תרגול 11
רגרסיה ליניארית פשוטה

רגרסיה, רגרסיה ליניארית, ורגרסיה ליניארית פשוטה

נניח שאספנו תצפיות של משתנה מקרי מסויים Y אותו אנחנו מעוניין לחקור. לתצפיות שלו יש פיזור, והשאלה היא **מה גורם לפיזור הזה**.

ההסבר שמציעה **שיטת הרגרסיה**: העובדה שהמשתנה Y קיבל ערכים שונים נובעת מכך שהערכים שנקבעו עבור קבוצת משתנים $X_1, X_2, \dots, X_n$ משפיעים על הערך של Y.

במילים אחרות: $Y \approx f(X_1, X_2, \dots, X_n)$.

Y נקרא משתנה מוסבר = תלוי = מנובא. **הוא משתנה מקרי.**

$X_1, X_2, \dots, X_n$ נקראים משתנים מסבירים = בלתי-תלויים = מנבאים. **הם בשליטתנו (לא מ"מ).**

הסיבה שבגללה אנו מקבלים רק **קירוב**, היא שגם כאשר כל משתנה מסביר מקבל ערך מסויים, עדיין עשויים לקבל ערכים שונים של המשתנה המוסבר. באופן כללי, הרגרסיה מייצרת פונקצייה f שמתארת את **התוחלת המותנית של Y**, בהינתן ערכי המשתנים המסבירים.

$$E[Y|X_1 = x_1, X_2 = x_2, \dots, X_n = x_n] = f(X_1, X_2, \dots, X_n)$$

הפער בין התוחלת המותנית של Y בהינתן ערכי המשתנים המסבירים לערך שהתקבל בפועל הוא **רעש אקראי**.

כאשר f היא פונקציה ליניארית, השיטה נקראת **רגרסיה ליניארית**.

כאשר קיים בנוסף רק משתנה מסביר אחד, השיטה נקראת **רגרסיה ליניארית פשוטה**.

מסמנים: β_0 – החותך של המשוואה הליניארית; β_1 – השיפוע של המשוואה הליניארית.

הנחות מודל רגרסיה ליניארית פשוטה

1. $E[Y|X = x] = \beta_0 + \beta_1 x$
2. עבור תצפית מסויימת: $Y_i = (\beta_0 + \beta_1 x) + \varepsilon_i$. ε_i מייצג את הרעש האקראי.
3. לכל תצפית i, $\varepsilon_i \sim N(0, \sigma^2)$ וב"ת ברעשים האחרים.
4. השונות אחידה לכל התצפיות, ללא תלות בערכי המשתנה המסביר.

סיכום ההנחות: $(Y|X = x) \sim N(\beta_0 + \beta_1 x, \sigma^2)$.

השאלות המרכזיות שנשאל את עצמנו: מהם המקדמים שיגרמו ל"הסבר הטוב ביותר" על סמך מדגם מסויים, האם ההסבר מספיק טוב, והאם הוא מובהק?

הבעיה: הפרמטרים של האוכלוסיה אינם ידועים...

אילו היינו מודדים את כל האוכלוסיה, היינו יכולים לחשב ולמצוא את הקו הישר $\beta_0+\beta_1 x$.

מכיוון שאנחנו מתבססים על **מדגם** מתוך האוכלוסיה נצטרך לאמוד את הפרמטרים β_0,β_1 על סמך n תצפיות, כל תצפית היא הזוג (x_i, y_i).

האומד ל- β_0 יסומן b_0, והאומד ל- β_1 יסומן b_1.

לכן קו הרגרסיה שיתקבל **מהמדגם**: $\boxed{\hat{y}_i = b_0 + b_1 x_i}$

<u>בניית משוואת רל"פ - שיטת הריבועים הפחותים (Ordinary Least Squares)</u>

הסטייה של תצפית במדגם ביחס לקו שאנחנו מייצרים: $e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i$

נגדיר כמדד את סכום ריבועי השגיאות: $Min\sum_{i=1}^{n} e_i^2 = Min\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2$

היינו רוצים למצוא אומדים b_0, b_1 שימזערו את סכום ריבועי השגיאות. ע"י גזירה חלקית מקבלים:

$$\boxed{b_1 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \equiv \frac{SS_{xy}}{SS_x} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \qquad b_0 = \bar{y} - b_1\bar{x}}$$

<u>תכונות קו הריבועים הפחותים</u>

- הקו עובר דרך נקודת הממוצעים $(\bar{x}, \bar{y})$.
- האומדים b_0 ו-b_1 הם חסרי הטיה. כלומר: $E(b_1)=\beta_1 \qquad E(b_0)=\beta_0$

<u>תרגיל 1 – מתוך מבחן</u>

מהנדס כימי דוגם את האפקט של הטמפרטורה על אחוז התוצרת המתקבלת בתהליך כימי:

טמפרטורה	100	110	120	130	140	150	160	170	180	190
אחוז תוצרת	45	51	54	61	66	70	74	78	85	89

א. חשב את מקדמי הרגרסיה

$n=10 \qquad \sum_{x_i} x_i = 1450 \qquad \sum_{y_i} y_i = 673 \qquad \sum_{x_i} x_i^2 = 218500 \qquad \sum x_i y_i = 101570$

$\bar{x}=145 \qquad \bar{y}=67.3$

$SS_{xy} = \sum x_i y_i - n\bar{x}\cdot\bar{y} = 101570 - 10\cdot 67.3\cdot 145 = 3985$

$SS_x = \sum_{x_i} x_i^2 - n\bar{x}^2 = 218500 - 10\cdot 145^2 = 8250$

$\boxed{b_1 = \frac{SS_{xy}}{SS_x} = \frac{3985}{8250} = 0.48303}$ $\qquad$ $\boxed{b_0 = \bar{y} - b_1\bar{x} = 67.3 - 0.48303\cdot 145 = -2.73939}$

משוואת הרגרסיה שנאמדה: $\boldsymbol{\hat{y} = -2.74 + 0.48\cdot x}$

משפט פירוק השונויות

סכום הסטיות של התצפיות מהממוצע שלהן $\bar{y}$

$$\boldsymbol{SST} \equiv SS_y = \sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

סך הסטייה מהממוצע $\bar{y}$ שהרגרסיה מצליחה להסביר ("סך השונות המוסברת")

$$\boldsymbol{SSR} \equiv \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \overset{\text{ב-רל"פ בלבד}}{=} b_1^2 SS_x$$

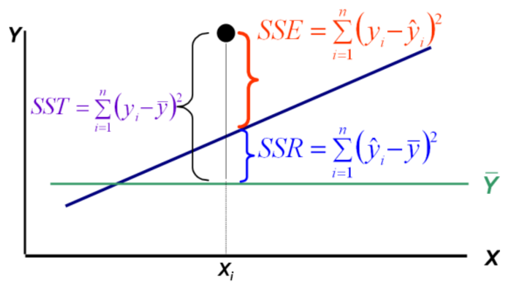
סך הסטייה מהממוצע $\bar{y}$ שהרגרסיה אינה מצליחה להסביר ("סך השונות הבלתי-מוסברת", "סך הרעש במודל")

$$\boldsymbol{SSE} \equiv \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

שימו לב: SST הוא מאפיין של **הנתונים**, בעוד ש-SSR ו-SSE יכולים להשתנות כתלות במשוואת הרגרסיה שאנחנו בונים!

ניתן להוכיח שמתקיים הקשר הבא:

$$\boxed{\boldsymbol{SST = SSE + SSR}}$$



- ככל שקו הרגרסיה מתאים יותר לנתוני המדגם, SSR יותר גדול ו-SSE יותר קטן.
- אם הצלחנו לייצר התאמה מושלמת, האומד לתצפית $\hat{y}_i$ שווה בדיוק לתצפית y_i עבור כל התצפיות, ואז: $SSR = SST, SSE = 0$.

- מגדירים את "מדד ההסבר" = **"אחוז השונות המוסברת"**: $\boldsymbol{R^2} \equiv \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

זהו מדד למידת ההתאמה של קו הרגרסיה לנתונים: R^2 מקבל ערכים בין 0 ל-1, כאשר 0 משמעו שקו הרגרסיה אינו תואם כלל לנתונים, ו-1 פירושו שהנתונים "יושבים" על הקו באופן מושלם.

- **המתאם בין שני משתנים מקריים:**

במדגם

$$\hat{\rho} \equiv r = \frac{SS_{xy}}{\sqrt{SS_x \cdot SS_y}}$$

$\rightarrow$

באוכלוסיה

$$\rho = \frac{cov(X,Y)}{\sigma_x\,\sigma_y}$$

ברגרסיה ליניארית __פשוטה__ מתקיים הקשר: $r^2 = R^2$

R^2 מכונה בד"כ "מקדם ההסבר" ו-r "מקדם המתאם".

- **פוטנציאל בלבול:** לעתים קוראים ל-R^2 "מקדם המתאם המרוּבֶּה".

<u>תרגיל 1 - המשך</u>

ב. חשב את מקדם ההסבר R^2. מה ניתן לומר על הקשר בין הטמפרטורה לתהליך הכימי?

$$SSR = b_1^2 \cdot SS_x = 0.48303^2 \cdot 8250 = 1924.876, \qquad SST = SS_y = 1932.1$$

$$\Rightarrow R^2 = \frac{SSR}{SST} = \frac{1924.876}{1932.1} = 0.996261$$

R^2 מעל 99%, הקשר חזק מאוד. קו הרגרסיה שנאמד מסביר את נתוני המדגם בצורה מצויינת.

ג. חשב את מקדם המתאם במדגם (r)

$$r = \frac{SS_{xy}}{\sqrt{SS_x \cdot SS_y}} = \frac{3985}{\sqrt{8250 \cdot 1932.1}} = 0.998129 = \sqrt{R^2}$$

<u>אמידת σ^2 ושונויות האומדים b_0, b_1</u>

אומד חסר הטיה ל- σ^2:

$$s^2 \equiv \hat{\sigma}^2 \equiv MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{SST - SSR}{n-2} = \frac{SS_y - b_1^2 SS_x}{n-2}$$

שונות האומדים b_0, b_1 תלויה ב- σ^2. ע"י הצבת s מתקבלים אומדים חסרי הטייה לשונות של b_0, b_1:

$$S_{b_1} = \frac{s}{\sqrt{SS_x}} \qquad S_{b_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_x}}$$

<u>איכות מודל הרגרסיה</u>

<u>בניית רווחי סמך לפרמטרים שנאמדו</u>

רב"ס דו-צדדי ל- β_0 ברמת סמך $1-\alpha$:

$$\beta_0 \in \left(b_0 - s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_x}}\, t_{1-\frac{\alpha}{2}}^{(n-2)} \quad , \quad b_0 + s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_x}}\, t_{1-\frac{\alpha}{2}}^{(n-2)} \right)$$

רב"ס דו-צדדי ל- β_1 ברמת סמך $1-\alpha$:

$$\beta_1 \in \left(b_1 - \frac{s}{\sqrt{SS_x}}\, t_{1-\frac{\alpha}{2}}^{(n-2)} \quad , \quad b_1 + \frac{s}{\sqrt{SS_x}}\, t_{1-\frac{\alpha}{2}}^{(n-2)} \right)$$

כדי לבנות רב"סים חד צדדיים, נחליף את $1-\frac{\alpha}{2}$ ל- $1-\alpha$ ונעדכן את הגבול המתאים ל $\pm\infty$

<u>תרגיל 1 - המשך</u>

ד. מצא רווח סמך ברמת סמך 95% ל- β_1

$$SS_y = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 1932.1 \Rightarrow s^2 = \frac{SS_y - b_1^2 SS_x}{n-2} = \frac{1932.1 - 0.48^2 \cdot 8250}{8} = 0.903$$

$$\beta_1 \in \left(b_1 \pm \frac{s}{\sqrt{SS_x}}\, t_{1-\frac{\alpha}{2}}^{n-2} \right) \Rightarrow 0.483 \pm 0.0105 \cdot 2.306 \Rightarrow \beta_1 \in (0.459, 0.507)$$

בדיקת השערות על ערכי β_0, β_1 ברמת מובהקות α

	מבחן חד צדדי		מבחן דו צדדי
מערכת ההשערות	$H_0 : \beta_0 = \mu_0$ $H_1 : \beta_0 > \mu_0$	$H_0 : \beta_0 = \mu_0$ $H_1 : \beta_0 < \mu_0$	$H_0 : \beta_0 = \mu_0$ $H_1 : \beta_0 \neq \mu_0$
אזור דחייה	$T_{b_0} = \frac{b_0 - \mu_0}{S_{b_0}} > t_{1-\alpha}^{n-2}$	$T_{b_0} = \frac{b_0 - \mu_0}{S_{b_0}} < -t_{1-\alpha}^{n-2}$	$\left\|T_{b_0}\right\| = \left\|\frac{b_0 - \mu_0}{S_{b_0}}\right\| > t_{1-\frac{\alpha}{2}}^{n-2}$

	מבחן חד צדדי		מבחן דו צדדי
מערכת ההשערות	$H_0 : \beta_1 = \mu_1$ $H_1 : \beta_1 > \mu_1$	$H_0 : \beta_1 = \mu_1$ $H_1 : \beta_1 < \mu_1$	$H_0 : \beta_1 = \mu_1$ $H_0 : \beta_1 \neq \mu_1$
אזור דחייה	$T_{b_1} = \frac{b_1 - \mu_1}{S_{b_1}} > t_{1-\alpha}^{n-2}$	$T_{b_1} = \frac{b_1 - \mu_1}{S_{b_1}} < -t_{1-\alpha}^{n-2}$	$\left\|T_{b_1}\right\| = \left\|\frac{b_1 - \mu_1}{S_{b_1}}\right\| > t_{1-\frac{\alpha}{2}}^{n-2}$

<u>בדיקת השערות על מובהקות הרגרסיה</u>

מערכת ההשערות:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

<u>דרך ראשונה: מבחן t על הפרמטר β_1</u>

סטטיסטי המבחן:

$$T_{b_1} = \frac{b_1 - 0}{S_{b_1}} = \frac{b_1}{s/\sqrt{SS_x}} \sim t(n-2)$$

כלל ההחלטה עבור רמת מובהקות α: דחה את H_0 אם $\left|T_{b_1}\right| > t_{1-\frac{\alpha}{2}}^{n-2}$

- דרך שקולה לביצוע מבחן t היא לבנות רב״ס דו״צ ל-β_1 ברמת אמינות $1-\alpha$.
  אם הערך 0 כלול ברב״ס, מקבלים את השערת האפס (כלומר, אין קשר ליניארי).

<u>דרך שנייה: מבחן F</u>

סטטיסטי המבחן:

$$F = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/(n-2)}$$

כלל ההחלטה עבור רמת מובהקות α: דחה את H_0 אם $F > f_{1-\alpha}^{1,n-2}$

עבור **רגרסיה פשוטה בלבד**, ברמת מובהקות נתונה α מתקיים הקשר:

$$\boxed{F = (T_{b_1})^2}$$

<u>תרגיל 1 - המשך</u>

ה. בדוק את ההשערה $H_0: \beta_1 = 0$ כנגד $H_1: \beta_1 \neq 0$ ברמת מובהקות 5%. איך נקראת בחינה זו?

בחינה זו נקראת בחינת מובהקות הרגרסיה.

$$T_{b_1} = \frac{b_1 - 0}{s / \sqrt{SS_x}} = \frac{0.48303}{s / \sqrt{8250}}$$

$$s = \sqrt{\frac{SS_y - b_1^2 SS_x}{n-2}} = \sqrt{\frac{SS_y - 0.48303^2 \cdot 8250}{8}}$$

$$SS_y = \sum_{y_i} y_i^2 - n\bar{y}^2 = 47225 - 10 \cdot 67.3^2 = 1932.1$$

$$\Rightarrow s = 0.950279 \Rightarrow T_{b_1} = 46.16897$$

$$t^{n-2}_{1-\frac{\alpha}{2}} = t^{8}_{0.975} = 2.306$$

$46.16897 > 2.306$ לכן הרגרסיה מובהקת ברמת מובהקות 5%.

ו. בצע מבחן F למובהקות הרגרסיה, ברמת מובהקות של 5%

$$F = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/(n-2)} = \frac{SSR}{s^2} = \frac{1924.876}{0.90303} = 2131.574$$

$$f^{1,n-2}_{1-\alpha} = f^{1,8}_{0.95} = 5.32 \qquad \Rightarrow F > f^{1,n-2}_{1-\alpha}$$

לכן הרגרסיה מובהקת ברמת מובהקות 5%.

נשים לב שאכן מתקיים: $F_{stat} = (T_{stat})^2 \Rightarrow 2131.574 = (46.16897)^2$

חיזוי ערכי המשתנה המוסבר

לאחר שמצאנו את קו הרגרסיה, התחזית שלנו לערכו של המשתנה המוסבר Y כאשר $X = x_p$ היא פשוט: $y_p = b_0 + b_1 \cdot x_p$.

תחזית נכונה יותר צריכה להגדיר תחום ערכים אפשרי ל-y_p בהסתברות $1 - \alpha$ - כלומר, רב"ס!

- **רווח בר סמך לאומדן y_p בהינתן $X = x_p$ ברמת סמך $1-\alpha$:**

$$y_p \in \left((b_0 + b_1 \cdot x_p) \pm t^{n-2}_{1-\frac{\alpha}{2}} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_x}} \right)$$

- **רווח בר-סמך לתוחלת (המותנית) של y_p בהינתן $X = x_p$ ברמת סמך $1-\alpha$:**

$$E(y_p) \in \left((b_0 + b_1 \cdot x_p) \pm t^{n-2}_{1-\frac{\alpha}{2}} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_x}} \right)$$

ז. מצא אומדן לתוחלת אחוז התוצרת בטמפ' של 210 מעלות ברמת סמך של 95%.

$$E(Y / X = x_p) \in \left((b_0 + b_1 x_p) \pm t^{n-2}_{1-\frac{\alpha}{2}} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_x}} \right)$$

$$E(Y / X = 210) \in \left((-2.739 + 0.48303 \cdot 210) \pm t^{8}_{0.975} \cdot 0.95 \cdot \sqrt{\frac{1}{10} + \frac{(210 - 145)^2}{8250}} \right)$$

$$E(Y / X = 210) \in 98.7 \pm 2.306 \cdot 0.95 \cdot \sqrt{0.6121} \Rightarrow E(Y / X = 210) \in 98.7 \pm 1.714$$

ח. מצא אומדן לאחוז התוצרת בטמפ' של 210 מעלות ברמת סמך של 95%.

$$(Y / X = x_p) \in \left((b_0 + b_1 x_p) \pm t^{n-2}_{1-\frac{\alpha}{2}} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_x}} \right)$$

$$(Y / X = 210) \in \left((-2.739 + 0.48303 \cdot 210) \pm t^{8}_{0.975} \cdot 0.95 \cdot \sqrt{1 + \frac{1}{10} + \frac{(210 - 145)^2}{8250}} \right)$$

$$(Y / X = 210) \in 98.7 \pm 2.306 \cdot 0.95 \cdot \sqrt{1.6121} \Rightarrow (Y / X = 210) \in 98.7 \pm 2.78$$

<u>תרגיל 2</u>

חברת שיווק בודקת קשר בין מכירות שבועיות y (באלפי יחידות) לבין הוצאות פרסום x (בעשרות אלפי ₪). מדגם מקרי שנאסף מנתוני 30 שבועות, הראה את התוצאות הבאות:

$$\sum_{i=1}^{30} x_i = 180,\ \ \sum_{i=1}^{30} y_i = 210,\ \ \sum_{i=1}^{30} y_i^2 = 2436,\ \ \sum_{i=1}^{30} x_i^2 = 1680,\ \ \sum_{i=1}^{30}(y_i-\bar{y})(x_i-\bar{x}) = 750$$

א. מצא את משוואת הרגרסיה לחיזוי המכירות השבועיות על סמך הוצאות הפרסום

$$b_1 = \frac{\sum_{i=1}^{30}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{30} x_i^2 - n\bar{x}^2} = \frac{750}{1680-30\cdot\left(\frac{180}{30}\right)^2} = 1.25$$

$$b_0 = \bar{y} - b_1\bar{x} = \frac{210}{30} - 1.25\cdot 6 = -0.5 \Rightarrow \qquad \hat{y} = b_0 + b_1 x = 1.25x - 0.5$$

ב. האם ניתן לומר ברמת מובהקות של 5% שתוספת של 10,000 ש"ח להוצאות הפרסום מעלה את המכירות ביותר מאלף יחידות?

[הערה: שימו לב ליחידות - אלף יחידות משמען $y=1$, ו-10,000 ₪ מיוצגים ע"י $x=1$!]

מערכת ההשערות: $H_0: \beta_1 \le 1$, $H_1: \beta_1 > 1$ **כלל ההחלטה:** דחה אם $T_{b_1} = \frac{b_1-\mu_1}{s/\sqrt{SS_x}} > t_{1-\alpha}^{n-2}$.

$$s^2 = \frac{SSE}{n-2} = \frac{SS_y - b_1^2 SS_x}{n-2} = \frac{\sum_{y_i} y_i^2 - n\bar{y}^2 - b_1^2\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}{n-2} =$$

$$\frac{2436-30\cdot 7^2 - 1.25^2\cdot(1680-30\cdot 6^2)}{28} = 1.018$$

$$T_{b_1} = \frac{b_1-\mu_1}{s/\sqrt{SS_x}} = \frac{1.25-1}{\sqrt{1.018}/\sqrt{1680-30\cdot 6^2}} = 5.129$$

$$t_{1-\alpha}^{n-2} = t_{1-0.05}^{30-2} = t_{0.95}^{28} = 1.7$$

$5.129 > 1.7$ ולכן נדחה את השערת האפס ברמת מובהקות 5%.

(עוברים עכשיו לדף הנפרד עם טבלת הרגרסיה מ-Excel)

<u>תרגיל 3</u>

הועלתה השערה כי כמות המכירות של מוצר מסוים עולה ליניארית עם מספר אנשי המכירות המועסקים בחברה. להלן תוצאות חלקיות של פלט אקסל של ניתוח רגרסיה בין Y – כמות מכירות (במאות $) ו-X – מספר אנשי המכירות המועסקים.

SUMMARY OUTPUT			
Regression statistics			
Multiple R			
R Square			
Standard Error			
Observations	10		
ANOVA			
	Df	*SS*	*MS*
Regression		860.051	
Residual			
Total		1004.525	
	Coefficients	*Standard Error*	*t Stat*
Intercept	10.528	3.745	
X	0.953		6.901

(1) השלימו את הערכים החסרים בטבלה.

מילוי הטבלה מבוצע עפ״י ההסברים לגבי ניתוח טבלאות רגרסיה (בדף הנפרד שחולק).

הסברים מפורטים (עפ״י האותיות העבריות שליד כל ערך) בעמוד הבא.

א. **עמודת דרגות החופש**

בשורה הראשונה מופיע מספר המשתנים המסבירים k – ברגרסיה פשוטה תמיד שווה ל-1.
בשורה האחרונה מופיע מספר התצפיות פחות מספר המשתנים המסבירים – ברגרסיה פשוטה זה תמיד שווה ל-(n-1). את מספר התצפיות n לוקחים מהתא שנקרא Observations (השורה הרביעית מלמעלה). מקבלים n-1)=10-1=9).
השורה האמצעית היא ההפרש בין השורה האחרונה לראשונה: 9-1)=8).

ב. **עמודת SS**

עפ"י משפט פירוק השונויות מעמוד 3 מתקבל $SSE = SST - SSR$, ולכן הערך החסר שווה ל: $SSE = 1004.525 - 860.051 = 144.47$.

ג. **עמודת MS**

עפ"י הגדרה, $MSR = \left(\frac{SSR}{k}\right)$. ברגרסיה פשוטה תמיד k=1 ולכן $MSR = SSR = 860.051$.

באותו אופן, $MSE = \left(\frac{SSE}{n-k-1}\right)$. כלומר, $MSE = \frac{SSE}{8} = 18.059$.

ד. **R^2 (R Square) ו-r (Multiple R)**

עפ"י ההגדרה, $R^2 = \left(\frac{SSR}{SST}\right)$. ע"י הצבת הערכים מעמודת SS מקבלים $R^2 = 0.856$.

ברגרסיה פשוטה מתקיים $(r)^2 = R^2$ ולכן $r = \sqrt{R^2} = 0.925$.

ה. **סטיית התקן של הרגרסיה** (Standard Error למעלה)

s^2 היא שונות הרגרסיה ומתקיים $s^2 = MSE$.

ע"י הצבת ערך MSE מעמודת MS מקבלים: $s = \sqrt{s^2} = \sqrt{MSE} = 4.25$

ו. **מקדמי הרגרסיה**

הערכים חושבו בעזרת הקשרים הבאים:

$$T_{b_0} = \frac{b_0}{S_{b_0}} = \frac{10.528}{3.745} = 2.811$$

$$T_{b_1} = \frac{b_1}{S_{b_1}} \Rightarrow S_{b_1} = \frac{b_1}{T_{b_1}} = \frac{0.953}{6.901} = 0.138$$

(2) בחנו את מובהקות רגרסיה ברמת מובהקות 0.05, והדגימו את הקשר: $T_{b_1} = \frac{b_1}{S_{b_1}} = \sqrt{F_{stat}}$

$$F = \frac{MSR}{MSE} = \frac{860.051}{18.06} = 47.62, \qquad f_{1-\alpha}^{1,n-2} = f_{0.95}^{1,8} = 5.32 \Rightarrow \qquad F > f_{1-\alpha}^{1,n-2}$$

הרגרסיה מובהקת. כמו כן: $\sqrt{47.62} = 6.901$.

תרגיל בית 11

<u>שאלה 1</u>

מוצר הדגל בקפיטריה של הפקולטה להנדסה הינו כריך עם גבינה בולגרית. מנהל הקפיטריה מאמין שקיים קשר ליניארי בין הפידיון מיחידת כריך בש"ח (Y) לאורך הלחמנייה ממנה עשוי הכריך, בס"מ (X). הוא הורה למהנדס התעשייה של הקפיטריה לבחון את הקשר באמצעות רגרסיה ליניארית פשוטה. לשם כך, אסף המהנדס 30 תצפיות והתקבל הממצא הבא: $\hat{y} = 10 + 0.1x$.
רק אז הבחין המהנדס כי בתצפיות שאסף, אורך הכריכים נמדד במ"מ במקום בס"מ, ולכן התקבל קו ניבוי שגוי.
מהי משוואת הרגרסיה שהייתה מתקבלת אילו המדידות היו מתבצעות בס"מ? הסבירו את תשובתכם.

<u>שאלה 2 - ממבחן</u>

החליטו לבדוק את הקשר בין כמות המטען שמובילה חברת תעופה לבין ההכנסות שלה מכך. לשם כך בדקו את כמות המטען וההכנסה ב-10 חברות תעופה גדולות בארצות הברית. נסמן ב- X את כמות המטען (במיליוני טון) וב- Y את ההכנסה מהובלת מטענים (במיליוני דולרים). התקבלו הממצאים הבאים:

$$\bar{x} = 432.2, \qquad \sum_{i=1}^{10} x_i^2 = 2{,}408{,}810 \quad , \quad \sum_{i=1}^{10} y_i^2 = 103{,}195$$

האומד לקו הרגרסיה חושב והתקבל: $\hat{y} = 5.821 + 0.195x$.

א. מהו ממוצע ההכנסות מהובלת המטענים במדגם? **(90.1)**

ב. חשב את ערכו של מקדם המתאם המרובה (r). **(0.966)**

ג. האם קיים קשר ליניארי חיובי בין כמות המטען לבין ההכנסות מהובלתו ברמת מובהקות 0.05? **(כן)**

<u>שאלה 3 - ממבחן</u>

להלן נתונים לגבי עלויות התפעול של דוד חשמלי כפונקציה של דרגות עומס:

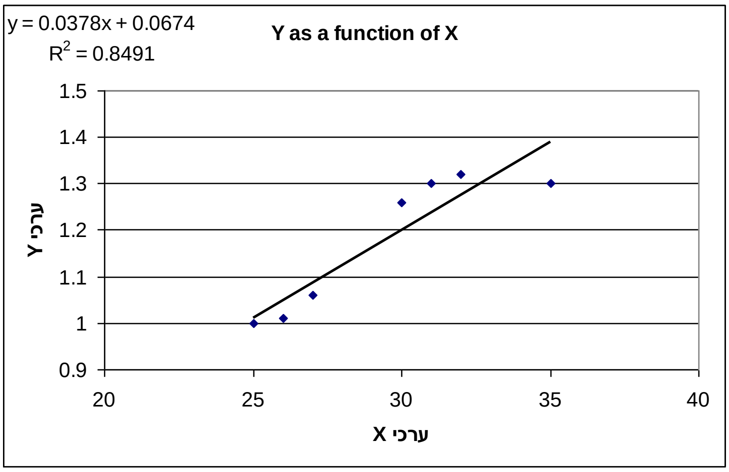
עומס	0	1	2	4	5	8
עלות	0.5	2	4.2	6	6.5	9.5

א. אמוד את קו הרגרסיה של העלות כפונקצית ליניארית של העומס. **($\hat{y} = 1.15 + 1.09x$)**

ב. האם ע"ס הנתונים ניתן לומר שיש קשר ליניארי בין העלות לעומס ברמת מובהקות 0.05? **(כן)**

<u>שאלה 4</u>

נתון הגרף הבא:



עוד נתון כי היו 7 תצפיות.

יתר על כן $SS_x = 77.714$.

א. האם הרגרסיה מובהקת בר"מ 1%? **(כן)**

ב. בדוק את ההשערה $H_0: \beta_1 \ge 0.03$ כנגד $H_1: \beta_1 < 0.03$ בר"מ 10%. **(מקבלים את H_0)**

<u>שאלה 5</u>

חוקר מעוניין לבדוק את הקשר בין עלויות התפעול של דוד חשמלי לדרגת העומס שלו. ההנחה היא שדרגת העומס של הדוד החשמלי משפיעה על עלות התפעול שלו. לצורך כך ביצע החוקר ניסוי. את התוצאות הוא ניתח באמצעות מודל הרגרסיה הליניארית ב-excel. אבוי לחוקר, הוא שכח לבצע גיבוי ובקובץ פגע וירוס רשע ומרושע. מה שהוא הצליח להציל מקובץ הפלט, היה הפלט הקטוע הבא:

SUMMARY OUTPUT						
Regression Statistics						
Multiple R						
R Square						
Standard Error						
Observations	6					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression				126.8087	0.00035429	
Residual			0.406			
Total						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept			2.774758	0.050086	-0.000701492	2.300701492
X		0.096794787			0.821254031	1.358745969

א. עזור לחוקר להשלים את הפלט הדרוש.

ב. בחן את ההשערה $H_0: \beta_1 \le 1$ כנגד האלטרנטיבה $H_1: \beta_1 > 1$ ברמת מובהקות 0.1 **(לא דוחים את השערת האפס)**

ג. מצא רב"ס לתוחלת עלות התפעול של דוד חשמלי כשדרגת העומס שלו היא 7, ברמת סמך 95%.

(8.78 ± 1.2234)