רגרסיה- תרגול 6- תשפ״ה

תזכורת

התפלגויות קשורות להתפלגות הנורמלית ומבוא להסקה סטטיסטית

Definition (Chi-square distribution). If $Z_1, Z_2, \ldots, Z_k \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, then the distribution of

$$Q = \sum_{j=1}^{k} Z_j^2$$

is called the Chi-square distribution with k degrees of freedom, and we denote $Q \sim \chi_k^2$ (in R: `pchisq()`, `qchisq()`, `rchisq()`).

Definition 5 (t-distribution). . If $Z \sim \mathcal{N}(0,1), V \sim \chi_k^2$, are independent random variables, then the distribution of

$$T = \frac{Z}{\sqrt{V/k}}$$

is called the t-distribution with k degrees of freedom, and we denote $T \sim t_k$ (in R: `pt ()`, `qt( )`, `rt( )`).

Definition 6 (F distribution). If $V_1 \sim \chi_{k_1}^2, V_2 \sim \chi_{k_2}^2$, are independent random variables, the distribution of

$$F = \frac{V_1/k_1}{V_2/k_2}$$

is called the F-*distribution with* k_1 *and* k_2 (numerator and denominator, respectively) *degrees of freedom*, and we denote $F \sim \mathrm{F}_{k_1,k_2}$.

שאלה- המשך מהתרגול הקודם

ו. אפיינו את ההתפלגות של

$$\frac{\left|\left|\hat{Y} - X\beta\right|\right|^2}{\left|\left|e\right|\right|^2}$$

ז. הסיקו מה ההתפלגות של הסטטיסטי $\frac{R^2}{1-R^2} = \frac{\frac{SSR}{SST}}{\frac{SSE}{SST}} = \frac{SSR}{SSE}$ במקרה שבו $Y = 1_n\mu + \epsilon$ כאשר $\epsilon \sim N(0, \sigma^2 I)$.

ז. נסמן ב-$s_j = (X^TX)^{-1}_{jj}$. אפיינו את ההתפלגות של $(\hat{\beta}_j - \beta_j)/\sqrt{(\widehat{\sigma^2} s_j)}$.

פתרון:

ו.

$$\hat{Y} - X\beta = P_X Y - X\beta = P_X X\beta + P_X\epsilon - X\beta = X\beta + P_X\epsilon - X\beta = P_X\epsilon$$

אזי

$$\left|\left|\hat{Y} - X\beta\right|\right|^2 = ||P_X Y - X\beta||^2 = \left|\left|P_X\epsilon\right|\right|^2$$

תוך שימוש בסעיף ב׳, זה מתפלג $\sigma^2\chi_{p+1}^2$.

כמו כן, הוכחנו כבר בשיעור (ושוב- תוך שימוש בסעיף ב׳) כי $\left|\left|e\right|\right|^2 \sim \sigma^2\chi_{n-p-1}^2$. לכן הסטטיסטי שבשאלה שקול ל:

$$T := \frac{\frac{\left|\left|\hat{Y} - X\beta\right|\right|^2}{\sigma^2}}{\frac{\left|\left|e\right|\right|^2}{\sigma^2}} = \frac{V_1}{V_2}$$

עבור $V_1 \sim \sigma^2\chi^2_{p+1}$ ואילו $V_2 \sim \sigma^2\chi^2_{n-p-1}$ ושני המשתנים בלתי תלויים (מהסעיף הקודם). לאחר חלוקה של המונה ב- $p+1$ ושל המכנה ב-$n-p-1$ נקבל בדיוק את ההגדרה של משתנה מקרי המפולג $F_{p+1,n-p-1}$. כלומר: $\frac{n-p-1}{p+1}T \sim F_{p+1,n-p-1} \Leftrightarrow T \sim \frac{p+1}{n-p-1}F_{p+1,n-p-1}$.

כעת:

$$\frac{R^2}{1-R^2} = \frac{SSR}{SSE} = \frac{\left|\left|\hat{Y} - 1_n\bar{Y}\right|\right|^2}{\left|\left|e\right|\right|^2} = \frac{\left|\left|\hat{Y} - 1_n\bar{Y}\right|\right|^2}{\left|\left|e\right|\right|^2} = \frac{\left|\left|P_{X\cap 1_n^\perp}Y\right|\right|^2}{\left|\left|e\right|\right|^2} = \frac{\left|\left|P_{X\cap 1_n^\perp}(1_n\mu + \epsilon)\right|\right|^2}{\left|\left|e\right|\right|^2} = \frac{\left|\left|P_{X\cap 1_n^\perp}\epsilon\right|\right|^2}{\left|\left|e\right|\right|^2}$$

$$= \frac{V_1}{V_2}, \qquad V_1 \sim \sigma^2\chi^2_p, \qquad V_2 \sim \sigma^2\chi^2_{n-p-1}$$

$$\Rightarrow \frac{R^2}{1-R^2} \sim \frac{p}{n-p-1}F_{p,n-p-1} \Rightarrow \boldsymbol{F_{STAT} = \frac{n-p-1}{p}\cdot\frac{R^2}{1-R^2}}$$

ז. תחת המודל הנורמלי מתקיים:

$$\hat{\beta}_j \sim N\left(\beta, \sigma^2 s_j\right)$$
$$(\hat{\beta}_j - \beta_j)/\sqrt{(\widehat{\sigma^2}s_j)} = [(\hat{\beta}_j - \beta_j)/\sigma\cdot\sqrt{s}_j]/\sqrt{(\widehat{\sigma^2}\,]/\sigma}$$

המונה מפולג נורמלי סטנדרטי בעוד שבמכנה ישנו שורש על פני $\frac{\left|\left|e\right|\right|^2}{\sigma^2\cdot n-p-1} = \widehat{\sigma^2}\,/\sigma^2 = \frac{\chi^2_{n-p-1}}{n-p-1}$ כלומר חי בריבוע המחולק בדרגות החופש שלו. מכאן שההתפלגות היא התפלגות t_{n-p-1}.

שאלה

א. מצאו את ההתפלגות של $\hat{\beta}_j$ ובנו רווח סמך ל-β_j כאשר σ^2 ידוע. חזרו על כך כאשר σ^2 איננו ידוע.

פתרון:

א. ע"פ השאלה הקודמת, ההתפלגות נורמלית עם הפרמטרים שהוזכרו. מכאן שרווח סמך ל-β_j:

$$\hat{\beta}_j \pm Z_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{(X^TX)_{jj}}$$

וכאשר σ^2 איננו ידוע:

$$\hat{\beta}_j \pm t_{n-p-1,1-\frac{\alpha}{2}}\cdot\frac{\hat{\sigma}}{(X^TX)_{jj}}$$

שאלה

```
Call:
lm(formula = TotalMurderRate ~ Population + Density + Ownership,
    data = guns)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2783 -1.3871 -0.3493  0.9758  5.9019

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.84193    1.02830   1.791   0.0797 .
Population   0.09146          A   2.238   0.0300 *
Density      1.91414    0.20841   9.184 4.63e-12 ***
Ownership          B    2.23611   1.258   0.2146
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.816 on 47 degrees of freedom
Multiple R-squared:  0.6835,    Adjusted R-squared:  C
F-statistic: 33.83 on D and 47 DF,  p-value: 8.432e-12
```

	intercept	Population	Density	Ownership
intercept	0.3205745	-0.0073787	-0.0360011	-0.6548395
Population	-0.0073787	0.0005063	0.0006838	0.0109158
Density	-0.0360011	0.0006838	0.0131685	0.0721368
Ownership	-0.6548395	0.0109158	0.0721368	1.5159148

3. **(10 נק')** נניח, היפותטית, שהיתה מדינה 52 שלא כלולה בקובץ הנתונים המקורי, ועבורה המשתנים המסבירים מקבלים את הערכים: Population = 9.53, density = 0.2, ownership = 0.63. מצאו אומד ליניארי חסר-הטייה בעל שונות מינימלית עבור תוחלת שיעור מקרי הרצח במדינה עם המאפיינים האלה (שהמודל הליניארי שמתאר את הקשר בין המשתנים עבור קובץ הנתונים המקוריים, מתאים גם לתצפית החדשה).
4. **(10 נק')** בנו ר"ס ברמת ביטחון 90% עבור הפרמטר שנאמד בסעיף הקודם. יש לציין אילו הנחות נדרשות על השגיאות ϵ_i כדי שרווח-הסמך אכן יהיה תקף.

5. ברמת מובהקות של 5%, בדקו את ההשערה כי ההשפעה של גודל האוכלוסיה על שיעור מקרי הרצח זהה לזו של צפיפות האוכלוסיה.

4. לפי פירוק לינארי כללי, $\theta = a^T\beta$, כאשר הנבדל הוא $\hat{\theta} = a^T\hat{\beta}$,
והשונות: $Var(\hat{\theta}) = Var(a^T\hat{\beta}) = \sigma^2 a^T[(X^TX)^{-1}]a$.
במקרה שלנו $a = (1, 9.53, 0.2, 0.63)^T$, והמטריצה $(X^TX)^{-1}$ נתונה.
נחשב:

$$a^T[(X^TX)^{-1}]a = \sum_{i=1}^{4}\sum_{j=1}^{4}[(X^TX)^{-1}]_{ij}\, a_i a_j =$$

$$= 0.32 - .0073 \cdot 9.53 - \overset{0.036}{.36} \cdot 0.2 - .654 \cdot .63 + \qquad (i=1)$$

$$+ 9.53\left(-.0073 + .0005 \cdot 9.53 + .0006 \cdot 0.2 + .011 \cdot 0.63\right) + \qquad (i=2)$$

$$+ 0.2\left(-\overset{0.036}{.36} + .0006 \cdot 9.53 + .0131 \cdot 0.2 + 0.72 \cdot 0.63\right) + \qquad (i=3)$$

$$+ 0.63\left(-.654 + .011 \cdot 9.53 + .072 \cdot 0.2 + 1.516 \cdot 0.63\right) \qquad (i=4)$$

$$= \boxed{0.14}$$

גיליון:

	1	9.53	0.2	0.63
1	0.32	-.0073	-.036	-.654
9.53	-.0073	.0005	.0006	$\overset{+}{-}$.011
0.2	-.036	.0006	.0131	.072
0.63	-.654	.011	.072	1.516

לדוגמא:

$$CI = \hat{\theta} \pm \hat{\sigma}\sqrt{a^T (X^T X)^{-1} a} \cdot t_{n-p-1;\, 1-\alpha/2}$$

$$= 4.865 \pm 1.816 * \sqrt{0.14} \cdot 1.678$$

$$= 4.865 \pm 1.14 = \boxed{(3.72, 6)}$$

הרווח בקצה התחתון עלול שלילי ב-ϵ (בגלל הפרעות בדגימה בלינארי הכללי: לינארי; שיטה שלא-חוקית ומחרבת לשקר בינתן).

משתנים קטגוריאליים ומשתני דמי, ובדיקת הנחות המודל הלינארי

שאלה

א. מה היא הפרשנות של האומד ל-β_1 במודל:

$$Y_i = \beta_0 + \beta_1 \cdot 1_{\{i \in A\}} + \beta_2 X_{1_i} + \epsilon_i$$

פתרון:

תוך שימוש בסימונים שהגדרנו בעבר, נסתכל על התוחלת של $\hat{\theta} := a^T \hat{\beta}_{OLS}$ ונקבל:

$$E(\widehat{\theta}) = \beta_0 + \beta_1 \cdot 1_{\{i \in A\}} + \beta_2 X_{1_i}$$

נפריד למקרים ונקבל:

$$E(\widehat{\theta} | i \in A) - E(\hat{\theta} | i \in B) = \beta_0 + \beta_1 + \beta_2 X_{1_i} - \beta_0 - \beta_2 X_{1_i} = \beta_1$$

זהו האומד להפרש בין החותך של פרטים מקבוצה A ופרטים מקבוצה B- בעבור כל רמה של X. כלומר הפרש התוחלות המותנות של הפרטים, הקבוע על פני כל ערכי X.

שאלה:
במבחן ברגרסיה נבחנים 90 סטודנטים. נניח שהציון במבחן (Y) הוא קומבינציה לינארית של זמן ההרצאות המוקלטות שראיתם (R), כמות השאלות שפתרתם (בעצמכם!) (Q) ומספר השעות שישנתם בלילה (H) שלפני הבחינה.
כמו כן, נניח שלצרכי מחקר ובאופן אקראי לחלוטין, בתחילת המבחן 30 סטודנטים קיבלו מהמתרגל שוקולד מריר (D), 20 קיבלו שוקולד חלב (M) ולשאר- נאמר "בהצלחה" (G).

א. כתבו במפורש את המודל הבודק את הקשר בין כל המשתנים לעיל ובין ציון הבחינה. בפרט, כתבו את המטריצה $X \in R^{n \times p+1}$.

ב. הציעו למתרגל מבחן סטטיסטי הבודק את ההשערה כי לשוקולד אין השפעה על ציון המבחן **מעבר ל"בהצלחה"**[1] כנגד האלטרנטיבה כי השוקולד אכן משפיע על כמות הנקודות לסטודנטים בעבור **כל רמת הכנה לבחינה וכל זמן שינה**.

ג. הציעו למתרגל מבחן סטטיסטי הבודק את ההשערה כי ההשפעה של שוקולד מריר ושל שוקולד חלב זהה (בעבור כל רמה של המשתנים האחרים).

פתרון:

א. נשתמש תחילה בקבוצה השלישית כקבוצת ה"ביקורת". באופן שקול היינו יכולים להשתמש בכל אחת מהקבוצות האחרות, אבל אז יש לשים לב לכך שפרשנות המקדמים הייתה משתנה:

$$Y_i = \beta_0 + \beta_1 \cdot R_i + \beta_2 \cdot Q_i + \beta_3 \cdot H_i + \beta_4 \cdot 1_{\{i \in D\}} + \beta_5 \cdot 1_{\{i \in M\}} + \epsilon_i$$

במקרה כזה הפרשנות של β_4 לדוגמא, היא ההפרש הקבוע בתוחלת הציון בעבור על רמה של המשתנים האחרים, בין הסטודנטים שקיבלו שוקולד מריר ובין הסטודנטים שקיבלו "בהצלחה". באופן שקול, יכולנו להסתכל על המודל:

$$Y_i = \gamma_0 + \gamma_1 \cdot R_i + \gamma_2 \cdot Q_i + \gamma_3 \cdot H_i + \gamma_4 \cdot 1_{\{i \in D\}} + \gamma_5 \cdot 1_{\{i \in G\}} + \epsilon_i$$

במקרה כזה המשמעות של האומדים $\beta_0, \beta_4, \beta_5$ הייתה משתנה, שכן כעת מדובר בהפרשים למול קבוצת הביקורת החדשה- הסטודנטים שקיבלו שוקולד חלב:

$$E_1(Y_i | R_i, Q_i, H_i, i \in M) = \beta_0 + \beta_1 \cdot R_i + \beta_2 \cdot Q_i + \beta_3 \cdot H_i + \beta_5$$
$$= E_2(Y_i | R_i, Q_i, H_i, i \in M) = \gamma_0 + \gamma_1 \cdot R_i + \gamma_2 \cdot Q_i + \gamma_3 \cdot H_i \Rightarrow \boldsymbol{\gamma_0 = \beta_0 + \beta_5}$$

באופן דומה:

---

[1] שימו לב שלא ניתן לבדוק את ההשערה האם אין אפקט לאף אחד מה"טיפולים". שכן אין במדגם מישהו שלא "טופל".

$$\gamma_0 + \gamma_4 = \beta_0 + \beta_4 \Rightarrow \boldsymbol{\gamma_4 = \beta_4 - \beta_5}$$

$$\gamma_0 + \gamma_5 = \beta_0 \Rightarrow \boldsymbol{\gamma_5 = -\beta_5}$$

כדי לבנות את המטריצה X נניח לצורך נוחות הכתיבה שיש רק 10 סטודנטים, והם מסודרים לפי הקבוצות:

$$\begin{matrix} 1 & R_1 & Q_1 & H_1 & 1 & 0 \\ 1 & R_2 & Q_2 & H_2 & 1 & 0 \\ 1 & R_3 & Q_3 & H_3 & 1 & 0 \\ 1 & R_4 & Q_4 & H_4 & 0 & 1 \\ 1 & R_5 & Q_5 & H_5 & 0 & 1 \\ 1 & R_6 & Q_6 & H_6 & 0 & 0 \\ 1 & R_7 & Q_7 & H_7 & 0 & 0 \\ 1 & R_8 & Q_8 & H_8 & 0 & 0 \\ 1 & R_9 & Q_9 & H_9 & 0 & 0 \\ 1 & R_{10} & Q_{10} & H_{10} & 0 & 0 \end{matrix}$$

שימו לב שלא נכניס את משתנה הדמי בעבור הקבוצה השלישית שכן במקרה כזה נקבל כי עמודת האחדות תהיה תלויה ב-3 העמודות של משתני הדמי ונקבל כי המטריצה $X^T X$ לא תהיה הפיכה.

ב. ההשערה המתאימה (במודל הראשון) היא:

$$H_0: \beta_4 = \beta_5 = 0$$
$$H_1: O.W$$

כדי לבדוק את ההשערה, נאמוד שני מודלים. האחד הוא המודל המאולץ (R), כלומר ללא המשתנים שמתאפסים תחת H_0, והשני הוא המודל המלא (U). באופן דומה להוכחה בשאלה הקודמת, ניתן להראות שבמקרה הכללי, לסטטיסטי תחת השערת ה-0:

$$\frac{R_U^2 - R_R^2}{1 - R_U^2} \cdot \frac{n-p-1}{q} \sim F_{q,n-p-1}$$

כאשר q הוא מספר האילוצים (תספרו סימני שיוויון) ונדחה את ההשערה אם מתקיים ש:

$$P\left(F_{q,n-p-1} \geq f\right) \leq \alpha$$

(כאן f הוא הערך שנצפה במדגם).

שאלה (משתנה אינטראקציה):

נניח שבידינו נתונים על שתי קבוצות, ומשתנה תוצאה Y.

א. הראו שהאמידות הבאות שקולות ובטאו כל אחד מהמקדמים:

1. אמידת שני מודלים נפרדים:

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

2. אמידת המודל המשותף:

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

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

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

ג. נניח שהרצנו את מודל מספר 2 על הנתונים מהשאלה הקודמת, אך כעת רק עם שעות שינה וסוג שוקולד אחד. הסבירו כיצד הייתם בודקים את ההשערה כי לשוקולד אפקט מעורר- כלומר אצל מי שקיבל שוקולד, השפעת שעות השינה על ציון המבחן נמוכה יותר. הסבירו כיצד הייתם בודקים את ההשערה כי לשוקולד אין כלל השפעה על ציון המבחן.

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:

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

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:

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

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.

ג. זה שקול לבדיקת ההשערה כי אין אינטראקציה בין השוקולד ובין שעות השינה. כלומר:

$$H_0: \gamma_3 = 0$$
$$H_1: O.W$$

ובדיקת ההשערות תהיה בדיקת השערות דו צדדית רגילה, תוך שימוש בהתפלגות t_{n-p-1}.

כדי לבדוק האם לשוקולד אין כלל השפעה, ההשערה היא:

$$H_0: \gamma_1 = \gamma_3 = 0$$
$$H_1: O.W$$

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 תקופת 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?

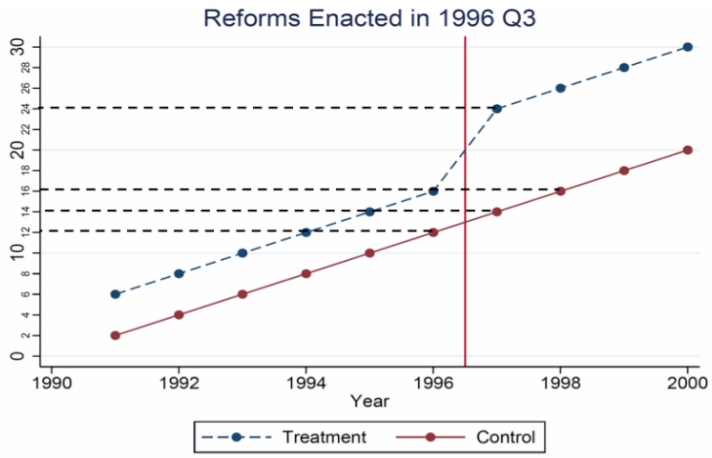
נסתכל על התוחלות של כל אחת מהקבוצות לפני ואחרי הטיפול:

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

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).

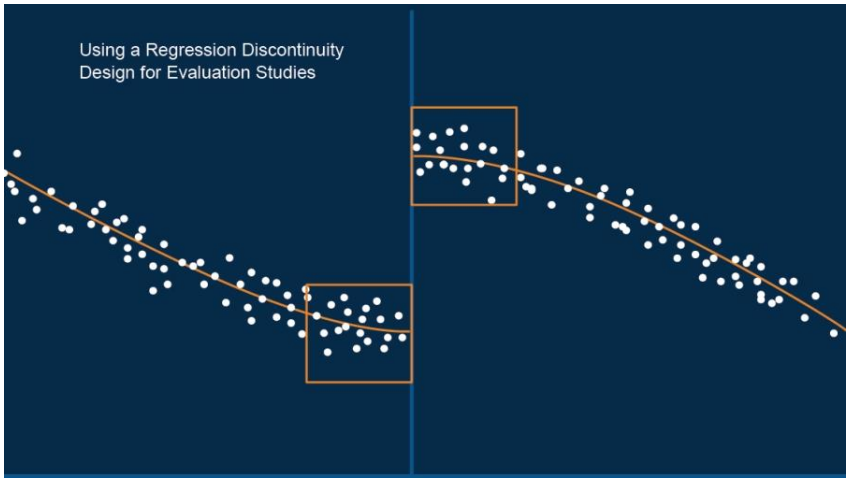
2. *Regression Discontinuity*.

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 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:

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

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.

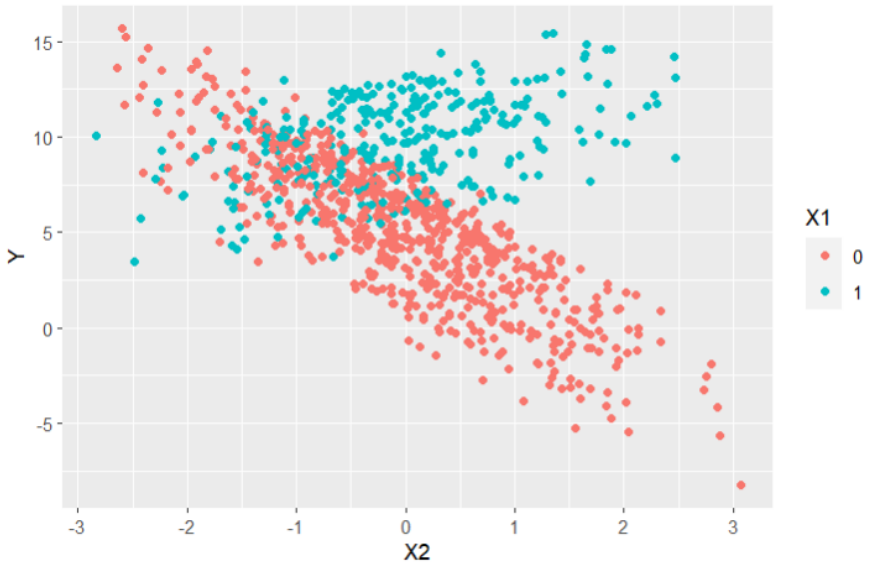


דוגמא:

מה היא ההשפעה של השתתפות בתכניות מחוננים בילדות על השכר/השכלה וכו'?
רבים מהתלמידים בישראל עוברים בכיתה ב'-ג' מבחן סיווג לתכניות מחוננים. 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 ההכנסה בגיל 30 (לדוגמא) והמשתנים המסבירים הם הציון בבחינה ומשתנה הדמי. אם המקדם יהיה מובהק- אז נגיד שהתכנית השפיעה על השכר בגיל 30 (לפחות בקרבת הסף), ואחרת לא.

שאלה

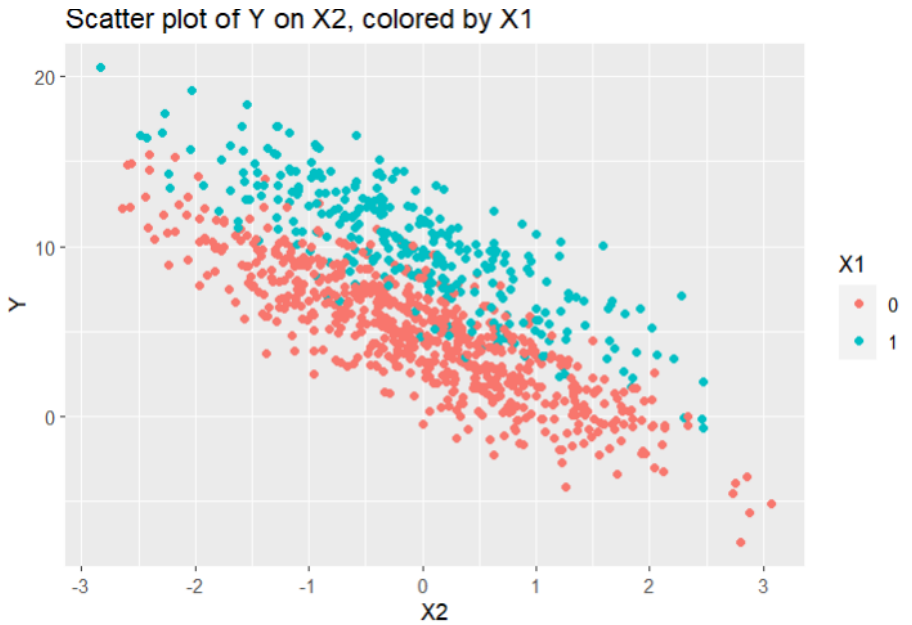
לפניכם תרשימי פיזור של Y על X_2, כאשר X_1 הוא משתנה דמי. התאימו לכל תרשים האם יש אינדיקציה להפרש בתוחלות, האם לאינטראקציה?



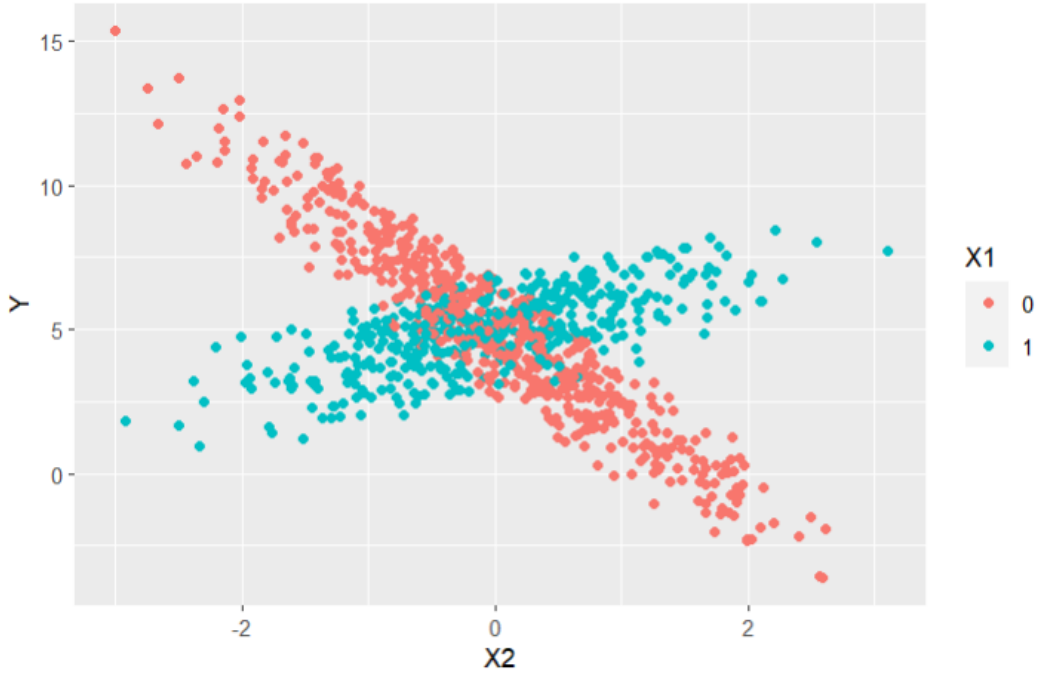
בזכות השאלה השניה בקובץ נוכל לפתור זאת כך :

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.

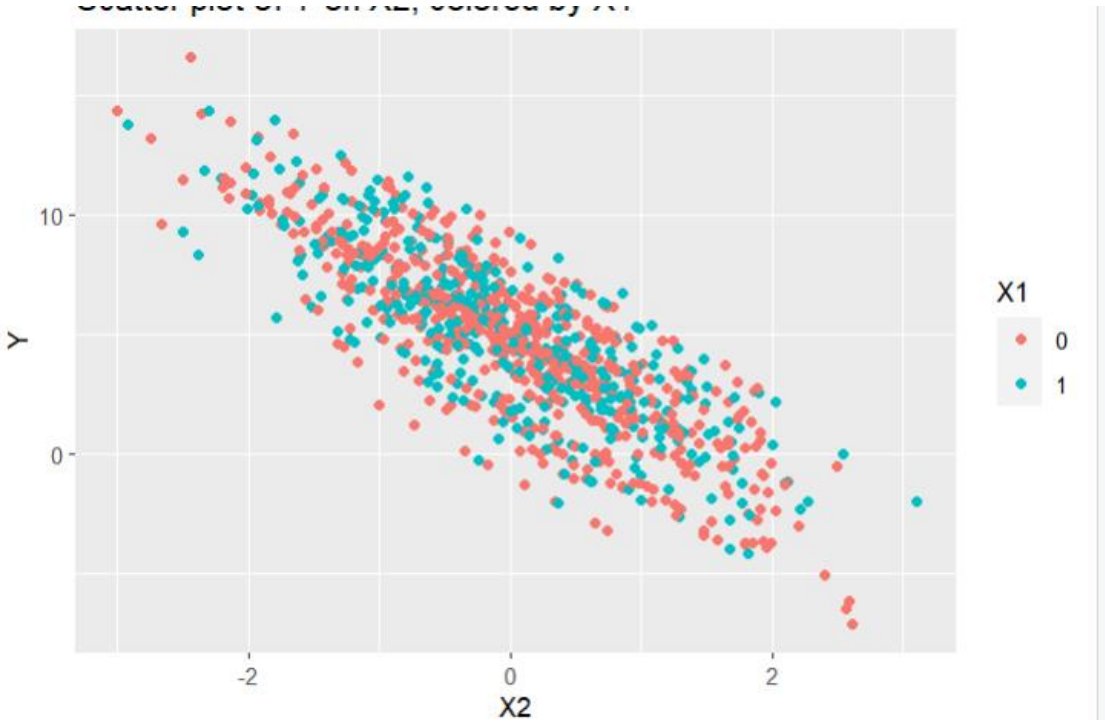
יש עדות לאינטראקציה- השיפועים שונים. יש עדות להפרש תוחלות- שימו לב מה קורה לקו ה"דימיוני" של הקבוצה הכחולה באיזור הנקודה 0. הוא גבוה יותר מזה של הקבוצה האדומה.



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.