

## רגרסיה ליניארית פשוטה

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

### מקדם המתאם:

$$r = \frac{\text{cov}(x, y)}{S_x \cdot S_y} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^n x_i^2 - n\bar{x}^2\right)} \cdot \sqrt{\left(\sum_{i=1}^n y_i^2 - n\bar{y}^2\right)}} = \frac{SXY}{\sqrt{SXX} \cdot \sqrt{SYY}}$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
 **המודל באוכלוסיה:**

כאשר:

$\beta_0$  הוא החותך

$\beta_1$  הוא שיפוע

$\varepsilon_i$  הינו גורם הטעות מסביב לקו הליניארי.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$
 **המודל הנאמד (על סמך מדגם):**

**נמחיש בדוגמא:**

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.

שתי ההנחות האלה מאפיינות את הקשר. אם נסמן את גודל הדירה ב-  $X$  ואת מחיר הדירה ב-  $Y$ , נוכל לכתוב באופן מתמטי כי  $y_i = \alpha + \beta x_i + \varepsilon$ . זהו המודל של המתווך.  $X$  ו-  $Y$  הם המשתנים של המודל.  $Y$  הוא המשתנה המוסבר של המודל.  $X$  הוא המשתנה המסביר של המודל (יכול להיות יותר ממשתנה מסביר אחד).  $\alpha$  ו-  $\beta$  הם הפרמטרים של המודל.  $\alpha$  נקרא חותך.  $\beta$ , או כל מקדם אחר של משתנה מסביר, נקרא שיפוע.  $\varepsilon$  מכונה הפרעה האקראית.

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

**מודל:**

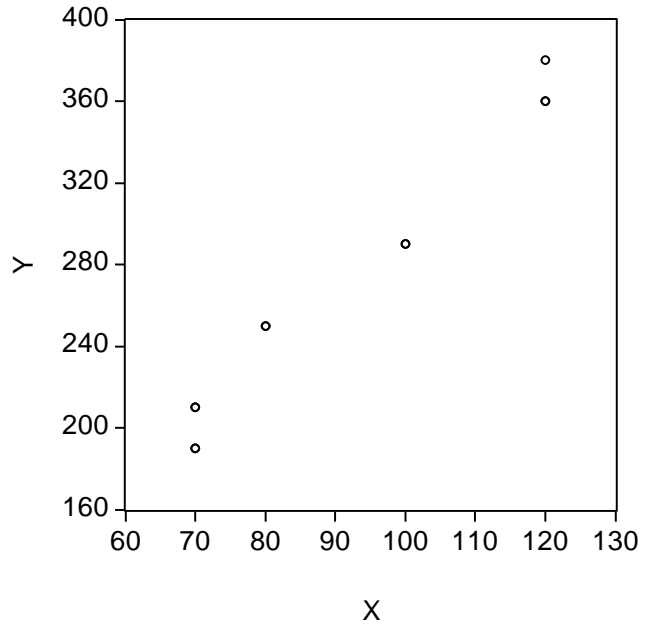
משתנה מוסבר
פרמטרים: חותך -  $\alpha$ 
משתנה מסביר
הפרעה אקראית

חותך -  $\alpha$ 
שיפוע -  $\beta$

אחרי הגדרת המודל המתווך אסף נתונים על 6 דירות, שנמכרו בחודש האחרון באותו איזור. זהו המדגם של המתווך. במדגם יש 6 תצפיות. נוהגים להציג את המודל כאשר לכל משתנה נוסף אינדקס  $y_i = \alpha + \beta x_i + \varepsilon_i$ . האינדקס מייצג את מספר התצפית.

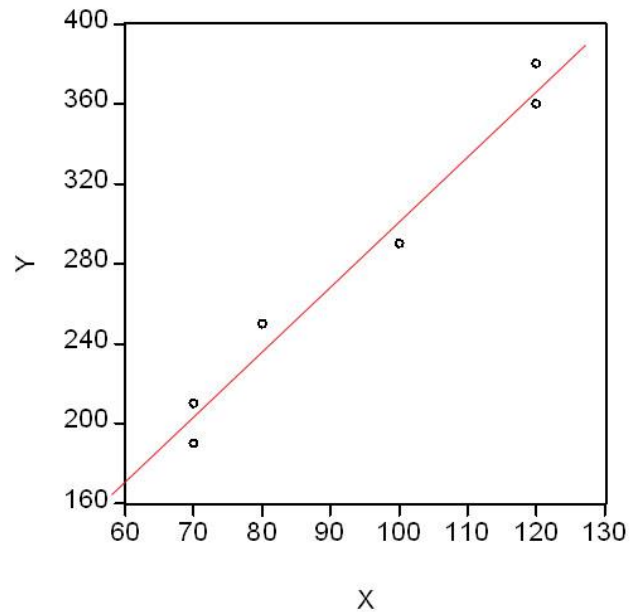
מספר הדירה	גודל הדירה במ"ר	מחיר הדירה באלפי דולרים
1	$X_1 = 70$	$Y_1 = 190$
2	$X_2 = 70$	$Y_2 = 210$
3	$X_3 = 80$	$Y_3 = 250$
4	$X_4 = 100$	$Y_4 = 290$
5	$X_5 = 120$	$Y_5 = 360$
6	$X_6 = 120$	$Y_6 = 380$

**נציג את 6 התצפיות בגרף:**



מהו הקו הישר המתאר את הקשר בין שני המשתנים בצורה הטובה ביותר? (הקו הוא ישר בגלל שהמתווך הניח לינאריות של המודל).

מסתבר שקו הרגרסיה הטוב ביותר הוא קו שחושב בשיטת הריבועים הפחותים (הסבר בהמשך):



הנוסחה של הקו היא:  $\hat{Y}_i = -27.32 + 3.29 X_i$ .

זהו כנראה לא הקו האמיתי, אך ממילא את הקו האמיתי אף פעם אי אפשר לדעת. סביר שקו זה הוא די קרוב לקו האמיתי.

לפי הנוסחה כל מ"ר נוסף שיש בדירה מעלה את מחירה ב-3,290 דולר.

מקו זה יודע המתווך להעריך מחירים של דירות. כשפנה אליו בעל דירה שגודלה 90 מ"ר ושאל אותו מה שווי הדירה, חישב המתווך לפי הנוסחה,  $-27.32 + 3.29 \cdot 90 = 268.78$ , והשיב לבעל הדירה: "המחיר שאתה יכול לקבל עליה הוא 268,780 דולר. אם יהיה לך מזל תקבל יותר, אבל יכול להיות שתצטרך למכור בפחות".

כלומר נוכל לומר כי אם יהיה לו מזל אז ההפרעה האקראית תהיה חיובית, ואם לא – היא תהיה שלילית.

## לסיכום:

1) במודל  $y_i = \alpha + \beta x_i + \varepsilon_i$ ,  $\alpha$  ו- $\beta$  הם מספרים קבועים אך לא ידועים. אנו יכולים להעריך אותם ולקבל אומדים (תהליך קבלת האומדנים נקרא אמידה).

2)  $\hat{\alpha}$  הוא האומד ל- $\alpha$ .  $\hat{\beta}$  הוא האומד ל- $\beta$ .

3) אומדי ריבועים פחותים (אר"פ) הם אומדים שחושבו בשיטת הריבועים הפחותים. אומדי הריבועים הפחותים מסומנים בד"כ ע"י 'כובע' -  $\hat{\beta}, \hat{\alpha}$ .

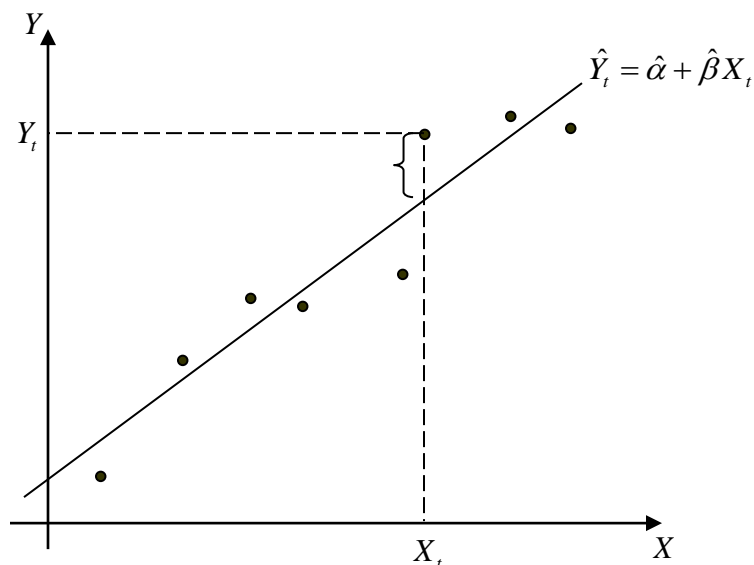
4) בעוד  $\alpha$  ו- $\beta$  הם קבועים,  $\hat{\alpha}$  ו- $\hat{\beta}$  הם משתנים מקריים. מדוע? מפני שבכל מדגם מתקבלים  $\hat{\alpha}$  ו- $\hat{\beta}$  אחרים.

5) את  $\alpha$  ו- $\beta$  אי אפשר לדעת, ולכן אי אפשר לדעת מהו הקו האמיתי, וכן אי אפשר לדעת את  $\varepsilon$ .

6) אפשר לדעת את  $e$ , שהיא הסטייה מקו הרגרסיה. נגדיר זאת באופן הבא:

\* עבור  $X_t$ , הערך הצפוי של המשתנה המוסבר ( $\hat{Y}_t$ ) 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:  $e_t = Y_t - \hat{Y}_t$



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

שיטת האמידה של  $\alpha$  ושל  $\beta$  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.

$$\min_{\hat{\alpha}\hat{\beta}} \sum e_i^2 = \min_{\hat{\alpha}\hat{\beta}} \sum (y_i - \hat{y}_i)^2 = \min_{\hat{\alpha}\hat{\beta}} \sum [y_i - (\hat{\alpha} + \hat{\beta}x_i)]^2 = ?$$

מתוך גזירת הפונקציה הזו מתקבלים האומדים  $\hat{\alpha}$  ו- $\hat{\beta}$ :

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

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

? על סמך הדוגמא הנ"ל חשבו את:

1. מקדם המתאם בין גודל הדירה למחיר הדירה. מה משמעותו?
2. קו הרגרסיה לניבוי מחיר הדירה באמצעות גודל הדירה ופרשו את משמעות המקדמים.
3. המחיר החזוי על פי קו הרגרסיה של דירה בגודל 100 מ"ר.

## 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 $\sigma^2$ שונות הטעויות:

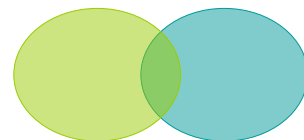
$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{(1-r^2)SST}{n-2}$$

## מבחן F

מבחן זה נעשה על מנת לבדוק האם משוואת הרגרסיה מובהקת.

מבחן F מתבסס על פירוק סכום הריבועים:

$$\underbrace{SST}_{S_Y^2} = \underbrace{SSR}_{r^2 S_Y^2} + \underbrace{SSE}_{(1-r^2) S_Y^2}$$



---

**טבלת ניתוח שונות (טבלת ANOVA)**

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

**כלל הכרעה:**

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

**? בהמשך לדוגמא הנ"ל:**

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

הערה: ניתן גם לשאול- האם משוואת הרגרסיה לניבוי מחיר הדירה על סמך גודלה מובהקת באוכלוסיה ברמת מובהקות של 1%?

**מבחן t**

מבחן זה נעשה על מנת לבדוק האם מקדם הרגרסיה מובהק.

**סטטיסטי המבחן:**

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

אם השערת האפס מתייחסת ל- $\beta=0$  (בדר"כ):

$$t_{stt} = \frac{r^2 \sqrt{n-2}}{\sqrt{1-r^2}}$$



**כלל הכרעה:**

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

**? בהמשך לדוגמא הנ"ל:**

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

אפשר גם לבקש: בצעו מבחן t לבדיקת הקשר בין גודל הדירה למחירה.

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

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

**\*\*שימו לב כי במודל של רגרסיה ליניארית פשוטה ערך ה-t סטטיסטי שהתקבל שווה בדיוק**

**לשורש של ערך F המחושב:**

$$t = \sqrt{F}$$

$$Pvalue = Pvalue$$

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

**מבחן F ?**

## רווח סמך לאמידת $\beta$ :

$$p(\text{גבול תחתון} \leq \beta \leq \text{גבול עליון}) = 1 - \alpha$$

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

? חשבו רב"ס לאמידת מקדם הרגרסיה ברמת סמך של 0.99. השוו עם תוצאות מבחן t.

## מדד טיב ההתאמה $R^2$ :

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

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

נרצה פרופורציית שונות מוסברת קרובה ככל האפשר ל-1.

אחוז השונות המוסברת:  $R^2 \cdot 100$

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

## תרגול מסכם

בפיצויית "שלמה המלך" חושדים כי מספר הלקוחות המבקרים בפיצויה 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	שבוע	שבוע	שבוע	שבוע	שבוע	שבוע	
9	10	11	12	13	14	15	מחיר הבירה
164	155	150	150	148	145	143	כמות הלקוחות

- 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. חשבו את אחוז השונות המוסברת. מה משמעותה?
- ה. בצעו חיזוי לכמות הלקוחות אם מחיר הבירה יהיה 16 ₪. האם להערכתכם ניתן להיסתמך על חיזוי זה?
- ו. בצעו מבחן  $F$  לבדיקה האם קיים קשר בין מחיר הבירה לבין כמות הלקוחות המבקרים בפיצויה ברמת מובהקות 5%.
- ז. בצעו מבחן  $t$  לבדיקה האם קיים קשר בין מחיר הבירה לבין כמות הלקוחות המבקרים בפיצויה ברמת מובהקות 5%. השוו את התוצאות.
- ח. אמדו את מקדם הרגרסיה ברמת סמך של 0.95. השוו את התוצאה עם הסעיף הקודם.

# רגרסיה פשוטה-פלטי SPSS

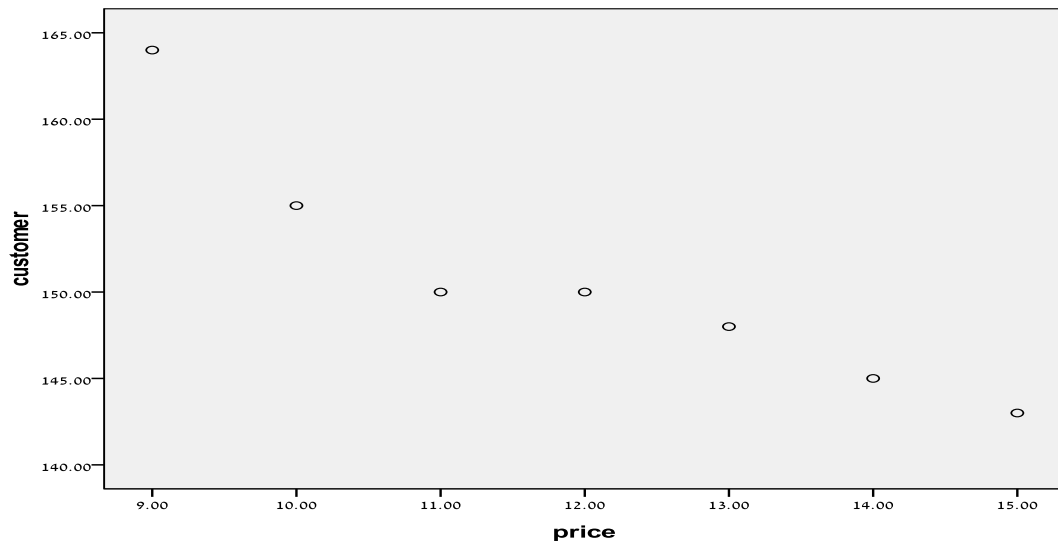
הבא נראה כיצד לקרוא פלטי SPSS ברגרסיה פשוטה.

על סמך הנתונים של שאלה מס' 1 :

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

התקבלו הפלטים הבאים:

**1) דיאגרמת הפיזור (scatter plot):**



**(2) סטטיסטיקה תיאורית (descriptive statistics):**

**Descriptive Statistics**

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

**(3) 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

**(4) פלט model summary:**

**Model Summary<sup>b</sup>**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.935 <sup>a</sup>	.873	.848	2.73470

a. Predictors: (Constant), price

b. Dependent Variable: customer

**(5) פלט ניתוח שונות (ANOVA):**

**ANOVA<sup>b</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	258.036	1	258.036	34.503	.002 <sup>a</sup>
	Residual	37.393	5	7.479		
	Total	295.429	6			

a. Predictors: (Constant), price

b. Dependent Variable: customer

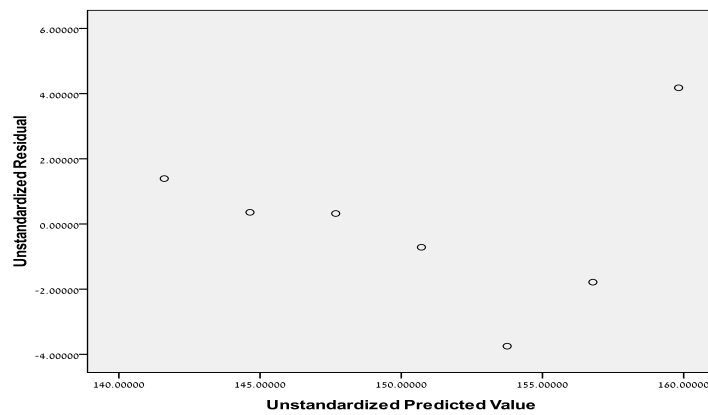
**(6) פלט מקדמי הרגרסיה (coefficients):**

**Coefficients<sup>a</sup>**

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

**(7) גרף ניתוח שאריות:**



**? על סמך הפלטים הנתונים :**

א. מהו מודל הרגרסיה שנאמד?

ב. מהו מקדם המתאם  $r_{xy}$ ?

ג. מהי השונות של שאריות 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- $p$ value של המבחנים הסטטיסטיים? מה משמעותו?

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