

סטטיסטיקה לניהול

פרק 28 - רגרסיה פשוטה

תוכן העניינים

1. כללי 1

הגרסיה פשוטה:

רקע:

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

$$r = \frac{\text{cov}(x, y)}{S_x \cdot S_y} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^n x_i^2 - n\bar{x}^2\right)} \cdot \sqrt{\left(\sum_{i=1}^n y_i^2 - n\bar{y}^2\right)}} = \frac{S_{XY}}{\sqrt{S_{XX}} \cdot \sqrt{S_{YY}}} : \text{מקדם המתאם}$$

המודל באוכלוסייה: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

כאשר:

β_0 הוא החותך.

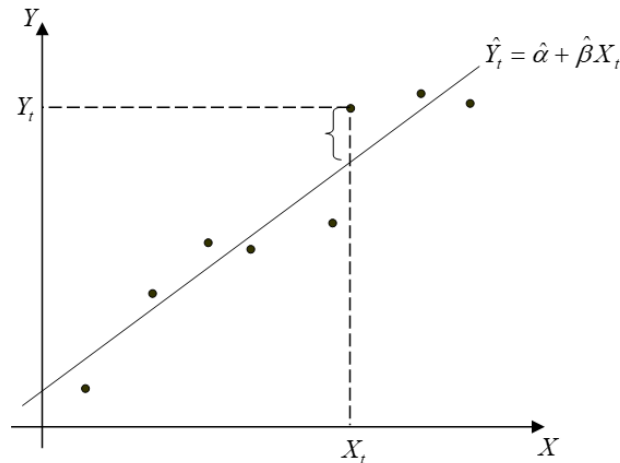
β_1 הוא שיפוע.

ε_i הינו גורם הטעות מסביב לקו הליניארי.

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

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 'כובעי' - $\hat{\beta}, \hat{\alpha}$.
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: $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$.
 - הסטייה של התצפית (Y_i) מהערך הצפוי לפי הרגרסיה (\hat{Y}_i) 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.

$$\min_{\hat{\alpha}, \hat{\beta}} \sum e_t^2 = \min_{\hat{\alpha}, \hat{\beta}} \sum (y_t - \hat{y}_t)^2 = \min_{\hat{\alpha}, \hat{\beta}} \sum [y_t - (\hat{\alpha} + \hat{\beta}x_t)]^2 = ?$$

ובתרגום מתמטי:

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

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

מבחני המובהקות:

$$H_0: \beta = 0$$

השערות:

$$H_1: \beta \neq 0$$

הרגרסיה 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 נדחת: אין הוכחה לקשר בין המשתנים X ו- Y , משוואת הרגרסיה איננה מובהקת וכך גם מקדם הרגרסיה.

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

אמידת שונות הטעויות:

מבחן F:

מבחן זה נעשה על מנת לבדוק האם משוואת הרגרסיה מובהקת.
 המבחן מתבסס על פירוק סכום הריבועים: $SST = SSR + SSE$
 $s_y^2 = r^2 s_y^2 + (1-r^2) s_y^2$

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.

אמידה נקודתית ואמידת מרווח ל- $E(Y)$ ול- Y :

אמידה נקודתית:

אמידה נקודתית עבור X_0 מסוים (תחזית) מחושבת על פי קו הרגרסיה במדגם:

$$\hat{Y}_0 = \hat{\alpha} + \hat{\beta} \cdot X_0$$

אמידת מרווח ל- $E(Y)$:

כאשר אנו מתבקשים לאמוד את התחזית באוכלוסייה עבור X_0 מסוים, נחשב רווח בר סמך לערך ממוצע של Y באוכלוסייה עבור X_0 מסוים $E(Y)$ ברמת סמך: $1-\alpha$.

$$\hat{Y} \pm t_{n-2, 1-\frac{\alpha}{2}} \hat{\sigma}_u \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2}} : \text{נוסחת הרבי"ס}$$

$$\hat{\sigma}_u = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}}, \quad \sum (X_i - \bar{X})^2 = S_{xx} = (n-1)S_x^2$$

$$p(\text{---} \leq E(Y) \leq \text{---}) = 1-\alpha : \text{רישום הרבי"ס}$$

אמידת מרווח ל- Y :
 כאשר אנו מתבקשים לאמוד את ערך בודד של Y באוכלוסייה עבור X_0 מסוים,
 נחשב רווח בר סמך לערך בודד של Y באוכלוסייה עבור X_0 מסוים (Y_0) ברמת
 סמך: $1-\alpha$.

$$\hat{Y} \pm t_{n-2, 1-\frac{\alpha}{2}} \hat{\sigma}_u \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2}} : \text{נוסחת הרב"ס}$$

$$. p(\text{---} \leq Y \leq \text{---}) = 1 - \alpha : \text{רישום הרב"ס}$$

התחזית מדויקת יותר (שונות התחזית קטנה יותר) כאשר :

1. n (גודל המדגם) גדול יותר.
2. שונות המשתנה המסביר X גדולה יותר.
3. X_0 קרוב יותר ל- \bar{X} .
4. האומד לשונות הטעויות - $\hat{\sigma}_u$ קטן יותר.

מדד טיב ההתאמה R^2 :

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

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

(מהשונות של Y).

נרצה פרופורציית שונות מוסברת קרובה ככל האפשר ל-1.

אחוז השונות המוסברת: $R^2 \cdot 100$.

מבחן t :

מבחן זה נעשה על מנת לבדוק האם מקדם הרגרסיה מובהק.

$$. t_{sr} = \frac{\hat{\beta}_1 - \beta_{1,0}}{\hat{s.e.}(\hat{\beta}_1)} \sim t_{c(n-2)} : \text{סטטיסטי המבחן}$$

$$s.e.(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{SXX}}$$

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

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

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

רווח סמך לאמידת β :

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

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

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:
- רק גודל הדירה משפיע על מחיר הדירה באופן שיטתי. כל שאר הדברים המשפיעים על מחיר הדירה הם אקראיים ולא ניתנים לחיזוי.
 - 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	$X_6 = 120$	$Y_6 = 380$

על סמך הדוגמא הנ"ל חשבו את:

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

מבחן F:

- (2) בהמשך לדוגמא הנ"ל: בצעו מבחן F לבדיקת הקשר בין גודל הדירה למחירה ברמת מובהקות של 1%.

ביצוע תחזיות:

- (3) במדגם של 30 דירות מושכרות לסטודנטים ברדיוס של עד 2 ק"מ מסביב למכללה נחקר הקרש בין שכר דירה למספר הסטודנטים הגרים בדירה. להלן התוצאות:

Descriptive Statistics			
	Mean	Std. Deviation	N
שכר הדירה	1386.7667	509.41027	30
מספר הסטודנטים	3.0000	1.31306	30

Model Summary ^b				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.602 ^a	.362	.339	414.05503

a. Predictors: (Constant), number of students
b. Dependent Variable: rent

ANOVA ^b						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2726579.520	1	2726579.520	15.904	.000 ^a
	Residual	4800363.847	28	171441.566		
	Total	7526943.367	29			

a. Predictors: (Constant), number of students
b. Dependent Variable: rent

Coefficients ^a					
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		

- א. חשב אומדן נקודתי לשכר הדירה אותו ישלמו סטודנטים החולקים את הדירה עם שותף אחד בלבד.
- ב. אמוד את שכר הדירה הממוצע שישלמו סטודנטים החולקים את הדירה עם שותף אחד בלבד, ברמת ביטחון של 95%.
- ג. אמוד את שכר הדירה שישלם סטודנט יחיד החולק את הדירה עם שותף אחד בלבד, ברמת ביטחון של 95%.

מדד טיב התאמה:

- (4) חשבו את אחוז השונות המוסברת של מחיר הדירה על ידי גודלה.

מבחן t:

- (5) 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?

רב"ס למקדם השיפוע:

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

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

תרגול מסכם:

8) בפיצוויית "שלמה המלך" חושדים כי מספר הלקוחות המבקרים בפיצווייה 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:

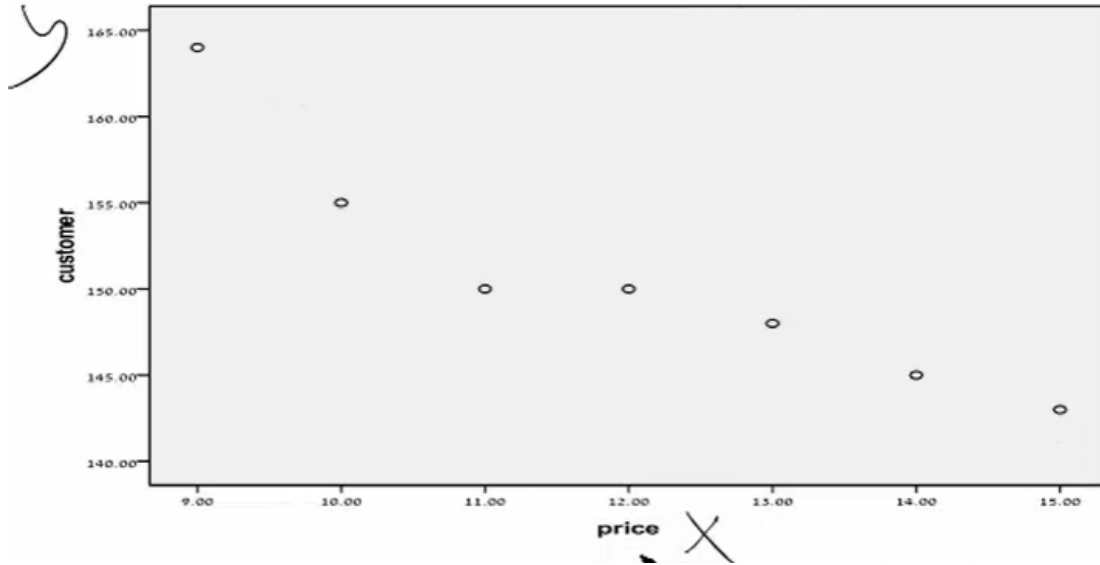
שבוע 7	שבוע 6	שבוע 5	שבוע 4	שבוע 3	שבוע 2	שבוע 1	
9	10	11	12	13	14	15	מחיר הבירה
164	155	150	150	148	145	143	כמות הלקוחות

- 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. בצעו בדיקה גראפית של אקראיות השאריות.
- ה. חשבו את אחוז השונות המוסברת. מה משמעותה?
- ו. בצעו חיזוי לכמות הלקוחות אם מחיר הבירה יהיה 16 ש"ח. 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 בפיצווייה 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:

9) על סמך הנתונים של השאלה הקודמת התקבלו הפלטים הבאים:

דיאגרמת הפיזור (scatter plot):



סטטיסטיקה תיאורית (descriptive statistics):

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

פלט מקדם המתאם (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

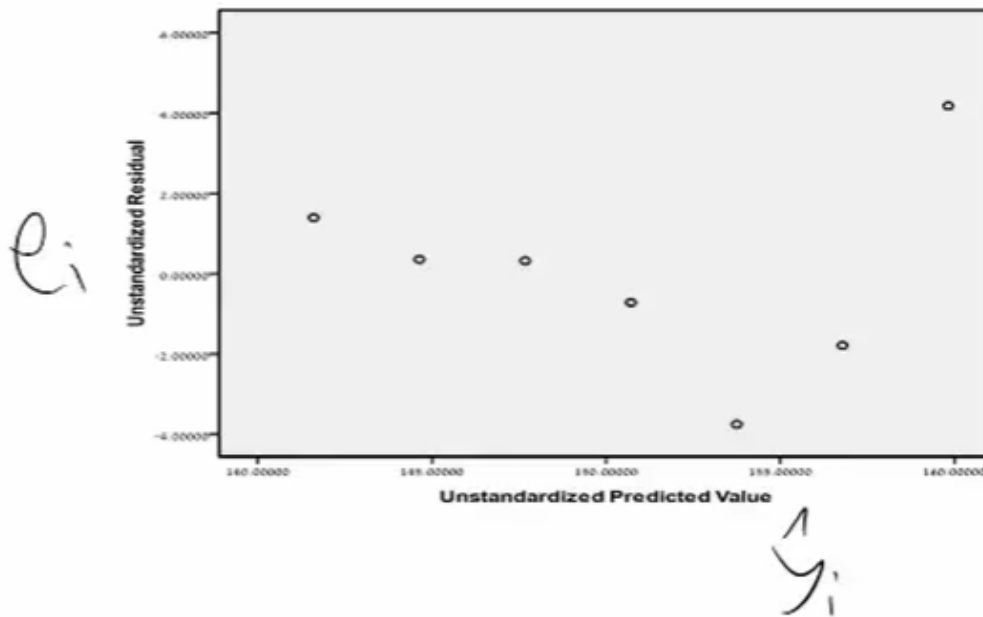
פלט מקדמי הרגרסיה (coefficients) :

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) יש עדות לקשר מובהק. $F_{st} = 21.198$
- (3) א. $\hat{Y}_0 = 1153.247$. ב. $p(957.4 \leq \mu_{Y_{X=2}} \leq 1349.08) = 0.95$. ג. $p(282.94 \leq Y_{X=2} \leq 2023.55) = 0.95$
- (4) $R^2 = 0.974$
- (5) א. $t_{stt} = 4.604$. ב. יש עדות לכך. ג. $pvalue < 0.001$
- (6) $t_{stt}^2 = 147$, $pvalue < 0.001$
- (7) $p(2.061 \leq \beta \leq 4.519) = 0.99$
- (8) 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. $p(-4.36 \leq \beta \leq -1.709) = 0.95$
- י. ראו סרטון.
- (9) א. $\hat{y}_i = 187.143 - 3.036x_i$. ב. $r_{xy} = 0.935$
- ג. $MSE = 7.479$. ד. לא. ה. $R^2 = 0.874$
- ו. $F = 34.503$. ז. $t = -5.874$. ח. $pvalue = 0.002$
- ט. ראו סרטון.