52571 - רגרסיה ומודלים סטטיסטיים - תשפ״ד סמסטר ב׳

בוחן אמצע — מועד א׳

<u>הנחיות כלליות</u>

1. זמן הבחינה: שעתיים אקדמיות
2. חומר פתוח
3. מותר לצטט כל תוצאה שראינו בכיתה, אלא אם השאלה מבקשת במפורש לפתח או להוכיח את התוצאה

נתון מדגם $(\boldsymbol{x}_i, Y_i)$ עבור $i = 1,...,n$, כאשר $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ וקטור משתנים מסבירים ו-Y_i משתנה תוצאה רציף. נסמן $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$, ונסמן

$$\boldsymbol{X} = [x_{ij}] \in \mathbb{R}^{n\times(p+1)}, \quad i = 1,...,n,\ j = 0,...,p$$

את מטריצת המודל, כאשר $x_{i0} = 1$. הניחו שהעמודות של $\boldsymbol{X}$ בלתי-תלויות ליניארית (בת״ל).

א. (10 נק׳) כתבו את הנחות המודל הליניארי (הכללי) ואת הנחות המודל הליניארי הנורמלי על הנתונים. יש לציין מהם הפרמטרים הלא ידועים של המודל.

ב. (15 נק׳) תחת המודל הליניארי, חשבו את הגדלים הבאים במונחי הפרמטרים של המודל, **וציינו את המימד** של כל אחד מהם: $\text{cov}[\hat{\boldsymbol{Y}}]$, $\mathbb{E}[\hat{\boldsymbol{Y}}]$, $\text{cov}(\boldsymbol{Y}, \hat{\boldsymbol{Y}})$.

ג. (15 נק׳) תחת המודל הליניארי הנורמלי, נניח שהנתונים כוללים $p = 5$ משתנים מסבירים. כתבו מבחן ברמת מובהקות α לבדיקת ההשערה

$$H_0 :\ \beta_1 - \beta_5 = 0 \quad \text{vs} \quad H_1 :\ \beta_1 - \beta_5 \neq 0\ .$$

יש לציין (i) סטטיסטי מבחן (ii) ערך קריטי (עבור α כללית).

ד. (15 נק׳) נסמן $M := \text{Im}(\boldsymbol{X})$. נסמן ב-$\tilde{\boldsymbol{X}}$ את המטריצה המתקבלת מ-$\boldsymbol{X}$ ע״י מחיקת חלק מהעמודות, ונסמן $L := \text{Im}(\tilde{\boldsymbol{X}})$. הראו שמתקיים $P_L\boldsymbol{Y} = P_L\hat{\boldsymbol{Y}}$ (כאשר $\hat{\boldsymbol{Y}}$ זה וקטור הערכים החזויים במודל עם מטריצת ה-$\boldsymbol{X}$ המקורית). האם נדרשות הנחות המודל הליניארי או המודל הליניארי הנורמלי כדי שהטענה תתקיים?

עבור הסעיפים הבאים, נסתכל על קובץ נתונים ספציפי שכולל את המשתנים המסבירים

weekly sport time = זמן שבועי (בדקות) המוקדש לפעילות גופנית

group: adults (A), children (C), elderly (E) = קבוצת גיל: מבוגרים, ילדים, קשישים

ואת משתנה התוצאה

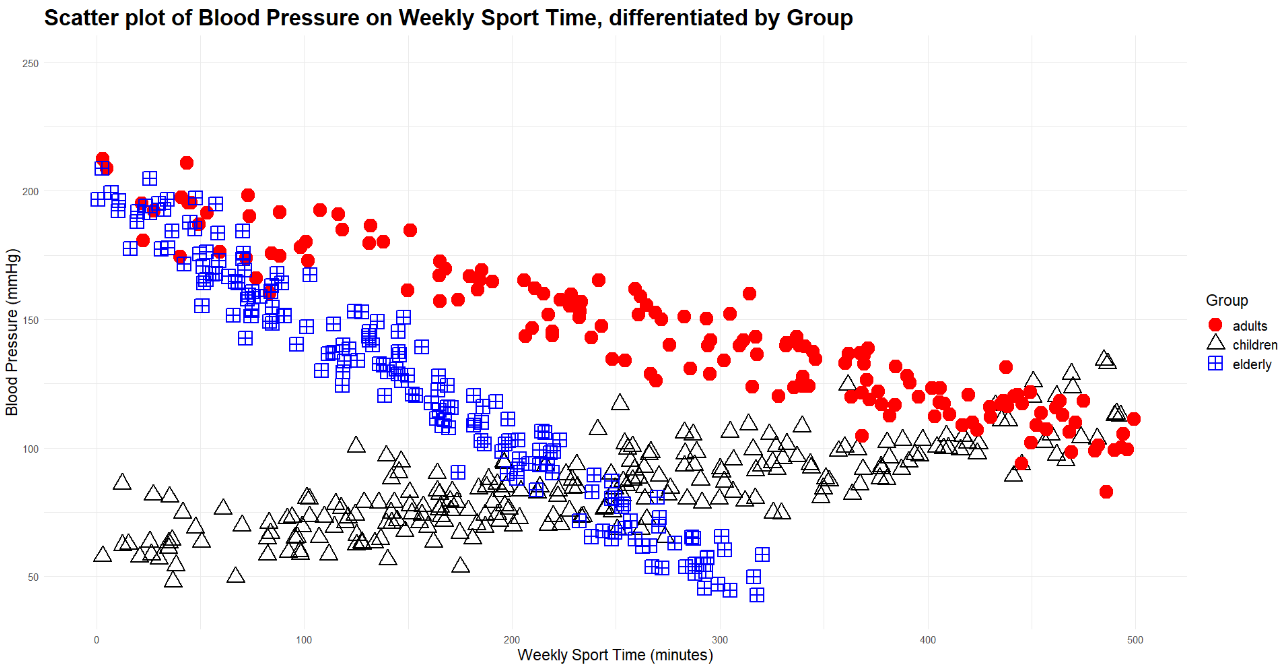
Y = blood pressure = לחץ דם

מצורף תרשים פיזור של הנתונים לפי קבוצת גיל.

ה. (15 נק׳) מהסתכלות ראשונית על התרשים בלבד: האם יש אינדיקציה ברורה לאינטראקציה בין קבוצת הגיל ובין זמן הפעילות הגופנית בדקות? האם יש אינדיקציה ברורה לחותך שונה עבור כל אחת מהקבוצות? הסבירו **בקצרה**.

ו. (15 נק׳) אנחנו רוצים לבדוק אם זמן הפעילות הגופנית משפיע על לחץ הדם של מבוגרים (A) וקשישים (E) באותו האופן, כלומר, שאותה עלייה בלחץ הדם לכל דקת פעילות נוספת צפויה עבור מבוגרים ועבור קשישים. מהי מטריצת X המתאימה במודל הליניארי?

ז. (15 נק׳) נסחו את השאלה שבה מתעניינים בסעיף ה׳ בתור השערת אפס פורמלית (במונחי הפרמטרים של המודל).



בהצלחה!

<u>רגרסיה ומודלים סטטיסטיים- פתרון בוחן האמצע- מועד א'</u>

א. המודל הלינארי הכללי:

$$Y = X\beta + \epsilon, \epsilon \sim (0, \sigma^2 I)$$

כאשר מוסיפים את הנחת הנורמליות:

$$\epsilon \sim N(0, \sigma^2 I)$$

הפרמטרים הלא ידועים של המודל הם β, σ^2.

ב. עבור $P_X = X(X^TX)^{-1}X^T \in R^{n\times n}$:

$$E(\hat{Y}) = E(P_XY) = P_XE(Y) = P_X[X\beta + E(\epsilon)] = P_XX\beta = X\beta \in R^n$$

$$cov(\hat{Y}) = cov(P_XY) = P_Xcov(Y)P_X^T = P_Xcov(X\beta + \epsilon)P_X^T = P_Xcov(\epsilon)P_X^T$$
$$= \sigma^2 P_XP_X^T \overset{P_X^T = P_X = P_X^2}{\cong} \sigma^2 P_X \in R^{n\times n}$$

$$cov(\hat{Y}, Y) = cov(P_XY, Y) = P_Xcov(Y) = \sigma^2 P_X \in R^{n\times n}$$

ג.

ראינו כי בעבור ו"מ Z, מתקיים:

$$cov(Z)_{ij} = Cov(Z_i, Z_j)$$
$$E(Z)_i = E(Z_i)$$

תחת הנחות המודל הלינארי הנורמלי:

$$\hat{\beta} \sim N_{p+1}(\beta, \sigma^2(X^TX)^{-1})$$

כפי שראינו זה גורר שההתפלגות השולית של כל כניסה היא נורמלית חד ממדית עם הפרמטרים לעיל.

לכן:

$$\hat{\beta}_1 - \hat{\beta}_5 \sim N_1(\beta_1 - \beta_5, \sigma^2((X^TX)^{-1}_{22} + (X^TX)^{-1}_{66} - 2\cdot(X^TX)^{-1}_{26})$$

תחת H_0, $\beta_1 - \beta_5 = 0$ ומכאן:

$$T = \frac{\hat{\beta}_1 - \hat{\beta}_5}{\sqrt{(\widehat{\sigma^2}\ (X^TX)^{-1}_{22} + (X^TX)^{-1}_{66} - 2\cdot(X^TX)^{-1}_{26})}} \sim t_{n-6}$$

וזאת כי $\frac{\sigma^2\chi^2_{n-6}}{n-6} \sim \widehat{\sigma^2} = \frac{||e||^2}{n-5-1}$ והמשתנה:

$$\frac{\hat{\beta}_1 - \hat{\beta}_5}{\sqrt{(\ (X^TX)^{-1}_{22} + (X^TX)^{-1}_{66} - 2\cdot(X^TX)^{-1}_{26})}} \overset{H_0}{\approx} \sigma\cdot N(0,1)$$

וכן כי העמודה והשורה הראשונה של המטריצה מתאימות לחותך.
האלטרנטיבה דו צדדית, לכן נדחה אם:

$$|T| > C_\alpha$$

כאשר $C_\alpha = t_{1-\frac{\alpha}{2}, n-6}$

ד. ללא שום הנחה מלבד אי התלות בין עמודות X, מתקיים:

$$P_L\hat{Y} = P_L P_M Y = P_L Y$$

כאשר המעבר השני נובע מהטענה שראיתם:

אם L, M תתי מרחבים כך שמתקיים $L \subseteq M$ אז $P_M P_L = P_L P_M = P_L$.

דרך נוספת:

$$P_L Y = P_L(P_M Y + (I - P_M)Y) = P_L(P_M Y + P_M^{\perp} Y) = P_L\hat{Y} + P_L P_M^{\perp} Y = \ \ P_L\hat{Y}$$

כי $L \subseteq M$ ו- $P_M^{\perp} Y$ שייך למשלים האורתוגונלי של L (מההגדרה $(P_M^{\perp} Y)^T u = 0, \forall u \in M$ ובפרט עבור $u \in L$).

ה. יש אינדיקציה ברורה לאינטראקציה בעבור כל שלוש הקבוצות- נראה שאם היינו אומדים בנפרד 3 קווי רגרסיה, אחד בעבור כל קבוצת גיל (ראינו שזה שקול), היינו מקבלים שיפוע שונה בכל קבוצה. אין אינדיקציה ברורה להבדלים בחותכים של הקווים בקבוצת המבוגרים והקשישים, אך כן חותך שונה ונמוך הרבה יותר בעבור קבוצת הילדים.

ו. ראינו בתרגול כי אמידה של שתי (או 3 במקרה הזה- כי יש 3 קבוצות בנתונים) רגרסיות נפרדות, שקולה לאמידת המודל עם אינטראקציה וחותך נפרד לכל קבוצה. לכן כל אחד מהמודלים הבאים יתקבל ומתאים:

1.

$$Y_i = \ \beta_0 + \ \beta_1 \cdot 1_{\{i\in E\}} + \beta_2 \cdot 1_{\{i\in C\}} + \beta_3 \cdot S_i + \beta_4 \cdot S_i \cdot 1_{\{i\in E\}} + \beta_5 \cdot S_i \cdot 1_{\{i\in c\}} + \epsilon_i$$

2.

$$Y_i = \ \beta_0 + \ \beta_1 \cdot 1_{\{i\in E\}} + \beta_2 \cdot 1_{\{i\in A\}} + \beta_3 \cdot S_i + \beta_4 \cdot S_i \cdot 1_{\{i\in E\}} + \beta_5 \cdot S_i \cdot 1_{\{i\in A\}} + \epsilon_i$$

המטריצה X המתאימה עבור המודל הראשון (נניח שישנם 3 קשישים ו-3 ילדים והמטריצה מסודרת לפי קשישים, מבוגרים, ילדים):

$$\begin{matrix} 1 & 1 & 0 & S_1 & S_1 & 0 \\ 1 & 1 & 0 & S_2 & S_2 & 0 \\ 1 & 1 & 0 & S_3 & S_3 & 0 \\ 1 & 0 & 0 & S_4 & 0 & 0 \\ 1 & 0 & 0 & S_5 & 0 & 0 \\ 1 & 0 & 0 & S_6 & 0 & 0 \\ 1 & 0 & 0 & S_7 & 0 & 0 \\ 1 & 0 & 1 & S_8 & 0 & S_8 \\ 1 & 0 & 1 & S_9 & 0 & S_9 \\ 1 & 0 & 1 & S_{10} & 0 & S_{10} \end{matrix}$$

ניקוד חלקי יינתן למי שיכתוב:

$$Y_i = \ \beta_0 + \beta_1 \cdot 1_{\{i\in E\}} + \ \beta_2 \cdot S_i + \beta_3 \cdot S_i \cdot 1_{\{i\in E\}} + \epsilon_i$$

או

$$Y_i = \ \beta_0 + \beta_2 \cdot S_i + \beta_3 \cdot S_i \cdot 1_{\{i\in E\}} + \beta_4 \cdot S_i \cdot 1_{\{i\in C\}} + \epsilon_i$$

ז. במודל 1, נבדוק את ההשערה:

$$H_0 : \beta_4 = 0 \ vs \ H_1 : \beta_4 \neq 0$$

במודל 2, נבדוק את ההשערה:

$$H_0 : \beta_5 - \beta_4 = 0 \ vs \ H_1 : \beta_5 - \beta_4 \neq 0$$