רגרסיה ומודלים סטטיסטיים- תרגיל 5

**שאלה 1- בניית רווחי סמך והסקה סטטיסטית**

קובץ הנתונים *Boston* מהספרייה *Mass* ב-*R* מכיל נתונים על מחירי בתים בעיירות בפרברי בוסטון. תוכלו לקרוא על תיאור המשתנים כאן. לאורך השאלה הניחו כי כל הנחות המודל הלינארי הנורמלי מתקיימות.
פלט הרגרסיה שהתקבל מהרצה תוך שימוש בחלק מהמשתנים:

```
Call:
lm(formula = medv ~ lstat + nox + dis + rm, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max 
-17.072  -3.228  -0.907   1.968  26.402 

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)    
(Intercept)  12.03600    3.99033   3.016              0.00269 ** 
lstat        -0.65865    0.05099 -12.917 < 0.0000000000000002 ***
nox         -14.53791    3.49991  -4.154          0.000038444 ***
dis          -0.96699    0.18018  -5.367          0.000000123 ***
rm            4.86340    0.43754  11.115 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.396 on 501 degrees of freedom
Multiple R-squared:  0.6585,	Adjusted R-squared:  0.6558 
F-statistic: 241.6 on 4 and 501 DF,  p-value: < 0.00000000000000022
```

והמטריצה $(X^TX)^{-1}$:

| | Intercept | lstat | nox | dis | rm |
|---|---|---|---|---|---|
| Intercept | 0.546920650 | -0.00300380644 | -0.267061900 | -0.01476321472 | -0.0481765600 |
| lstat | -0.003003806 | 0.00008930473 | -0.001871825 | 0.00004125515 | 0.0004384591 |
| nox | -0.267061900 | -0.00187182489 | 0.420746305 | 0.01463589900 | 0.0002890320 |
| dis | -0.014763215 | 0.00004125515 | 0.014635899 | 0.00111507017 | 0.0003008943 |
| rm | -0.048176560 | 0.00043845907 | 0.000289032 | 0.00030089435 | 0.0065757967 |

לנוחותכם, קוד $R$ לבנייתה:

```
values <- c(
 0.546920650, -0.00300380644, -0.267061900, -0.01476321472, -0.0481765600,
 -0.003003806, 0.00008930473, -0.001871825, 0.00004125515, 0.0004384591,
 -0.267061900, -0.00187182489, 0.420746305, 0.01463589900, 0.0002890320,
 -0.014763215, 0.00004125515, 0.014635899, 0.00111507017, 0.0003008943,
 -0.048176560, 0.00043845907, 0.000289032, 0.00030089435, 0.0065757967)
XTX_inv_matrix <- matrix(values, nrow = 5, ncol = 5, byrow = TRUE)
rownames(XTX_inv_matrix) <- colnames(XTX_inv_matrix) <- c("Intercept", "lstat", "nox",
"dis", "rm")
```

א. על בסיס אומד לינארי חסר הטיה בעל שונות מינימלית, בנו רווח סמך ברמת סמך 0.9 למחיר החציוני של בתים בעיירה שלא כלולה בנתונים, בה ריכוז פליטות הפחמן הוא 0.5 (חלקיקים ל-10 מיליון), מספר החדרים הממוצע הוא 4.5, המרחק הממוצע מאיזורי תעסוקה הוא 3 ק״מ, וכ-13% מהאוכלוסייה בה שייכים למעמד הנמוך.

ב. ברמת מובהקות של 5 אחוזים, בדקו את ההשערה כי ההשפעה של שיעור השייכים למעמד הנמוך באוכלוסייה ושל המרחק הממוצע מאיזורי תעסוקה, על המחיר החציוני של הבתים באיזור, זהה.

ג. פרטו מי הוא סטטיסטי המבחן לבדיקת ההשערה כי כל המקדמים $\beta_1, \beta_2, \beta_3, \beta_4$ שווים ל-0. האם תדחו את ההשערה ברמת מובהקות של 5%?

ד. הניחו שאנו משתמשים ברמת מובהקות של 5%. הסבירו מה המשמעות של מקרה (ללא קשר לנתוני השאלה) בו ערך ה-$P$ של הסטטיסטי מהסעיף הקודם קטן מ-5%, אך ערך ה-$P$ תחת $H_0: \beta_j = 0$ של כל אחד מהסטטיסטיים:

$$\frac{\hat{\beta}_j}{\sqrt{\widehat{\sigma^2} \cdot (X^T X)^{-1}_{(j+1)(j+1)}}}$$

גדול מ-5%.

ה. הניחו את העיירה מנתוני סעיף א'.
בפני ועד העיירה עומדות שתי אלטרנטיבות- להקים איזור תעסוקה שמרחקו הממוצע מהבתים בעיירה עומד על 2 ק"מ, לעומת הוספת חדר לכל בית בעיירה.
בשל עלויות הקמת איזור התעסוקה, נקבע לבחור באפשרות הראשונה אם ורק אם ניתן לומר ברמת מובהקות של 5% שהקמתו תעלה את תוחלת מחירי הבתים בעיירה <u>ביותר מפי 2</u> מאשר אם תיבחר האפשרות השנייה. עזרו לחברי הועד להגיע להחלטה.

באוכלוסייה ושל המרחק הממוצע מאיזורי תעסוקה, על המחיר החציוני של הבתים באיזור, זהה.

פתרון:

אומד לינארי חסר הטיה בעל שונות מינימלית הוא אומד $OLS$ ע"פ גאוס מרקוב. על פי תוצאות הרגרסיה האומד הנקודתי יהיה:

$$a^T\hat{\beta} = 12.03 - 0.658 \cdot 13 - 14.54 \cdot 0.5 - 0.96 \cdot 3 + 4.9 \cdot 4.5 = 15.376$$

כדי לבנות רווח סמך נצטרך את האומד ס"ת של האומד. על פי פלט הרגרסיה:

$$\widehat{\sigma^2} = 5.396^2$$

והאומד לשונות יהיה:

$$\hat{Var}(a^T\hat{\beta}) = a^T \hat{Var}(\hat{\beta}) a = \widehat{\sigma^2}\, a^T (X^T X)^{-1} a = 0.7739957$$

כאשר החישוב:

```
> a = c(1,13,0.5,3,4.5)
> 5.396^2 * t(a) %*% XTX_inv_matrix%*%a
          [,1]
[1,] 0.7739957
```

מכאן שנקבל שרווח הסמך הוא:

$$15.376 \pm t_{0.975,501} \cdot \sqrt{0.774} = [13.64751, 17.10449]$$

ב. סטטיסטי המבחן יהיה:

$$T = \left| \frac{\widehat{\beta_1} - \widehat{\beta_3}}{\sqrt{V\hat{a}r(\widehat{\beta_1} - \widehat{\beta_3})}} \right| = \left| \frac{\widehat{\beta_1} - \widehat{\beta_3}}{\sqrt{V\hat{a}r(\widehat{\beta_1}) + V\hat{a}r(\widehat{\beta_3}) - 2co\hat{v}(\widehat{\beta_1}, \widehat{\beta_3})}} \right|$$

כאשר חישוב הסטטיסטי באופן דומה לסעיף הקודם. התפלגותו תחת השערת ה-0 היא $t_{501}$ ולכן נדחה אם הוא גדול מ-1.96.

ג. כפי שהוכחנו בתרגול סטטיסטי המבחן לבדיקת ההשערה הזו הוא סטטיסטי $F$ הנתון בפלט. תחת השערת ה-0 הוא מתפלג $F_{4,501}$ וערכו הוא 241.6>*qf(0.95,4,501)* - ניתן גם היה להסתכל על ה-$Pvalue$.

ד. המשמעות של מקרה כזה היא שהמודל בכללותו מובהק- כלומר ניתן לדחות את ההשערה שלא כל המקדמים הם אפס, וכי יש קשר בין $X$ ל-$Y$. למרות זאת, לא ניתן להצביע מי מהמקדמים אינו 0. כלומר "משהו" משפיע, אך לא ניתן להצביע בדיוק על מה ועל איך.
בהמשך נראה שבדרך כלל מקרה כזה ינבע ממולטיקולינאריות גבוהה במודל- כלומר מתאם גבוה בין המשתנים המסבירים, כך שתהיה שונות גבוהה לאומד.

ה. נתון שנעדיף את האפשרות הראשונה (להקטין את המרחק הממוצע ביחידה) על השנייה (להגדיל את מספר החדרים ביחידה) אם ורק אם תוחלת המחיר תעלה ביותר מפי 2 בעקבות השינוי הראשון לעומת השינוי השני.
ננסח את בדיקת ההשערות:

$$H_0: -\beta_3 \leq 2\beta_4$$
$$H_1: -\beta_3 \geq 2\beta_4$$

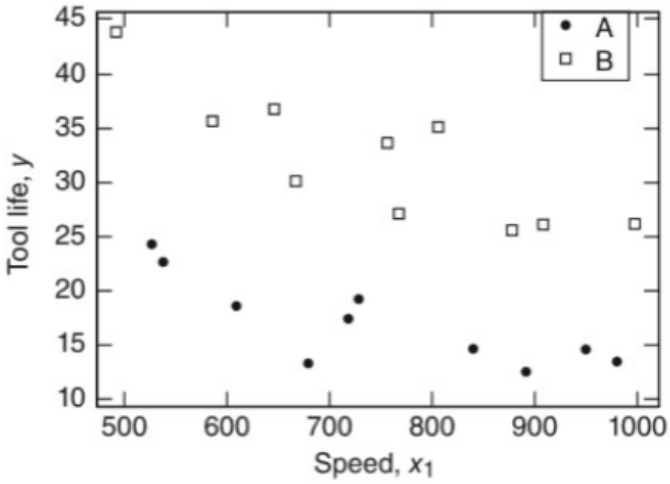
סטטיסטי המבחן המתאים:

$$T = \frac{2\widehat{\beta_4} + \widehat{\beta_3}}{\sqrt{v\hat{a}r(\widehat{\beta_3}) + 4v\hat{a}r(\widehat{\beta_4}) + 4co\hat{v}(\widehat{\beta_3}, \widehat{\beta_4})}}$$

ונדחה אם $T < qt(0.05,501)$. ניתן לחשב את ערכו באופן דומה לסעיף הקודם, אך ניתן גם לשים לב שהמונה חיובי (והמכנה תמיד חיובי) ואילו הערך הקריטי שלילי, ולכן לא נדחה את ההשערה.

**שאלה 2- משתני דמי**

תרשים הפיזור המצורף מראה נתונים על משך זמן חיים (, בשעות) של מחרטה במפעל כנגד מהירות המחרטה (, בסיבובים לדקה) עבור שני סוגים שונים של מחרטות (A, B). מעוניינים לנתח, בשיטות של רגרסיה ליניארית, את הקשר שבין משך החיים הממוצע של מחרטה במפעל לבין מהירות המכונה.




א. הסבירו מה תאמוד רגרסיה ליניארית עבור המודל

$$Y = \beta_0 + \beta_1 X_1 + \epsilon,$$

כלומר הסבירו מה תהיה הפרשנות של האומדים עבור המקדמים במודל הזה.

ב. איזה מודל ליניארי אפשר להתאים לנתונים אם אנחנו רוצים לאמוד (בבת-אחת, כלומר עם רגרסיה בודדת) את האפקט של מהירות המחרטה על תוחלת חיים, בנפרד עבור מחרטה מסוג A ועבור מחרטה מסוג B? השתמשו בסימון $\gamma_j$ עבור המקדמים של המודל שאתם מציעים בסעיף הזה, זאת אומרת, כתבו $Y = \gamma_0 + \gamma_1 X_1 + \cdots + \gamma_p X_p + \epsilon$ עבור $p$ מתאים ועבור משתנים מסבירים מתאימים $X_2, ..., X_p$, בנוסף ל-$X_1$, שאותם אתם צריכים להגדיר באופן מדוייק ומפורש (זכרו שכל המשתנים המסבירים חייבים לקבל ערכים מספריים).

הסבירו במפורש את המשמעות של כל אחד מהמקדמים.

ג. מתוך הסתכלות על התרשים בלבד, אילו מהמקדמים מהסעיף הקודם צפויים להיות קרובים ל-0 ואילו שונים מ-0? השוו זאת למודל מסעיף א'. ממה נובע ההבדל?

ד. הניחו כעת כי ישנם סוגים רבים של מחרטות והן מחולקות ל-4 קבוצות על פי מחירן. הקבוצות מסומנות בתור 1,2,3,4:

1- עד 2,000 ₪
2- 2,000-5,000 ₪
3- 5,000-9,000 ₪
4- 9,000 ₪ ומעלה

הצעה: מכיוון שמדובר במשתנים אורדינליים בהם יש משמעות למספרי הקבוצות, וכיוון שאנו לא יודעים את המחירים המדויקים של המחרטות, נוכל להכניס לרגרסיה מסעיף א' את סימון הקבוצה כמשתנה מסביר וכך לקבל אומד לאפקט של מהירות המחרטה ($\beta_1$) המתחשב במחיר שלה, מבלי לייצר משתני דמי.
הסבירו בקצרה מה הבעיה בהצעה זו תוך התייחסות למשמעות המקדמים ברגרסיה לינארית.

א). $\hat{\beta}_1$ יכול להיות שונה מאפס בגלל שלא לקחנו בחשבון את ההטרוגניות של המחירים אבל אולי יש בסלב (מה יותר). בנוסף, הכוונה שאין הטיה של מדד המחירים. $\hat{\beta}_1$ יהיה אומד ל- baseline בלבד, כלומר זה אומד לקבוצת הבית שסביר שהם בעלי הכנסה נמוכה ניגודי ($\hat{\beta}_1 X_1$) של שלם ימין בחיים לפי סלב.

ב). נגדיר את המשתנה הבא:

$$X_2 := \begin{cases} 1, & \text{אם } A \\ 0, & \text{אם } B \end{cases}$$

וכן את המשתנה האינטראקציה:

$$X_3 := X_1 \cdot X_2 .$$

אז המודל הסטנדרטי לעניין יהיה:

$$Y = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3 + \epsilon$$

ג). אם נבדוק מה יהיה בערך קו הרגרסיה הפחתית אבור אחוזים A (ישראלים בלבד), ובנפרד מה יהיה בערך קו הרגרסיה הפחתית אבור אחוזים B, אז לנו נראה שהשיפוע של שני הקווים האלה שונים אצלנו. כלומר שלנו נראה שהאינטראקציה "מתבטאת".
פורמלית, נעשה זה באמצעות:

אבור $X_2 = 1$ :

$$Y = \gamma_0 + \gamma_1 X_1 + \gamma_2 + \gamma_3 X_1$$
$$= (\gamma_0 + \gamma_2) + (\gamma_1 + \gamma_3) X_1$$

אבור $X_2 = 0$ :

$$Y = \gamma_0 + \gamma_1 X_1 + \epsilon$$

כלומר שהשיפועים של שני הקווים האלה הם שונים אם $\gamma_3 \neq 0$, כלומר אם המקדם של אינטראקציה $\neq 0$. אם כן אפשר לבדוק בעזרת ה- t-statistic שלמשתנה $X_3$ במודל שהגדרנו.

ד. משמעות השיפוע ברגרסיה לינארית היא בכמה עליה של יחידה אחת במשתנה המסביר משנה את הערך של המשתנה המוסבר. כלומר $\frac{dY}{dX_j} = \beta_j$. כיוון שהשיפוע קבוע לכל $X_j$, חייב להתקיים שאם $X_{ij} = 4$ ו- $X_{kj} = 1$, אז אכן ליחידה $j$ יש פי 4 יותר "משהו" מאשר ליחידה $k$, אותו המשתנה מייצג. במקרה הנוכחי, המרווחים אינם אחידים, וגם מתפרשים על פני טווח רחב של מחירים, ולכן לא מתקיים הקשר הזה.