רגרסיה- תרגול 7- תשפ"ה

שאלה (משתנה אינטראקציה):

נניח שבידינו נתונים על שתי קבוצות, ומשתנה תוצאה Y.

א. הראו שהאמידות הבאות שקולות (תחת הנחות המודל הלינארי) ובטאו כל אחד מהמקדמים :

1. אמידת שני מודלים נפרדים :

$$Y_i^j = \beta_0^j + \beta_1^j {X_1}_i^j + \epsilon_i^j, j \in \{1,2\}$$

2. אמידת המודל המשותף :

$$Y_i = \gamma_0 + \gamma_1 \cdot 1_{\{i\in A\}} + \gamma_2 {X_1}_i + \gamma_3 X_{1i} \cdot 1_{\{i\in A\}}$$

ב. הסבירו מדוע לא ניתן לייצג אף אחד מהמקדמים במודלים לעיל על ידי מקדמי המודל :

$$Y_i = \alpha_0 + \alpha_1 \cdot 1_{\{i\in A\}} + \alpha_2 X_{1i} + \epsilon_i$$

פתרון :

א. נזכור כי תחת הנחות המודל הלינארי אומדי OLS הם חסרי הטיה. כיוון שלכל פרט בקבוצה יש ייצוג מלא בשתי הדרכים נקבל כי לכל ערך X_{i1} חייב להתקיים :

$$(1,1,X_{i1},X_{i1})(\gamma_0,\gamma_1,\gamma_2,\gamma_3)^T = \gamma_0 + \gamma_1 + \gamma_2 {X_1}_i + \gamma_3 X_{1i} = E(Y_i | i\in A) = E(Y_i^A)$$
$$= (1,X_{i1})(\beta_0^A,\beta_1^A) = \beta_0^A + \beta_1^A X_{i1}$$

ניקח $X_{i1} = 0$ ונקבל $\beta_0^A = \gamma_0 + \gamma_1$. נציב ונקבל כי $\beta_1^A = \gamma_2 + \gamma_3$.

באותו האופן בעבור קבוצה B:

$$(1,0,X_{i1},0)(\gamma_0,\gamma_1,\gamma_2,\gamma_3)^T = \gamma_0 + \gamma_2 {X_1}_i = E(Y_i | i\in A) = E(Y_i^B) = (1,X_{i1})(\beta_0^B,\beta_1^B)$$
$$= \beta_0^B + \beta_1^B X_{i1}$$

נקבל כי $\beta_0^B = \gamma_0$ וכן $\beta_1^B = \gamma_2$.

ב. זאת מכיוון שאין ייצוג מלא. במודל הזה אין אפשרות לשיפוע שונה בעבור כל קבוצה, לכן אין דרך לייצג את מקדמי השיפועים באמצעות מקדמי המודל השלישי.

שאלה- בוחן אמצע- תשפ"ד:

זמן שבועי (בדקות) המוקדש לפעילות גופנית = weekly sport time

קבוצת גיל: מבוגרים, ילדים, קשישים = group: adults (A), children (C), elderly (E)

ואת משתנה התוצאה

לחץ דם = Y = blood pressure

מצורף תרשים פיזור של הנתונים לפי קבוצת גיל.

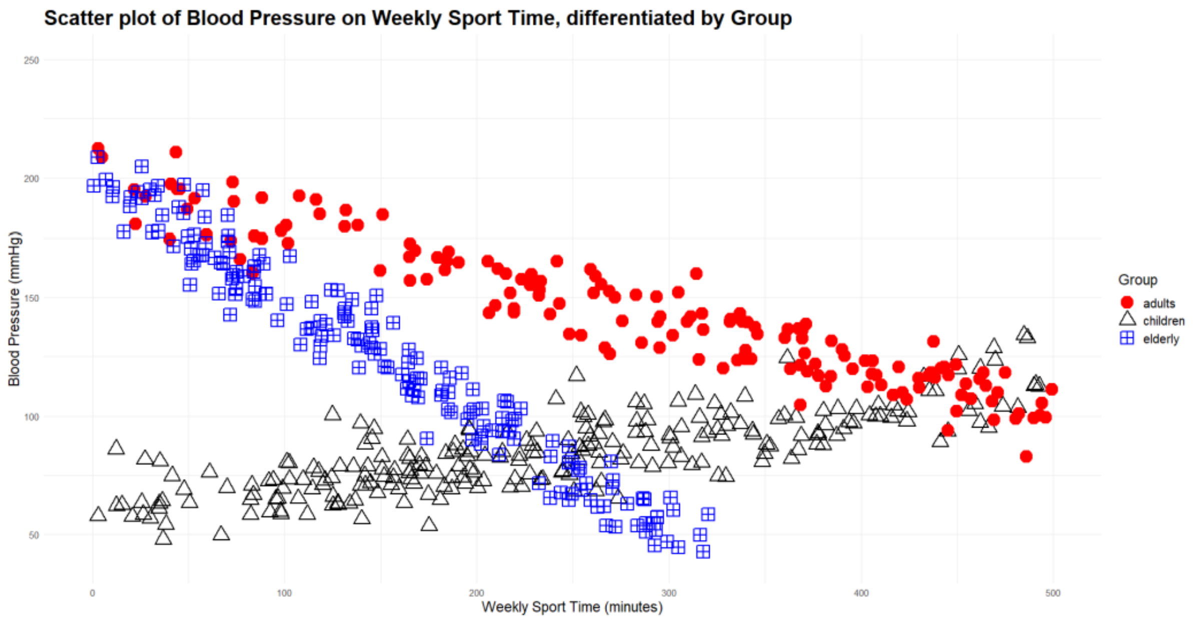
ה. (15 נק׳) מהסתכלות ראשונית על התרשים בלבד: האם יש אינדיקציה ברורה לאינטראקציה בין קבוצת הגיל ובין זמן הפעילות הגופנית בדקות? האם יש אינדיקציה ברורה לחותך שונה עבור כל אחת מהקבוצות? הסבירו **בקצרה**.

ו. (15 נק׳) אנחנו רוצים לבדוק אם זמן הפעילות הגופנית משפיע על לחץ הדם של מבוגרים (A) וקשישים (E) באותו האופן, כלומר, שאותה עלייה בלחץ הדם לכל דקת פעילות נוספת צפויה עבור מבוגרים ועבור קשישים. מהי מטריצת $\boldsymbol{X}$ המתאימה במודל הליניארי?

ז. (15 נק׳) נסחו את השאלה שבה מתעניינים בסעיף ה׳ בתור השערת אפס פורמלית (במונחי הפרמטרים של המודל).

ח. הניחו מודל חלופי בו הגיל נתון באופן רציף- ללא חלוקה לקבוצות. מה הפרשנות של כל אחד מהמקדמים במודל:

$$Y_i = \beta_0 + \beta_1 \cdot Sport_i + \beta_2 \cdot Age_i + \beta_3 \cdot Sport_i \cdot Age_i + \epsilon_i$$



פתרון:

ה. יש אינדיקציה ברורה לאינטראקציה בעבור כל שלוש הקבוצות- נראה שאם היינו אומדים בנפרד 3 קווי רגרסיה, אחד בעבור כל קבוצת גיל (ראינו שזה שקול), היינו מקבלים שיפוע שונה בכל קבוצה. אין אינדיקציה ברורה להבדלים בחותכים של הקווים בקבוצת המבוגרים והקשישים, אך כן חותך שונה ונמוך הרבה יותר בעבור קבוצת הילדים.

ו. ראינו בתרגול כי אמידה של שתי (או 3 במקרה הזה- כי יש 3 קבוצות בנתונים) רגרסיות נפרדות, שקולה לאמידת המודל עם אינטראקציה וחותך נפרד לכל קבוצה. לכן כל אחד מהמודלים הבאים יתקבל ומתאים:

.1

$$Y_i = \beta_0 + \beta_1 \cdot 1_{\{i\in E\}} + \beta_2 \cdot 1_{\{i\in C\}} + \beta_3 \cdot S_i + \beta_4 \cdot S_i \cdot 1_{\{i\in E\}} + \beta_5 \cdot S_i \cdot 1_{\{i\in c\}} + \epsilon_i$$

.2

$$Y_i = \beta_0 + \beta_1 \cdot 1_{\{i\in E\}} + \beta_2 \cdot 1_{\{i\in A\}} + \beta_3 \cdot S_i + \beta_4 \cdot S_i \cdot 1_{\{i\in E\}} + \beta_5 \cdot S_i \cdot 1_{\{i\in A\}} + \epsilon_i$$

המטריצה X המתאימה עבור המודל הראשון (נניח שישנם 3 קשישים ו-3 ילדים והמטריצה מסודרת לפי קשישים, מבוגרים, ילדים):

$$\begin{matrix} 1 & 1 & 0 & S_1 & S_1 & 0 \\ 1 & 1 & 0 & S_2 & S_2 & 0 \\ 1 & 1 & 0 & S_3 & S_3 & 0 \\ 1 & 0 & 0 & S_4 & 0 & 0 \\ 1 & 0 & 0 & S_5 & 0 & 0 \\ 1 & 0 & 0 & S_6 & 0 & 0 \\ 1 & 0 & 0 & S_7 & 0 & 0 \\ 1 & 0 & 1 & S_8 & 0 & S_8 \\ 1 & 0 & 1 & S_9 & 0 & S_9 \\ 1 & 0 & 1 & S_{10} & 0 & S_{10} \end{matrix}$$

ז. במודל 1, נבדוק את ההשערה:

$$H_0: \beta_4 = 0 \ vs \ H_1: \beta_4 \neq 0$$

במודל 2, נבדוק את ההשערה:

$$H_0: \beta_5 - \beta_4 = 0 \ vs \ H_1: \beta_5 - \beta_4 \neq 0$$

ח. β_0- לחץ הדם החזוי עבור אדם בגיל 0, שלא מתאמן כלל. חסר משמעות מיידית.
β_1- ההשפעה של דקת ספורט על לחץ הדם החזוי, עבור אדם בגיל 0- כלומר ההשפעה שלא תלויה בגיל כלל.

β_2-ההשפעה של תוספת שנת חיים ללחץ הדם החזוי, עבור אדם שלא מתאמן כלל.
β_3-<u>השינוי</u> בהשפעה של תוספת שנת חיים ללחץ הדם החזוי עם כל דקה נוספת של ספורט. או באופן סימטרי- השינוי בהשפעת דקת אימון עם כל שנה של הגיל.

דוגמא: נניח $\beta_0 = 70, \beta_1 = -0.01, \beta_2 = 0.8, \beta_3 = -0.005$. נקבל:

Age	Sport_Minutes	Predicted_BP
20	0	86
20	30	82.7
20	60	79.4
60	0	118
60	30	108.7
60	60	99.4
100	0	150
100	30	134.7
100	60	119.4

כלומר β_3 במקרה הזה הוא האפקט "הממתן" של השפעת הגיל על לחץ הדם עבור כל דקת אימון.

שאלה - העשרה:

יישומים של משתני דמי בניסויים טבעיים:

1. *Difference-in-Differences* (ישנו יישום דומה גם בניסוי מבוקר):

נניח שאנו רוצים לבדוק אפקט של טיפול/אירוע מסויים, ויש לנו תצפיות על קבוצת הטיפול ועל קבוצת הביקורת לאורך זמן, לפני ואחרי הטיפול. הקבוצות לא בהכרח חייבות להיות זהות אך המגמות בטרם הטיפול זהות, והנחה נדרשת היא שאילולא הטיפול הן היו ממשיכות להיות זהות.
המטרה בשיטה זו היא לבודד את השפעת הזמן שעשוייה להיווצר ולהשפיע (הנחה) באופן זהה על שתי הקבוצות, ואת ההטייה שעשויה להתעורר מעצם המאפיינים הייחודיים בין שתי הקבוצות.
לשם פשטות, נניח כאן שישנה תקופה אחת לפני הטיפול ותקופה אחת לאחריה. אז במקרה כזה נוכל להגדיר את המודל:

$$Y_{it} = \beta_0 + \beta_1 \cdot Treated_i + \beta_2 \cdot Post_t + \beta_3 \cdot Post_t \times Treated_i + \epsilon_{it}$$

כאשר:

$Treated_i$ - משתנה דמי המקבל 1 אם התצפית שייכת לקבוצת הטיפול (לפני או אחרי שקיבלה את הטיפול).

$Post_t$ - משתנה דמי המקבל 1 אם התצפית שייכת לתקופה שאחרי שניתן הטיפול (גם אם בקבוצת הטיפול וגם אם בקבוצת הביקורת).

מה יהיה האומד לאפקט של הטיפול?

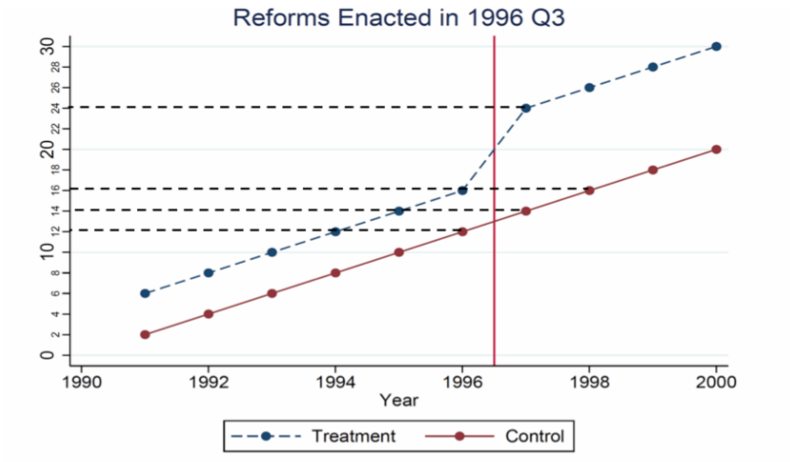
נסתכל על התוחלות של כל אחת מהקבוצות לפני ואחרי הטיפול:

$$E(Y_{10}) = \beta_0 + \beta_1$$
$$E(Y_{11}) = \beta_0 + \beta_1 + \beta_2 + \beta_3$$
$$E(Y_{00}) = \beta_0$$
$$E(Y_{01}) = \beta_0 + \beta_2$$

ואם נסתכל על "הפרש ההפרשים"- כלומר על ההפרש שבין קבוצת הטיפול לפני ואחרי הטיפול (שינקה את המאפיינים הקבועים בין אותה הקבוצה לעצמה) ובין אותו ההפרש בקבוצת הביקורת (שינקה את השפעת הזמן), נקבל:

$$E(Y_{11}) - E(Y_{10}) - \left(E(Y_{01}) - E(Y_{00}) \right) = \beta_2 + \beta_3 - (\beta_2) = \beta_3$$

כלומר $\widehat{\beta_3}$ יהיה האומד לאפקט של הטיפול.



דוגמא מפורסמת (נובל!) לשימוש בשיטה:
האם שכר המינימום משפיע על שיעור האבטלה? (תקציר כאן).

2. *Regression Discontinuity*.

נניח שרוצים לבדוק את האפקט של טיפול מסוים, אך הקצאת הטיפול איננה אקראית, אלא נקבעת על פי ערך סף מסויים של משתנה רציף X, שהחל ממנו רוב הפרטים שמעל הסף מקבלים את הטיפול, ומתחת אליו לא מקבלים. הרעיון הוא שהפרטים שמעט מעל הסף ומעט מתחת לסף דומים מאוד במאפיינים שלהם, ורק בגלל השרירותיות של הסף, מטופלים אחרת. כך, קבוצת הטיפול היא אלו שמעט מעל הסף, וקבוצת הביקורת הם אלו שמעט מתחת לסף. כמו כן, מניחים שמשתנה התוצאה הוא פונקציה רציפה של המשתנה X.

המודל (הבסיסי) הוא:

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \beta_2 \cdot 1_{\{i\ is\ above\ cutoff\}}$$

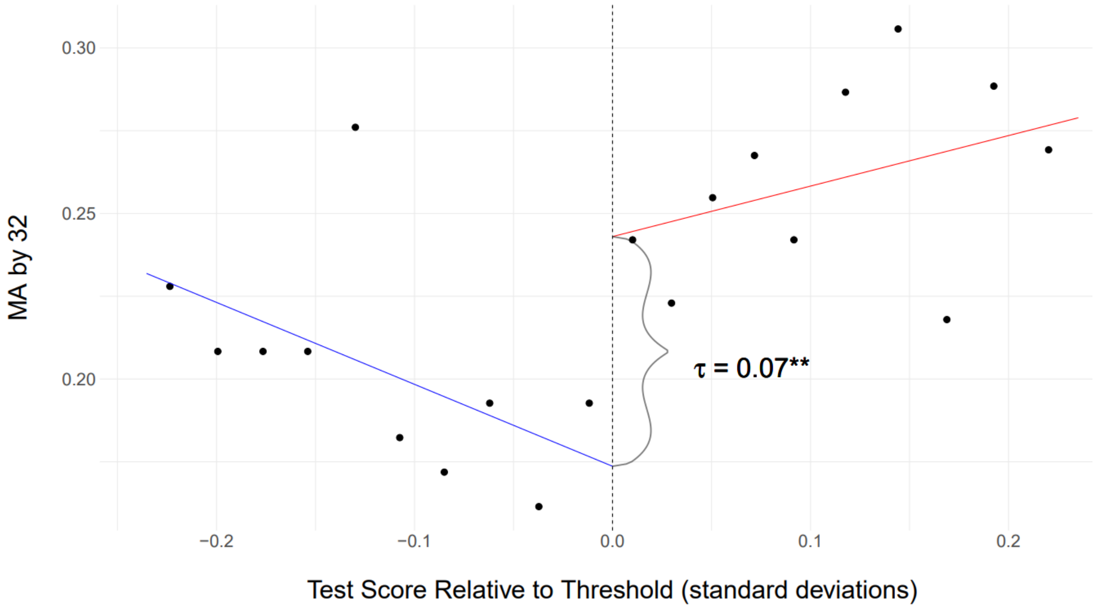
שימו לב שכיוון שהמשתנה Y רציף ב-X, נקבל שאם המקדם $\beta_2 \neq 0$ (באופן מובהק), (ובהינתן שהקבוצות באמת דומות בכל המאפיינים האחרים), נוכל להגיד כי ישנה "אי רציפות" שנובעת מכך שהפרט שייך לקבוצת הטיפול, וזה יהיה האומד לאפקט של הטיפול.

דוגמא:

מה היא ההשפעה של השתתפות בתכניות מחוננים בילדות על השכר/השכלה וכו'?
רבים מהתלמידים בישראל עוברים בכיתה ב'-ג' מבחן סיווג לתכניות מחוננים (מבחן IQ). הסף נקבע כך שבקירוב כ-2.5% מהתלמידים מאותרים כמחוננים וזכאים להצטרף לתכניות/כיתות מחוננים. אלו שלא עברו את הבחינה, לא מורשים להשתתף.
נניח שאנחנו מתעניינים בשכר החזוי כמשתנה תוצאה. כיוון שמראש התלמידים שנחשבים למחוננים צפויים להרוויח משמעותית יותר מאלו שאינם מחוננים, ללא קשר לתכנית אלא מעצם היותם אינטיליגנטיים במיוחד, קשה לבדוק את האפקט של השתתפות בתכנית. כדי להתגבר על כך משתמשים במשתנה הדמי "האם עבר את הבחינה", ואומדים את המודל בו המשתנה המוסבר הוא ההסתברות ללמוד בתואר שני בגיל 32, והמשתנים המסבירים הם ציון ה-IQ ומשתנה הדמי. אם המקדם יהיה מובהק- אז נגיד שהתכנית השפיעה על ההסתברות ללמוד בתואר שני בגיל 32, ואחרת לא.

שאלה:

כיצד הייתם בודקים את ההשערה כי אצל מחוננים הקשר בין ה-IQ ובין ההסתברות להשלים תואר שני שונה מאלו שאינם מחוננים? מה עשוי לקרות אם נתעלם מהבדלים אלו?



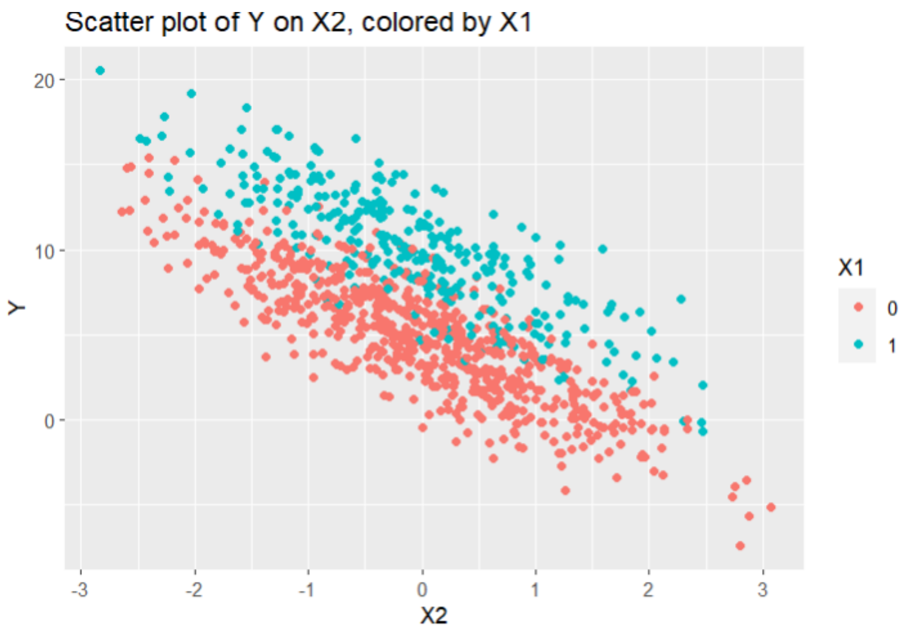
שאלה

לפניכם תרשימי פיזור של Y על X_2, כאשר X_1 הוא משתנה דמי. התאימו לכל תרשים האם יש אינדיקציה להפרש בתוחלות, האם לאינטראקציה?

בזכות השאלה הראשונה בקובץ נוכל לפתור זאת כך :

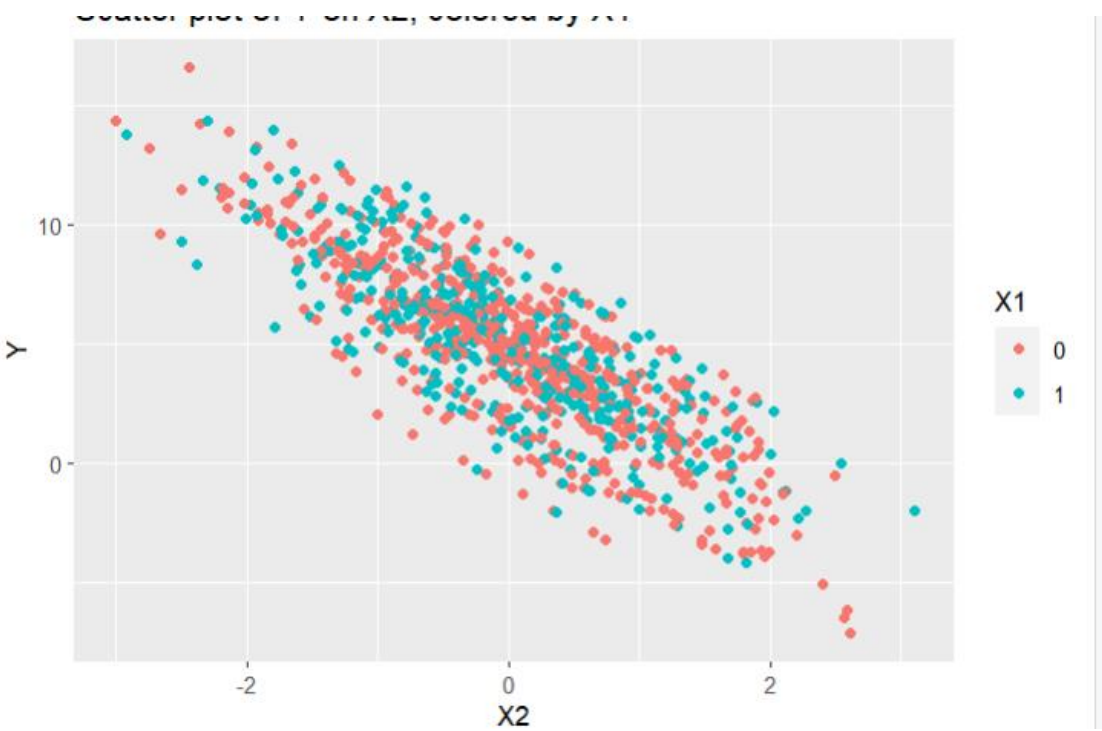
נדמה שני קווי רגרסיה נפרדים- האחד בעבור הקבוצה האדומה והשני בעבור הקבוצה הכחולה. אם נראה שהחותך שונה- נגיד שיש עדות להפרש קבוע בתוחלות. אם נראה שהשיפוע שונה- הרי שזו אינדיקציה לאינטראקציה.

יש עדות לאינטראקציה- השיפועים שונים. יש עדות להפרש תוחלות- שימו לב מה קורה לקו ה"דימיוני" של הקבוצה הכחולה באיזור הנקודה 0. הוא גבוה יותר מזה של הקבוצה האדומה.



אין עדות לאינטראקציה כיוון שהשיפועים נראים זהים. יש עדות להפרש קבוע בתוחלות- נראה שהקו של התצפיות הכחולות מקבל ערך קבוע ב-X הגבוה מזה של התצפיות האדומות.

יש עדות לאינטראקציה, אין עדות להפרש קבוע בתוחלות. סביב 0 הקווים היו נחתכים.



אין עדות לאינטראקציה וגם לא להפרש תוחלות.

בדיקת הנחות המודל הלינארי

תזכורת: תחת המודל הלינארי אנו מניחים:

$$Y = X\beta + \epsilon, \epsilon \sim (0, \sigma^2 I)$$

נחלק זאת ל-3 הנחות נפרדות:

1. לינאריות: $E(Y) = X\beta$. מתקיים אס"ם $E(\epsilon|X) = 0$.
2. שיוויון שונויות: $var(\epsilon_i) = \sigma^2\ \forall_i$
3. חוסר מתאם בין השגיאות: $Var(\epsilon) = diag(\sigma_1^2, \dots, \sigma_n^2)$.

תחת הנחות אלו מובטח כי $\hat{\theta} = a^T\hat{\beta}_{OLS}$ הוא האומד הלינארי חסר ההטיה בעל השונות המינימלית. (*BLUE*)

בשבועיים האחרונים הוספנו גם את הנחת הנורמליות:

$$\epsilon \sim N(0, \sigma^2 I)$$

הוספנו את הנחה זו על מנת לבצע הסקה סטטיסטית. (כמו כן, תחת הנחה זו מובטח גם כי אומד OLS הוא $UMVUE$- האומד חסר ההטיה בעל השונות המינימלית).

היינו רוצים לבדוק את הנחות אלו, אך שימו לב שעבור כולן אנו נדרשים לדעת את התכונות של ϵ, שאינו נצפה. מה עושים? ננצל את תכונת ה**עקיבות** של אומד OLS.

תזכורת/הגדרה: אומד $\hat{\theta}$ יקרא עקיב לפרמטר θ אם מתקיים שלכל $\delta > 0$:

$$\lim_{n\to\infty} P\left(\left|\hat{\theta} - \theta\right| \geq \delta\right) = 0$$

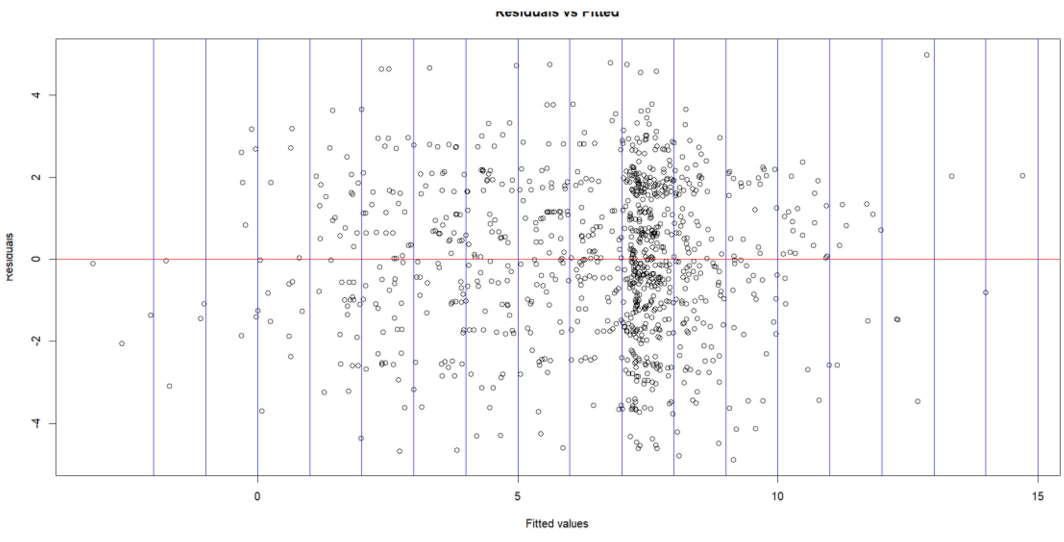
באופן שקול, ניתן להראות שאם אומד הוא חסר הטיה, אז הוא עקיב לפרמטר אס"ם השונות שלו מתכנסת ל-0 כאשר גודל המדגם הולך לאינסוף.

מדוע אומד OLS הוא עקיב? לא נוכיח זאת אך ניתן אינטואיציה- את אומד OLS בעצמו ניתן לראות כ"החלקה" של ממוצעי Y בכל "תא קטן" עם ערך x מסויים. כלומר ה"ממוצע המותנה" של $Y|X$. מ-$WLLN$, הממוצע עקיב לתוחלת, ולכן האומד עקיב לתוחלת המותנה של $Y|X$.

מסיבות אלו, אנו מסיקים כי אם המדגם גדול מספיק, נוכל לבדוק את ההנחות על ידי החלפת התוחלת של Y ב-$\hat{Y}$, ועל ידי החלפת ϵ ב-e, על אף שזה לא מדויק.

בדיקת הנחת הלינארית

שימו לב ש**תמיד** מתקיים כי $E(\bar{e}) = 0$ וזאת ללא קשר לשום הנחה. אך כיוון שמההנחה נובע כי $E(\epsilon|X) = 0$, אז אם היא נכונה מתקיים שגם $E(e|X) = 0$.
מכאן, שכדי לבדוק את ההנחה נרצה ליצור את אותם ה"תאים הקטנים" שהוזכרו לעיל, ולבדוק האם הממוצע, כאומד לתוחלת, של השאריות, הוא 0 בכל תא:



כדי שתוכלו להתאמן על כך בעצמכם :

קוד לייצור פלטים דומים (לא לגמרי זהים) :

הנחת הליניאריות מתקיימת וגם הנחת שיוויון השונויות:

```
X1 <- rbinom(1000, 1, 0.4)
X2 <- rnorm(1000, mean = 0, sd = 1)
X3 <- rexp(1000, 2)
epsilon <- rnorm(1000, 0, sd =2)
Y <- 5 + 2 * X1 - 3 * X2 + 3 * X2 * X1 + X3 + epsilon
```

```
Y <- 5 + 2 * X1 - 3 * X2 + X2^2+ X3^2+epsilon
```

הנחת הליניאריות מתקיימת אך הנחת שיוויון השונויות לא מתקיימת:

```
errors <- rnorm(1000, 0, sd = 1 +  2* abs(X2))
Y <- 5 + 2 * X1 - 3 * X2 + 3 * X2 * X1 + X3 + errors
```

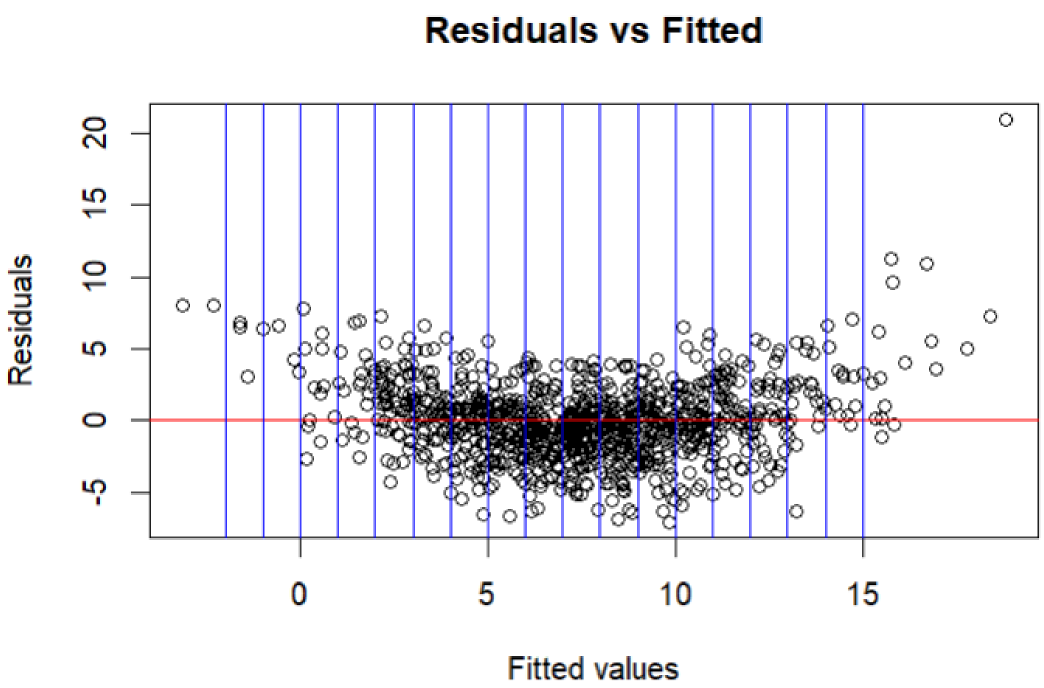
שתי ההנחות לא מתקיימות:

```
Y <- 5 + 2 * X1 - 3 * X2 + X2^2+ X3^2+0.02*exp(X3)+errors
```

אמידת המודל והשרטוט:

```
reg <- lm(Y ~ X1 * X2 + X3)
e <- reg$residuals
f <- reg$fitted.values
plot(f, e, xlab = "Fitted values", ylab = "Residuals", main = "Residuals vs Fitted")
abline(h = 0, col = "red")
```

הפרה של הנחת הליניאריות (ללא הפרת שיוויון שונויות) :



בדיקת הנחת שיוויון השונויות:

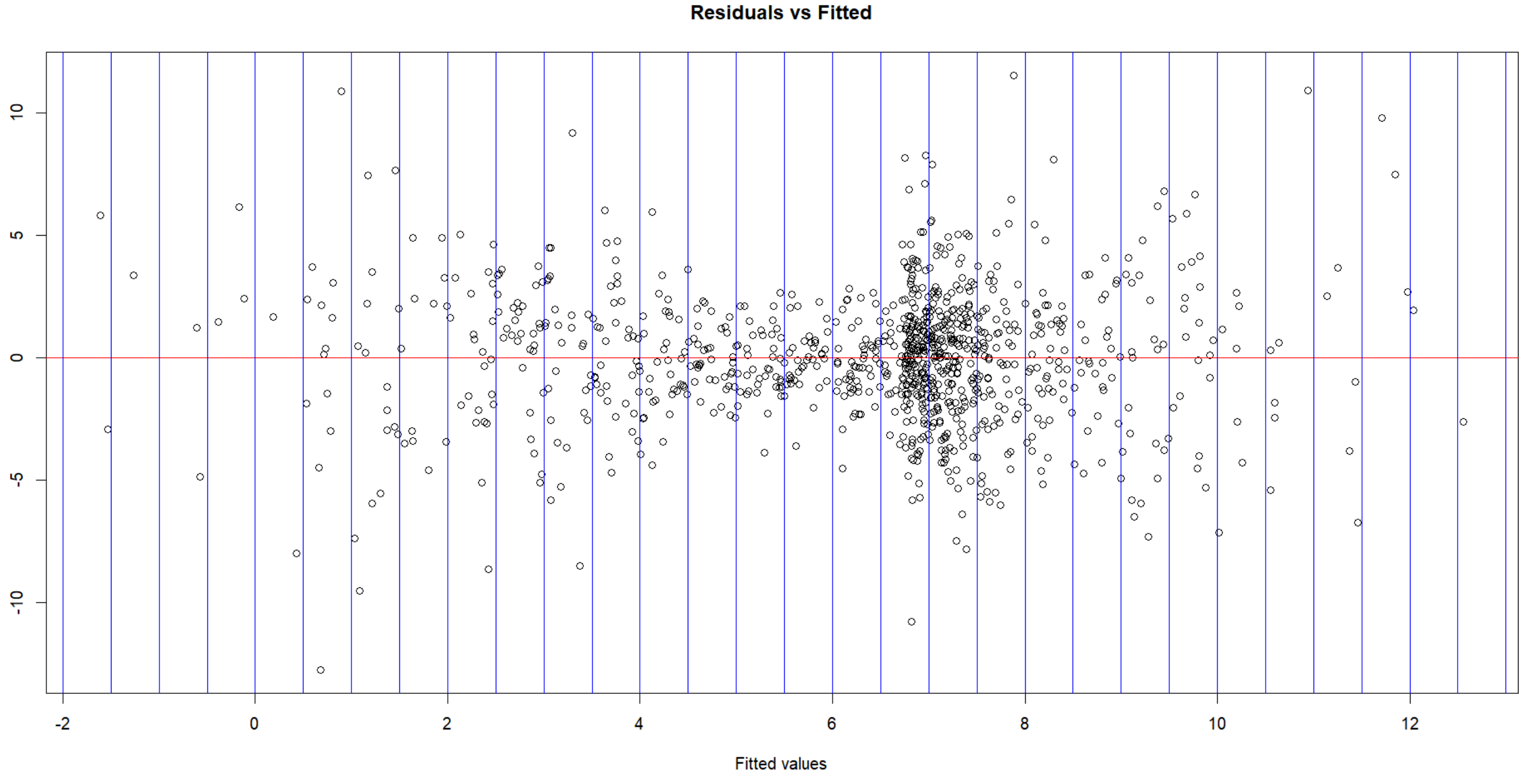
$$Cov(e) = Cov\big((I - P_X)Y\big) = (I - P_X)Var(Y)(I - P_X) = \sigma^2(I - P_X)$$

וכאשר המדגם גדול, נוכל לראות כי:

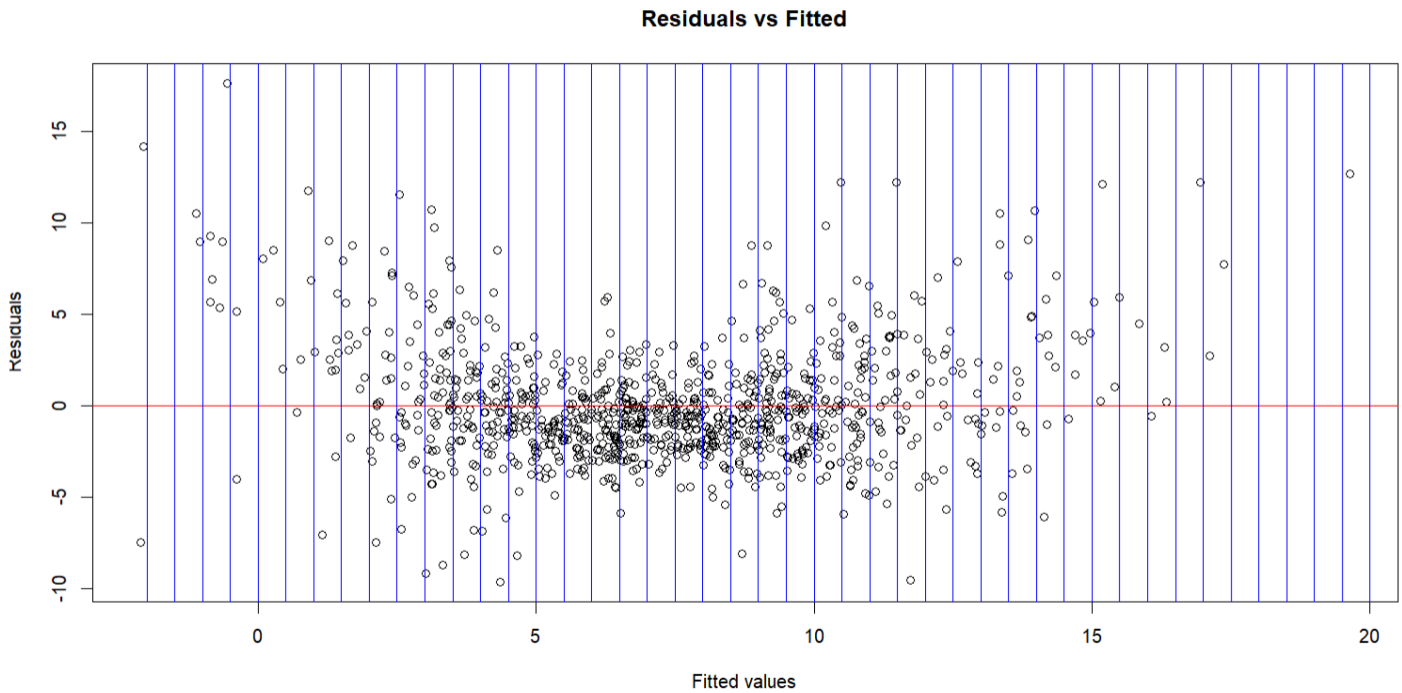
$$I - P_X \to I$$

כלומר שוב נשתמש בקירוב $e \approx \epsilon$ ונבדוק האם מתקיים ש- $Var(e|X) = \sigma^2$ - כלומר קבוע על פני ערכי X.

הפרה של שיוויון השונויות אך ללא הפרת הלינאריות:



הפרה של שתי ההנחות:



<u>בדיקת הנחת הנורמליות, התמודדות עם הפרת הנחות המודל ותצפיות חריגות</u>

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:

$$\hat{Q}_p = X_{\lceil n\cdot p\rceil}$$

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:

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

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

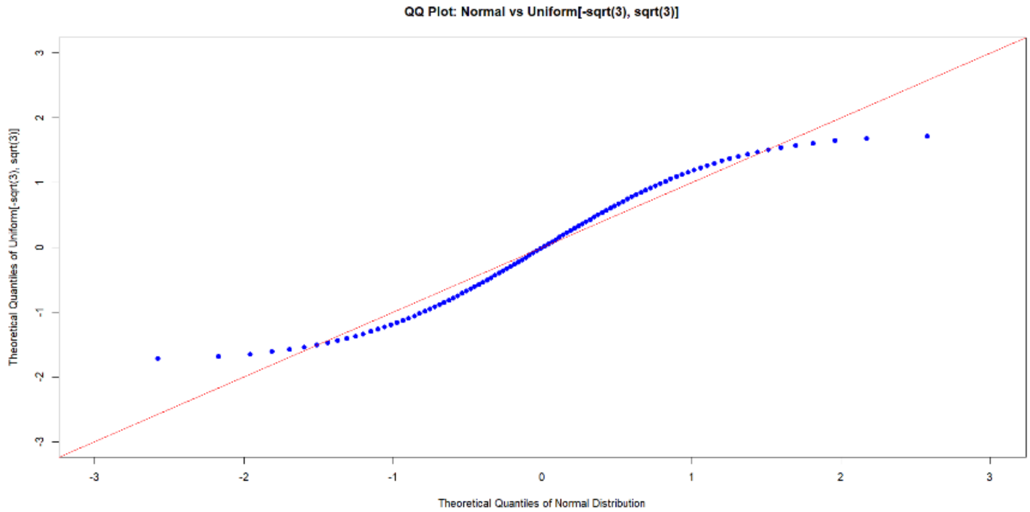
$F^{-1}(p)$ (כמו $qnorm(p)$ למשל.

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

נסמן $F^{-1}(0.025) = q$. אז:

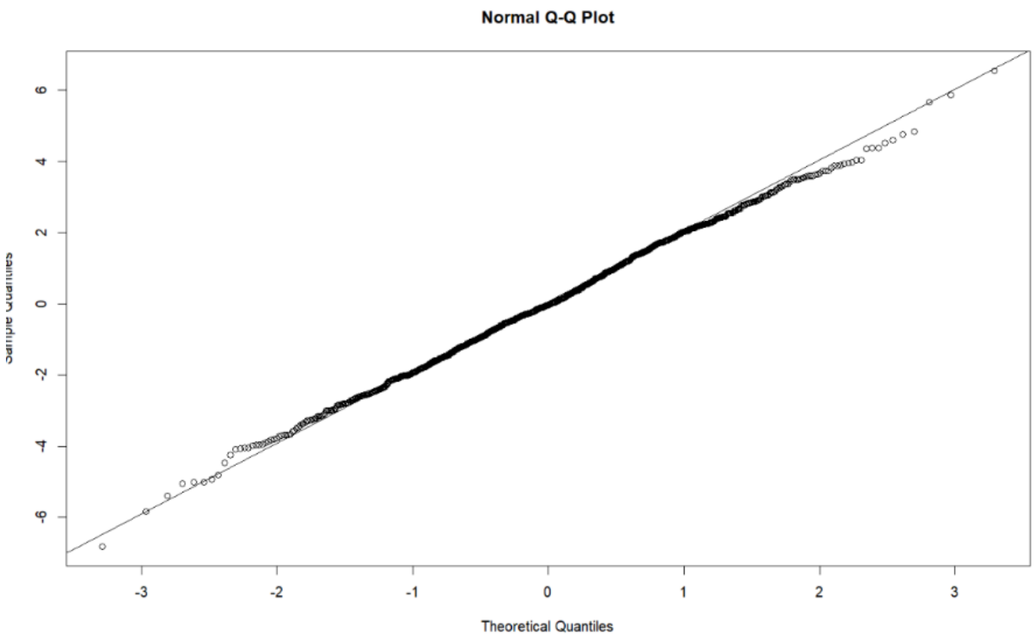
$$0.025 = P(X_1 \le q) = \frac{q+\sqrt{3}}{2\sqrt{3}} \Leftrightarrow q = -1.645448$$

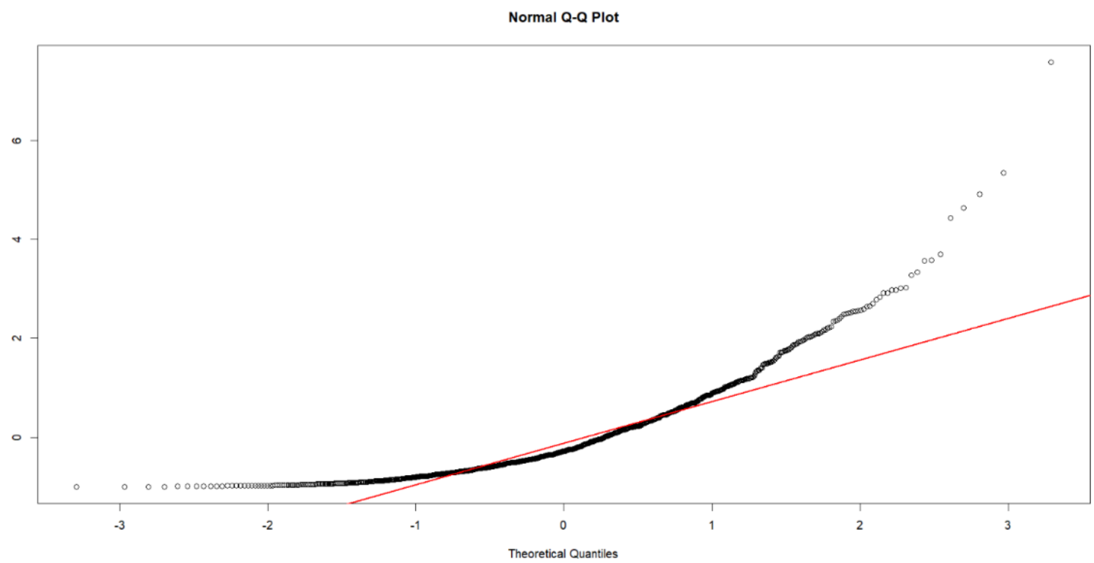
כלומר ה-2.5 אחוזונים הראשונים נצברים מוקדם יותר בהתפלגות הנורמלית מאשר בהתפלגות היוניפורמית וזאת כיוון שהסתברותם של ה"זנבות" בהתפלגות היוניפורמית אפסית, לעומת ההתפלגות הנורמלית שצפיפותה חיובית לכל נקודה בישר הממשי. בהקשר של בדיקת הנחת הנורמליות נצפה לראות שיש לנו "חוסר" בתצפיות בזנבות בהתפלגות היוניפורמית-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

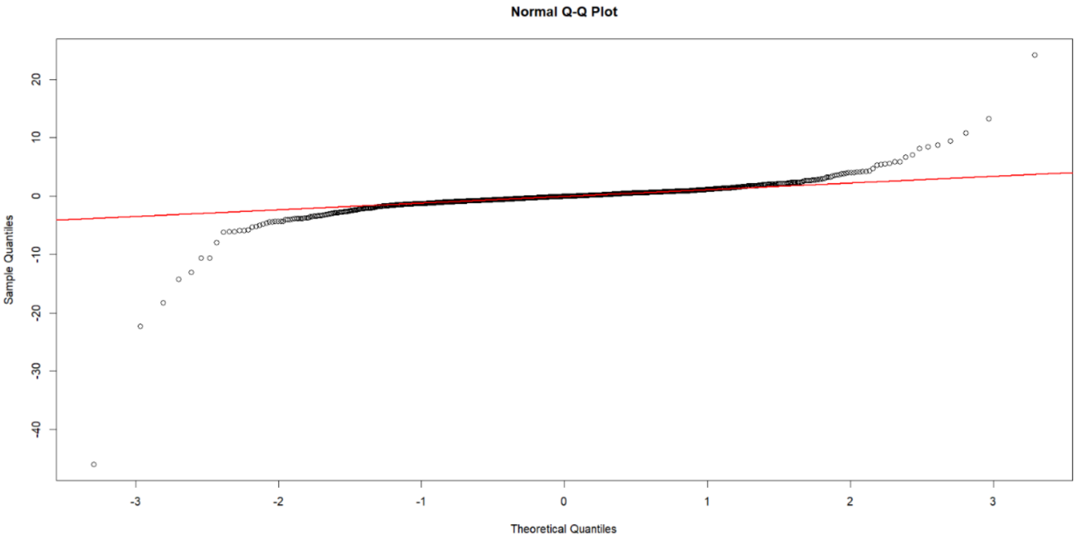


שאלה

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







Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
20
10
0
-10
-20
-30
-40
-3
-2
-1
0
1
2
3

פתרון :