

כרכסיה וניתוח

שונות

חברת תרפיה

להלן תצפיות על הכנסה (X) וחטכון (Y) של חמש משפחות.

הכנסה חטכון	3000 100	5000 250	6400 500	5900 300-	6100 220
----------------	-------------	-------------	-------------	--------------	-------------

- מהו הניבוי לחטכון של משפחה הנלקחת באקראי מחמשת המשפחות?
- מהי הטעות הצפויה בניבוי זה?
- איזו משפחה תורמת תרומה שלילית לשונות המשותפת?
- איזו משפחה תורמת את התרומה הגדולה ביותר לשונות של ההכנסה, איזו לשונות של החטכון ואיזו לשונות המשותפת? יש להראות מה קורה לכל אחד מהמדדים אם אנו "מסלקים" את המשפחה בעלת התרומה המקסימלית מההתפלגות.

תשובה לשאלה 1 :

משפחה	X	Y	$X=x-5280$	$Y=y-154$	XY
1	3000	100	-2280	-54	300000
2	5000	250	-280	96	1250000
3	6400	500	1120	346	3200000
4	5900	-300	620	-454	-1770000
5	6100	220	820	66	1342000

הטבר :
 5280 הינו ממוצע של הכנסה.
 154 הינו ממוצע של חטכון.
 בכום המכפלות הינו : 4322000

- הניבוי יהיה 154 שקל שהוא החטכון הממוצע.
- הטעות בניבוי תהיה סטיית התקן 1231.9 ש"ח.
- המשפחה הרביעית שהכנסתה מעל לממוצע ויחטכויותיה" ב 300 ש"ח מתחת לממוצע.
- התצפית בעלת התרומה המקסימלית לשונות ההכנסה היא המשפחה הראשונה (הפער בין הכנסתה להכנסה הממוצעת הוא 2280 שקל.
 המשפחה השלישית תורמת את התרומה המרבית לשונות החטכון.

רצונו של אדם

שאלה 2

להלן נתונים על הקשר בין כמות הקורסים שהאדם לומד (X) לגובה המוטיבציה שלו (Y).
(ציוני המוטיבציה נעים בין 0 ל-15).

Y	10	7	10	4	8	8	6	7	9	11
X	12	11	14	6	10	7	9	11	10	10

- מצא את ערכי a ו- b לניבוי רמת מוטיבציה מתוך כמות הקורסים.
- שרטט את קו הרגרסיה המתאים.
- אדם שקבל 10 ב-X מה תנבא לו ב-Y?
- קיימים במדגם 3 אנשים שקבלו את הציון 10 ב-X. מדוע ציונם איננו זהה לציון שנבאת להם בסעיף ג'?

פתרון שאלה 2

א. בניית קו הרגרסיה טבלת סכומים, חישובי עזר לצורך חישוב ממוצעים סטיות תקן ושונותיות משותפות לצרכי חישוב מתאם פירסון.

Y^2	X^2	$X \cdot Y$	Y	X	
100	144	120	10	12	
49	121	77	7	11	
100	196	140	10	14	
16	36	24	4	6	
64	100	80	8	10	
64	49	56	8	7	
36	81	54	6	9	
49	121	77	7	11	
81	100	90	9	10	
121	100	110	11	10	
680	1048	828	80	100	סכומים

$$Sx = \frac{\sqrt{n \cdot \sum_{i=1}^n xi^2 - \left(\sum_{i=1}^n Xi\right)^2}}{n} = \frac{\sqrt{10 \cdot 1048 - 100^2}}{10} = \frac{\sqrt{10480 - 10000}}{10} = \frac{\sqrt{480}}{10} = \frac{21.908}{10} = 2.1908$$

$$Sy = \frac{\sqrt{n \cdot \sum_{i=1}^n yi^2 - \left(\sum_{i=1}^n Yi\right)^2}}{n} = \frac{\sqrt{10 \cdot 680 - 80^2}}{10} = \frac{\sqrt{6800 - 6400}}{10} = \frac{\sqrt{400}}{10} = \frac{20}{10} = 2.00$$

$$\bar{x} = \frac{\sum_{i=1}^n Xi}{n} = \frac{100}{10} = 10 \quad \bar{y} = \frac{\sum_{i=1}^n yi}{n} = \frac{80}{10} = 8$$

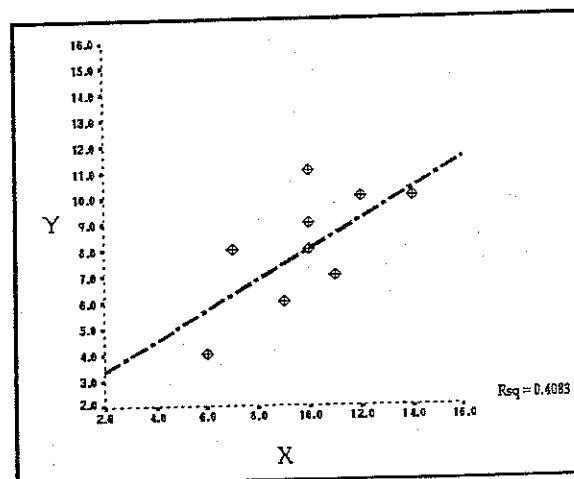
$$COV(x, y) = \frac{n \sum_{i=1}^n xi \cdot yi - \left(\sum_{i=1}^n xi\right) \cdot \left(\sum_{i=1}^n yi\right)}{n^2} = \frac{10 \cdot 828 - 100 \cdot 80}{10^2} = \frac{8280 - 8000}{100} = \frac{280}{100} = 2.8$$

$$r_{x, y} = \frac{cov(x, y)}{SxSy} = \frac{2.8}{2.1908 \cdot 2} = 0.639$$

$$b = r_{x, y} \cdot \frac{sy}{sx} = \frac{0.639 \cdot 2}{2.1908} = 0.581 \quad a = \bar{y} - b\bar{x} = 8 - (0.581) \cdot 10 = 2.19$$

$$\tilde{y} = a + bxi = 2.19 + 0.581 \cdot Xi$$

ב. תיאור גרפי של קו הרגרסיה



ג. ניבוי ערך ספציפי של Y לפי ערך ידוע של X. הצבת הערך במשוואה הכללית שהתקבלה.

$$\tilde{y} = a + bxi = 2.19 + 0.581 \cdot Xi$$

$$Xi = 10 \ggggggg 2.19 + 0.581 \cdot 10 = 8$$

עבור אדם שלומד 10 קורסים נובא רמת מוטיבציה של 8. במקרה זה אין אפילו צורך בחישוב. 10 זה ממוצע רמת הקורסים לכן הממוצע של X מובא את הממוצע של Y. זו הנקודה היחידה שהיא זהה גם בקו הניבויים (ראו שרטוט בסעיף ב') וגם בתצפיות, הנקודה היחידה שלגביה אין טעות ניבוי. ממוצעי המשתנים תמיד נפגשים על הקו בדיוק ללא טעות.

ד. הסיבה שלא כל אדם שלומד 10 קורסים יש לו רמת מוטיבציה של 8 נובעת מכך שהמתאם בין שני המשתנים אינו מושלם, לכן גם טיב הניבוי אינו מושלם. בהעדר קשר מושלם (מתאם שנמוך מ-1 או 1-) הניבוי אינו מדויק, כלומר לא כל הנקודות על קו הניבוי זהות לנקודות האמיתיות על פי התצפיות. קו הרגרסיה הוא הניבוי הטוב ביותר האפשרי על פי עקרון הריבועים הפחותים אך הוא מנבא על פי הממוצע ולכן אינו יכול להיות מדויק עבור כל תצפית ותצפית.

שאלה 3

נתונים: אחידות הגובה בס"מ והמשקל בק"ג של 10 אנשים.

	1	2	3	4	5	6	7	8	9	10
Y גובה	175	167	170	167	157	160	174	150	168	156
X משקל	55	60	59	50	49	55	65	38	70	50

- חשב את קו הניבוי ליבוי המשקל באמצעות הגובה.
- ידוע שמשקלו של אדם 67 ק"ג, מה יהיה גובהו המנובא?
- שרטט את הנתונים על גרף והעבר את קו הרגרסיה.
- מהי השונות המנובאת ליבוי הגובה באמצעות המשקל?

פתרון שאלה 3

א. בניית קו הרגרסיה טבלת סכומים, חישובי עזר לצורך חישוב ממוצעים סטיות תקן ושונויות משותפות לצרכי חישוב מתאם פירסון.

Y ²	X ²	X*Y	Y	X	
3025	30625	9625	55	175	
3600	27889	10020	60	167	
3481	28900	10030	59	170	
2500	27889	8350	50	167	
2401	24649	7693	49	157	
3025	25600	8800	55	160	
4225	30276	11310	65	174	
1444	22500	5700	38	150	
4900	28224	11760	70	168	
2500	24336	7800	50	156	
31,101	270,888	91,088	551	1644	סכומים

$$S_x = \frac{\sqrt{n \cdot \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}}{n} = \frac{\sqrt{10 \cdot 270888 - 1644^2}}{10} = \frac{\sqrt{2708880 - 2702736}}{10} = \frac{\sqrt{6144}}{10} = \frac{78.383}{10} = 7.838$$

$$S_y = \frac{\sqrt{n \cdot \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i\right)^2}}{n} = \frac{\sqrt{10 \cdot 31010 - 551^2}}{10} = \frac{\sqrt{310100 - 303601}}{10} = \frac{\sqrt{7409}}{10} = \frac{86.075}{10} = 8.607$$

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{1644}{10} = 164.4 \quad \bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{551}{10} = 55.1$$

$$COV(x, y) = \frac{n \sum_{i=1}^n x_i \cdot y_i - \left(\sum_{i=1}^n x_i\right) \cdot \left(\sum_{i=1}^n y_i\right)}{n^2} = \frac{10 \cdot 91088 - 551 \cdot 1644}{10^2} = \frac{910880 - 905844}{100} = \frac{5036}{100} = 50.36$$

$$r_{x, y} = \frac{cov(x, y)}{S_x S_y} = \frac{50.36}{7.838 \cdot 8.607} = 0.746$$

$$b = r_{x, y} \cdot \frac{s_y}{s_x} = \frac{0.746 \cdot 8.607}{7.838} = 0.819 \quad a = \bar{y} - b\bar{x} = 55.1 - (0.819) \cdot 164.4 = -79.543$$

$$\tilde{y} = a + bxi = -79.543 + 0.819 \cdot Xi$$

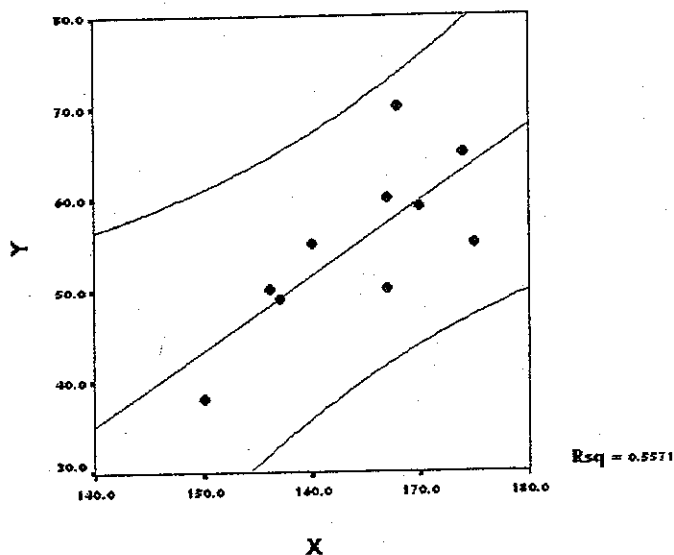
ב. ניבוי גובהו של אדם שמשקלו 67 ק"ג. ניבוי ערך ספציפי של X לפי ערך ידוע של Y. הצבת הערך במשוואה הכללית שהתקבלה. לשם כך צריך למצוא את משוואת הרגרסיה לניבוי X לפי Y.

$$b = r_{x,y} \cdot \frac{s_x}{s_y} = \frac{0.746 \cdot 7.838}{8.607} = 0.682 \quad a = \bar{y} - b\bar{x} = 164.4 - (0.682) \cdot 55.1 = 126.822$$

$$\tilde{x} = a + by_i = 126.822 + 0.682 \cdot y_i \gggggg y_i = 67 \ggggg \tilde{x}_{67} = 126.822 + 0.682 \cdot 67 = 172.51$$

עבור אדם ששוקל 67 ק"ג נבא גובהו של 172.51 ס"מ.

ג. תיאור גרפי של פיזור הנתונים (דיאגרמת פיזור) ושרטוט קו הרגרסיה



ד. חישוב שונות הניבויים כאשר Y מנבא את X

$$S_{reg}^2 = r^2 S_x^2 = 0.746^2 \cdot 7.838^2 = 34.17$$

שאלה 4

נתונים הנתונים הבאים:

$$cov(x, y) = 20.7 \quad \bar{x} = 66.83 \quad S_x^2 = 109.2$$

$$\bar{y} = 339.45 \quad S_y^2 = 806.56$$

- א. חשב את a ו-b לניבוי X מתוך Y.
- ב. חשב את a ו-b לניבוי Y מתוך X.
- ג. הסבר מדוע הניבוי איננו זהה למרות שהמתאם ביניהם הוא סימטרי (המתאם בין X ל Y זהה למתאם בין Y ל X).

פתרון שאלה 4

א. חישוב מקדם השיפוע והקבוע לניבוי X באמצעות Y.

$$r_{x,y} = \frac{cov(x, y)}{S_x S_y} = \frac{20.7}{\sqrt{109.2 \cdot 806.56}} = 0.069$$

$$b = r_{x,y} \cdot \frac{S_x}{S_y} = \frac{0.069 \cdot \sqrt{109.2}}{\sqrt{806.56}} = 0.025 \quad a = \bar{x} - b\bar{y} = 66.83 - (0.025) \cdot 339.45 = 58.344$$

$$\tilde{x} = a + by_i = 58.344 + 0.025 \cdot Y_i$$

ב. חישוב מקדם השיפוע והקבוע לניבוי Y באמצעות X.

$$r_{x,y} = \frac{\text{cov}(x,y)}{S_x S_y} = \frac{20.7}{\sqrt{109.2 * 806.56}} = 0.069$$

$$b = r_{x,y} * \frac{S_y}{S_x} = \frac{0.069 * \sqrt{806.56}}{\sqrt{109.2}} = 0.187 \quad a = \bar{y} - b\bar{x} = 339.45 - (0.187) * 66.83 = 326.953$$

$$\bar{y} = a + bxi = 326.953 + 0.187 * Xi$$

ג. יעילות הניבוי (טיב הניבוי), מכיון שטיב הניבוי נמדד באמצעות ריבוע מתאם פירסון, ומתאם פירסון הוא מדד קשר סימטרי. אחוז שונות מוסברת (היחס בין שונות הניבויים לשונות הכללית של Y), זהה בשני המקרים אולם הניבוי עצמו הוא ביחס להתפלגות הערכים אותם מנבאים והוא ינוע סביב הממוצע, מכיון שערכי הממוצעים שונים כך גם הניבויים. במלים אחרות מתאם פירסון אמנם סימטרי אך מודל הרגרסיה לא.

שאלה 5

נערך מחקר לבדיקת הקשר בין אורך השביתה לבין אורך המאמר על השביתה שהתפרסם בעיתון "הארץ" מיד עם סיומה. התקבלו הנתונים הבאים על 5 שביתות:

משך השביתה (X) (בימים)	אורך המאמר (Y) (במילים)
6	4000
4	3000
7	2000
8	1000
10	0

- א. חשב את מקדם המתאם בין אורך השביתה לבין אורך המאמר. הסבר את התוצאה שהתקבלה.
- ב. מהו אורך המאמר שתנבא לאחר 9 ימי שביתה?

פתרון שאלה 5

א. חישוב מתאם פירסון, שוב צריך את אותה טבלת חישובי עזר לצורך חישוב ממוצעים סטיות תקן ושונות משותפת בין שני המשתנים.

Y ²	X ²	X*Y	Y	X	סכומים
16000000	36	24000	4000	6	
9000000	16	12000	3000	4	
4000000	49	14000	2000	7	
1000000	64	8000	1000	8	
0	100	0	0	10	
30,000,000	265	58000	10000	35	

$$Sx = \frac{\sqrt{n \cdot \sum_{i=1}^n xi^2 - \left(\sum_{i=1}^n Xi\right)^2}}{n} = \frac{\sqrt{5 \cdot 265 - 35^2}}{5} = \frac{\sqrt{1325 - 1225}}{5} = \frac{\sqrt{100}}{5} = \frac{10}{5} = 2$$

$$Sy = \frac{\sqrt{n \cdot \sum_{i=1}^n yi^2 - \left(\sum_{i=1}^n Yi\right)^2}}{n} = \frac{\sqrt{5 \cdot 30,000,000 - 10,000^2}}{5} = \frac{\sqrt{150,000,000 - 100,000,000}}{5} = \frac{\sqrt{50,000,000}}{5} = 1414.213$$

$$COV(x, y) = \frac{n \sum xi * yi - \left(\sum xi * \sum yi\right)}{n^2} = \frac{5 \cdot 58000 - 35 \cdot 10000}{5^2} = \frac{290000 - 350000}{25} = \frac{-60000}{25} = -2400$$

$$rx, y = \frac{cov(x, y)}{SxSy} = \frac{-2400}{2 \cdot 1414.213} = -0.848$$

התקבל מתאם שלילי גבוה מאד. משמעות המתאם יש קשר שלילי בין אורך המאמר לאורך השביתה. ככל שהשביתה ארוכה יותר המאמר שנכתב עליה קצר יותר.

ב. ניבוי אורך מאמר באמצעות מספר ימי שביתה. קודם כל צריך לבנות את המשוואה ולאחר מכן להציב את מספר הימים המבוקש לקבלת ניבוי אורך המאמר.

$$b = rx, y \cdot \frac{Sy}{Sx} = \frac{-0.848 \cdot 1414.213}{2} = -599.626 \quad a = \bar{y} - b\bar{x} = 2000 - (-599.626) \cdot 7 = 6197.38$$

$$\bar{y} = a + bxi = 6197 - 599.626 \cdot Xi >>>>>>> 6197 - 599.626 \cdot 9 = 800.746$$

לשביתה בת 9 ימים נבא מאמר באורך של 800.746 מלים.

1. שאלה 1

2. במטרה לבחון את הקשר בין ציון במבחן קבלה (x) לבין הערכת המרצה את הסטודנט בתום שנה א' (y) נערך מחקר על עשרה נבדקים.

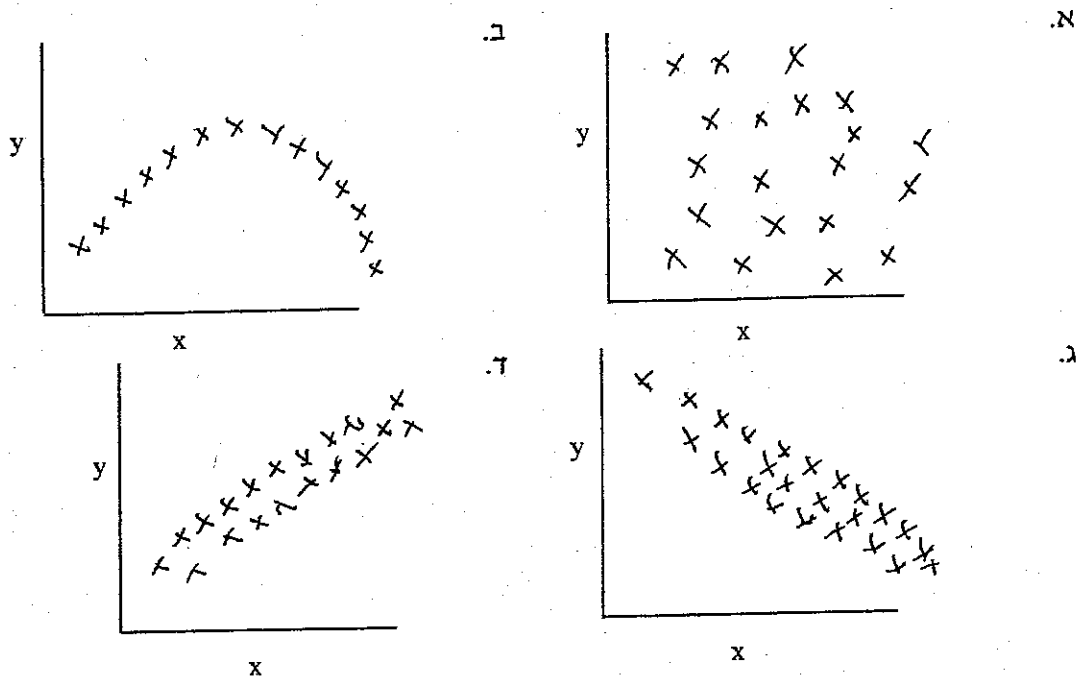
להלן הנתונים:

x	y	מס' נבדק
70	8	1
90	6	2
80	9	3
50	4	4
60	7	5
75	8	6
80	9	7
90	9	8
75	6	9
60	6	10

א.	חשב את ערכי a ו-b לניבוי y מתוך x.
ב.	האם ערך b מובהק ב-0.05.
ג.	סטודנט שקיבל 10 ע"י המנחה מה יהיה ציונו המנובא במבחן הקבלה?

שאלה 2

לפניך שרטוטים המתארים קשרים הפוטנטיים בין שני משתנים:
 מידת האדיבות במגע בין אישי (x) ומידת ההצלחה בעבודה כפקיד קבלה (y).



לגבי כל אחד משרטוטים אלה:

- א. פרש את הקשר המשורטט במילים.
- ב. לגבי אלו מהשרטוטים נכון לעשות רגרסיה ולגבי אלו אין מקום לכך? הסבר בקצרה מדוע.
- ג. אילו חושב מתאם פרסון לגבי כל אחד מהשרטוטים שלעיל, מה היה הערך הסביר שהיה מתקבל: חיובי, שלילי, קרוב ל-0 או קרוב ל-1.

שאלה 3

במחקר אודות הקשר בין גיל התחלת לימוד פסנתר (x) לבין רמת הנגינה לאחר 5 שנים של אמון (y), נלקח מדגם מקרי של 100 תלמידים שלמדו 5 שנים פסנתר. לכל תלמיד ניתן מבחן פסנתר, ועל סמך ביצועו במבחן נשפטה רמת נגינתו. להלן התוצאות:

$$\sum x_i = 1475$$

$$\sum y_i = 12890$$

$$\sum x_i^2 = 24459$$

$$\sum y_i^2 = 1714421$$

$$\sum x_i y_i = 186659$$

- א. האם זוהי רגרסיה פשוטה או מרובת?
 ב. מצא את ערך ה-b.
 ג. מצא את ערך ה-a.
 ד. האם ה-b מובהק ב- $\alpha = 0.05$?
 ה. הנח כי תלמיד מסוים התחיל את לימודי הפסנתר שלו בגיל 35. מה תהיה רמתו המנובאת בגיל 40?

חלק א שרטוט פירוש מילולי	חלק ב רגרסיה (לינארית)	חלק ג מתאם פירסון
א	אפשרית. שונות טעויות גבוהה	נמוך חיובי
ב	בלתי אפשרית	0
ג	אפשרית	קרוב ל -1
ד	אפשרית	קרוב ל 1

שאלה 3
חלק א

מדובר ברגרסיה פשוטה. יש רק משתנה מנבא אחד (גיל התחלת לימוד פסנתר).

חלק ב + ג (חישוב b, חישוב a)

כדי לחשב את a ואת b, נבצע חישובים מקדימים

$$S_x = \sqrt{\frac{n \sum x_i^2 - (\sum x_i)^2}{n^2}} = \sqrt{\frac{100 * 24459 - 1475^2}{100^2}} = \sqrt{\frac{270275}{10000}} \approx 5.199$$

$$S_y = \sqrt{\frac{n \sum y_i^2 - (\sum y_i)^2}{n^2}} = \sqrt{\frac{100 * 1714421 - 12890^2}{100^2}} = \sqrt{\frac{5290000}{10000}} \approx 23$$

$$\text{cov}(x, y) = \frac{\sum x_i y_i}{n} - \bar{x} \bar{y} = \frac{186659}{100} - 14.75 * 128.9 = -34.685$$

$$\Gamma = \frac{\text{cov}(x, y)}{S_x S_y} = \frac{-34.685}{5.199 * 23} \approx -0.29$$

חישוב b ולפיו חישוב a

$$b = \frac{\Gamma S_y}{S_x} = \frac{-0.29 * 23}{5.199} \approx -1.283 \Rightarrow a = \bar{y} - b \bar{x} = 128.9 - (-1.283) * 14.75 \approx 147.824$$

חלק ד - מובהקות b

לצורך חישוב המובהקות נזדקק לחישובים נוספים. את SS_x נחץ מתוך נתון ידוע וכן נחשב את SS_{res}

$$S_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} \Rightarrow \sum (x_i - \bar{x})^2 = S_x^2 n \cap \sum (x_i - \bar{x})^2 = SS_x \Rightarrow$$

$$SS_x = 5.199^2 * 100 = 2702.96$$

$$SS_{res} = (1 - \Gamma^2) S_y^2 n = (1 - (-0.29^2)) * 23^2 * 100 = 48451.11$$

וגם את Sb

$$S_b = \frac{\sqrt{SS_{res}}}{\sqrt{(n-2) \sum SS_x}} = \frac{\sqrt{48451.11}}{\sqrt{98 * 2702.96}} \cong 0.4277$$

כמו בשאלה ראשונה השערת האפס היא $H_0: \beta=0$

$$t = b/S_b = -1.283/0.4277 \cong -3$$

על פי לוח התפלגות t עבור $\alpha=0.025$ (דו צדדי) ולפי דרגת חופש 100 (ערכים דומים ל 90) - מתקבל +1.984.

מכאן שהשיפוע מובהק.

חלק ה - גיבוי לגבי תלמיד

גיל התלמיד 35.

רמת נגינה מנובאת לאחר 5 שנים (כשיגיע לגיל 40) - מחושבת כך:

$$\tilde{y} = \bar{a} + b \bar{x} = 147.824 - 1.283 * 35 = 102.919$$

מחיר של שעת עבודה במוסך ה"גולף" נקבע ל 150 שקל לשעת עבודה מלבד שכום קבוע של מאה שקלים. לקוחות רבים התלוננו על כך שלא תמיד עובדים שעה מלאה ובכל זאת המחיר הוא לפי שעה. לכן הוחלט לגבות את המחיר על פי מספר דקות עבודה. מהו המודל למחיר עבודה בכל אחד מהמקרים?

תשובה :

כאשר המחיר נקבע לפי שעה "עגולה" המודל הוא : $Y=100+150X$, כאשר Y הוא המחיר בשקלים ו-X הוא מספר השעות.

כאשר המחיר נקבע על פי מספר הדקות המודל הוא : $Y=100+2.5X$ כאשר X הפעם הוא מספר הדקות. לפי המודל הראשון כאשר השרות נמשך ארבעים דקות המחיר הוא 250 שקלים ($100+150$) שכס ארבעים הדקות עוגלו לשעה, לפי המודל השני המחיר הוא 150 שקלים ($100+2.5*40$).

וכיצד הגענו ל 2.5 ?

חלקנו את המחיר לשעה עגולה ל 60 דקות. הרי מעתה המחיר הוא לפי דקות.