תרגול 10
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 מ–סויים, 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**: $\boxed{\hat{y}_i = b_0 + b_1 x_i}$

בניית משוואת רל"פ - שיטת הריבועים הפחותים (Ordinary Least Squares)

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")

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

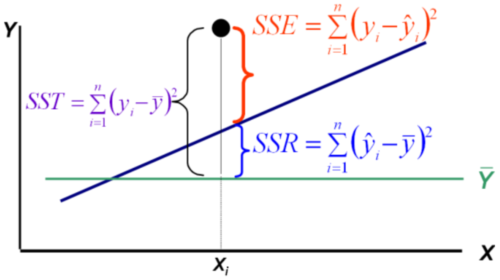
סך הסטייה מהממוצע $\bar{y}$ שהרגרסיה אינה מצליחה להסביר ("סך השונות הבלתי-מוסברת", "סך הרעש במודל")

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

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, האומדן 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		**באוכלוסיה**
$\hat{\rho} \equiv r = \frac{SS_{xy}}{\sqrt{SS_x \cdot SS_y}}$	$\rightarrow$	$\rho = \frac{cov(X,Y)}{\sigma_x\,\sigma_y}$

ברגרסיה ליניארית <u>פשוטה</u> 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 המרובֶּה".

<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	$T_{b_0}=\dfrac{b_0-\mu_0}{S_{b_0}}>t_{1-\alpha}^{n-2}$	$T_{b_0}=\dfrac{b_0-\mu_0}{S_{b_0}}<-t_{1-\alpha}^{n-2}$	$\left\|T_{b_0}\right\|=\left\|\dfrac{b_0-\mu_0}{S_{b_0}}\right\|>t_{1-\frac{\alpha}{2}}^{n-2}$

	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	$T_{b_1}=\dfrac{b_1-\mu_1}{S_{b_1}}>t_{1-\alpha}^{n-2}$	$T_{b_1}=\dfrac{b_1-\mu_1}{S_{b_1}}<-t_{1-\alpha}^{n-2}$	$\left\|T_{b_1}\right\|=\left\|\dfrac{b_1-\mu_1}{S_{b_1}}\right\|>t_{1-\frac{\alpha}{2}}^{n-2}$

<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.

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

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

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

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

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

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

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

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

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

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:

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

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 השערות: $H_0: \beta_1 \le 1$, $H_1: \beta_1 > 1$ **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%.