רגרסיה- תרגול 8

בדיקת הנחת הנורמליות, התמודדות עם הפרת הנחות המודל ותצפיות חריגות

שאלה- היסטוגרמה ואחוזון:

א. יהיו $X_1, \dots, X_n \sim F$ מ"מ ב"ת ש"ה. לשם פשטות הניחו כי קובעים את תאי ההיסטוגרמה מראש ב-K נקודות ידועות: $x_1 < x_2 < \dots < x_K$. נגדיר ב-n_j את מספר התצפיות בתא ה-j. הראו כי לוקטור:

$(n_1, \dots, n_k)^T$ יש התפלגות מולטינומית ומצאו את הפרמטרים שלו.

ב. נגדיר את האחוזון האמפירי ה-p:

$$\hat{Q}_p = X_{[n \cdot p]}$$

כלומר סטטיסטי הסדר הקטן ביותר שמקיים ש- $n \cdot p$ מהתצפיות קטנות ממנו או שוות לו.

הסבירו מדוע זהו אומד הגיוני לאחוזון ה-p.

ג. הניחו כי $F = unif(-\sqrt{3}, \sqrt{3})$, הניחו כי $p = 0.025$. חשבו את $\Phi^{-1}(0.025)$ והשוו זאת ל-$F^{-1}(0.025)$. הסבירו כיצד זה מתקשר לבדיקת הנחת הנורמליות על ידי $QQ - plot$.

פתרון:

א. נגדיר $x_0 := -\infty$. במקרה הזה מספר התצפיות בתא ה-j יהיה $n_j = \sum_{i=1}^{n} 1_{\{X_i \in (x_{i-1}, x_i)\}}$. נשים לב שבמקרה כזה מדובר בקטעים זרים ומתקיים $\sum_{j=1}^{k} n_j = n$.
נסמן $b_{ij} = 1_{\{X_i \in (x_{i-1}, x_i)\}}$. אז כיוון שהקטעים זרים ו-X_iים ב"ת, נקבל כי זהו ניסוי עם n חזרות ו-k קטגוריות אפשריות כך שמתקיים ש- $b_{ij} \sim Ber\left(p\left(X_i \in (x_{i-1}, x_i)\right)\right)$. ולכן $n_j = \sum_{i=1}^{n} b_{ij}$ מפולג בינומי, והוקטור מפולג מולטינומי. הפרמטרים הם $n, P\left(\left(X_i \in (x_{i-1}, x_i)\right)\right) = F(x_i) - F\left(x_{(i-1)}\right)$.

ב. פורמלית, האחוזון האמפירי מקיים:

$$\hat{Q}_p = \inf\left\{t \left| \frac{\sum_{i=1}^{n} 1_{X_i \le t}}{n} \ge p \right.\right\}$$

מהחוק החלש זהו אומד עקיב ל- $\inf\{t | P(X_i \le t) \ge p\}$, כלומר לאחוזון האמיתי (במקרה הרציף:

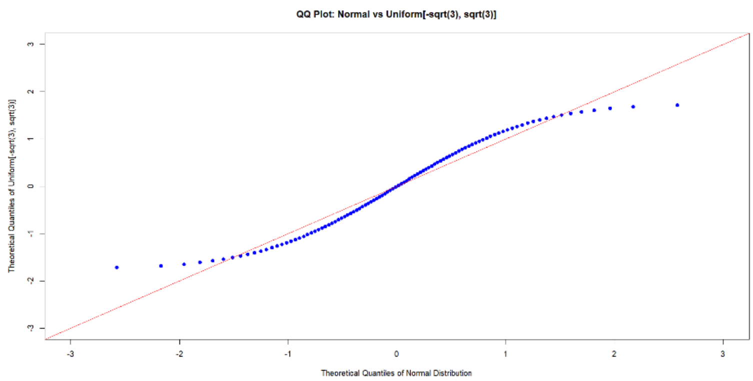
$F^{-1}(p)$ כמו $qnorm(p)$ למשל.

ג. $qnorm(0.025) = -qnorm(0.975) = -1.96$

נסמן $F^{-1}(0.025) = q$. אז:

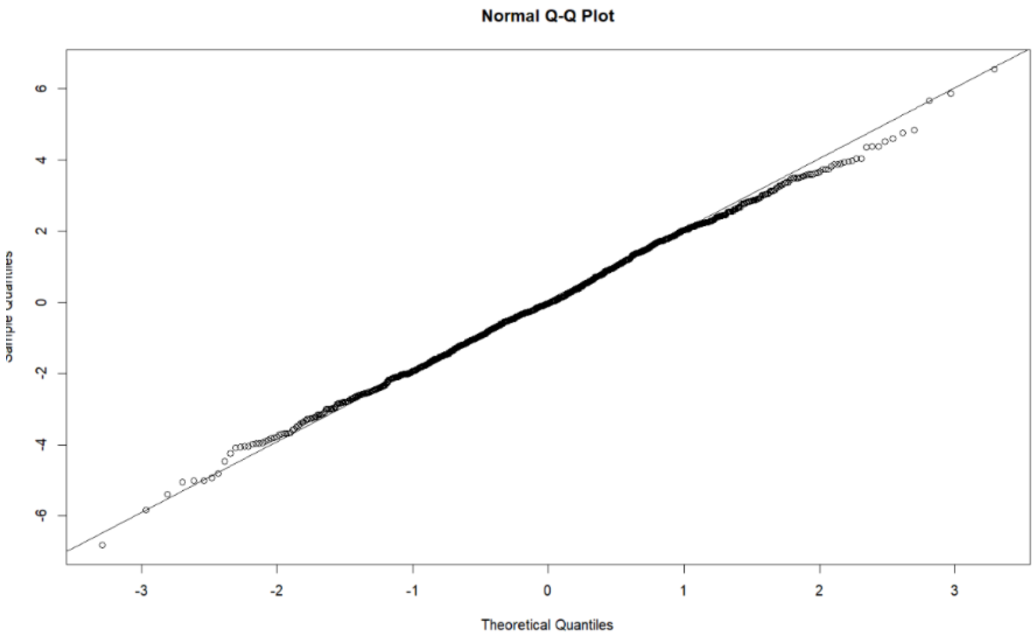
$$0.025 = P(X_1 \le q) = \frac{q + \sqrt{3}}{2\sqrt{3}} \Leftrightarrow q = -1.645448$$

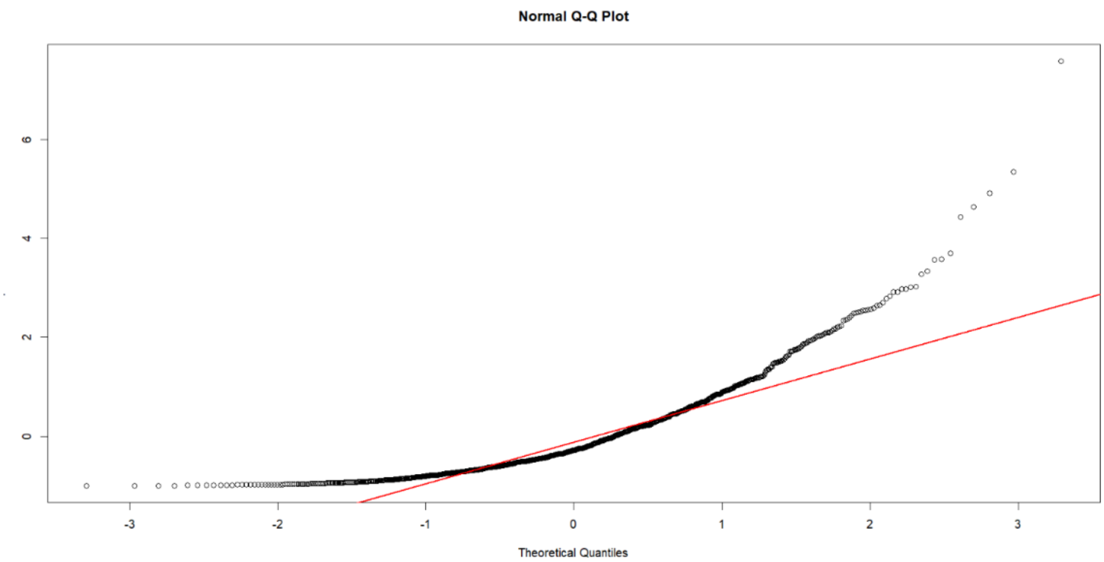
כלומר ה-2.5 אחוזונים הראשונים נצברים מוקדם יותר בהתפלגות הנורמלית מאשר בהתפלגות היוניפורמית וזאת כיוון שהסתברותם של ה"זנבות" בהתפלגות היוניפורמית אפסית, לעומת ההתפלגות הנורמלית שצפיפותה חיובית לכל נקודה בישר הממשי. בהקשר של בדיקת הנחת הנורמליות נצפה לראות שיש לנו "חוסר" בתצפיות בזנבות בהתפלגות היוניפורמית- כלומר בעבור ערכי ה-X הקטנים נהיה מעל הישר $Y = \frac{1}{\sigma} * X$ ואילו בערכים הגדולים נצפה לראות ריכוז תצפיות מתחת לישר:

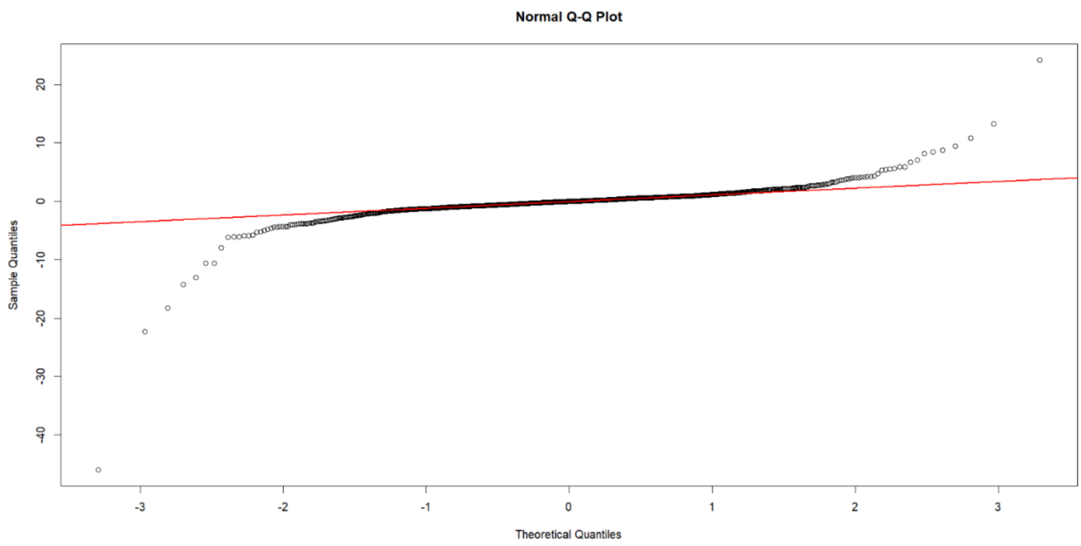


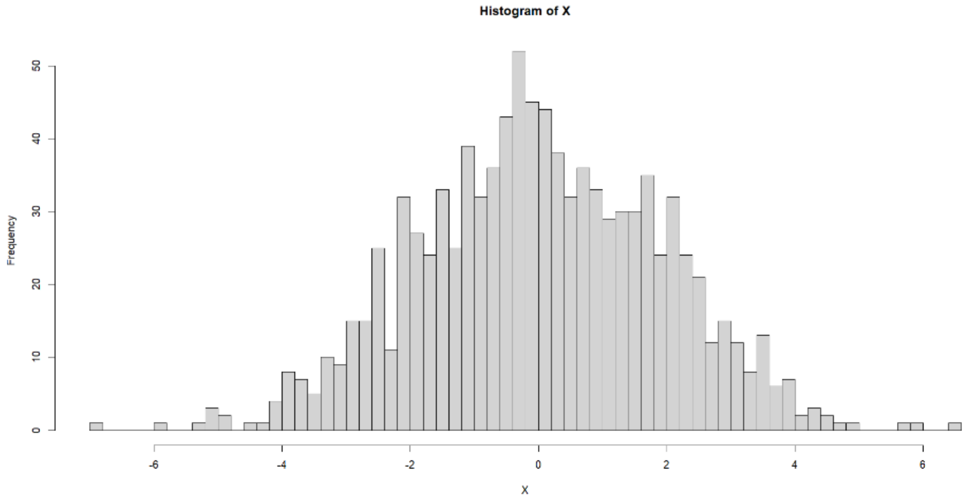
שאלה

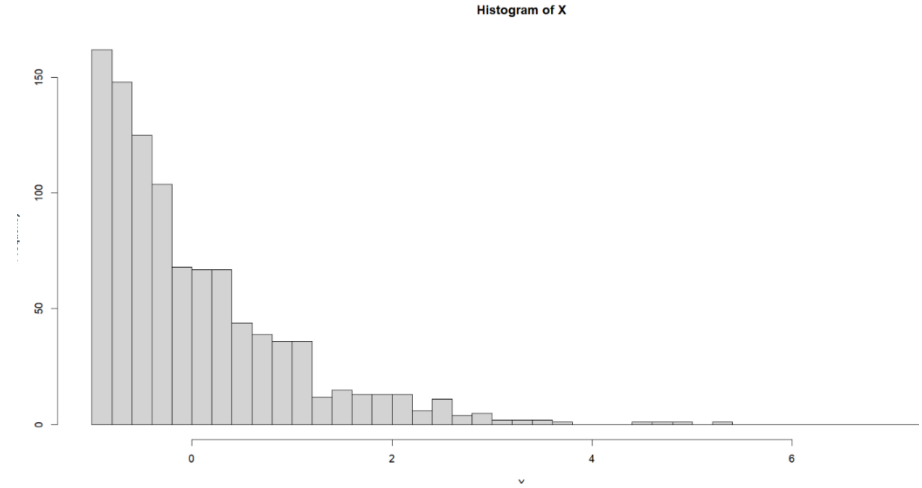
לפניכם תרשימי $QQ-plot$ של נתונים מהתפלגויות שונות. תארו כיצד תיראה ההיסטוגרמה של כל אחד מהמדגמים ביחס להיסטוגרמה של נתונים שהגיעו מהתפלגות נורמלית :

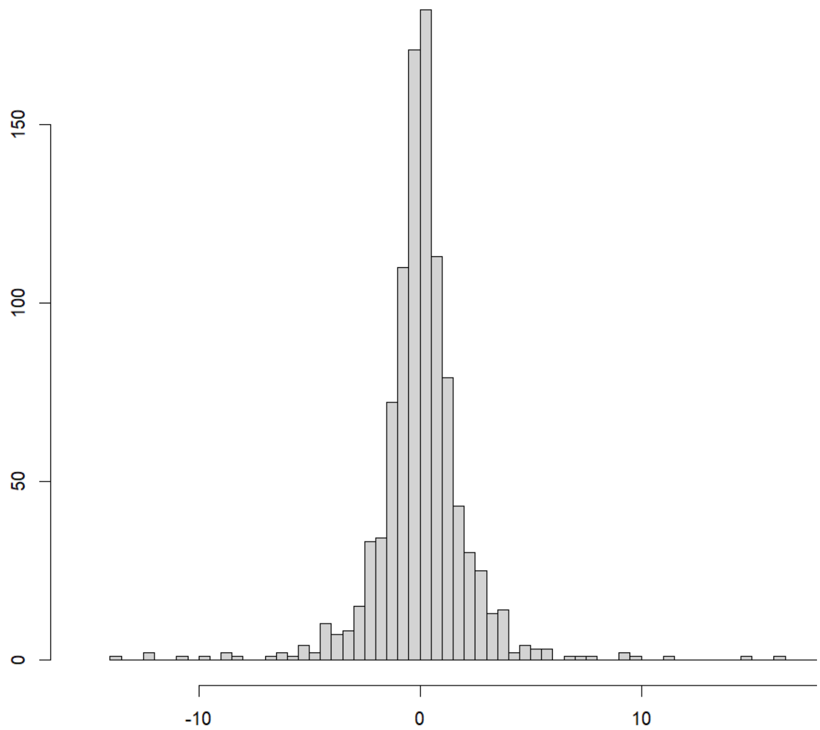







Histogram of X
Frequency
0
10
20
30
40
50
-6
-4
-2
0
2
4
6
X


Histogram of X
0
50
100
150
0
2
4
6


0
50
100
150
-10
0
10

גישות לתיקון הנחות המודל

הפרת הנחת הלינאריות:

דוגמא/שאלה

נניח שאנו מעוניינים לאמוד את המודל הבא:

Y_i- מספר הצרכנים שרכשו מוצרים בחודש ה-i.

X_{1i}— תקציב השיווק החודשי של החנות.

X_{2_i}- מספר ימי הגשם בחודש.

יהיה הגיוני להניח כי מתקיים:

$$Y \sim Pois(\lambda), \qquad \lambda = e^{X\beta}$$

כלומר מודל מהצורה:

$$Y_i = e^{X_i^T\beta}$$

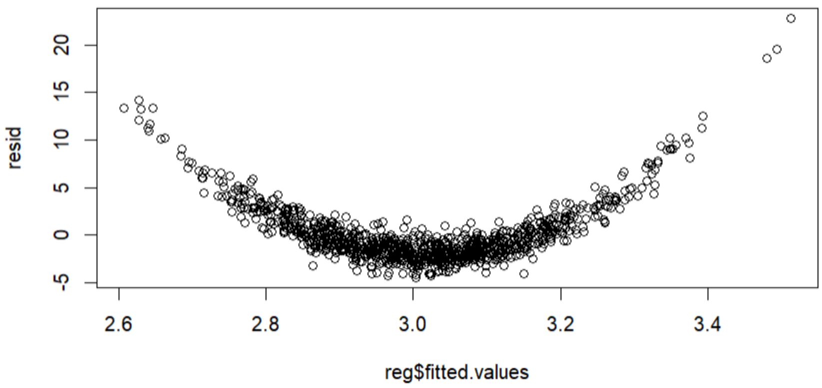
א. מה היא משמעות המקדמים במודל? בפרט, נניח שמספר ימי הגשם בחודש עלה מ-0 ל-1. כיצד הדבר צפוי להשפיע על מספר הצרכנים בחודש?
ב. הצעה: כיוון ש-$E(Y|X) = e^{X\beta}$, נאמוד מודל לוג-לינארי: $\log(Y) = X\beta$. מה דעתכם?
ג. הראו שהנחת שיוויון השונויות מופרת והציעו טרנספורמציה מייצבת שונות לתקנה.

הערה: לפעמים משתמשים בטרנספורמציית לוג כדי לתת פרשנות של שינוי באחוזים עבור כל שינוי ביחידה של X. שימו לב שזה נכון רק בעבור מקדמים קטנים יחסית (נובע מקירוב טיילור):

	b	exp_b
1	0.00	0.00
2	0.02	2.02
3	0.04	4.08
4	0.06	6.18
5	0.08	8.33
6	0.10	10.52
7	0.12	12.75
8	0.14	15.03
9	0.16	17.35
10	0.18	19.72
11	0.20	22.14
12	0.22	24.61
13	0.24	27.12
14	0.26	29.69
15	0.28	32.31
16	0.30	34.99

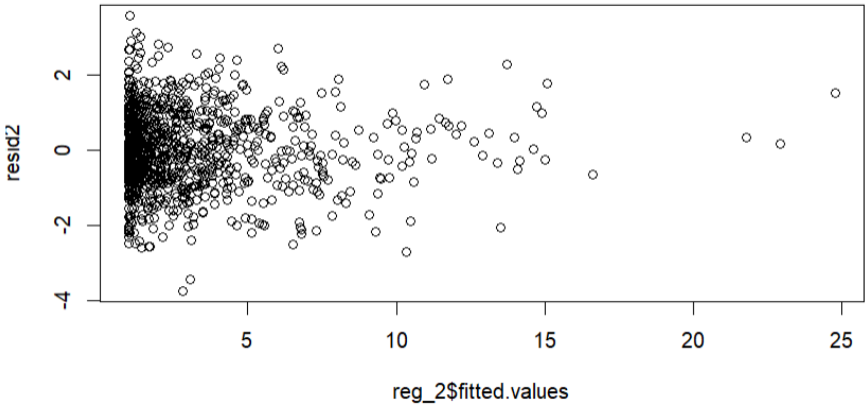
שאלה

מרגרסיה לינארית רגילה של Y על X_1 התקבל גרף השאריות על הערכים החזויים הבא:



א. האם יש אינדיקציה להפרה של הנחת הלינאריות? האם של הנחת שיוויון השונויות?

ב. הציעו טרנספורמציה לתיקון ההפרה מהסעיף הקודם.



שאלה- טרנספורמציה מייצבת שונות:

א. הוכיחו כי אם Y בינארי, אז הנחת שיוויון השונויות בהכרח מופרת.

ב. הציעו מקרה בו הנחת השונויות מופרת, אך ניתן לבצע טרנספורמציה מייצבת שונות, כלומר לאמוד משתנה שלאחר הטרנספורמציה עליו, מתקבלות שונויות שוות.

שאלה - תצפיות חריגות

תזכורת:

מדד המינוף (*leverage*) של התצפית ה-i הינו:

$$cov(Y_i, \hat{Y}_i)$$

א. הראו שניתן לכתוב $\hat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j$ כך שמתקיים $cov\left(Y_j, \hat{Y}_i\right) = \sigma^2 \cdot h_{ij} = \sigma^2 \cdot \frac{\partial \hat{Y}_i}{\partial Y_j}$.

ב. הראו שמתקיים: $1 \geq h_{ii} \geq 0$

ג. תנו פרשנות לתוצאות מסעיפים א' וב'.

ד. הראו שברגרסיה פשוטה מתקיים:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2}$$

פתרון:

א.

מרחק Cook

בעבור התצפית ה-i נגדיר את מרחק Cook :

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

בתרגיל תראו :

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

א. תנו פרשנות למדד? כיצד הוא עשוי לסייע באיתור תצפיות חריגות?

שאלה ממבחן:

2. מדען החוקר את השפעת צריכת הקוקאין של עכברים על התנהגותם ערך ניסוי עבור $n = 19$ עכברים ובו נתן לכל עכבר i כמות של $x_i = i$ מיקרוגרם קוקאין. המדען מדד את מספר הקפיצות של העכברים בדקה y_i , קיבל את התצפיות $(x_1, y_1), ..., (x_{19}, y_{19})$ והתאים מודל רגרסיה לינארית פשוטה **עם חותך** $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ לתצפיות.

ידוע שהתפלגות מספר הקפיצות בדקה של עכברים y_i תלויה בכמות הסם במיקרוגרם אותו צרכו x_i באופן הבא: $y_i \sim Binom(10, \frac{x_i}{20})$ וכן ה־y_i עבור עכברים שונים הם בלתי תלויים.

(א) [6 נקודות] האומד לשיפוע $\hat{\beta}_1$ (עבור הערך האמיתי $\beta_1 = \frac{1}{2}$)הוא:

i. חסר הטיה ותמיד אי שלילי
ii. חסר הטיה אך יכול להיות שלילי
iii. מוטה ותמיד אי שלילי
iv. מוטה ויכול להיות שלילי

(ב) [6 נקודות] עבור ההנפות של תצפיות מספר $1, 10$:

i. ההנפה של תצפית 10 גדולה יותר
ii. ההנפה של תצפית 1 גדולה יותר
iii. שתי ההנפות שוות
iv. לא ניתן לדעת ללא עוד נתונים

(ג) [5 נקודות] עבור מרחקי $Cook$ של תצפיות מספר $1, 10$:

i. מרחק ה־$Cook$ של תצפית 10 גדול יותר
ii. מרחק ה־$Cook$ של תצפית 1 גדול יותר
iii. שני מרחקי ה־$Cook$ שווים
iv. לא ניתן לדעת ללא עוד נתונים

(ד) [8 נקודות] לאחר שצפה בפאודה, עוזר המחקר הערני שם לב שעכברים מספר $1,19$ הם בעצם גרבילים מסתעכברים ולכן החליט להסיר אותם מהמחקר ולנתח את הנתונים מחדש בלעדיהם, כלומר עבור טבלת נתונים חדשה עם 17 עכברים. סמנו עבור כל נתון כיצד ישתנה בעקבות ההסרה:

i. $\bar{x}$: בהכרח יגדל/בהכרח לא ישתנה/בהכרח יקטן/ אחרת
ii. $\bar{y}$: בהכרח יגדל/בהכרח לא ישתנה/בהכרח יקטן/ אחרת
iii. $\hat{\beta}_1$: בהכרח יגדל/בהכרח לא ישתנה/בהכרח יקטן/ אחרת
iv. ההנפה הממוצעת : בהכרח תגדל/בהכרח לא תשתנה/בהכרח תקטן/ אחרת

(ה) [7 נקודות] חשבו טרנספורמציה לייצוב השונות עבור מספר הקפיצות של העכברים. תוכלו להשתמש בנוסחא: $\int \frac{1}{\sqrt{t(1-at)}}dt = \frac{2}{\sqrt{a}}sin^{-1}\left(\sqrt{at}\right)$

ב.

שאלה- ריבועים פחותים מוכללים:

3. עבור וקטור מקרי $Y \in R^n$ נניח את המודל הלינארי:

$$Y = X\beta + \epsilon, \quad \epsilon \sim N(0, \Sigma)$$

עבור מטריצה X קבועה וידועה מדרגה $p<n$, ו ϵ המקיים $E[\epsilon] = \mathbf{0}$ ו $Cov[\epsilon] = \Sigma$.
(Σ מוגדרת חיובית).

א. ניתן לבטא את וקטור השאריות $Y - X\hat{\beta}_{OLS} = H \cdot Y$ עבור H מטריצת הטלה nxn.
מצאו את H, הוכיחו שזו מטריצת הטלה, והראו שדרגתה n-p.

ג. הניחו ש $\Sigma = \sigma^2 I_n$.
מצאו את התפלגות של $\left\|Y - X\hat{\beta}_{OLS}\right\|^2$. ניתן להשתמש בתוצאות של סעיף א׳, ב׳.

ד. כעת הניחו Σ מטריצה כללית.
הוכיחו שההתפלגות של $T = \left|\left|Y - X\hat{\beta}_{GLS}(\Sigma)\right|\right|_\Sigma$ איננה תלויה בΣ.
(שימו לב שמדובר בנורמת מהלנוביס לפי מטריצת Σ)

הערה : נורמת מהלנוביס בין שני וקטורים Z, W לפי המטריצה V מוגדרת להיות :

$$(Z - W)^T V^{-1} (Z - W)$$

בעבור V מטריצה סימטרית חיובית.

במקרה הספציפי בו V היא המטריצה $Cov(Z, W)$ נקבל שזהו מדד למרחק בין שני וקטורים הלוקח בחשבון את המתאם ביניהם. כלומר אם מתואמים חיובית, המרחק שיתקבל ביניהם יחשב לקטן יותר.

פתרון :