רגרסיה ומודלים סטטיסטיים - בוחן 2

<u>שאלה 1</u>:

בשאלה זו נחזור לעסוק במודל WLS ונרשום פונקציה המחזירה מחזירה אומדי רגרסיה ומקבלת:

- בסיס נתונים שבו קיים משתנה מוסבר רציף Y ו־p משתנים מסבירים X
- קבוע δ שיהווה תנאי עצירה
- מספר מקסימלי של איטרציות

על הפונקציה לבצע את השלבים הבאים:

1. התאימו מודל רגרסיה לינארית $Y = X\beta + \epsilon$ תחת ההנחה של־ϵ_i יש שונות שווה . חלצו את וקטור השאריות e.
2. כעת הניחו כי $\epsilon_i = a_i\xi_i$ כאשר $\xi_i \sim N(0,1)$ ו־ $a_i = \exp\left(\sum_{j=0}^{p}\gamma_j X_{ij}\right)$. העריכו את הפרמטרים γ_j על ידי מזעור

$$\sum_{i=1}^{n}\left[|e_i| - \exp\left(\sum_{j=0}^{p}\gamma_j X_{ij}\right)\right]^2$$

לשם כך תצטרכו להשתמש באלגוריתם ניוטון־רפסון.

3. אמדו מחדש את β על ידי מזעור

$$\sum_{i=1}^{n}\left[|e_i| - \exp\left(\sum_{j=0}^{p}\hat{\gamma}_j X_{ij}\right)\right]^{-2}\left[Y_i - \sum_{j=0}^{p}\beta_j X_{ij}\right]^2$$

ועדכנו בהתאם את וקטור השאריות e.

4. חזרו על שלבים 2 ו־3 עד ש־ $\max_j|\beta_j^{(m)} - \beta_j^{(m-1)}| \leq \delta$ או עד שתבצעו את מספר האיטרציות המקסימלי שנקבע.

<u>סעיפי השאלה:</u>

א. רשמו את הפונקציה לפי ההנחיות מעלה.

ב. הפעילו את הפונקציה על בסיס הנתונים mtcars מתרגיל 1, כאשר המשתנה המוסבר הוא mpg והמשתנים המסבירים הם wt, disp, qseq, drat. קבעו $\delta = 1e-4$ והגבילו את מספר האיטרציות ב־15. דווחו את האומדים שקיבלתם.

<u>שאלה 2</u>:

בשאלה זו נעסוק במודל ניתוח שונות חד כיווני עם שונויות שונות:

$$Y_{ij} = \mu_i + \epsilon_{ij}, \qquad i = 1, \dots I, \quad j = 1, \dots n_i$$

תחת ההנחות כי $\epsilon_{ij} \sim N\left(0, \sigma_i^2\right)$, ו־ ϵ_{ij} בלתי תלויים לכל i, j. לצורך שאלה זו הניחו גם כי σ_i ידועים לכל i.

נגדיר:

$$\overline{Y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$

$$w_i = \frac{n_i}{\sigma_i^2}, \quad W = \sum_{i=1}^{I} w_i, \quad \pi_i = \frac{w_i}{W}$$

$$\overline{Y}_{..w} = \sum_{i=1}^{I} \pi_i \overline{Y}_{i.}$$

$$\mathrm{SSB}_w = \sum_{i=1}^{I} w_i \left(\overline{Y}_{i.} - \overline{Y}_{..w}\right)^2$$

הוכיחו כי תחת השערת האפס $H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$ מתקיים: $\mathrm{SSB}_w \sim \chi^2_{I-1}$.

<u>רמזים</u>:

1. השתמשו לשם כך במשפט שהוכחנו בכיתה: תהי P מטריצה סימטרית מגודל $n \times n$ מדרגה k המקיימת $P^2 = P$ ויהי $Z \sim N(0, I)$, אזי $\|PZ\|^2 \sim \chi^2_k$.
2. נסחו את הבעיה במונחי רגרסיה, חשבו מה הן המטריצות המתאימות והעזרו בהן לחישוב הסטטיסטי המתאים.

<u>שאלה 3</u>:

בשאלה זו נעסוק ברגרסיה לוגיסטית וננסה לבנות מודל אשר חוזה מי מנוסעי הטיטניק יינצל. הנתונים זמינים בקובץ titanic.csv הנמצא במודל.

קראו את ההסבר לגבי המשתנים בקובץ Titanic Data Dictionary.pdf. המשתנה המוסבר הוא Survived.

חלקו את הנתונים שלכם ל־85% שבהם תשתמשו לבניית המודל (train) ו־ 15% נוספים שעליהם תבדקו את תפקודו בהמשך (test). עבדו מעתה על ה־ train.

א. בצעו ניתוח נתונים ראשוני לפי השלבים שלמדנו והפעילו שיקול דעת לגבי האם ישנם דברים נוספים שתרצו לבדוק. מטרת הניתוח שלכן צריכה להיות להבין אילו משתנים צריכים להיכלל במודל ולהסביר מהם המאפיינים הבולטים של מי ששרד. בפרט, התייחסו לתלויות - האם הסיכויים לשרוד עבור קבוצה כלשהי שנקבעת לפי משתנה אחד משתנים כתלות במשתנה אחר? רשמו את מסקנותיכן. הקפידו לרשום הסברים מפורטים.

ב. בחרו 5 משתנים מסבירים ובנו מודל רגרסיה לוגיסטית באמצעותם. הסבירו מדוע בחרתם במשתנים אלו.

ג. בצעו ניתוח מולטיקולינאריות במודל שבניתם.

ד. בדקו תלויות גם בין המשתנים הרציפים למשתנים הקטגוריאליים. רשמו את הממצאים.

ה. בצעו ניתוח שאריות והסבירו את הממצאים. הסבירו מה מאפיין תצפיות שבהן הטעות יחסית גדולה.

ו. בחנו האם יש טרנספורמציות שניתן לעשות על המשתנים: התייחסו גם למשתני אינטראקציה וגם לאפשרות לאחד משתנים (כמו למשל על ידי סכימה או חיסור בינהם). בצעו את הטרנספורמציות שאתם מציעים והשוו את המודלים המתקבלים איתן ובלעדיהן באמצעות AIC. הסבירו את התוצאות.

ז. כעת בחרו את המודל הטוב ביותר שקיבלתם. נתייחס אליו כאל מסווג: חשבו את תחזיות המודל על ה־test (השתמשו לשם כך בפונקציה predict(model, newdata=test)). בדקו עבור איזה אחוז מהתצפיות המודל שבניתם מספק תחזית נכונה. לשם כך החשיבו כל תצפית שההסתברות שנחזית עבורה גדולה מ־0.5 כתצפית שעבורה חזיתם שהנוסע יינצל. דווחו את התוצאה.