תרגיל בית 13

<u>שאלה 1 – שאלה מסכמת בנושא רגרסיה ליניארית</u>

במטרה לחקור את הגורמים הקובעים את רמת השימוש בחשמל בארץ, נאספו נתונים על פני 28 תקופות.
לצורך בניית המודל נקבעו המשתנים הבאים:

Y – צריכת חשמל במיליוני קוט״ש

X_1 – הכנסה ריאלית לנפש

X_2 – גודל האוכלוסיה (באלפים)

D – משתנה דמי המקבל את הערך 1 בעונת החורף ואת הערך 0 בכל עונה אחרת

א. בשלב ראשון נאמדה משוואת הרגרסיה הבאה:

$$\hat{Y} = 7578 + 5.16X_1$$
$$(580) \quad (0.413)$$
$$\sum_i e_i^2 = 2615$$

הערכים בסוגריים הם האומדנים לסטיות התקן של המקדמים.
האם הרגרסיה מובהקת ברמת מובהקות 0.05?

ב. בשלב שני נאמדה המשוואה הבאה <u>מתוך אותם הנתונים</u>:

$$\hat{Y} = 181 + 0.847X_1 + 2.46X_2$$
$$(114) \quad (0.06) \quad (0.035)$$
$$\sum_i e_i^2 = 13$$

חשב את מקדם ההסבר המרובה R^2.

ג. בשלב שלישי נאמדה המשוואה הבאה <u>מתוך אותם הנתונים</u>:

$$\hat{Y} = 82.68 + 0.78X_1 + 2.56X_2 + 39.59D$$
$$(77) \quad (0.04) \quad (0.024) \quad (6.93)$$
$$\sum_i e_i^2 = 5.6$$

האם ההבחנה בין עונת החורף לשאר עונות השונה הוסיפה הסבר מובהק להשתנות של צריכת החשמל מעבר להסבר של ההכנסה ושל גודל האוכלוסייה ברמת מובהקות 0.05? נמק.

ד. בחן בו זמנית את ההשערה שגודל האוכלוסייה ועונות השנה אינם תורמים להסבר השונות של צריכת החשמל מעבר להסבר שהתקל ע״י ההכנסה ברמת מובהקות 0.05.

תשובה

שתי הבחנות שילוו אותנו לאורך הפתרון:

- מכיוון שהמודלים שמוצגים בכל הסעיפים נבנו על בסיס אותו סט נתונים, בכולם SST זהה (משום ש-SST הוא מאפיין של ערכי המשתנה המוסבר בנתונים ולא של הרגרסיה).
- $\sum_i e_i^2 = \text{SSE}$ (בהגדרה)

א. זוהי רגרסיה ליניארית פשוטה. ניתן לבדוק את המובהקות שלה או עם מבחן t על המקדם b_1 או באמצעות מבחן F. מכיוון שגם המקדם b_1 וגם סטיית התקן שלו נתונים לנו, קל להשתמש פה במבחן t:

$$T_{b_1} = \frac{b_1}{s_{b_1}} = \frac{5.16}{0.413} = 12.49 > t_{0.975}^{26} = 2.056$$

הרגרסיה מובהקת בר״מ 0.05.

ב. נחשב את SST לפי הנתונים ממודל הרגרסיה בסעיף א׳:

$$0.413 = s_{b_1} = \sqrt{\frac{SSE/_{n-2}}{SS_x}} = \sqrt{\frac{\frac{2615}{26}}{SS_x}} \rightarrow SS_x = 589.655$$

$$SSR = b_1^2 SS_x = 5.16^2 \cdot 589.655 = 15699.924$$

$$SST = SSR + SSE = 15699.924 + 2615 = 18314.924$$

זה SST של המודל מסעיף א׳, אבל כאמור הוא זהה לכל המודלים.

לכן, גם במודל מסעיף ב׳ מתקיים $SST = 18314.924$. נתון $SSE = 13$.

לכן: $R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{13}{18314.924} = 0.9993$.

ג. השאלה היא בעצם האם התרומה של המשתנה D למודל מובהקת. נבין זאת באמצעות מבחן t:

$$T_D = \frac{39.59}{6.93} = 5.7 > 2.064 = t_{0.975}^{28-3-1}$$

ולכן דוחים את השערת האפס – כלומר ההבחנה בין עונת החורף לשאר העונות הוסיפה הסבר מובהק בר״מ 0.05.

ד. כעת עוסקים בתוספת התרומה של קבוצת משתנים מקריים (D ו-X_2) להסבר השונות, ואת זה ניתן לבחון באמצעות מבחן F חלקי:

$$F = \frac{(2615 - 5.6)/_2}{5.6/_{24}} = 5591.57 > 3.4 = f_{0.95}^{2,24}$$

ולכן דוחים את השערת האפס, כלומר תוספת התרומה מובהקת בר״מ 0.05.

שאלה 2

בעל רשת בתי קולנוע מבקש להעריך את הכנסתו השבועית כפונקצייה של הוצאות הפרסום. יש להניח כי כל ההנחות הקלאסיות של המודל תקפות. לצרכי התרגיל נלקח מדגם בן 8 שבועות בלבד:

הכנסה (1000$)	פרסום בטלוויזיה (1000$)	פרסום בעיתון (1000$)
96	5	1.5
90	2	2
95	4	1.5
92	2.5	2.5
95	3	3.3
94	3.5	2.3
94	2.5	4.2
94	3	2.5

א. חשב את משוואת הרגרסיה הליניארית הפשוטה להכנסה, כאשר הפרסום בטלוויזיה הוא המשתנה הבלתי תלוי היחיד. חשב באופן ידני או ב r($\boldsymbol{y = 88.64 + 1.603x}$)

ב. כיצד תעריך את ההכנסה אם ההוצאות על פרסום בטלוויזיה הן 3500$ ועל פרסום בעיתון 1800$?

(100)

ג. עבור המודל מסעיף ב' נמצאו הנתונים הבאים: $SSR = 23.435, SST = 25.5$.

(1) חשב באופן ידני את R^2 ואת $R^2_{adjusted}$. **($\boldsymbol{R^2 = 0.9190, R^2_{adj.} = 0.8866}$)**

(2) במודל מסעיף א' מתקיים $R^2 = 0.653, R^2_{adjusted} = 0.595$. האם כדאי להשתמש ברגרסיה הפשוטה או ברגרסיה המרובה להערכת ההכנסה? הסבר.

ד. האם המודלים מסעיף א' ומסעיף ב' מובהקים? בדוק בר"מ 0.01. **(שניהם מובהקים)**

שאלה 2

א. נחשב באופן ידני (בשביל התרגול...):

$$SS_{xy} = \sum_i x_i y_i - n\bar{x}\bar{y} = 2401 - 8 \cdot 93.75 \cdot 3.1875 = 10.375$$

$$SS_x = \sum_i x_i^2 - n(\bar{x})^2 = 87.75 - 8 \cdot 3.1875^2 = 6.46875$$

$$b_1 = \frac{SS_{xy}}{SS_x} = 1.603$$

$$b_0 = \bar{y} - b_1\bar{x} = 88.64$$

לשם השוואה, פלט הרגרסיה מאקסל:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.807807408
R Square	0.652552809
Adjusted R Square	0.594644943
Standard Error	1.215175116
Observations	8

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	16.64009662	16.64009662	11.26881134	0.015288079
Residual	6	8.859903382	1.476650564		
Total	7	25.5			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	88.63768116	1.582367131	56.01587609	2.174E-09	84.76576827	92.50959405	84.76576827	92.50959405
TV	1.603864734	0.47778079	3.356905024	0.015288079	0.434777257	2.772952212	0.434777257	2.772952212

ב. נבצע רגרסיה מרובה כאשר המשתנה התלוי הוא ההכנסה והמשתנים הב"ת הם הוצאות פרסום בטלוויזיה והוצאות פרסום בעיתון:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.958663444
R Square	0.9190356
Adjusted R Square	0.88664984
Standard Error	0.642587303
Observations	8

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	23.43540779	11.7177039	28.37776839	0.001865242
Residual	5	2.064592208	0.412918442		
Total	7	25.5			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	83.23009169	1.573868952	52.88247894	4.57175E-08	79.18433275	87.27585063
TV	2.290183621	0.304064556	7.531899313	0.000653232	1.508560796	3.071806446
Newspaper	1.300989098	0.320701597	4.056696662	0.009760798	0.476599398	2.125378798

$$\hat{y}(\mathrm{X}_1 = 3.5, \mathrm{X}_2 = 1.8) = 83.23 + 3.5 \cdot 2.29 + 1.3 \cdot 1.8 \cong 100$$

סעיף ג'

(1)

$$R^2 = \frac{SSR}{SST} = \frac{23.435}{25.5} = 0.9190$$

$$R^2_{adjusted} = 1 - \frac{\frac{SSE}{n-k-1}}{\frac{SST}{n-1}} = 1 - \frac{0.4129}{3.643} = 0.8866$$

(2) גם R^2 וגם $R^2_{adjusted}$ של מודל הרגרסיה המרובה גבוהים יותר בהשוואה למקביליהם ממודל הרגרסיה הפשוטה. לכן ההסבר שמספקת הרגרסיה המרובה טוב יותר (גם כאשר "קונסים" על תוספת משתנה ההחלטה), ולכן עדיף להשתמש ברגרסיה המרובה.

ד. נבדוק את *P-Values* של מבחני *F*.

ברגרסיה הפשוטה, *PV=0.015* ולכן הרגרסיה לא מובהקת בר"מ 0.01.

ברגרסיה המרובה, *PV=0.0018* ולכן הרגרסיה כן מובהקת בר"מ 0.01.