תרגול 13 : חזרה למבחן

.1

X מתפלג אחיד רציף $X \sim U(a,b)$. בהינתן n ריאליזציות של X :

א. אמוד את a ו-b לפי שיטת הנראות המקסימלית.

פתרון

א. שיטת הנראות המקסימלית :

$$f(x_i) = \frac{1}{b-a} \qquad a \le x_i \le b$$

$$L(a,b) = \prod_{i=1}^{n} f(x_i) = \left(\frac{1}{b-a}\right)^n \qquad a \le x_i \le b$$

$$\ln\left(L(a,b)\right) = \ln\left(\frac{1}{b-a}\right)^n = n\ln(\frac{1}{b-a}) = -n\ln(b-a)$$

גזירה חלקית לפי a ו-b :

$$\frac{\partial n\, ln(\,b-a)}{\partial a} = -n \cdot \frac{1}{b-a} \cdot (-1) = \frac{n}{b-a}$$

$$\frac{\partial n\, ln(\,b-a)}{\partial b} = -n \cdot \frac{1}{b-a} \cdot 1 = -\frac{n}{b-a}$$

הנגזרת אינה מתאפסת, לא לפי a ולא לפי b!

אז מה עושים?

הבחנה (1) : כל התצפיות במדגם נמצאות בין a ל-b. כלומר : $a \le x_i \;\; \forall i$ ובפרט $a \le \min\{x_i\}$.

באופן דומה, $b \ge x_i \;\; \forall i$ ובפרט $b \ge \max\{x_i\}$.

הבחנה (2) : הנגזרת לפי a חיובית לכל a, לכן L ב-a.

לכן המקסימום של L מתקבל בערך הגדול ביותר האפשרי של $\frac{\partial n\, ln(b-a)}{\partial a}$, שהוא $a^* = \min\{x_i\}$.

הבחנה (3) : הנגזרת לפי b שלילית לכל b, לכן L יורדת ב-b.

לכן המקסימום של L מתקבל בערך הנמוך ביותר האפשרי של $\frac{\partial n\, ln(b-a)}{\partial b}$, שהוא $b^* = \max\{x_i\}$.

סיכום: $\hat{b} = \max\{x_i\} \quad \hat{a} = \min\{x_i\}$

.2

כדי לאמוד את יעילותה של תוכנית ניסיונית להגדלת ההיגיינה, הופעלה התוכנית ב-9 מפעלים. בכל מפעל חושב המספר השנתי של מספר שעות העבודה שאבדו בגלל ימי מחלה. מספר זה נרשם בכל אחד מהמפעלים שהשתתפו במחקר, בשנה שלפני הפעלת התוכנית ובשנה שלאחריה, והתקבלו התוצאות הבאות :

מס' שעות העבודה שאבדו	
לפני הפעלת התוכנית	**אחרי הפעלת התוכנית**
50	42
87	76
37	36
141	130
59	60
65	55
24	26
88	86
25	29

האם התוכנית הוכיחה את עצמה כיעילה ברמת מובהקות 0.05 מבחינת הקטנת אובדן שעות עבודה, בהנחה שמספר השעות מתפלג נורמלית?

<u>פתרון</u>

מדובר במדגמים תלויים מכיוון שהמדגם בוצע על <u>אותם</u> מפעלים.

מס' שעות עבודה שאבדו		
לפני הפעלת התוכנית - X_i	אחרי הפעלת התוכנית - Y_i	ההפרש - $D_i = X_i - Y_i$
50	42	8
87	76	11
37	36	1
141	130	11
59	60	-1
65	55	10
24	26	-2
88	86	2
25	29	-4

μ_1 - תוחלת אובדן השעות במצב המקור. μ_2 - תוחלת אובדן השעות לאחר הפעלת התוכנית.

ההשערות הן :
$$H_0 : \mu_1 = \mu_2 \qquad H_1 : \mu_1 > \mu_2$$
או
$$H_0 : \mu_1 - \mu_2 = 0 = d_0 \qquad H_1 : \mu_1 - \mu_2 > 0 = d_0$$

כלל ההחלטה : דחה אם $\bar{d} > d_0 + t_{1-\alpha}^{(n-1)} \cdot \frac{s_d}{\sqrt{n}}$

דחה אם : $\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{8+11+1+...}{9} = 4 \,,\; s_d^2 = \frac{\sum_{i=1}^{n}(d_i-\bar{d})^2}{9-1} = \frac{\sum_{i=1}^{n} d_i^2 - n\bar{d}^2}{8} = 36$

$\bar{d} > 0 + t_{1-0.05}^{(9-1)} \cdot \frac{\sqrt{36}}{\sqrt{9}} = 1.86 \cdot 2 = 3.72$.

התקבל $\bar{d}$ שווה ל-4 לכן נדחה את השערת האפס - התוכנית אכן מפחיתה את אובדן שעות העבודה ברמת מובהקות 5%.

.3

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>:

$$\hat{Y} = 181 + 0.847X_1 + 2.46\mathrm{X}_2$$
$$(114) \quad (0.06) \quad (0.035)$$
$$\sum_i e_i^2 = 13$$

חשב את מקדם ההסבר המרובה R^2.

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

$$\hat{Y} = 82.68 + 0.78X_1 + 2.56\mathrm{X}_2 + 39.59D$$
$$(77) \quad (0.04) \quad (0.024) \quad (6.93)$$
$$\sum_i e_i^2 = 5.6$$

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.

פתרון

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 (X_2 ו-D) 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.