סטטיסטיקה יישומית

פרק 7 - רגרסיה פשוטה

תוכן העניינים

רגרסיה פשוטה:

רקע:

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

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

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

כאשר:

β_0 הוא החותך.

β_1 הוא שיפוע.

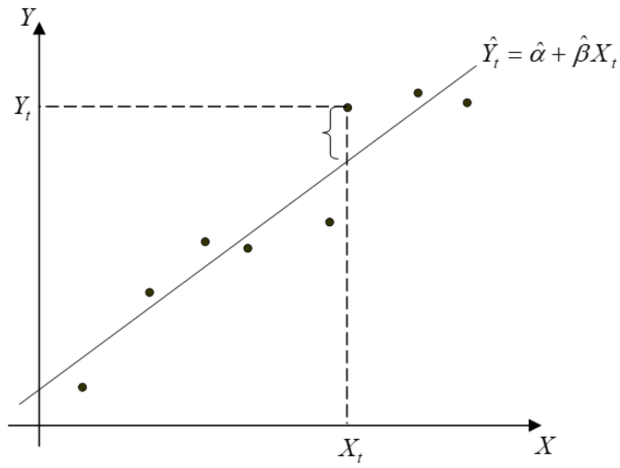
ε_i הינו גורם הטעות מסביב לקו הליניארי.

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

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

מבחני המובהקות:

השערות: $\begin{matrix} H_0 : \beta = 0 \\ H_1 : \beta \neq 0 \end{matrix}$.

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: $\underset{S_Y^2}{SST} = \underset{r^2 S_Y^2}{SSR} + \underset{(1-r^2)S_Y^2}{SSE}$.

טבלת ניתוח שונות (טבלת ANOVA):

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

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	$t_{statistic}=\dfrac{\hat{\beta}_1-\beta_{1,0}}{\overset{\wedge}{s.e.}\left(\hat{\beta}_1\right)=\sqrt{\dfrac{\hat{\sigma}^2}{SXX}}}=\dfrac{r^2\sqrt{n-2}}{\sqrt{1-r^2}}$		
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	$P\left(t_{n-2}>t_{statistic}\right)$	$P\left(t_{n-2}>t_{statistic}\right)$	$2*P\left(t_{n-2}>\lvert t_{statistic}\rvert\right)$

- 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) $=1-\alpha$.

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

מדד טיב ההתאמה R^2:

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

(X מסביר את כל השונות של Y): $0\leq R^2=\dfrac{SSR}{SST}=1-\dfrac{SSE}{SST}\leq 1$ (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 : $\hat{\mu}_0=\hat{y}_0=\hat{\alpha}+\hat{\beta}x_0$.

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 :

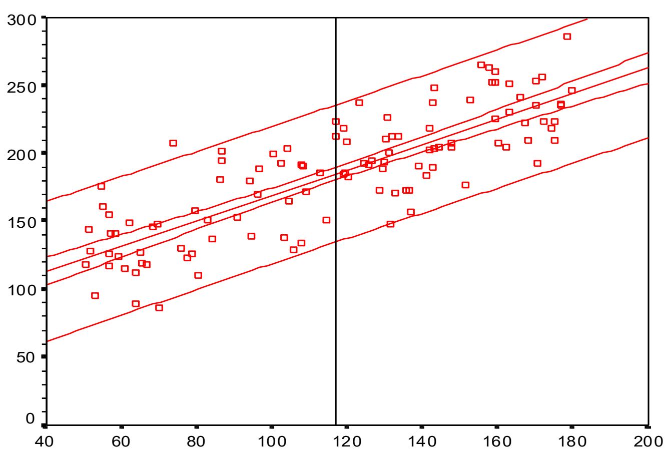
רב"ס שמטרתו לאמוד את כל טווח ערכי Y (y_0) עבור ערך X מסוים (x_0).

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 מסי 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 B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B 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.

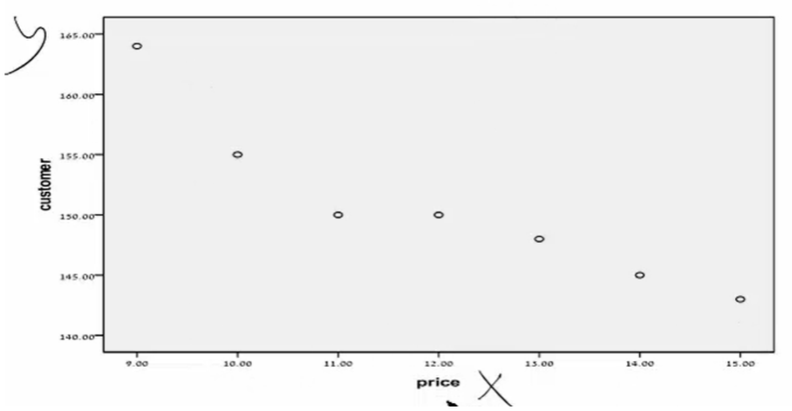
ט. אמדו את מקדם הרגרסיה ברמת סמך של 0.95. השוו את התוצאה עם הסעיף הקודם.

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

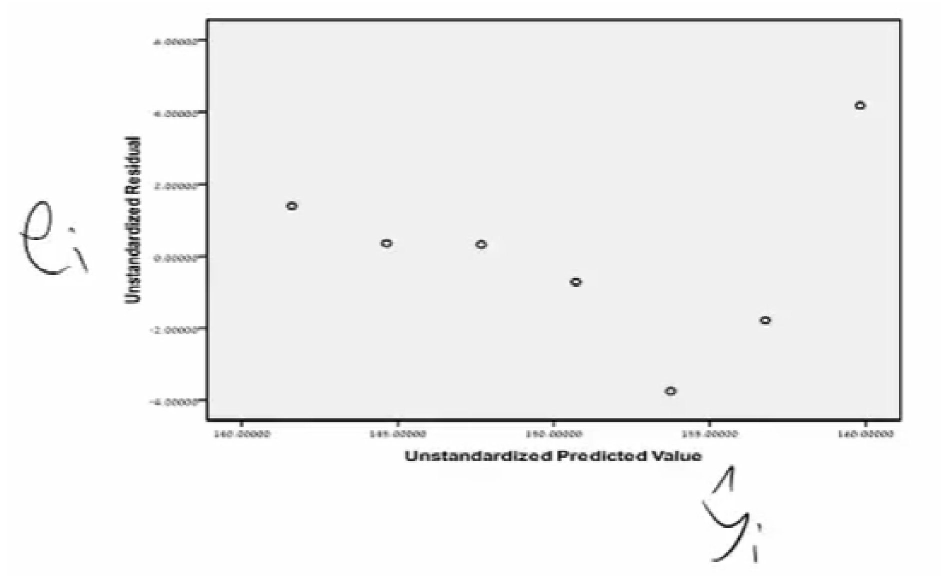
פלט מקדמי הרגרסיה (coeffitients):

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.