רגרסיה ומודלים סטטיסטיים- תרגיל 7

שאלה 1

ידוע שהשכר (Y) בהינתן שנות הלימוד X_1 ושנות הניסיון X_2 מפולג מעריכית כך שמתקיים:

$$E(Y_i) = \exp(X_i^T \beta)$$

א. הראו שמופרת הנחת שיוויון השונויות, והשתמשו בשיטת הדלתא על מנת למצוא טרנספורמציה מייצבת שונות מתאימה g.

ב. האם הנחת הלינאריות מתקיימת לאחר הטרנספורמציה? כלומר האם מתקיים ש- $E\big(g(Y)\big) = X\beta$?

שאלה 2

א. הניחו את המודל:

$$Y = X\beta + \epsilon, \epsilon_i \sim N\big(0, \sigma_i^2\big), \qquad cov\big(\epsilon_i, \epsilon_j\big) = 0 \; \forall i \neq j$$

מצאו את $E\big(\hat{\beta}_{OLS}\big), cov\big(\hat{\beta}_{OLS}\big)$ במקרה הזה.

ב.

הניחו כי במקום לאמוד את $\vec{\beta}$ על ידי אומדי הריבועים הפחותים כפי שנעשה בכיתה, נרצה לאמוד את $\vec{\beta}$ על ידי מזעור של ריבועים ממשוקלים:

$$S_w(b) = \sum_{i=1}^{n} w_i \left(y_i - \sum_{j=1}^{p} X_{ij} b_j \right)^2$$

עבור ערכים חיוביים נתונים w_i. הוכיחו שהפתרון נתן על ידי: $\hat{\beta}_w = (X^T W X)^{-1} X^T W Y$ כאשר W היא מטריצה אלכסונית וערכי האלכסון הם w_i. הסבירו מתי נרצה לבצע אמידה באופן זה.

ג. הראו כי תחת הנחות המודל מסעיף א' $\hat{\beta}_{OLS}$ אינו $BLUE$.

ד. הניחו כעת את המודל:

$$Y = X\beta + \epsilon, \qquad \epsilon_1 = Z_1, \epsilon_i = 0.5 \cdot Z_{i-1} + Z_i, \qquad Z_i \sim N(0,1) \; i.i.d$$

מצאו את האומד הלינארי חסר ההטיה ל-β הטוב ביותר במודל הזה.

שאלה 3

נסמן ב-X_i את השורה ה-i של המטריצה X וב-$X_{(i)}$ את המטריצה X ללא השורה ה-i. ובאופן דומה את $Y_i, Y_{(i)}$.

ובהתאמה נסמן את $\hat{Y}_{j(i)}$ להיות הכניסה ה-j של וקטור הערכים החזויים המתקבל ללא הזוג (X_i, Y_i)

בתרגול הגדרנו את מרחק $Cook$:

$$D_i = \frac{\sum_{j=1}^{n} \big(\hat{Y}_j - \hat{Y}_{j(i)}\big)^2}{(p+1) \cdot \widehat{\sigma^2}}$$

בשאלה זו נראה כי:

$$D_i = \frac{e_i^2}{(p+1)\cdot \widehat{\sigma^2}} \cdot \frac{(P_{X_{ii}})}{\left(1-P_{X_{ii}}\right)^2}$$

1. מדוע שנעדיף לחשב את מרחק $Cook$ בדרך השניה ולא הראשונה?
2. הוכיחו את הטענה על פי ההדרכה הבאה:

א. הראו כי:

$$\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2 = \left|\left|\hat{Y} - \hat{Y}_{(i)}\right|\right|^2 = \left(\hat{\beta}_{(i)} - \hat{\beta}\right)^T X^T X\left(\hat{\beta}_{(i)} - \hat{\beta}\right)$$

ב. הראו כי:

$$X^T Y = X_{(i)}^T Y_{(i)} + X_i Y_i$$
$$X^T X = X_{(i)}^T X_{(i)} + X_i X_i^T$$

ג. **תזכורת (נוסחת שרמן-מוריסון):**

עבור A מטריצה הפיכה ו- uv^T מטריצה מדרגה 1:

$$(A - uv^T)^{-1} = A^{-1} + \frac{A^{-1}uv^T A^{-1}}{1 - v^T A^{-1} u}$$

השתמשו בנוסחה זו על מנת להראות:

$$\left(X_{(i)}^T X_{(i)}\right)^{-1} = (X^T X)^{-1} + \frac{(X^T X)^{-1} X_i X_i^T (X^T X)^{-1}}{1 - X_i (X^T X)^{-1} X_i}$$

ד. סמנו $a := (X^T X)^{-1} X_i \in R^{p+1}$, הציבו בסעיף הקודם וקבלו כי:

$$\left(X_{(i)}^T X_{(i)}\right)^{-1} X_i = \frac{a}{1 - X_i^T a} = \frac{(X^T X)^{-1} X_i}{1 - P_{X_{ii}}}$$

ה. באמצעות הסעיף הקודם, פתחו את הביטוי $\left(\hat{\beta}_{(i)} - \hat{\beta}\right)$ והראו כי:

$$\left(\hat{\beta}_{(i)} - \hat{\beta}\right) = \left(X_{(i)}{}^T X_{(i)}\right)^{-1} X_{(i)}^T Y_{(i)} - \hat{\beta} = \frac{(X^T X)^{-1} X_i}{1 - P_{X_{ii}}}\left(X_i^T \hat{\beta} - Y_i\right)$$

ו. הציבו את הביטוי שקיבלתם בביטוי מסעיף א' כדי לקבל:

$$\left(\hat{\beta}_{(i)} - \hat{\beta}\right)^T X^T X\left(\hat{\beta}_{(i)} - \hat{\beta}\right) = e_i^2 \cdot \frac{P_{Xii}}{(1 - P_{Xii})^2}$$

והסיקו את מה שהיה צריך להוכיח.

שאלה 4

הריצו את הסימולציה הבאה ב-R:

```
set.seed(123)

n <- 100

x <- rnorm(n, mean = 50, sd = 5)

y <- 5 + 2 * x + rnorm(n, sd = 5)

x[c(99, 100)] <- c(50, 20)

y[c(99, 100)] <- c(10, 50)

model <- lm(y ~ x)
```

א. חשבו את המטריצה P_X. האם אתם מאתרים תצפיות שעשויות להיות חריגות על פי ההנפה (מינוף) שלהן?

ב. שרטטו גרף פיזור ואת ישר הרגרסיה של Y על X. מה דעתכם על טיב ההתאמה של המודל? האם אתם מאתרים תצפיות שעשויות להיות חריגות?

ג. חשבו את מרחק $Cook$ של כל אחת מהתצפיות. האם אתם מאתרים תצפיות חריגות? האם הן זהות לתצפיות מסעיף א'?

ד. חזרו על סעיף ב' פעם אחת ללא התצפית ה-99 ופעם אחת ללא התצפית ה-100. האם טיב ההתאמה של המודל השתפר? הסבירו את התוצאות תוך התייחסות לערכים שחישבתם בסעיפים א' ו-ג'.

ה. הניחו כי X הוא מספר המשלוחים שהוזמנו מחברה מסויימת בשעה, ו-Y הוא זמן המשלוח הכולל לכל ההזמנות הללו בדקות. האם הייתם מורידים את התצפיות החשודות כחריגות? הניחו כי X הוא מספר המוצרים שמכרו חברות ביום ו-Y הוא סך ההכנסות שלהן במאות שקלים. האם כעת הייתם מורידים את התצפיות החשודות כחריגות? נמקו.