רגרסיה- תרגיל 8 תשפ"ה

שאלה 1

א. קראו את הקובץ המצורף: "ex8_data.csv". הפרידו את המשתנה Y משאר המשתנים המסבירים. בדקו והציגו את כל 4 האינדיקציות למולטיקולינאריות בין המשתנים המסבירים. האם ישנה עדות למולטיקולינאריות בנתונים? בין אילו משתנים? הסבירו בקצרה.

ב. אמדו את המודל כאשר Y הוא המשתנה המוסבר על כל המשתנים המסבירים ושמרו את התוצאות. אילו מהמשתנים יצאו מובהקים? האם התוצאה מפתיעה?
בשלב הבא הוסיפו משתנה חדש:

$$Y_{new} = Y + rnorm(300,0,1)$$

אמדו את המודל המסביר את Y_{new} על כל שאר המשתנים המסבירים (ללא Y).

ג. הציגו $subplot$ המכיל שני גרפים: האחד $barplot$ כפול, כאשר בצבע אחד מופיעים המקדמים שהתקבלו מהמודל הראשון, ובצבע אחר המקדמים שהתקבלו מאמידת המודל השני.
השני, מאה נקודות שבחרתם מתוך וקטורי הערכים החזויים מכל אחת מהאמידות (כלומר 2 ערכים חזויים לכל תצפית- $\hat{Y}_i$, $\widehat{Y_{new_i}}$). בעבור כל אינדקס בין 1-100, הציגו בצבעים שונים את $\hat{Y}_i$, $\widehat{Y_{new_i}}$.
בכותרת של כל גרף, בהתאמה, הקפידו לציין את הגדלים:

$$\frac{\left\|\hat{\beta} - \widehat{\beta_{new}}\right\|^2}{\left\|\hat{\beta}\right\|^2}, \frac{\left\|\hat{Y} - \widehat{Y_{new}}\right\|^2}{\left\|\hat{Y}\right\|^2}$$

הסבירו את התוצאות ואת ההבדלים בין הגדלים.

שאלה 2:

בשאלה הזאת נמצא פרשנות למדד VIF: אנחנו נראה ש-VIF עבור משתנה מסביר j זה למעשה היחס בין השונות של המקדם (הנאמד) ברגרסיה מרובה לבין השונות של המקדם ברגרסיה פשוטה.

כרגיל, אנחנו מסמנים ב-$\boldsymbol{X}^{(j)}$ את העמודה במטריצה $\boldsymbol{X}$ שמתאימה למשתנה המסביר ה-j, כך ש: $\boldsymbol{X} = [\boldsymbol{X}^{(0)}\ \boldsymbol{X}^{(1)}\ \cdots \boldsymbol{X}^{(p)}] \in \mathbb{R}^{n\times(p+1)}$. נסמן $\boldsymbol{X}_{-j}$ את המטריצה $\boldsymbol{X}$ ללא העמודה $\boldsymbol{X}^{(j)}$, עבור $j = 1, ..., p$. כזכור, אם $SST(j),\ SSE(j)$ ו-R_j^2 מסמנים, בהתאמה, את SST, SSE ו-R^2 ברגרסיה מרובה של $\boldsymbol{X}^{(j)}$ על $\boldsymbol{X}_{-j}$, אז: $\mathrm{VIF} = \frac{1}{1-R_j^2}$.

א. רגרסיה פשוטה <u>ללא חותך</u> מתאימה לנתונים את המודל $Y_i = \beta X_i + \epsilon_i$ כלומר זה אכן מודל רגרסיה עם משתנה מסביר בודד וללא האיבר הקבוע (שאותו אנחנו רגילים לסמן β_0). הראו שהאומד של β ברגרסיה פשוטה ללא חותך נתון ע"י

$$\hat{\beta} = \frac{1}{\|\boldsymbol{x}\|^2}\boldsymbol{x}^\top \boldsymbol{Y}$$

כאשר $\boldsymbol{x} = (X_1, ..., X_n)^\top$.

עובדה כללית (*): ברגרסיה מרובה של $\boldsymbol{Y}$ על $\boldsymbol{X}$ אפשר לקבל את $\hat{\beta}_j$ ע״י רגרסיה פשוטה ללא חותך של $\boldsymbol{Y}$ על הוקטור

$$\boldsymbol{z} = (\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}^{(j)}$$

כאשר

$$\boldsymbol{P}_{-j} = \boldsymbol{X}_{-j}(\boldsymbol{X}_{-j}^\top \boldsymbol{X}_{-j})^{-1}\boldsymbol{X}_{-j}^\top$$

היא מטריצת ההיטל על מרחב העמודות של $\boldsymbol{X}_{-j}$.

ב. הראו/הסבירו שברגרסיה פשוטה (עם חותך) של $\boldsymbol{Y}$ על $\boldsymbol{X}^{(j)}$, כלומר עבור j נתון ותחת המודל $Y_i = \beta_0 + \beta_j X_{ij} + \epsilon_i$, ניתן להציג

$$\hat{\beta}_j = \frac{1}{\|\boldsymbol{w}\|^2}\boldsymbol{w}^\top \boldsymbol{Y}$$

כאשר

$$\boldsymbol{w} = (\boldsymbol{I} - \boldsymbol{P}_0)\boldsymbol{X}^{(j)}, \qquad \boldsymbol{P}_0 := \mathbf{1}_n(\mathbf{1}_n^\top \mathbf{1}_n)^{-1}\mathbf{1}_n^\top = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$$

הדרכה: הסתכלו על הרגרסיה הפשוטה (עם חותך) בתור רגרסיה מרובה על המטריצה $n \times 2$ שנתונה ע״י $\boldsymbol{X} = [\mathbf{1}_n \ \boldsymbol{X}^{(j)}]$, והשתמשו בעובדה הכללית (*).

ג. בעובדה כללית (*) הגדרנו $\boldsymbol{z} = (\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}^{(j)}$. הראו שבביטוי הזה ניתן להחליף את $\boldsymbol{X}^{(j)}$ ב-$\boldsymbol{w}$, כלומר שבאופן כללי מתקיים:

$$\boldsymbol{z} = (\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{w}$$

עבור הוקטור $\boldsymbol{w}$ שהוגדר בסעיף ב׳, והסיקו בעזרת עובדה כללית (*) שברגרסיה מרובה של $\boldsymbol{Y}$ על $\boldsymbol{X}$, אפשר להציג:

$$\hat{\boldsymbol{\beta}}_j = \frac{1}{\|(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{w}\|^2}[(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{w}]^\top \boldsymbol{Y}$$

ד. את הסעיפים הקודמים אפשר לסכם באופן הבא:

ברגרסיה פשוטה (עם חותך): $$\hat{\boldsymbol{\beta}}_j = \frac{1}{\|\boldsymbol{w}\|^2}\boldsymbol{w}^\top \boldsymbol{Y},$$

ברגרסיה מרובה: $$\hat{\boldsymbol{\beta}}_j = \frac{1}{\|(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{w}\|^2}[(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{w}]^\top \boldsymbol{Y},$$

כאשר $\boldsymbol{w}$ זה הוקטור שהוגדר בסעיף ב׳. חשבו את $\mathrm{Var}(\hat{\beta}_j)$ בכל אחד משני המקרים. בנוסף, הראו שהיחס בין $\mathrm{Var}(\hat{\beta}_j)$ ברגרסיה מרובה לבין $\mathrm{Var}(\hat{\beta}_j)$ ברגרסיה פשוטה נתון ע״י:

$$\frac{\|\boldsymbol{w}\|^2}{\|(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{w}\|^2}.$$

ה. הסבירו שמתקיים:

$$\|(\boldsymbol{I}-\boldsymbol{P}_0)\boldsymbol{X}^{(j)}\|^2 = \|\boldsymbol{w}\|^2 = SST(j)$$

$$\|(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{X}^{(j)}\|^2 = \|(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{w}\|^2 = SSE(j)$$

ו. הסיקו:

$$VIF = \frac{1}{1-R_j^2} = \frac{\|\boldsymbol{w}\|^2}{\|(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{w}\|^2},$$

כלומר זה בדיוק היחס בין השונויות שחושב בסעיף ד׳.

שאלה 3:

בשאלה זו נניח כרגיל שהנתונים הם $\boldsymbol{X} \in \mathbb{R}^{n\times(p+1)}$, מטריצה דטרמיניסטית (לא מקרית), ו- $\boldsymbol{Y} \in \mathbb{R}^n$, וקטור מקרי שהתפלגותו תלויה ב- $\boldsymbol{X}$, אבל המודל עבור $\boldsymbol{Y}$ יהיה

$$\boldsymbol{Y} = \boldsymbol{\mu} + \varepsilon, \quad \varepsilon \sim (0, \sigma^2 I_n), \tag{4}$$

כאשר $\boldsymbol{\mu} \in \mathbb{R}^n$ דטרמיניסטי ויכול להיות פונקציה של $\boldsymbol{X}$.
יהי $\hat{\boldsymbol{Y}} \in \mathbb{R}^n$ וקטור ערכים חזוים שהתקבלו מהנתונים $(\boldsymbol{X}, \boldsymbol{Y})$ בשיטה <u>כלשהי</u> (לא בהכרח מריבועים פחותים). נגדיר את תוחלת השגיאה הריבועית בחיזוי (MSPE) המתאימה בתור $MSPE := \mathbb{E}\left\|\hat{\boldsymbol{Y}} - \boldsymbol{\mu}\right\|^2$.
כאן נתרכז בשיטות חיזוי שעבורן $\hat{\boldsymbol{Y}} = \boldsymbol{P}\boldsymbol{Y}$, כאשר $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ היא מטריצת ההטלה על איזשהו תת-מרחב של $\mathbb{R}^n$ ממימד $0 \le r \le n$. $\boldsymbol{P}$ דטרמיניסטית, ויכולה - אך לא חייבת - להיות תלויה ב- $\boldsymbol{X}$.

א. הוכיחו את התוצאה הכללית הבאה. יהיו $Z, W \in \mathbb{R}^n$ זוג וקטורים מקריים בעלי משותף כלשהו. הראו כי מתקיים:

$$tr[\mathbb{E}[ZW^T]] = tr[Cov(Z, W)] + tr[\mathbb{E}(Z) \cdot [\mathbb{E}(W)]^T]$$

ב. נגדיר $SSE(\boldsymbol{P}) = \|\boldsymbol{Y} - \boldsymbol{P}\boldsymbol{Y}\|^2$. הוכיחו:

$$MSPE = \mathbb{E}\left[SSE(\boldsymbol{P}) + 2\sigma^2 r - n\sigma^2\right]$$

הדרכה - בדומה למה שראינו בכיתה, הציגו $MSPE = \mathbb{E}\left\|\hat{\boldsymbol{Y}} - \boldsymbol{\mu}\right\|^2 = \mathbb{E}\left\|(\hat{\boldsymbol{Y}} - \boldsymbol{Y}) + (\boldsymbol{Y} - \boldsymbol{\mu})\right\|^2$, הרחיבו את הביטוי בעזרת כללים אלמנטריים ($\|a+b\|^2 = \|a\|^2 + \|b\|^2 + 2a^T b$), והשתמשו בסעיף א' כדי לסיים.

ג. תהי $\boldsymbol{Y}^*$ התממשות חדשה ממודל (4) ($\boldsymbol{Y}, \boldsymbol{Y}^*$ הם iid). הראו:

$$\mathbb{E}\left\|\boldsymbol{Y}^* - \hat{\boldsymbol{Y}}\right\|^2 = MSPE + n\sigma^2 \tag{5}$$

$$\mathbb{E}\left\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\right\|^2 = \mathbb{E}[SSE(\boldsymbol{P})] \tag{6}$$

והסיקו

$$\mathbb{E}\left\|\boldsymbol{Y}^* - \hat{\boldsymbol{Y}}\right\|^2 - \mathbb{E}\left\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\right\|^2 = MSPE - \mathbb{E}[SSE(\boldsymbol{P})] \ge 0 \tag{7}$$

<u>הערה:</u> את הביטוי ב-(7) אפשר להבין בתור תוחלת ההפרש בין שגיאת החיזוי "מחוץ למדגם" ("$Out - Of - Sample$") לבין שגיאת החיזוי "בתוך המדגם" ("$In - Sample$"). אם בגישה נאיבית היינו מצפים שתוחלת ההפרש הזה היא אפס, אז (7) מראה שבאופן כללי, תוחלת ההפרש הזה חיובית (כלומר, $\left\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\right\|^2$ זה אומד "אופטימי" (optimistic) עבור $\mathbb{E}\left\|\boldsymbol{Y}^* - \hat{\boldsymbol{Y}}\right\|^2$).