מטלת פרויקט 3

מטרת משימה 3 היא ליישם את השיטות שלמדתם כדי לבחון את הקשר בין משתנים מסבירים (X) למשתנה מוסבר (Y).

ודאו כי ההסברים שלכם מלווים בתוצאות המוצגות בקובץ ההגשה, במידת הצורך ניתן להציג בשנית תוצאות ממשימה קודמת. הפעילו שיקול דעת ובחרו בעצמכם כיצד להציג את התוצאות של הסעיפים השונים. יש לספק הסברים מפורטים לתוצאות המתקבלות.

חלק א – רגרסיה ליניארית

בחלק זה, בחרו לפחות 3 משתנים מסבירים מתוכם לפחות אחד רציף ואחד בדיד (בלומר, קטגוריאלי או בינארי) ומשתנה מוסבר אחד שהוא רציף. אפשר להתייחס למשתנה בדיד שהוא אורדינלי ויש בו הרבה ערכים (לדוגמה גיל) כמשתנה רציף.

1. הגדירו את שאלת המחקר (מהי השפעה של המשתנים המסבירים על המשתנה המוסבר).
2. בחרו באופן אקראי תת-מדגם בגודל 200, אלו הנתונים שאיתם נעבוד בחלק זה. לצורך למידה, נתייחס למדגם המקורי (לפני שלקחנו ממנו תת-מדגם) כמייצג את האוכלוסייה ולתת-המדגם כמייצג את הנתונים שאנחנו חשופים אליהם.
3. <u>הצגת הנתונים וניתוח ראשוני:</u> הסתכלו על הנתונים.
   א. הציגו Boxplot/היסטוגרמה לכל משתנה רציף וטבלה לכל משתנה בדיד.
   ב. שאלו את עצמכם: האם יש נתונים חסרים? נתונים חריגים? האם ההתפלגות של המשתנים סימטרית? האם תוכלו להעריך מאיזו התפלגות הנתונים לקוחים?
   ג. עבור כל משתנה מסביר רציף X, ציירו גרף של Y כפונקציה של X. כיצד נראה הקשר בין X ל-Y? לאיזה משתנה מסביר X נראה שיש הכי הרבה השפעה על Y?
   ד. עבור כל משתנה בדיד עם מעט ערכים, ציירו Boxplot/היסטוגרמה של Y כפונקציה של ערכי X השונים.

   חשבו כיצד להציג את התוצאות של סעיף זה היות וישנם לא מעט גרפים. אם גרף לא מועיל לכם, אל תציגו אותו וציינו זאת בתוספת הסבר.
4. חשבו את וקטור המקדמים $\hat{\beta}$ והסבירו את משמעות הערכים המתקבלים (מה ההשפעה של עליה ביחידה אחת של כל משתנה על המשתנה המוסבר).
5. כתבו את טבלת ה-ANOVA, בדקו את מבחן F וחשבו את R^2, R^2_{adj}.
6. חשבו רווח סמך למקדמים $\beta_0^*, \dots, \beta_k^*$. חשבו כעת גם את $\hat{\beta}$ על המדגם המקורי. נרצה שרווחי הסמך על סמך 200 התצפיות יכילו את האומד הנקודתי על סמך כלל הנתונים. בדקו האם רווחי הסמך שקיבלתם עבור התת-מדגם מכילים את ערכי $\widehat{\beta_j}$ שקיבלתם במדגם המקורי.
7. בדקו בעזרת מבחן סטטיסטי אם כל אחד מהמקדמים $\beta_0^*, \dots, \beta_k^*$ שונה מאפס. נסחו את ההשערות והסבירו מהו הסטטיסטי.
8. השתמשו בגרף שאריות ובכל כלי אחר כדי לקבוע אם הנחת הלינאריות סבירה, אם ההנחה של שוויון שונויות סבירה ואם סביר שהתפלגות הרעש הינה נורמלית.

9. עבור 1000 נקודות מהמדגם המקורי שאינן מופיעות בתת-מדגם בגודל 200, נחשוב על כל נקודה כזו ב- (X_{new}, Y_{new}):
   א. חזו את ערך הנקודה בהינתן X_{new}.
   ב. חשבו רווח סמך לנקודה (ולא לתוחלת) ברמת סמך מקורבת של 95%.
   ג. חשבו את אחוז רווחי הסמך שמכילים את הערך האמיתי Y_{new}. האם הוא קרוב ל-95%? מה זה אומר לדעתכם?

   בסעיף זה יש להציג פלט רק עבור סעיף ג'.
10. הוסיפו למודל את האינטראקציות מסדר ראשון. במילים אחרות, נניח כי המשתנים מהסבירים הם X_1, X_2, X_3, הוסיפו את המשתנים החדשים
$$Z_{12} = X_1 \cdot X_2,\ Z_{13} = X_1 \cdot X_3,\ Z_{23} = X_2 \cdot X_3$$
אמדו את וקטור המקדמים במודל זה. איזה מודל עדיף? השתמשו בשתי שיטות שונות לבחירת מודל כדי להצדיק את תשובתכם.

חלק שני – רגרסיה לוגיסטית

בחלק זה, בחרו לפחות 3 משתנים מסבירים מתוכם לפחות אחד רציף ואחד בדיד ומשתנה מוסבר אחד שהוא בינארי.

1. הגדירו את שאלת המחקר (מהי השפעה של המשתנים המסבירים על המשתנה המוסבר).
2. בחרו באופן אקראי תת-מדגם בגודל 200. אלו הנתונים שאיתם נעבוד בחלק זה.
3. הצגת הנתונים וניתוח ראשוני: אם אתם משתמשים במשתנים שלא השתמשתם בהם קודם, הסתכלו על הנתונים:
   א. הציגו Boxplot/היסטוגרמה לכל משתנה רציף וטבלה לכל משתנה בדיד.
   ב. שאלו את עצמכם: האם יש נתונים חסרים? נתונים חריגים? האם ההתפלגות של המשתנים סימטרית?
   ג. עבור כל משתנה מסביר רציף X, ציירו Boxplot/היסטוגרמה של X כפונקציה של ערכי Y השונים. לאיזה משתנה מסביר X נראה שיש את ההשפעה הגדולה ביותר על Y?
4. חשבו את וקטור המקדמים $\hat{\beta}$ והסבירו את משמעות הערכים המתקבלים (מה ההשפעה של עליה ביחידה אחת של כל משתנה על המשתנה המוסבר).
5. חשבו רווח סמך למקדמים $\beta_0^*, \dots, \beta_k^*$. חשבו כעת גם את $\hat{\beta}$ על המדגם המקורי. בדקו האם רווחי הסמך שקיבלתם עבור תת המדגם מכילים את ערכי $\widehat{\beta_j}$ שקיבלתם במדגם המקורי.
6. הוסיפו למודל את האינטראקציות מסדר ראשון. במילים אחרות, נניח כי המשתנים המסבירים הם X_1, X_2, X_3, הוסיפו את המשתנים החדשים
$$Z_{12} = X_1 \cdot X_2,\ Z_{13} = X_1 \cdot X_3,\ Z_{23} = X_2 \cdot X_3 .$$
אמדו את וקטור המקדמים במודל זה. איזה מודל עדיף? השתמשו במבחן יחס נראות וכן בשיטה נוספת מבין השיטות שנלמדו כדי להצדיק את תשובתכם.
7. כעת נממש פרוצדורות Forward/Backward Stepwise Regression לבחירת המודל הטוב ביותר. התייחסו כעת לכל המשתנים המסבירים שיש בקובץ הנתונים שלכם (שימו לב לא לבחור משתנים שמעידים באופן ודאי על המשתנה המוסבר, לצורך העניין אם מספר הסיגריות ביום הוא חיובי, בבירור מדובר באדם מעשן). אם מספר המשתנים

המסבירים האפשריים קטן מ-30, הוסיפו אברי אינטראקציה מסדר ראשון לבחירתכם עד שתגיעו למספר זה של משתנים מסבירים. בחרו את המודל הטוב ביותר מבין המודלים הקיימים, תוך שימוש בשתי השיטות (Forward/Backward) עם מדד ציון מודלים מבין אלו שנלמדו. האם קיבלתם תוצאות זהות בשתי השיטות?

8. בצעו את סעיף 7 פעם נוספת אך כעת עם מדד ציון אחר. האם השתנתה ההחלטה שלכם לגבי המודל הטוב ביותר? נסו להסביר את התוצאה שקיבלתם.