כלים כמותיים מתקדמים של תכן סטטיסטי לאיכות

תוכן העניינים

1. מקדם המתאם (מדד קשר) הלינארי ומובהקותו 1
2. רגרסיה פשוטה 24
3. רגרסיה מרובה 36
4. רגרסיה - שאלות ממבחנים 41

כלים כמותיים מתקדמים של תכן סטטיסטי לאיכות

פרק 1 - מקדם המתאם (מדד קשר) הליניארי ומובהקותו

תוכן העניינים

1. מקדם המתאם הלנארי (פירסון) 1

2. חישוב מקדם המתאם הלנארי (פירסון) 12

3. בדיקת השערות על מקדם המתאם הלנארי 17

4. בדיקת השערות על מקדם המתאם הלנארי באמצעות טבלה של ערכים קריטיים 21

מקדם המתאם (מדד קשר) הלינארי ומובהקותו

מדד הקשר הלינארי (פירסון) – מבוא

מעוניינים לבדוק עד כמה קיים קשר מסוג קשר לינארי (קו ישר) בין שני משתנים.
שני המשתנים שאנו בודקים לגביהם קשר צריכים להיות משתנים כמותיים.
מבחינת סולמות מדידה כל משתנה נחקר צריך להיות מסולם רווחים או מנה.
בדרך כלל המשתנה המוצג כ- Y הוא המשתנה התלוי והמשתנה המוצג ב- X הוא המשתנה הבלתי תלוי.
תיאור גרפי לנתונים נעשה על ידי דיאגרמת פיזור. בדיאגרמת פיזור אנחנו מסמנים כל תצפית בנקודה לפי שיעור ה- X ושיעור ה- Y שלה.
דיאגרמת הפיזור נותנת אינדיקציה גרפית על הקשר בין שני המשתנים.

דוגמה (פתרון בהקלטה) :

בבניין 8 דירות בדקו לכל דירה את מספר החדרים שלה וכמו כן את מספר הנפשות הגרות בדירה. להלן התוצאות שהתקבלו :

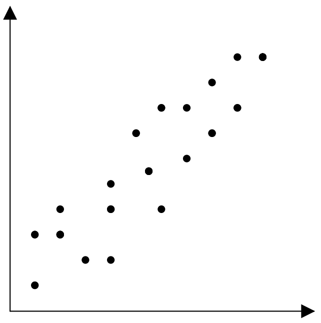
מספר חדרים בדירה	2	2	3	2	3	3	4	4
מספר הנפשות בדירה	0	1	2	2	3	4	4	5

1) כמה תצפיות ישנן בדוגמה?
2) כמה משתנים ישנם בדוגמה, מי הם?
3) שרטטו לנתונים דיאגרמת פיזור.
4) מי המשתנה התלוי ומיהו המשתנה הבלתי תלוי?

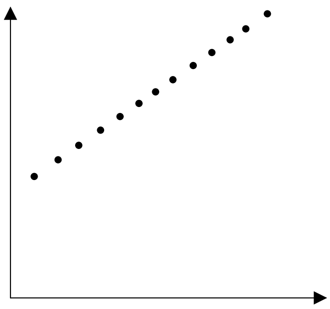
© כל הזכויות שמורות לברק קנדל

דיאגרמות פיזור לקשר בין משתנים וניתוחם

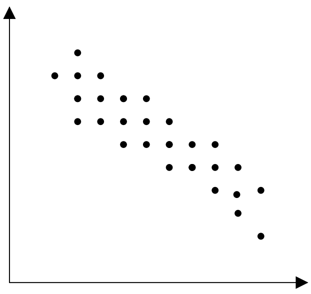
קשר לנארי חיובי חלקי



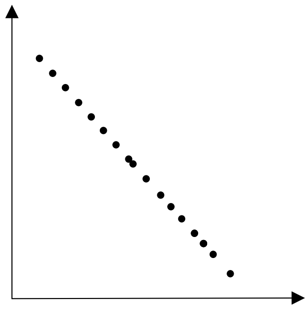
קשר לנארי חיובי מלא



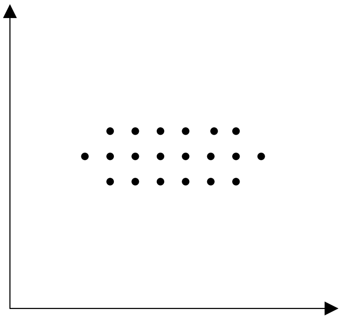
קשר לינארי שלילי חלקי



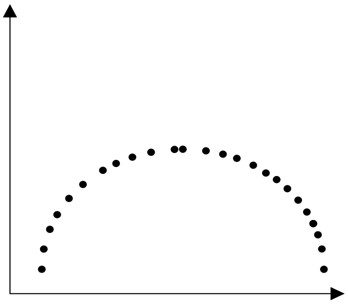
קשר לינארי שלילי מלא



אין קשר



אין קשר לינארי



משמעות מקדם המתאם:

כדי לבדוק עד כמה קיים קשר לינארי בין שני המשתנים ישנו מדד קשר שנקרא גם מקדם המתאם הלינארי הידוע גם בשם מקדם המתאם של פירסון.

מקדם מתאם זה מקבל ערכים בין 1 ל-1-.

-1   0   1

© כל הזכויות שמורות לברק קנדל

מקדם מתאם 1-או 1 אומר שקיים קשר לינארי מלא בין המשתנים שניתן לבטאו על ידי נוסחה של קו ישר: $y = ax + b$.

מתאם חיובי מלא (מקדם מתאם 1):

קיים קשר לנארי מלא בו השיפוע a יהיה חיובי ואילו מתאם שלילי (מקדם מתאם-1) מלא אומר שקיים קשר לנארי מלא בו השיפוע a שלילי.

מתאם חיובי חלקי:

ככל שמשתנה אחד עולה לשני יש נטייה לעלות בערכו אבל לא קיימת נוסחה לינארית שמקשרת את X ל- Y באופן מוחלט ואילו מתאם שלילי חלקי אומר שככל שמשתנה אחד עולה לשני יש נטייה לרדת אבל לא קיימת נוסחה לינארית שמקשרת את X ל- Y באופן מוחלט. ככל שמקדם המתאם קרוב לאפס עוצמת הקשר יותר חלשה וככל שהמדד רחוק יותר מהאפס העוצמה יותר חזקה. לסיכום, מקדם המתאם בודק את עוצמת הקשר הלנארי, ואת כיוון הקשר.

מקדם המתאם הלינארי אינו מושפע מיחידות המדידה. כל שינוי ביחידות המדידה של המשתנים, לא ישנה את מקדם המתאם.

מדד הקשר הלינארי באוכלוסייה, שנקרא גם מקדם המתאם של פירסון או מדד הקשר של פירסון באוכלוסייה מסומן ב: ρ - פרמטר המאפיין את עוצמת הקשר הלינארי באוכלוסייה וכיוונו בין שני המשתנים הנחקרים. כאשר:

r - מדד הקשר הלינארי במדגם שמהווה אומד לפרמטר ρ.

קיומו של מתאם בין שני משתנים אינו מצביע על סיבתיות בהכרח. למשל, אם נמצא מתאם חיובי בין כמות הסוכרזית שאדם אוכל לבין במשקל שלו אין זה אומר שהסיבה להשמנה היא הסוכרזית. מדד הקשר של פירסון הוא מדד קשר סימטרי, כלומר אם נחליף את X ב- Y התוצאה תהיה זהה.

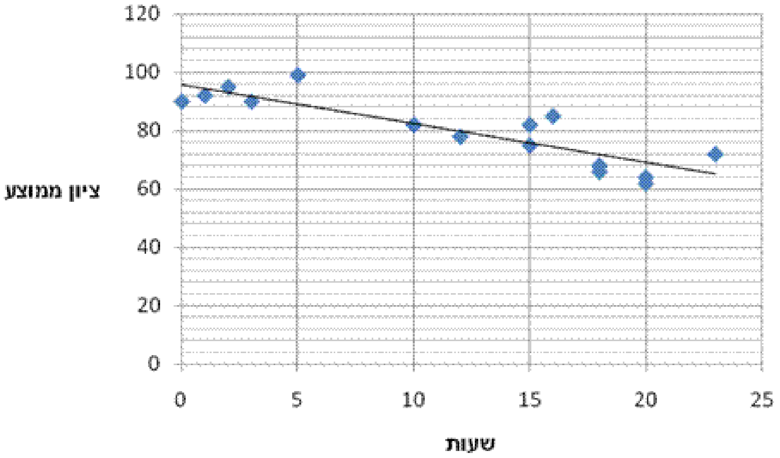
דוגמה (פתרון בהקלטה):

- מה ניתן להגיד על מקדם המתאם של שני המשתנים על סמך דיאגרמת הפיזור ששרטטנו?
- אם היינו משנים את השרטוט כך שבציר האנכי היה המשתנה "מספר החדרים" ובציר האופקי היה "מספר הנפשות", האם הדבר היה משפיע על מדד הקשר של פירסון?

© כל הזכויות שמורות לברק קנדל

שאלות

1) חוקר רצה לאפיין את הקשר בין מספר השעות בשבוע שסטודנט מקדיש לבילויים לבין הציון הממוצע שלו בסוף הסמסטר. לשם כך הוא אסף נתונים של 15 סטודנטים ויצר דיאגרמת פיזור:



א. מיהו המשתנה הבלתי תלוי?

ב. מה ניתן לומר על כיוון הקשר בין מספר שעות הבילוי השבועיות לבין הציון הממוצע של הסמסטר? מה ניתן להגיד על עוצמת הקשר?

2) להלן טבלה המסכמת את מקדמי המתאם הלינארי בין ציוני מבחנים שונים שהתקבלו עבור תלמידים בכיתה מסוימת:

	ספורט	לשון	מתמטיקה
ספורט	?	-0.7	?
לשון	?	?	0.6
מתמטיקה	-0.1	?	?

א. השלימו את מקדמי המתאם שמסומנים בסימן שאלה בטבלה.

ב. בין אילו שני ציוני מקצועות שונים קיים מתאם בעל העוצמה החזקה ביותר?

3) במחקר נתבקשו לבדוק את הקשר בין מספר שעות התרגול של קורס לבין הציון הסופי שלו. להלן תוצאות מדגם שהתקבל:

שעות תרגול	ציון סופי
20	90
25	90
30	95
15	60
30	90
20	85
10	50

א. מיהו המשתנה התלוי ומיהו המשתנה הבלתי תלוי בדוגמה זו?

ב. שרטטו דיאגרמת פיזור לנתונים.

ג. מה ניתן לומר על הקשר בין המשתנים במדגם?

ד. מסתבר שבסופו של דבר נתנו פקטור של 5 נקודות לציון הסופי. כיצד הדבר היה משנה את מקדם המתאם של המדגם?

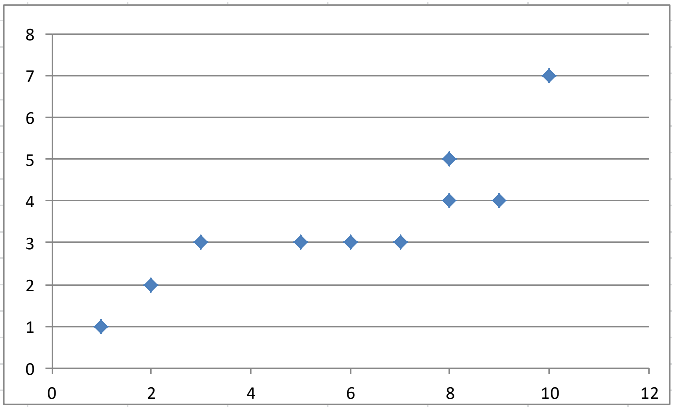
© כל הזכויות שמורות לברק קנדל

4) בתחנה המטאורולוגית רצו לבדוק את הקשר שבין הטמפרטורה במעלות צלזיוס לכמות המשקעים במ"מ. הם אספו נתונים על 10 ימים במהלך חודש ינואר. המתאם שהתקבל היה 0.8.

א. השלימו את המשפט :
בחודש ינואר ככל שהטמפרטורה היומית נוטה לרדת, כך כמות המשקעים נוטה ________.

ב. הוחלט להעביר את הטמפרטורה למעלות פרנהייט על מנת שיוכלו להשוות אותה לנתונים מארה"ב. נוסחת המעבר היא $F^0=32+\frac{9}{5}C^0$.
כיצד הדבר ישפיע על מקדם המתאם בין הטמפרטורה במעלות פרנייט לכמות המשקעים במ"מ?

5) להלן דיאגרמת פיזור המראה קשר בין שני משנים :



א. השלימו : ניתן לראות שהקשר הוא ליניארי _________ (מלא\ חלקי) כיוון הקשר הוא (חיובי \שלילי).

ב. השלימו : אם היינו מוסיפים תצפית שערך ה- X שלה הוא 4 וערך ה- Y שלה הוא 7, מקדם המתאם של פירסון היה _______ (גדלו\ קטן\ לא משתנה).

שאלות רב ברירה (יש לבחור את התשובה הנכונה):

6) חוקר אקלים דגם כמה ימים בשנה ומדד את הטמפרטורה בטורונטו שבקנדה ואת הטמפרטורה בסידני שבאוסטרליה באותו היום. הוא חישב ומצא מקדם מתאם שלילי בין הטמפרטורה היומית בטורונטו לבין הטמפרטורה היומית בסידני. משמעות מקדם המתאם השלילי במדגם :

א. אין קשר בין הטמפרטורה בטורונטו לבין הטמפרטורה בסידני בימים שנדגמו.

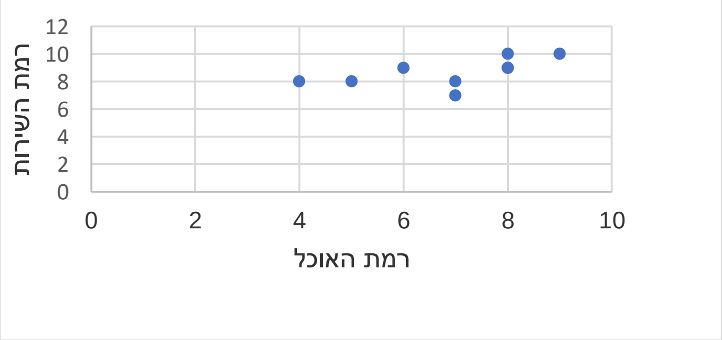
ב. במדגם, רוב הטמפרטורות בטורונטו היו שליליות.

ג. ההפרש בין הטמפרטורה בטורונטו לבין הטמפרטורה באוסטרליה, במדגם זה, הוא שלילי.

ד. במדגם יש נטייה שהטמפרטורה יורדת בטורונטו לטמפרטורה לעלות בסידני.

© כל הזכויות שמורות לברק קנדל

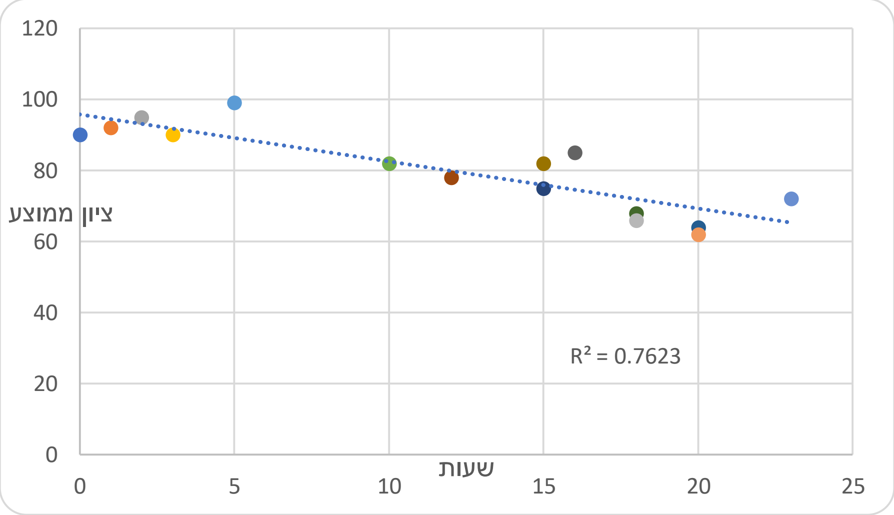
7) בסקר שביעות רצון שנערך בבית הקפה "פת לחם" התבקשו הלקוחות לדרג את מידת שביעות הרצון שלהם (בסולם 1-10) בשני נושאים: רמת האוכל ורמת השירות.



מה יהיה ערכו של מקדם המתאם (r)?

א. $r=-0.3$
ב. $r=0$
ג. $r=1.125$
ד. $r=0.593$

8) חוקר רצה לאפיין את הקשר בין מספר השעות בשבוע שסטודנט מקדיש לבילויים לבין הציון הממוצע שלו בסוף הסמסטר. לשם כך הוא אסף נתונים של 15 סטודנטים ויצר דיאגרמת פיזור.

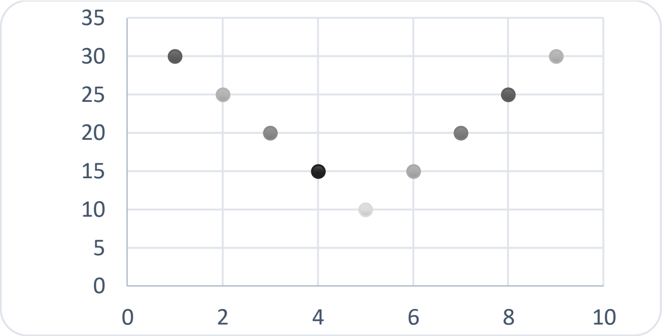


מה ניתן לומר על כיוון הקשר במדגם בין מספר שעות הבילוי השבועיות לבין הציון הממוצע של הסמסטר?

א. ככל שמבלים יותר הציון נוטה לרדת.
ב. אין קשר בין שעות הבילוי לציון.
ג. ככל שמבלים פחות הציון נוטה לרדת.
ד. ככל שהציון נוטה לרדת הסטודנט מבלה פחות.

© כל הזכויות שמורות לברק קנדל

9) התרשים הבא מתאר קשר בין שני משתנים, איזה מהמתאמים הבאים הוא המתאים ביותר לתיאור הקשר בין שני המשתנים?



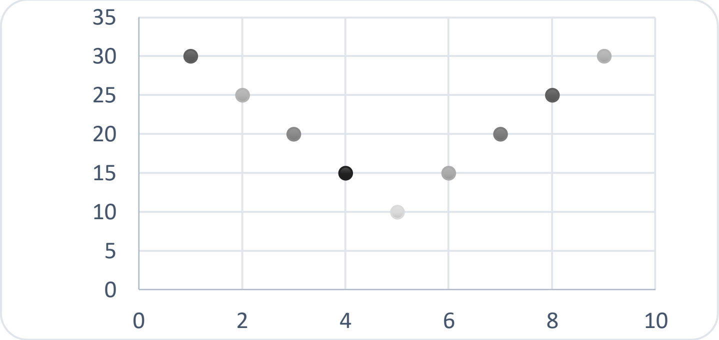
א. $r = 1$ היות ושני המשתנים יוצרים קוים ישרים.

ב. $r = 2$ היות ויש שני קוים בעלי קשר מושלם.

ג. $r = 0$ היות והקו יורד ואחר כך עולה באותו האופן.

ד. $r = \pm1$ היות ויש קו עולה וגם קו יורד.

10) התרשים הבא מתאר דיאגרמת פיזור.



איזו טענה נכונה?

א. בתרשים מוצג הקשר בין שני משתנים.

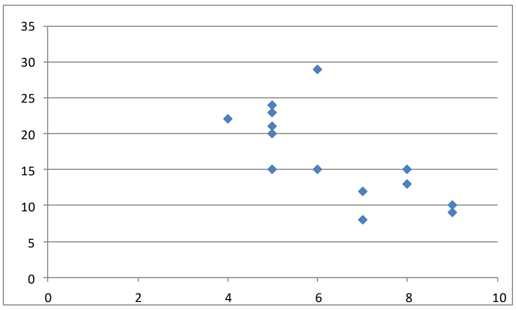
ב. בתרשים מוצג הקשר בין 9 משתנים.

ג. בתרשים מוצג הקשר בין 10 משתנים.

ד. אין לדעת כמה משתנים מוצגים בתרשים.

© כל הזכויות שמורות לברק קנדל

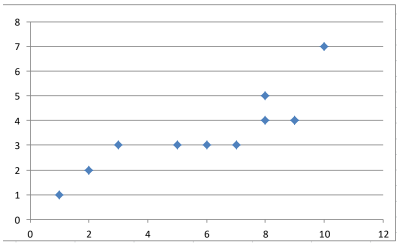
בגרף הבא מתוארת דיאגרמת פיזור של שני משתנים :
X - (משתנה בלתי תלוי בציר האופקי)
ו- Y (משתנה תלוי).
במדגם התקבל $r^2 = 0.52$.



11) לאור הנתונים המופיעים בדיאגרמה, איזה מבין הערכים הבאים מתאים להיות התוצאה של r ?
א. -0.52
ב. 0.72
ג. -0.72
ד. 0.52

12) אם מקדם המתאם בין שני משתנים הוא 1, אזי :
א. הערכים של המשתנים הם חיוביים.
ב. עבור כל תצפית ערך של משתנה אחד שווה לערך של המשתנה השני.
ג. הקשר הליניארי הוא בעוצמה חזקה.
ד. אף אחת מהתשובות לא בהכרח נכונה.

13) להלן דיאגרמת פיזור :
מה יהיה מקדם המתאם בין שני המשתנים?



א. 1
ב. 0.85
ג. 0.15
ד. 0

14) בבדיקת קשר בין שני משתנים התקבל : $r = -1$.
א. קיימת נוסחה לינארית הקושרת בין כל התצפיות.
ב. לא קיים קשר בין שני המשתנים.
ג. ככל שמשתנה אחד נוטה לרדת גם לשני יש נטייה לרדת.
ד. קיים קשר בין שני המשתנים, אך לא ניתן לדעת מאיזה סוג.

www.gool.co.il © כל הזכויות שמורות לברק קנדל

15) לפי הפתגם ״רחוק מהעין, רחוק מהלב״, יש קשר ____ בין קרבה פיזית לקרבה נפשית.

א. חיובי
ב. שלילי
ג. אפסי
ד. לא ניתן לדעת.

16) מבחן אמי״ר הינו מבחן מיון באנגלית של המרכז הארצי לבחינות והערכה. הציון המינימלי בבחינה הינו 150 והמקסימלי הינו 250. בקורס הכנה למבחן השתתפו 19 תלמידים. להלן הציונים שלהם על פי פלט שהתקבל:

		159
		170
		180
		185
		204
		224
		236
		212
		168
		189
		195
		163
		187
		206
		201
		223
		242
		203
		205
197.47	AVERAGE	
536.25	VARPA	

יש להוסיף עמודה נוספת לצד עמודת הציונים שתראה לכל תלמיד כמה נקודות חסרות לו כדי להשלים לציון המקסימלי בבחינה.

מה יהיה מקדם המתאם בין שתי העמודות (כלומר, מקדם המתאם בין הציון לבין הנקודות החסרות)?

א. -1
ב. 1
ג. -0.5
ד. 0.5

17) מקדם המתאם בין שטחי דירה למחיר שלהם חושב ונמצא 1.2. מה נובע מכך?

א. ככל שהדירה גדולה יותר בשטחה כך היא יקרה יותר.
ב. ככל שהדירה קטנה יותר בשטחה כך היא זולה יותר.
ג. לא קיים קשר בין שטח הדירה למחיר הדירה.
ד. מצב כזה שמתואר הנתונים לא אפשרי.

18) אם ניקח 10 אנשים ונרשום לכל אדם את הגובה במטר וכמו כן את הגובה בס״מ. מה יהיה מקדם המתאם בין גובה האדם במטר לגובה האדם בס״מ?

א. 1
ב. 0
ג. -1
ד. לא ניתן לדעת.

www.gool.co.il © כל הזכויות שמורות לברק קנדל

19) נמצא מתאם חיובי בעוצמה גבוהה בין X – ציון בבגרות בלשון ל Y – ציון בבגרות במתמטיקה. אילו מהמשפטים הבאים נכון?

א. ניתן לומר שאחת מהסיבות להבדלים שיש לסטודנטים במתמטיקה נובעים מההבדלים שיש להם בלשון.

ב. קיימת נוסחה של קו ישר שקושרת בין ציון בבגרות במתמטיקה לציון בבגרות בלשון.

ג. ללא יוצא מן הכלל, ניתן להגיד שכל תלמיד שמצליח יותר מתלמיד אחר בלשון גם יצליח יותר מאותו תלמיד במתמטיקה.

ד. אף אחד מהטענות שהוצגו אינה בהכרח נכונה.

20) עבור סדרה של תצפיות מדדו את X ואת Y. נמצא שעבור כל התצפיות שהערך של Y ירד הערך של X בהכרח ירד ללא יוצא מן הכלל. מקדם המתאם של פירסון יהיה בהכרח:

א. 1

ב. -1

ג. 0

ד. אף אחת מהתשובות.

© כל הזכויות שמורות לברק קנדל
www.gool.co.il

תשובות סופיות

1) א. שעות בילוי.
ב. הקשר חלקי, כיוון הקשר שלילי .

2) א. להלן טבלה :
ב. ספורט ולשון.

	ספורט	לשון	מתמטיקה
ספורט	1	-0.7	0.1
לשון	-0.7	1	0.6
מתמטיקה	-0.1	0.6	1

3) א. ב"ת- מסי שעות התרגול, תלוי- ציון.
ב. ראה גרף בפתרון וידאו.
ג. קשר לינארי חיובי חלקי.
ד. מקדם המתאם לא היה משתנה.

4) א. לעלות.
ב. לא ישפיע על מקדם המתאם.

5) א. חלקי, חיובי.
ב. קטן.

6) ד'. 7) ד'. 8) א'. 9) ג'. 10) א'.

11) ג'. 12) ד'. 13) ב'. 14) א'. 15) א'.

16) א'. 17) ד'. 18) א'. 19) ד'. 20) ד'.

© כל הזכויות שמורות לברק קנדל

מדדי קשר – מדד הקשר הלינארי (פירסון) – רקע

המטרה היא לבדוק האם קיים קשר (קורלציה, מתאם) של קו ישר בין שני משתנים כמותיים. מבחינת סולמות המדידה קשר בין סולמות רווחים ומנה.
בדרך כלל, X הוא המשתנה המסביר (הבלתי תלוי) ו- Y הוא המשתנה המוסבר (התלוי).

דוגמה:

נרצה להסביר כיצד השכלה של אדם הנמדדת בשנות לימוד – X מסבירה את ההכנסה שלו Y. במקרה זה שנות ההשכלה זהו המשתנה המסביר (או הבלתי תלוי) ואנחנו מעוניינים לבדוק כיצד שינויים בשנות ההשכלה של אדם יכולים להסביר את השינויים שלו בהכנסה, ולכן רמת ההכנסה זהו המשתנה המוסבר התלוי במשתנה המסביר אותו.

שלב ראשון: נהוג לשרטט דיאגרמת פיזור. זו דיאגרמה שנותנת אינדיקציה ויזואלית על טיב הקשר בין שני המשתנים.

דוגמה:

בבניין של 5 דירות בדקו את הנתונים הבאים:
X -מס' חדרים בדירה. Y-מס' נפשות הגרות בדירה.
להלן התוצאות שהתקבלו:

Y	X	מס' דירה
2	3	1
2	2	2
3	4	3
3	3	4
4	5	5

נשרטט מנתונים אלה דיאגרמת פיזור (הדיאגרמה המלאה בסרטון). נתבונן בכמה מקרים של דיאגרמות פיזור וננתח אותן (הדיאגרמות המלאות בסרטון).

שלב שני: מחשבים את מקדם המתאם (מדד הקשר) שבודק עד כמה קיים קשר לינארי בין שני המשתנים. המדד (ניקרא גם מדד הקשר של פירסון) מכמת את מה שניראה בשלב הראשון רק בעין.
המדד בודק את כיוון הקשר (חיובי או שלילי) ואת עוצמת הקשר (חלש עד חזק).
מקדם מתאם זה מקבל ערכים בין 1- ל-1.
מקדם מתאם 1- או 1 אומר שקיים קשר לינארי מוחלט ומלא בין המשתנים שניתן לבטאו על ידי הנוסחה: $y = bx + a$.

מתאם חיובי מלא (מקדם מתאם 1):

קיים קשר לנארי מלא בו השיפוע b יהיה חיובי ואילו מתאם שלילי מלא אומר שקיים קשר לנארי מלא בו השיפוע b שלילי (מקדם מתאם 1-).

 © כל הזכויות שמורות לברק קנדל

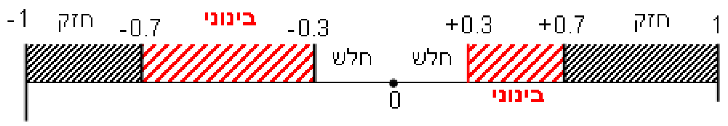
מתאם חיובי חלקי:

ככל שמשתנה אחד עולה לשני יש נטייה לעלות בערכו אבל לא קיימת נוסחה ליניארית שמקשרת את X ל- Y באופן מוחלט.

מתאם שלילי חלקי:

ככל שמשתנה אחד עולה לשני יש נטייה לרדת אבל לא קיימת נוסחה ליניארית שמקשרת את X ל- Y באופן מוחלט.

ככל שערך מקדם המתאם קרוב לאפס נאמר שעוצמת הקשר חלשה יותר וככל שמקדם המתאם רחוק מהאפס נאמר שעוצמת הקשר חזקה יותר:



מקדם המתאם יסומן באות r.

כדי לחשב את מקדם המתאם, יש לחשב את סטיות התקן של כל משתנה ואת השונות המשותפת.

שונות משותפת: $COV(x,y)=\dfrac{\sum(x-\bar{x})(y-\bar{y})}{n}=\dfrac{\sum xy}{n}-\bar{x}\cdot\bar{y}$

שונות של המשתנה X: $s_X^2=\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}=\dfrac{\sum_{i=1}^{n}x_i^2}{n}-\bar{x}^2$

שונות המשתנה Y: $S_Y^2=\dfrac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n}=\dfrac{\sum_{i=1}^{n}y_i^2}{n}-\bar{y}^2$

מקדם המתאם הליניארי: $r_{xy}=\dfrac{COV(x,y)}{S_x\cdot S_y}$

© כל הזכויות שמורות לברק קנדל www.gool.co.il

שאלות

1) להלן נתונים לגבי שישה תלמידים שנגשו למבחן. בדקו לגבי כל תלמיד את הציון שלו בסוף הקורס וכמו כן את מספר החיסורים שלו מהקורס.

מספר חיסורים	2	1	0	2	3	4
ציון	80	90	90	70	70	50

א. שרטטו דיאגראמת פיזור לנתונים. מה ניתן להסיק מהדיאגרמה על טיב הקשר בין מספר החיסורים של תלמיד לציונו? מיהו המשתנה הבלתי תלוי ומיהו המשתנה התלוי?

ב. חשבו את מדד הקשר של פירסון. האם התוצאה מתיישבת עם תשובתך לסעיף א'?

ג. הסבירו, ללא חישוב, כיצד מקדם המתאם היה משתנה אם היה מתווסף תלמיד שהחסיר 4 פעמים וקיבל ציון 80?

2) במחקר רפואי רצו לבדוק האם קיים קשר בין רמת ההורמון X בדם החולה לרמת ההורמון Y שלו. לצורך כך מדדו את רמת ההורמונים ההלו עבור חמישה חולים. להלן התוצאות שהתקבלו:

X	Y
10	12
14	15
15	15
18	17
20	21

א. מה הממוצע של כל רמת הורמון?

ב. מהו מקדם המתאם בין ההורמונים? ומה משמעות התוצאה?

3) נסמן ב- X את ההכנסה של משפחה באלפי ₪. נסמן ב- Y את ההוצאות של משפחה באלפי ₪. נלקחו 20 משפחות והתקבלו התוצאות הבאות:

$$\sum_{i=1}^{20} X_i = 240 \qquad \sum_{i=1}^{20} Y_i = 200$$

$$\sum_{i-1}^{20} (X_i - \bar{X})^2 = 76 \qquad \sum_{i-1}^{20} (Y_i - \bar{Y})^2 = 76$$

$$\sum_{i=1}^{20} (X_i - \bar{X})(Y - \bar{Y}) = 60.8$$

א. חשב את מדד הקשר הליניארי בין X ל- Y. מיהו המשתנה התלוי?

ב. מה המשמעות של התוצאה שקיבלת בסעיף א?

© כל הזכויות שמורות לברק קנדל

4) נסמן ב- X את ההכנסה של משפחה באלפי ₪. נסמן ב- Y את ההוצאות של משפחה באלפי ₪. נלקחו 20 משפחות והתקבלו התוצאות הבאות:

$$\sum_{i=1}^{20} X_i = 240 \qquad \sum_{i=1}^{20} Y_i = 200$$

$$\sum_{i-1}^{20} X_i^2 = 2960 \qquad \sum_{i=1}^{20} Y_i^2 = 2080$$

$$\sum_{i=1}^{20} X_i Y_i = 2464$$

חשבו את מדד הקשר הליניארי בין X ל- Y.

5) במוסד אקדמי ציון ההתאמה מחושב כך: מכפילים את הציון הממוצע בבגרות ב-3 ומפחיתים 2 נקודות. ידוע שעבור 40 מועמדים סטיית התקן של ממוצע הציון בבגרות הייתה 2.
מה מקדם המתאם בין ציון ההתאמה לציון הממוצע בבגרות שלהם?

6) להלן רשימת טענות, לגבי כל טענה קבעו נכון/לא נכון ונמקו.
א. מתווך דירות המיר מחירי דירות מדולר לשקל. נניח שדולר אחד הוא 3.5₪. אם מתווך הדירות יחשב את מדד הקשר של פירסון בין מחיר הדירה בשקלים למחיר הדירה בדולרים הוא יקבל 1.
ב. לסדרה של נתונים התקבל $\bar{X}=\bar{Y}=6$, $S_x = S_y = 1$.
לכן, מדד הקשר של פירסון יהיה 1.
ג. אם השונות המשותפת של X ושל Y הינה 0 אז בהכרח גם מקדם המתאם של פירסון יהיה 0.

שאלות רב-ברירה:

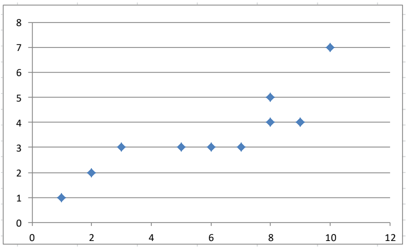
7) נמצא שקיים מקדם מתאם שלילי בין הציון בעברית לציון בחשבון בבחינה לכן:
א. הדבר מעיד שהציונים בכתה היו שליליים.
ב. ככל שהציון של תלמיד יורד בחשבון יש לו נטייה לרדת בעברית.
ג. ככל שהציון של תלמיד עולה בחשבון יש לו נטייה לרדת בעברית.
ד. אף אחת מהתשובות לא נכונה.

© כל הזכויות שמורות לברק קנדל

8) נלקחו 20 מוצרים ונבדק ביום מסוים המחיר שלהם בדולרים והמחיר שלהם בש״ח (באותו היום ערך הדולר היה-4.2₪).
מהו מקדם המתאם בין המחיר בדולר למחיר בש״ח?

א. 1
ב. 0
ג. 4.2
ד. לא ניתן לדעת.

9) להלן דיאגרמת פיזור:
מה יהיה מקדם המתאם בין שני המשתנים?



א. 1
ב. 0.85
ג. 0.15
ד. 0

תשובות סופיות

1) א. משתנה תלוי: ציון, משתנה ב״ת: מס׳ חיסורים. ראה דיאגרמה בוידאו. ניתן להסיק שקיים קשר לינארי שלילי וחלקי בין מספר החיסורים לציון התלמיד.
ב. -0.9325. ג. הקשר יישאר לינארי שלילי חלקי אך עוצמתו תחלש.

2) א. $\bar{y}=16$, $\bar{x}=15.4$ ב. $r_{xy}=0.96$.

3) א. 0.8.

4) 0.8.

5) 1.

6) א. נכון. ב. לא נכון. ג. נכון.

7) ג׳.

8) א׳.

9) ב׳.

www.gool.co.il © כל הזכויות שמורות לברק קנדל

בדיקת השערות על מקדם המתאם הליניארי – רקע

מדד הקשר הליניארי באוכלוסייה, שנקרא גם מקדם המתאם של פירסון או מדד הקשר של פירסון באוכלוסייה מסומן ב: ρ - פרמטר המאפיין את עוצמת הקשר הליניארי וכיוונו בין שני המשתנים הנחקרים באוכלוסייה. כאשר:
r - מדד הקשר הליניארי במדגם שמהווה אומד לפרמטר ρ.

השערת האפס: תהיה שבאוכלוסייה לא קיים כלל קשר ליניארי בין שני המשתנים $H_0:\rho=0$.

ההנחה שעליה אנו מתבססים בתהליך היא ששני המשתנים הנחקרים מתפלגים דו נורמלית.

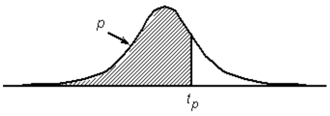
סטטיסטי המבחן: $t=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\sim t(n-2)$

סטטיסטי זה מתפלג t עם $n-2$ דרגות חופש.

השערת האפס:	$H_0:\rho=0$	$H_0:\rho=0$	$H_0:\rho=0$
השערת המחקר:	$H_1:\rho\neq 0$	$H_1:\rho<0$	$H_1:\rho>0$
כלל ההכרעה: אזור דחייה של השערת האפס	$t\geq t_{1-\alpha}$ או $t\leq -t_{1-\alpha}$	$t\leq -t_{1-\alpha}$	$t\geq t_{1-\alpha}$

© כל הזכויות שמורות לברק קנדל
www.gool.co.il

טבלת ערכים קריטיים של t - נספח: טבלת התפלגות T



	P						
דרגות חופש	0.75	0.90	0.95	0.975	0.99	0.995	0.9995
1	1.000	3.078	6.314	12.709	31.821	63.657	636.619
2	0.816	1.886	2.920	4.303	6.965	9.925	31.598
3	0.765	1.638	2.353	3.182	4.541	5.841	12.941
4	0.741	1.533	2.132	2.776	3.747	4.604	8.610
5	0.727	1.476	2.015	2.571	3.365	4.032	6.859
6	0.718	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	1.415	1.895	2.365	2.998	3.499	5.405
8	0.706	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	1.363	1.796	2.201	2.718	3.106	4.437
12	0.695	1.356	1.782	2.179	2.681	3.055	4.318
13	0.694	1.350	1.771	2.160	2.650	3.012	4.221
14	0.692	1.345	1.761	2.145	2.624	2.977	4.140
15	0.691	1.341	1.753	2.131	2.602	2.947	4.073
16	0.690	1.337	1.746	2.120	2.583	2.921	4.015
17	0.689	1.333	1.740	2.110	2.567	2.898	3.965
18	0.688	1.330	1.734	2.101	2.552	2.878	3.922
19	0.688	1.328	1.729	2.093	2.539	2.861	3.883
20	0.687	1.325	1.725	2.086	2.528	2.845	3.850
21	0.686	1.323	1.721	2.080	2.518	2.831	3.819
22	0.686	1.321	1.717	2.074	2.508	2.819	3.792
23	0.685	1.319	1.714	2.069	2.500	2.807	3.767
24	0.685	1.318	1.711	2.064	2.492	2.797	3.745
25	0.684	1.316	1.708	2.060	2.485	2.787	3.725
26	0.684	1.315	1.706	2.056	2.479	2.779	3.707
27	0.684	1.314	1.703	2.052	2.473	2.771	3.690
28	0.683	1.313	1.701	2.048	2.467	2.763	3.674
29	0.683	1.311	1.699	2.045	2.462	2.756	3.659
30	0.683	1.310	1.697	2.042	2.457	2.750	3.646
40	0.681	1.303	1.684	2.021	2.423	2.704	3.551
60	0.679	1.296	1.671	2.000	2.390	2.660	3.460
120	0.677	1.289	1.658	1.980	2.358	2.617	3.373
∞	0.674	1.282	1.645	1.960	2.326	2.576	3.291

© כל הזכויות שמורות לברק קנדל www.gool.co.il

שאלות

1) להלן נתונים על הוותק בעבודה (בשנים) ועל השכלה (בשנים) במדגם של 10 עובדים :

נחקר	1	2	3	4	5	6	7	8	9	10
X -וותק	13	18	2	8	16	9	5	28	17	24
Y -השכלה	12	14	17	8	11	12	13	8	12	15

מקדם המתאם חושב והתקבל : -0.31.

א. האם קיים מתאם בין וותק העובד להשכלתו? בדקו ברמת מובהקות של 5%?

ב. אם הוותק של העובד היה נמדד בחודשים האם התשובה לסעיף א' הייתה משתנה?

2) מחקר התעניין לבדוק את הקשר בין גיל נשים בהריון לרמת ההמוגלובין שלהן בדם בזמן הריון. נדגמו 7 נשים והתקבלו התוצאות הבאות :

נחקרת	1	2	3	4	5	6	7
המוגלובין	14.7	13.5	9.7	12	10.8	13	10.3
גיל	39	34	30	29	28	26	23

במדגם חושב מדד הקשר של פירסון להיות 0.7.

א. האם ניתן לומר שבמדגם אם אישה היא יותר מבוגרת אזי בהכרח יש לה יותר המוגלובין בדם?

ב. האם ניתן לומר, ברמת מובהקות של 5%, שקיים מתאם בין גיל האישה שבהריון לבין רמת ההמוגלובין שלה בדם?

3) בתחנה המטאורולוגית רצו לבדוק את הקשר שבין הטמפרטורה במעלות צלזיוס לכמות המשקעים במ״מ. הם אספו נתונים על 10 ימים במהלך חודש ינואר. המתאם שהתקבל היה -0.8.

א. בדקו ברמת מובהקות של 2.5% האם קיים קשר לינארי שלילי בחודש ינואר בין הטמפרטורה במעלות צלזיוס לבין המשקעים במעלות צלזיוס.

ב. כיצד הייתה משתנה התשובה לסעיף א אם הינו מוסיפים עוד תצפיות למדגם?

ג. על סמך טבלת T המצורפת עבור אילו רמות מובהקות ניתן להחליט שקיים קשר לינארי שלילי מובהק?

4) מתווך דירות חישב את מקדם המתאם בין שטח דירה במרכז תל אביב לבין המחיר של הדירה עבור 17 דירות. מקדם המתאם שקיבל היה 0.6.

א. בדוק ברמת מובהקות של 5% האם ניתן להגיד שקיים קשר ישר עולה בין שטח הדירה לבין מחיר הדירה במרכז תל אביב?

ב. מהי מובהקות התוצאה לבדיקת ההשערה שקיים קשר ישר עולה בין שטח הדירה לבין מחיר הדירה בתל אביב.

© כל הזכויות שמורות לברק קנדל
www.gool.co.il

תשובות סופיות

1) א. לא נדחה את H_0 . ב. לא תשתנה.

2) א. לא ב. לא נדחה את H_0 .

3) א. נדחה את H_0 . ב. לא ניתן לדעת.
ג. לפחות 0.005.

4) א. נדחה את H_0 . ב. $0.005 < P_V < 0.01$.

© כל הזכויות שמורות לברק קנדל

בדיקת השערות על מקדם המתאם הליניארי (באמצעות טבלה של ערכים קריטיים) – רקע

מדד הקשר הליניארי באוכלוסייה, שנקרא גם מקדם המתאם של פירסון או מדד הקשר של פירסון באוכלוסייה מסומן ב: ρ - פרמטר המאפיין את עוצמת הקשר הליניארי וכיוונו בין שני המשתנים הנחקרים באוכלוסייה. כאשר:
r - מדד הקשר הליניארי במדגם שמהווה אומד לפרמטר ρ.

השערת האפס: תהיה שבאוכלוסייה לא קיים כלל קשר ליניארי בין שני המשתנים: $H_0:\rho=0$.

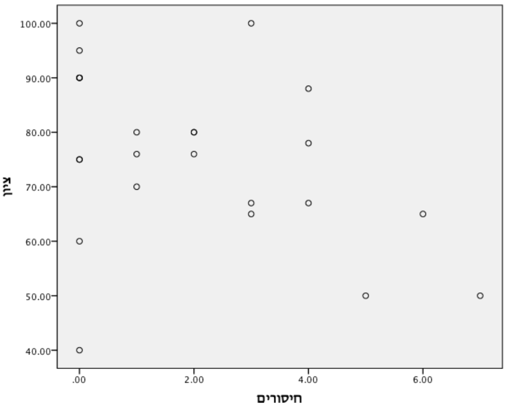
ההנחה שעליה אנו מתבססים בתהליך היא ששני המשתנים הנחקרים מתפלגים דו-נורמלית.

את מקדם המדגם הקריטי, שנסמן ב- r_c, נוציא מתוך טבלה של ערכים קריטיים שמצורפת בהמשך.

השערת האפס:	$H_0:\rho=0$	$H_0:\rho=0$	$H_0:\rho=0$
השערת המחקר:	$H_1:\rho\neq0$	$H_1:\rho<0$	$H_1:\rho>0$
כלל ההכרעה: אזור דחייה של השערת האפס	$r\geq r_c$ או $r\leq -r_c$	$r\leq -r_c$	$r\geq r_c$

דוגמה (פתרון בהקלטה) :

הדיקן ביקש לדגום סטודנטים כדי לבדוק את הקשר בין ציון הסטודנט בקורס למספר הפעמים שהוא החסיר שיעור בקורס.
דיאגרמת הפיזור שהתקבלה במדגם שבוצע:



מיהו המשתנה התלוי ומיהו המשתנה הבלתי תלוי במחקר?

מה ניתן לראות לגבי הקשר הליניארי בין המשתנים שהתקבל במדגם?

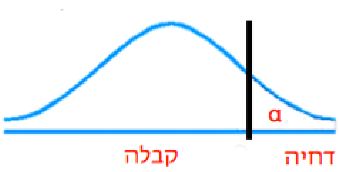
חושב האומד למקדם המתאם הליניארי על סמך 24 הסטודנטים שנדגמו והתקבל: -0.389.

מה משמעות של מקדם המתאם שהתקבל במדגם?

האם ניתן להגיד ברמת מובהקות של 5% שקיים מתאם ליניארי שלילי בין מספר החיסורים של הסטודנטים מהקורס לבין הציון של הסטודנטים בקורס?

www.gool.co.il © כל הזכויות שמורות לברק קנדל

טבלת ערכים קריטיים של מקדם המתאם הלינארי



0.0005	0.005	0.025	0.05	α / n
0.999	0.990	0.950	0.900	4
0.991	0.959	0.878	0.805	5
0.974	0.917	0.811	0.729	6
0.951	0.875	0.754	0.669	7
0.925	0.834	0.707	0.621	8
0.898	0.798	0.666	0.582	9
0.872	0.765	0.632	0.549	10
0.847	0.735	0.602	0.521	11
0.823	0.708	0.576	0.497	12
0.801	0.684	0.553	0.476	13
0.780	0.661	0.532	0.458	14
0.760	0.641	0.514	0.441	15
0.742	0.623	0.497	0.426	16
0.725	0.606	0.482	0.412	17
0.708	0.590	0.468	0.400	18
0.693	0.575	0.456	0.389	19
0.679	0.561	0.444	0.378	20
0.665	0.549	0.433	0.369	21
0.652	0.537	0.423	0.360	22
0.640	0.526	0.413	0.352	23
0.629	0.515	0.404	0.344	24
0.618	0.505	0.396	0.337	25
0.607	0.496	0.388	0.330	26
0.597	0.487	0.381	0.323	27
0.588	0.479	0.374	0.317	28
0.579	0.471	0.367	0.311	29
0.570	0.463	0.361	0.306	30
0.532	0.430	0.334	0.283	35

© כל הזכויות שמורות לברק קנדל

שאלות

1) להלן נתונים על הוותק בעבודה (בשנים) ועל השכלה (בשנים) במדגם של 10 עובדים:

נחקר	1	2	3	4	5	6	7	8	9	10
X - וותק	13	18	2	8	16	9	5	28	17	24
Y - השכלה	12	14	17	8	11	12	13	8	12	15

מקדם המתאם חושב והתקבל: 0.31-.

א. האם קיים מתאם בין וותק העובד להשכלתו? בדקו ברמת מובהקות של 5%.

ב. אם הוותק של העובד היה נמדד בחודשים האם התשובה לסעיף א' הייתה משתנה?

2) מחקר התעניין לבדוק את הקשר בין גיל נשים בהריון לרמת ההמוגלובין שלהן בדם בזמן הריון. נדגמו 7 נשים והתקבלו התוצאות הבאות:

נחקרת	1	2	3	4	5	6	7
המוגלובין	14.7	13.5	9.7	12	10.8	13	10.3
גיל	39	34	30	29	28	26	23

במדגם חושב מדד הקשר של פירסון להיות 0.7.

א. האם ניתן לומר שבמדגם אם אישה היא יותר מבוגרת אזי היא בהכרח יש לה יותר המוגלובין בדם?

ב. האם ניתן לומר, ברמת מובהקות של 5%, שהמתאם בין גיל האישה שבהריון לבין רמת ההמוגלובין שלה בדם הוא חיובי?

3) בתחנה המטאורולוגית רצו לבדוק את הקשר שבין הטמפרטורה במעלות צלזיוס לכמות המשקעים במ"מ. הם אספו נתונים על 10 ימים במהלך חודש ינואר. המתאם שהתקבל היה 0.8-.

א. בדקו ברמת מובהקות של 2.5% האם קיים קשר לינארי שלילי בחודש ינואר בין הטמפרטורה במעלות צלזיוס לבין המשקעים במ"מ?

ב. כיצד הייתה משתנה התשובה לסעיף א אם הינו מוסיפים עוד תצפיות למדגם?

תשובות סופיות

1) א. לא נדחה את H_0. ב. לא תשתנה.

2) א. לא. ב. נדחה את H_0.

3) א. נדחה את H_0. ב. לא ניתן לדעת.

© כל הזכויות שמורות לברק קנדל

כלים כמותיים מתקדמים של תכן סטטיסטי לאיכות

פרק 2 - רגרסיה פשוטה

תוכן העניינים

1. כללי....24

רגרסיה פשוטה:

רקע:

רגרסיה ליניארית פשוטה מסתמכת על המתאם הליניארי בין המשתנה התלוי (המנובא) לב"ת (המנבא).

מקדם המתאם: $\frac{SXY}{\sqrt{SXX}\cdot\sqrt{SYY}}=\frac{\sum_{i=1}^{n}x_iy_i-n\overline{xy}}{\sqrt{\left(\sum_{i=1}^{n}x_i^2-n\overline{x}^2\right)}\cdot\sqrt{\left(\sum_{i=1}^{n}y_i^2-n\overline{y}^2\right)}}=\frac{\operatorname{cov}(x,y)}{S_x\cdot S_y}=r$.

המודל באוכלוסייה: $y_i=\beta_0+\beta_1x_i+\varepsilon_i$

כאשר:

β_0 הוא החותך.

β_1 הוא שיפוע.

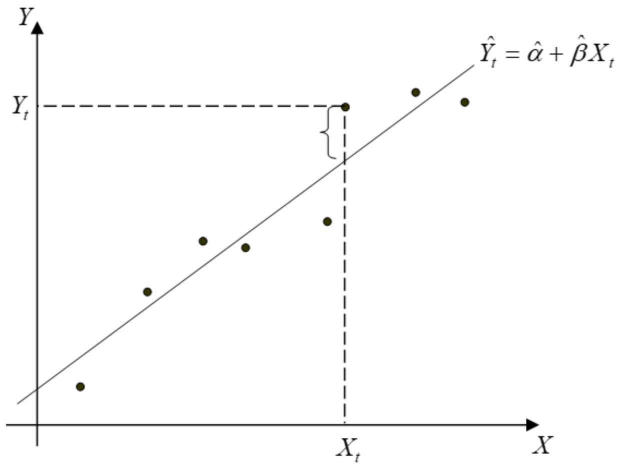
ε_i הינו גורם הטעות מסביב לקו הליניארי.

המודל הנאמד (על סמך מדגם): $\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_i$.

לסיכום:

1. במודל $y_i=\alpha+\beta x_i+\varepsilon_i$, α ו- β הם מספרים קבועים אך לא ידועים. אנו יכולים להעריך אותם ולקבל אומדים (תהליך קבלת האומדנים נקרא אמידה).
2. $\hat{\alpha}$ הוא האומד ל- α. $\hat{\beta}$ הוא האומד ל- β.
3. אומדי ריבועים פחותים (אר"פ) הם אומדים שחושבו בשיטת הריבועים הפחותים. אומדי הריבועים הפחותים מסומנים בד"כ ע"י 'כובע' - $\hat{\alpha}$, $\hat{\beta}$.
4. בעוד α ו- β הם קבועים, $\hat{\alpha}$ ו- $\hat{\beta}$ הם משתנים מקריים. מדוע? מפני שבכל מדגם מתקבלים $\hat{\alpha}$ ו- $\hat{\beta}$ אחרים.
5. את α ו- β אי אפשר לדעת, ולכן אי אפשר לדעת מהו הקו האמיתי, וכן אי אפשר לדעת את ε.
6. אפשר לדעת את e, שהיא הסטייה מקו הרגרסיה. נגדיר זאת באופן הבא:

- עבור X_i, הערך הצפוי של המשתנה המוסבר $(\hat{Y}_i)$ המתקבל לפי הרגרסיה הוא: $\hat{Y}_t=\hat{\alpha}+\hat{\beta}X_t$.
- הסטייה של התצפית (Y_t) מהערך הצפוי לפי הרגרסיה $(\hat{Y}_t)$ היא: $e_i=Y_i-\hat{Y}_i$.



האומדים של הרגרסיה ($\hat{\alpha}, \hat{\beta}$):

שיטת האמידה של α ושל β נקראת שיטת הריבועים הפחותים
.Ordinary Least Squares (OLS)
השאלה הנשאלת בשיטת אמידה זו היא:
איזה $\hat{\alpha}$ ו- $\hat{\beta}$ יביאו למינימום את סכום ריבועי טעויות האמידה.

ובתרגום מתמטי: $\min_{\hat{\alpha}\hat{\beta}} \sum e_t^2 = \min_{\hat{\alpha}\hat{\beta}} \sum (y_t - \hat{y}_t)^2 = \min_{\hat{\alpha}\hat{\beta}} \sum \left[y_t - \left(\hat{\alpha} + \hat{\beta} x_t \right) \right]^2 = ?$.

מתוך גזירת הפונקציה הזו מתקבלים האומדים $\hat{\alpha}$ ו- $\hat{\beta}$: $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$.

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{SXY}{SXX} = \frac{COV(X,Y)}{V(X)} = r\frac{S_Y}{S_X} .$$

מבחני המובהקות:

השערות: $H_0 : \beta = 0$, $H_1 : \beta \neq 0$.

ברגרסיה פשוטה בה יש לנו רק מנבא אחד: ניתן לבצע מבחן F למובהקות משוואת הרגרסיה או מבחן T למובהקות מקדם הרגרסיה (הביטא).
משמעות דחיית השערת האפס: משוואת הרגרסיה מובהקת, מקדם הרגרסיה מובהק, הקשר בין X ל- Y מובהק.
ולהיפך – אם השערת האפס לא נדחית: אין הוכחה לקשר בין המשתנים X ו- Y, משוואת הרגרסיה איננה מובהקת וכך גם מקדם הרגרסיה.

אמידת שונות הטעויות: $\hat{\sigma}^2 = MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{(1-r^2)SST}{n-2}$.

מבחן F:

מבחן זה נעשה על מנת לבדוק האם משוואת הרגרסיה מובהקת.

המבחן מתבסס על פירוק סכום הריבועים: $SST = SSR + SSE$.

$S_Y^2 \quad r^2S_Y^2 \quad (1-r^2)S_Y^2$

טבלת ניתוח שונות (טבלת ANOVA):

מקור	סכום ריבועים SS	דרגות חופש $d.f.$	ממוצע סכום ריבועים $MS = \frac{SS}{d.f.}$	F
מודל הרגרסיה	SSR	1	$MSR = \frac{SSR}{1}$	$\frac{MSR}{MSE}$
שאריות	SSE	$n-2$	$MSE = \frac{SSE}{n-2}$	
סה"כ	SST	$n-1$		

כלל הכרעה:

אם: $F_{st} > F_c\alpha(1, n-2)$ נדחה את השערת האפס.

מבחן t:

מבחן זה נעשה על מנת לבדוק האם מקדם הרגרסיה מובהק.

סטטיסטי המבחן: $t_{st} = \frac{\hat{\beta}_1 - \beta_{1,0}}{\overset{\wedge}{s.e.}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{SXX}}} \sim t_{c(n-2)}$.

אם השערת האפס מתייחסת ל- $\beta_0 =$ (בדר"כ): $t_{stt} = \frac{r^2\sqrt{n-2}}{\sqrt{1-r^2}}$.

כלל הכרעה:

	השערה חד צדדית ימנית $H_1:\beta_1>\beta_{1,0}$	השערה חד צדדית שמאלית $H_1:\beta_1<\beta_{1,0}$	השערה דו צדדית $H_1:\beta_1\neq\beta_{1,0}$
סטטיסטי המבחן	$t_{statistic}=\frac{\hat{\beta}_1-\beta_{1,0}}{\overset{\wedge}{s.e.}(\hat{\beta}_1)=\sqrt{\frac{\hat{\sigma}^2}{SXX}}}=\frac{r^2\sqrt{n-2}}{\sqrt{1-r^2}}$		
אזור דחייה	$t_{statistic}\geq t_{n-2,1-\alpha}$	$t_{statistic}\leq -t_{n-2,1-\alpha}$	$\lvert t_{statistic}\rvert\geq t_{n-2,1-\alpha/2}$
P-VALUE	$P(t_{n-2}>t_{statistic})$	$P(t_{n-2}>t_{statistic})$	$2*P(t_{n-2}>\lvert t_{statistic}\rvert)$

- שימו לב כי במודל של רגרסיה ליניארית פשוטה ערך ה-t סטטיסטי שהתקבל שווה בדיוק לשורש של ערך F המחושב: $t=\sqrt{F}$, $Pvalue=Pvalue$.

רווח סמך לאמידת β:

p (גבול תחתון $\leq\beta\leq$ גבול עליון) $=1-\alpha$.

$\hat{\beta}_1\pm t_{n-2,1-\frac{\alpha}{2}}\cdot\overset{\wedge}{s.e.}(\hat{\beta}_1)$.

מדד טיב ההתאמה R^2:

מדד שנותן את פרופורציית השונות המוסברת. כמה מהשונות של Y מוסבר על ידי השונות של X:

X מסביר את כל השונות של Y) $0\leq R^2=\frac{SSR}{SST}=1-\frac{SSE}{SST}\leq 1$ (X לא מסביר כלום מהשונות של Y).

נרצה פרופורציית שונות מוסברת קרובה ככל האפשר ל-1.

אחוז השונות המוסברת: $R^2\cdot 100$.

רב"ס שמטרתו לאמוד את תוחלת ערכי המשתנה התלוי (μ_0) עבור ערך מסוים של המשתנה הב"ת (x_0). במילים אחרות אנו מתבקשים לאמוד את הניבוי באוכלוסייה עבור ערך מסוים של X.

האומד הנקודתי (הסטטיסטי) סביבו בנוי הרב"ס הוא הניבוי במדגם עבור אותו

ה- X : $\hat{\mu}_0=\hat{y}_0=\ddot{\alpha}+\hat{\beta}x_0$.

נוסחת הרב"ס : $\hat{\mu}_0 \pm t_{n-2}\left(\frac{\alpha}{2}\right)\cdot\sqrt{MSRES\cdot\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{SSX}\right)}$.

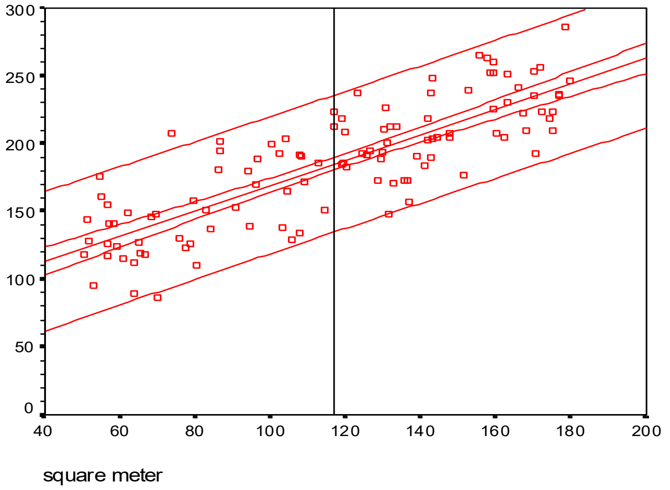
טעות התקן/גודל הרב"ס מושפעים מ-4 גורמים :

1. $MSRES$ - האומד לשונות הטעויות. ככל שגדל, טעות התקן/הרב"ס גדלים ולהפך.
2. n - גודל המדגם. ככל שגדל, טעות התקן/הרב"ס קטנים ולהפך.
3. SSX - מונה השונות של X (קשור לתופעת קיצוץ תחום). ככל שגדל, טעות התקן/הרב"ס קטנים ולהפך.
4. $(x_0-\bar{x})$ - הסטייה של ערך X המסוים מהממוצע של X. ככל שגדלה טעות התקן/הרב"ס גדלים ולהפך.

רב"ס לערכי Y עבור ערך מסוים של X :
רב"ס שמטרתו לאמוד את כל טווח ערכי Y (y_0) עבור ערך X מסוים (x_0).

נוסחת הרב"ס : $\hat{\mu}_0 \pm t_{n-2}\left(\frac{\alpha}{2}\right)\cdot\sqrt{MSRES\cdot\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{SSX}\right)}$.

ניתן לראות כי גם רב"ס זה בנוי סביב האומדן הנקודתי לתוחלת ערכי Y עבור ערך ה- X המסוים ($\hat{\mu}_0$).
ההבדל בין רב"ס לערכי Y לבין הרב"ס לתוחלת ערכי Y בא לידי ביטוי בטעות התקן. ניתן לראות כי טעות התקן של הרב"ס לערכי Y גדולה יותר מטעות התקן של הרב"ס לתוחלת ערכי Y. כאשר כל יתר הפרמטרים נשארים קבועים רב"ס זה יהיה רחב יותר מן הרב"ס לתוחלת.
התרשים הבא מתאר רב"ס לתוחלת ולערך המשתנה התלוי וממחיש זאת בבירור :



שאלות:

קו הרגרסיה:

1) מתווך דירות בתל אביב רצה לבדוק איך משפיע גודלה של דירה על המחיר שבו היא נמכרת. הוא הניח 2 הנחות מקדימות:

1. רק גודל הדירה משפיע על מחיר הדירה באופן שיטתי. כל שאר הדברים המשפיעים על מחיר הדירה הם אקראיים ולא ניתנים לחיזוי.
2. ההשפעה של גודל הדירה על מחיר הדירה היא ליניארית.

גודל הדירה הינו X ומחיר הדירה הינו Y. מודל המתווך: $y_i = \alpha + \beta x_i + \varepsilon$.

המתווך אסף נתונים על 6 דירות, שנמכרו בחודש האחרון באותו אזור:

מספר הדירה	גודל הדירה במ"ר	מחיר הדירה באלפי דולרים
1	$X_1 = 70$	$Y_1 = 190$
2	$X_2 = 70$	$Y_2 = 210$
3	$X_3 = 80$	$Y_3 = 250$
4	$X_4 = 100$	$Y_4 = 290$
5	$X_5 = 120$	$Y_5 = 360$
6	$X_6 = 120$	$Y_6 = 380$

א. מקדם המתאם בין גודל הדירה למחיר הדירה. מה משמעותו?

ב. קו הרגרסיה לניבוי מחיר הדירה באמצעות גודל הדירה ופרשו את משמעות המקדמים.

ג. המחיר החזוי על פי קו הרגרסיה של דירה בגודל 100 מ"ר.

מבחן F:

2) בצעו מבחן F לבדיקת הקשר בין גודל הדירה למחירה ברמת מובהקות של 1%.

מבחן t:

3) בהמשך לדוגמא הנ"ל:

א. בצעו מבחן t למובהקות מקדם הרגרסיה ברמת מובהקות של 1%.

ב. בדקו את הטענה כי עליה במ"ר אחד תעלה את מחיר הדירה ביותר מ-2000$.

ג. מהו ה-pvalue של מובהקות הקשר בין גודל הדירה למחירה. מה משמעותו?

קשר בין מבחן F למבחן t:

4) חשבו את סטטיסטי המבחן F על סמך סטטיסטי המבחן t שקיבלתם.
מה ה-pvalue של מבחן F?

5) חשבו רב״ס לאמידת מקדם הרגרסיה ברמת סמך של 0.99.
השוו עם תוצאות מבחן t.

6) חשבו את אחוז השונות המוסברת של מחיר הדירה על ידי גודלה.

רווח בר סמך לתוחלת:

7) השאלה מבוססת על נתוני דוגמא מס׳ 2 (ראו סרטון) והפלטים הבאים :

Case Summaries

	N	Mean	Std. Deviation	Minimum	Maximum
SIZE square meter	112		39.13942	50.46	179.76
PRICE thousands $	112	185.0664	44.45345	86.20	286.56

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)					.000	60.979	91.015
	SIZE square meter		.062	.823	15.173	.000		

a. Dependent Variable: PRICE thousands $

Descriptive Statistics

	Mean	Std. Deviation	N
SIZE square meter	116.740	39.139	112
PRICE thousands $	185.066	44.453	112
PRE_1 Unstandardized Predicted Value	185.066	36.568	112
RES_1 Unstandardized Residual	.000	25.277	112

חשב רב״ס ברמת סמך של 95% לתוחלת מחיר הדירה כאשר שטח הדירה הוא 100 מ״ר.

רווח בר סמך לערכי נעלם:

8) חשב רב״ס ברמת ביטחון של 95% למחיר הדירה עבור שטח דירה של 100 מ״ר.
מה ההבדל בין רב״ס זה לרב״ס הקודם?

תרגול מסכם:

9) בפיצוציית ״שלמה המלך״ חושדים כי מספר הלקוחות המבקרים בפיצוצייה תלוי במחיר המכירה של הבירה במקום. לשם בדיקת הנושא ערכו ניסוי בו בכל שבוע שינו את מחיר הבירה במקום ומנו את מספר הלקוחות שהגיעו במשך אותו שבוע. משך הניסוי 7 שבועות עוקבים. להלן נתוני הניסוי:

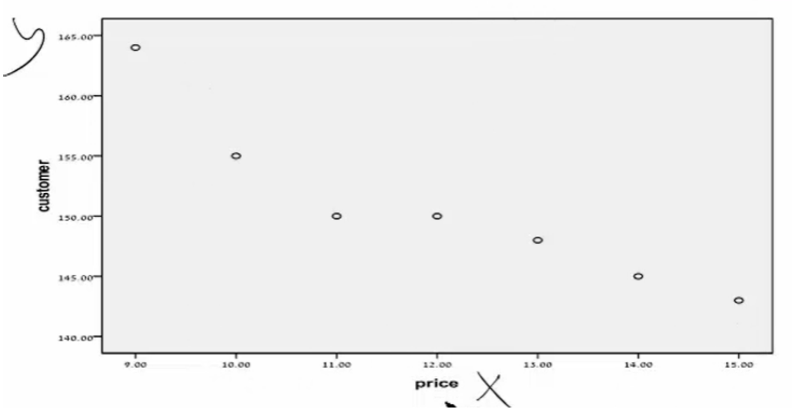
	שבוע 1	שבוע 2	שבוע 3	שבוע 4	שבוע 5	שבוע 6	שבוע 7
מחיר הבירה	15	14	13	12	11	10	9
כמות הלקוחות	143	145	148	150	150	155	164

א. אמדו את מודל הרגרסיה ע״י חישוב מקדמי הרגרסיה.
ב. חשבו את מקדם המתאם r_{xy}.
ג. אמדו את השונות של שאריות המודל.
ד. בצעו בדיקה גראפית של אקראיות השאריות.
ה. חשבו את אחוז השונות המוסברת. מה משמעותה?
ו. בצעו חיזוי לכמות הלקוחות אם מחיר הבירה יהיה 16 ₪.
האם להערכתכם ניתן להסתמך על חיזוי זה?
ז. בצעו מבחן F לבדיקה האם קיים קשר בין מחיר הבירה לבין כמות הלקוחות.
ח. בצעו מבחן t לבדיקה האם קיים קשר בין מחיר הבירה לבין כמות הלקוחות המבקרים בפיצוצייה ברמת מובהקות 5%. השוו את התוצאות.
ט. אמדו את מקדם הרגרסיה ברמת סמך של 0.95. השוו את התוצאה עם הסעיף הקודם.
י. כתבו דו״ח קצר על הממצאים.

קריאת פלטים של SPSS:

(10 על סמך הנתונים של השאלה הקודמת התקבלו הפלטים הבאים :

דיאגרמת הפיזור (scatter plot):



סטטיסטיקה תיאורית (descriptive statistics):

Descriptive Statistics

	Mean	Std. Deviation	N
customer	150.7143	7.01699	7
Price	12.0000	2.16025	7

פלט מקדם המתאם (correlations):

Correlations

		customer	Price
Pearson Correlation	customer	1.000	-.935
	price	-.935	1.000
Sig. (1-tailed)	customer	.	.001
	price	.001	.
N	customer	7	7
	price	7	7

פלט model summary:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.935[a]	.873	.848	2.73470

a. Predictors: (Constant), price

b. Dependent Variable: customer

פלט ניתוח שונות (ANOVA):

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	258.036	1	258.036	34.503	.002[a]
	Residual	37.393	5	7.479		
	Total	295.429	6			

a. Predictors: (Constant), price

b. Dependent Variable: customer

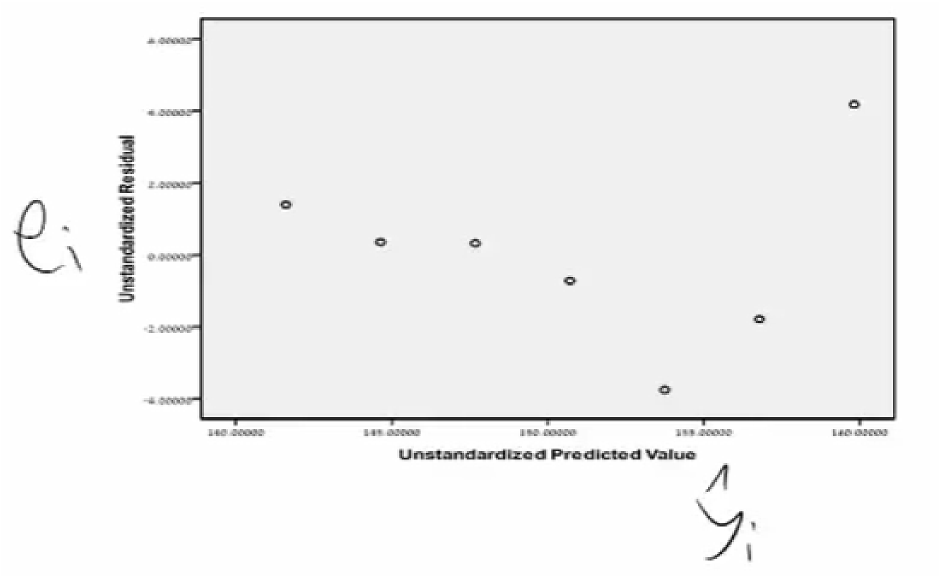
פלט מקדמי הרגרסיה (coeffitients):

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	187.143	6.287		29.765	.000
	Price	-3.036	.517	-.935	-5.874	.002

a. Dependent Variable: customer

גרף ניתוח שאריות:



על סמך הפלטים הנתונים:

א. מהו מודל הרגרסיה שנאמד?

ב. מהו מקדם המתאם r_{xy} ?

ג. מהי השונות של שארית המודל?

ד. האם נמצא דפוס מיוחד בשאריות?

ה. מהו אחוז השונות המוסברת?

ו. על פי מבחן F: האם קיים קשר בין מחיר הבירה לבין כמות הלקוחות המבקרים בפיצוצייה ברמת מובהקות 5%?

ז. על פי מבחן t: האם קיים קשר בין מחיר הבירה לבין כמות הלקוחות המבקרים בפיצוצייה ברמת מובהקות 5%? השוו את התוצאות.

ח. מה ה-pvalue של המבחנים הסטטיסטיים? מה משמעותו?

ט. בדקו האם קיים קשר חיובי מובהק בין המשתנים ברמת מובהקות 5%?

תשובות סופיות:

1) א. $r = 0.987$. ב. $\hat{Y}_t = -27.32 + 3.29X_t$. ג. 301.68 אלף דולר.

2) $F_{st} = 21.198$, יש עדות לקשר מובהק.

3) א. $t_{stt} = 4.604$. ב. יש עדות לכך. ג. $\text{pvalue} < 0.001$.

4) $t_{stt}^2 = 147$, $\text{pvalue} < 0.001$.

5) $p(2.061 \le \beta \le 4.519) = 0.99$.

6) 97.4%.

7) $p(163.889 \le \mu_{100} \le 174.24) = 0.95$.

8) $p(119.036 \le Y_{100} \le 219.09) = 0.95$, רחב יותר.

9) א. $\hat{y}_i = 187.143 - 3.0357x_i$. ב. $r = -0.93457$. ג. $\hat{\sigma}^2 = 7.4785$.
ד. ראו סרטון. ה. $R^2 = 0.873$. ו. $\hat{y} = 138.5714$, כן.
ז. $F_{stt} = 34.5 > F_c 0.05(1,5) = 6.6$, יש עדות לכך. ח. $t_{st} = -5.87395$.
ט. $p(-4.36 \le \beta \le -1.709) = 0.95$. י. ראו סרטון.

10) א. $\hat{y}_i = 187.143 - 3.036x_i$. ב. $r_{xy} = 0.935$. ג. $\text{MSE} = 7.479$.
ד. לא. ה. $R^2 = 0.874$. ו. $F = 34.503$.
ז. $t = -5.874$. ח. $\text{pvalue} = 0.002$. ט. ראו סרטון.

כלים כמותיים מתקדמים של תכן סטטיסטי לאיכות

פרק 3 - רגרסיה מרובה

תוכן העניינים

1. כללי.. 36

רגרסיה מרובה:

רקע:

ניבוי המשתנה התלוי באמצעות יותר ממשתנה ב"ת אחד.

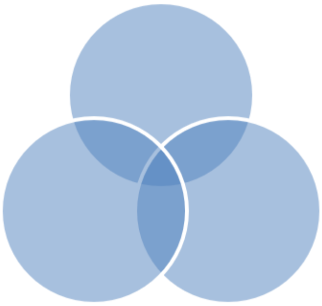
המודל באוכלוסייה: $y=\alpha+\beta_1x_1+\beta_2x_2+...+\beta_nx_n$.

מקדמי מודל הרגרסיה המרובה:

α = חותך אחד שמשמעותו: הציון המנובא כאשר כל המשתנים הב"ת = 0.

$\beta_1...\beta_J$ = מקדמי השיפוע. מס' הבטות = למספר המשתנים הב"ת במודל.

משמעות מקדם השיפוע β_j: ההשפעה הייחודית של המשתנה הב"ת המסוים לניבוי המשתנה התלוי, בניכוי השפעתם של כל יתר המשנים הב"ת האחרים המצויים במשוואת הרגרסיה.



אמידת מודל הרגרסיה המרובה:

ברגרסיה מרובה, כמו ברגרסיה פשוטה, שיטת האמידה הטובה ביותר היא שיטת הריבועים הפחותים. כלומר, נרצה להביא את סכום הטעויות בניבוי למינימום.

מפתרון פונקציית הריבועים הפחותים נקבל את אומדי הרגרסיה: $\hat{\alpha},\hat{\beta}_1...\hat{\beta}_J$.

מבחני מובהקות:

1. מבחן F למובהקות הרגרסיה:
   בדיקה האם קיים קשר ליניארי בין המשתנה התלוי Y לבין לפחות אחד מהמשתנים המסבירים.

   ההשערות הן:
   $$H_0:\beta_1=\beta_2=...=\beta_k=0$$
   $$H_1:Not\quad H_0=at\ least\ one\ of\ the\ \beta's\ is\ not\ 0$$

טבלת ניתוח שונות (ANOVA):

מקור	סכום ריבועים SS	דרגות חופש $d.f.$	ממוצע סכום ריבועים $MS=\frac{SS}{d.f.}$	$F_{st}\sim F_{k,n-k-1}$
מודל הרגרסיה	SSR	K	$MSR=\frac{SSR}{K}$	$F_{st}=\frac{MSR}{MSE}$
שאריות	SSE	$n-k-1$	$MSE=\frac{SSE}{(n-k-1)}$	
סה"כ	TSS	$n-1$		

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

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

חישוב סכומי הריבועים:

$$TSS=\sum_{i=1}^{n}y_i^2-n\bar{y}^2$$

$$SSR=R^2\cdot TSS \, .$$

$$SSE=(1-R^2)TSS$$

פרופורציית השונות המוסברת R^2:

$$R^2=\frac{SSR}{SST}=1-\frac{SSE}{SST} \, .$$

ברגרסיה מרובה אומד זה לפרופורציית השונות המוסברת הוא בעייתי שכן הוא מושפע ממספר המשתנים הב"ת במודל. אומד זה יכול רק לגדול בהוספת משתנים ב"ת למודל ולכן לא ייתן לנו אינדיקציה האם כדאי היה לי להוסיף אותם למודל
או לא.

האומד המתוקן לפרופורציית השונות המוסברת $AdjR^2$:

$$\bar{R}^2=1-\left[\frac{(1-R^2)(n-1)}{n-k-1}\right] \, .$$

בניגוד ל- R^2 לוקח בחשבון את מספר המשתנים הב"ת במודל. יכול שלא לגדול ואף לקטון בהוספת משתנה ב"ת שלא תורם תרומה משמעותית לניבוי.

2. מבחן t למובהקות משתנה ב"ת יחיד:

השערות: $H_0: \beta_j = 0$, $H_1: else$.

סטטיסטי המבחן וכלל הכרעת השערת האפס:

$$\left|t_{\hat\beta_i}\right| = \left|\frac{\hat\beta_i}{S_{\hat\beta_i}}\right| > t_{\left(T-k-1;1-\frac{\alpha}{2}\right)}.$$

רווח בר סמך לאמידת ה- β_j: $\hat\beta_j \pm t_{n-k-1,1-\frac{\alpha}{2}} s.e.\left(\hat\beta_j\right)$.

שאלות:

1) לצורך בדיקת ההשערה שקיים קשר בין מספר המוניות בעיר באר שבע (y) לבין מספר התושבים בעיר באלפים (x_1) ומספר הרכבים הפרטיים באלפים (x_2).

הוחלט לבנות מודל רגרסיה מהצורה: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$, על סמך הנתונים הבאים: $\sum_i y_i = 1673$, $\sum_i y_i^2 = 338,657$, $MSE = 119.789$.

א. ע"ס הנתונים הנ"ל, השלימו את טבלת ניתוח השונות הבאה.
איזו השערה ניתן לבדוק באמצעותה? כתוב את ההשערה ובחן אותה.

SOURCE	SS	DF	MS	F
Regression				
Error				
Total		8		

ב. חשבו את מדד טיב ההתאמה. הסבר את משמעותו.

ג. נתונה טבלת המקדמים (החלקית) הבאה:

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-511.727	114.9476				
X 1	9.208785		3.732167			
X 2	-8.79921	4.420456				

i. רשמו את האומדן למשוואת הרגרסיה ופרשו את מקדמיה.

ii. בחנו את ההשערה כי קיים קשר בין מספר הרכבים הפרטיים לבין מספר המוניות ברמת מובהקות של 5%.

iii. בנו רווח סמך למקדם של מספר התושבים בעיר

iv. ענה ללא חישוב (על סמך הסעיפים הקודמים) – האם קיים קשר בין מספר התושבים לבין מספר המוניות ברמת מובהקות 5%?

v. מהי תחזית מס׳ המוניות בבאר שבע עבור 100,000 תושבים ו-52,000 מכוניות פרטיות?

vi. האם ניתן לסמוך על תחזית זאת?

תרגול מסכם:

2) מעוניינים למצוא קשר בין מחיר הדירה (ב-$) לבין ארבעה משתנים מסבירים:
(1) שטח הדירה ו-(2) גודל שטח האמבטיה (ב-Sqft) וכן (3) מרחק הדירה מהים ו-(4) מהאוניברסיטה (במיילים).
לשם כך נדגמו מספר דירות והריצו רגרסיה אשר בה המשתנה המוסבר הוא מחיר הדירה.
להלן פלט הרגרסיה שהתקבל:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.952[a]			

a. Predictors: (Constant), Sea_Dist, Apartment, Bath

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression					.000[a]
	Residual					
	Total	1940484.615	25			

a. Predictors: (Constant), Univ_Dist, Bath, Sea_Dist, Apartment

b. Dependent Variable: Price

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-265.514	146.673		-1.810	.085
	Apartment		.449	.722	6.572	
	Bath	4.256		.297	2.687	.014
	Sea_Dist	-32.114	11.090	-.223		.009
	Univ_Dist	11.746	9.439	.095	1.244	.227

a. Dependent Variable: Price

ענו על הסעיפים הבאים:

א. מלאו את התאים החסרים בטבלה (אם לא ניתן למלא את כל התאים החסרים באופן מלא נמקו באופן מפורש מדוע לא ניתן).

ב. כתבו את האומדן למשוואת מחיר הדירה בצורה מפורשת על סמך הפלט הנ"ל. פרשו את מקדמי הרגרסיה.

ג. בדקו האם ארבעת הגורמים ביחד אכן מסבירים את מחיר הדירה. הסבירו את המסקנה שהגעתם אליה. השתמשו ברמת מובהקות 5%.

ד. הסבירו מהו ערך ה-Pvalue ומה ניתן להסיק ממנו לגבי המשתנים המסבירים?

ה. בנו רווח סמך למקדם גודל שטח האמבטיה. השתמשו ברמת מובהקות של 2%.

ו. ברמת מובהקות של 5% יש לבדוק האם המרחק מהאוניברסיטה אכן משפיע על מחיר הדירה.

ז. האם במודל הרגרסיה הנוכחי ניתן לוותר על גורם המרחק מהים? השתמשו ברמת מובהקות 1%

ח. בדקו את ההשערה כי קיים קשר חיובי בין גודל הדירה למחירה ברמת מובהקות של 5%.

תשובות סופיות:

1) א. השערה:
$$H_0: \beta_1 = \beta_2 = 0$$
$$H_1: at\ least\ one\ of\ the\ \beta's\ is\ not\ 0$$

SOURCE	SS	DF	MS	F
Regression	26945.784	2	13472.892	112.414
Error	718.733	6	119.789	
Total	27664.517	8		

ב. 97.4%. ג. i. $\hat{y}_i = -511.727 + 9.208x_{1i} - 8.799x_{2i}$.

ii. אין עדות לכך. iii. $p(3.17 \le \beta_1 \le 15.24) = 0.95$.

iv. כן. v. 321. vi. כן.

2) א. ראו סרטון. ב. $\hat{y}_i = -256.514 + 2.95x_{1i} + 4.256x_{2i} - 32.114x_{3i} + 11.746x_{4i}$.

ג. לפחות אחד מהמשתנים הב"ת שונה מאפס באוכלוסייה.

ד. ראו סרטון. ה. $p(1.016 \le \beta_2 \le 7.496) = 0.98$.

ו. לא. ז. לא. ח. יש עדות לכך.

כלים כמותיים מתקדמים של תכן סטטיסטי לאיכות

פרק 4 - רגרסיה - שאלות ממבחנים

תוכן העניינים

1. מבחן 1 41
2. מבחן 2 45

מבחן 1:

שאלות:

1) להלן תוצאות הרצת רגרסיה של Y בתלות ב-X עבור 10 תצפיות (חלק מהנתונים הושמטו בכוונה מהפלט, אך ניתנים לחישוב על ידך).

נתון כי: $\sum(X_i-\bar{X})^2=1475.6$.

מקור	סכום ריבועים SS
רגרסיה	SSR = 2148.6
שאריות	SSE = ?
סה"כ	SST = ?

משתנה	מקדם b_i	טעות תקן S_{bi}	ערך סטטיסטי(מתוקנן) t	p-value מובהקות
קבוע (חותך)	-24.7	11.3	?	?
X	1.20	?	**10.5**	

א. מהו SST?
   i. לא ניתן לקבוע.
   ii. 1994.42.
   iii. 2304.1.
   iv. 1629.78.

ב. האם הרגרסיה מובהקת? בדוק לפי p value.
   i. הרגרסיה מובהקת.
   ii. הרגרסיה אינה מובהקת.

2) לפניך פלט רגרסיה פשוטה (ממנו הושמטו נתונים שבאפשרותך להשלים), המתאר את ציון המבחן כפונקציה של מספר התרגילים שהגיש הסטודנט במהלך הסמסטר, ידוע כי כל הנחות המודל תקפות.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.842105598
R Square	0.709141839
Adjusted R Square	0.688366256
Standard Error	5.315523758
Observations	16

ANOVA

	Df	*SS*	*MS*	*F*	*Significance F*
Regression	1	964.4329004	964.4329004	34.13343	4.28E-05
Residual	14	395.5670996	28.25479283		
Total	15	1360			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	50.99134199	4.318918368	11.80650747	1.15E-08
מספר התרגילים	4.086580087	0.699471573	5.842381939	4.28E-05

א. ע"פ הנתונים, אחוז השונות של ציוני המבחן המוסברת ע"י מספר התרגילים שהגיש הסטודנט, היא_________. אם נוסיף משתנים נוספים, אחוז השונות המוסברת _________, ו-Radjusted _________.

i. 84%, יגדל, לא ניתן לקבוע ללא נתונים נוספים.
ii. 84%, יקטן, יגדל.
iii. 70.9%, יגדל, לא ניתן לקבוע ללא נתונים נוספים.
iv. 70.9%, יגדל, יקטן.
v. 68.8%, יגדל, יגדל.
vi. 68.8%, יקטן, יקטן.

ב. מהו הרב"ס של שיפוע הרגרסיה β_1 ? (בר"מ של 1%).

i. $2.58<\beta<5.58$.
ii. $2<\beta<6.17$.
iii. $1.74<\beta<6.88$.
iv. $2.86<\beta<5.3$.

במטרה לנבא בצורה טובה יותר את הצלחת הסטודנטים בבחינה, החליט החוקר להוסיף 2 משתנים נוספים לניתוח הרגרסיה.
מספר השיעורים בהם נכח הסטודנט, ומספר השעות שלמד לבחינה.
לפניכם הפלט החסר:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.880163577
R Square	0.774687922
Adjusted R Square	0.718359902
Standard Error	5.053253294
Observations	16

ANOVA

	df	SS	MS	F
Regression	3	?	351.1918579	13.75315
Residual	12	306.4244263	25.53536886	
Total	15	?		

	Coefficients	Standard Error	t Stat	P-value
Intercept	37.08959571	8.726136945	4.250402663	0.001127
מספר השיעורים	2.610000755	1.429904588	1.82529714	0.092932
מספר התרגילים	3.198068014	1.239162836	2.58082951	0.024061
שעות לימוד לבחינה	-0.108373802	1.021202927	-0.106123669	0.917238

ג. בדוק את ההשערה כי הרגרסיה מובהקת לכל אחד מהמשתנים המסבירים בר"מ של 1%.

i. הרגרסיה אינה מובהקת לכל המשתנים שנבדקו.

ii. לא ניתן לקבוע מהנתונים האם הרגרסיה מובהקת.

iii. הרגרסיה מובהקת למשתנה מספר התרגילים, אך אינה מובהקת למשתנים מספר השיעורים ומספר שעות הלימוד לבחינה.

iv. הרגרסיה מובהקת למשתנה מספר התרגילים ומספר השיעורים, אך אינה מובהקת למשתנה שעות הלימוד לבחינה.

ד. מהו SSR של הרגרסיה המרובה?

i. $SSR = 964.4$.

ii. $SSR = 1053.57$.

iii. $SSR = 694.57$.

iv. $SSR = 853.57$.

v. לא ניתן לחשב את SSR מהנתונים שהתקבלו.

תשובות סופיות:

1) א. iii. ב. i.

2) א. iii. ב. ii. ג. i. ד. ii.

מבחן 2:

שאלות:

1) הועלתה השערה שהוצאות האחזקה של מערכת לעיבוד נתונים קשורות למספר שעות השימוש השבועיות במערכת. להלן תוצאות חלקיות של פלט EXCEL של ניתוח רגרסיה בין Y הוצאות אחזקה שנתיות (במאות $) ו-X מספר שעות השימוש השבועיות.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	
R Square	
Adjusted R Square	
Standard Error	
Observations	10

ANOVA

	df	SS	MS	F	Significance F
Regression		860.051			0.00012
Residual					
Total		1004.525			

	Coefficients	Standard Error	t Stat	P-value
Intercept	10.528	3.745		0.023
שעות שימוש	0.953		6.901	

א. לאור התוצאות ה-p-value בבדיקת ההשערה: $H_0: \beta_1 \le 0.5$, $H_1: \beta_1 > 0.5$.

i. $Pv < 0.005$.

ii. $0.005 < Pv < 0.01$.

iii. $0.01 < pv < 0.025$.

iv. $0.025 < pv < 0.05$.

v. $Pv > 0.05$.

ב. אם היינו מריצים רגרסיה בה Y מספר שעות השימוש השבועיות ואילו המשתנה המסביר X, הוצאות אחזקה שנתיות (במאות $), אזי השיפוע של קו הרגרסיה יהיה:

i. 0.953.

ii. 1.049.

iii. 10.528.

iv. 0.095.

v. 0.898.

2) הועלתה השערה שמספר התקלות ברכב Y קשורה לגיל הנהג X. לשם כך נלקח מדגם של 10 נהגים.

כמו כן חושבו הסכומים הבאים:

$\sum X_i^2 = 14,227$, $\sum X_i = 363$, $\sum Y_i = 13$, $\sum Y_i^2 = 29$, $\sum X_i Y_i = 366$.

משוואת קו הרגרסיה נתונה על ידי: $\hat{Y} = 4.96 - 0.1X_i$.

א. ערכו של מקדם המתאם הליניארי בין מספר התקלות לבין גיל הנהג הוא:

i. ‎-10.59.

ii. ‎-0.9395.

iii. 0.8826.

iv. ‎-0.1.

החוקר לא היה מרוצה מעוצמת הקשר ולכן החליט להוסיף לרגרסיה את המשתנים הבאים: מספר הק"מ שהמכונית נסעה (באלפי ק"מ) וסוג הרכב.
במדגם נכללו 2 סוגי רכבים: A ו-B כאשר סוג B קודד כערך 0 וסוג רכב A קודד בערך 1.
להלן פלט הרגרסיה המרובה. שימו לב כי חלק מהערכים בפלט חסרים:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	
R Square	
Adjusted R Square	
Standard Error	0.204047
Observations	10

ANOVA

	df	SS	MS	F
Regression	3			0.00002
Residual	6	0.249811		
Total	9			

	Coefficients	Standard Error	t Stat	P-value
Intercept	1.407943	0.71094		
גיל הנהג X1	-0.03094	0.014611		
סוג הרכב X2	0.239574	0.152994		
ק"מ באלפים X3	0.027666	0.005329		

ב. להלן מספר טענות לגבי מובהקות הרגרסיה ומובהקות המשתנה סוג הרכב ברמת מובהקות 0.1:

i. הרגרסיה מובהקת והמשתנה סוג הרכב מובהק ברמת מובהקות 0.1.

ii. הרגרסיה מובהקת, אך המשתנה סוג הרכב אינו מובהק ברמת מובהקות 0.1.

iii. הרגרסיה אינה מובהקת, אך המשתנה סוג הרכב מובהק ברמת מובהקות 0.1.

iv. הרגרסיה והמשתנה סוג הרכב אינם מובהקים ברמת מובהקות 0.1.

ג. SST בפלט הרגרסיה המרובה שווה ל:

i. 12.1.

ii. 16.0.

iii. 20.0.

iv. אין מספיק נתונים לחשב.

ד. לאור התוצאות, רב"ס למקדם המשתנה מספר הקילומטרים, בר"מ 5%:

i. (0.015,0.04).

ii. (0.017,0.038).

iii. (0.02,0.035).

iv. אין מספיק נתונים לחשב.

ה. אם היינו מקודדים את סוג רכב B כערך 1 וסוג רכב A קודד בערך 0 משוואת הרגרסיה הייתה:

i. נשארת ללא שינוי.

ii. $\hat{Y} = 1.4079 - 0.0309X_{1i} - 0.2395X_{2i} + 0.0276X_{3i}$.

iii. $\hat{Y} = 1.6475 - 0.0309X_{1i} - 0.2395X_{2i} + 0.0276X_{3i}$.

iv. לא ניתן לדעת ללא הרצה מחדש.

ו. החוקר רצה להוסיף משתנה מסביר נוסף X_4 מספר השנים שחלפו מאז קבלת רישיון הנהיגה. להלן פלט הרגרסיה (חלק מהנתונים חסרים) עם 4 המשתנים המסבירים:

	Coefficients	Standard Error	t Stat	P-value
Intercept	1.5	0.71		
X1 גיל הנהג	5.1	14.1		
X2 סוג הרכב	0.25	0.2		
X3 ק"מ באלפים	0.02	0.008		
מס' השנים שחלפו מאז קבלת רשיון X4	-5.13094	0.8		

על סמך נתונים אלו:

i. חשש סביר למולטיקוליניאריות.

ii. התחזית הנקודתית של מספר התקלות תהיה דומה לזו של הרגרסיה הקודמת (עם 3 המשתנים המסבירים).

iii. קיימת קורלציה גבוהה בין חלק מהמשתנים המסבירים.

iv. כל התשובות נכונות.

3) במפעל מסוים הורץ מודל של רגרסיה ליניארית פשוטה על 8 עובדים כאשר Y תפוקת העובד ו-X גיל העובד. נמצא שהחלק המוסבר על ידי הרגרסיה הוא 74. הטעויות שהתקבלו מופיעות בחלקן בטבלה שלהלן:

e_8	e_7	e_6	e_5	e_4	e_3	e_2	e_1
0	2	-2	2	?	2	-3	0

אחת מהטענות שלהלן נכונה:

א. מקדם ההסבר בין X ל-Y הוא 0.74.

ב. לא ניתן לחשב את מקדם המתאם המרובה.

ג. $SSR = 18$.

ד. $SSE = 74$.

ה. אף אחת מהטענות איננה נכונה.

תשובות סופיות:

1) א. ii. ב. v.
2) א. ii. ב. i. ג. i. ד. i. ה. ii. ו. iv.
3) א'.