תרגיל בית 4

הנחיות להגשת התרגיל

- יש להגיש שני קבצים נפרדים, אחד עבור הקוד (קובץ ipynb) ואחד לפתרון החלק היבש. בתוך קובץ פתרון החלק היבש הוסיפו גם את הפלט של הקוד.

שאלה 1

א. הוכיחו כי $OR = \frac{odds(E=1|D=1)}{odds(E=1|D=0)}$

ב. השתמשו בסעיף א' כדי למצא אומד ל-OR במדגם Case Control

שאלה 2

הראו והסבירו מה ההשפעה של עליה ביחידה אחת ב-X_1 במודלים הבאים.

1. $\mathbb{E}[Y|X] = \pi\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{1,2} X_1 X_2\right)$
2. $\mathbb{E}[Y|X] = \pi(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2)$

שאלה 3

הנתונים איתם נעבוד בתרגיל זה מתארים ספירה של מקרי פציעות ושברים שהתרחשו במכרות פחם באזור האפלצ'ים במערב וירג'יניה. בסך הכול נאספו 44 תצפיות על מכרות באזור זה. נמדדו ארבעה משתנים הקשורים למאפייני המכרות:

- X1 – Inner burden thickness in feet (INB)
- X2 – Percent extraction of the lower previously mined seam (EXTRP)
- X3 – lower seam height in feet (HEIGHT)
- X4 – Time that the mine has been opened in years (TIME)

הנתונים מצורפים לתרגיל הבית בקובץ mine_fracture.csv

א. כתבו משוואת מודל רגרסיה פואסונית עבור הקשר בין מספר הפציעות לבין ארבעת המשתנים המסבירים האלה.

ב. השתמשו ב statsmodels.GLM או בכל פונקציית ספריה אחרת כדי לאמוד את מקדמי הרגרסיה הפואסונית.

ג. פי כמה משתנה תוחלת מספר הפציעות כאשר גובה השכבה התחתונה (X3) גדל ביחידה אחת? מצאו אומד נקודתי לגודל זה ורווח סמך.

ד. מצאו תת מודל בגודל 2 (מודל עם שני משתנים מסבירים מבין X1, X2, X3, X4) עבורו מתקבל מדד ה-AIC הטוב ביותר. אם אתם נעזרים בפונקציית ספריה, שימו לב כיצד ה-AIC מוגדר בספריה זו ואם צריך למזער או למקסם אותו. לפי מדד AIC, האם עדיף לבחור בתת המודל שמצאתם בסעיף זה או במודל המלא מסעיף א'?

ה. בצעו מבחן שבודק אם המודל המלא מסעיף א' עדיף על המודל המצומצם שמצאתם בסעיף ד'.