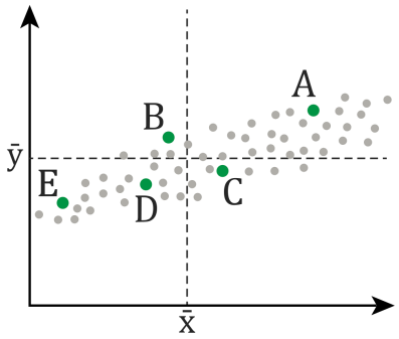
מקדם המתאם הלינארי
גרפי

מקדם המתאם הלינארי
גרפי-אלגברי

1. לפניכם דיאגרמת פיזור, עליה מסומנים קווי הממוצעים ו- 5 נקודות.

א. השלימו:

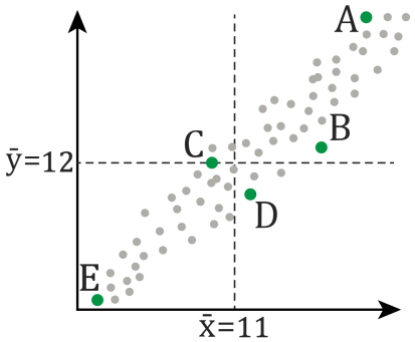
- ערך ה-x של נקודה A גדול/קטן מ-$\bar{x}$,
  ערך ה-y שלה גדול/קטן מ-$\bar{y}$,
  לכן היא תתרום ערך חיובי/שלילי למונה הנוסחה.
- ערך ה-x של נקודה B גדול/קטן מ-$\bar{x}$,
  ערך ה-y שלה גדול/קטן מ-$\bar{y}$,
  לכן היא תתרום ערך חיובי/שלילי למונה הנוסחה.



ב. קבעו עבור כל אחת מהנקודות האחרות אם היא תתרום ערך חיובי / שלילי / אפס למונה נוסחת מקדם המתאם.

2. לפניכם דיאגרמת פיזור שמסומנים בה קווי הממוצעים ו-5 נקודות.

א. קבעו עבור כל אחת מהנקודות,
אם היא תתרום ערך חיובי/ שלילי / אפס
כאשר היא תוצב במונה של נוסחת מקדם המתאם.



להלן שיעורי הנקודות:

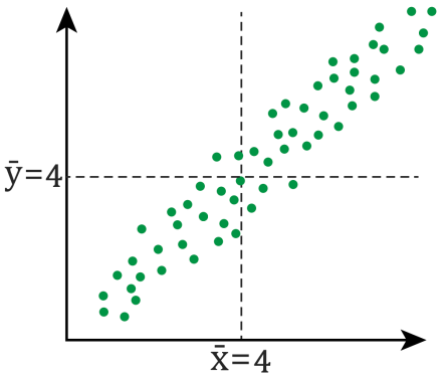
$A(22,24)$ $B(20,15)$ $C(8,12)$ $D(13,7)$ $E(2,2)$

ב. חשבו עבור כל נקודה מה יהיה הערך שיתקבל עבורה כאשר תוצב בביטוי $(x_1-\bar{x})(y_1-\bar{y})$.

ג. איזו מהנקודות תתרום את הערך הגדול ביותר למונה נוסחת מקדם המתאם r?

רשות: האם יש קשר בין תרומת הנקודה למונה הנוסחה לבין גודל המרחק שלה מנקודת הממוצעים?

3. לפניכם דיאגרמת פיזור, עליה מסומנים קווי הממוצעים.



א. הוסיפו לגרף נקודה כך שתתרום ערך חיובי למונה נוסחת מקדם המתאם.

ב. הוסיפו לגרף נקודה כך שתתרום ערך שלילי למונה נוסחת מקדם המתאם.

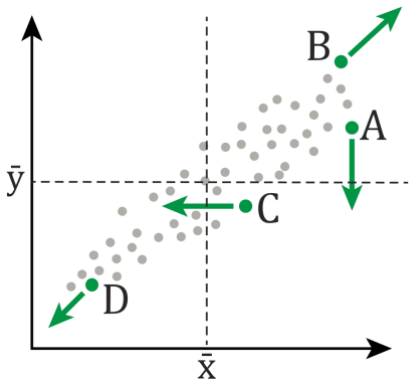
ג. הוסיפו לגרף נקודה כך שתתרום ערך חיובי, וערך ה-x שלה גדול מ-$\bar{x}$.

ד. הוסיפו לגרף נקודה כך שתתרום ערך חיובי, וערך ה-x שלה קטן מ-$\bar{x}$.

ה. הוסיפו לגרף נקודה כך שתתרום ערך שלילי, וערך ה-y שלה גדול מ-$\bar{y}$.

ו. הוסיפו לגרף נקודה כך שתתרום ערך שלילי, וערך ה-y שלה קטן מ-$\bar{y}$.

4. לפניכם דיאגרמת פיזור.



לאחר בדיקה התברר כי יש לשנות את מיקום הנקודות A, B, C, D בהתאם לחיצים.

כל אחת זזה ממקומה <u>לקצה החץ שמתואר</u>.

א. קבעו עבור כל נקודה אם התזוזה שלה מחזקת או מחלישה את עוצמת הקשר.

A - מחזקת / מחלישה B - מחזקת/ מחלישה

C - מחזקת / מחלישה D - מחזקת/ מחלישה

ב. עבור כל נקודה קבעו האם הערך שהיא תורמת למונה נוסחת מקדם המתאם חיובי/שלילי/אפס.

בדקו לפני התזוזה שלה ואחריה.

A - תרמה לפני התזוזה ערך _____ אחרי התזוזה תרמה ערך _____

B - תרמה לפני התזוזה ערך _____ אחרי התזוזה תרמה ערך _____

C - תרמה לפני התזוזה ערך _____ אחרי התזוזה תרמה ערך _____

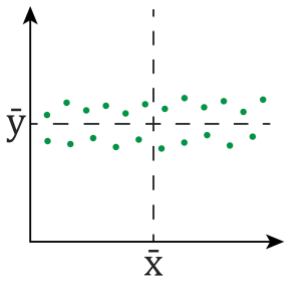
D - תרמה לפני התזוזה ערך _____ אחרי התזוזה תרמה ערך _____

5. לפניכם דיאגרמות פיזור, בכל אחת מהן ענו על הסעיפים הבאים:

א. סמנו + על כל הנקודות שתורמות ערך חיובי לנוסחת מקדם המתאם.

ב. סמנו – על כל הנקודות שתורמות ערך שלילי לנוסחת מקדם המתאם.

i

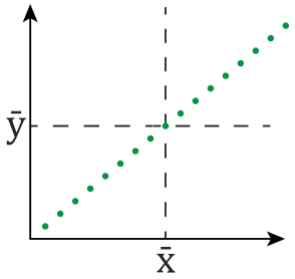


לאחר הסימון, מה תוכלו לומר על עוצמת הקשר הלינארי?

_____

_____

ii

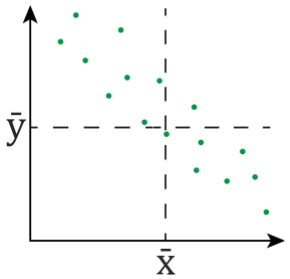


לאחר הסימון, מה תוכלו לומר על עוצמת הקשר הלינארי?

_____

_____

iii



לאחר הסימון, מה תוכלו לומר על עוצמת הקשר הלינארי?

_____

_____

חישוב ממוצע

חישוב סטיית תקן 1

חישוב סטיית תקן 2

מקדם המתאם בסיס 1

מקדם המתאם בסיס 2

חישוב ממוצע מתוך טבלת נתונים

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_n f_n}{n}$$

1. בטבלאות הנתונים הבאות, חשבו את הממוצע עבור כל אחד מהמשתנים: x, y.

א

x	y
3	22
5	26
9	45
7	32
8	13
$\bar{x} =$ ____	$\bar{y} =$ ___

ב

x	y
10.2	87
11.6	45
13	39
14.1	40
17.1	32
18	75
$\bar{x} =$ ____	$\bar{y} =$ ___

ג

x	y
4500	20
3240	14
7600	26
8400	24
8400	28
6560	20
$\bar{x} =$ ____	$\bar{y} =$ ___

2. להלן טבלאות נתונים, בכל אחת מהם יש ערך שחסר. חשבו אותו.

א

x	y
34	2.1
46	4.5
X	1.7
27	Y
46	6.5
$\bar{x} = 41$	$\bar{y} = 3.4$

ב

x	y
X	5
420	4
520	7
460	3
370	8
430	Y
$\bar{x} = 420$	$\bar{y} = 6$

חישוב סטיית תקן מתוך טבלת נתונים

$$S=\sqrt{\frac{(x_1-\bar{x})^2\cdot f_1+(x_2-\bar{x})^2\cdot f_2+\cdots+(x_n-\bar{x})^2\cdot f_n}{n}}$$

3. בטבלאות הנתונים הבאות, חשבו את הממוצע ואת סטית התקן עבור כל אחד מהמשתנים: x, y.

א

y	x
25	2
27	4
32	6
36	8
40	10
$\bar{y}=32$	$\bar{x}=6$
$S_y=$ ___	$S_x=$ ___

ב

y	x
36	3
38	4
44	7
46	8
56	12
62	14
$\bar{y}=47$	$\bar{x}=8$
$S_y=$ ___	$S_x=$ ___

ג

y	x
2.5	15
3.2	13
1.8	14
2.7	10
4.3	11
3.5	18
$\bar{y}=$ ___	$\bar{x}=$ ___
$S_y=$ ___	$S_x=$ ___

חישוב מקדם המתאם

$$r=\frac{(x_1-\bar{x})(y_1-\bar{y})+\cdots+(x_n-\bar{x})(y_n-\bar{y})}{n\cdot S_x\cdot S_y}$$

4. בטבלאות הנתונים הבאות, נתונים הערכים של כל אחד מהמשתנים.

חשבו עבור כל אחת מהם את מקדם המתאם.

א

y	x
24	63
27	61
30	59
33	57
36	55
$\bar{y}=30$	$\bar{x}=59$
$S_y=4.243$	$S_x=2.828$
$r=$ _____	

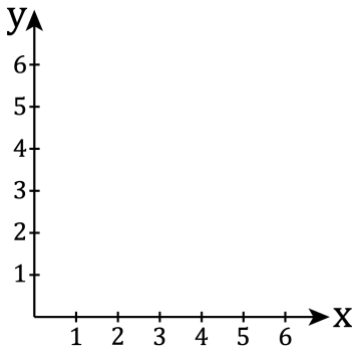
ב

y	x
12.2	42
14.4	38
16.6	26
18.8	28
21	31
23.2	33
$\bar{y}=$ ___	$\bar{x}=33$
$S_y=3.757$	$S_x=$ ___
$r=$ _____	

ג

y	x
12	4
18	6
17	9
19	2
11	5
22	10
$\bar{y}=$ ___	$\bar{x}=$ ___
$S_y=$ ___	$S_x=$ ___
$r=$ _____	

5. בית ספר בדק את הקשר בין מספר שעות הקריאה בשבוע של תלמידי כיתה ג׳ לבין מספר שגיאות הכתיב בעמוד, הנתונים רוכזו בטבלה הבאה:



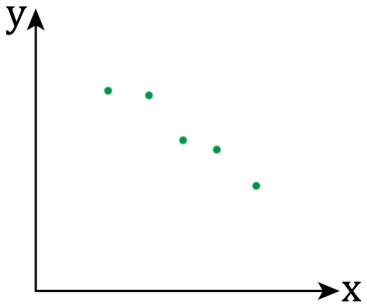
שעות קריאה (משתנה x)	שגיאות הכתיב (משתנה y)
3	5
5	2
4	3
2	6
1	4

א. שרטטו דיאגרמת פיזור מתאימה לנתונים.

ב. חשבו את מקדם המתאם, וקבעו אם הקשר שנמצא חיובי או שלילי.

ג. שני תלמידים נוספים נדגמו, שניהם קוראים 6 שעות בשבוע. כשהוסיפו את הנתונים שלהם נמצא שאחד מהם חיזק את עוצמת הקשר ואחד מהם החליש אותו.
הציעו ערך אפשרי למספר שגיאות הכתיב של כל אחד מהם.

6. בית ספר בדק את הקשר בין מספר התלמידים בכיתה לבין הציון הממוצע בהיסטוריה, הנתונים מוצגים בטבלה הבאה, ובדיאגרמת הפיזור:



מספר התלמידים (משתנה x)	הציון בהיסטוריה (משתנה y)
31	74
20	83
26	82
36	72
42	65

א. התבוננו בדיאגרמת הפיזור, ושערו אילו מבין ערכי r הבאים יכול להתאים לגרף:

$$r_1 = 0.875 \qquad r_2 = -0.875 \qquad r_3 = 0.972$$
$$r_4 = -0.972 \qquad r_5 = -0.152 \qquad r_6 = -1$$

ב. חשבו את מקדם המתאם הלינארי.

ג. הוכנסה כיתה נוספת למדגם, בה מספר התלמידים הוא 43 והציון הממוצע בהיסטוריה בכיתה הוא 90.
קבעו על כל טענה אם היא נכונה, נמקו.

i. ממוצע התלמידים בכיתה לא השתנה.

ii. סטיית התקן של הציונים בהיסטוריה השתנתה.

iii. עוצמת הקשר השתנתה.

יחידה שלישית

קשר בטבלאות נתונים

קשר בטבלאות נתונים

עד כה למדנו לחשב את נוסחת המתאם, וכן לקבוע את קיומו של קשר על פי צורת הנקודות בדיאגרמת פיזור. כעת נלמד לקרוא טבלת נתונים עם מספר קטן יחסית של נתונים, ולקבוע ללא חישוב מהי מגמת הקשר שמופיעה בה. אומנם אנחנו כבר יודעים לחשב את r ולקבוע בצורה מדויקת את ערכו, אבל במקרים שבהם נתון מספר נמוך של תצפיות ניעזר בעקרונות אחרים כדי לקבוע אם קיים קשר בין המשתנים.

הערה: בעולם האמיתי רוב הזמן נעשה שימוש במאות ואלפי נתונים והטבלאות הן ארוכות מאוד, אבל במצבים כאלו חישוב הקשר ייעשה בתוכנה המיועדת לכך. במקרים שלפנינו, ננסה להבין מתוך התבוננות בנתונים האם יש או אין קשר, ולא ניעזר לשם כך בחישוב ממשי של מקדם המתאם.

מבט אופקי לעומת אנכי

נתחיל בשתי דוגמאות פשוטות. בטבלה שלפנינו קיים קשר פשוט וברור בין ערכי x לערכי y. מאחר שהוא מתקיים בצורה אחידה עבור כל התצפיות, אין יוצא מהכלל, נוכל לקבוע שהקשר הוא מושלם.

במקרה הזה כדי לנבא את y מתוך x צריך לכפול את x פעמיים ולהגיע ל-y. זה נכון עבור כל התצפיות. במילים אחרות הניבוי מדויק והקשר מושלם.

משתנה X	משתנה Y
2	4
3	6
4	8
5	10

x2 x2 x2 x2

אבל מה נעשה כשקשה יותר לעבור מ-x ל-y?

למשל בטבלה הבאה פחות ברור הקשר בין שני המשתנים.

לכן, במקום לנסות למצוא את הקשר האופקי בין המשתנים, נתבונן אנכית על "הקפיצות" ב-x וב-y, ונשווה ביניהן:

ניתן לראות שערכי x גדלים בצורה קבועה ב-2, ואילו ערכי y גדלים בצורה קבועה ב-4. כלומר "הקפיצות" בערכי x ובערכי y מתואמות ביניהן בצורה מושלמת. המשמעות היא שלמרות שקצת יותר קשה לראות את זה במבט ראשון, גם במקרה הזה הקשר הוא מושלם, r=1.

משתנה X	משתנה Y
10	18
12	22
14	26
16	30

+2 +2 +2 +4 +4 +4

קשר חיובי מושלם r = 1

בקשר מושלם נצפה לתיאום מלא בין הקפיצות ב-x לבין הקפיצות ב-y.

אבל שימו לב שכאשר אנחנו מתיחסים לתיאום מושלם בין "הקפיצות", הן לא בהכרח צריכות להיות זהות אחת לשניה.

למשל בדוגמה הבאה, הקפיצות ב-x מתואמות בצורה מושלמת עם "הקפיצות" ב-y, אבל לא מדובר בקפיצות זהות אחת לשניה ב-x, אלא שהתיאום בינן לבין "הקפיצות" במשתנה y הוא מושלם. r=1.

משתנה X	משתנה Y
9	21
12	27
14	31
19	41

+3 +2 +5 +6 +4 +10

הערה חשובה: כדי לעבוד בשיטת "הקפיצות", עלינו ראשית לסדר את משתנה x לפי סדר עולה של הערכים, כדי שיהיה קל לראות את הקשר בין הקפיצות בשני המשתנים.

קשר חיובי שאינו מושלם

$$0 < r < 1$$

כאשר הקשר הוא חיובי אבל לא מושלם, נראה "קפיצות" בכיוון חיובי גם במשתנה x וגם במשתנה y, אבל הקפיצות לא ישמרו על יחס קבוע ביניהן, למשל בדוגמה הבאה:

משתנה X	משתנה Y
10	21
12	28
15	30
16	32

+2 +3 +1 (X) | +7 +2 +2 (Y)

למעשה, גם אם לא כל "הקפיצות" חיוביות, אבל רובן המכריע בכיוון חיובי- עדיין הקשר יהיה חיובי. בקשר שלפנינו נוספה תצפית חמישית, שמציגה מגמה שונה ממה שציפינו, "הקפיצה" שלה ב-x חיובית, ואילו "הקפיצה" ב-y שלילית. כיון שהשינוי מינורי יחסית, עדיין הקשר יישאר חיובי (חלש יותר מאשר בדוגמה שלמעלה)

<u>חשוב להדגיש שמקרים מורכבים יותר מהדיאגרמה שלפנינו לא יידרשו ב-4 יחל.</u>

משתנה X	משתנה Y
10	21
12	28
15	30
16	32
18	29

+2 +3 +1 +2 (X) | +7 +2 +2 +2 -3 (Y)

קשרים שליליים

$$r = -1$$

בקשרים שליליים, נצפה שככל שערכי ה-x יגדלו, ערכי ה-y יקטנו. כלומר אם "הקפיצות" ב-x חיוביות, "הקפיצות" ב-y יהיו שליליות. בטבלה שלפנינו ניתן לראות קשר שלילי מושלם: הקפיצות החיוביות ב-x מתואמות בצורה מושלמת עם הקפיצות השליליות במשתנה y.

משתנה X	משתנה Y
21	10
23	8
27	4
32	-1

+2 +4 +5 (X) | -2 -4 -5 (Y)

$$-1 < r < 0$$

בטבלה שלפנינו ניתן לראות שככל שערכי x גדלים (קפיצות חיוביות) ערכי y קטנים (קפיצות שליליות).

משתנה X	משתנה Y
10	10
13	9
17	7
20	2

+3 +4 +3 (X) | -1 -2 -5 (Y)

היעדר קשר

$$r = 0$$

במצב שבו אין קשר נוכל לראות שינוי בערכי x שאינו מתואם כלל עם השינויים בערכי y. למשל בדוגמה שלפנינו, ניתן לראות עליה בערכי x, ומנגד, ערכי y עולים ויורדים לסירוגין ללא כל קשר לשינוי שחל בערכי x. במצב כזה הקשר הוא 0.

משתנה X	משתנה Y
21	2
26	4
31	2
36	4
41	2

+5 +5 +5 +5 (X) | +2 -2 +2 -2 (Y)

הסקה על קשר מטבלאות נתונים

1. לפניך טבלאות נתונים שונות.

קבע לגבי כל אחת מהן, מה יתאר אותה בצורה הטובה ביותר.

r לא מוגדר	$r=-1$	$-1<r<0$	$r=0$	$0<r<1$	$r=1$

א

x	y
2	22
4	44
5	55
7	77
9	99

ב

x	y
1	7
2	8
3	9
4	10
5	11

ג

x	y
7	3
8	6
9	9
10	12
11	15

ד

x	y
1	20
2	18
3	16
4	14
5	12

ה

x	y
5	10
6	9
7	8
8	7
9	6

ו

x	y
1	1
10	10
3	3
10	10
12	12

ז

x	y
13	7
16	7
17	7
18	7
21	7

ח

x	y
1	3
2	11
3	3
4	11
5	3

ט

x	y
1	4
1	4
1	4
2	5
4	7

י

x	y
2	32
4	10
6	32
8	10
10	32

יא

x	y
1	1
2	7
3	12
6	14
7	12

יב

x	y
3	19
6	18
8	11
11	7
17	1

יג

x	y
3	71
12	68
18	60
24	54
32	56

יד

x	y
13.5	9
13.5	11
13.5	12
13.5	13
13.5	17

טו

x	y
0	1
10	2
20	3
30	4
40	5

טז

x	y
101	7
103	9
112	13
120	14
121	18

2. לפניך טבלאות נתונים. סדרו את נתוני הטבלאות, וקבעו לגבי כל אחת מהן מה מהבאים יתאר בצורה הטובה ביותר את הקשר בין המשתנים:

$r = 1$	$0 < r < 1$	$r = 0$	$-1 < r < 0$	$r = -1$	r לא מוגדר

א

x	y
13	32
14	38
10	20
7	2
9	12

ב

x	y
25	10
31	22
40	40
22	4
34	28

ג

x	y
6	6
3	14
1	18
2	16
5	8

ד

x	y
12.8	45
13.4	30
12.2	60
12.6	50
13	40

3. לפניך טבלאות נתונים שונות.

לכל אחת מהטבלאות הוסיפו את <u>אחת מבין הנקודות</u> C,B,A.

קבעו עבור כל נקודה אם התוספת שלה תחזק / תחליש / לא תשנה את עוצמת הקשר.

א

x	y
1	9
2	11
3	13
4	15
5	17

A	$y = 19 \quad x = 6$	תחזק / תחליש / לא תשנה
B	$y = 21 \quad x = 6$	תחזק / תחליש / לא תשנה
C	$y = 27 \quad x = 10$	תחזק / תחליש / לא תשנה

ב

x	y
21	31
24	34
26	36
29	39
30	40

A	$y = 43 \quad x = 33$	תחזק / תחליש / לא תשנה
B	$y = 50 \quad x = 41$	תחזק / תחליש / לא תשנה
C	$y = 10 \quad x = 20$	תחזק / תחליש / לא תשנה

ג

x	y
10	23
11	20
12	20
13	19
14	17

A	$y = 14 \quad x = 15$	תחזק / תחליש / לא תשנה
B	$y = 12 \quad x = 16$	תחזק / תחליש / לא תשנה
C	$y = 23 \quad x = 16$	תחזק / תחליש / לא תשנה

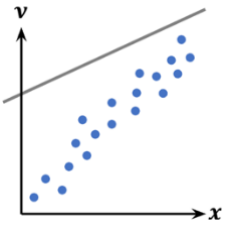
יחידה רביעית

חישוב קו הרגרסיה

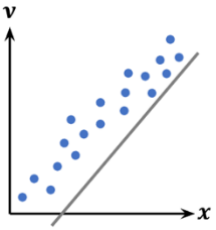
עד כה דנו הרבה במקדם המתאם הלינארי, שסייע לנו לקבוע האם קיים או לא קיים קשר לינארי בין שני משתנים. כעת נעבור לשאלה מעשית וחשובה אחרת – אם קיים קשר בין המשתנים, כיצד נוכל להשתמש בקשר הזה בצורה מעשית כדי לנבא את ערכי y מתוך ערכי x?

או במילים אחרות, מהו הקו שמתאר בצורה הטובה ביותר את הקשר בין המשתנים?

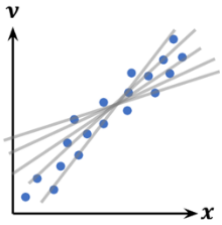
כדי למצוא את קו הרגרסיה אנחנו מחפשים קו שיענה על כמה דרישות: שיעבור בקרבת עננת התצפיות, עדיף ממש בתוכה, ושהנקודות יהיו קרובות אליו ככל האפשר. כלומר הקו צריך לעבור במרכז העננה:



ברור שהקו לא יעבור כאן...



וגם לא כאן...

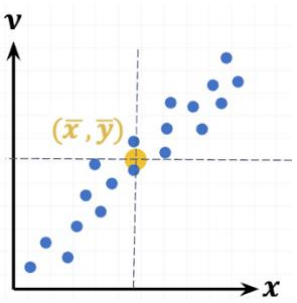


הוא יעבור בצורה כלשהי
במרכז ענן התצפיות

אז כיצד נחשב את הקו הזה? ננסה לחפש **נקודה ושיפוע** כדי לחשב את הקו.

לגבי נקודה, זה קל - הקו צריך לעבוד במרכז העננה, כלומר נחפש נקודה שנמצאת בדיוק במרכז עננת התצפיות. ואיזו נקודה נמצאת במרכז עננת התצפיות? קללל שנקודת הממוצעים!

ואכן קו הרגרסיה עובר בנקודת הממוצעים:



אבל מה לגבי שיפוע הקו?

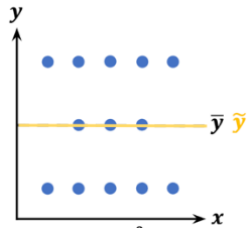
מתברר שיש שלושה גורמים שמשפיעים על שיפוע הקו:

- חוזק הקשר בין המשתנים (r)
- הפיזור של משתנה y (S_y)
- הפיזור של משתנה x (S_x)

נראה כל אחד מהגורמים הללו בצורה גרפית.

השפעת r על שיפוע קו הרגרסיה

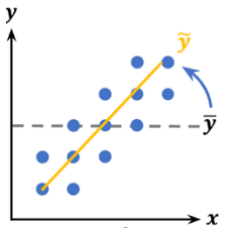
כזכור, כאשר הקשר הוא r=0 קו הניבוי הוא למעשה קו הממוצע של y, כלומר השיפוע של קו הניבוי גם הוא 0. אבל ככל שמתחזק הקשר בין המשתנים קו הניבוי הולך "ונפרד" מקו הממוצע, והשיפוע של קו הרגרסיה הולך וגדל. במילים אחרות - ככל ש-r גדול יותר, גם שיפוע קו הרגרסיה גדול יותר. ככל ש-r קטן יותר, גם שיפוע קו הרגרסיה יקטן:



$r = 0$

קו הרגרסיה מתלכד עם קו הממוצע $\bar{y}$

$m = 0$



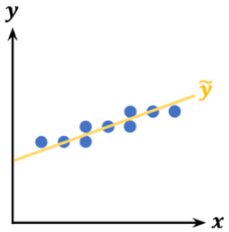
$r > 0$

קו הרגרסיה "נפרד" מקו הממוצע $\bar{y}$

$m > 0$

השפעת S_y על שיפוע קו הרגרסיה

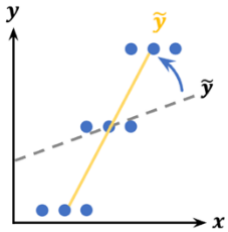
ככל שפיזור המשתנה y גדל, כך הנקודות רחוקות יותר זו מזו ומפוזרות יותר לאורך הערכים בציר y. במצב כזה, כדי להישאר קרוב לנקודות של הדיאגרמה, קו הרגרסיה "יתקרב" לכיוון הנקודות שכעת מפוזרות יותר לאורכו של ציר y. לכן אם S_y גדלה, גם שיפוע הקו יגדל, כפי שניתן לראות בשרטוט הבא:



פיזור אנכי קטן

קטנה S_y

קטן m



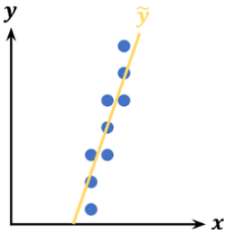
פיזור אנכי גדול

גדולה S_y

גדול m

השפעת S_x על שיפוע קו הרגרסיה

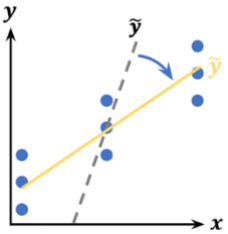
ככל שפיזור המשתנה x גדל, כך הנקודות מתרחקות זו מזו לאורך ציר x. שימו לב שבמצב כזה, כדי שקו הרגרסיה יישאר קרוב לנקודות של הדיאגרמה, עליו "להתקרב" לכיוון הנקודות שמפוזרות לאורכו של ציר x. לכן אם S_x גדלה, שיפוע הקו דווקא ילך ויקטן, כפי שניתן לראות בשרטוט הבא:



פיזור אופקי קטן

קטנה S_x

גדול m



פיזור אופקי גדול

גדולה S_x

קטן m

שלושת הגורמים שמשפיעים על שיפוע קו הרגרסיה מסתכמים בנוסחה הפשוטה הבאה:

$$m = r \cdot \frac{S_y}{S_x}$$

מאחר שיש לנו כעת נקודה ושיפוע, נוכל לחשב את קו הרגרסיה עצמו: $y - \bar{y} = r \cdot \frac{S_y}{S_x}(x - \bar{x})$

דגשים

- סימן השיפוע תמיד זהה לסימנו של הקשר.
- כאשר r=0 הוא אפס גם השיפוע הוא אפס m=0.
- אם הקשר לא מושלם, לעולם לא נוכל לדעת אם הערך שניבאנו ל-y על פי x הוא זהה או שונה מהערך שיתקבל בפועל עבור תצפית עם אותו ערך x.
  כלומר, זה שיש לנו ניבוי לא אומר שזה הערך האמיתי של התצפית! ניבוי הוא רק ניבוי...
- נוסחת קו הרגרסיה שחישבנו מתאימה רק לניבוי ערכי y על ידי הצבה של ערכי x. לא ניתן לעשות בה את הפעולה ההפוכה ולהציב את ערכי y ולקבל מתוכם את ערכי x המנובאים (למורים, ראו הרחבה בהערות למורה).

הערות למורה בלבד

1. כאשר נבחר ערכי x להציב בנוסחת הקו ולחשב עבורם ערכי y מנובאים, נוכל כמובן לבחור ערכים שלא נדגמו בפועל במדגם, אלא ערכים שונים. למעשה, נוכל גם לבחור ערכים מחוץ לטווח הערכים ששימש אותנו במדגם אותו חקרנו, אך סמוכים יחסית אליו. חשוב להדגיש כי לפעמים יש מגבלות על בחירת ערכי x שמותר להשתמש בהם (למשל, ערכים מאוד רחוקים מטווח התצפיות בו השתמשנו במדגם) <u>אך נושאים אלו אינם נכללים בתוכנית הלימוד ל-4 יחל</u>. עבור תלמידי 4 יחל, ניתן להציב בקו הרגרסיה את כל ערכי X של המשתנה.
2. במסגרת תוכנית 471 אנחנו לומדים רק את ניבוי y מתוך x. אולם מחוץ למסגרת התוכנית ניבוי y מתוך x שקול למעשה לניבוי x מתוך y והבחירה במי מהם להשתמש תלויה במטרות החוקר והיא ענין של נוחות. חשוב לזכור שמקדם המתאם הוא מדד סימטרי לעוצמת הקשר, וכפועל יוצא, ניתן לנבא את משתנה y לפי x, או להיפך, לפי בחירת החוקר.
3. אומנם r הוא מדד סימטרי לעוצמת הקשר, אולם לא ניתן להשתמש באותו הקו שמשמש אותנו לניבוי y מתוך x כדי לבצע את הניבוי של x מתוך y.
   הנוסחה של קו הרגרסיה לניבוי x מתוך y היא : $\tilde{x} = r \cdot \frac{S_x}{S_y} + b$. ניתן לראות ששיפוע הקו אומנם מתנהג לפי חוקים דומים לשיפוע הקו לניבוי y, אולם למרות זאת הוא שונה ממנו- וגם אינו הופכי לו, כפי שאולי היינו מצפים.
4. אם נבטא את דיאגרמת הפיזור בציוני תקן – ולא בציוני גלם - או אז תתבטא במלואה תכונת הסימטריות של r, וקווי הרגרסיה לניבוי y ולניבוי x יהיו הפוכים זה לזה בצורה סימטרית לגמרי.

שיפוע קו הרגרסיה

שיפוע m של ישר הרגרסיה: $m = r \cdot \frac{S_y}{S_x}$ משוואת ישר הרגרסיה: $y - \bar{y} = m(x - \bar{x})$

1. חשבו את שיפוע ישר הרגרסיה בהתאם לנתונים הבאים:

	s_x	s_y	r	m
א	1.23	12.42	0.324	
ב	32	24.5	-0.713	
ג	141.3	203.24	0.866	
ד	4.87	9.69	-0.932	

2. חשבו את הנתון החסר בהתאם לנתונים הידועים:

	s_x	s_y	r	m
א	16.51		-0.882	-0.41
ב		1.61	0.341	0.514
ג	1.265	1.14		0.859
ד		0.65	-0.268	-0.8

3. מצאו את משוואת ישר הרגרסיה בהתאם לנתונים הבאים:

	s_x	s_y	r	m	$\bar{x}$	$\bar{y}$	משוואת הישר
א	6.709	2.83	0.93		45	7.5	
ב	5.478	10.61	-0.24		24	83	
ג	22.1	13.09	0.67		150	400	
ד	1.18	3.231	-0.881		8.5	12.2	

4. בטבלה הבאה מרוכזים נתונים אודות קשר לינארי בין שני משתנים.
בכל סעיף חלק מהנתונים חסרים. הסתמכו על משוואת הישר הנתונה והשלימו את החסר.

	s_x	s_y	r	m	$\bar{x}$	$\bar{y}$	משוואת הישר
א	10	5				15	$\tilde{y} = 0.45x + 1.5$
ב	6		-0.72		20		$\tilde{y} = -1.44x + 108.8$
ג	25	30			100		$\tilde{y} = 1.2x + 130$
ד		2	-0.65			9.5	$\tilde{y} = -0.3x + 11$

5. נמצא כי קיים קשר חזק בין מספר העובדים בחברה לבין כמות המיילים שנשלחים בה ביום. חושב קו הרגרסיה לניבוי כמות המיילים היומית מתוך כמות עובדי החברה, והתקבל שהוא $\tilde{y} = 5.5x - 38.5$.
ידוע כי כמות המיילים הממוצעת בקרב החברות שהשתתפו בסקר היא 132 מיילים ליום.

א. עבור חברה שיש בה 35 עובדים, מהי כמות המיילים היומית המנובאת?

ב. מהי כמות המיילים המדויקת שנשלחת בחברה שיש בה 35 עובדים?
בדיוק 154 / פחות מ- 154 / יותר מ-154 / לא ניתן לדעת.

ג. מהי כמות העובדים הממוצעת בקרב החברות שהשתתפו בסקר?

6. מחקר גריאטרי בדק את הקשר בין מספר התלונות הרפואיות של אנשים מבוגרים (משתנה x) לבין רמת שמחת החיים שלהם (משתנה y).
נמצא כי קיים קשר הפוך בין כמות התלונות לבין רמת שמחת החיים: ככל שאנשים מדווחים על מספר גבוה יותר של בעיות רפואיות, כך שמחת החיים שלהם יורדת: $r = -0.57$.
כמו כן נמצא שהממוצע של מספר התלונות הרפואיות הוא 3, עם סטית התקן של 1.8.
ידוע שקו הרגרסיה לניבוי רמת שמחת החיים מתוך כמות התלונות הרפואיות הוא:

$$\tilde{y} = -1.14x + 11.42$$

א. מהו ממוצע שמחת החיים של הנבדקים במחקר?

ב. מהי סטית התקן של שמחת החיים של הנבדקים במחקר?

ג. מהי רמת שמחת החיים המנובאת לאדם שכמות התלונות הרפואיות שלו היא 2?

ד. במרפאה הרופא פגש אדם שהתלונן 6 פעמים.
האם ניתן לקבוע כי שמחת החיים שלו היא 4.58?

ה. במרפאה הרופא פגש אדם שהתלונן 3 פעמים.
האם ניתן לקבוע כי שמחת החיים שלו היא 8?

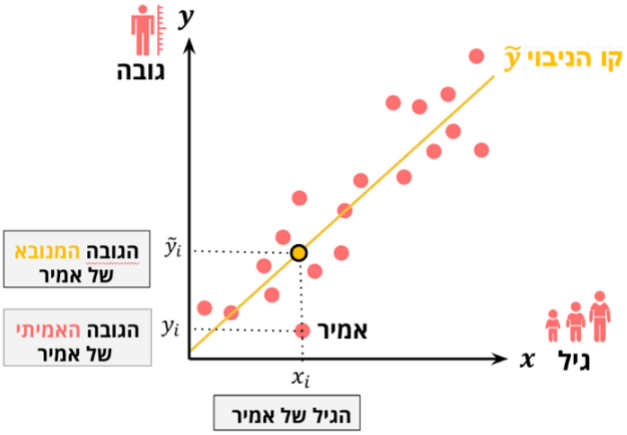
יחידה חמישית

תצפיות אמיתיות ותצפיות מנובאות

תצפיות מנובאות

עד עכשיו הכרנו את התצפיות האמיתיות בדיאגרמת הפיזור. כלומר אנשים אמיתיים שדגמנו, וחישבנו על בסיס הנתונים שלהם את עוצמת הקשר ואת קו הרגרסיה. אבל כעת, כשיש לנו קו רגרסיה ביד, למעשה אנחנו יכולים בעזרתו לנבא לכל ערך x (בין אם דגמנו אותו במדגם שלנו, ובן אם לא דגמנו אותו) את ערך y הצפוי לו על פי קו הרגרסיה.

נדגים בעזרת דיאגרמת פיזור את ההבדל בין הגובה האמיתי של אמיר, לגובה שמנובא לו על ידי קו הרגרסיה:

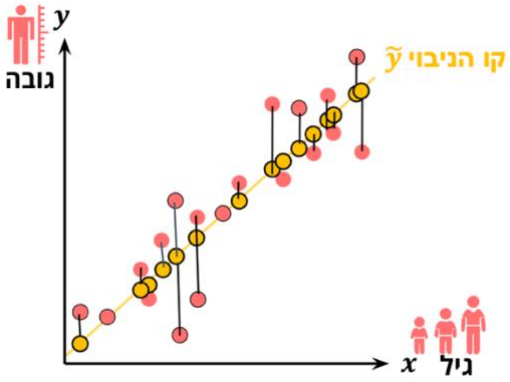


ניתן לראות שבדיאגרמה מופיעה באדום **התצפית האמיתית** של אמיר, כלומר הגיל שלו - x_i, והגובה שלו - y_i.

בנוסף, מופיעה **התצפית המנובאת** לאמיר על פי קו הרגרסיה, שהיא התצפית הצהובה על קו הניבוי.

אם נתבונן בתצפית הצהובה נראה שערך ה-x שלה זהה לערך ה-x של אמיר, ואילו ערך ה-y שלה הוא ערך תיאורטי, שיתקבל אם נציב את ערך ה-x של אמיר בנוסחת קו הרגרסיה.

למעשה במצב כזה נוצרים בדיאגרמת הפיזור שני "סוגים" של תצפיות: תצפיות אמיתיות, אותן דגמנו במדגם, ותצפיות מנובאות, שיתקבלו אם נציב את ערכי x שדגמנו בנוסחת קו הרגרסיה.



בדיאגרמה הבאה נוכל לראות שקו הרגרסיה מאפשר להתאים לכל תצפית אמיתית במדגם, את התצפית המנובאת שלה, וכך נוצרות למעשה "זוגות" של תצפיות: כל זוג תצפיות מורכב **מהתצפית האמיתית**, **ומהתצפית המנובאת** על ידי קו הרגרסיה.

ערך ה-y של התצפית האמיתית יסומן y_i וערך ה-y של התצפית המנובאת יסומן $\tilde{y}_i$. כמובן, שבכל זוג תצפיות, ערך ה-x זהה.

בדוגמאות אלו עסקנו בערכים מנובאים עבור תצפיות שנדגמו במדגם המקורי. אולם חשוב להדגיש שלאחר שחושב קו הרגרסיה ניתן להציב בו ערכי x שונים, לאו דווקא כאלו שנדגמו במדגם המקורי, ולנבא עבורם את ערכי y הצפויים על פי קו הרגרסיה.

לגבי תצפיות אמיתיות ומנובאות חשוב לזכור מספר עקרונות מנחים:

הקשר בין ערך ה-y האמיתי לבין ערך y המנובא

כאשר הקשר אינו מושלם, התצפית שקו הרגרסיה מנבא כמובן לא חייבת להיות בהכרח זהה לתצפית האמיתית שהתקבלה במדגם. נכון לומר שבאופן כללי, ככל שחוזק הקשר גדול יותר, כך התצפיות במדגם קרובות יותר לקו ישר, ולכן גם קרובות יותר לתצפיות המנובאות. אבל עבור תצפית בודדת, לעולם לא נוכל לדעת מה מרחקה מהערך המנובא שלה.

מכאן, שהתצפית האמיתית עשויה להיות גדולה יותר, קטנה יותר, או זהה לתצפית המנובאת.

ניתן לסמן את יחסי הגודל בין התצפית האמיתית למנובאת באופן הבא:

1. התצפית המנובאת גדולה מהתצפית האמיתית $\tilde{y}_i > y_i$
2. התצפית המנובאת קטנה מהתצפית האמיתית $\tilde{y}_i < y_i$
3. התצפית המנובאת שווה מהתצפית האמיתית. $\tilde{y}_i = y_i$

חשוב לזכור שיש מקרה אחד ויחיד שבו נוכל לקבוע בוודאות את הקשר בין הערך האמיתי והערך המנובא – בקשר שבו כל התצפיות האמיתיות נמצאות על קו הרגרסיה, כלומר בקשר מושלם. אם הקשר מושלם התצפיות האמיתיות והמנובאות תהיינה זהות אחת לשניה... כלומר לא תהיינה טעויות בניבוי (:

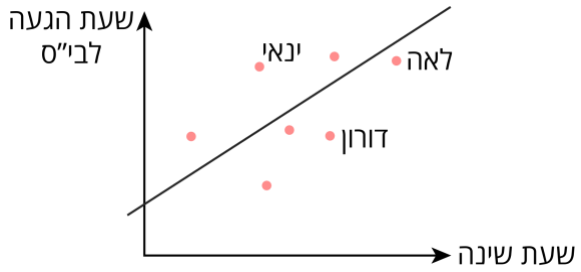
איך נדע מי מהתצפיות בגרף אמיתית ומי מנובאת?

בחלק מהגרפים מובאות זו עם זו תצפיות אמיתיות ומנובאות. במצב כזה עלינו לשים לב להיזהר לא ליפול במכשולים הבאים:

- לא כל התצפיות שנמצאות על קו הרגרסיה הן בהכרח תצפיות מנובאות!
  יתכן שתהיה תצפית אמיתית שנמצאת על קו הרגרסיה גם אם הקשר לא מושלם.
- אם התצפית <u>לא נמצאת</u> על קו הרגרסיה - בהכרח מדובר בתצפית אמיתית!
  זאת מכיוון שלא תיתכן תצפית מנובאת שלא נמצאת על קו הרגרסיה.

תצפיות אמיתיות ותצפיות מנובאות

1. לפניך דיאגרמות פיזור המתארת את הקשר בין שעת השינה של ילדים בכיתה י"ב לבין שעת ההגעה שלהם לבית הספר בבוקר שאחרי.



א. סמנו בנקודה על קו הרגרסיה את שעת ההגעה לבית הספר הצפויה ללאה.

ב. שעת ההגעה לבית הספר הצפויה ללאה <u>מוקדמת/מאוחרת</u> יותר משעת ההגעה שלה בפועל.

ג. סמנו בנקודה על קו הרגרסיה את שעת ההגעה לבית הספר הצפויה עבור ינאי.

ד. שעת ההגעה לבית הספר הצפויה לינאי <u>מוקדמת/מאוחרת</u> יותר משעת ההגעה שלו בפועל.

יאיר, תמר, נעמי וזיו הם תלמידים בכיתה י"ב, שהנתונים שלהם לא מוצגים בגרף.

ה. יאיר הגיע לבית הספר מוקדם יותר משעת ההגעה המנובאת לו ע"י קו הרגרסיה.
סמנו בגרף נקודה אפשרית עבורו.

ו. תמר הגיעה לבית הספר בדיוק בשעה שניבא לה קו הרגרסיה.
סמנו בגרף נקודה אפשרית עבורה.

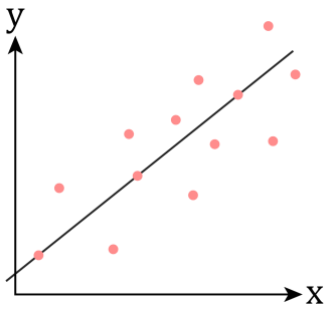
ז. ישר הרגרסיה מתאר עבור נעמי כי שעת ההגעה שלה תהיה 10:02.

- האם יתכן כי נעמי תגיע לבית הספר בשעה 10:02? נמקו.
- האם יתכן כי נעמי לא תגיע לבית הספר בשעה 10:02? נמקו.

ח. (רשות) ידוע כי זיו הלך לישון מאוחר יותר מדורון, והגיע לבית הספר מוקדם יותר ממנו.

- סמנו נקודה אפשרית עבור זיו.
- קבעו האם נכון או לא נכון:
  זיו הגיע מוקדם יותר משעת ההגעה המנובאת לו על פי קו הרגרסיה.

2. לפניך דיאגרמה המתארת תצפיות שנאספו וקו רגרסיה.



א. הקיפו בעיגול את הנקודות שבהן קו הרגרסיה מנבא ערך גבוה מהערך האמיתי שהתקבל, (כלומר הנקודות בהן $\tilde{y}_i > y_i$).

כמה נקודות כאלו יש בדיאגרמה? _____

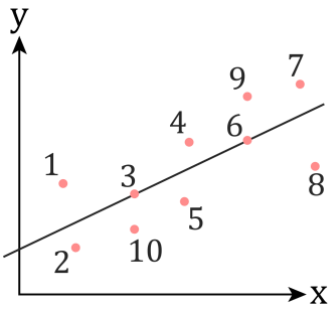
ב. סמנו × על הנקודות שבהן קו הרגרסיה מנבא ערך נמוך יותר מהערך האמיתי שהתקבל, (כלומר הנקודות בהן $\tilde{y}_i < y_i$).

כמה נקודות כאלו יש בדיאגרמה? _____

ג. סמנו ✓ על הנקודות שבהן קו הרגרסיה מנבא ניבוי זהה לערך האמיתי, (כלומר הנקודות בהן $\tilde{y}_i = y_i$).

כמה נקודות כאלו יש בדיאגרמה? _____

3. בגרף שלפניך מסומנות תצפיות וקו רגרסיה.



חלק מהתצפיות אמיתיות וחלקן מנובאות.

קבעו לגבי כל אחד מההיגדים הבאים האם נכון / לא נכון / לא ניתן לדעת.

א. תצפיות 7 ו-9 הן תצפיות אמיתיות.

ב. תצפיות 2 ו-10 הן תצפיות מנובאות.

ג. תצפיות 3 ו-6 הן תצפיות אמיתיות.

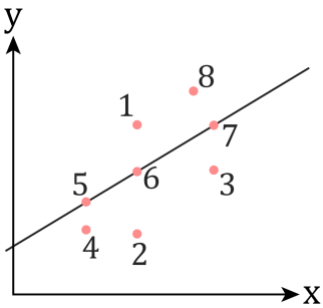
ד. תצפיות 3 ו-6 הן תצפיות מנובאות.

ה. תצפיות 5 ו-8 הן תצפיות מנובאות.

ו. יתכן כי תצפית 3 היא התצפית המנובאת של תצפית 10.

ז. יתכן כי תצפית 9 היא התצפית המנובאת של תצפית 6.

4. בגרף שלפניך התערבבו תצפיות אמיתיות ומנובאות,



חלק מהתצפיות אמיתיות וחלק מנובאות.

א. מה התצפית המנובאת לתצפית 3?

ב. מה התצפית המנובאת לתצפית 2?

ג. נתון כי תצפית 5 מנובאת ע"י קו הרגרסיה.

מה עשויה להיות התצפית האמיתית המתאימה לה?

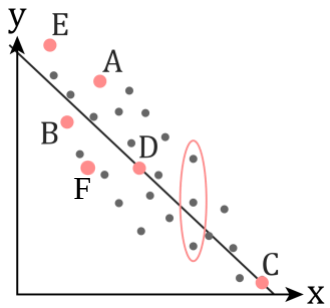
ד. נתון כי תצפית 6 מנובאת ע"י קו הרגרסיה.

מה עשויות להיות התצפיות האמיתיות המתאימות לה?

ה. החוקר דגם למחקרו אדם נוסף והוסיף אותו כנקודה בגרף. האם יתכן שערכה יהיה זהה לנקודה 3?

ו. החוקר דגם למחקרו אדם נוסף והוסיף אותו כנקודה בגרף. האם יתכן שערכה יהיה זהה לנקודה 7?

5. לפניך דיאגרמת פיזור.



מבין הנקודות המסומנות A,B,C,D,E,F

א. מהן הנקודות שבהן ערך ה-$\tilde{y}$ המנובא גדול מערך ה-y האמיתי? ___

ב. מהן הנקודות שבהן ערך ה-$\tilde{y}$ המנובא קטן מערך ה-y האמיתי? ___

ג. מהן הנקודות שבהן ערך ה-$\tilde{y}$ המנובא זהה לערך ה-y האמיתי? ___

התבוננו ב-3 הנקודות המוקפות בעיגול.

קבעו לגבי כל אחת מההיגדים הבאים האם נכון / לא נכון/ לא ניתן לדעת.

ד. ערך ה-x של כל הנקודות זהה.

ה. ערך ה-y של כל הנקודות זהה.

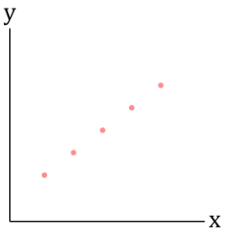
ו. ערך ה-$\tilde{y}$ של כל הנקודות זהה.

ז. עבור חלק מהתצפיות ערך ה-$\tilde{y}$ המנובא גדול מערך ה- y האמיתי.

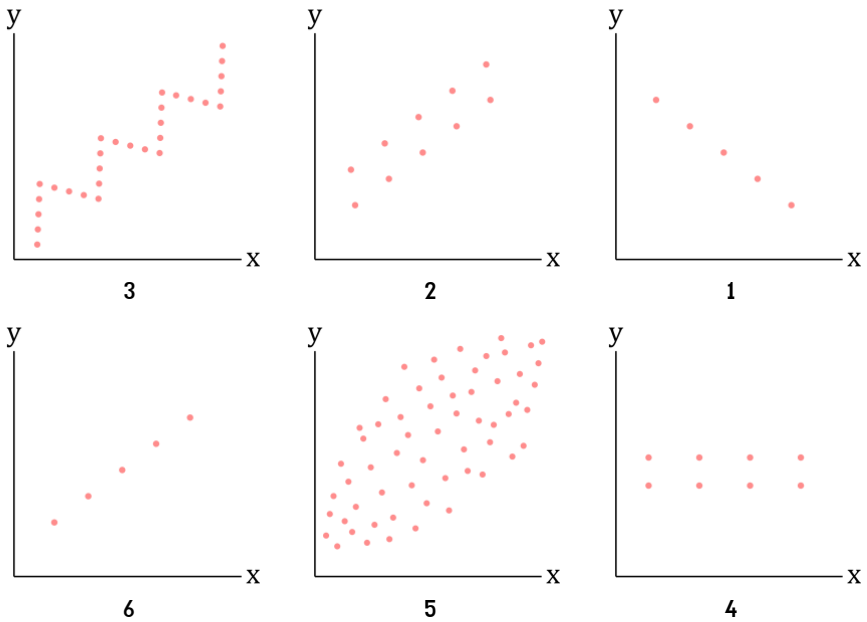
ח. עבור חלק מהתצפיות ערך ה-$\tilde{y}$ המנובא קטן מערך ה- y האמיתי.

ט. עבור חלק מהתצפיות ערך ה-$\tilde{y}$ המנובא שווה לערך ה- y האמיתי.

6. לפניכם דיאגרמה שנתונות בה תצפיות מנובאות בלבד.

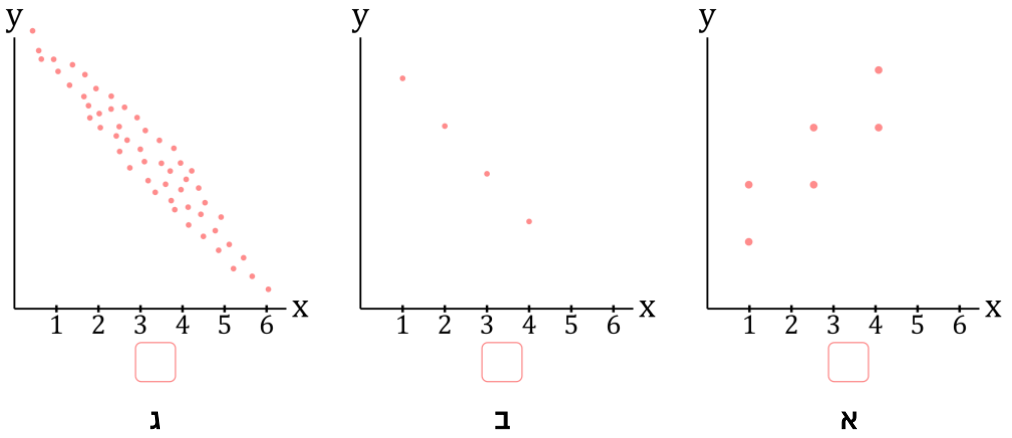


א. מבין הגרפים הבאים סמנו ☑ על כל הגרפים שיכולים להיות הגרפים של התצפיות האמיתיות.



ב. האם ניתן לקבוע בוודאות מי מהגרפים המתוארים הוא דיאגרמת הפיזור האמיתית?

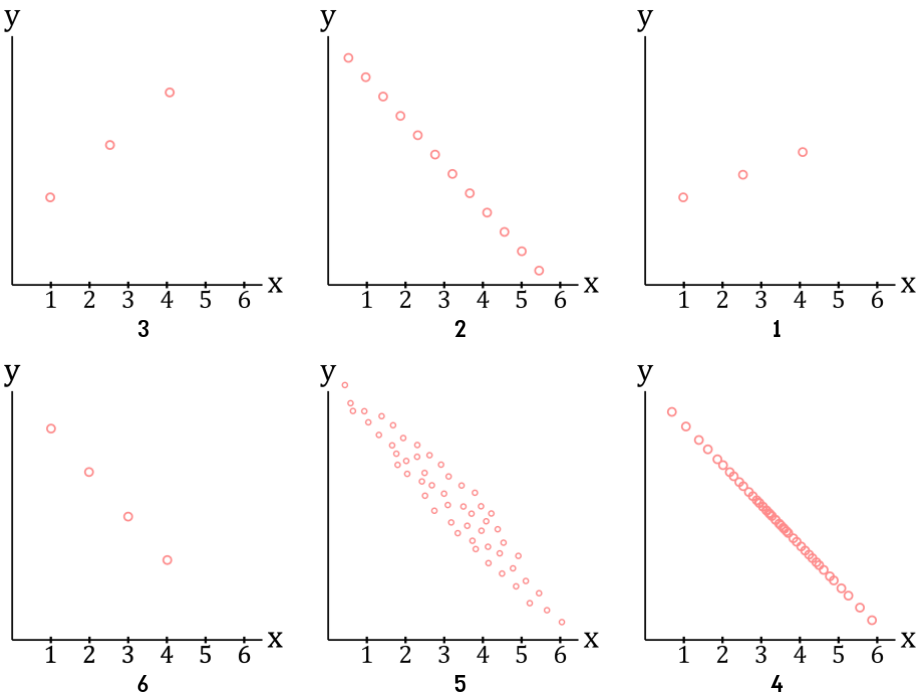
7. שאלת אתגר: לפניכם 3 דיאגרמת פיזור שונות המתארות תצפיות אמיתיות.



לכל אחת מהנקודות בגרפים שלמעלה חושבה תצפית מנובאת אחת המתאימה לה בהתאם לקו הרגרסיה, התצפיות המנובאות שהתקבלו הוצגו בגרף נפרד.

לפניכם 6 גרפים המציגים תצפיות מנובאות. מצאו ביניהם את שלושת הגרפים בהם סומנו רק התצפיות המנובאות המתאימות לתצפיות האמיתיות שבגרפים למעלה.

כתבו תחת כל אחת מהדיאגרמות א-ג את מספר דיאגרמת התצפיות המנובאות המתאימה לה.



יחידה שישית

טרנספורמציות על קו הרגרסיה

הכרנו טרנספורמציות לינאריות בלימודי הסטטיסטיקה התיאורית בכיתה י', ובהתפלגות הנורמלית בכיתה יא', וכעת נבחן כיצד טרנספורמציות משפיעות על דיאגרמת הפיזור וקו הרגרסיה.

נזכיר: טרנספורמציה לינארית היא פעולה מתמטית (חיבור, חיסור, כפל וחילוק) שמתרחשת על כל ערכי המשתנה (למשל, כפל פי 2 של כל הערכים).

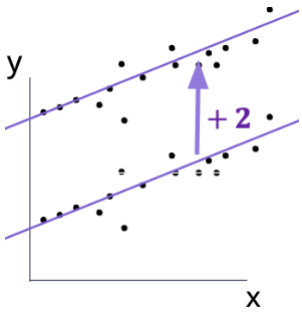
לפני שנתחיל שתי הערות חשובות:

1. בתוכנית הלימוד של 4 יח"ל נתמקד אך ורק בטרנספורמציות על משתנה y. טרנספורמציות על משתנה x הן מחוץ לתחום. (בעולם האמיתי טרנספורמציות מתרחשות גם על משתנה x והחוקים במקרה כזה זהים למה שנלמד כאן).
2. בטרנספורמציות לינאריות מקדם המתאם לא משתנה.

הוספה והפחתה

נבדק הקשר בין כמות האוכל שחתולים אוכלים לבין משקלם. אולם בדיעבד, התברר, שהמשקל שבו נשקלו החתולים לא היה מכויל והפחית 2 ק"ג ממשקלו של כל חתול. לכן כדי לבטא את משקלם האמיתי של החתולים עלינו להוסיף 2 ק"ג למשקלו של כל חתול. כלומר יש להוסיף 2 לכל ערכי ה-y.

אם נוסיף לכל אחד מערכי y 2 ק"ג, דיאגרמת התצפיות כולה תזוז ב-2 כלפי מעלה. התוספת תשפיע על קו הרגרסיה שגם הוא יזוז ב-2 כלפי מעלה, ויהיה כעת:



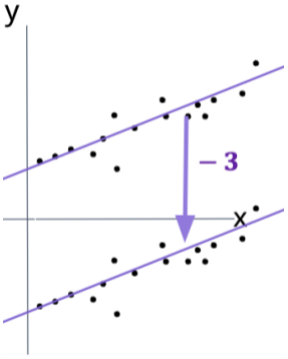
$$\tilde{y} = mx + b + 2$$

בנוסף לדיאגרמת הפיזור ולקו הרגרסיה, נוכל לקבוע אילו מדדים ישתנו ואילו לא:

- ממוצע y יגדל ב-2.
- סטיות התקן של משתנה y לא תשתנה, כי טרנספורמציה של הוספה/הפחתה לא תשפיע עליה.
- סטית התקן של משתנה x לא תשתנה כי לא ערכנו שינוי במשתנה x.
- שיפוע קו הרגרסיה לא ישתנה כי ערכי סטיות התקן והמתאם לא השתנו.

כפי שאמרנו בתחילה: מקדם המתאם r לא ישתנה בטרנספורמציה לינארית.

מה היה קורה אם במקום להוסיף 2 היינו נדרשים להפחית מכל הערכים 3? במצב כזה כל ערכי y היו יורדים 3 יחידות למטה, דיאגרמת התצפיות תזוז ב-3 יחידות כלפי מטה. ההפחתה תשפיע על קו הרגרסיה בצורה דומה להוספה, והקו ויהיה כעת:

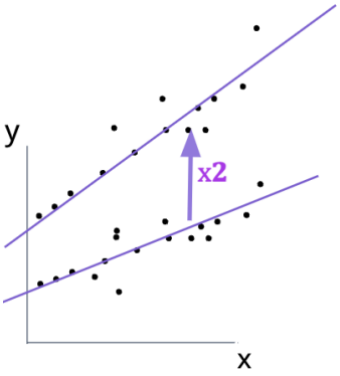


$$\tilde{y} = mx + b - 3$$

כמו בדיאגרמה הקודמת גם כאן סטית התקן של y, ממוצע x, שיפוע קו הרגרסיה ומקדם המתאם – לא ישתנו.

הכפלה וחלוקה

כעת נבדוק מה קורה כאשר אנחנו כופלים את כל התצפיות במספר כלשהו.



למשל, נכפול את כל ערכי התצפיות במשתנה y פי 2. במצב כזה ניתן לראות שכל הנקודות בגרף "עלו" כלפי מעלה. אלא שהפעם העליה שלהם כלפי מעלה אינה אחידה, אלא תלויה בערך y המקורי שלהן: ככל שהנקודות גבוהות יותר בערך ה-y, הן גם "יקפצו" כלפי מעלה הרבה יותר מאשר נקודות "נמוכות" בערך ה-y. לכן נוכל לראות ששיפוע קו הרגרסיה השתנה יחד עם השינוי בפיזור הנקודות.

ההכפלה תשפיע על קו הרגרסיה שגם הוא יזוז פי 2 כלפי מעלה, ויהיה כעת:

$$\tilde{y} = 2 \cdot (mx + b)$$

במקרה של הכפלה חשוב שנבין קצת יותר טוב גם מבחינה אלגברית מדוע קו הרגרסיה הוכפל פי 2. לשם כך נבדוק מה קרה לממוצע ולסטית התקן ב-y.

לאור העובדה שכל הערכים הוכפלו, נוכל לקבוע כי גם הממוצע יגדל פי 2 וגם סטית התקן תגדל פי 2.

כלומר הממוצע החדש יהיה $2 \cdot \bar{y}$ וסטית התקן החדשה תהיה $2 \cdot S_y$.

אם נציב את סטית התקן החדשה בנוסחת שיפוע קו הרגרסיה נקבל:

$$m = r \cdot \frac{2S_y}{S_x}$$

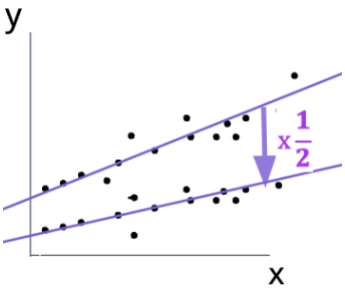
כלומר ניתן לראות ששיפוע הקו הוכפל פי 2. מאחר שגם ממוצע y הוכפל פי 2, כאשר נציב אותו בנוסחה לחישוב קו הרגרסיה, נראה שלמעשה כל הנוסחה הוכפלה פי 2. (ניתן לדעת את זה גם לפי העובדה ש-b שהוא נקודת החיתוך עם ציר y, בהכרח הוכפל פי 2, בדומה לכל נקודה אחרת בגרף. אם גם b וגם השיפוע הוכפלו פי 2 – כל הנוסחה הוכפלה פי 2).

נסכם:

- ממוצע y יגדל פי 2.
- סטית התקן של משתנה y תגדל פי 2 (כי טרנספורמציה של הכפלה תשפיע גם עליה).
- סטית התקן של משתנה x לא תשתנה (כי לא ערכנו שינוי במשתנה x).
- שיפוע הקו יגדל פי 2.

וכמובן: מקדם המתאם r לא ישתנה תחת טרנספורמציה לינארית.

מה היה קורה אם במקום להכפיל פי 2 היינו מחלקים ב-2?



במצב כזה כל ערכי y היו מחולקים ב-2, ויורדים כלפי מטה. גם שיפוע הקו היה משתנה, ולמעשה כל קו הרגרסיה היה מחולק ב-2:

$$\tilde{y} = \frac{1}{2} \cdot (mx + b)$$

כמו בדיאגרמה הקודמת גם כאן סטית התקן של y וממוצע y היו משתנים בהתאם לטרנספורמציה על הערכים, כלומר היו קטנים פי 2.

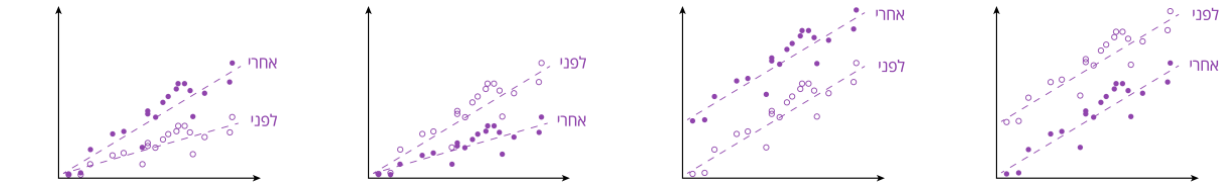
הערה חשובה: עד היום תלמידים לא התבקשו בבגרות לחשב את קווי הרגרסיה החדשים לאחר טרנספורמציה לינארית, אלא רק לקבוע אם חל שינוי בשיפוע הקו ובמקדם המתאם. לפעמים הופיעה בנוסף שאלה על שינוי בממוצע או בסטית תקן כתוצאה מהשינוי בערכים.

במועד חורף 2025, לראשונה, הופיע סעיף בו תלמידים התבקשו לכתוב מהו קו הרגרסיה החדש, בטרנספורמציה של הפחתה. לכן, החוברת כוללת בכל אחת מהטרנספורמציות את חישוב קו הרגרסיה לאחר השינוי.

טרנספורמציות על קו הרגרסיה

1. משרד התחבורה בדק את הקשר בין מספר העיכובים השבועי של הרכבת (X), לבין מספר התלונות השבועי של הנוסעים (Y). לאחר שנמצא קשר לינארי בין המשתנים, התברר כי בכל אחת מהדגימות נשמטו שתי תלונות, ולכן הוכנה דיאגרמת פיזור חדשה.

א. מה מבין הבאים יכול להיות הגרף המתאים לתיאור הנתונים לפני השינוי ואחריו?



1 2 3 4

ב. השלימו את הנתונים בטבלה לאחר השינוי:

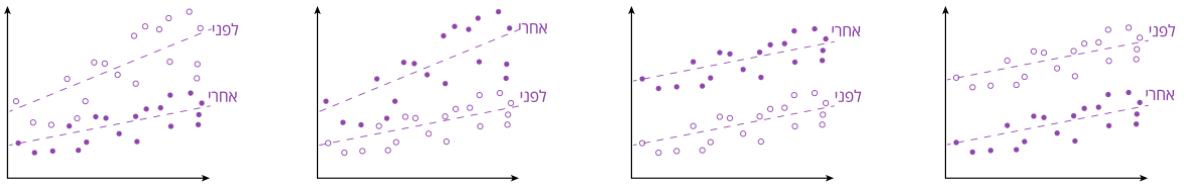
	$\bar{x}$	$\bar{y}$	S_x	S_y	r	m	קו הרגרסיה
לפני השינוי	4	6	2	25			$\tilde{y} = x + 3$
אחרי השינוי							

ג. עבור 6 עיכובים של הרכבת כמה תלונות ינובאו לפני השינוי? $\tilde{y} =$ _____

עבור 6 עיכובים של הרכבת כמה תלונות ינובאו אחרי השינוי? $\tilde{y} =$ _____

2. בבדיקה שערך משרד התיירות על מספר התיירים בערים שונות, נמצא קשר לינארי בין מספר התיירים שמבקרים בעיר (x) לבין מספר המלונות שבה (y). יועץ חיצוני שהועסק על ידי המשרד קבע כי יש לכלול בבדיקה גם את דירות האירוח, וידוע כי בכל עיר מספר דירות האירוח גדול פי 2 ממספר המלונות.

א. מה מבין הבאים יכולים להיות הגרפים המתאימים לתיאור הנתונים לפני השינוי ולאחריו?



1 2 3 4

ב. הקיפו בעיגול את הגדלים שלא יושפעו מהשינוי?

$$\bar{x},\quad \bar{y},\quad S_x,\quad S_y,\quad r,\quad m,\quad \tilde{y}$$

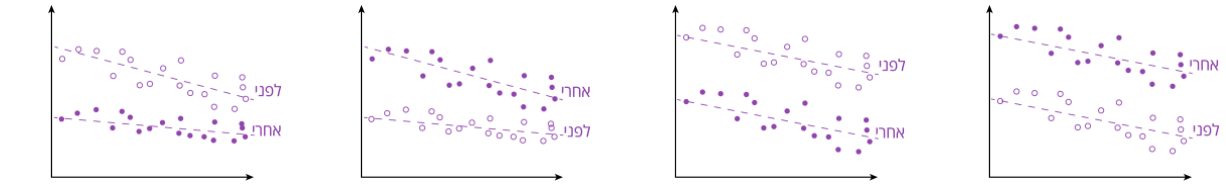
ג. לגבי כל אחד מהגדלים שלא הקפתם בעיגול, תארו במילים מהו השינוי שנגרם.

Classit
info@classit.co - 073-3744093

3. לאור הצתות חוזרות של גזם יבש בשטחים פתוחים ברשויות המקומיות, המשרד לאיכות הסביבה בדק האם איסוף הגזם על ידי הרשויות מפחית את מספר ההצתות. נערך מחקר שבדק את הקשר בין כמות הגזם שנאסף (המשתנה X, נמדד במאות טונות) לבין מספר ההצתות (המשתנה Y).

בבדיקה חוזרת של הנתונים התגלה כי $\frac{1}{4}$ מההצתות למעשה לא היו הצתות, אלא שריפות שפרצו כתוצאה ממזג האויר. הנתונים עודכנו.

א. מה מבין הבאים יכול להיות הגרף המתאים לפני השינוי ולאחריו?



1 2 3 4

ב. השלימו את הנתונים בטבלה לאחר השינוי:

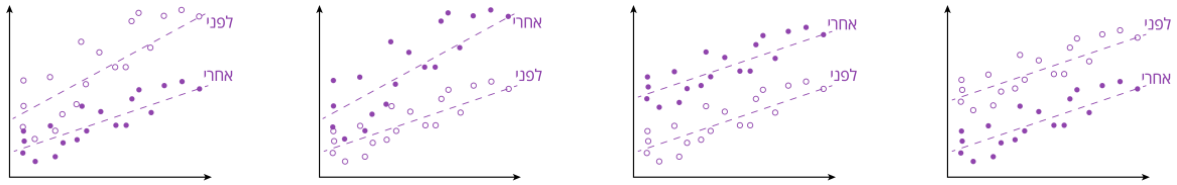
	$\bar{x}$	$\bar{y}$	S_x	S_y	r	m	קו הרגרסיה
לפני השינוי	2540	5740	160	230			$\tilde{y} = -0.46x + 6908$
אחרי השינוי							

ג. לנקודה $x = 1000$ ינובא לפני השינוי הערך $\tilde{y} =$ _____

לנקודה $x = 1000$ ינובא אחרי השינוי הערך $\tilde{y} =$ _____

4. משרד החקלאות מצא קשר לינארי בין מספר עצי הפרי ביישוב (x) לבין כמות הדבש שהופקה מהכוורות באזור (y). לאחר בדיקת איכות הדבש הוחלט לגרוע מכמות הדבש 2.5 ק"ג מכל כוורת.

א. מה מבין הבאים יכול להיות הגרף המתאים לפני השינוי ולאחריו?



1 2 3 4

ב. השלימו את הנתונים בטבלה לאחר השינוי:

	$\bar{x}$	$\bar{y}$	S_x	S_y	r	m	
לפני השינוי	120	50	43	15			$\tilde{y} = 0.3x + 14$
אחרי השינוי							

ג. ליישוב שבו נטועים 150 עצי פרי תנובא לפני השינוי כמות דבש $\tilde{y} =$ _____

ליישוב שבו נטועים 150 עצי פרי תנובא אחרי השינוי כמות דבש $\tilde{y} =$ _____

	$r=1$	$0<r<1$	$r=0$	$-1<r<0$	$r=-1$	r לא מוגדר
1.	ה	ו, ז, יב	ד, ט, י	א, ב, יא	אין	ג, ח

2. א, ג, ה, ו, ז, ח, י, יב.

3.

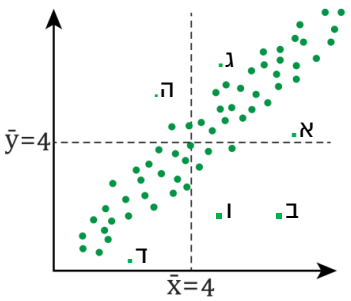
1	$r_1 > r_2$ $\lvert r_1\rvert < \lvert r_2\rvert$	2	$r_1 = r_2$ $\lvert r_1\rvert = \lvert r_2\rvert$	
3	$r_1 > r_2$ $\lvert r_1\rvert > \lvert r_2\rvert$	4	$r_1 > r_2$ $\lvert r_1\rvert < \lvert r_2\rvert$	
5	$r_1 < r_2$ $\lvert r_1\rvert = \lvert r_2\rvert$	6	$r_1 > r_2$ $\lvert r_1\rvert < \lvert r_2\rvert$	
7	$r_1 > r_2$ $\lvert r_1\rvert = \lvert r_2\rvert$	8	$r_1 = r_2$ $\lvert r_1\rvert = \lvert r_2\rvert$	
9	$r_1 < r_2$ $\lvert r_1\rvert < \lvert r_2\rvert$	10	$r_1 > r_2$ $\lvert r_1\rvert < \lvert r_2\rvert$	

מקדם המתאם הלינארי

1. א. נקודה A גדול, גדול, ערך חיובי.
   נקודה B קטן, גדול, ערך שלילי.
   ב. נקודה C שלילי
   נקודה D חיובי
   נקודה E חיובי

2. א. חיובי: E,B,A שלילי: D אפס: C
   ב. A=132, B=27, C=0, D=-10, E=90
   ג. A

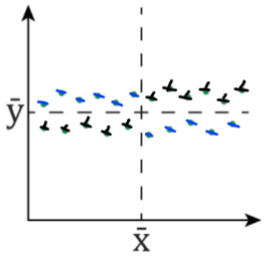
3.



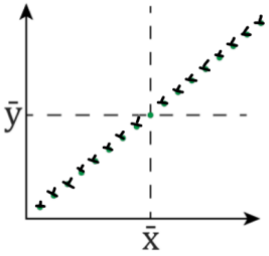
4. א. מחזקת: B, C, D מחלישה: A
   ב. A לפני: חיובי אחרי: שלילי
   B לפני: חיובי אחרי: חיובי
   C לפני: שלילי אחרי: חיובי
   D לפני: חיובי אחרי: חיובי

Classit
info@classit.co - 073-3744093

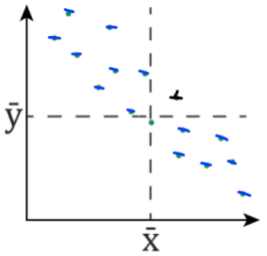
5.



הקשר שואף לאפס, אפשר לראות שהתרומה החיובית והשלילית דומה.

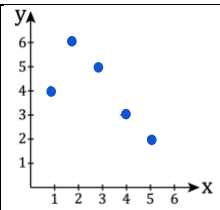


כאשר הקשר חיובי מושלם – אין נקודות שתורמות ערך שלילי לנוסחת מקדם המתאם



כאשר יש קשר שלילי – יש יותר תרומות שליליות.

חישוב מקדם המתאם

1.	א.	$\bar{x}=6.4$	$\bar{y}=27.6$	ב.	$\bar{x}=14$	$\bar{y}=53$	ג.	$\bar{x}=6450$ $\bar{y}=22$
2.	א.	$x=52$	$y=2.2$	ב.	$x=320$	$y=9$		
3.	א.	$s_x=2.828$	$s_y=5.55$	ב.	$s_x=3.958$	$s_y=9.292$	ג.	$\bar{x}=13.5$ $\bar{y}=3$ $s_x=2.63$ $s_y=0.792$
4.	א.	$r=-1$		ב.	$s_x=5.538$ $\bar{y}=17.7$ $r=-0.564$		ג.	$\bar{x}=6$ $\bar{y}=16.5$ $s_x=2.769$ $s_y=3.862$ $r=0.436$
5.	א.		ב. $r=-0.7$ הקשר שלילי				ג.	חיזק: 1 החלישו: 6
6.	א.	$r_2,\ r_4$	ב. $r=-0.972$	ג. i לא נכונה, הוספת גודל גדול מהממוצע מגדילה אותו. ii נכונה, הוספת מספר רחוק מהממוצעים ולכן הפיזור יגדל. iii נכונה, הנקודה תורמת ערך חיובי, בשונה מכיוון הקשר, עצמת הקשר תקטן.				

הסקה על קשר מטבלאות נתונים

	$r=1$	$0<r<1$	$r=0$	$-1<r<0$	$r=-1$	r לא מוגדר
1.	א, ב, ג, ו,ט, טו	יא, טז	ח, י	יב,יג,	ד, ה	ז, יד,

2. א. $0<r<1$ ב. $r=1$ ג. $-1<r<0$ ד. $r=-1$

3. א. A לא תשנה
B תחליש
C לא תשנה

ב. A לא תשנה
B תחליש
C תחליש

ג. A תחזק
B תחזק
C תחליש

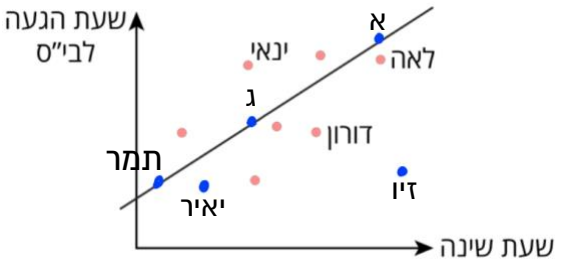
Classit
info@classit.co - 073-3744093

מציאת משוואת ישר רגרסיה

1	א. $m = 3.271$	ב. $m = -0.545$	ג. $m = 1.2456$	ד. $m = -1.854$
2	א. $s_y = 7.674$	ב. $s_x = 1.068$	ג. $r = 0.953$	ד. $s_x = 0.21775$
3	א. $m = 0.392$ $\tilde{y} = 0.392x - 10.14$	ב. $m = -0.464$ $\tilde{y} = -0.464x + 94.13$	ג. $m = 0.396$ $\tilde{y} = 0.396x + 340.6$	ד. $m = -2.412$ $\tilde{y} = -2.412x + 32.70$
4.	א. $m = 0.45$ $r = 0.9$ $\bar{x} = 30$	ב. $m - 1.44$ $s_y = 12$ $\bar{y} = 80$	ג. $m = 1.2$ $r = 1$ $\bar{y} = 250$	ד. $m = -0.3$ $s_x = 4.33$ $\bar{x} = 5$
5.	א. 154	ב. לא ניתן לדעת	ג. $\bar{x} = 31$	
6.	א. $\bar{y} = 8$	ב. $s_y = 3.6$	ג. $\tilde{y} = 9.14$	ד. לא ה. לא

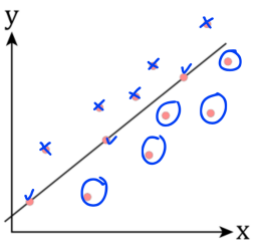
תצפיות אמיתיות ותצפיות מנובאות

1.



ב. מאוחרת ד. מוקדמת ז. כן, יתכנו שני המצבים ח. נכון.

2.



א. 5 ב. 5 ג. 3

3. נכונות: א, ו לא נכונות: ב, ה, ז לא ניתן לדעת: ג, ד

4. א. 7 ב. 6 ג. 4 ד. 2,1 ה. כן ו. כן.

5. א. B,F ב. A,E ג. D,C.

ד. נכון ה. לא נכון ו. נכון. ז. נכון ח. נכון ט. לא נכון.

6. א. גרפים: 2,3,5,6 ב. לא ניתן לקבוע.

7. א. 3 ב. 6 ג. 4.

1. **א.** גרף 2

ב-ג	$\bar{x}$	$\bar{y}$	S_x	S_y	r	m	קו הרגרסיה	$x=6$
לפני	4	6	2	25	0.08	1	$\tilde{y}=x+3$	$\bar{y}=9$
אחרי	4	8	2	25	0.08	1	$\tilde{y}=x+5$	$\bar{y}=11$

2. **א.** גרף 3

ב. $\bar{x},\ S_x,\ r$ - לא ישתנו

ג. $\bar{y},\ S_y$ יגדלו פי 3. $\tilde{y},\ m$

3. **א.** גרף 4

ב-ג	$\bar{x}$	$\bar{y}$	S_x	S_y	r	m	קו הרגרסיה	$x=1000$
לפני	2540	5740	160	230	−0.32	−0.46	$\tilde{y}=-0.46x+6908$	6448
אחרי	2540	4305	160	172.5	−0.32	−0.345	$\tilde{y}=-0.345x+5181$	4836

4. **א.** גרף 1

ב-ג	$\bar{x}$	$\bar{y}$	S_x	S_y	r	m	קו הרגרסיה	$x=150$
לפני	120	50	43	15	0.86	0.3	$\tilde{y}=0.3x+14$	59
אחרי	120	47.5	43	15	0.86	0.3	$\tilde{y}=0.3x+11.5$	56.5